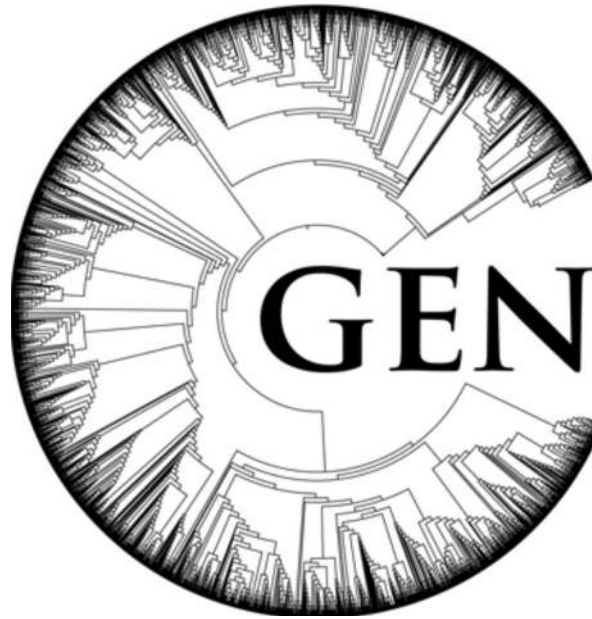


(A VERY SHORT INTRO TO)

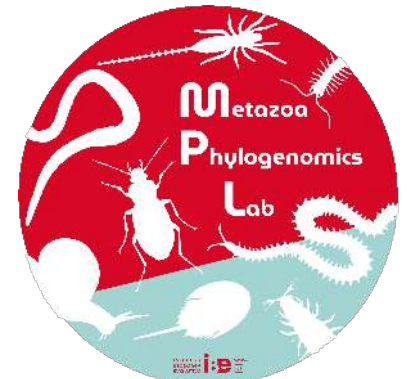
PHYLO GENOMICS



Rosa Fernández

Institute of Evolutionary Biology (CSIC-UPF)

rosa.fernandez@ibe.upf-csic.es



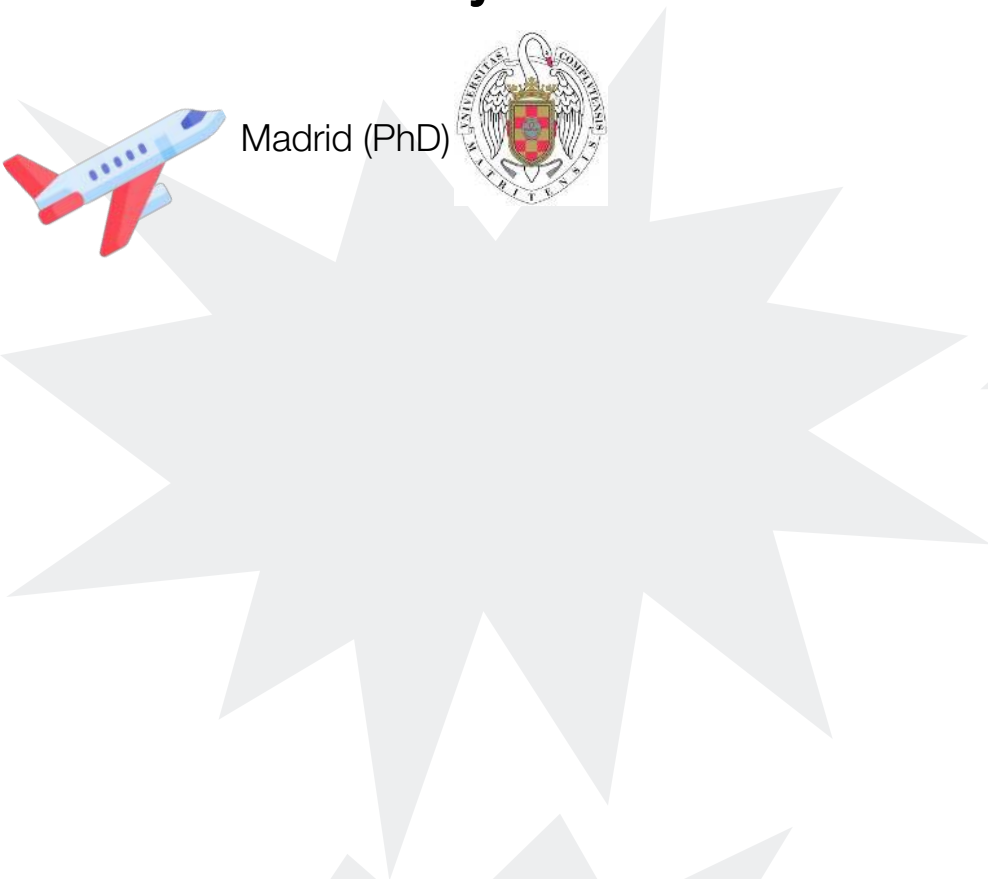
A little bit about myself



A little bit about myself



Madrid (PhD)



A little bit about myself



Madrid (PhD)



Boston (1st postdoc)



HARVARD
UNIVERSITY

A little bit about myself



Madrid (PhD)



Boston (1st postdoc)



Barcelona (2nd postdoc & my lab)



A little bit about myself



Madrid (PhD)



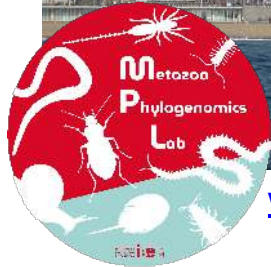
Boston (1st postdoc)



HARVARD
UNIVERSITY



Barcelona (2nd postdoc & my lab)



www.metazomics.com



@Rosamygale
@metazomics



A little bit about myself



Madrid (PhD)



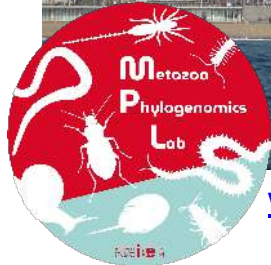
Boston (1st postdoc)



HARVARD
UNIVERSITY



Barcelona (2nd postdoc & my lab)



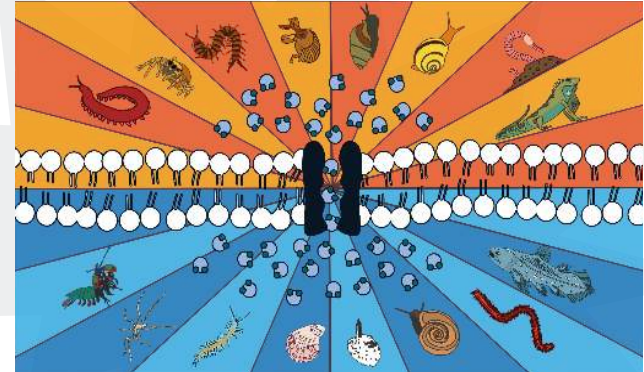
www.metazomics.com



@Rosamygale
@metazomics



Main lines of research:



A little bit about myself



Madrid (PhD)



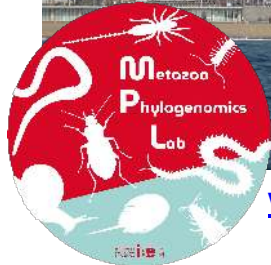
Boston (1st postdoc)



HARVARD
UNIVERSITY



Barcelona (2nd postdoc & my lab)



www.metazomics.com



@Rosamygale
@metazomics



Main lines of research:



A little bit about myself



Madrid (PhD)



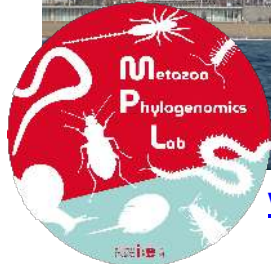
Boston (1st postdoc)



HARVARD
UNIVERSITY



Barcelona (2nd postdoc & my lab)



www.metazomics.com



@Rosamygale

@metazomics



Main lines of research:



Fun Facts:

I'm a zoologist by training, I did not jump into the world of genomics until I was a postdoc

A little bit about myself



Madrid (PhD)



Boston (1st postdoc)



HARVARD
UNIVERSITY



Barcelona (2nd postdoc & my lab)

INSTITUT de BIOLOGIA EVOLUTIVA **ibe** CSIC upf

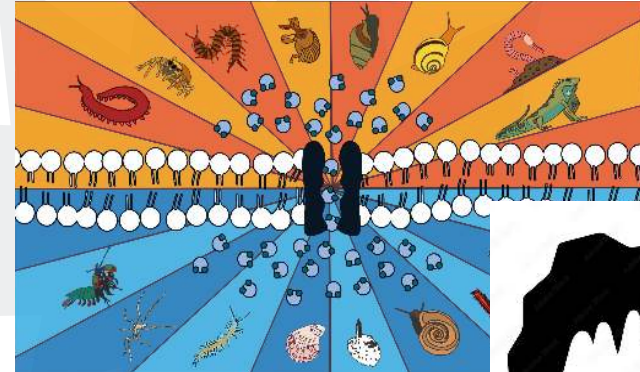


**Metazoa
Phylogenomics
Lab**

www.metazomics.com

 @Rosamygale
@metazomics

Main lines of research:



Fun Facts:

I'm a zoologist by training, I did not jump into the world of genomics until I was a postdoc



I did my PhD on earthworms

A little bit about myself



Madrid (PhD)



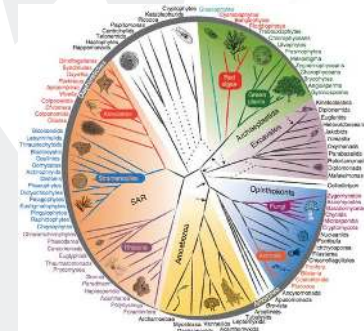
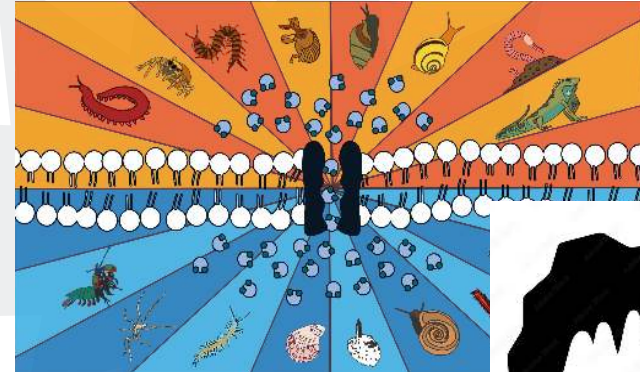
Boston (1st postdoc)



Barcelona (2nd postdoc & my lab)

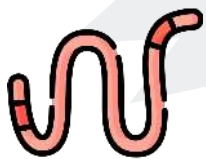


Main lines of research:



Fun Facts:

I'm a zoologist by training, I did not jump into the world of genomics until I was a postdoc



I did my PhD on earthworms

Also organizing the Workshop on Phylogenomics

A little bit about myself



Madrid (PhD)



Boston (1st postdoc)



Barcelona (2nd postdoc & my lab)



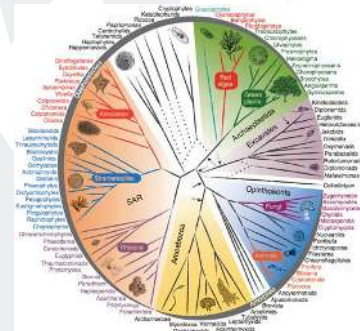
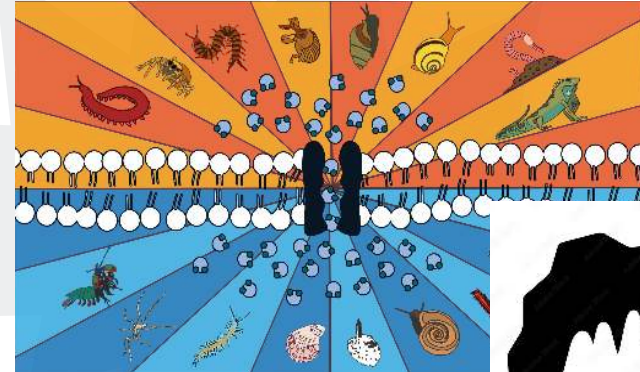
Fun Facts:

I'm a zoologist by training, I did not jump into the world of genomics until I was a postdoc



I did my PhD on earthworms

Main lines of research:



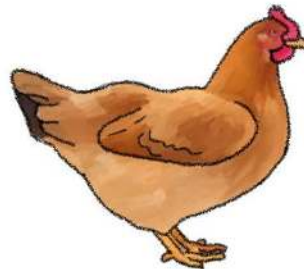
Also organizing the Workshop on Phylogenomics

SPOILER ALERT: there will be BEARS!!



Which came first, the chicken or the egg?

Which came first, the chicken or the egg?

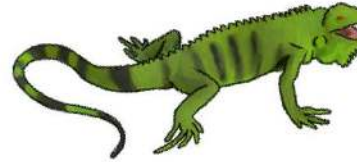


Birds
(Chickens)

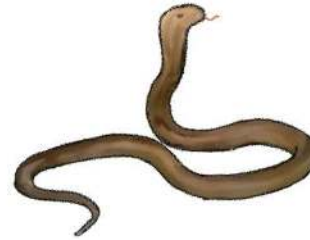
Which came first, the chicken or the egg?



Turtles



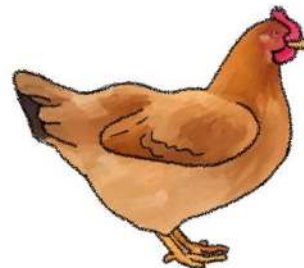
Lizards



Snakes

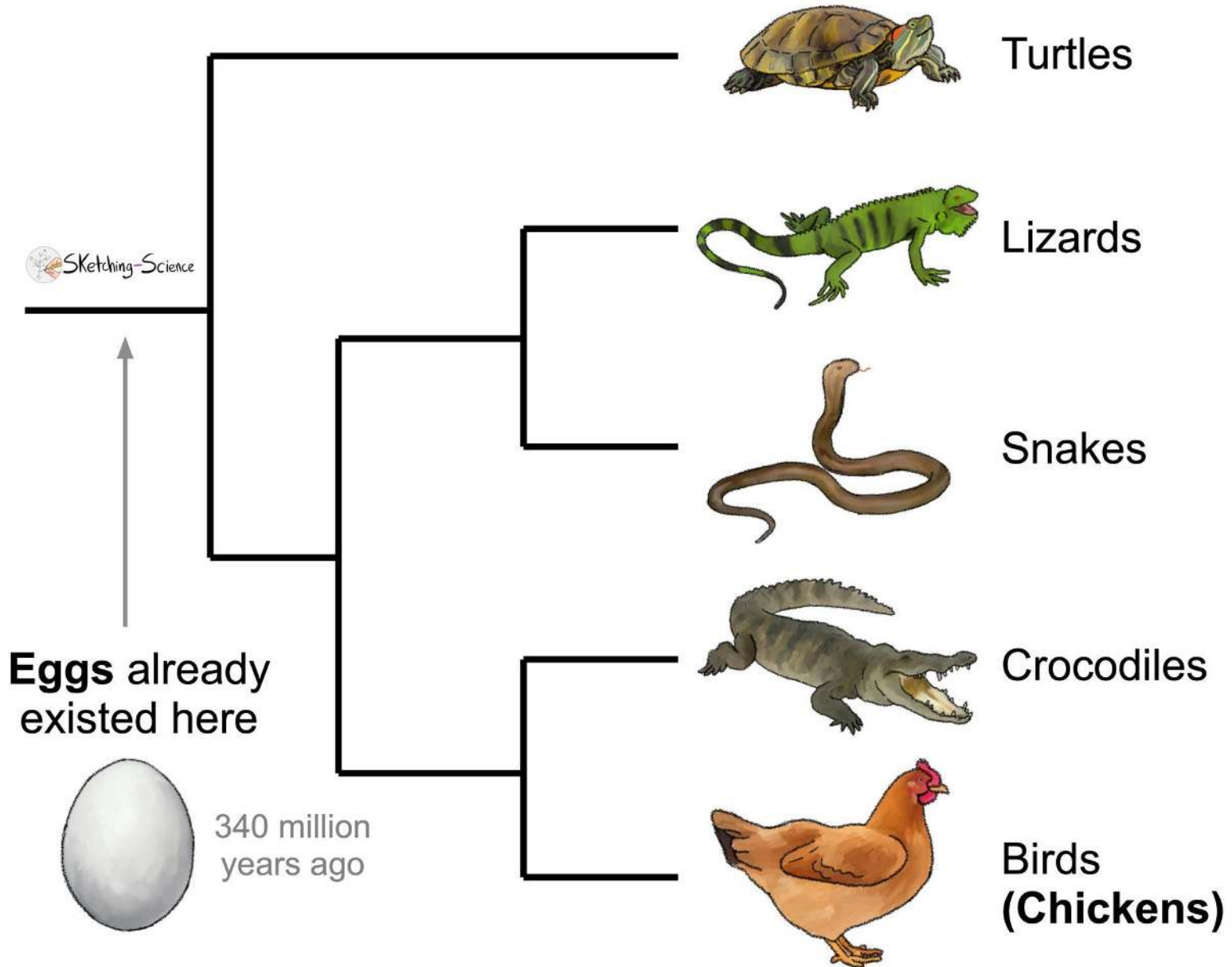


Crocodiles



Birds
(Chickens)

Which came first, the chicken or the egg?



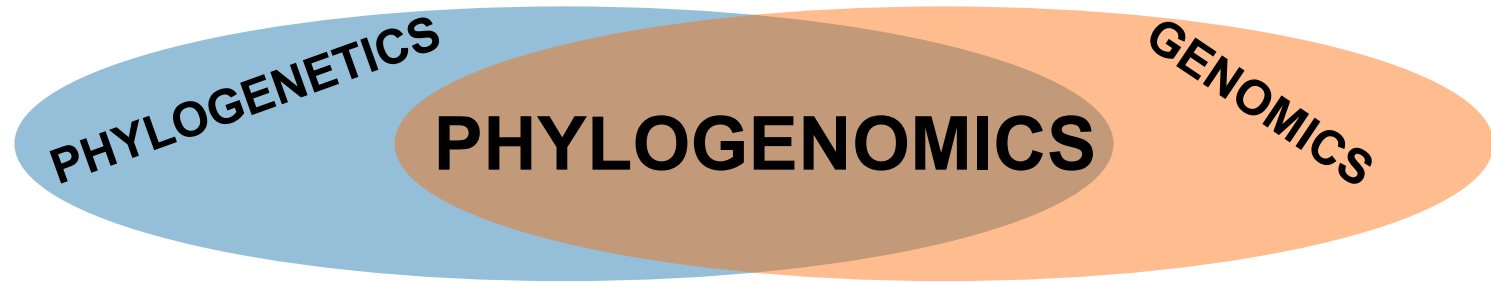
Content today's lecture

Intro de Phylogenomics

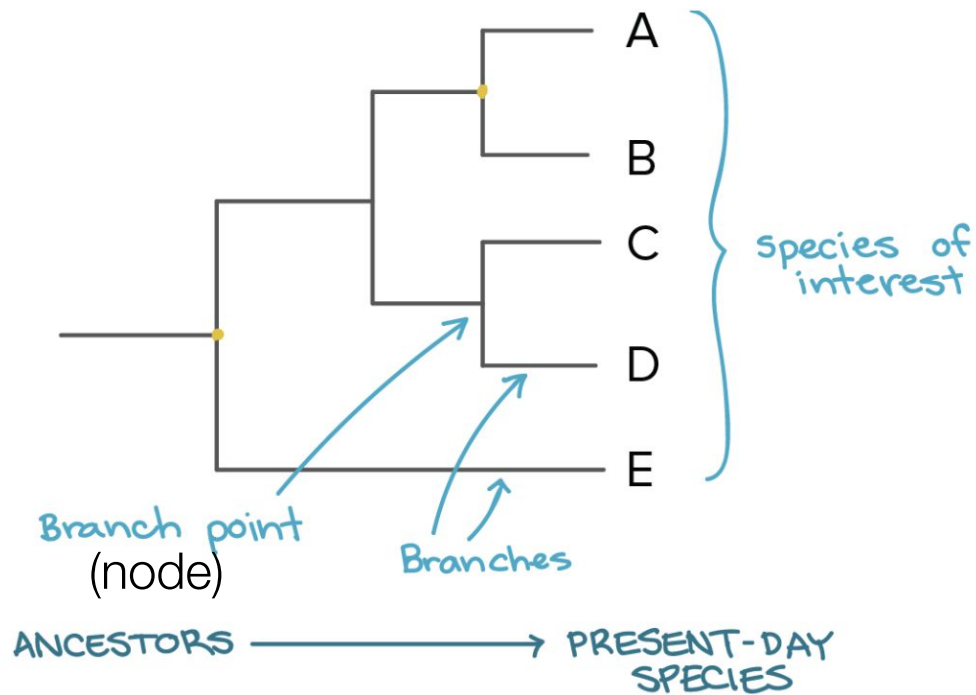
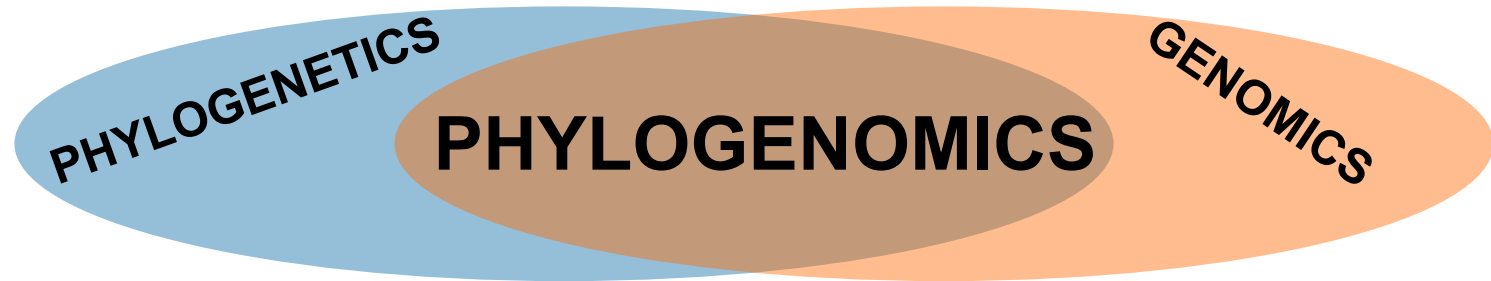
Hands-on species tree reconstruction & sensitivity analysis

Intro to Phylogenomics

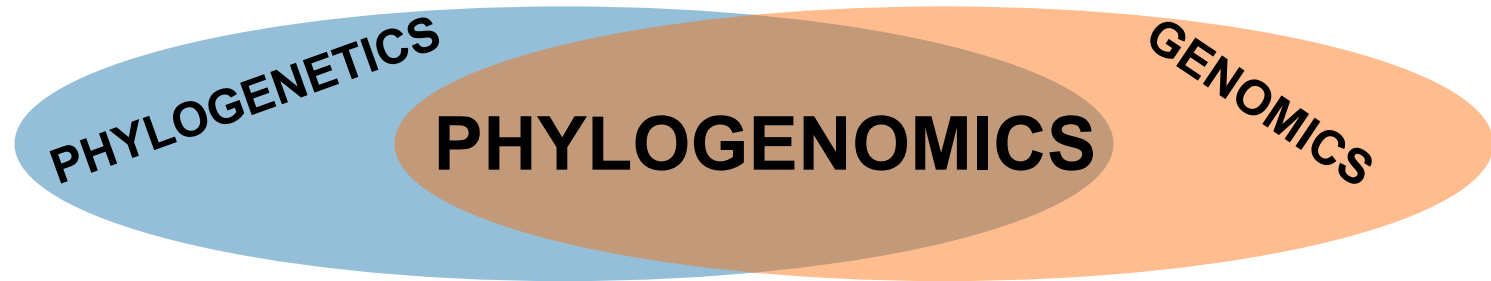
Intro to Phylogenomics



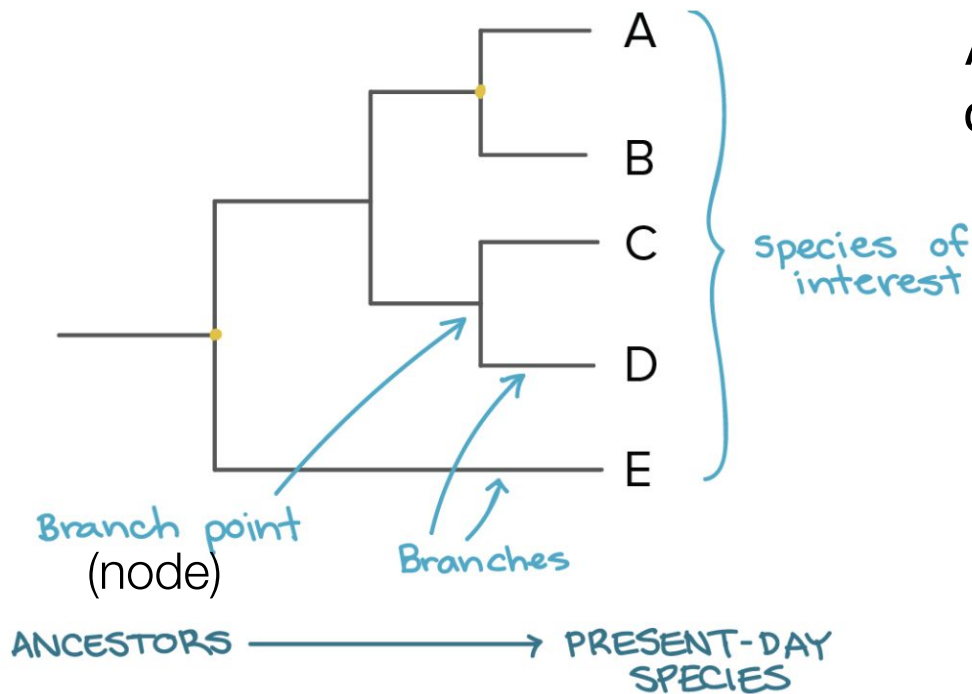
Intro to Phylogenomics



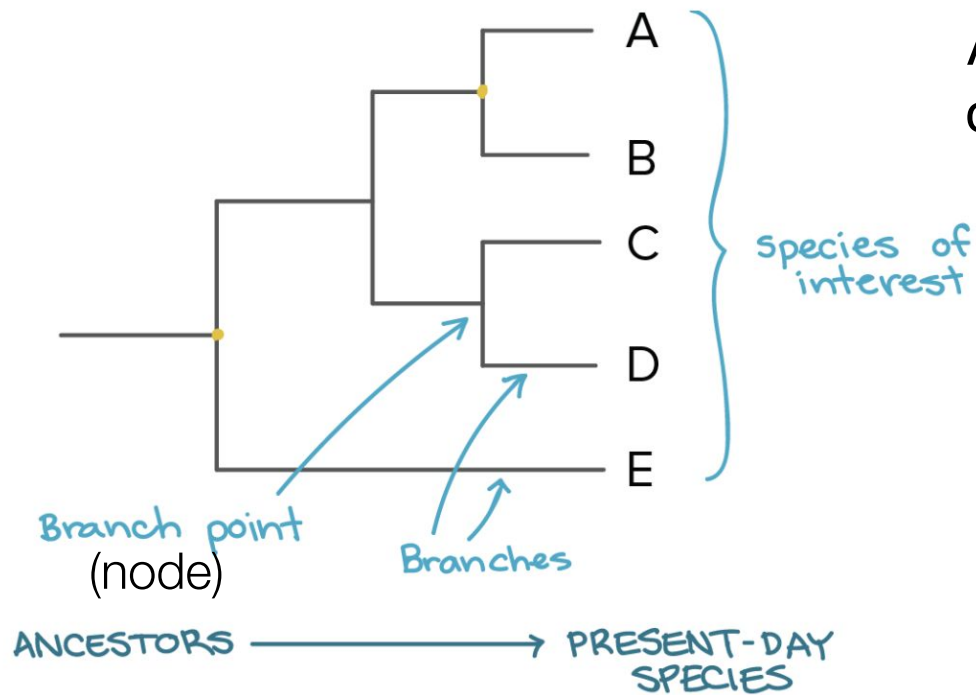
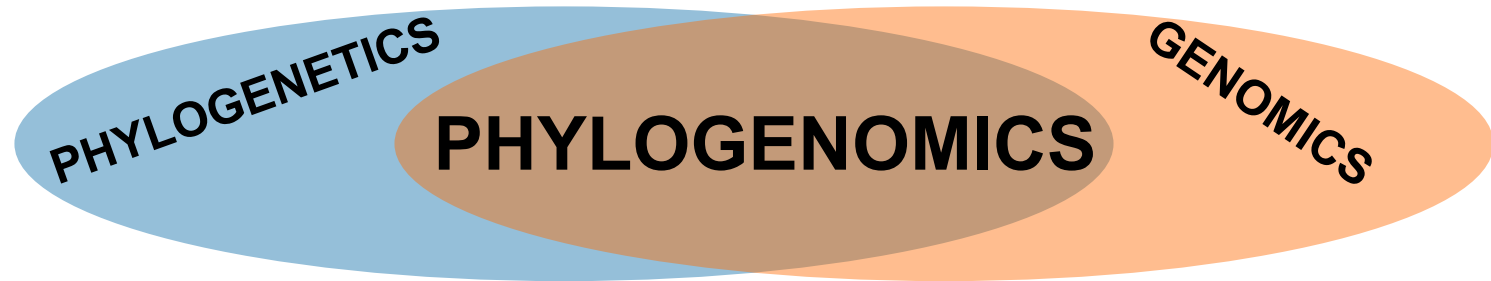
Intro to Phylogenomics



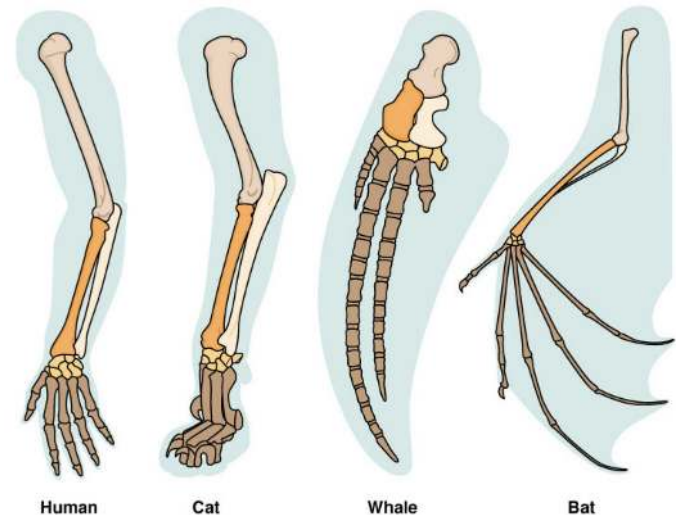
A **phylogenetic tree** is a hypothesis of how species or genes are related through evolution



Intro to Phylogenomics

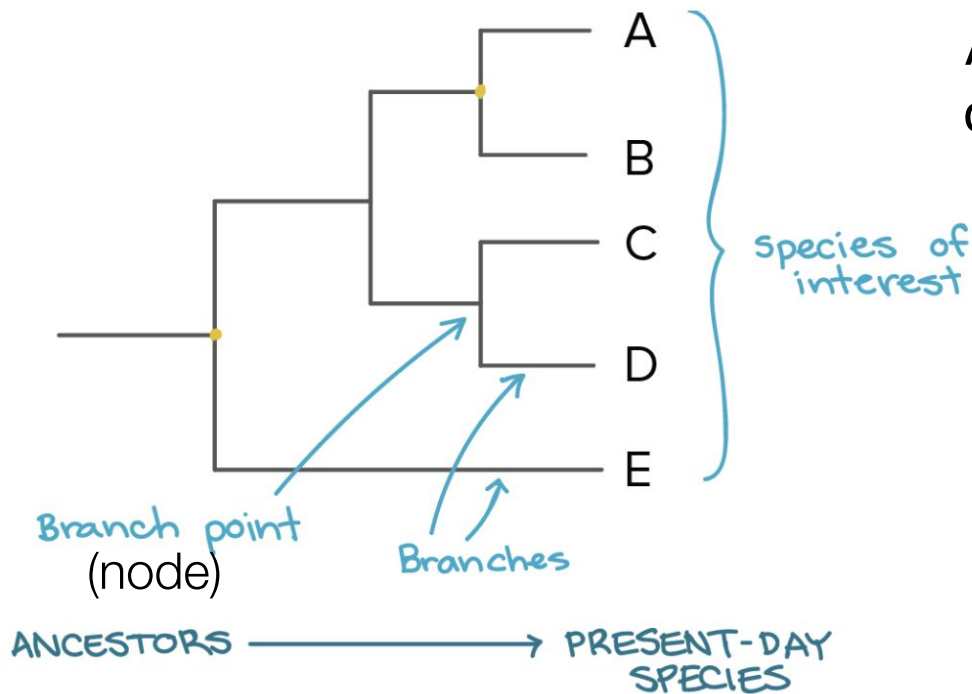
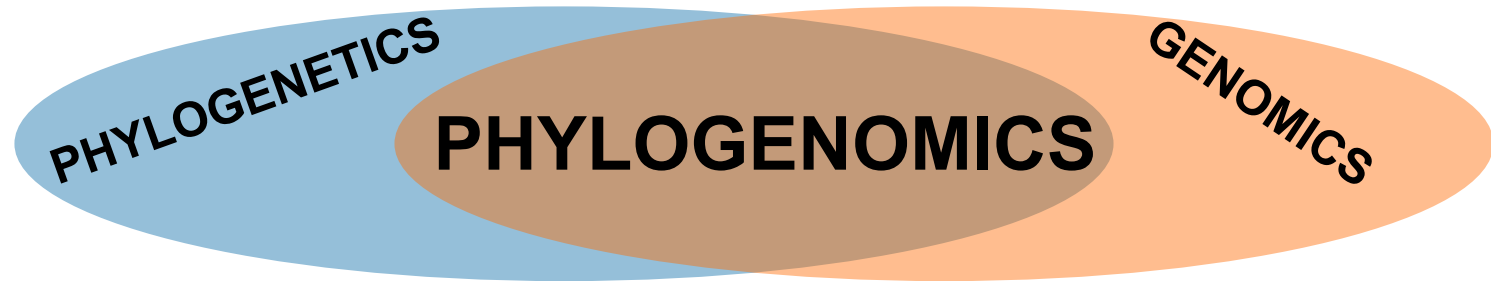


A **phylogenetic tree** is a hypothesis of how species or genes are related through evolution

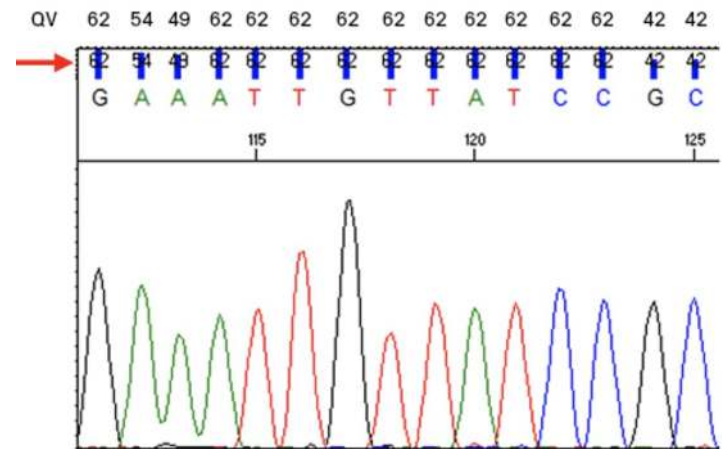


Morphological traits

Intro to Phylogenomics

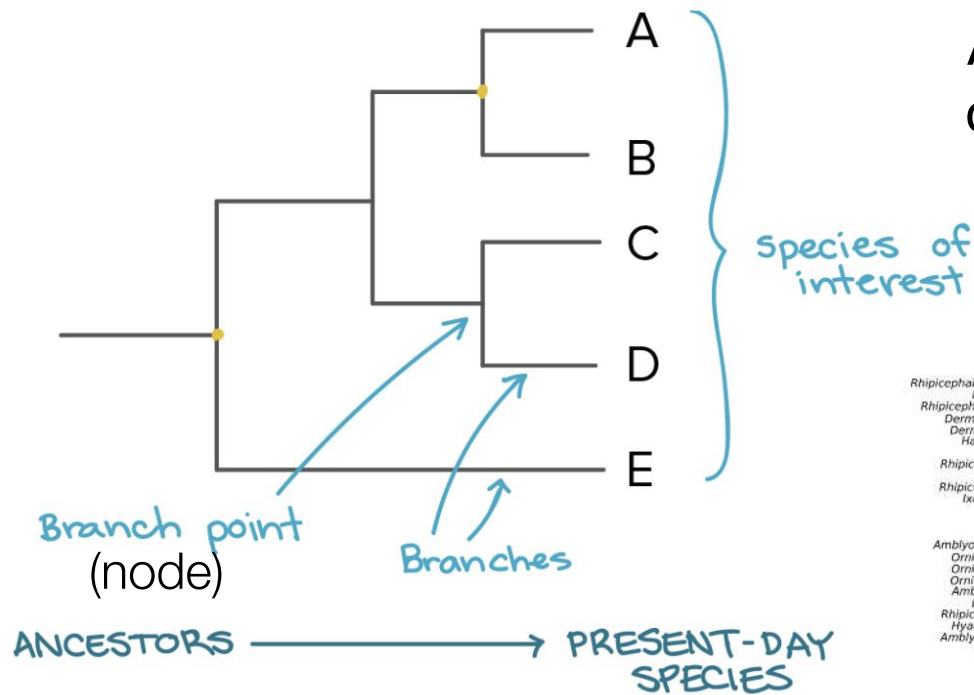
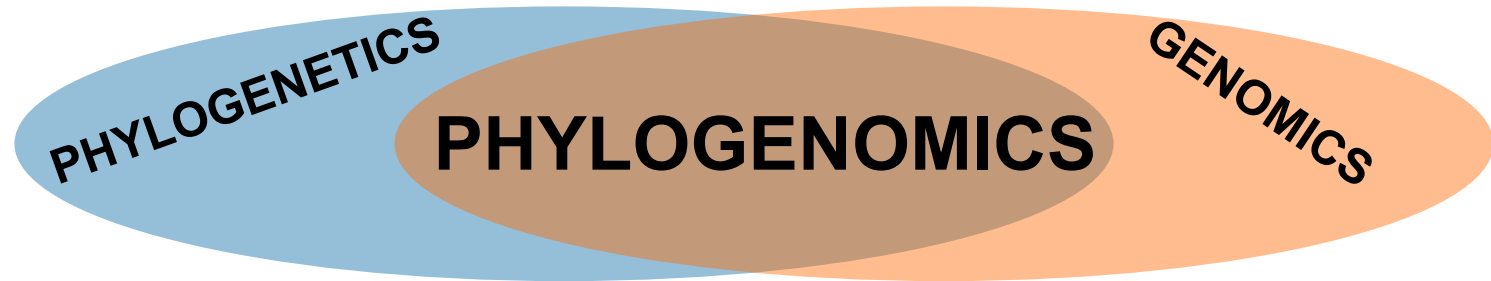


A **phylogenetic tree** is a hypothesis of how species or genes are related through evolution

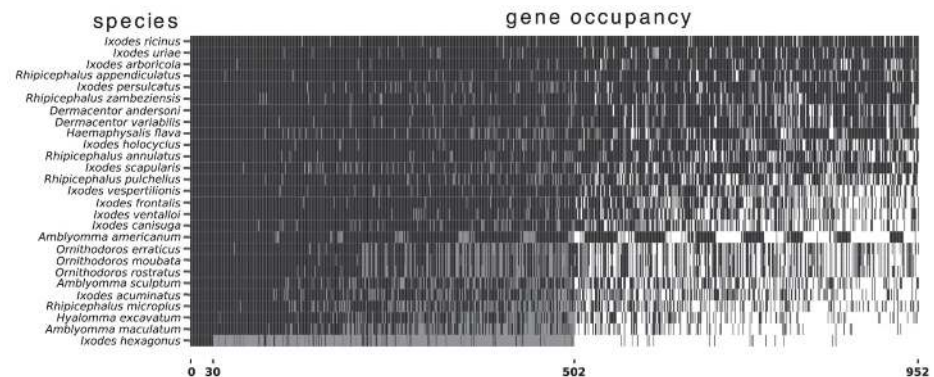


A few genes (eg, COI)

Intro to Phylogenomics

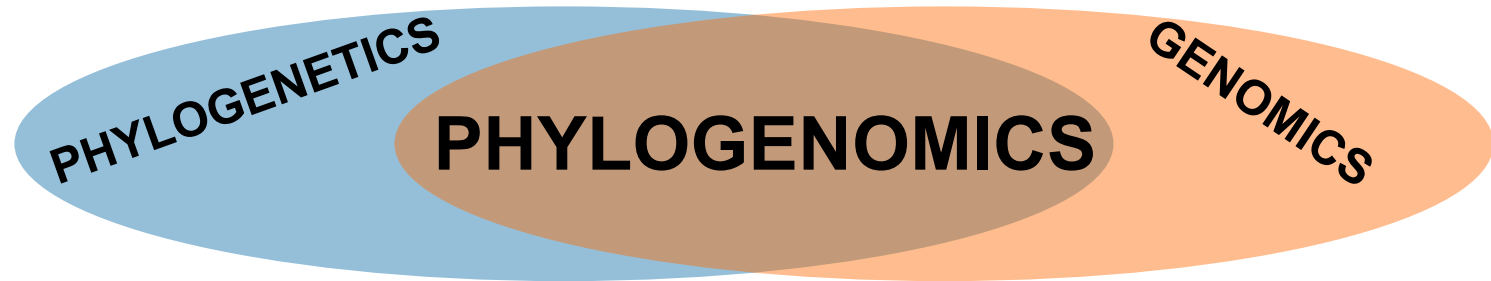


A **phylogenetic tree** is a hypothesis of how species or genes are related through evolution



100s / 1,000s of genes

Intro to Phylogenomics



8:163-167 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 \$5.00; www.genome.org

GENOME RESEARCH 163

Insight/Outlook

Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis

Jonathan A. Eisen¹

Department of Biological Sciences, Stanford University, Stanford, California 94305-5020 USA

The ability to accurately predict gene function based on gene sequence is an important tool in many areas of biological research. Such predictions have become particularly important in the genomics age in which numerous gene sequences are generated with little or no accompanying experimentally determined functional information. Almost all functional prediction methods rely on the identification, characterization,

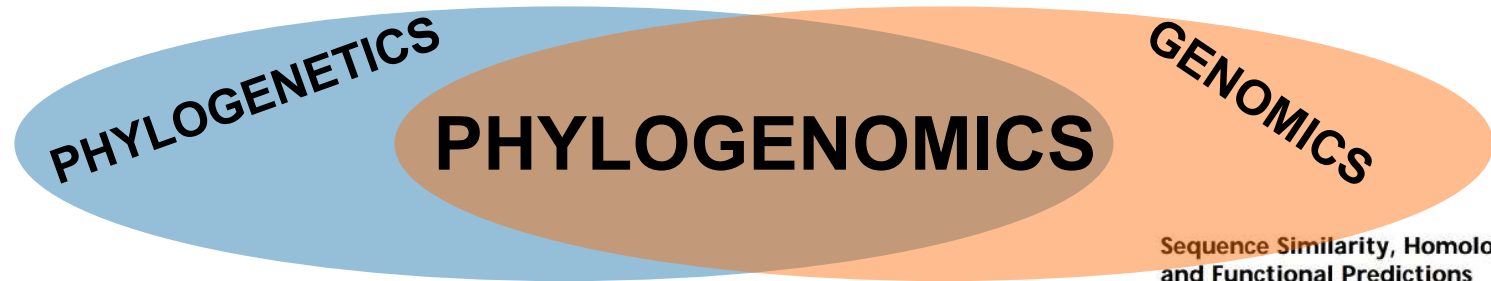
(e.g., Altschul et al. 1989; Goldman et al. 1996). In this commentary, I discuss the use of evolutionary information in the prediction of gene function. To appreciate the potential of a *phylogenomic approach* to the *prediction of gene function*, it is necessary to first discuss how gene sequence is commonly used to predict gene function and some general features about gene evolution.

convergence (the exact threshold for such an inference is not well established).

Improvements in database search programs have made the identification of likely homologs much faster, easier, and more reliable (Altschul et al. 1997; Henikoff et al. 1998). However, as discussed above, in many cases the identification of homologs is not sufficient to make specific functional predictions be-

Phylogenomics: prediction of gene function and gene family evolution

Intro to Phylogenomics



8:163-167 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 \$5.00; www.genome.org

GENOME RESEARCH 163

Insight/Outlook

Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis

Jonathan A. Eisen¹

Department of Biological Sciences, Stanford University, Stanford, California 94305-5020 USA

The ability to accurately predict gene function based on gene sequence is an important tool in many areas of biological research. Such predictions have become particularly important in the genomics age in which numerous gene sequences are generated with little or no accompanying experimentally determined functional information. Almost all functional prediction methods rely on the identification, characterization,

(e.g., Altschul et al. 1989; Goldman et al. 1996). In this commentary, I discuss the use of evolutionary information in the prediction of gene function. To appreciate the potential of a *phylogenomic approach* to the *prediction of gene function*, it is necessary to first discuss how gene sequence is commonly used to predict gene function and some general features about gene evolution.

convergence (the exact threshold for such an inference is not well established).

Improvements in database search programs have made the identification of likely homologs much faster, easier, and more reliable (Altschul et al. 1997; Henikoff et al. 1998). However, as discussed above, in many cases the identification of homologs is not sufficient to make specific functional predictions be-

Sequence Similarity, Homology, and Functional Predictions

To make use of the identification of sequence similarity between genes, it is helpful to understand how such similarity arises. Genes can become similar in sequence either as a result of *convergence* (similarities that have arisen without a common evolutionary history) or *descent with modification* from a common ancestor (also known as *homology*). It is imperative to recognize that *sequence similarity* and *homology* are not interchangeable terms. Not all homologs are similar in sequence (i.e., homologous genes can diverge so much that similarities are difficult or impossible to detect) and not all similarities are due to homology (Reeck et al. 1987; Hillis 1994). Similarity due to convergence, which is likely limited to small regions of genes, can be useful for some functional predictions (Henikoff et al. 1997). However, most sequence-based functional predictions are based on the identification (and subsequent analysis) of similarities that are thought to be due to homology. Because homology is a statement about common ancestry, it cannot be proven directly from sequence similarity. In these cases, the inference of homology is made based on finding levels of sequence similarity that are thought to be too high to be due to

Phylogenomics: prediction of gene function and gene family evolution

Intro to Phylogenomics

PHYLOGENETICS

PHYLOGENOMICS

GENOMICS

8:163-167 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 \$5.00; www.genome.org

GENOME RESEARCH 163

Insight/Outlook

Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis

Jonathan A. Eisen¹

Department of Biological Sciences, Stanford University, Stanford, California 94305-5020 USA

The ability to accurately predict gene function based on gene sequence is an important tool in many areas of biological research. Such predictions have become particularly important in the genomics age in which numerous gene sequences are generated with little or no accompanying experimentally determined functional information. Almost all functional prediction methods rely on the identification, characterization,

(e.g., Altschul et al. 1989; Goldman et al. 1996). In this commentary, I discuss the use of evolutionary information in the prediction of gene function. To appreciate the potential of a *phylogenomic approach* to the *prediction of gene function*, it is necessary to first discuss how gene sequence is commonly used to predict gene function and some general features about gene evolution.

convergence (the exact threshold for such an inference is not well established).

Improvements in database search programs have made the identification of likely homologs much faster, easier, and more reliable (Altschul et al. 1997; Henikoff et al. 1998). However, as discussed above, in many cases the identification of homologs is not sufficient to make specific functional predictions be-

Phylogenomics: prediction of gene function and gene family evolution

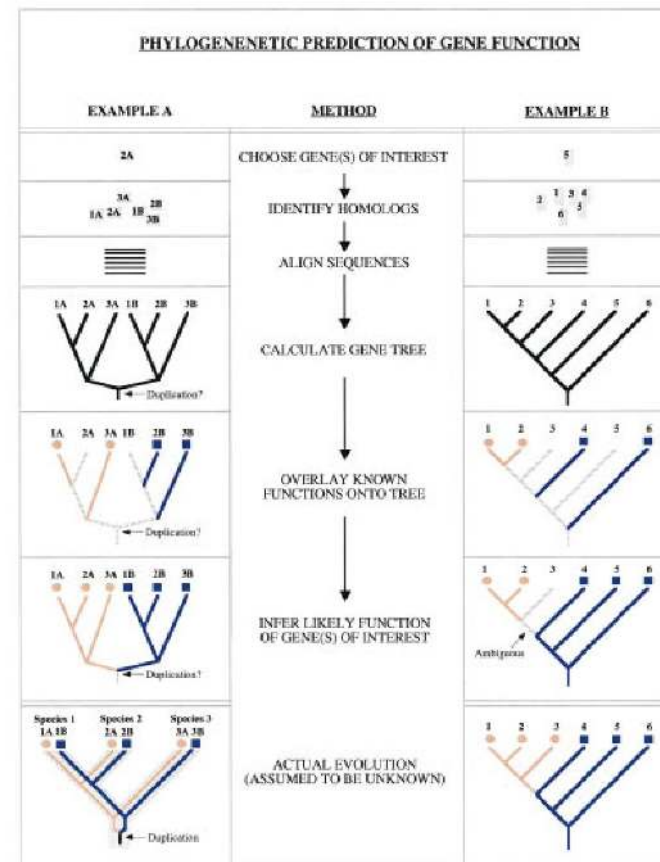
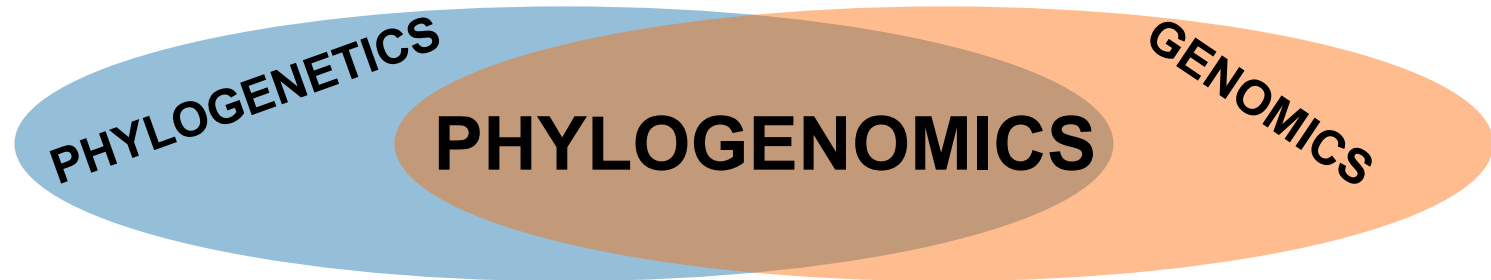


Figure 1 Outline of a phylogenomic methodology. In this method, information about the evolutionary relationships among genes is used to predict the functions of uncharacterized genes (see text for details). Two hypothetical scenarios are presented and the path of trying to

Intro to Phylogenomics



1414–1419 | PNAS | February 5, 2002 | vol. 99 | no. 3

www.pnas.org/cgi/doi/10.1073/pnas.032662799

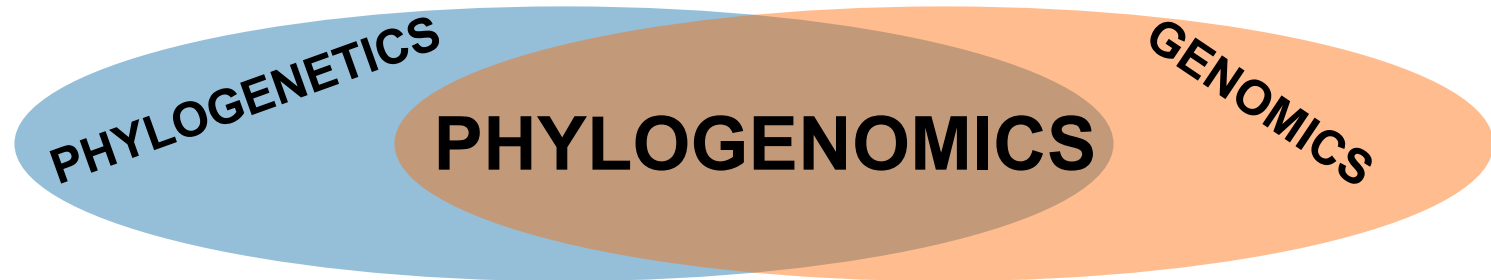
The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*

Eric Baptiste*, Henner Brinkmann†, Jennifer A. Lee‡, Dorothy V. Moore‡, Christoph W. Sensen§, Paul Gordon¶, Laure Duruflé*, Terry Gaasterland‡, Philippe Lopez*, Miklós Müller‡, and Hervé Philippe*||

The phylogenetic relationships of amoebae are poorly resolved. To address this difficult question, we have sequenced 1,280 expressed sequence tags from *Mastigamoeba balamuthi* and assembled a large data set containing 123 genes for representatives of three phenotypically highly divergent major amoeboid lineages: Pelobionta, Entamoebidae, and Mycetozoa. Phylogenetic reconstruction was performed on ≈25,000 aa positions for 30 species by using maximum-likelihood approaches. All well-established eukaryotic groups were recovered with high statistical support, validating our approach. Interestingly, the three amoeboid lineages strongly clustered together in agreement with the Conosa hypothesis [as defined by T. Cavalier-Smith (1998) *Biol. Rev. Cambridge Philos. Soc.* 73, 203–266]. Two amitochondriate amoebae, the free-living *Mastigamoeba* and the human parasite *Entamoeba*, formed a significant sister group to the exclusion of the mycetozoan *Dictyostelium*. This result suggested that a part of the reductive process in the evolution of *Entamoeba* (e.g., loss of typical mitochondria) occurred in its free-living ancestors. Applying this inexpensive expressed sequence tag approach to many other lineages will surely improve our understanding of eukaryotic evolution.

Phylogenomics: species tree inference

Intro to Phylogenomics



1414–1419 | PNAS | February 5, 2002 | vol. 99 | no. 3

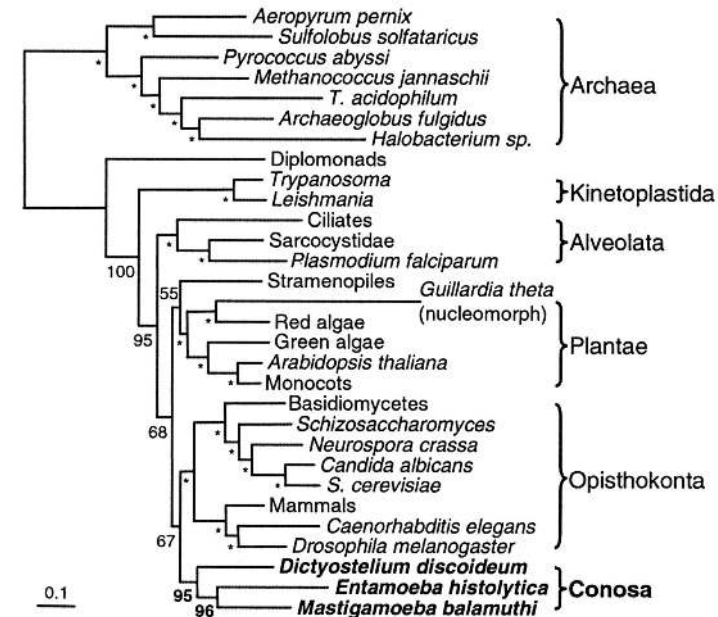
www.pnas.org/cgi/doi/10.1073/pnas.032662799

The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*

Eric Baptiste*, Henner Brinkmann†, Jennifer A. Lee‡, Dorothy V. Moore‡, Christoph W. Sensen§, Paul Gordon¶, Laure Duruflé*, Terry Gaasterland‡, Philippe Lopez*, Miklós Müller‡, and Hervé Philippe*||

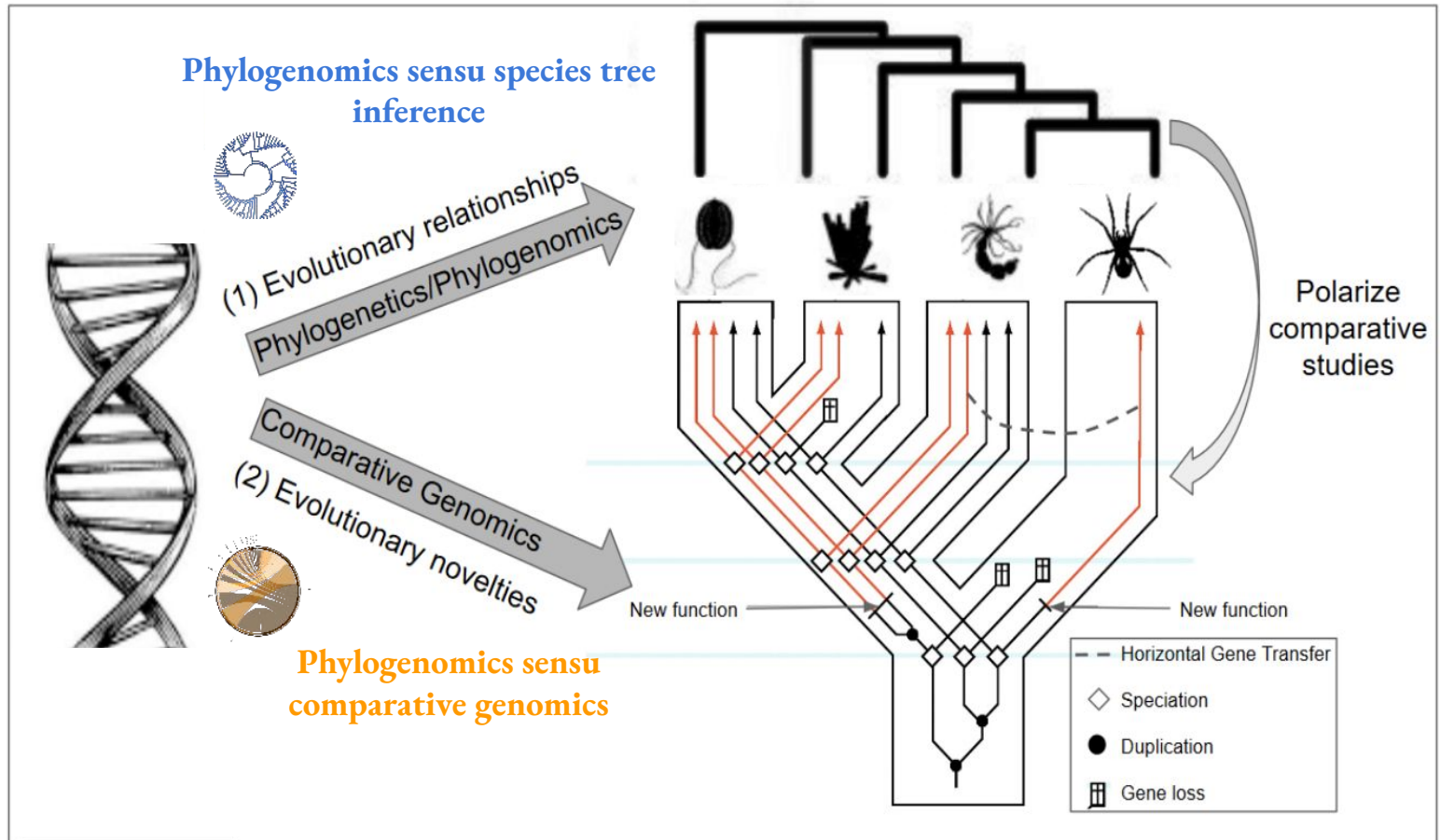
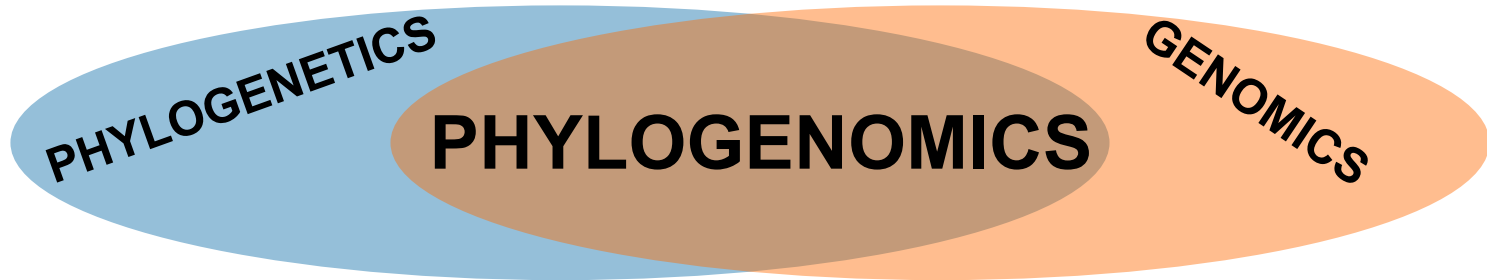
The phylogenetic relationships of amoebae are poorly resolved. To address this difficult question, we have sequenced 1,280 expressed sequence tags from *Mastigamoeba balamuthi* and assembled a large data set containing 123 genes for representatives of three phenotypically highly divergent major amoeboid lineages: Pelobionta, Entamoebidae, and Mycetozoa. Phylogenetic reconstruction was performed on $\approx 25,000$ aa positions for 30 species by using maximum-likelihood approaches. All well-established eukaryotic groups were recovered with high statistical support, validating our approach. Interestingly, the three amoeboid lineages strongly clustered together in agreement with the Conosa hypothesis [as defined by T. Cavalier-Smith (1998) *Biol. Rev. Cambridge Philos. Soc.* 73, 203–266]. Two amitochondriate amoebae, the free-living *Mastigamoeba* and the human parasite *Entamoeba*, formed a significant sister group to the exclusion of the mycetozoan *Dictyostelium*. This result suggested that a part of the reductive process in the evolution of *Entamoeba* (e.g., loss of typical mitochondria) occurred in its free-living ancestors. Applying this inexpensive expressed sequence tag approach to many other lineages will surely improve our understanding of eukaryotic evolution.

Phylogenomics: species tree inference

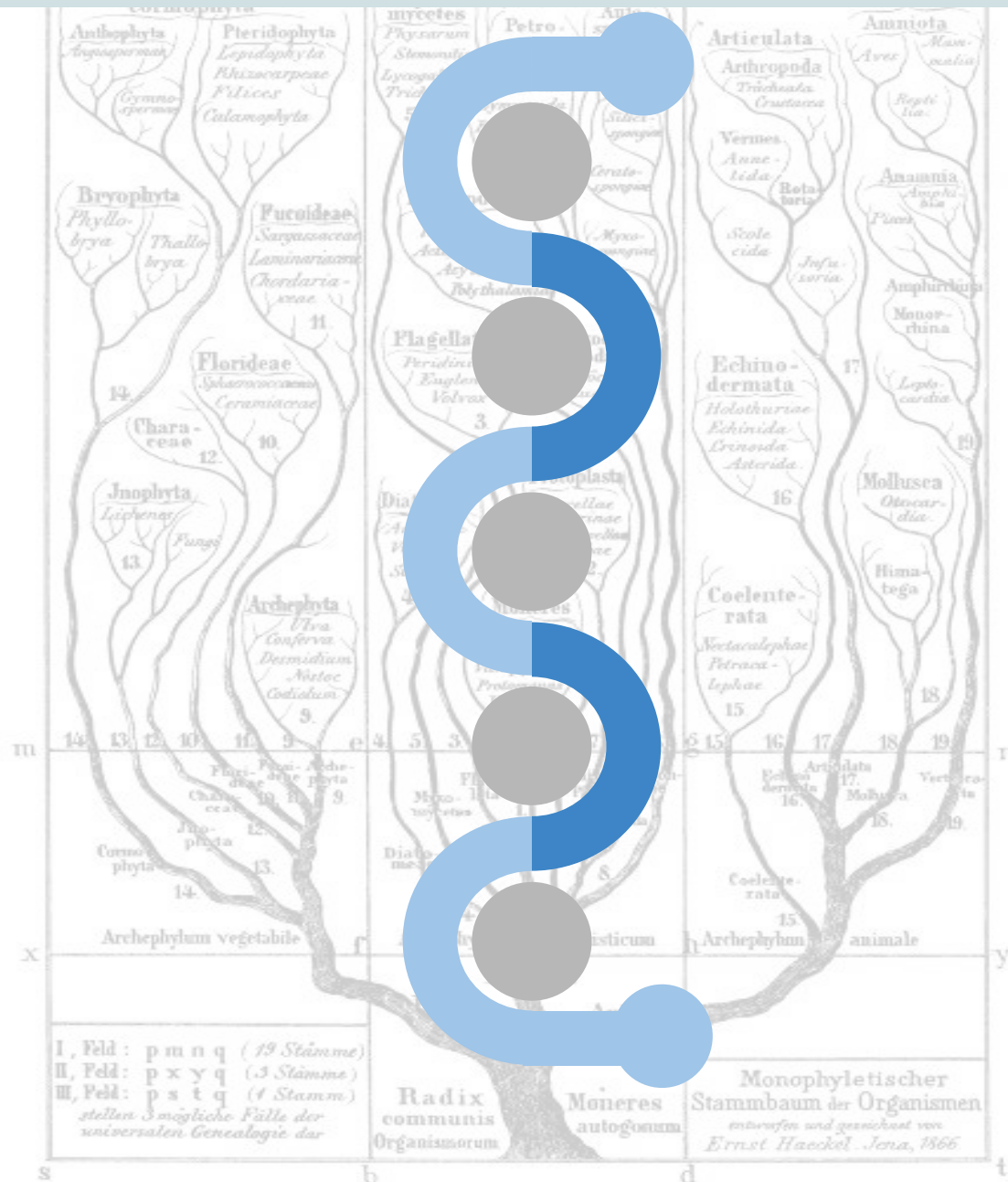


ML tree based on 25,032 aa positions. * Indicates a constrained node. We used the JTT model, without taking into account among-sites rate variation. The branch lengths have been computed on the concatenated sequences. BVs were obtained by bootstrapping the 123 genes.

Intro to Phylogenomics

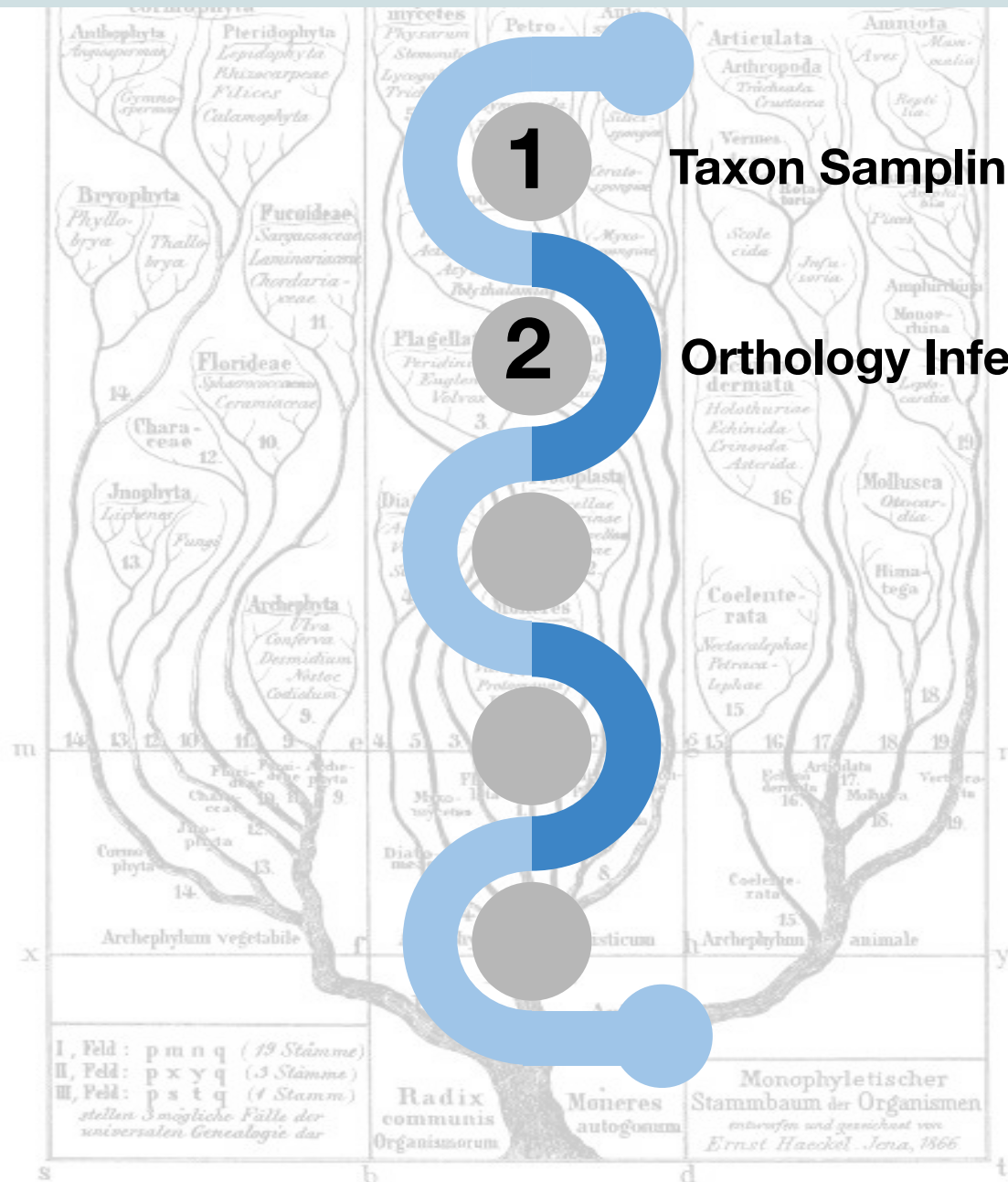


How to infer a species tree

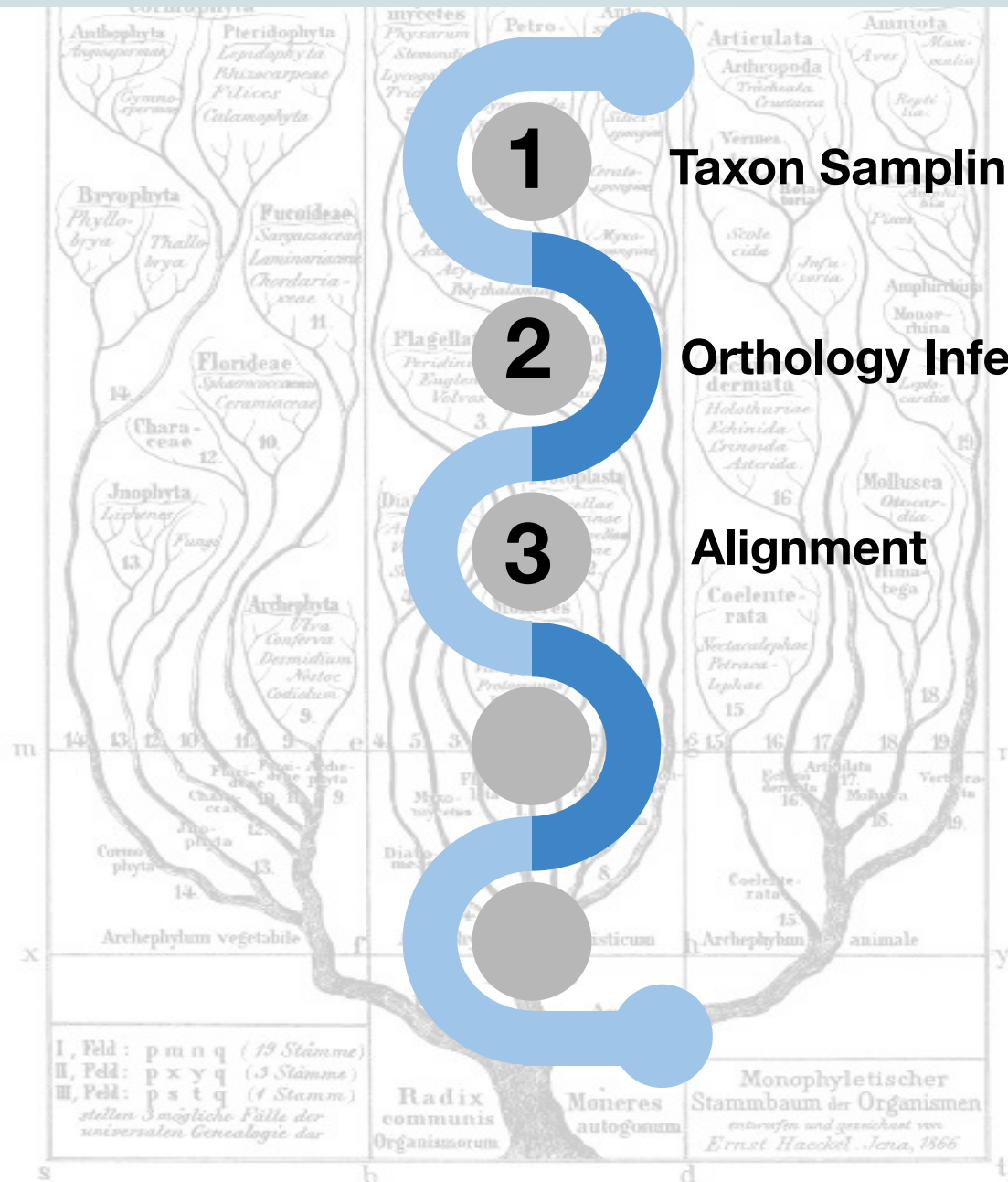


Taxon Sampling

How to infer a species tree



How to infer a species tree



1 Taxon Sampling

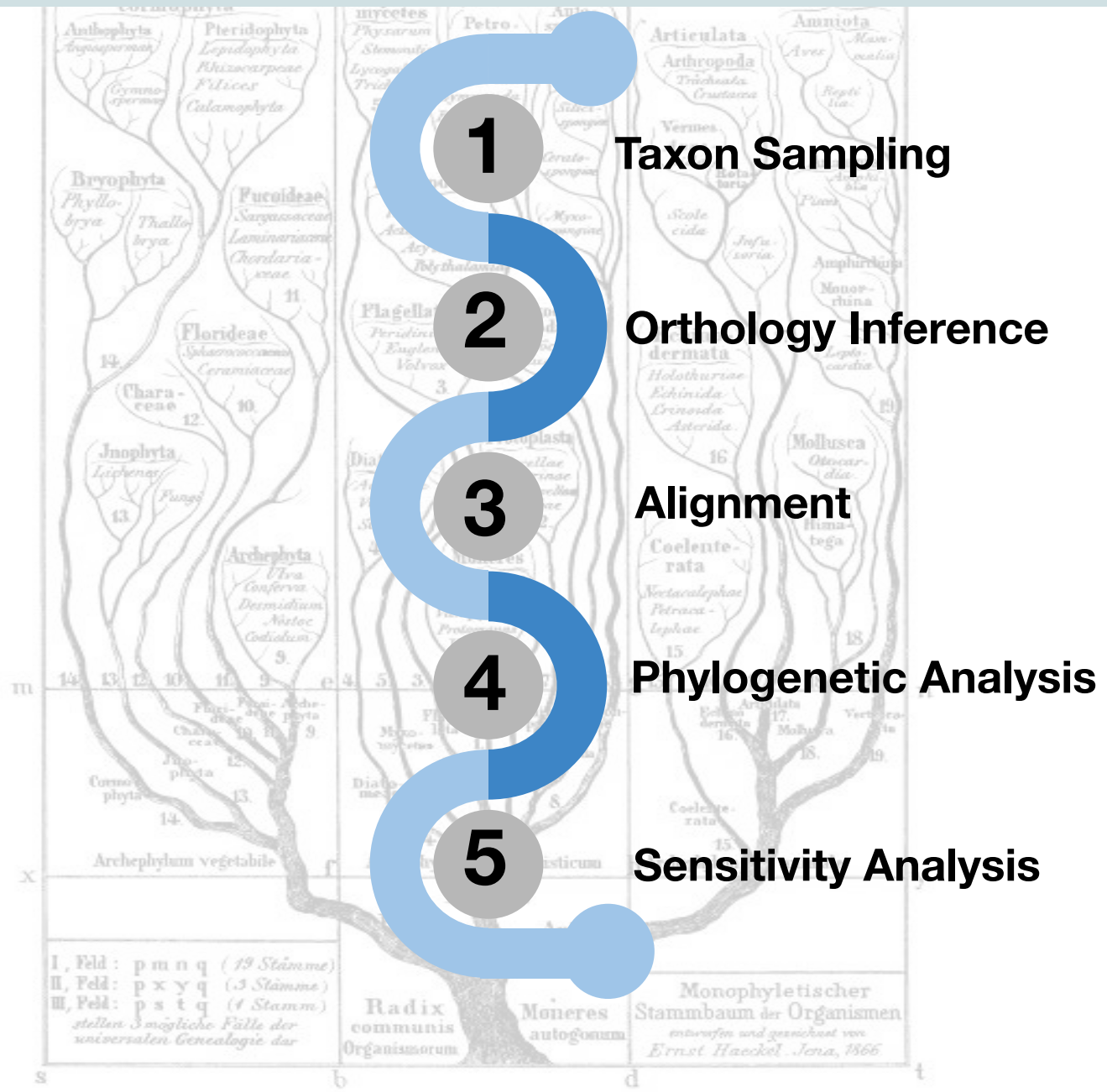
2 Orthology Inference

3 Alignment

4 Phylogenetic Analysis

Phylogenetic Analysis

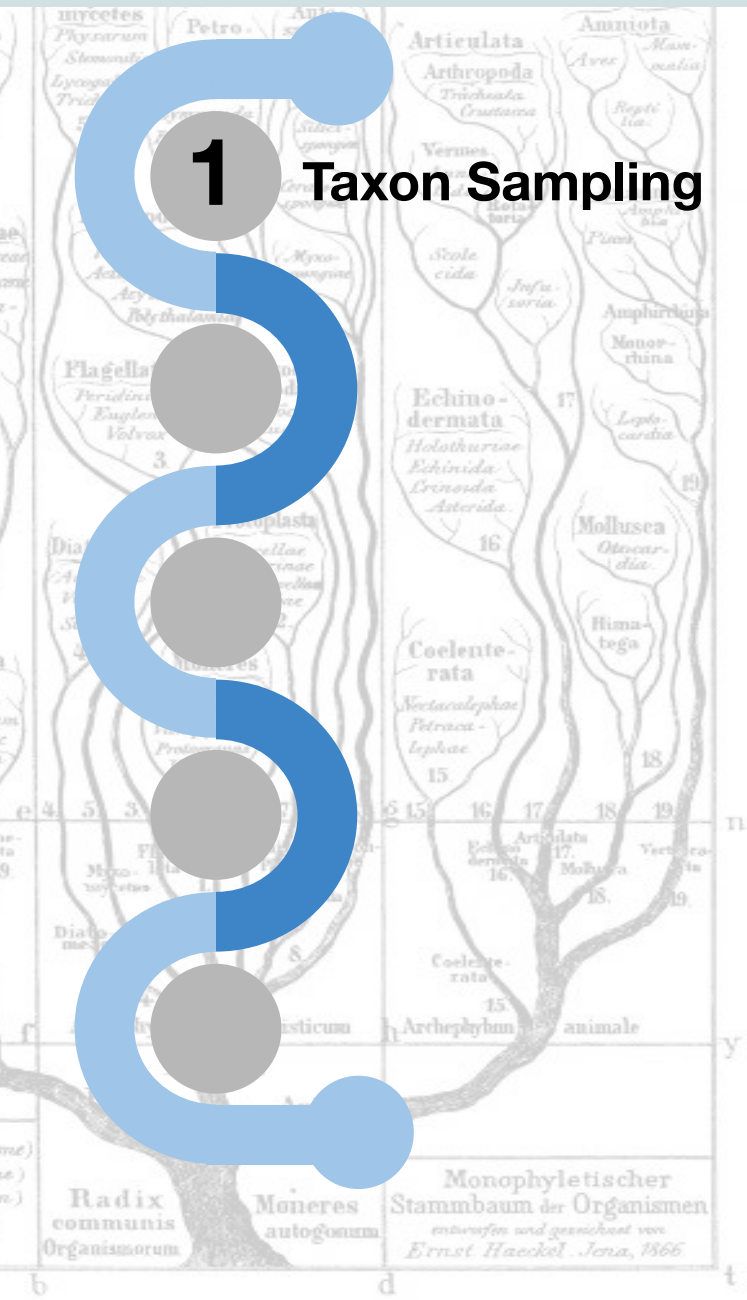
How to infer a species tree



How to infer a species tree

1

Taxon Sampling



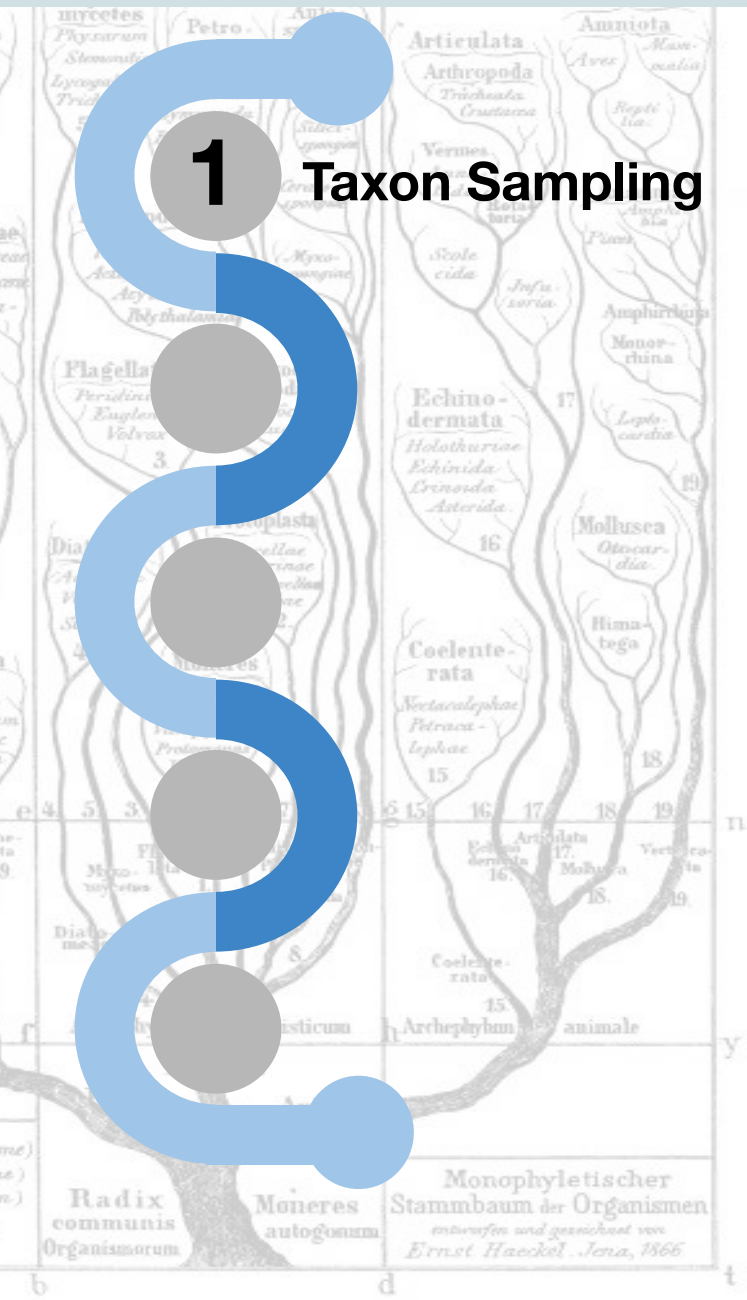
How to infer a species tree

1

Taxon Sampling

Key message: taxon sampling matters a lot

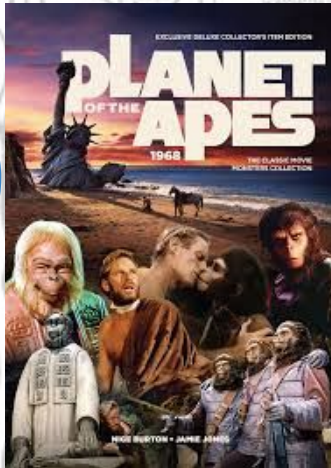
Incomplete, biased, or improper taxon sampling can lead to misleading results in reconstructing evolutionary relationships.



How to infer a species tree

1

Taxon Sampling



Key message: taxon sampling matters a lot

Incomplete, biased, or improper taxon sampling can lead to misleading results in reconstructing evolutionary relationships.

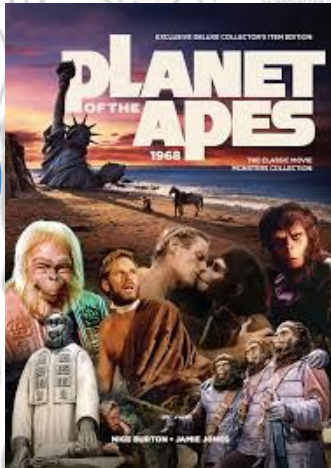
How to infer a species tree

1

Taxon Sampling

Key message: taxon sampling matters a lot

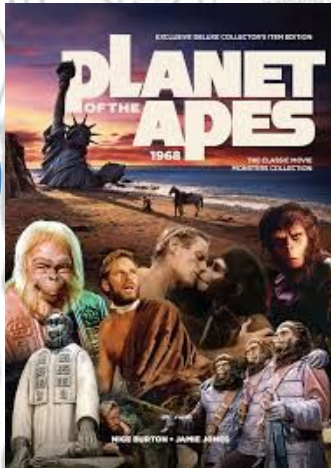
Incomplete, biased, or improper taxon sampling can lead to misleading results in reconstructing evolutionary relationships.



How to infer a species tree

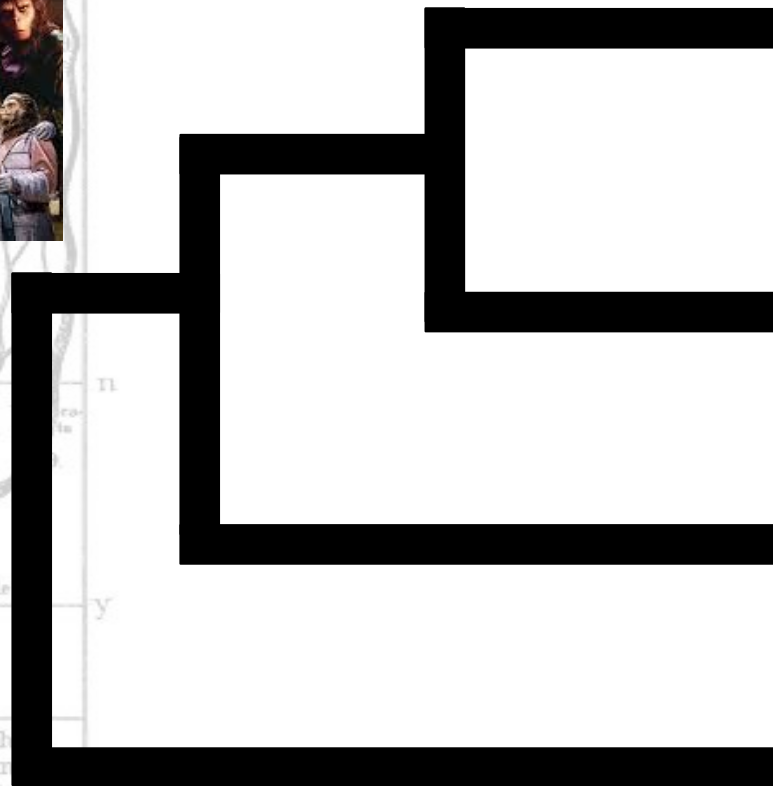
1

Taxon Sampling



Key message: taxon sampling matters a lot

Incomplete, biased, or improper taxon sampling can lead to misleading results in reconstructing evolutionary relationships.



How to infer a species tree

1

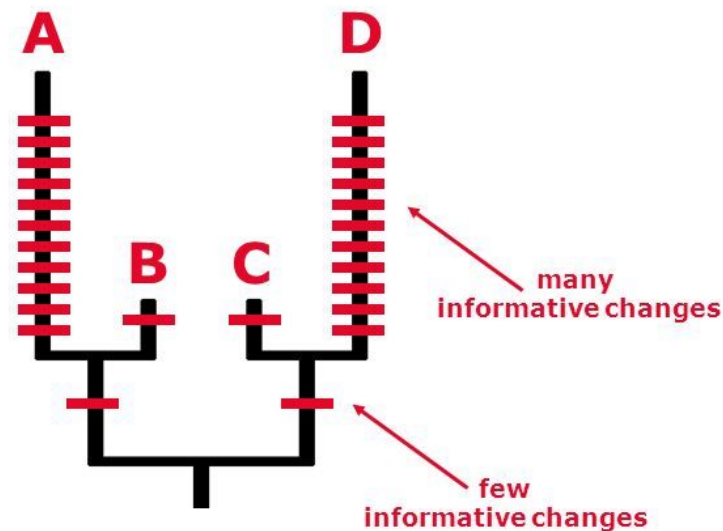
Taxon Sampling

Key message: taxon sampling matters a lot

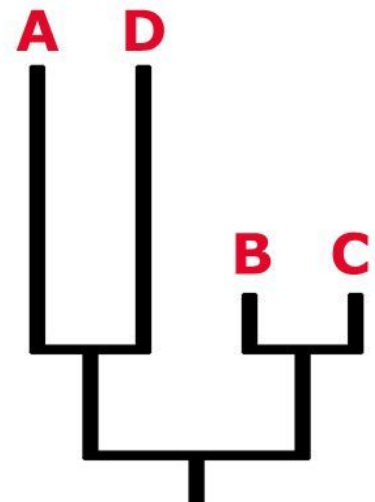
Incomplete, biased, or improper taxon sampling can lead to misleading results in reconstructing evolutionary relationships.

Long Branch Attraction

True Tree



Reconstructed Tree



How to infer a species tree

1

Taxon Sampling

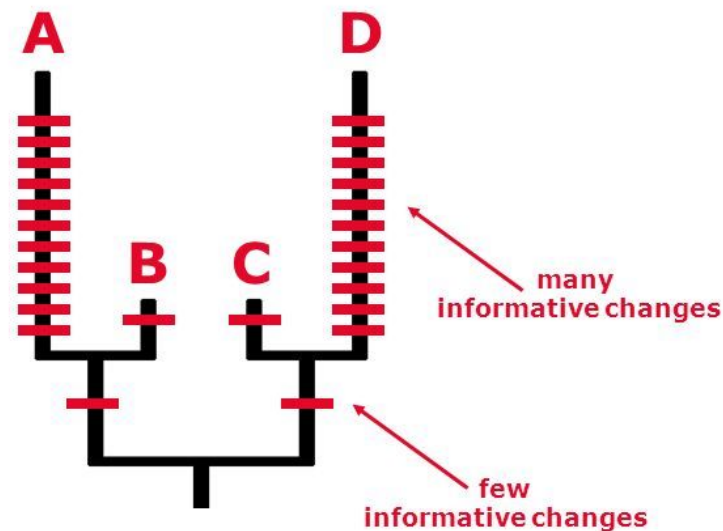
Key message: taxon sampling matters a lot

Incomplete, biased, or improper taxon sampling can lead to misleading results in reconstructing evolutionary relationships.

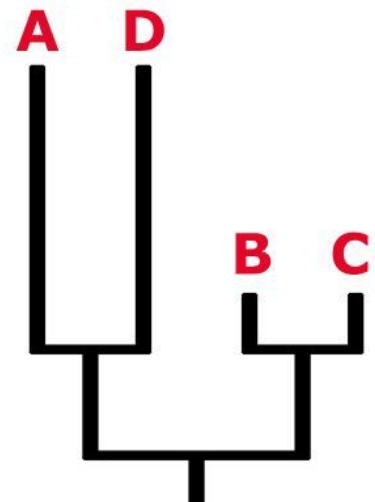
Long Branch Attraction

Outgroups / Fast-evolving lineages / Compositional heterogeneity

True Tree



Reconstructed Tree



How to infer a species tree

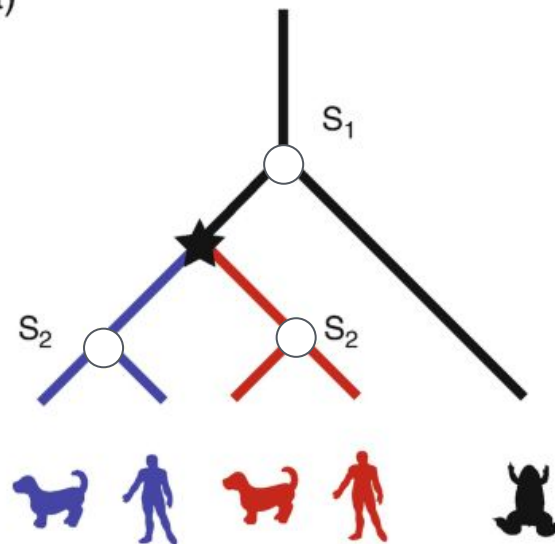
1

Taxon Sampling

2

Orthology Inference

a)



How to infer a species tree

1

Taxon Sampling

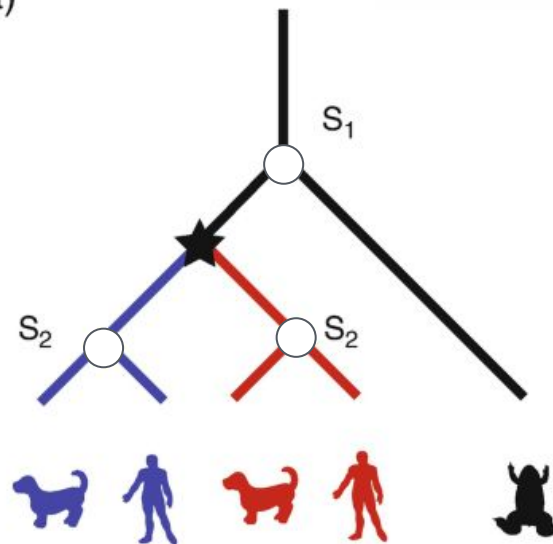
2

Orthology Inference

Definitions

- Two genes are **orthologs** if their MRCA is a **speciation**: ○
- Two genes are **paralogs** if their MRCA is a **duplication**: ★

a)



How to infer a species tree

1

Taxon Sampling

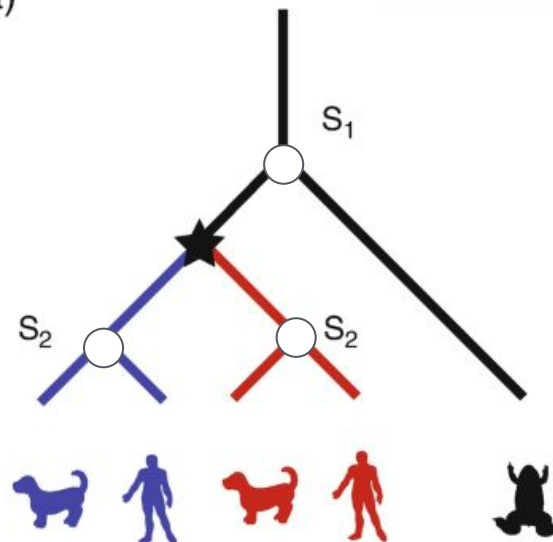
2

Orthology Inference

Definitions

- Two genes are **orthologs** if their MRCA is a **speciation**: ○
- Two genes are **paralogs** if their MRCA is a **duplication**: ★

a)



Orthology relationships are inferred pairwise

How to infer a species tree

1

Taxon Sampling

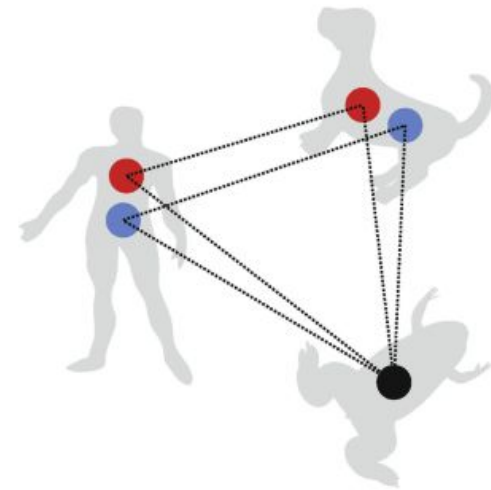
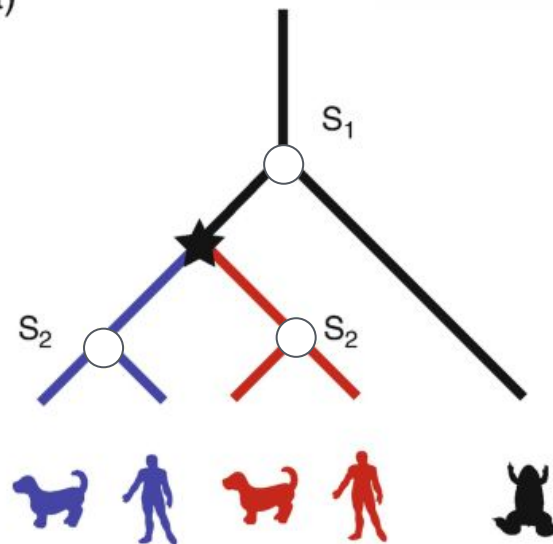
2

Orthology Inference

Definitions

- Two genes are **orthologs** if their MRCA is a **speciation**: ○
- Two genes are **paralogs** if their MRCA is a **duplication**: ★

a)



Orthology relationships are inferred pairwise

How to infer a species tree

1

Taxon Sampling

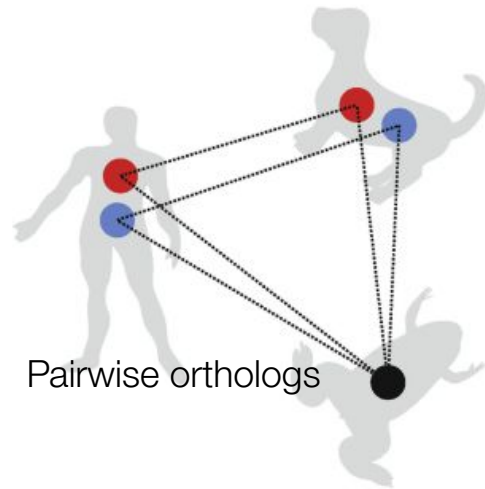
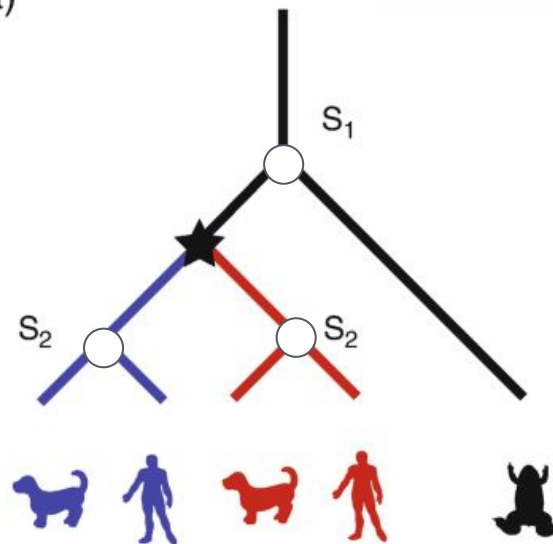
2

Orthology Inference

Definitions

- Two genes are **orthologs** if their MRCA is a **speciation**: ○
- Two genes are **paralogs** if their MRCA is a **duplication**: ★

a)



Pairwise orthologs

Orthology relationships are inferred pairwise

How to infer a species tree

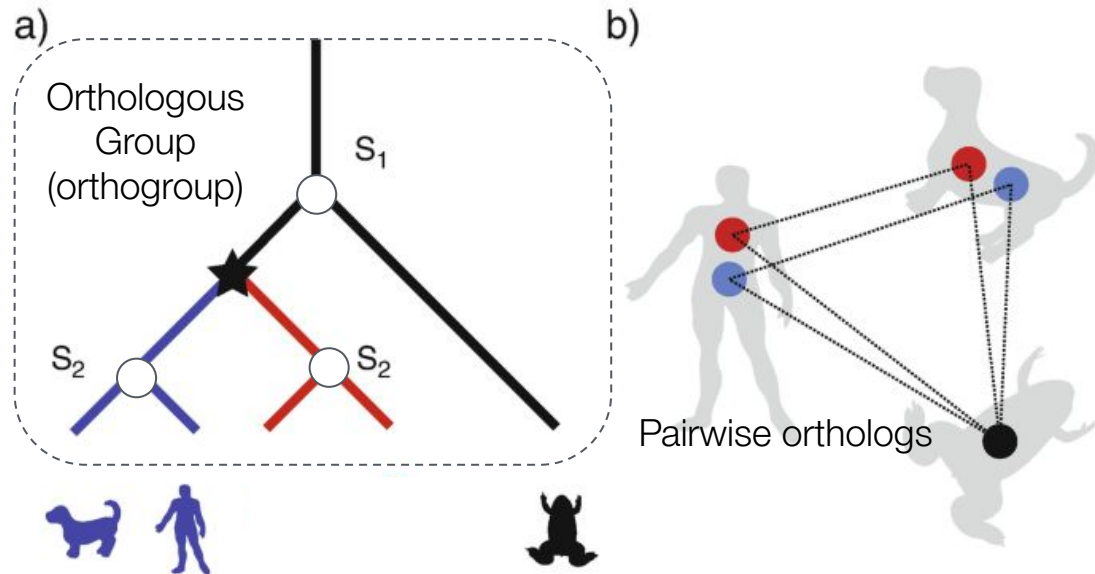
1

Taxon Sampling

2

Orthology Inference

For phylogenetic inference, our starting point are the Orthologous Groups, (OGs) since they are derived from speciation events



Orthology relationships are inferred pairwise

How to infer a species tree

1

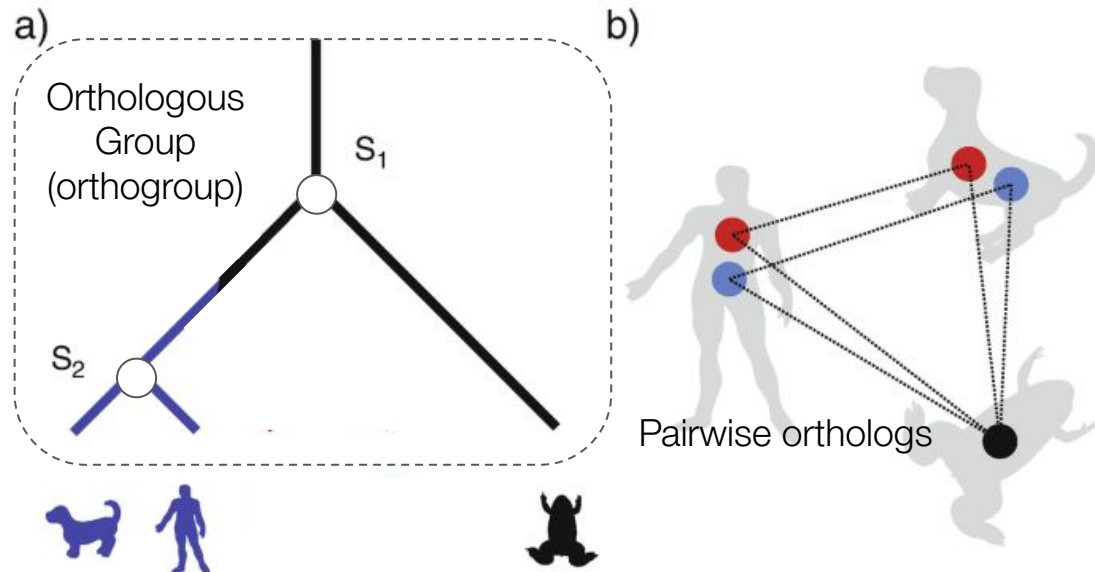
Taxon Sampling

2

Orthology Inference

For phylogenetic inference, our starting point are the Orthologous Groups, (OGs) since they are derived from speciation events

- **single copy OGs** (1:1 OGs: only one gene per species. Ready to continue with the next step!



Orthology relationships are inferred pairwise

How to infer a species tree

1

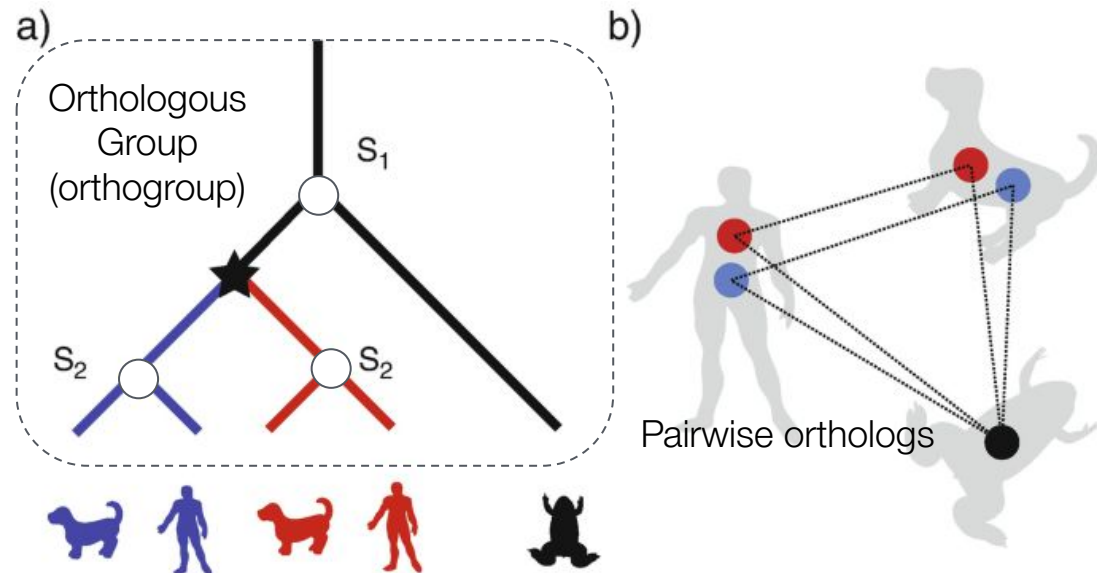
Taxon Sampling

2

Orthology Inference

For phylogenetic inference, our starting point are the Orthologous Groups, (OGs) since they are derived from speciation events

- single copy OGs (1:1 OGs: only one gene per species. Ready to continue with the next step!
- **OGs with duplicates** (1:many, many:many OGs): pruning is necessary to create your matrix



Orthology relationships are inferred pairwise

How to infer a species tree

1

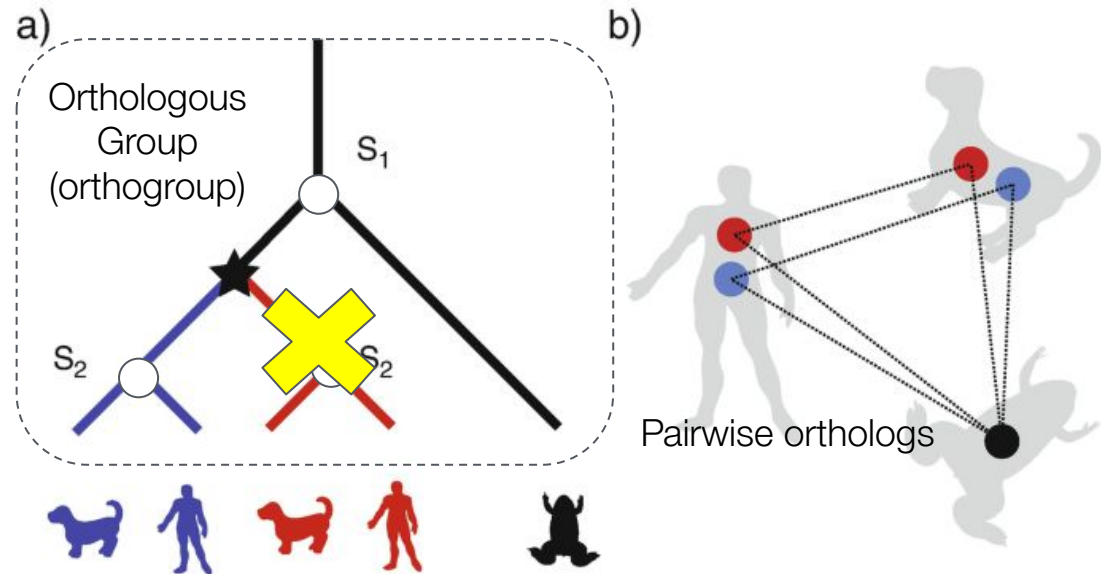
Taxon Sampling

2

Orthology Inference

For phylogenetic inference, our starting point are the Orthologous Groups, (OGs) since they are derived from speciation events

- single copy OGs (1:1 OGs: only one gene per species. Ready to continue with the next step!
- **OGs with duplicates** (1:many, many:many OGs): pruning is necessary to create your matrix



Orthology relationships are inferred pairwise

How to infer a species tree

1

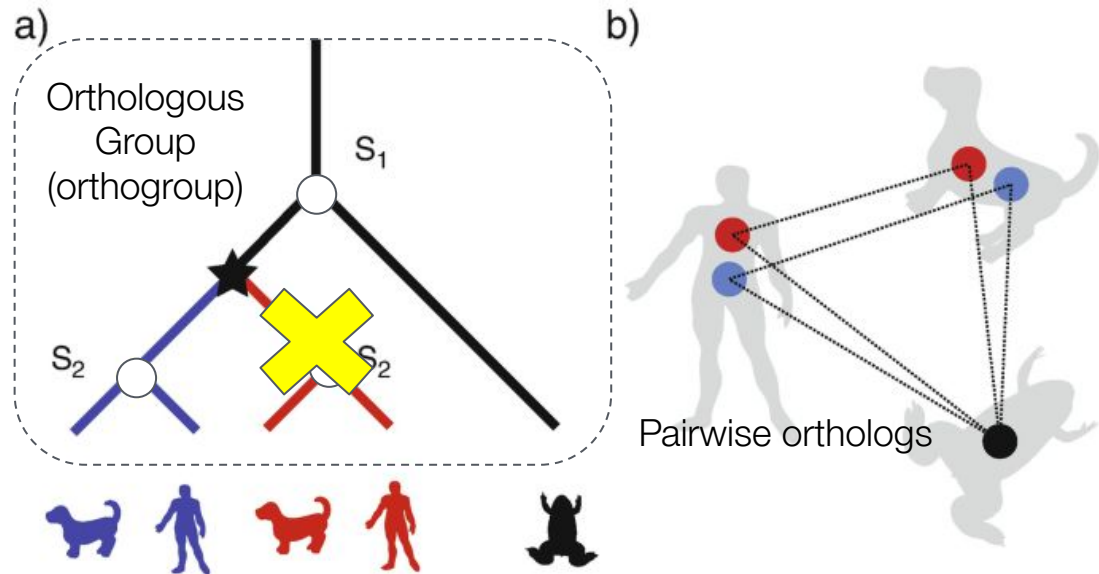
Taxon Sampling

2

Orthology Inference

For phylogenetic inference, our starting point are the Orthologous Groups, (OGs) since they are derived from speciation events

- single copy OGs (1:1 OGs: only one gene per species. Ready to continue with the next step!
- **OGs with duplicates** (1:many, many:many OGs): pruning is necessary to create your matrix



Key message: the selection of OGs for further analysis matters a lot

How to infer a species tree

Covered last week by Rayan

1

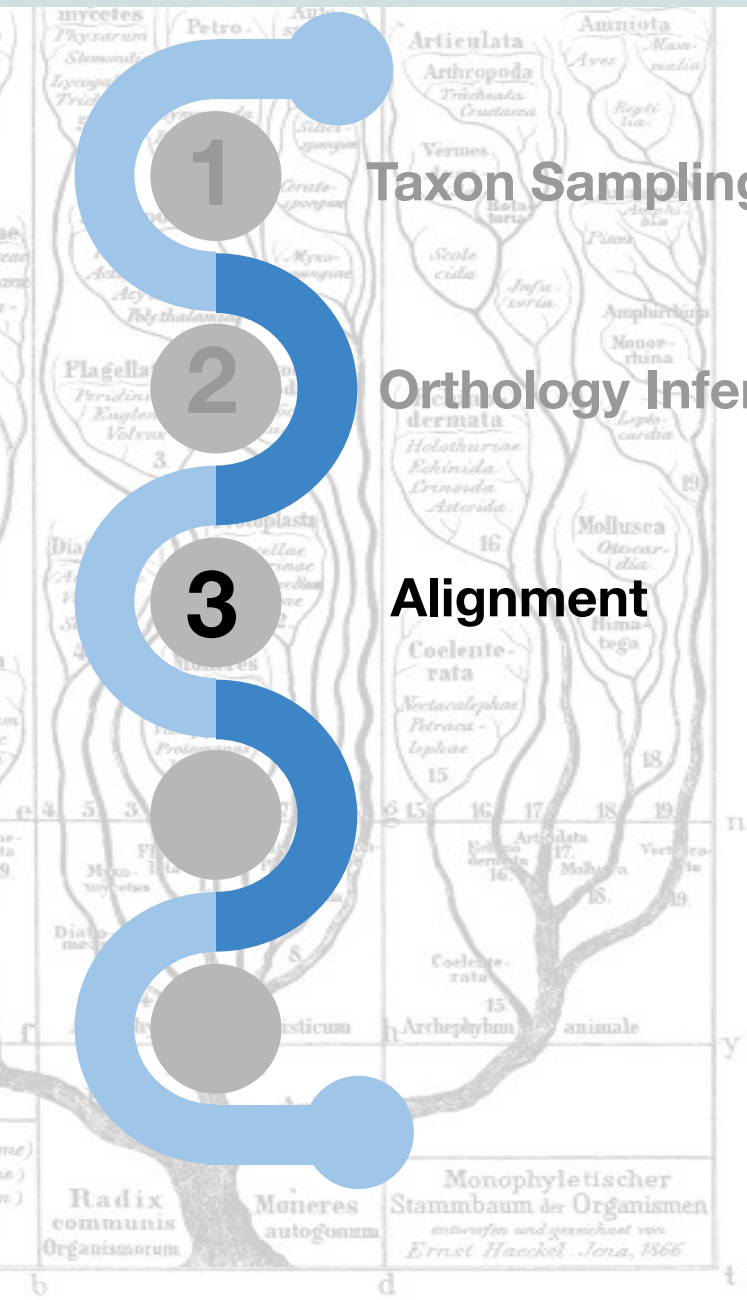
Taxon Sampling

2

Orthology Inference

3

Alignment



How to infer a species tree

Covered last week by Rayan

1

Taxon Sampling

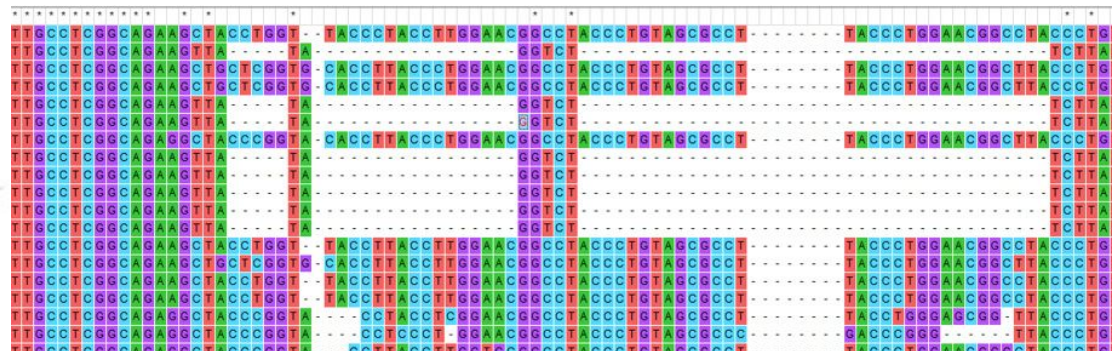
2

Orthology Inference

3

Alignment

If the sequences are poorly aligned, you may want to consider trimming the poorly aligned areas. There are several tools for it:



How to infer a species tree

Covered last week by Rayan

1

Taxon Sampling

2

Orthology Inference

3

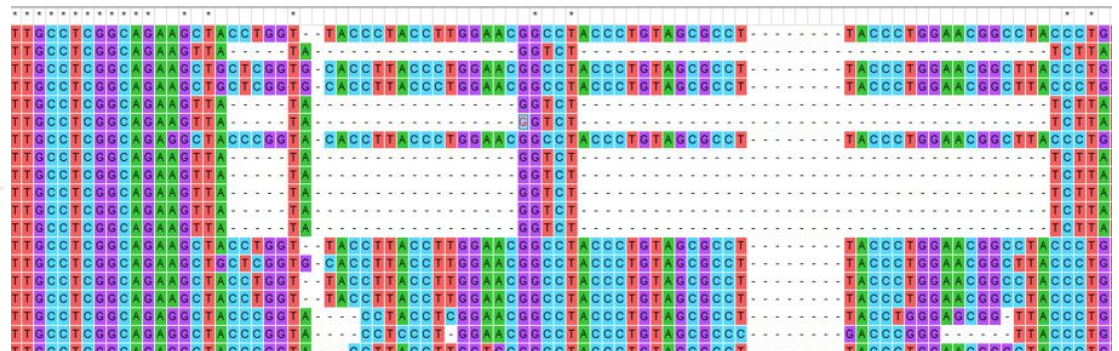
Alignment

If the sequences are poorly aligned, you may want to consider trimming the poorly aligned areas. There are several tools for it:



PREQUAL

(prealignment quality filter)



How to infer a species tree

1

Taxon Sampling

2

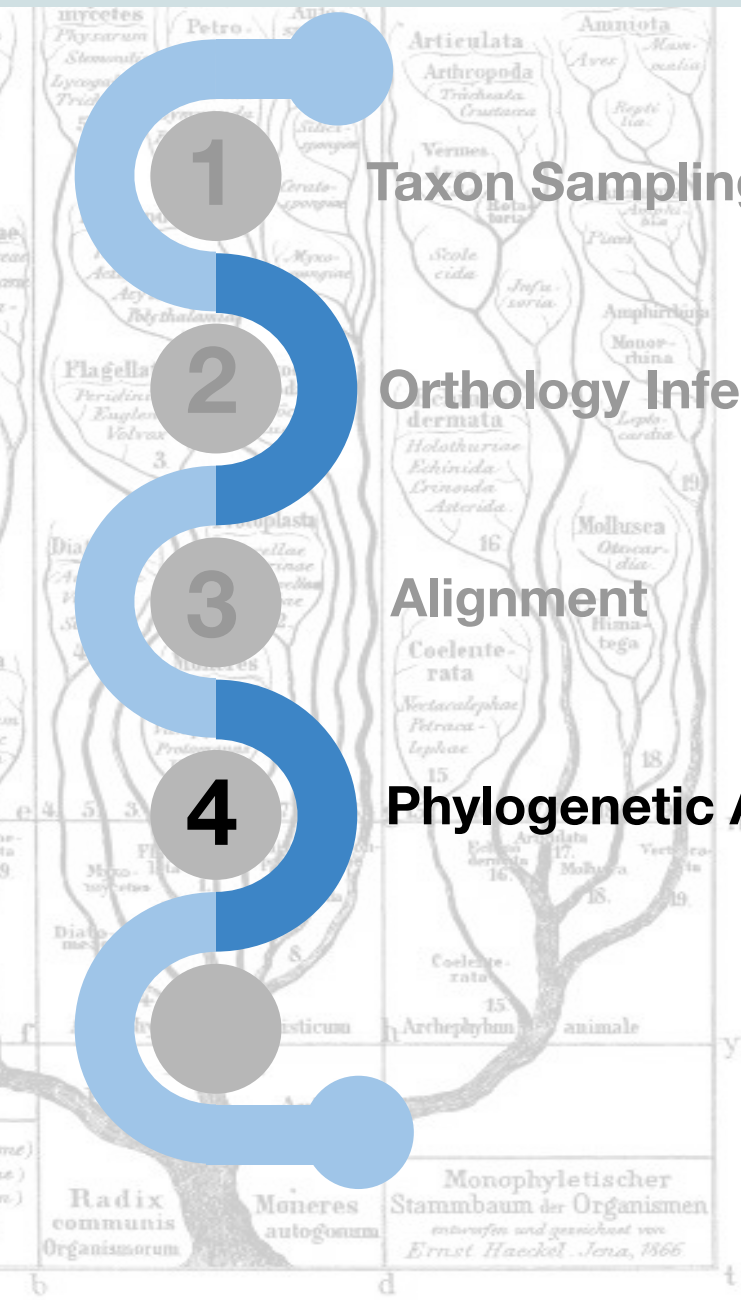
Orthology Inference

3

Alignment

4

Phylogenetic Analysis



How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

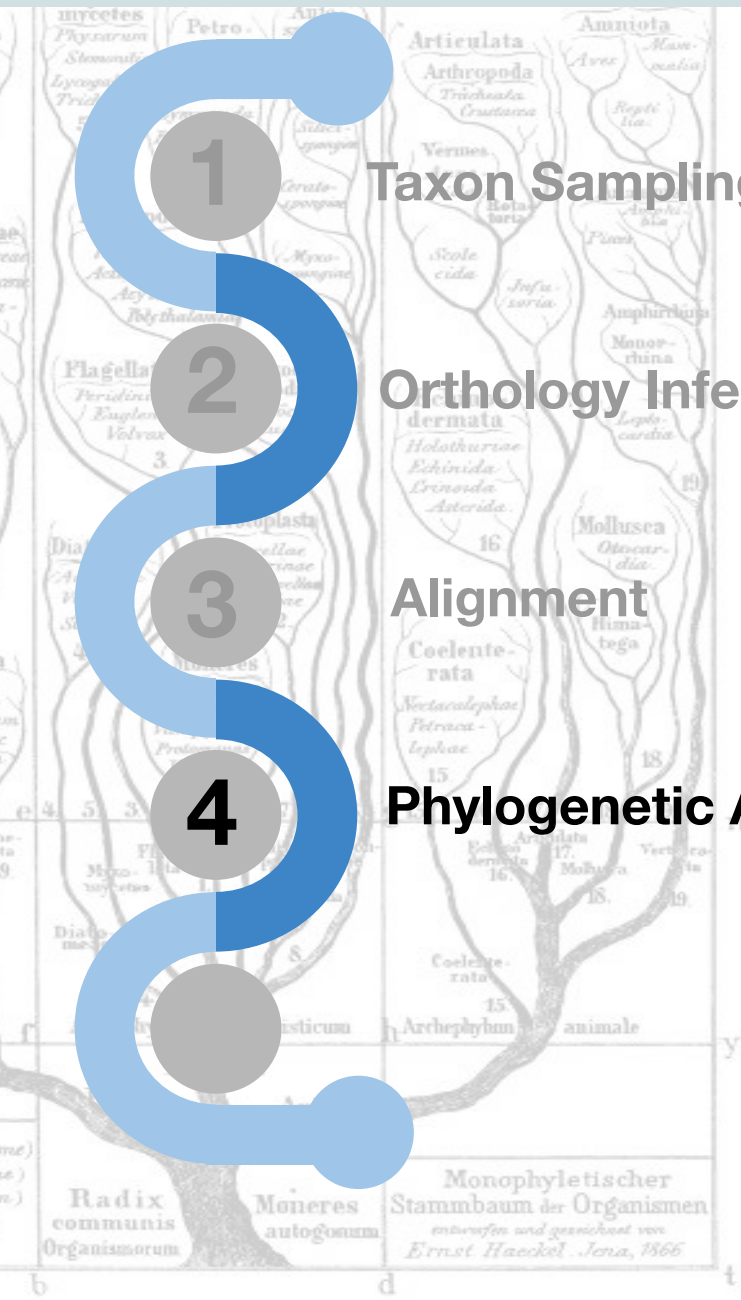
Alignment

4

Phylogenetic Analysis

```
AAATATATTCGGGTAAACCGAGATGAGAACGATGAGCCCATTTGAA  
AAATATATT. GGGTAACCGAGATGAGAACGATGAGCCCATTTGAA  
AAATATATTCGGGTAAACCGAGATGAGAACGATGAGCCCATTTGAA  
AAATATATTCGGGTAAACCGAGATGAGAACGATGAGCCCATTTGAA  
AAATATATTCGGGTAAACCGAGATGAGAACGATGAGCCCATTTGAA  
AAATATATTCGGGTAAACCGAGATGAGAACGATGAGCCCATTTGAA
```

DATA



How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylogenetic Analysis



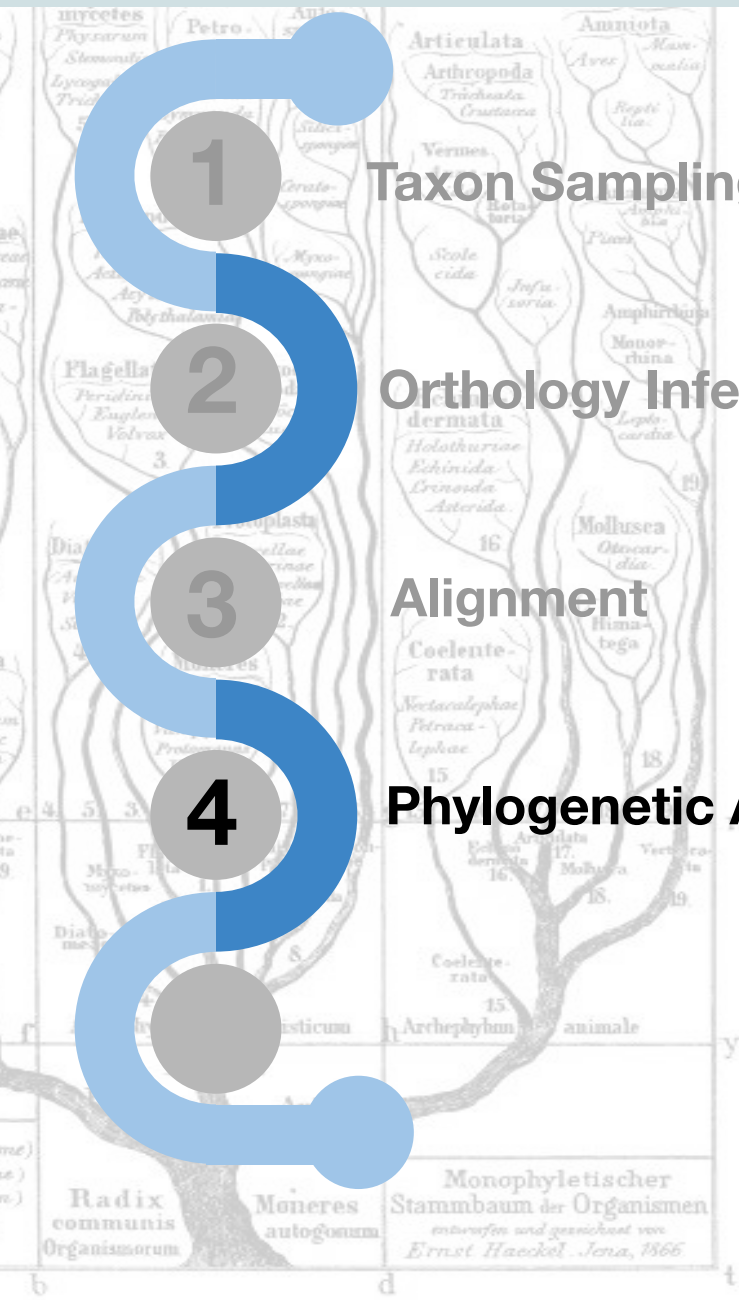
```
AATATATTCCGGTAACCGAGATGAGAACGATGAGCCCATTTGAA
AATATATT. GGGTAACCGAGATGAGAACGATGAGCCCATTTGAA
AATATATTCCGGTAACCGAGATGAGAACGATGAGCCCATTTGAA
AATATATTCCGGTAACCGAGATGAGAACGATGAGCCCATTTGAA
AATATATTCCGGTAACCGAGATGAGAACGATGAGCCCATTTGAA
```

DATA



MODEL OF EVOLUTION

describes the relative rates of different changes



How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylogenetic Analysis



```
AATATATTCCGGTAACCGAGATGAGAACGATGAGCCCATTTGAA
AATATATT. GGGTAACCGAGATGAGAACGATGAGCCCATTTGAA
AATATATTCCGGTAACCGAGATGAGAACGATGAGCCCATTTGAA
AATATATTCCGGTAACCGAGATGAGAACGATGAGCCCATTTGAA
AATATATTCCGGTAACCGAGATGAGAACGATGAGCCCATTTGAA
```

DATA



MODEL OF EVOLUTION

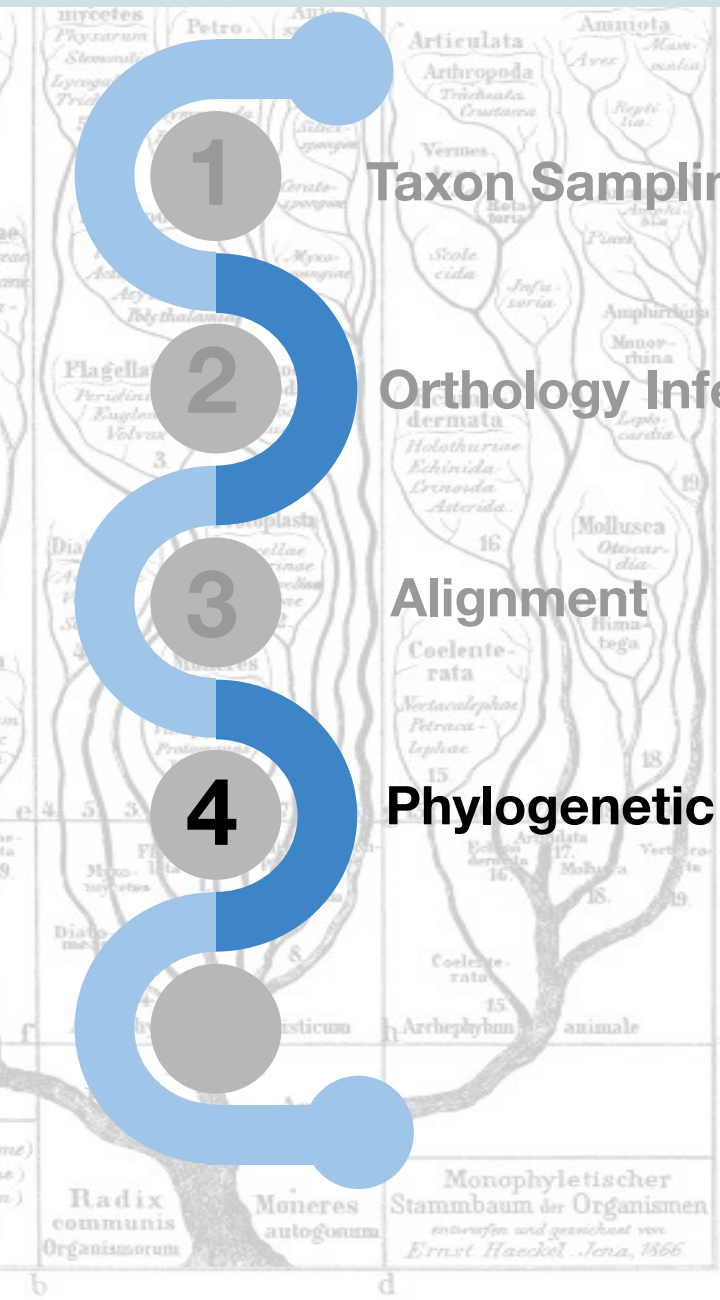
describes the relative rates of different changes

Seq1 ATGGCA

Seq2 ACGCCG

Seq3 AGGGCC

How to infer a species tree



1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylogenetic Analysis



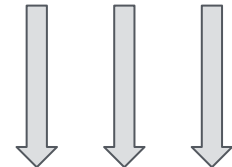
DATA



MODEL OF EVOLUTION

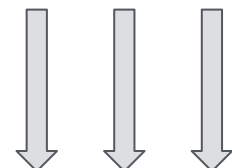
describe the relative rates of different changes

Seq1 ATGGCA



3 changes

Seq2 ACGCCG



3 changes

Seq3 AGGGCC

2 changes

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylogenetic Analysis



DATA



MODEL OF EVOLUTION

describe the relative rates of different changes

Seq1 ATGGCA

Seq2 ACGCCG

Seq3 AGGGCC

3 changes

3 changes

Jukes and
Cantor 1969
(JC69)

2 changes

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylogenetic Analysis

AATATATTCCGGTAACCGAGATGAGAACGATGAGCCCATTTGAA
AATATATT. GGGTAACCGAGATGAGAACGATGAGCCCATTTGAA
AATATATTCCGGTAACCGAGATGAGAACGATGAGCCCATTTGAA
AATATATTCCGGTAACCGAGATGAGAACGATGAGCCCATTTGAA
AATATATTCCGGTAACCGAGATGAGAACGATGAGCCCATTTGAA
AATATATTCCGGTAACCGAGATGAGAACGATGAGCCCATTTGAA

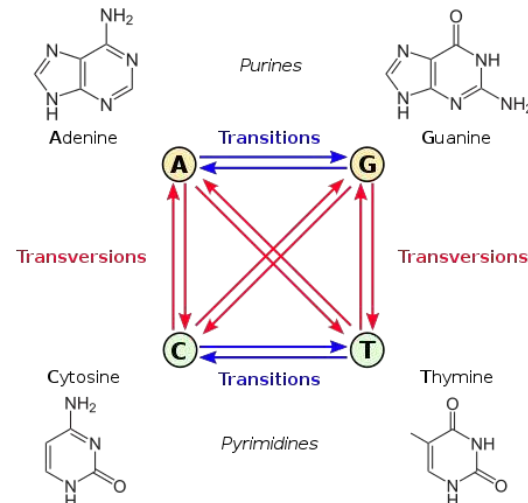
DATA



MODEL OF EVOLUTION

describes the relative rates of different changes

Eg, mutational biases and purifying selection favoring conservative changes are probably responsible for the relatively high rate of **transitions** compared to **transversions** in evolving sequences



How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylogenetic Analysis



DATA



MODEL OF EVOLUTION

describes the relative rates of different changes

Eg, mutational biases and purifying selection favoring conservative changes are probably responsible for the relatively high rate of **transitions** compared to **transversions** in evolving sequences

Seq1 ATGGCA

Seq2 ACGCCG

Seq3 AGGGCC

3 changes (1 transition, 2 transversions)

3 changes (3 transversions)

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylogenetic Analysis



DATA



MODEL OF EVOLUTION

describes the relative rates of different changes

Eg, mutational biases and purifying selection favoring conservative changes are probably responsible for the relatively high rate of **transitions** compared to **transversions** in evolving sequences

Seq1 ATGGCA

Seq2 ACGCCG

Seq3 AGGGCC

3 changes (1 transition, 2 transversions)

3 changes (3 transversions)

Kimura 1980 (K80)

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylogenetic Analysis



DATA

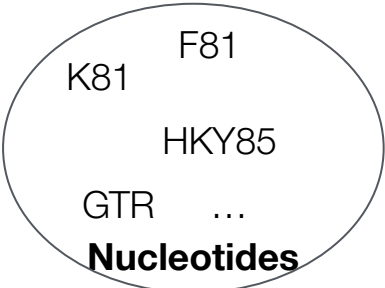


MODEL OF EVOLUTION

describes the relative rates of different changes

Eg, mutational biases and purifying selection favoring conservative changes are probably responsible for the relatively high rate of **transitions** compared to **transversions** in evolving sequences

There are many models that take into account other factors that can influence the rate of changes, eg time, reversibility, etc.



F81
K81
HKY85
GTR ...
Nucleotides

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylogenetic Analysis



DATA



MODEL OF EVOLUTION

describes the relative rates of different changes

Eg, mutational biases and purifying selection favoring conservative changes are probably responsible for the relatively high rate of **transitions** compared to **transversions** in evolving sequences

There are many models that take into account other factors that can influence the rate of changes, eg time, reversibility, etc.

K81 F81

HKY85

GTR ...

Nucleotides

PAM

JTT

WAG

LG

...

Amino Acids

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylogenetic Analysis



DATA



MODEL OF EVOLUTION

describes the relative rates of different changes

Eg, mutational biases and purifying selection favoring conservative changes are probably responsible for the relatively high rate of **transitions** compared to **transversions** in evolving sequences

K81 F81

HKY85

GTR ...

Nucleotides

PAM

JTT

WAG

LG

...

Amino Acids

} Same model for the complete gene / partition

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylogenetic Analysis



DATA



MODEL OF EVOLUTION

describe the relative rates of different changes

Eg, mutational biases and purifying selection favoring conservative changes are probably responsible for the relatively high rate of **transitions** compared to **transversions** in evolving sequences

Mixture models

C10 C20
C60 CAT

per-site
variation of
the model

F81
HKY85
GTR ...

Nucleotides

PAM
JTT
WAG LG ...

Amino Acids

Same model for
the complete
gene / partition

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylogenetic Analysis



DATA



MODEL OF EVOLUTION

describe the relative rates of different changes

Eg, mutational biases and purifying selection favoring conservative changes are probably responsible for the relatively high rate of **transitions** compared to **transversions** in evolving sequences

So... how do I select a model for my data?

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylogenetic Analysis



DATA



MODEL OF EVOLUTION

describe the relative rates of different changes

Eg, mutational biases and purifying selection favoring conservative changes are probably responsible for the relatively high rate of **transitions** compared to **transversions** in evolving sequences

So... how do I select a model for my data?

Don't worry, most phylogenetic programs have a tool to infer the model that better fits your data :-)

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylogenetic Analysis



DATA



MODEL OF EVOLUTION

describe the relative rates of different changes

Eg, mutational biases and purifying selection favoring conservative changes are probably responsible for the relatively high rate of **transitions** compared to **transversions** in evolving sequences

So... how do I select a model for my data?

Don't worry, most phylogenetic programs have a tool to infer the model that better fits your data :-)

If you're dealing with a difficult phylogenetic problem, mixture models are probably a good idea

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylogenetic Analysis



DATA



MODEL OF EVOLUTION

describe the relative rates of different changes

Eg, mutational biases and purifying selection favoring conservative changes are probably responsible for the relatively high rate of **transitions** compared to **transversions** in evolving sequences

So... how do I select a model for my data?

Key message: the selection of the model matters a lot

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

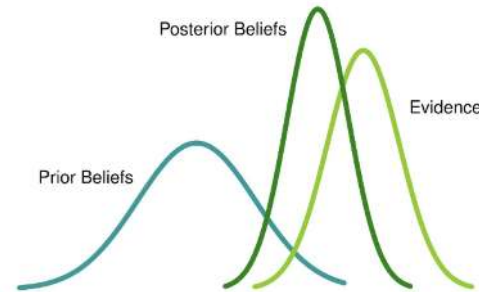
Phylo. Analysis



DATA + MODEL OF EVOLUTION



METHOD



How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

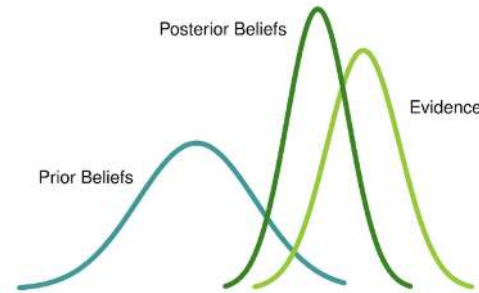
3

Alignment

4

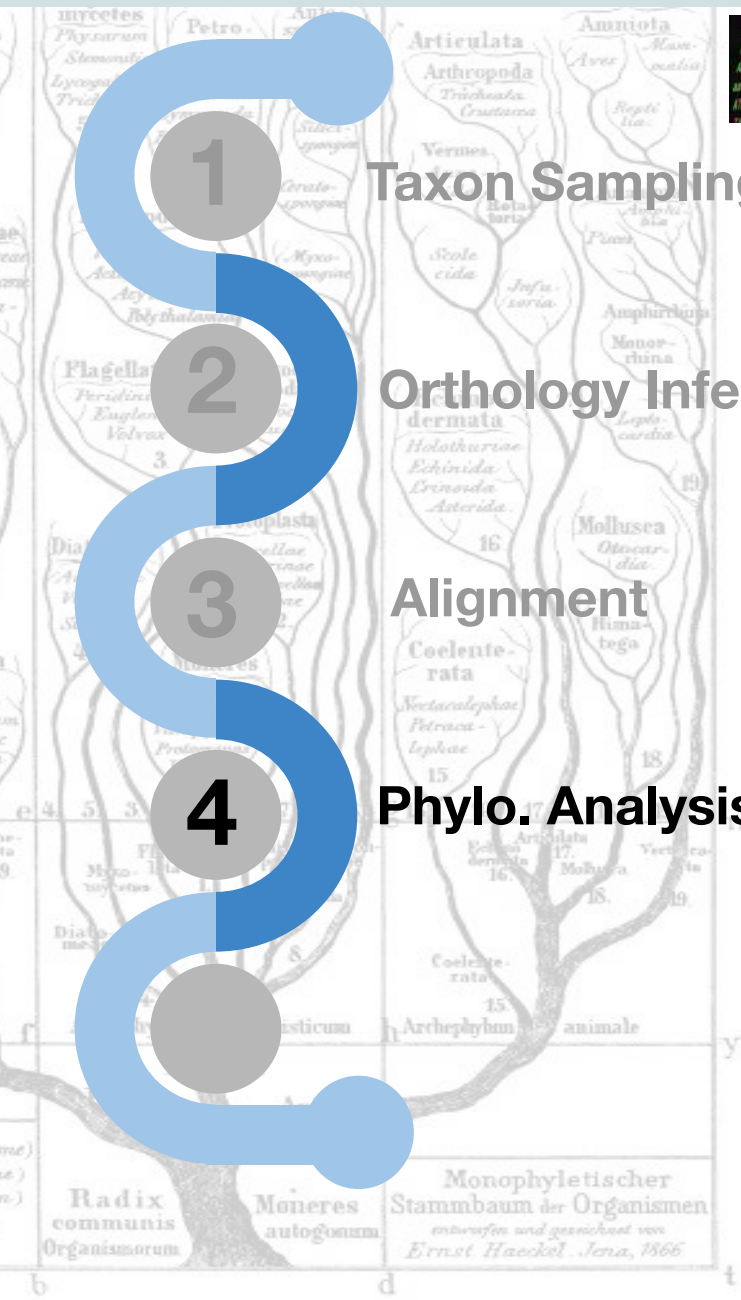
Phylo. Analysis

DATA + MODEL OF EVOLUTION + METHOD



Two main methods:

Maximum Likelihood (ML) and **Bayesian Inference (BI)**



How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

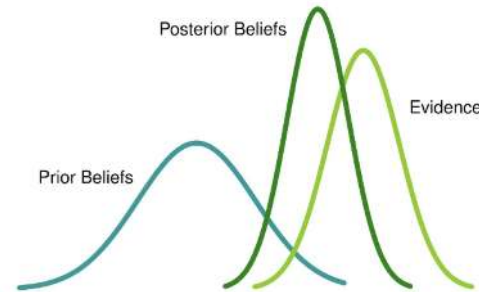
3

Alignment

4

Phylo. Analysis

DATA + MODEL OF EVOLUTION + METHOD



Two main methods:

Maximum Likelihood (ML) and **Bayesian Inference (BI)**

Basic question in BI:

'What is the probability that this model (T) is correct, given the data (D) that we have observed?'

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

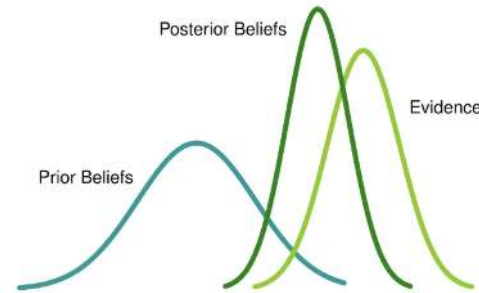
3

Alignment

4

Phylo. Analysis

DATA + MODEL OF EVOLUTION + METHOD



Two main methods:

Maximum Likelihood (ML) and **Bayesian Inference (BI)**

Basic question in BI:

'What is the probability that this model (T) is correct, given the data (D) that we have observed?'

Basic question in ML:

'What is the probability of seeing the observed data (D) given that a certain model (T) is true?'

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

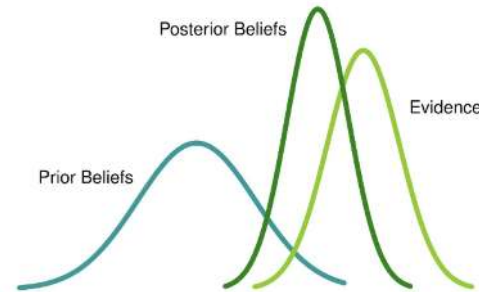
Phylo. Analysis



DATA + MODEL OF EVOLUTION



METHOD



Two main methods:

Maximum Likelihood (ML) and **Bayesian Inference (BI)**

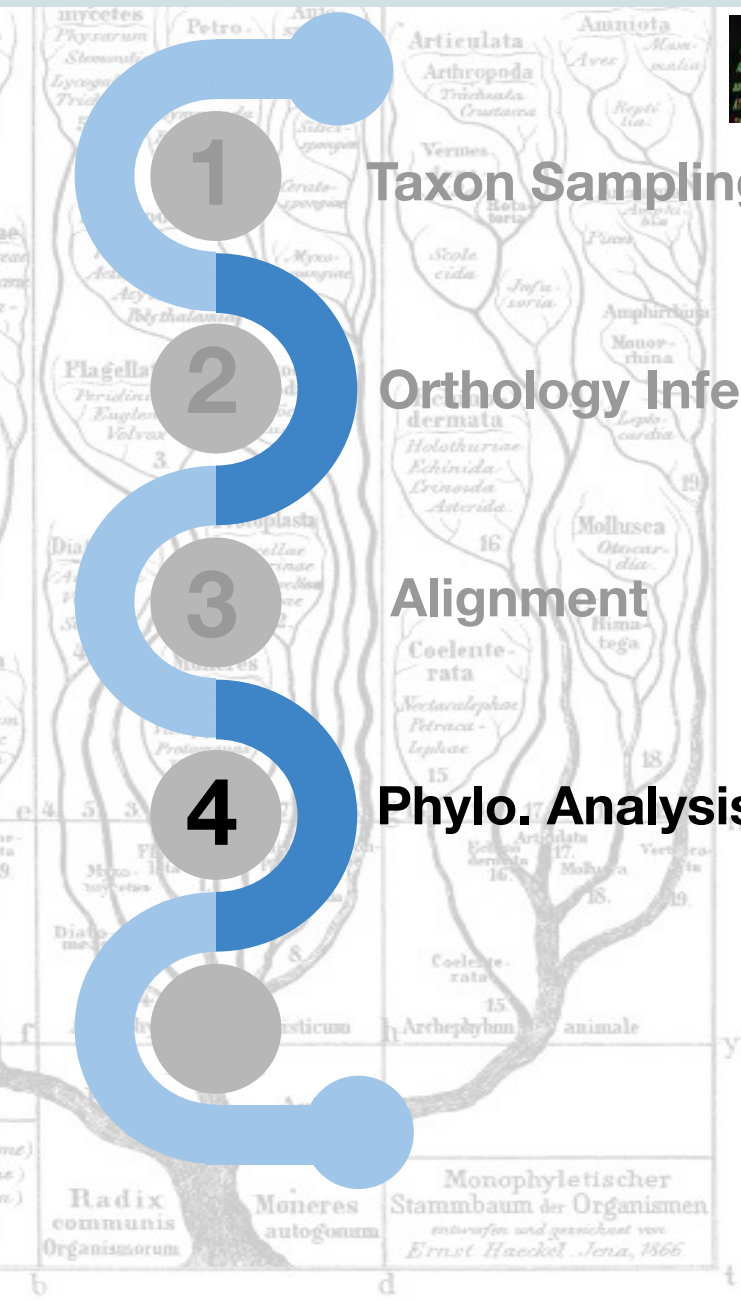
Basic question in BI:

'What is the probability that this model (T) is correct, given the data (D) that we have observed?'

Basic question in ML:

'What is the probability of seeing the observed data (D) given that a certain model (T) is true?'

BI seeks $P(T|D)$, while ML maximizes $P(D|T)$



How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

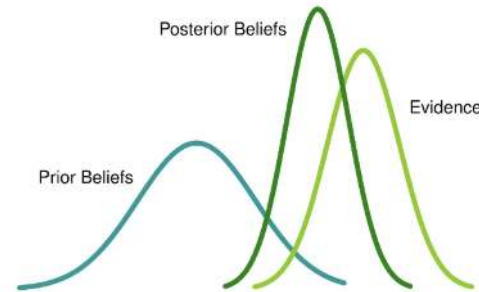
Phylo. Analysis



DATA + MODEL OF EVOLUTION



METHOD



Two main methods:

Maximum Likelihood (ML) and **Bayesian Inference (BI)**

RAxML



IQ-TREE



RevBayes

fasrc/phylobayes

A Bayesian Monte Carlo Markov Chain (MCMC) sampler for phylogenetic reconstruction and molecular dating.

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

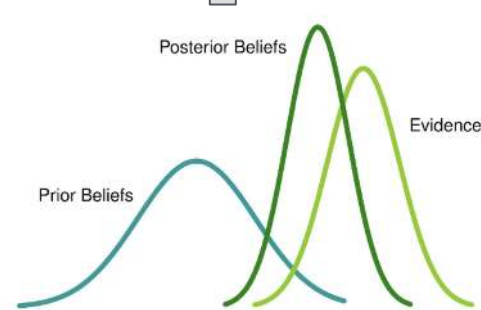
Alignment

4

Phylo. Analysis



DATA + MODEL OF EVOLUTION



+
METHOD

Two main methods:

Maximum Likelihood (ML) and **Bayesian Inference (BI)**

Which one should I choose?

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

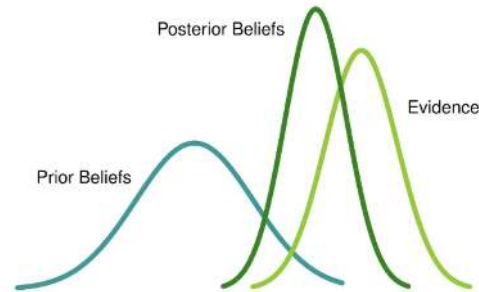
3

Alignment

4

Phylo. Analysis

DATA + MODEL OF EVOLUTION + METHOD



Two main methods:

Maximum Likelihood (ML) and **Bayesian Inference (BI)**

Which one should I choose?



How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

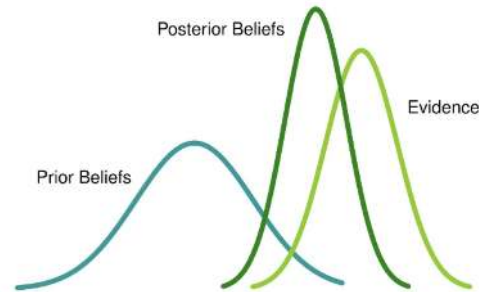
3

Alignment

4

Phylo. Analysis

DATA + MODEL OF EVOLUTION + METHOD



Two main methods:

Maximum Likelihood (ML) and **Bayesian Inference (BI)**

Which one should I choose?



Factors to consider: running time, availability of 'complex' models, etc.

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

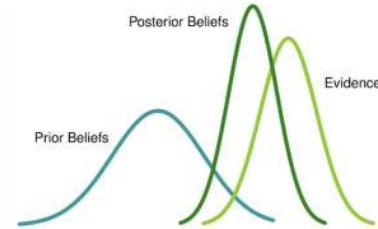
Alignment

4

Phylo. Analysis



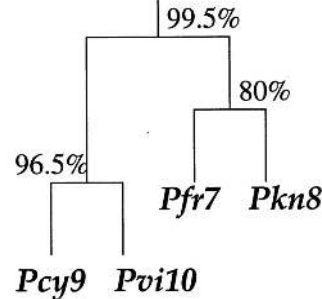
DATA + MODEL OF EVOLUTION



+
METHOD



A WAY TO ASSESS
HOW GOOD OUR
HYPOTHESIS IS



How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

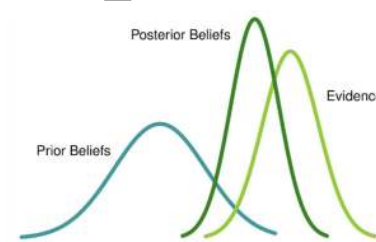
Alignment

4

Phylo. Analysis

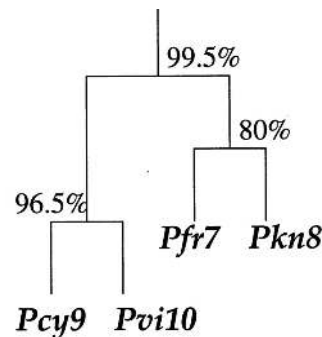


DATA + MODEL OF EVOLUTION



+
METHOD
+

**A WAY TO ASSESS
HOW GOOD OUR
HYPOTHESIS IS**



- ML: standard nonparametric bootstrap (100 reps), approximate likelihood ratio test (1,000 reps), ultrafast bootstrap (1,000 reps)(between 1 and 100)
 - you 'believe' in a clade with > 80-90% bootstrap support and/or ultrafast bootstrap > 95% and/or approx. LRT > 80-90%.
- BI: posterior probability (between 0 and 1)
 - you 'believe' in a clade with > 0.9 pp

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

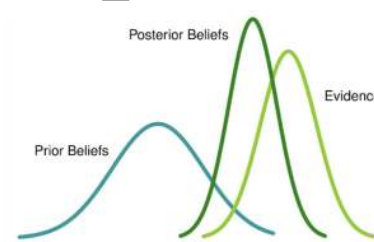
Alignment

4

Phylo. Analysis

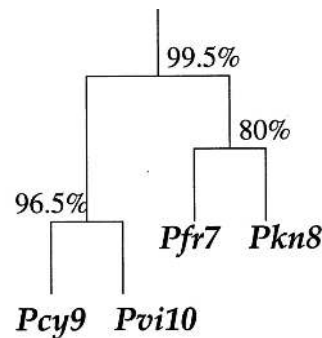


DATA + MODEL OF EVOLUTION



+
METHOD
+

**A WAY TO ASSESS
HOW GOOD OUR
HYPOTHESIS IS**



- New metrics (highly recommended in large matrices):
 - [concordance factor](#) (IQ-TREE): for every branch of a reference tree, the concordance factor is defined as the percentage of “decisive” gene trees containing that branch.
 - [internode certainty](#) (RAxML): 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](#) (RAxML): the sum of all the internode certainty across all internodes.

How to infer a species tree

- 1 Taxon Sampling
- 2 Orthology Inference
- 3 Alignment
- 4 Phylo. Analysis
- 5 Sensitivity Analysis

Lies, damn lies and phylogenomics

**Lies, damn lies and
phylogenomics**

Taxon Sampling

Orthology Inference

Alignment

Phylo. Analysis

Sensitivity Analysis

How to infer a species tree

1

Taxon Sampling

2

Orthology Inference

3

Alignment

4

Phylo. Analysis

5

Sensitivity Analysis

**Lies, damn lies and
phylogenomics**

inspired by

**Lies, damn lies, and
genomics**

you, your data, your perceptions and
reality

Christopher West Wheat



**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
does not mean that you
can trust your tree!!**

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

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

Wait, what?? And WHY is that?

Understanding your data (and the errors it may trigger in downstream analyses)

Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

Missing data

Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

Missing data

	Gene 1	Gene 2	Gene 3		Gene n
Species A					
Species B					
Species C					
Species D					

Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

Missing data

	Gene 1	Gene 2	Gene 3		Gene n
Species A					
Species B					
Species C					
Species D					

“Gruyère effect”



Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

Missing data

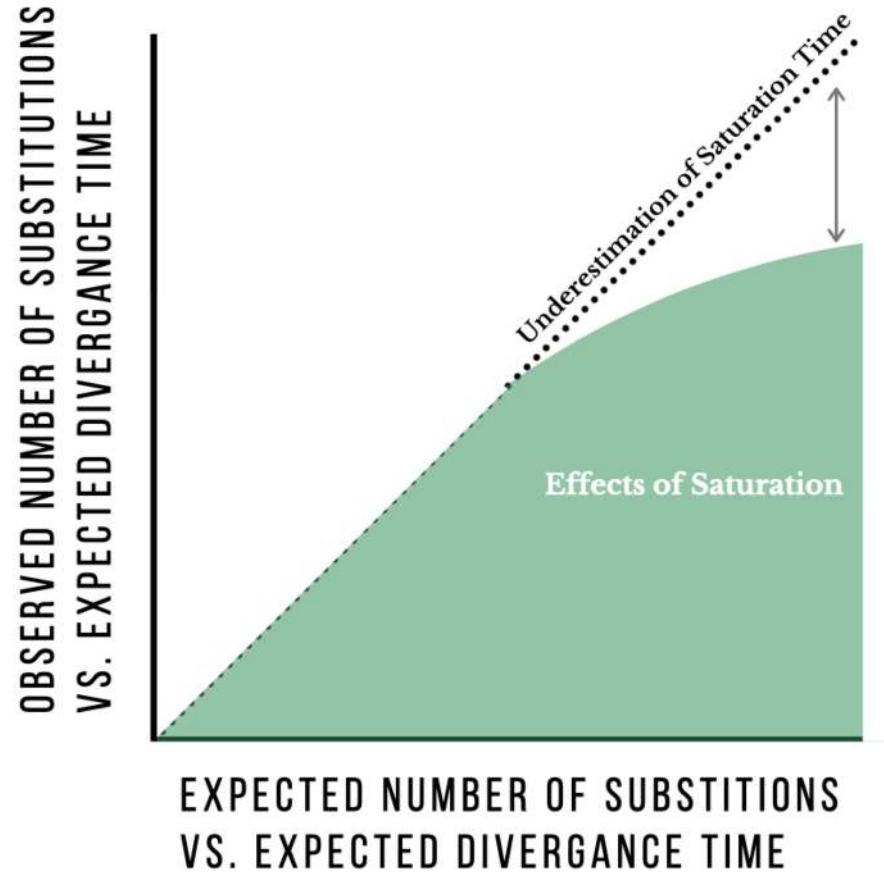
Saturation

Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

Missing data

Saturation



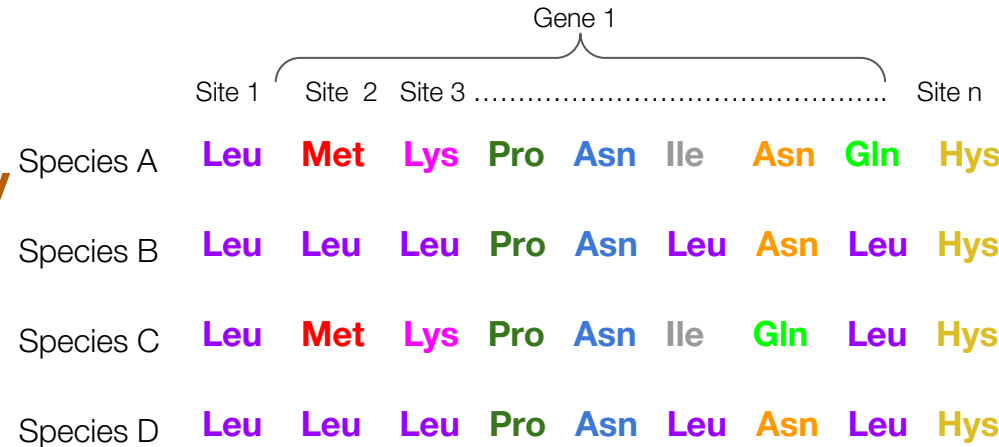
Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity



Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity

	Site 1	Gene 1							Site n
	Site 1	Site 2	Site 3						Site n
Species A	Leu	Met	Lys	Pro	Asn	Ile	Asn	Gln	Hys
Species B	Leu	Leu	Leu	Pro	Asn	Leu	Asn	Leu	Hys
Species C	Leu	Met	Lys	Pro	Asn	Ile	Gln	Leu	Hys
Species D	Leu	Leu	Leu	Pro	Asn	Leu	Asn	Leu	Hys

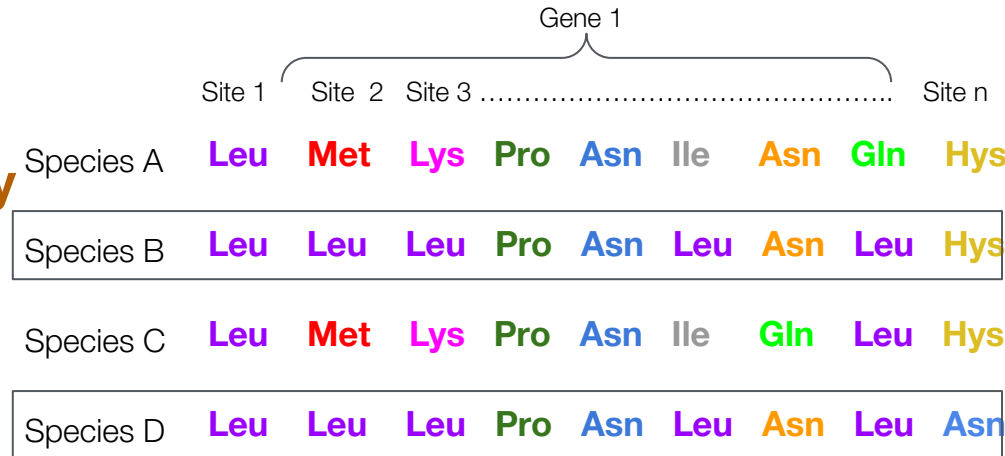
Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity



Understanding your data (and the errors it may trigger in downstream analyses)

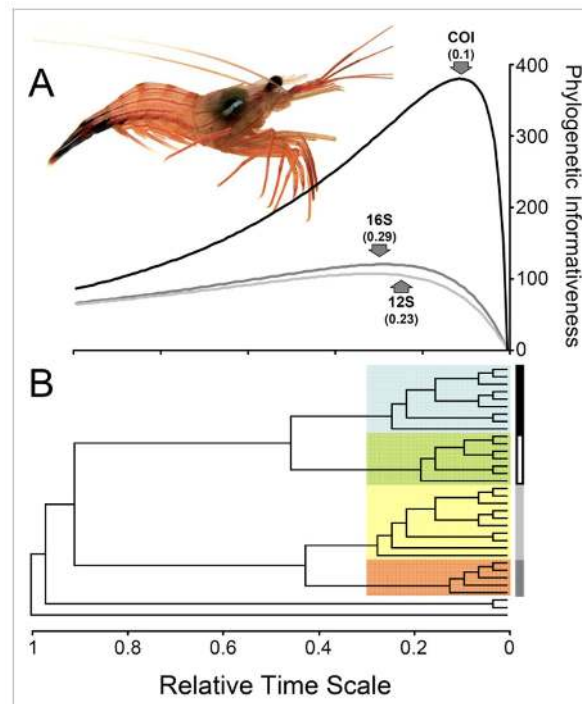
(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest



Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

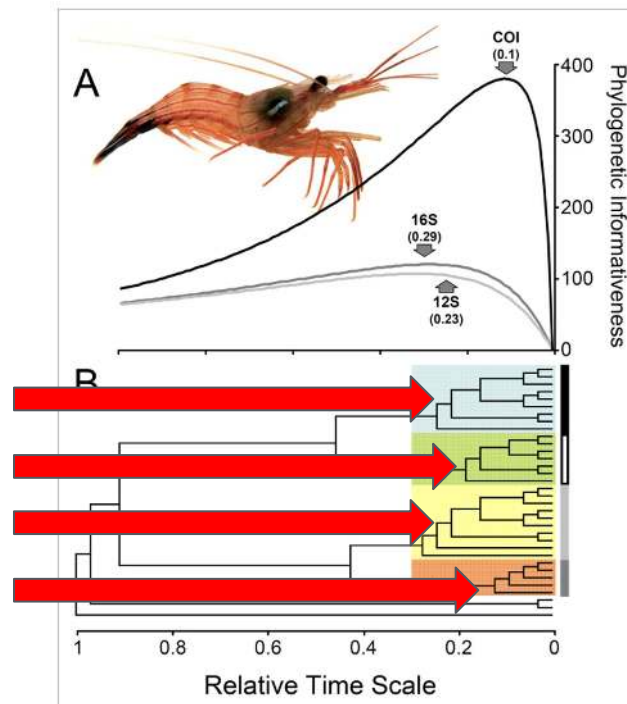
Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest

Good information to resolve these nodes



Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

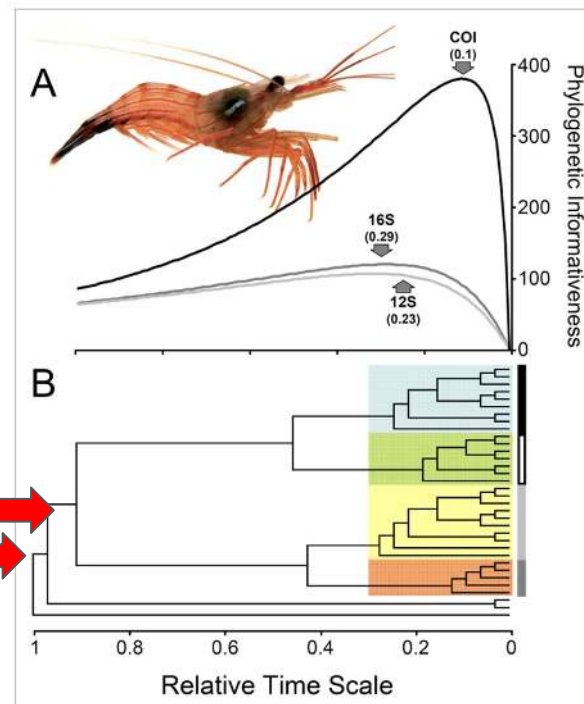
Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest

Not enough information
to resolve these nodes



Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest

etc.

Understanding your data (and the errors it may trigger in downstream analyses)

(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 species tree

Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity

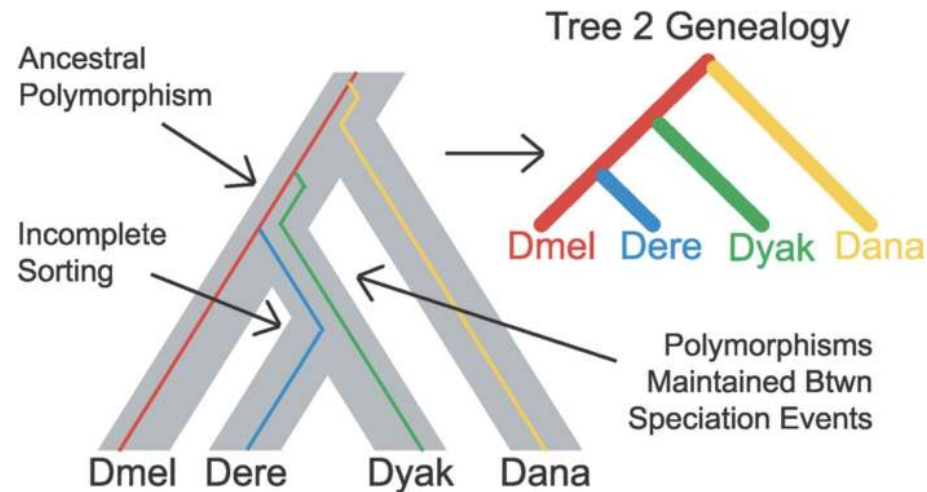
Lack of phylogenetic signal for your node of interest

etc.

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

Incomplete lineage sorting

Understanding your data (and the errors it may trigger in downstream analyses)



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 species tree

Incomplete lineage sorting

Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest

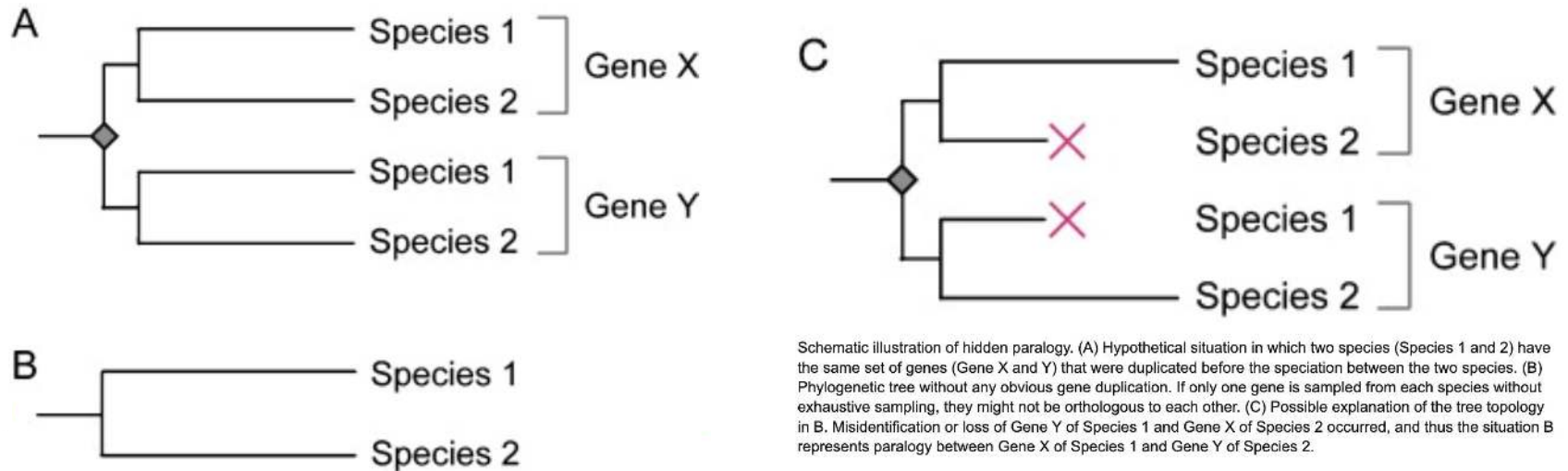
etc.

(2) **Conflict between individual gene trees and the species tree**

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Understanding your data (and the errors it may trigger in downstream analyses)



Kuraku 2013

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

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Understanding your data (and the errors it may trigger in downstream analyses)

(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 species tree**

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

Understanding your data (and the errors it may trigger in downstream analyses)

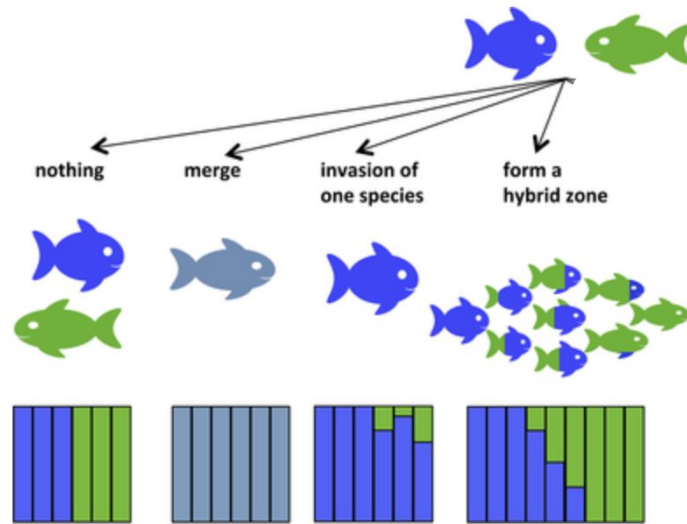


Fig 2.

Schematic representation of homoploid and allopolyploid hybrid speciation.

Runemark et al. 2019

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

Incomplete lineage sorting

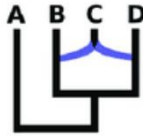
Gene loss (eg, hidden paralogy)

Hybridization

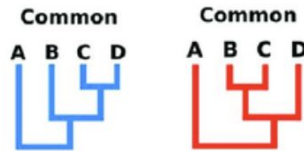
Understanding your data (and the errors it may trigger in downstream analyses)

Hybrid speciation

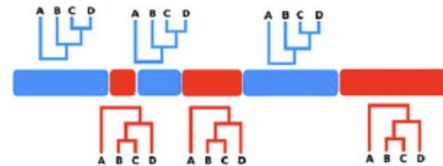
1. Species network



2. Contained species trees



3. Genomic histories



Folk et al. 2018

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

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

Understanding your data (and the errors it may trigger in downstream analyses)

(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 species tree**

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

Introgression

Understanding your data (and the errors it may trigger in downstream analyses)

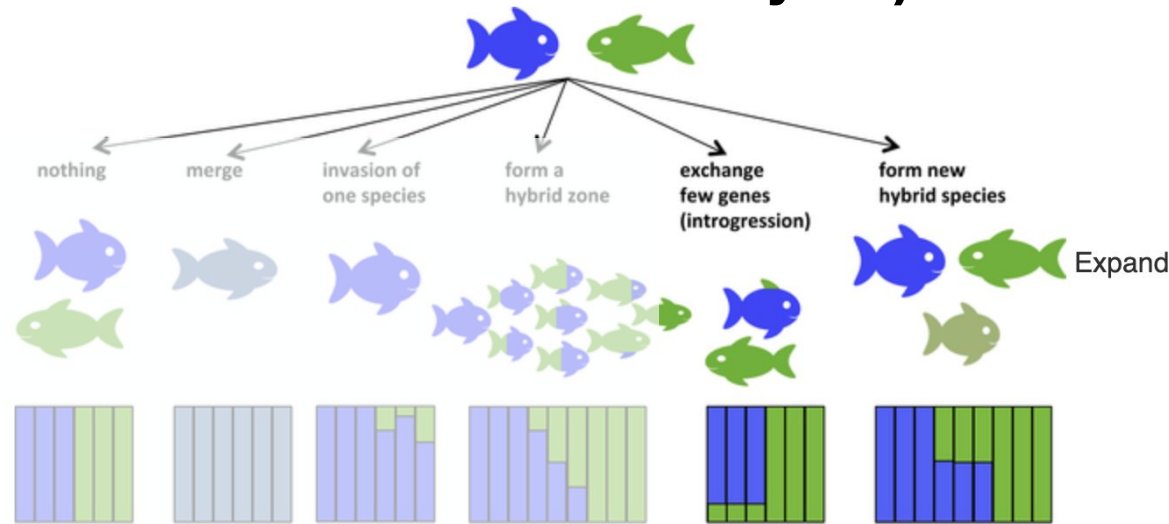


Fig 2.

Schematic representation of homoploid and allopolyploid hybrid speciation.

Runemark et al. 2019

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

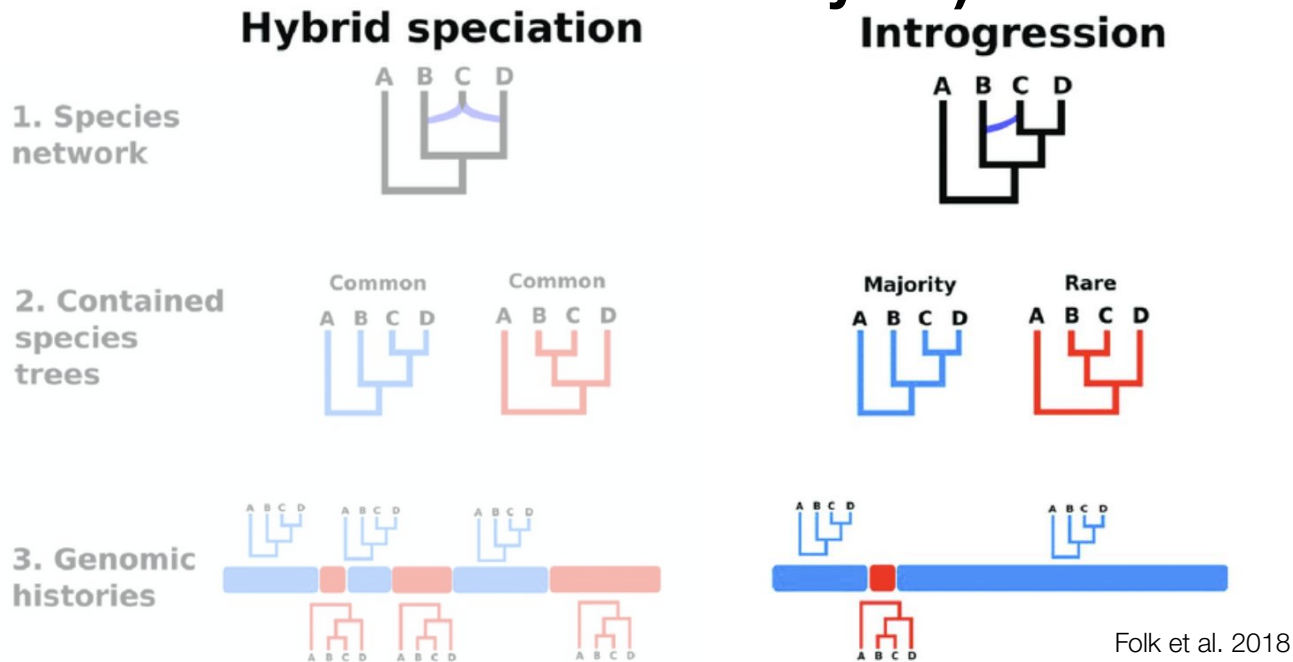
Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

Introgression

Understanding your data (and the errors it may trigger in downstream analyses)



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

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

Introgression

Understanding your data (and the errors it may trigger in downstream analyses)

(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 species tree**

Incomplete lineage sorting

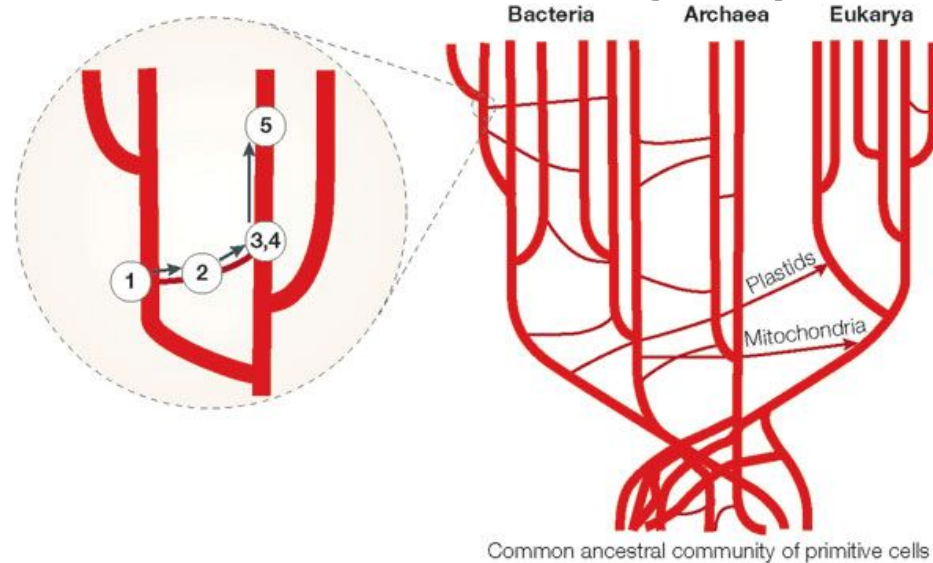
Gene loss (eg, hidden paralogy)

Hybridization

Introgression

Horizontal gene transfer

Understanding your data (and the errors it may trigger in downstream analyses)



Copyright © 2005 Nature Publishing Group
Nature Reviews | Microbiology

Smets and Barkay 2005

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

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

Introgression

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

Systematic Error vs Random Error

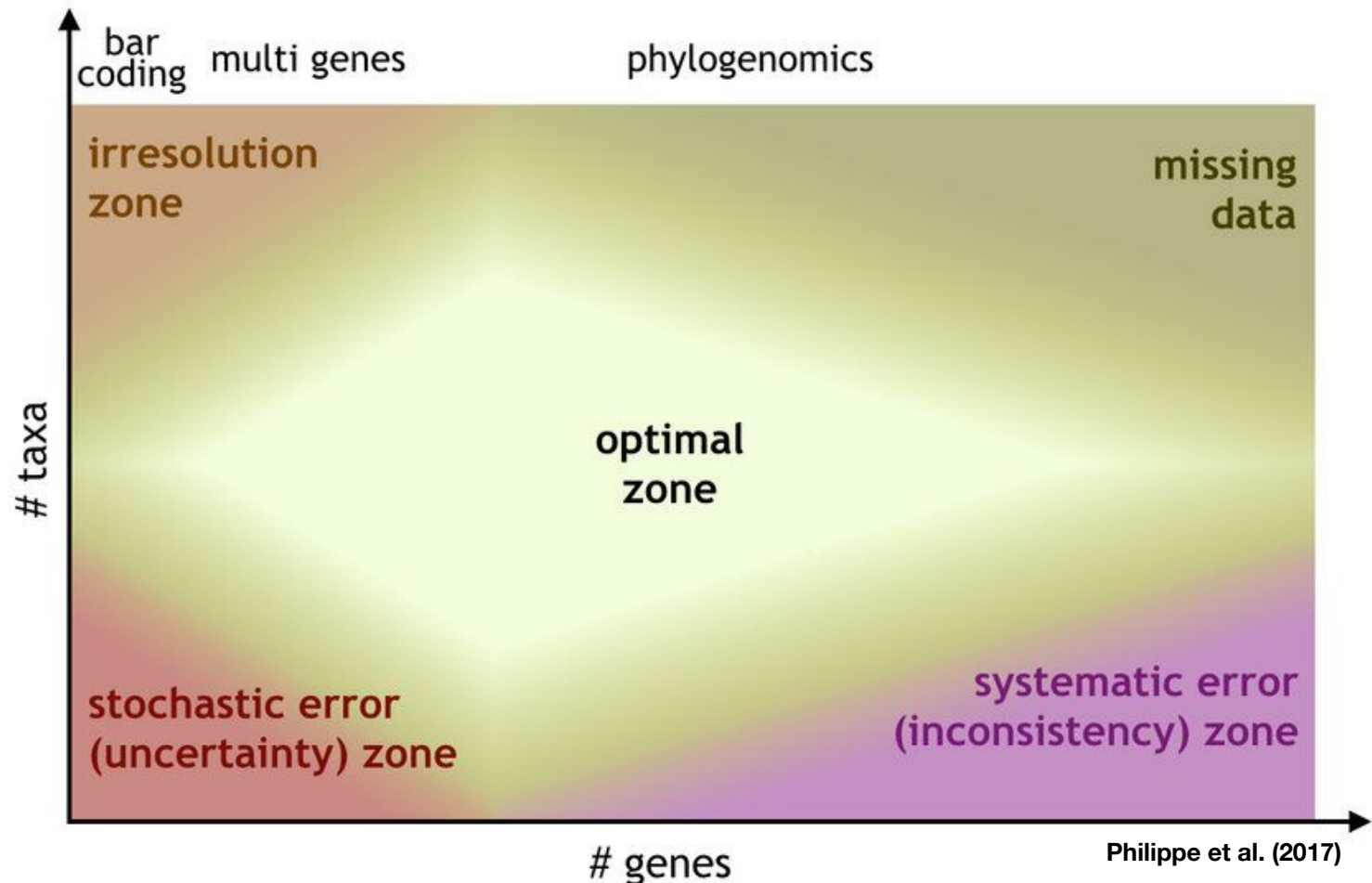
Systematic Error Measurements may be precise, but not accurate.  A target with five concentric red and white rings. Four black dots are clustered tightly together in the lower-left quadrant, representing high precision but low accuracy. • Using a stretched measuring tape • Scale that always reads too high or low • Reading an indicator from a poor angle	Random Error Measurements lack precision, but cluster around accurate value.  A target with five concentric red and white rings. Four black dots are scattered across the target, but they are generally clustered around the center, representing low precision but high accuracy. • Timing depends on reaction time • People take turns taking readings • Rounding values up or down.
---	--

sciencenotes.org

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
- 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 heterogeneity 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 heterogeneity 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 heterogeneity in the gene sequence to correctly infer/apply a substitution model

Eg 2, no recombination

Eg 3, genes evolved through duplication and not through speciation

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

Because of:

1) Systematic error

2) Model violation

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

Eg 2, no recombination

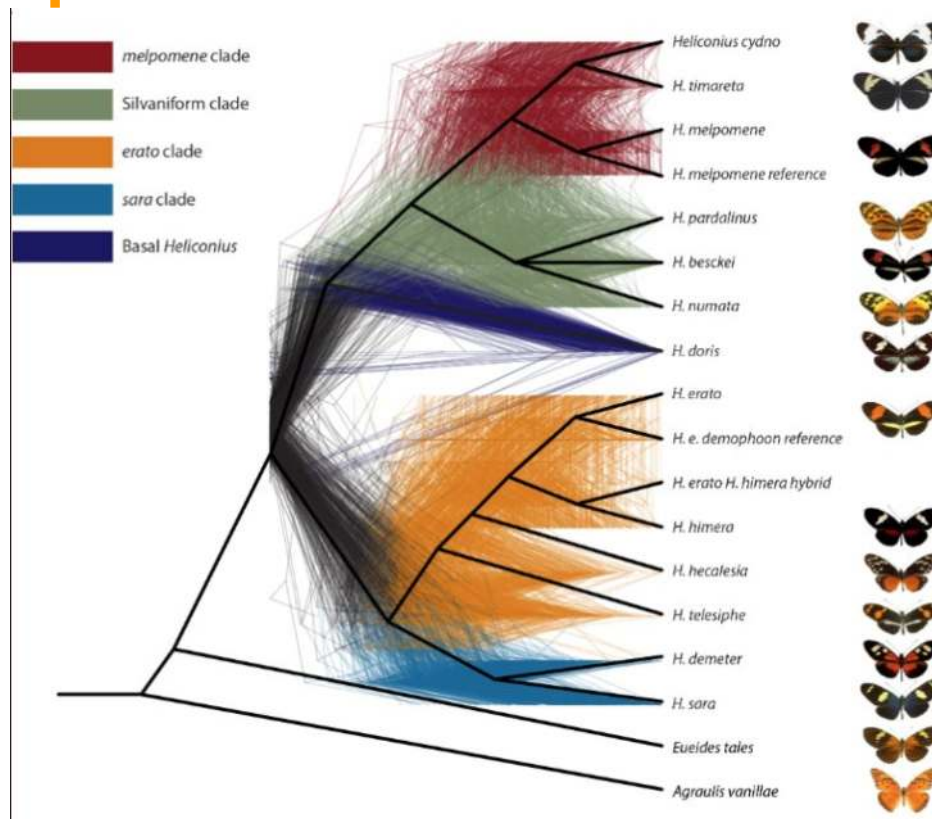
Eg 3, genes evolved through duplication and not through speciation

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



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?

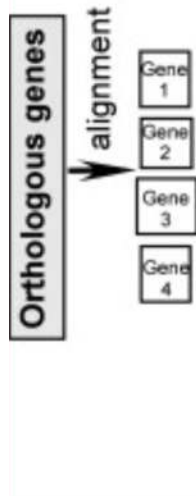
- 1) Build different subsets of your data through a subsampling strategy selecting genes with different properties
- 2) **Run different analyses that rely on different assumptions and/or apply different models**

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 and/or apply different models
- 3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*

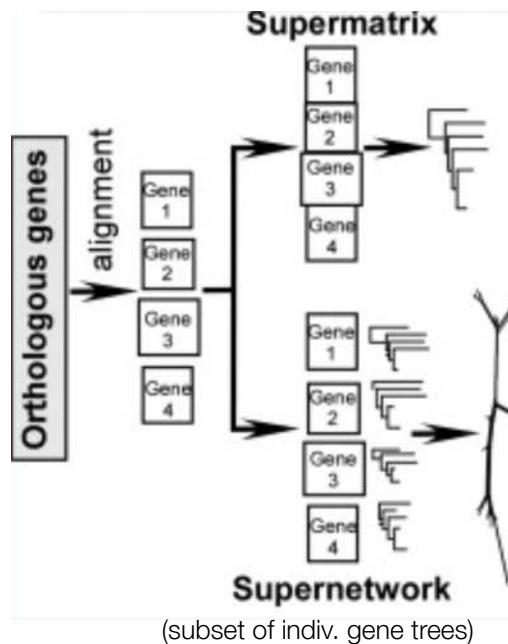
So... what do we do to test the robustness of our tree?

- 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 and/or apply different models
- 3) **Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees***



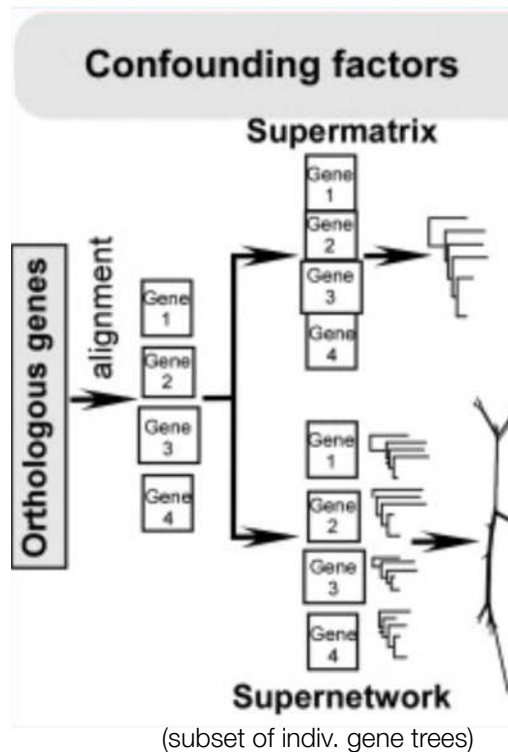
So... what do we do to test the robustness of our tree?

- 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 and/or apply different models
- 3) **Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees***



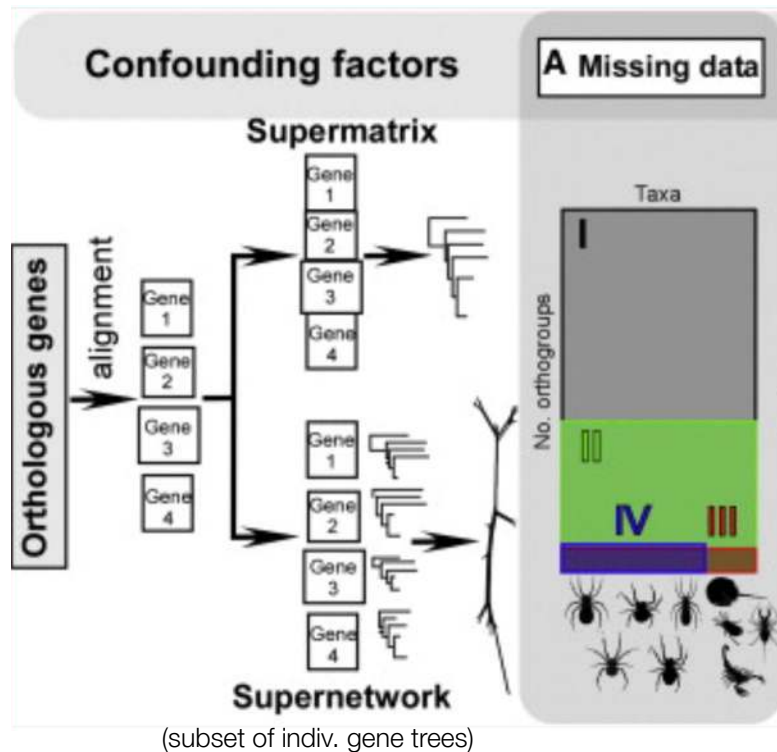
So... what do we do to test the robustness of our tree?

- 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 and/or apply different models
- 3) **Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees***



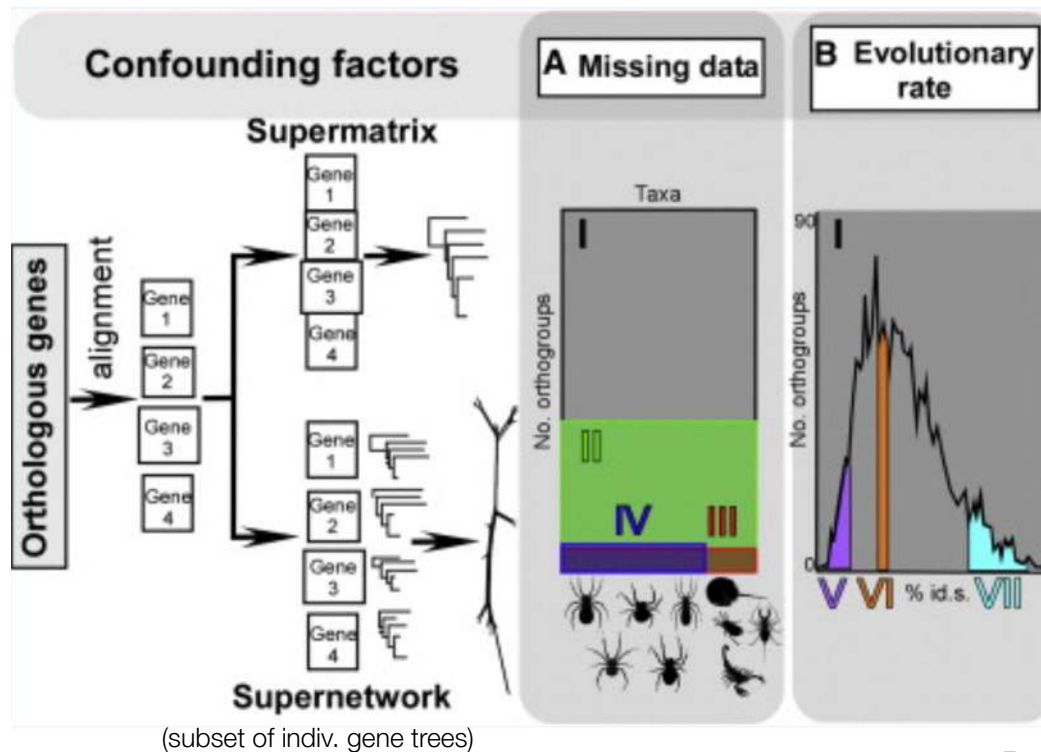
So... what do we do to test the robustness of our tree?

- 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 and/or apply different models
- 3) **Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees***



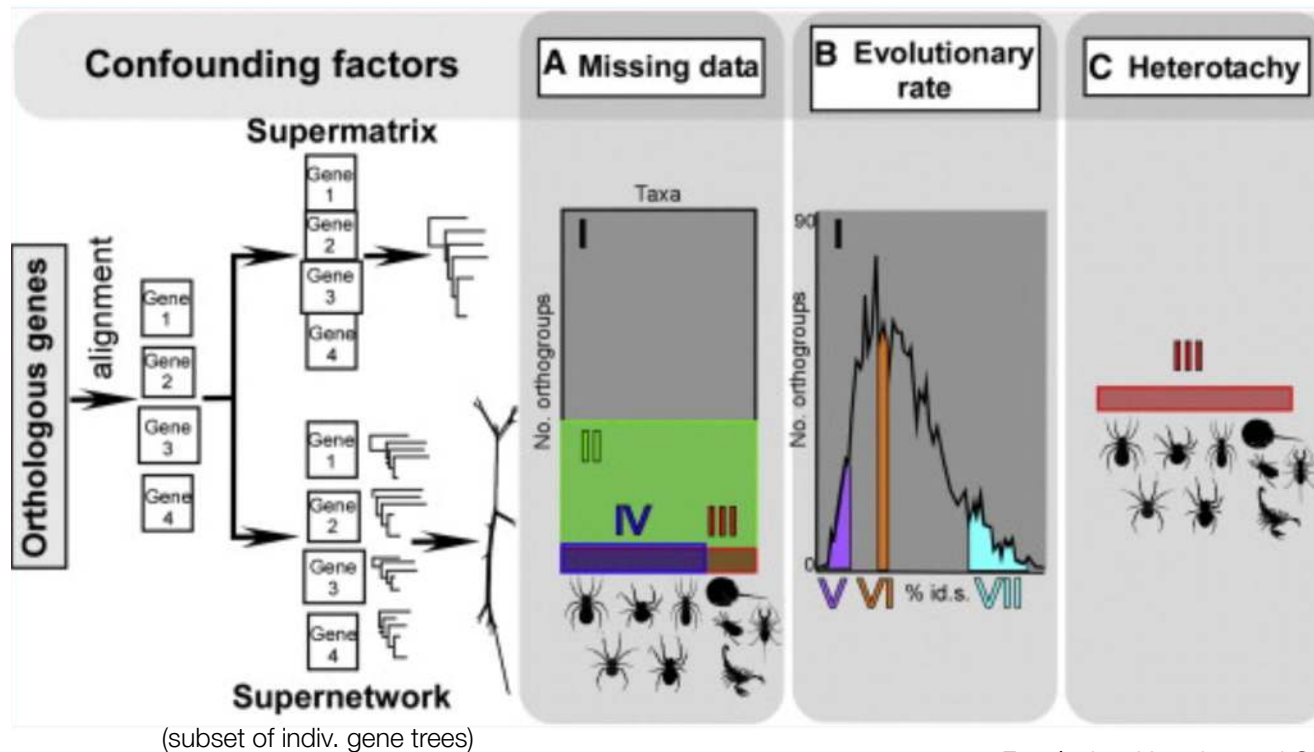
So... what do we do to test the robustness of our tree?

- 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 and/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?

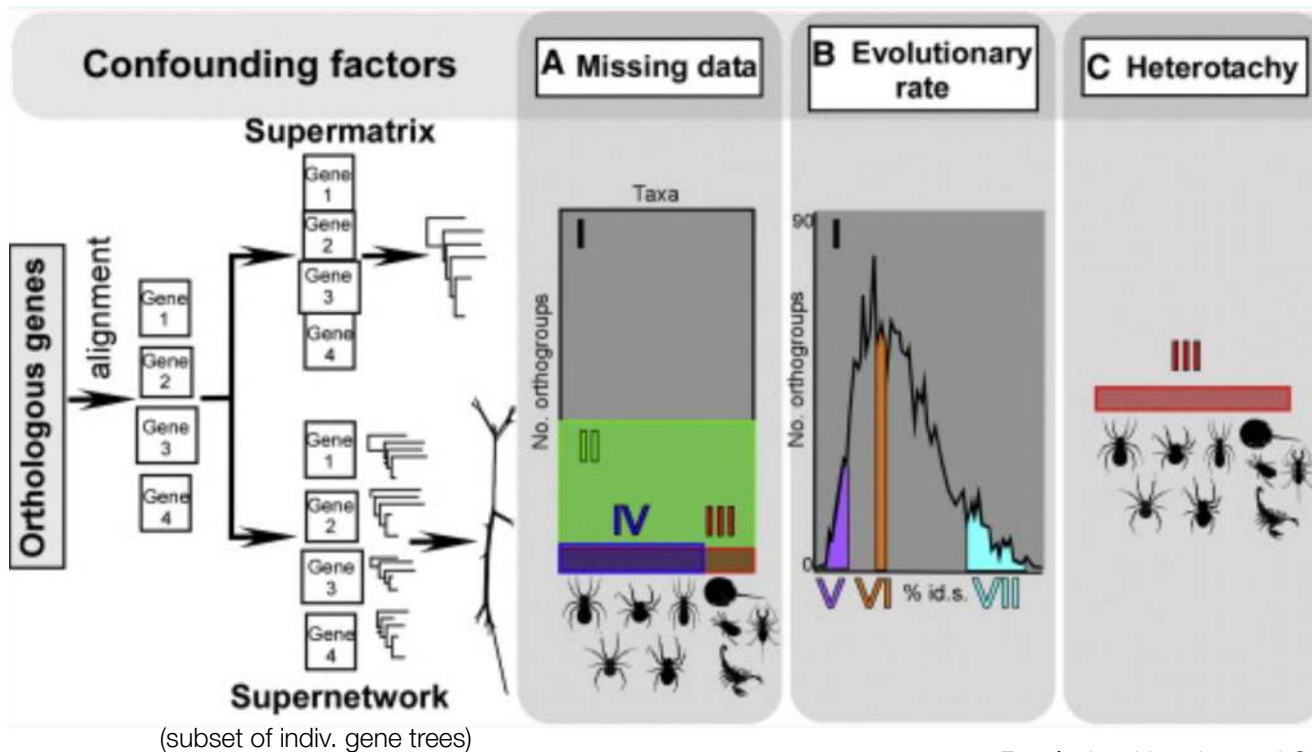
- 1) Build different subsets of your data through a subsampling strategy selecting genes with different properties
- 2) Run different analyses that rely on different assumptions and/or apply different models
- 3) **Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees***



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

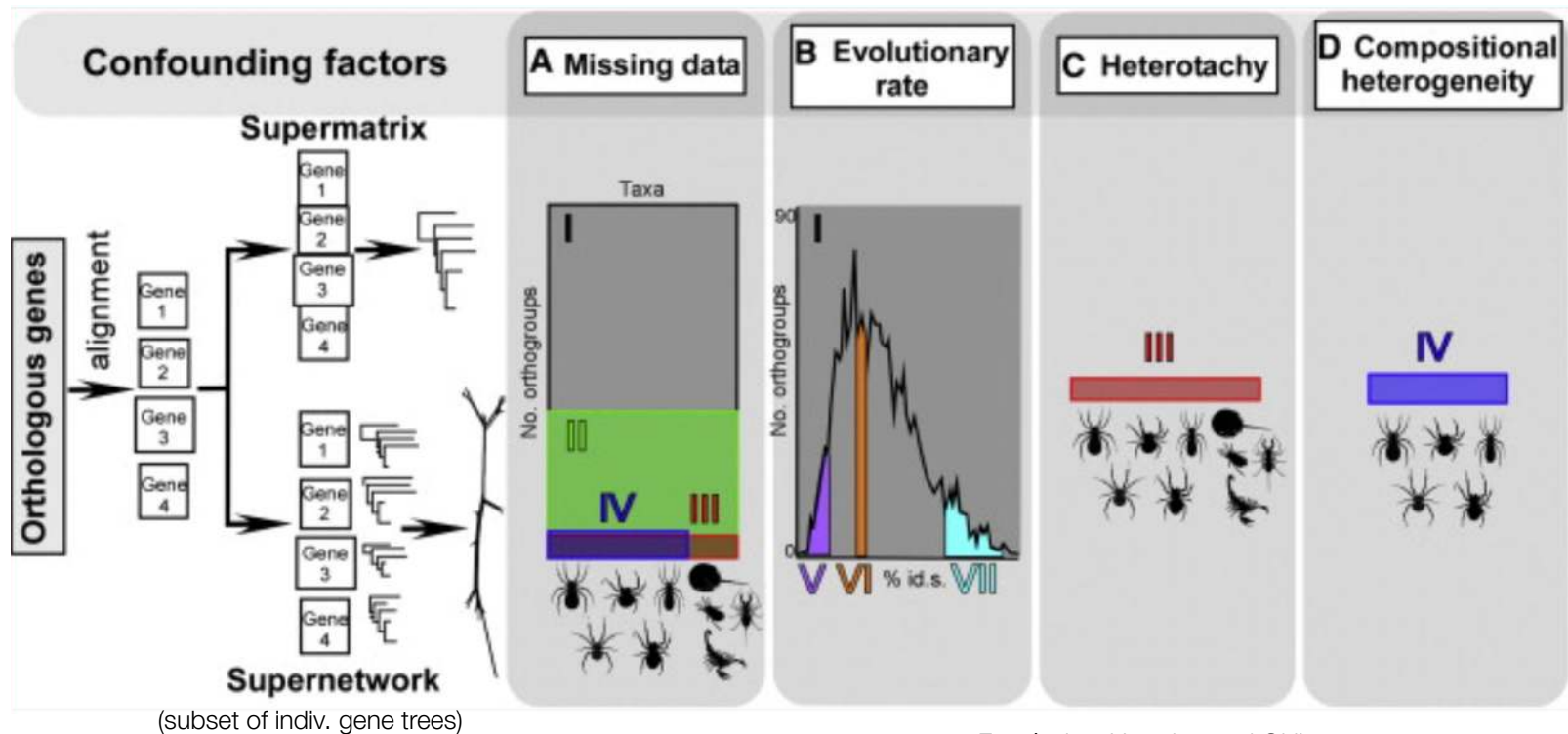
- 1) Build different subsets of your data through a subset of properties
- 2) Run different analyses that rely on different assumptions
- 3) **Do 1) and 2) both at the level of *supermatrix* and *supernetwork***

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)



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 and/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*

And... how do I *choose* a subset of genes to run these analyses?

And... how do I *choose* a subset of genes to run these analyses?

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

And... how do I *choose* 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 choose the ones that behave 'well' (eg, discard the outliers).

And... how do I *choose* 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 choose the ones that behave 'well' (eg, discard the outliers).
 - > Custom scripts (eg, select genes with less 50% of missing data)

And... how do I *choose* 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 choose 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 *choose* 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 choose the ones that behave 'well' (eg, discard the outliers).

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

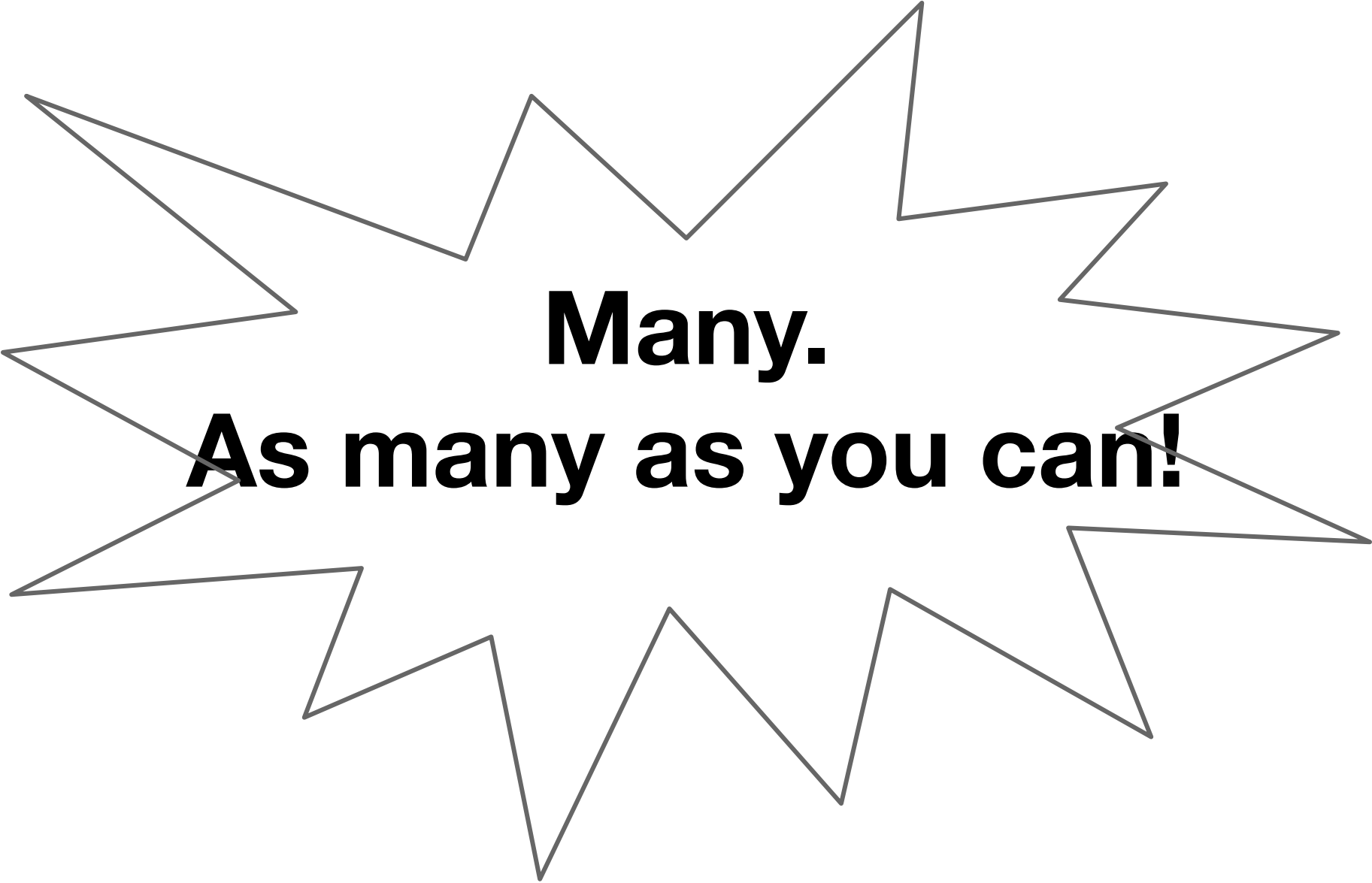
-> Software to measure some of these properties (eg, compositional heterogeneity, saturation, etc.)



We will be doing this today in our hands-on session

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

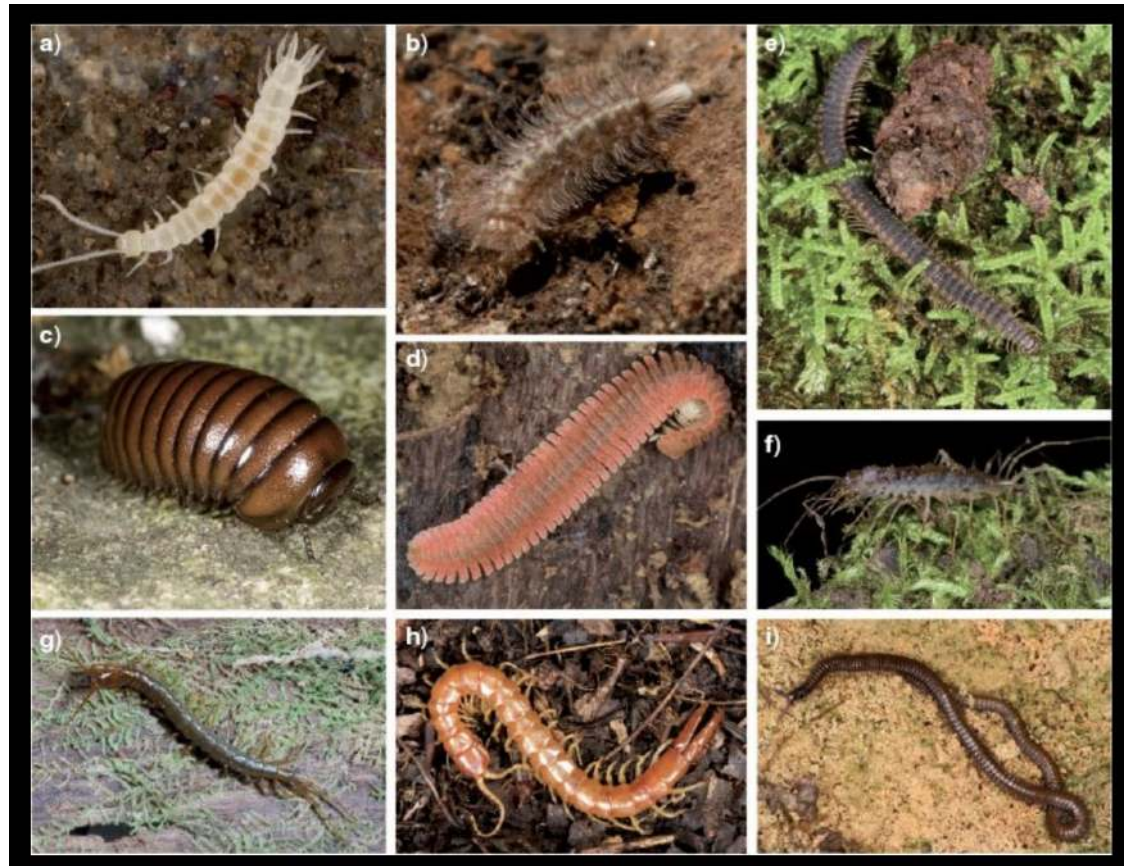
So... how many matrices/subsets/analyses should I analyze?



Many.

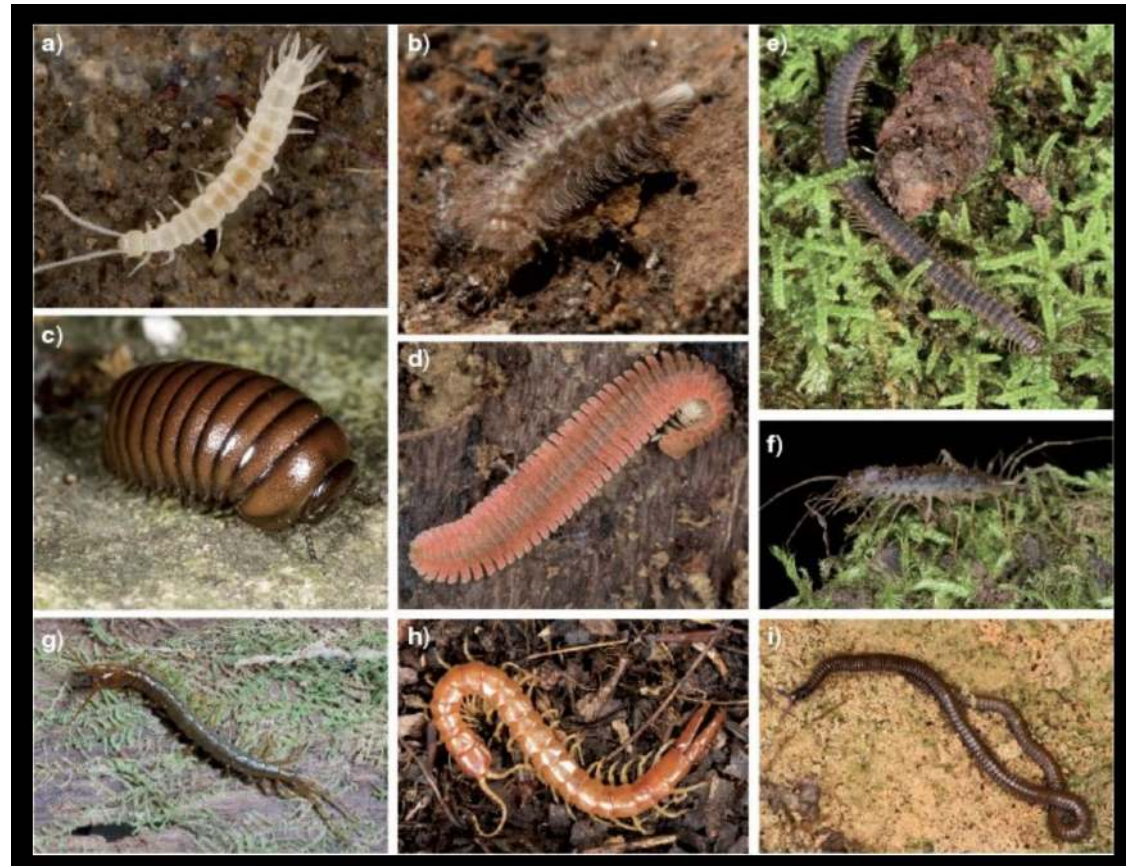
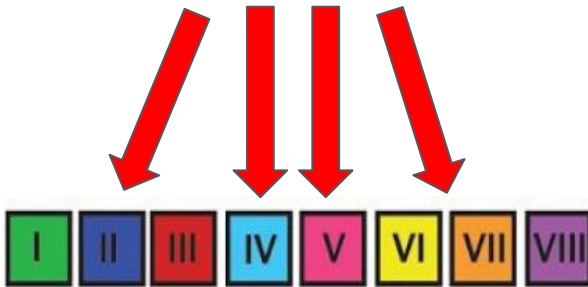
As many as you can!

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



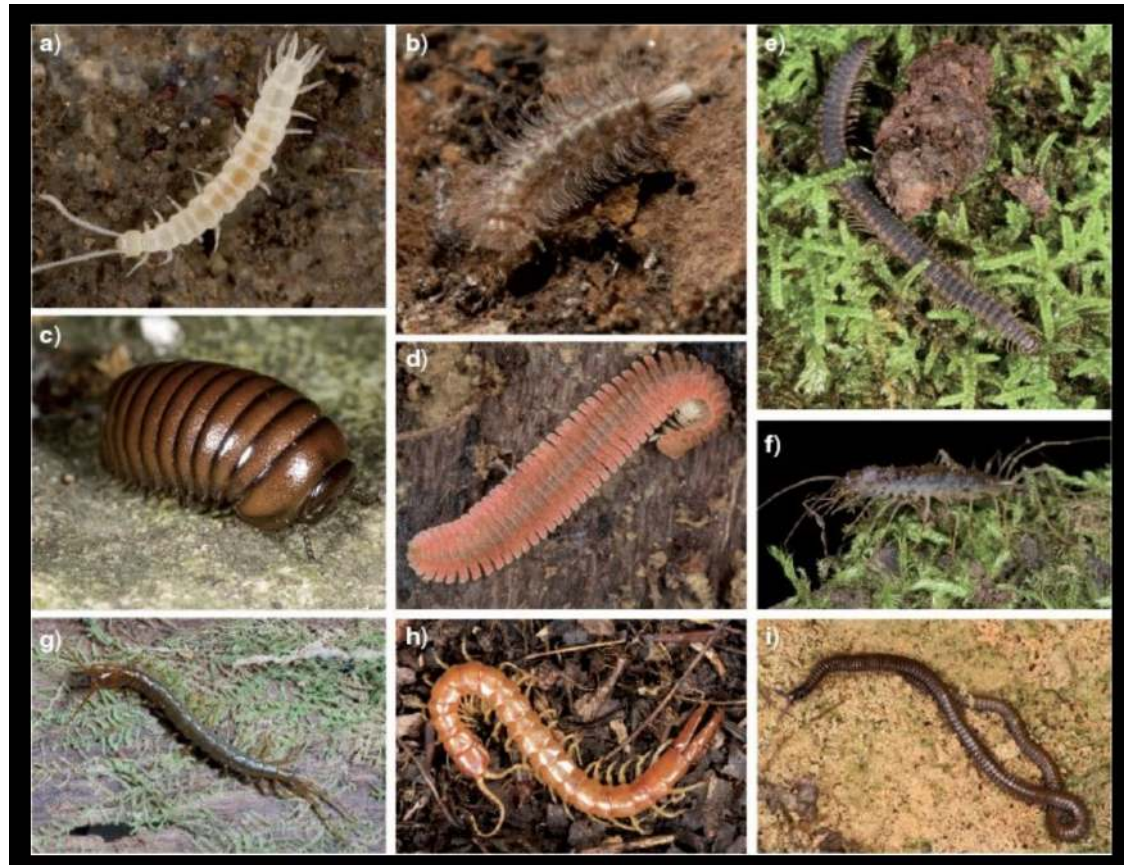
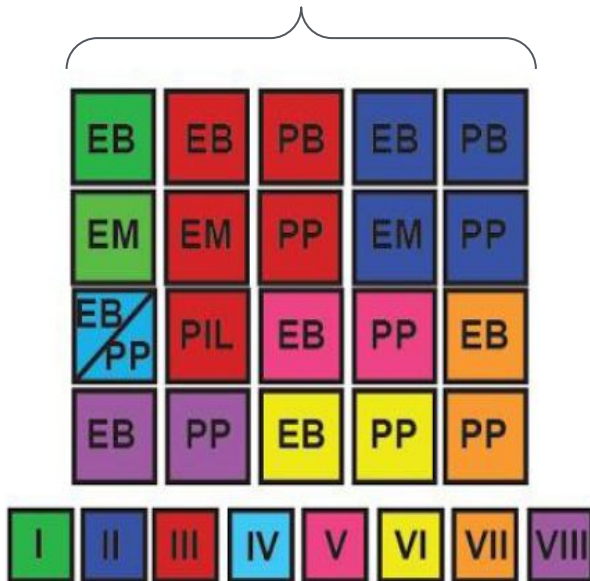
So... how many matrices/subsets/analyses should I analyze?

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

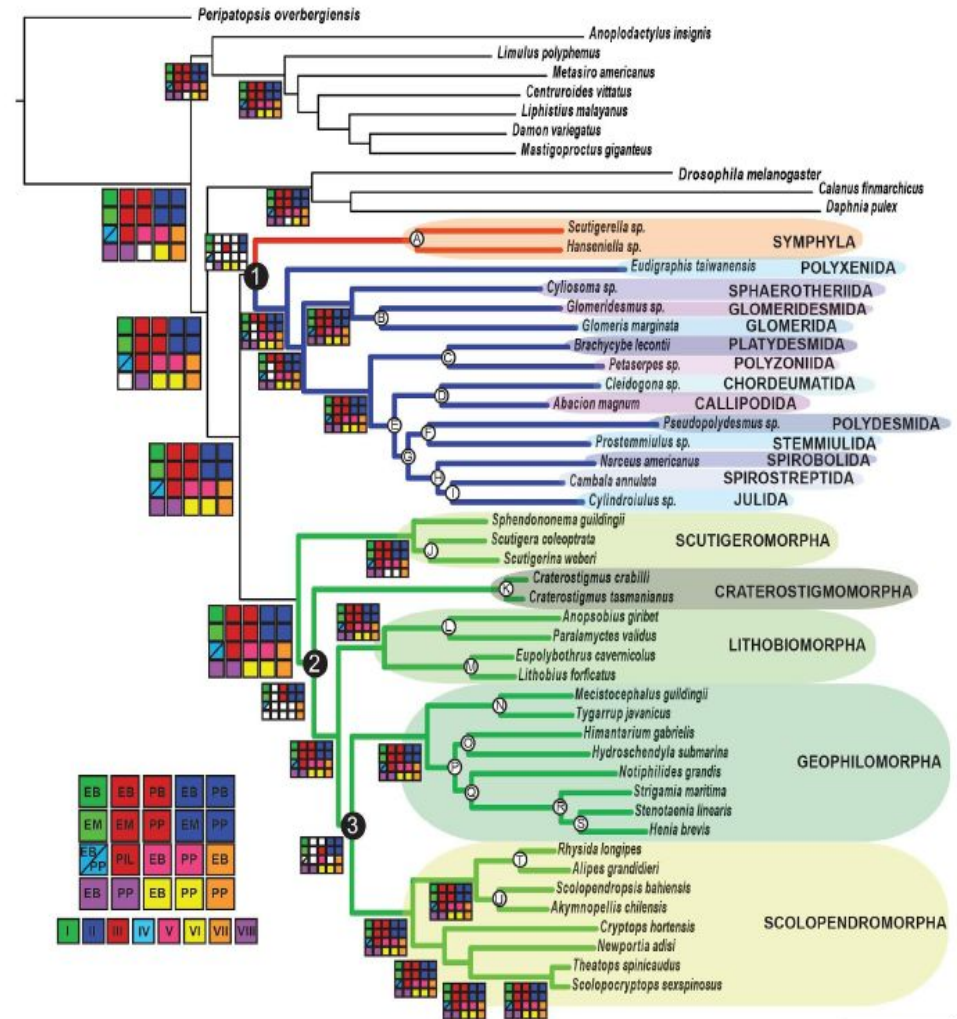
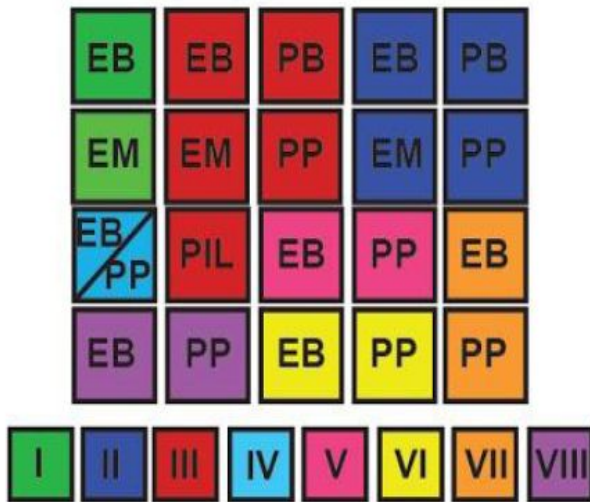


So... how many matrices/subsets/analyses should I analyze?

These are **analyses**



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



Fernández, Edgecombe & Giribet (2016) Syst Biol

Generating phylogenomic data matrices: hands-on session

Generating phylogenomic data matrices: hands-on session



Generating phylogenomic data matrices: hands-on session



The Český Krumlov town hall decides to fund a project to understand whether the brown bear is more closely related to the polar bear or the American black bear



Generating phylogenomic data matrices: hands-on session



(Important piece of information (shared by Scott): Český Krumlov locals used to refer to the workshop participants as '**molekulos**')

The Český Krumlov town hall decides to fund a project to understand whether the brown bear is more closely related to the polar bear or the American black bear



Generating phylogenomic data matrices: hands-on session



(Important piece of information (shared by Scott): Český Krumlov locals used to refer to the workshop participants as '**molekulos**')

The Český Krumlov town hall decides to fund a project to understand whether the brown bear is more closely related to the polar bear or the American black bear



Let's ask the '**molekulos**' for help!!

Generating phylogenomic data matrices: hands-on session

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



Giant Panda
Ailuropoda melanoleuca
(outgroup)



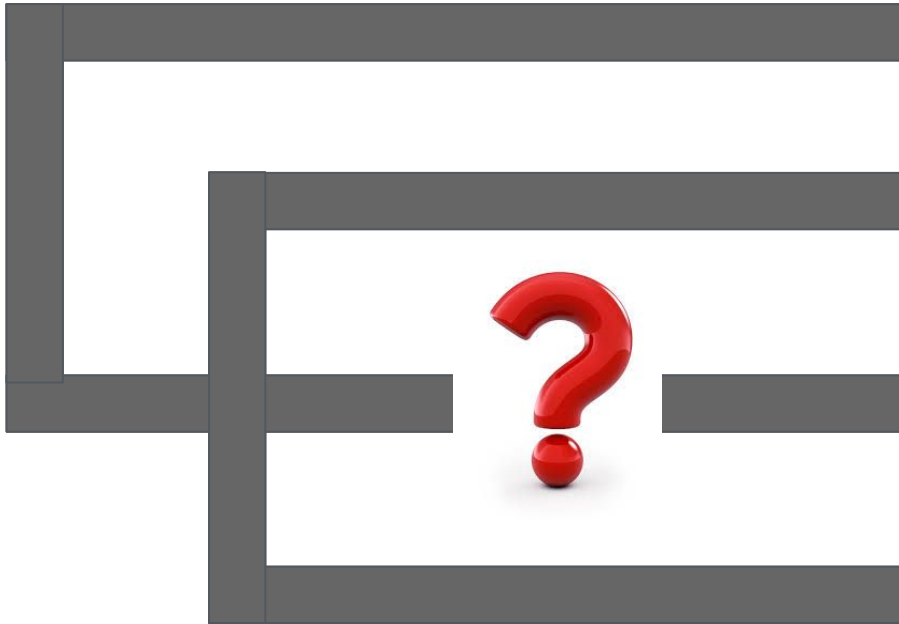
Black bear
Ursus americanus



Polar bear
Ursus maritimus



Brown bear
Ursus arctos



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



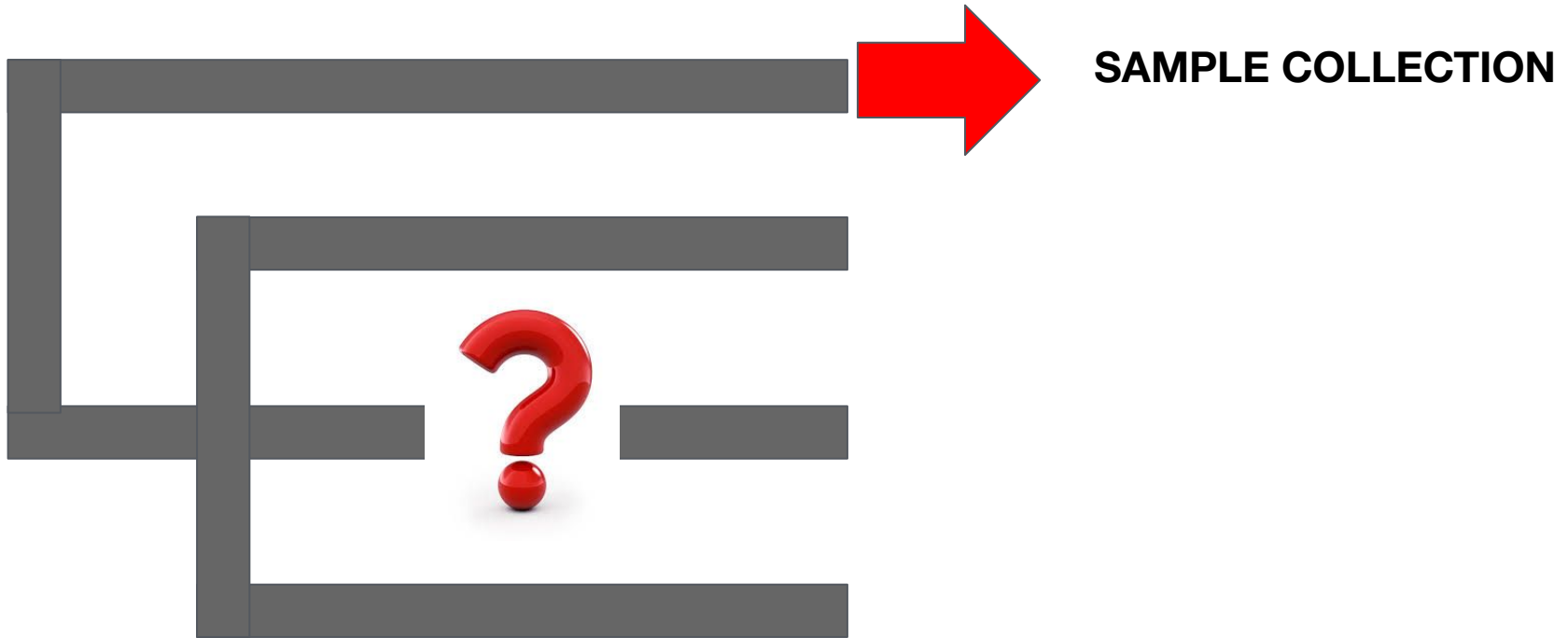
Joseph
Margaret
Maripepa
Maria



Amparo
Paco
Adelaide
Margo

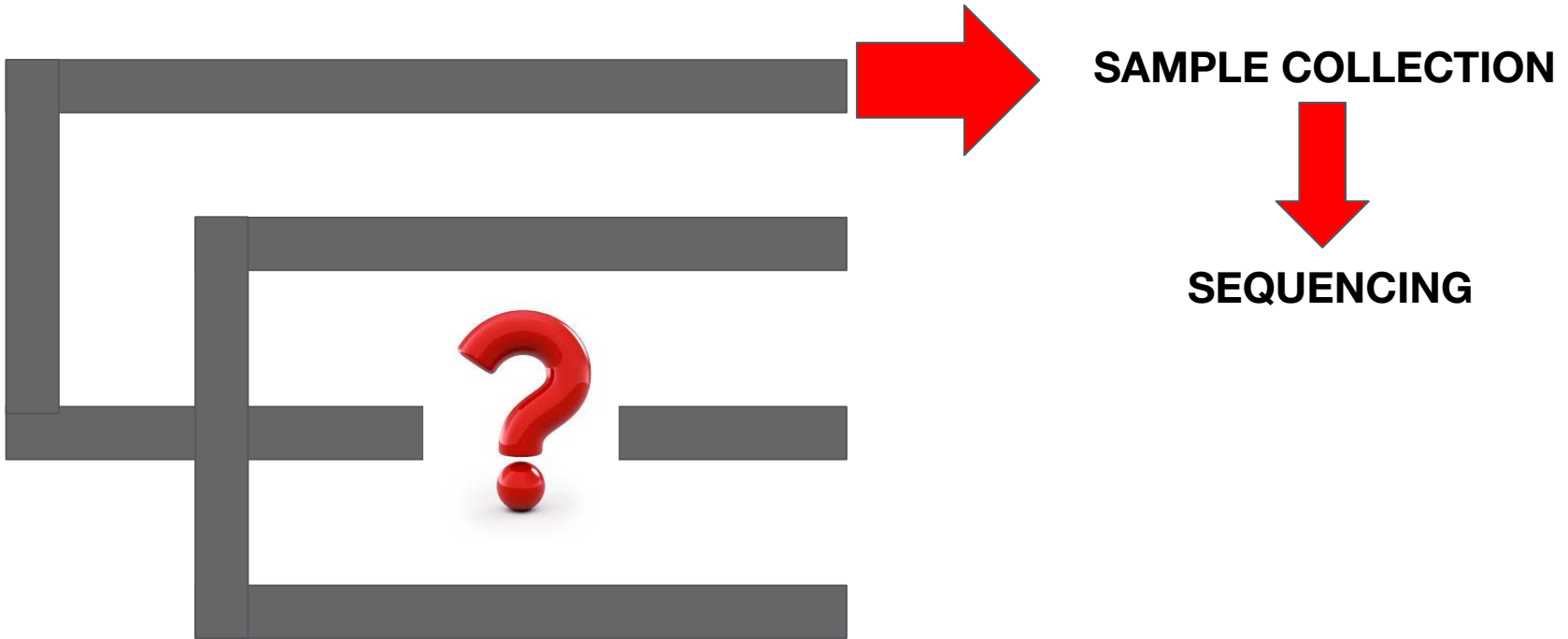
Generating phylogenomic data matrices: hands-on session

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



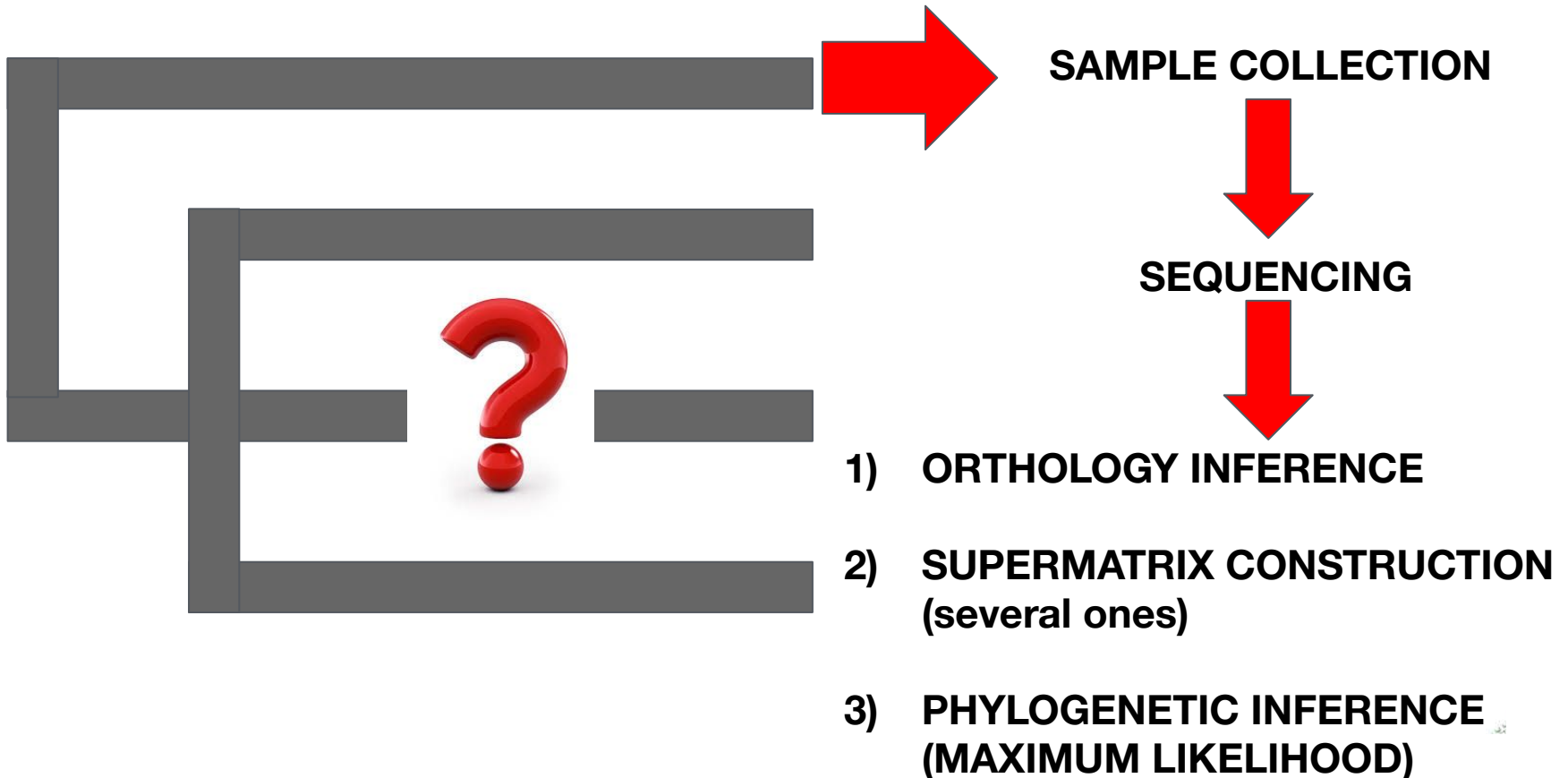
Generating phylogenomic data matrices: hands-on session

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



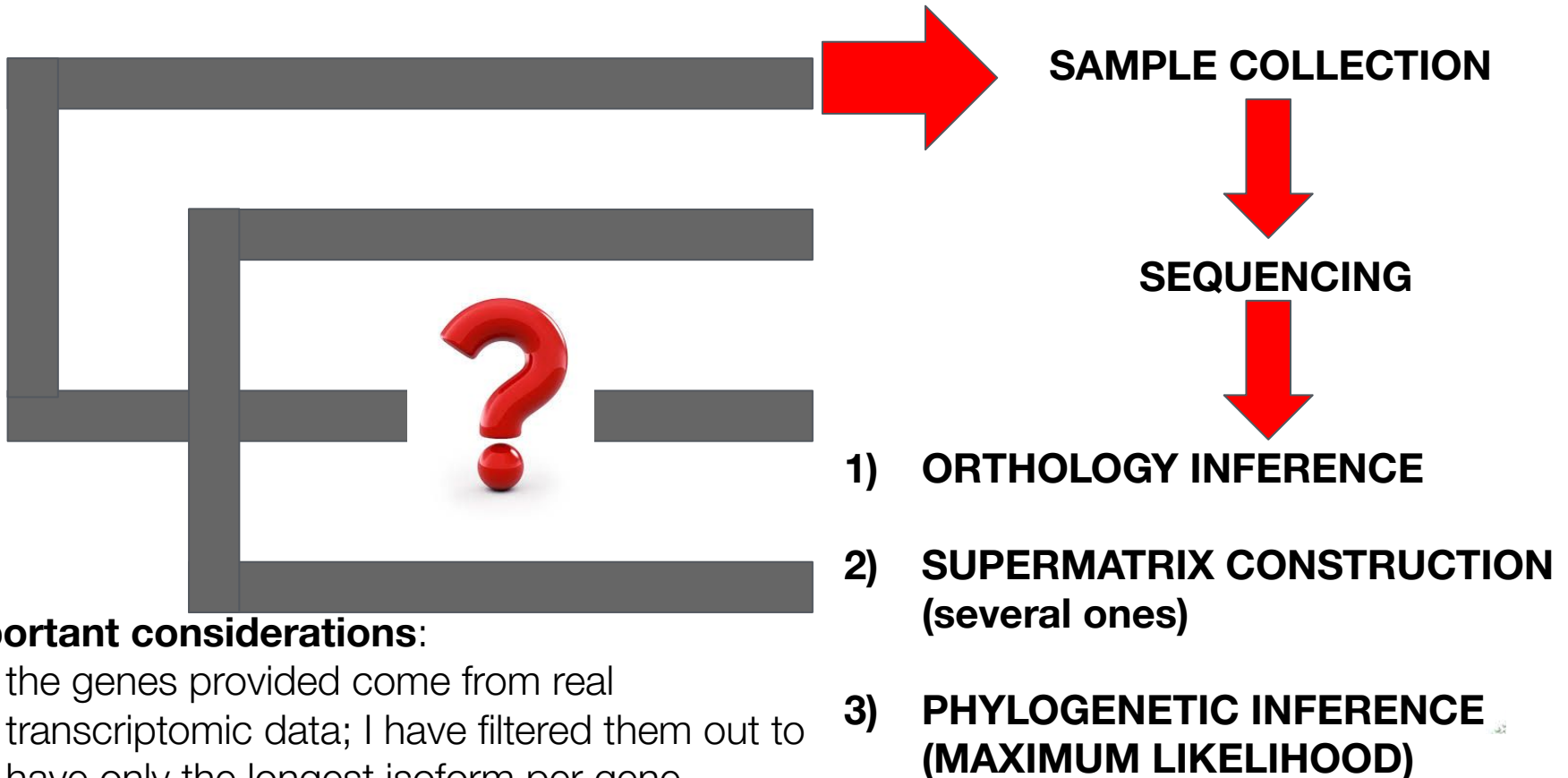
Generating phylogenomic data matrices: hands-on session

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



Generating phylogenomic data matrices: hands-on session

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

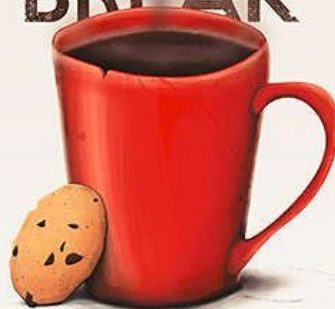


(Important considerations:

- the genes provided come from real transcriptomic data; I have filtered them out to have only the longest isoform per gene
- we are going to work at the level of amino acid data, not nucleotide (TransDecoder)



**KEEP
CALM
AND
TAKE
A
BREAK**



Generating phylogenomic data matrices: hands-on session

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

During the first part of the tutorial, you have...

- 1) Inferred orthologous groups with OrthoFinder.
- 2) Created a supermatrix selecting genes based on:
 - a) their amount of missing data
 - b) their 'decisiveness' (ie, representation among species)
 - c) removing outliers accounting for a bunch of confounding factors (eg, compositional heterogeneity, saturation, etc.).
- 3) Inferred ML phylogenetic trees from these matrices

Which topology are your analyses supporting? Is this topology robust, or are the analyses showing any

Generating phylogenomic data matrices: hands-on session

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



**Phylogenomics illuminate the interrelationships of the genus *Ursus*
and supports the brown bear as sister group to the polar bear**

Authors: Participants of the 2023 Workshop on Genomics český Krumlov ('molekulos')

Generating phylogenomic data matrices: hands-on session

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

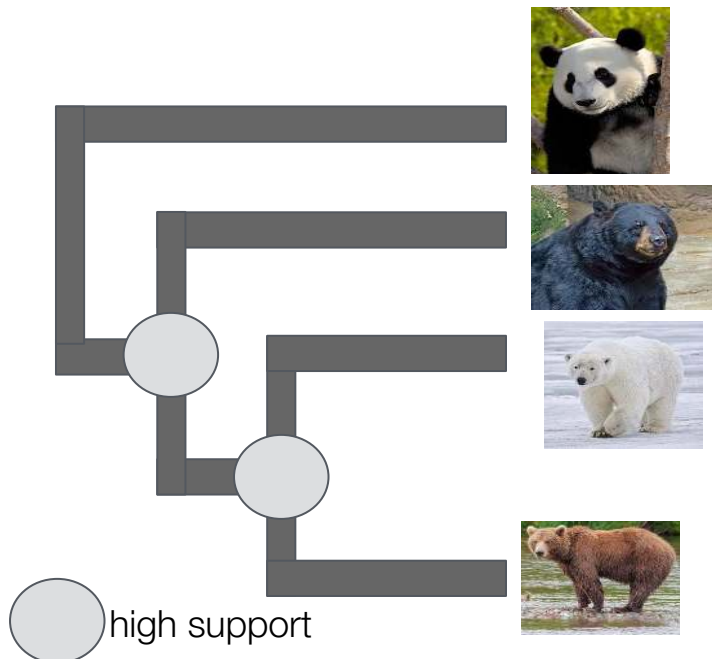


Phylogenomics illuminate the interrelationships of the genus *Ursus* and supports the brown bear as sister group to the polar bear

Authors: Participants of the 2023 Workshop on Genomics český Krumlov ('molekulos')

Abstract:

The interrelationships within the genus *Ursus* have been contentious based on the analysis of a limited amount of molecular markers. Here, we sequenced full genomes of 16 specimens of the American black bear, brown bear, polar bear and giant panda and explored their phylogenetic relationships through a phylogenomic spyglass. **Our results, based on the analysis of multiple supermatrices to account for the effect of missing data, compositional heterogeneity and other confounding factors, strongly support a sister relationship of the brown bear to the polar bear.** Our findings pave the road towards understanding bear evolution at a deeper level.



Generating phylogenomic data matrices: hands-on session

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



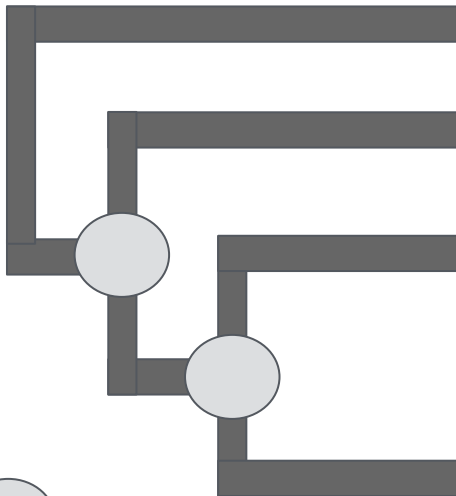
Phylogenomics illuminate the interrelationships of the genus *Ursus* and supports the brown bear as sister group to the polar bear

Authors: Participants of the 2023 Workshop on Genomics český Krumlov ('molekulos')

Abstract:

The interrelationships of the genus *Ursus* have been a contentious issue for decades. Recent molecular data from thousands of specimens (including American black, brown, and giant pandas) analyzed through a phylogenomic approach, using multiple supermatrices to account for the effect of missing data, compositional heterogeneity and other confounding factors, strongly support a sister relationship of the brown bear to the polar bear. Our findings pave the road towards understanding bear evolution at a deeper level.

REJECTED



high support



Generating phylogenomic data matrices: hands-on session

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



Phylogenomics illuminate the interrelationships of the genus *Ursus* and supports the brown bear as sister group to the polar bear

Authors: Participants of the 2023 Workshop on Genomics český Krumlov ('molekulos')

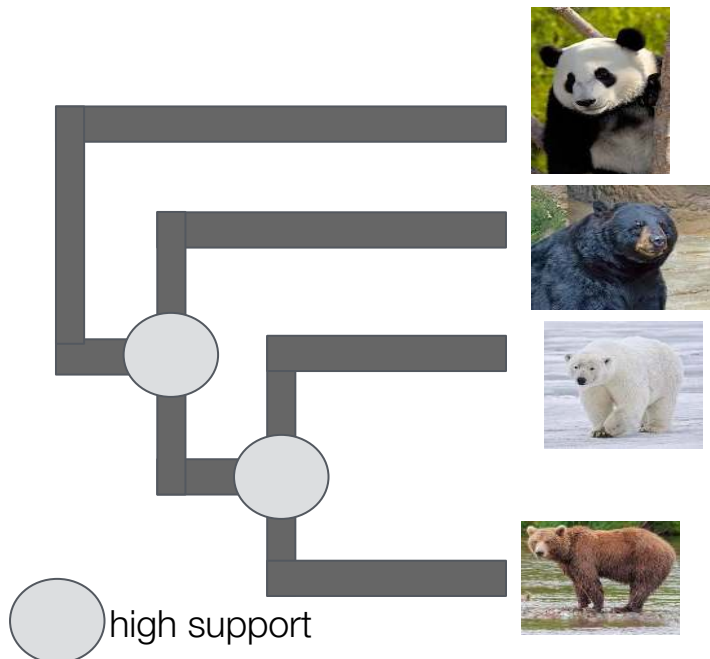
Abstract:

The interrelationships of the genus *Ursus* have been contentious since the first molecular phylogenetic analysis of specimens collected in the 19th century. Through a phylogenomic analysis of multiple supermatrices to account for the effect of

REJECTED

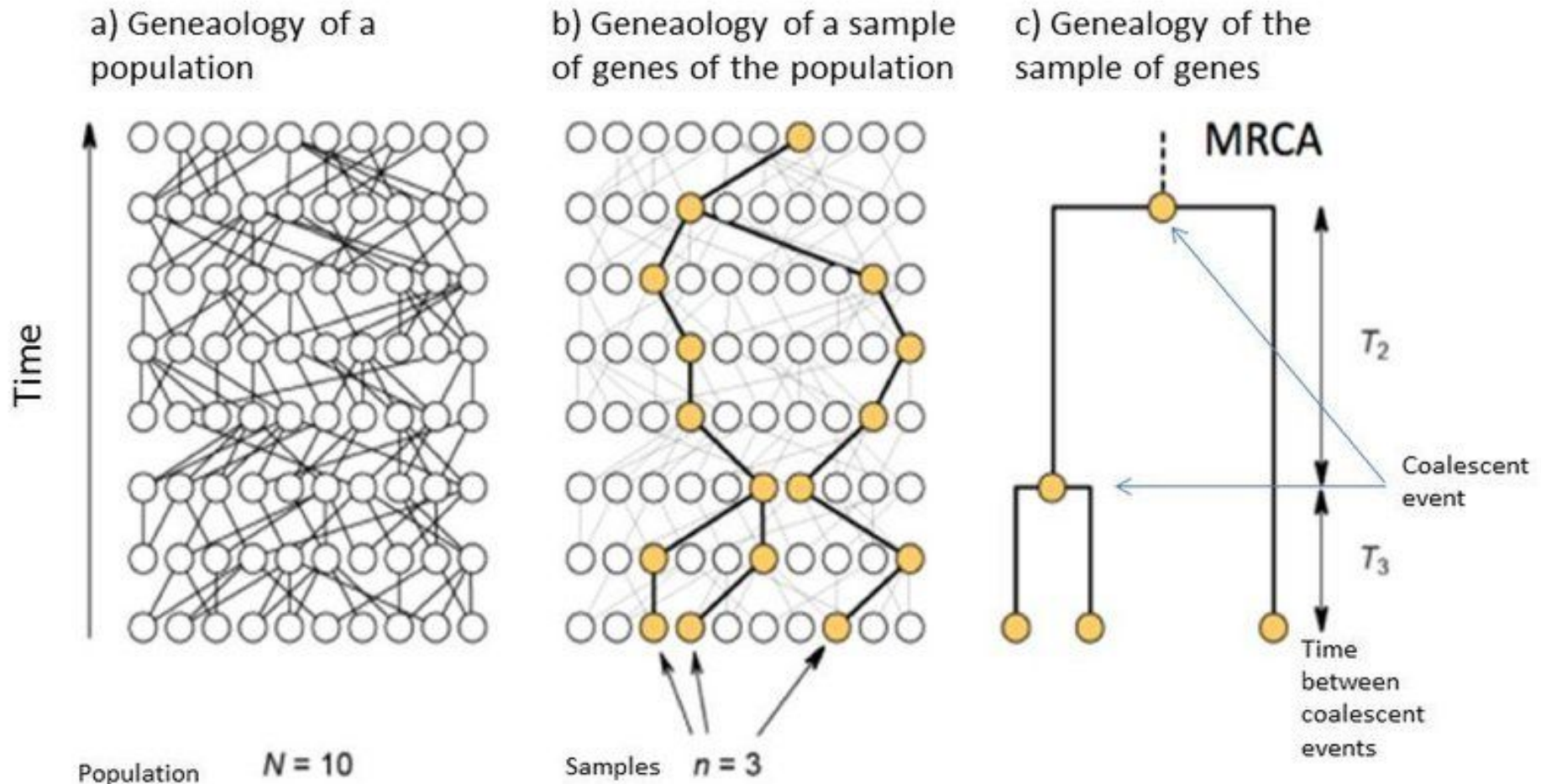
been
unt of
nes of 16
r, polar bear
relationships
sed on the
effect of

Reviewer #3: although I appreciate the efforts of the authors to account for confounding factors and test the robustness of their results, they failed to test whether their hypothesis was driven by incongruence between individual gene evolutionary trajectories.

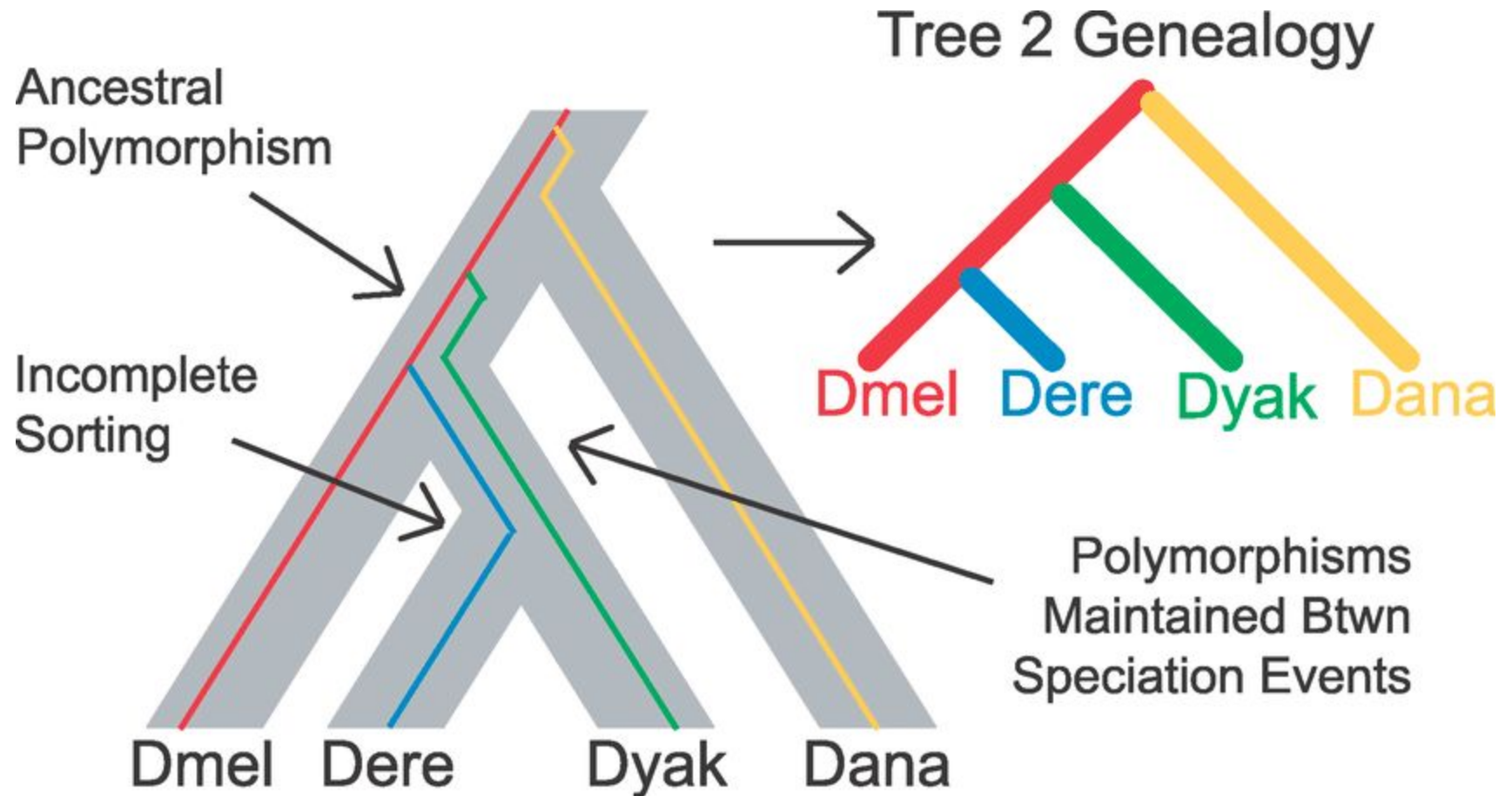


Analyzing gene tree/species tree conflict: hands-on session

Brief introduction to coalescent theory

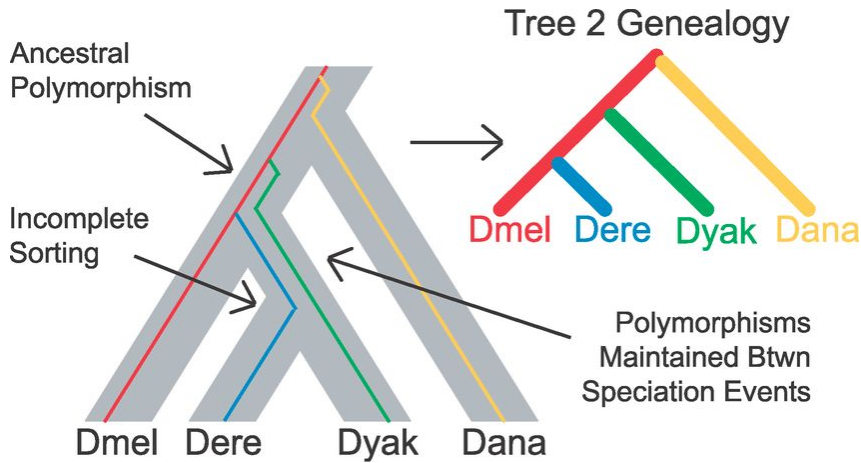


Brief introduction to coalescent theory

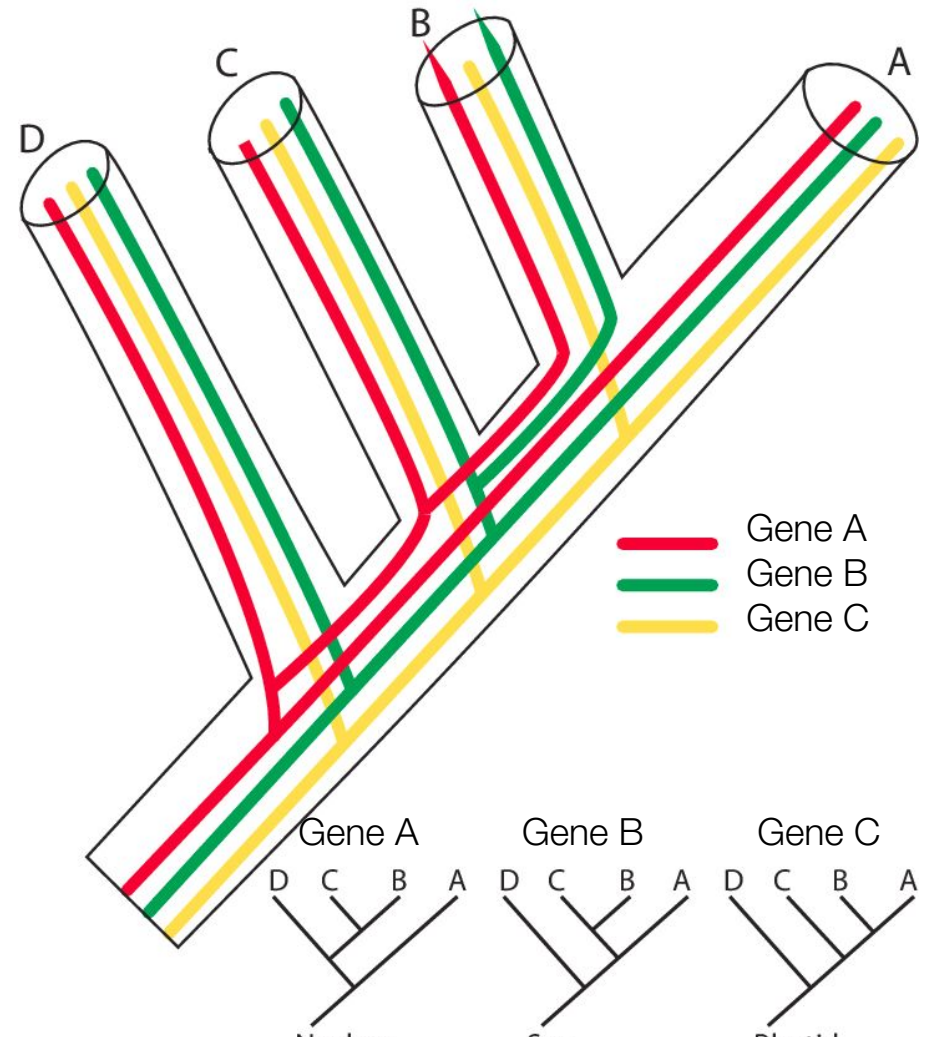


Analyzing gene tree/species tree conflict: hands-on session

Brief introduction to coalescent theory



Pollard et al. 2006



Naciri and Li 2015

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?

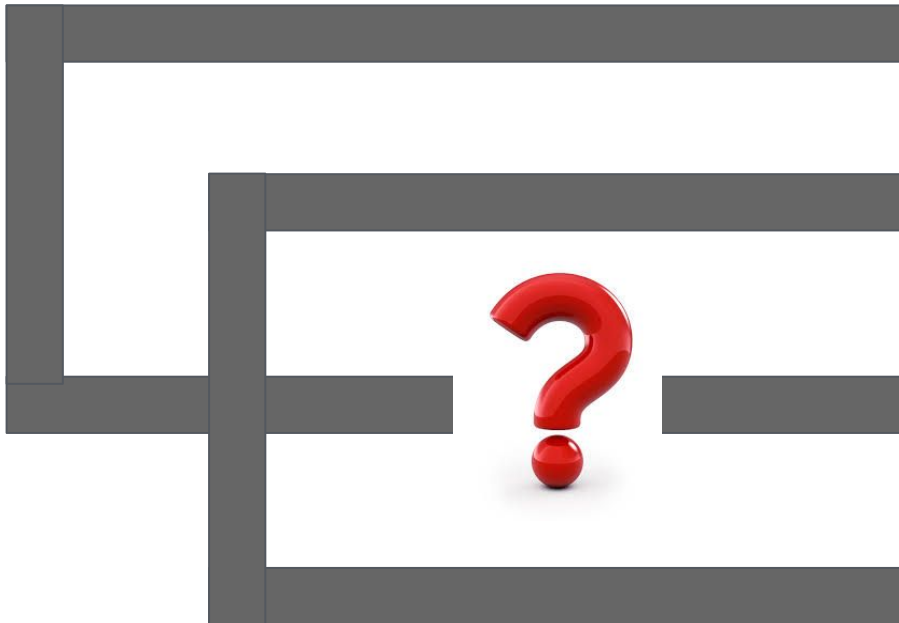
ASTRAL is a tool for estimating an unrooted species tree given a set of unrooted gene trees.

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

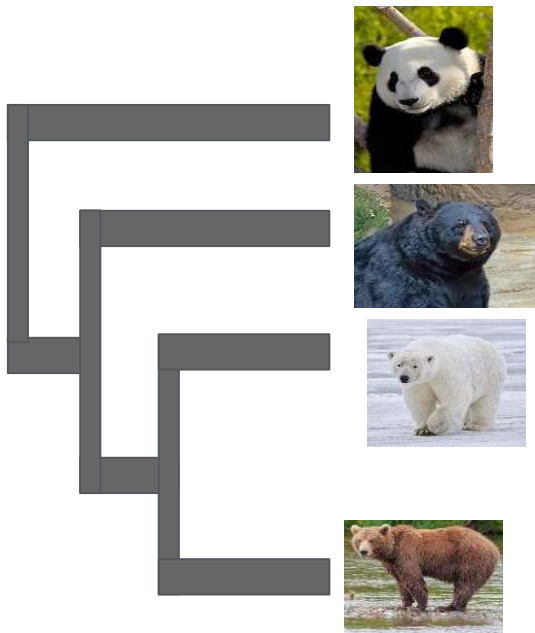
Is the polar bear the sister group to the American black bear or the brown bear?



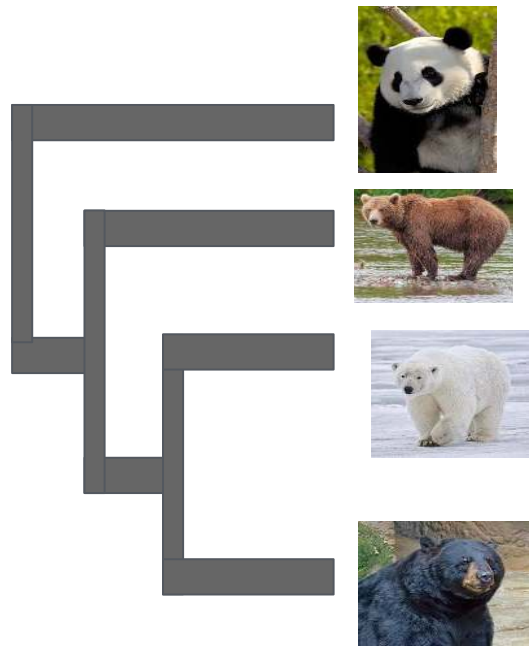
Analyzing gene tree/species tree conflict: hands-on session

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

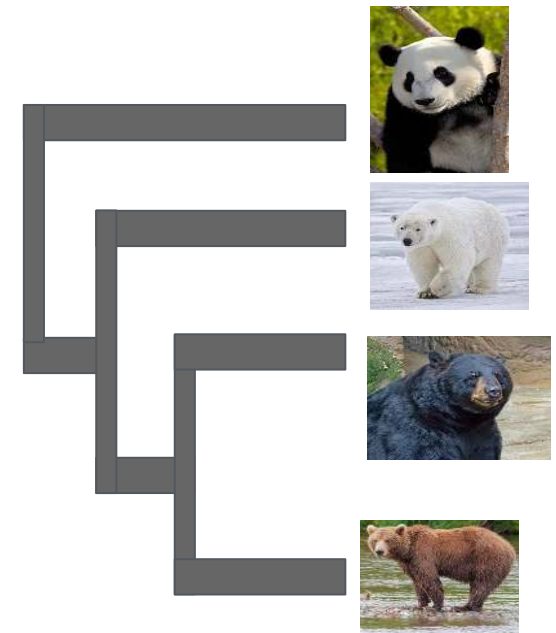
Second part of the tutorial: analyzing how the evolutionary trajectory of each individual orthogroup supports each of the three topologies.



Species Tree 1



Species Tree 2



Species Tree 3

Generating phylogenomic data matrices: hands-on session

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



**Phylogenomics illuminate the interrelationships of the genus *Ursus*
and supports the brown bear as sister group to the polar bear**

Authors: Participants of the 2023 Workshop on Genomics český Krumlov ('molekulos')

Generating phylogenomic data matrices: hands-on session

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

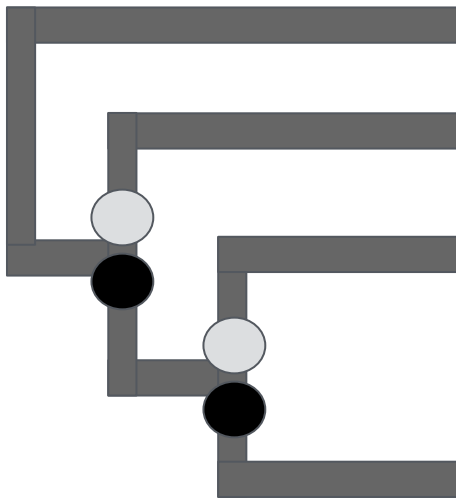


Phylogenomics illuminate the interrelationships of the genus *Ursus* and supports the brown bear as sister group to the polar bear

Authors: Participants of the 2023 Workshop on Genomics český Krumlov ('molekulos')

Abstract:

The interrelationships of the species within the genus *Ursus* has been contentious based on the analysis of a limited amount of molecular markers. Here, we sequenced full genomes of 16 specimens of the American black bear, brown bear, polar bear and giant panda and explored their phylogenetic relationships through a phylogenomic spyglass. Our results, based on the analysis of multiple supermatrices to account for the effect of missing data, compositional heterogeneity and other confounding factors, **as well as accounting for incongruence between individual gene trees under the multispecies coalescent model**, strongly support a sister relationship of the brown bear to the polar bear. Our findings pave the road towards understanding bear evolution at a deeper level.



high support supermatrix



high support indiv. gene trees (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?



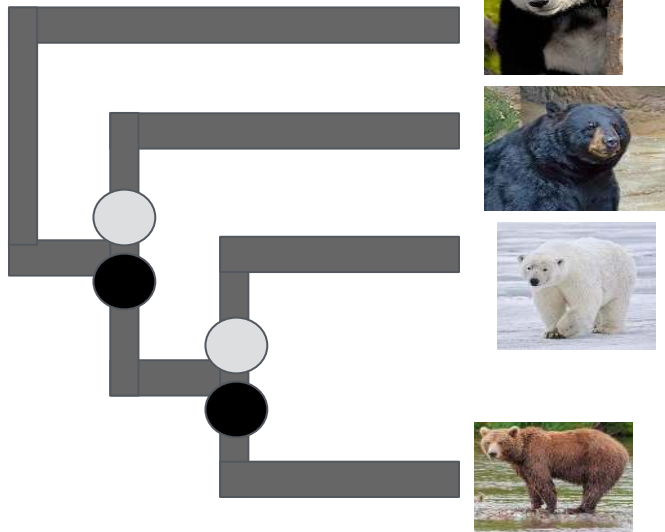
Phylogenomics illuminate the interrelationships of the genus *Ursus* and supports the brown bear as sister group to the polar bear

Authors: Participants of the 2023 Workshop on Genomics český Krumlov ('molekulos')

Abstract:

The interrelationships of the genus *Ursus* has been limited by the amount of genomic data available. We used a multispecies coalescent model to generate a full bear phylogeny, including the American black bear, brown bear, and polar bear. Our results, based on a multispecies coalescent model, strongly support a sister relationship of the brown bear to the polar bear. Our findings pave the road towards understanding bear evolution at a deeper level.

ACCEPTED



high support supermatrix



high support indiv. gene trees (multispecies coalescent)