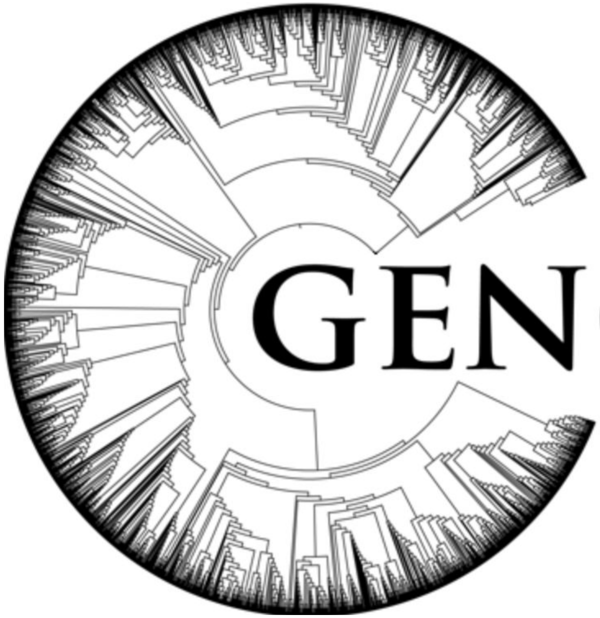
(A VERY SHORT INTRO TO)

PHYLOGENOMICS



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)

HARVARD
UNIVERSITY

CRG
Centre
for Genomic
Regulation

A little bit about myself

Madrid (PhD)

⇩

Boston (1st postdoc)

⇩

Barcelona (2nd postdoc & my lab)

www.metazomics.com

@Rosamygale

@metazomics

A little bit about myself

Madrid (PhD)

Boston (1st postdoc)

Barcelona (2nd postdoc & my lab)

HARVARD UNIVERSITY

CRG Centre for Genomic Regulation

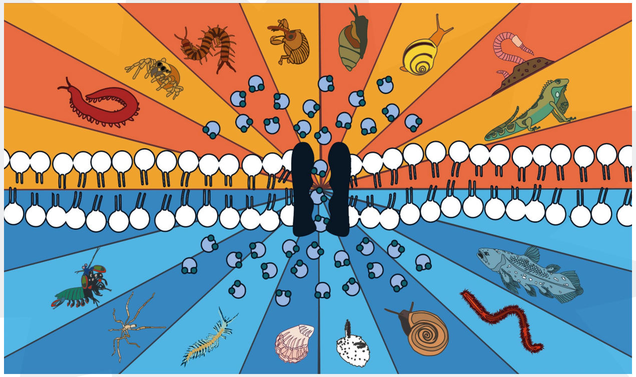
INSTITUT de BIOLOGIA EVOLUTIVA iBE CSIC upf

Metazoa Phylogenomics Lab

www.metazomics.com

@Rosamygale
@metazomics

Main lines of research:



A little bit about myself

Madrid (PhD)

Boston (1st postdoc)

Barcelona (2nd postdoc & my lab)

HARVARD UNIVERSITY

CRG Centre for Genomic Regulation

INSTITUT de BIOLOGIA EVOLUTIVA iBE CSIC upf

Metazoa Phylogenomics Lab

www.metazomics.com

@Rosamygale
@metazomics

Main lines of research:

A little bit about myself

Madrid (PhD)
⇩
Boston (1st postdoc)
⇩
Barcelona (2nd postdoc & my lab)

HARVARD UNIVERSITY

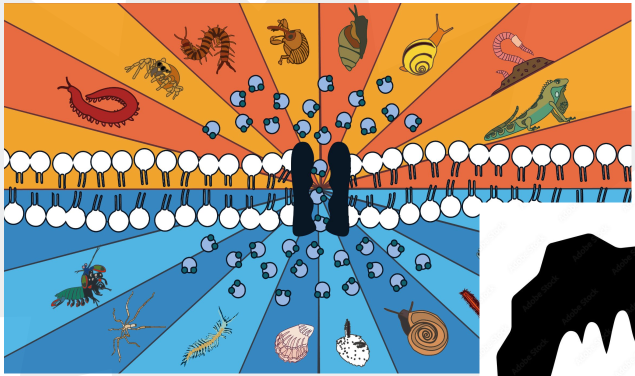
CRG Centre for Genomic Regulation

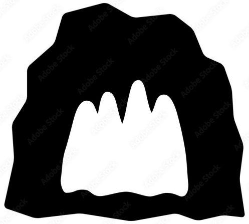
Metazoa Phylogenomics Lab

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

A little bit about myself

Madrid (PhD)

⇩

Boston (1st postdoc)

⇩

Barcelona (2nd postdoc & my lab)

HARVARD UNIVERSITY

CRG Centre for Genomic Regulation

INSTITUT de BIOLOGIA EVOLUTIVA

CSIC upf.

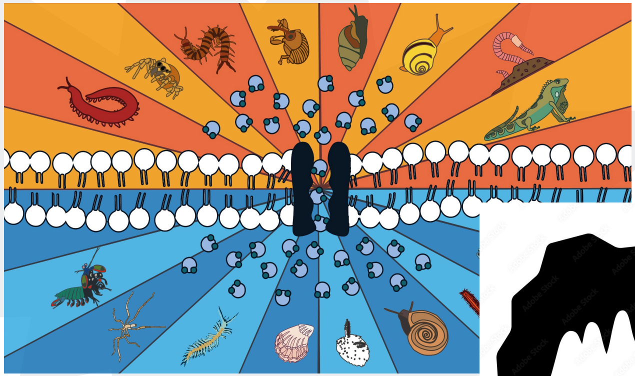
Metazoa Phylogenomics Lab

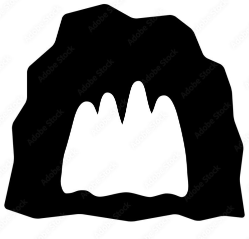
www.metazomics.com

@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)

HARVARD UNIVERSITY

CRG Centre for Genomic Regulation

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

I LOVE Trdelník (but I don't know how to pronounce it!)

A little bit about myself

Madrid (PhD)

⇩

Boston (1st postdoc)

⇩

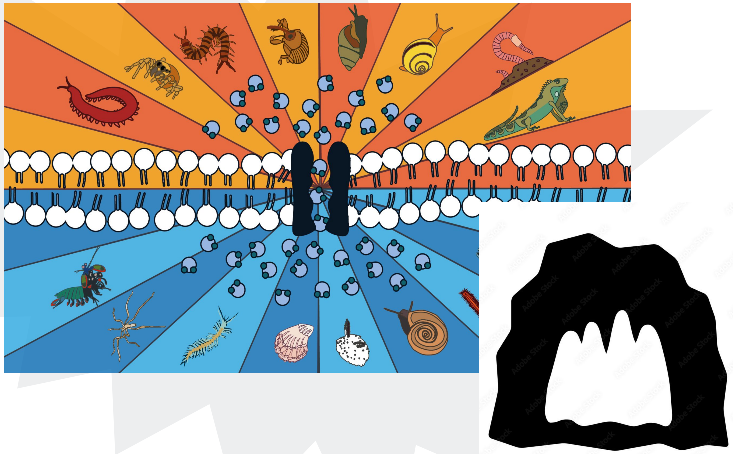
Barcelona (2nd postdoc & my lab)

Metazoa Phylogenomics Lab

www.metazomics.com

@Rosamygale
@metazomics

Main lines of research:



I've been here before :-)
(Workshop on Phylogenomics 2017)

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

I LOVE Trdelník (but I don't know how to pronounce it!)

A little bit about myself

Madrid (PhD)

⇩

Boston (1st postdoc)

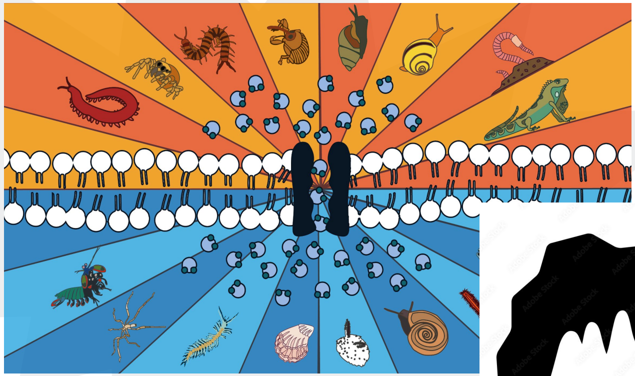
⇩

Barcelona (2nd postdoc & my lab)

www.metazomics.com

@Rosamygale
@metazomics

Main lines of research:



SPOILER ALERT: there will be BEARS!!

I've been here before :-)
(Workshop on Phylogenomics 2017)

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

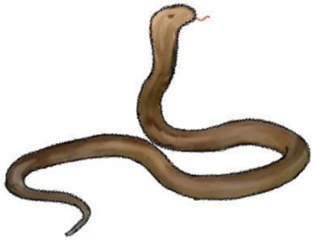
I LOVE Trdelník (but I don't know how to pronounce it!)

Which came first, the chicken or the egg?

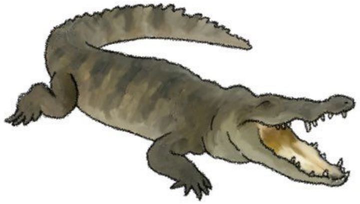
Which came first, the chicken or the egg?

Turtles

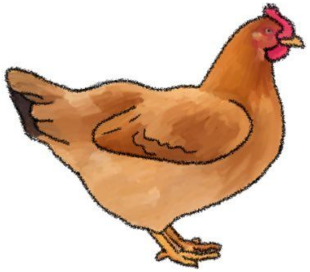
Lizards



Snakes

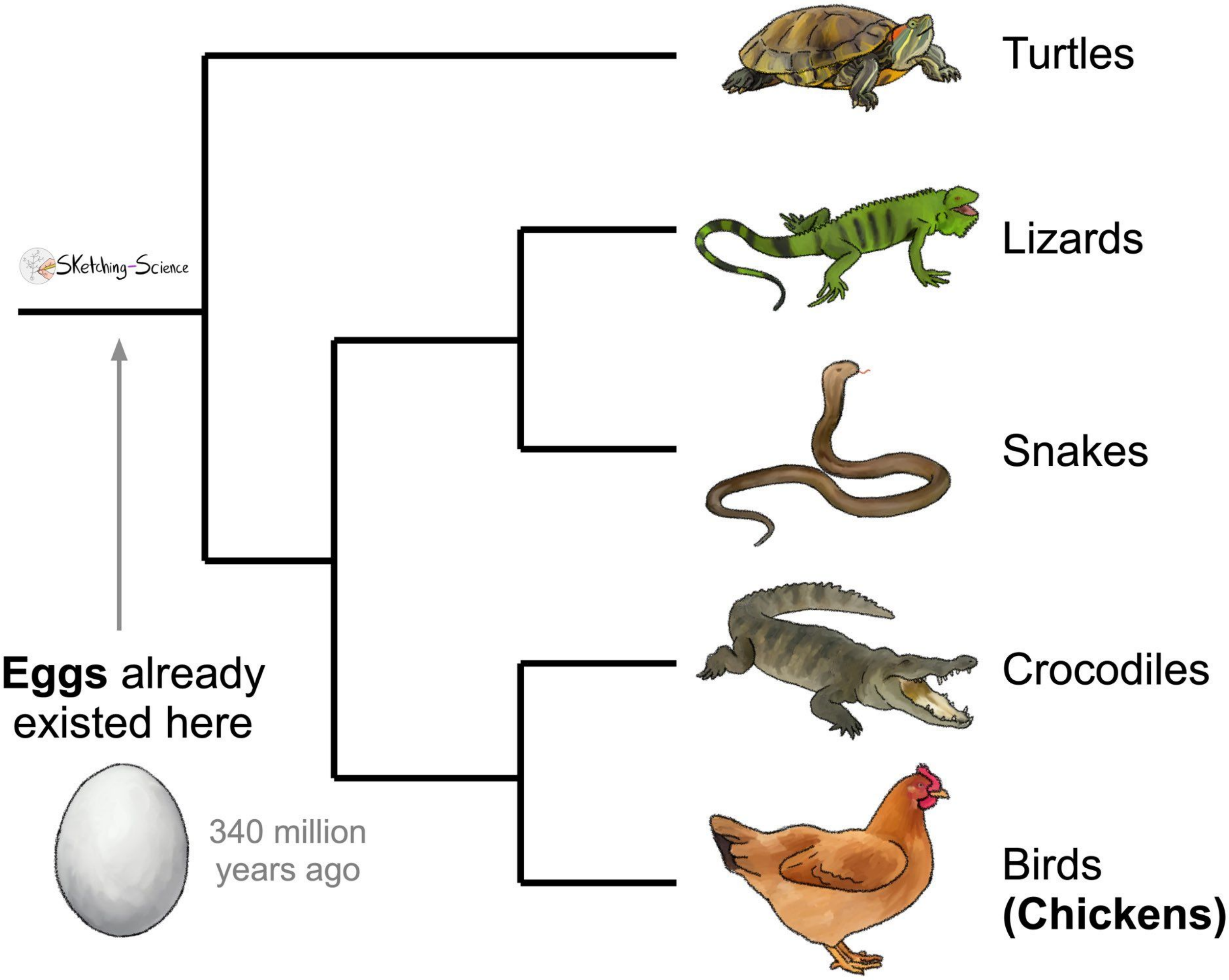


Crocodiles



Birds
(Chickens)

Which came first, the chicken or the egg?



SKetching-Science

Turtles

Lizards

Snakes

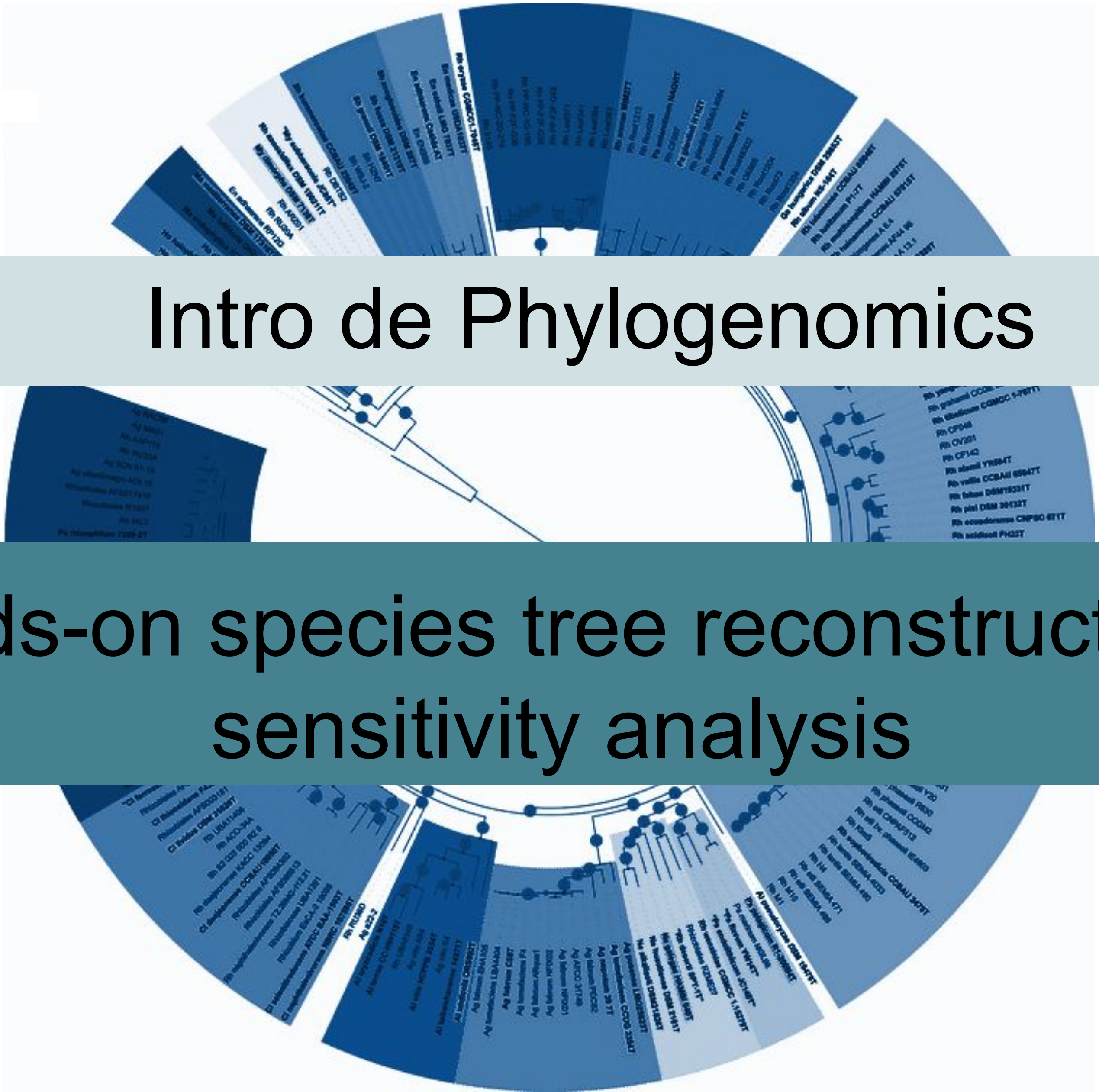
Crocodiles

Birds
(Chickens)

Eggs already existed here

340 million years ago

Content today's lecture

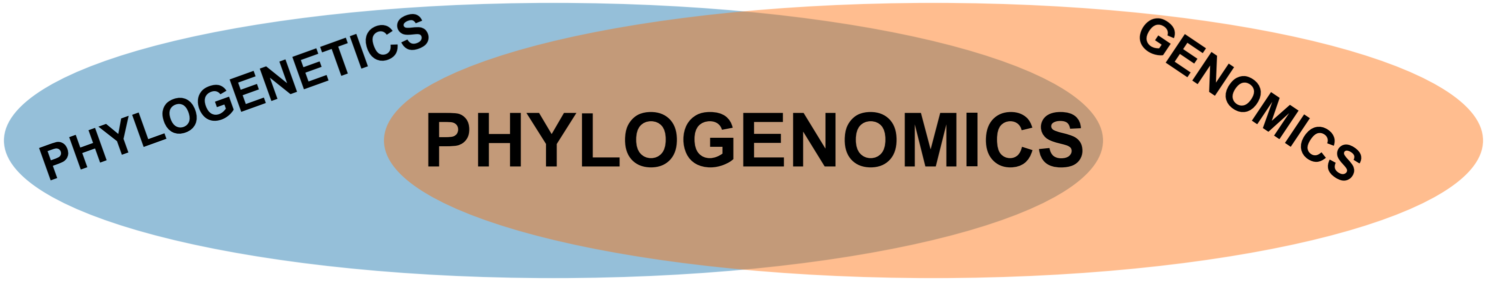


Intro de Phylogenomics

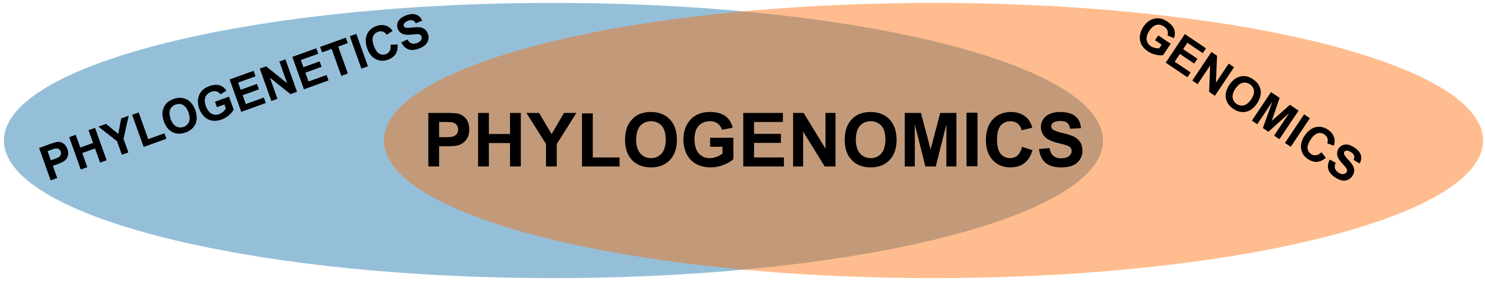
Hands-on species tree reconstruction & sensitivity analysis

Intro to Phylogenomics

Intro to Phylogenomics



Intro to Phylogenomics



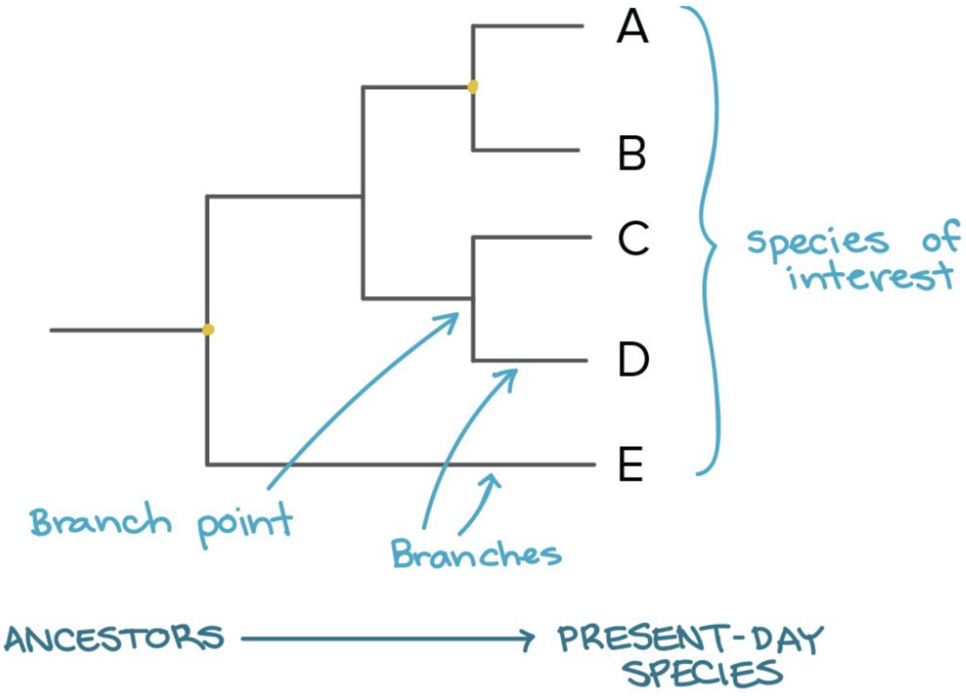
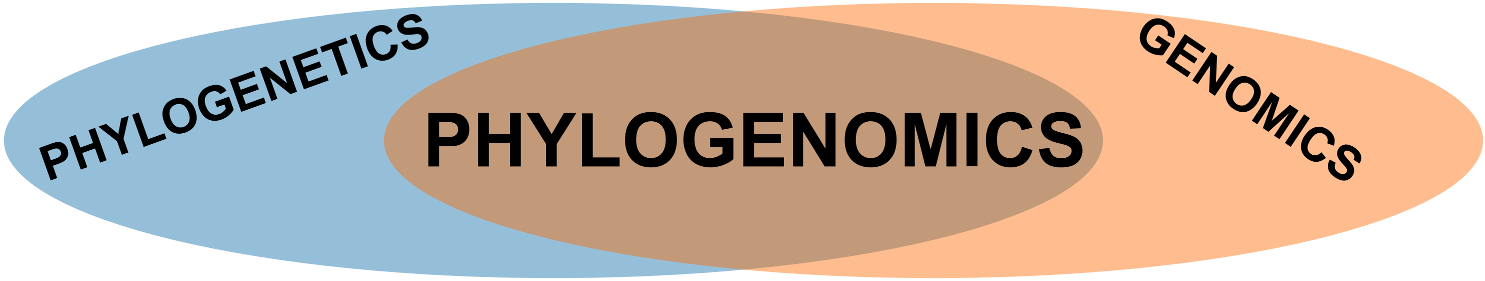


Image modified from *Taxonomy and phylogeny: Figure 2* by Robert Bear et al., CC BY 4.0

Intro to Phylogenomics



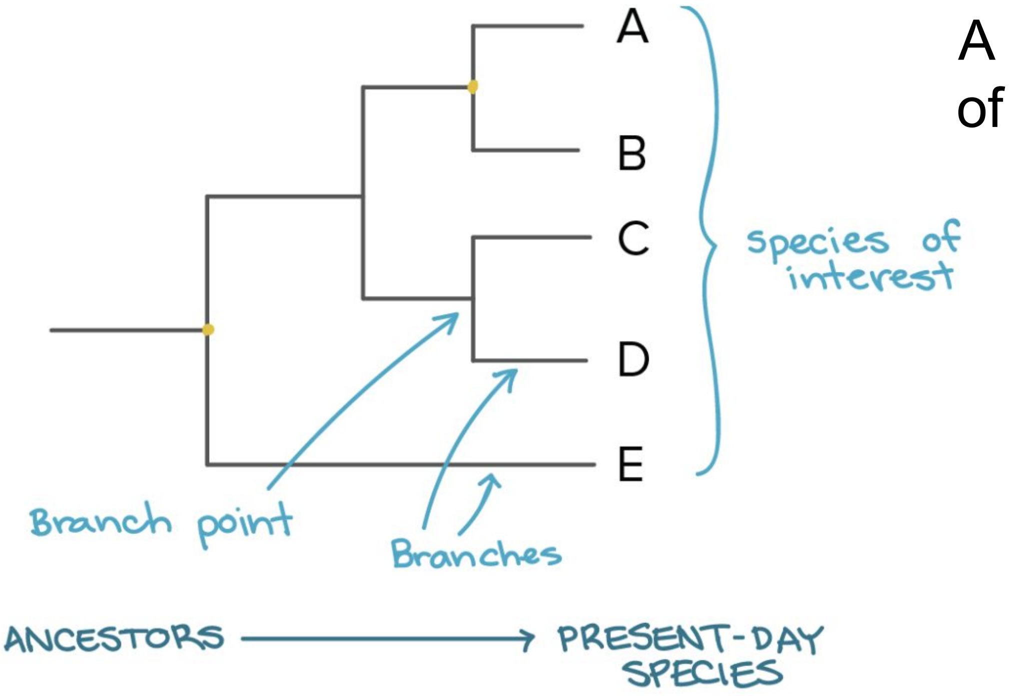
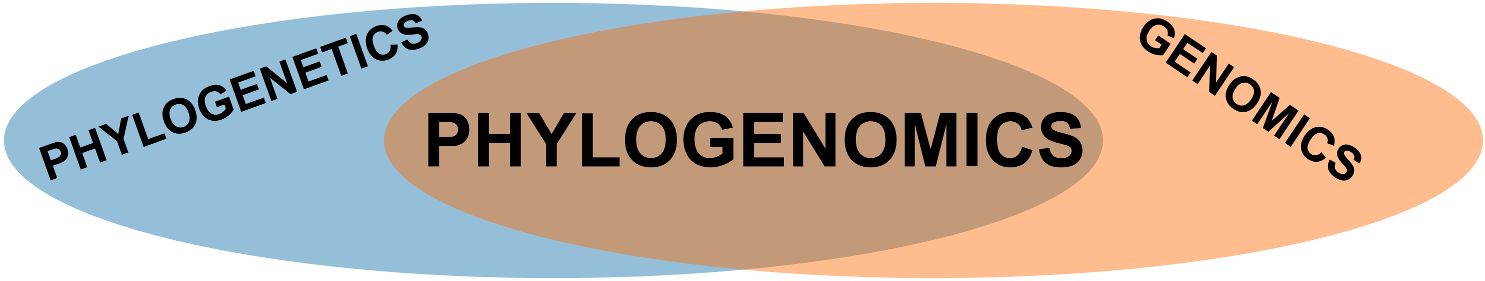


Image modified from *Taxonomy and phylogeny: Figure 2* by Robert Bear et al., CC BY 4.0

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

Intro to Phylogenomics



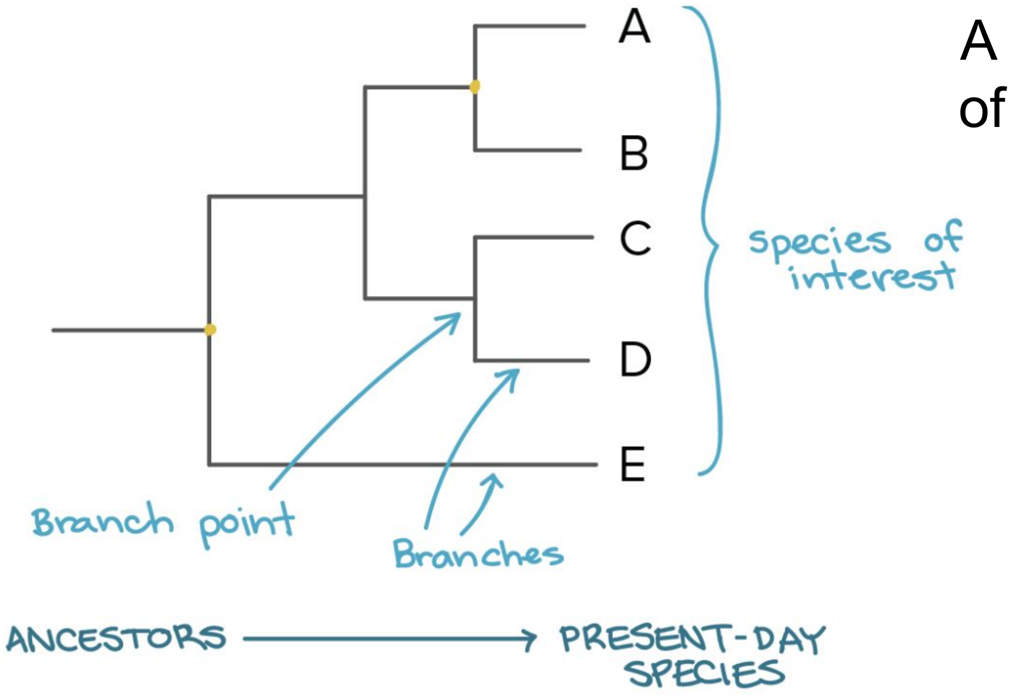
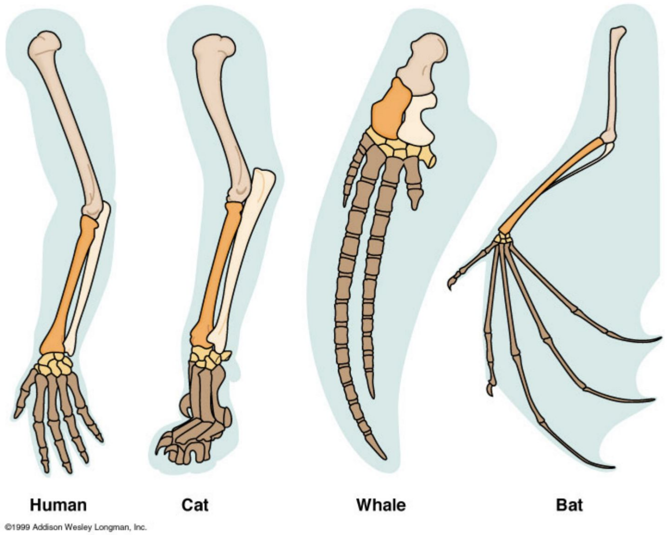


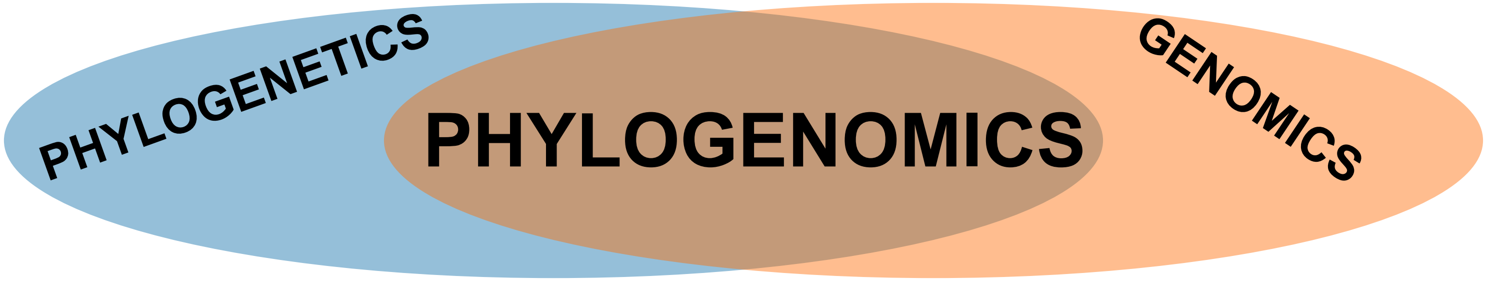
Image modified from Taxonomy and phylogeny: Figure 2 by Robert Bear et al., CC BY 4.0

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



Morphological traits

Intro to Phylogenomics



PHYLOGENETICS

PHYLOGENOMICS

GENOMICS

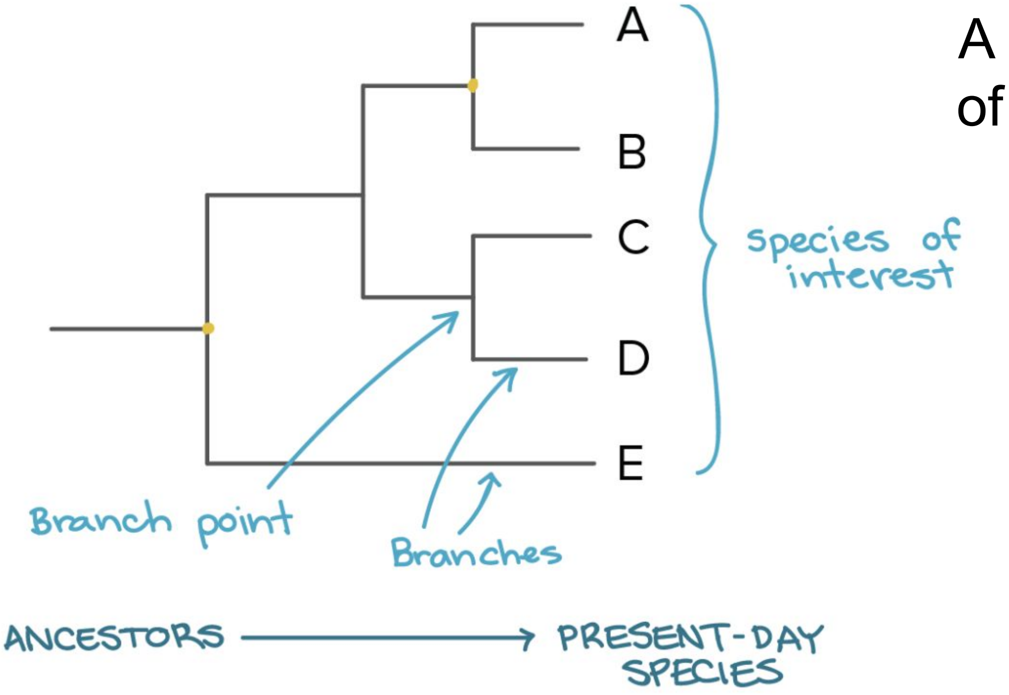
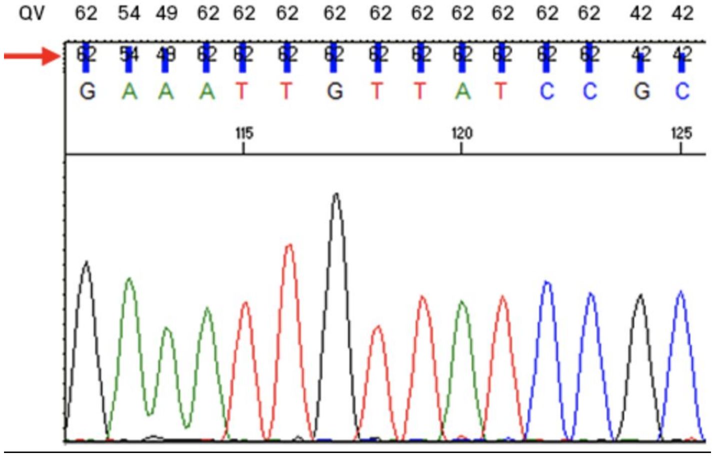


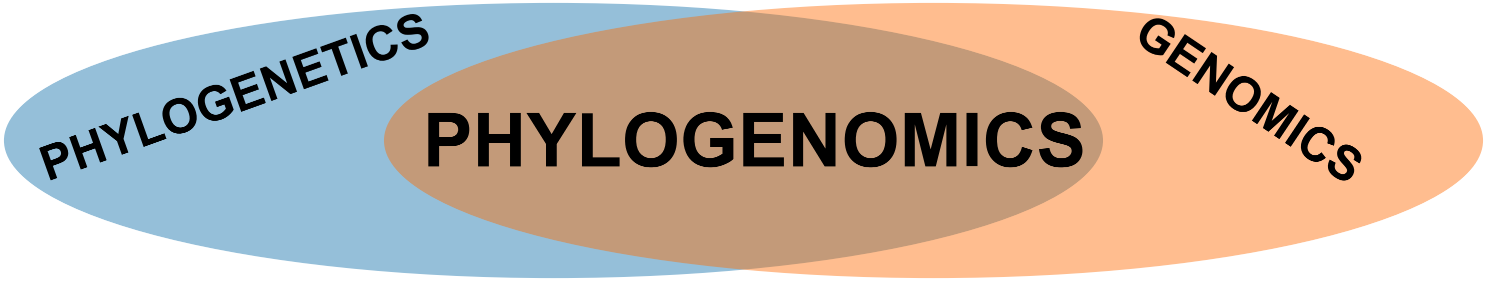
Image modified from *Taxonomy and phylogeny: Figure 2* by Robert Bear et al., CC BY 4.0

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



A few genes (eg, COI)

Intro to Phylogenomics



PHYLOGENETICS

PHYLOGENOMICS

GENOMICS

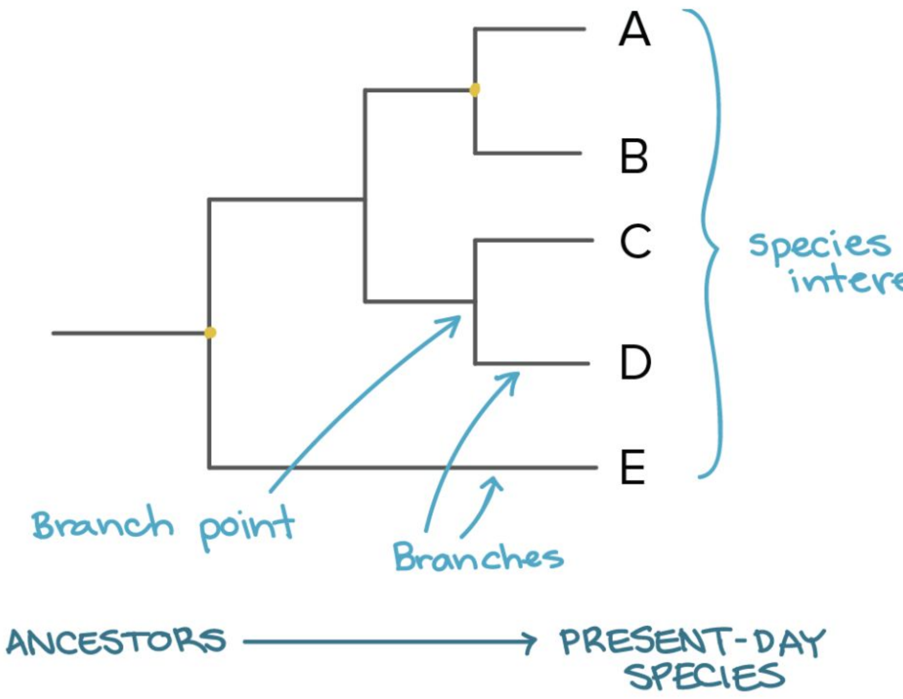
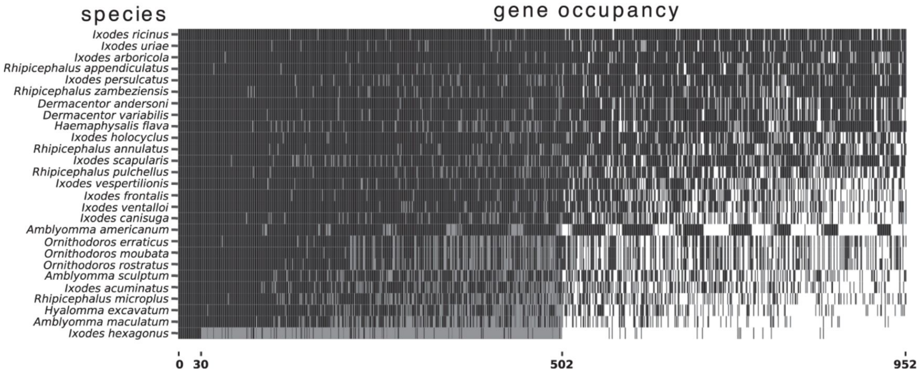


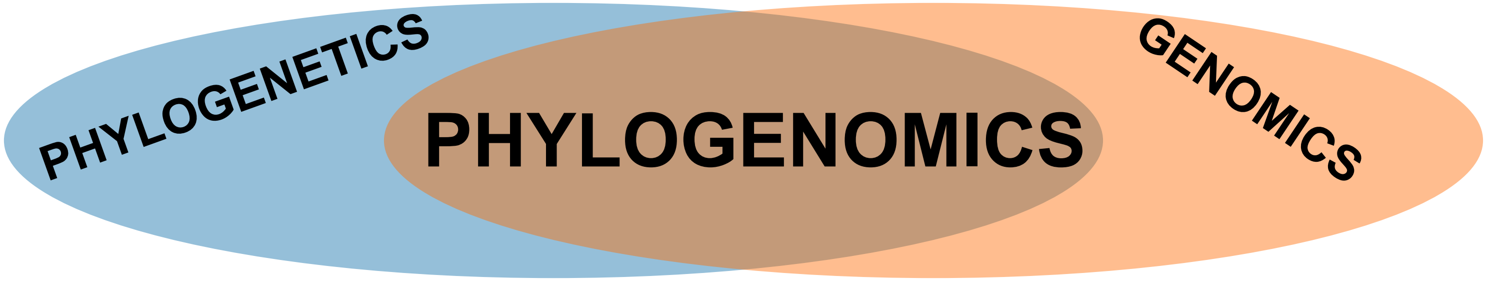
Image modified from Taxonomy and phylogeny: Figure 2 by Robert Bear et al., CC BY 4.0

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



100s / 1,000s of genes

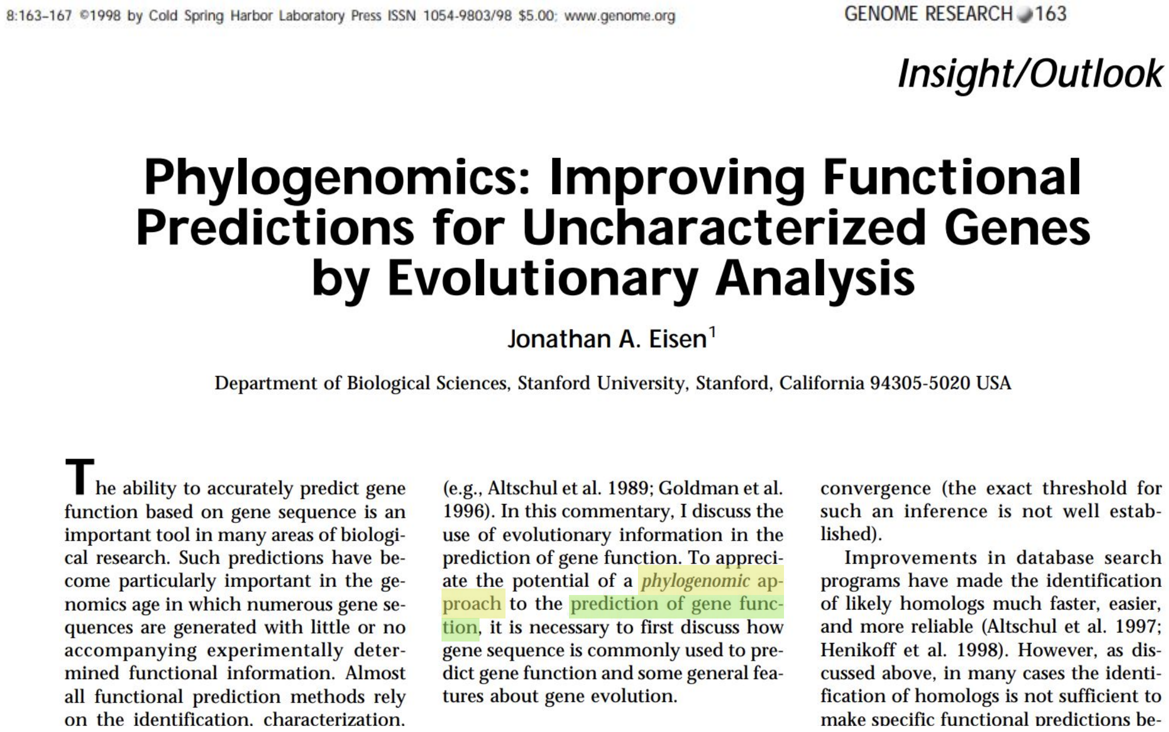
Intro to Phylogenomics



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. Eisen[1]

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,

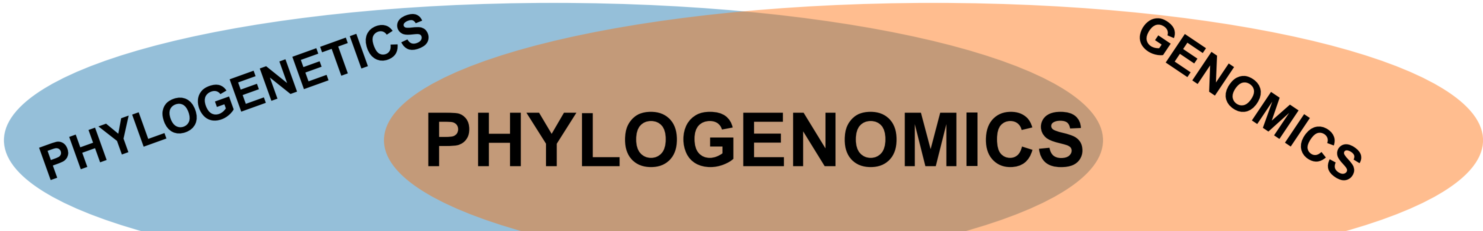
(e.g., Altschul et al. 1989; Goldman et al. 1996). In this commentary, I discuss the use of evolutionary information in the prediction of gene function. To appreciate the potential of a *phylogenomic* approach to the prediction of gene function, it is necessary to first discuss how gene sequence is commonly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

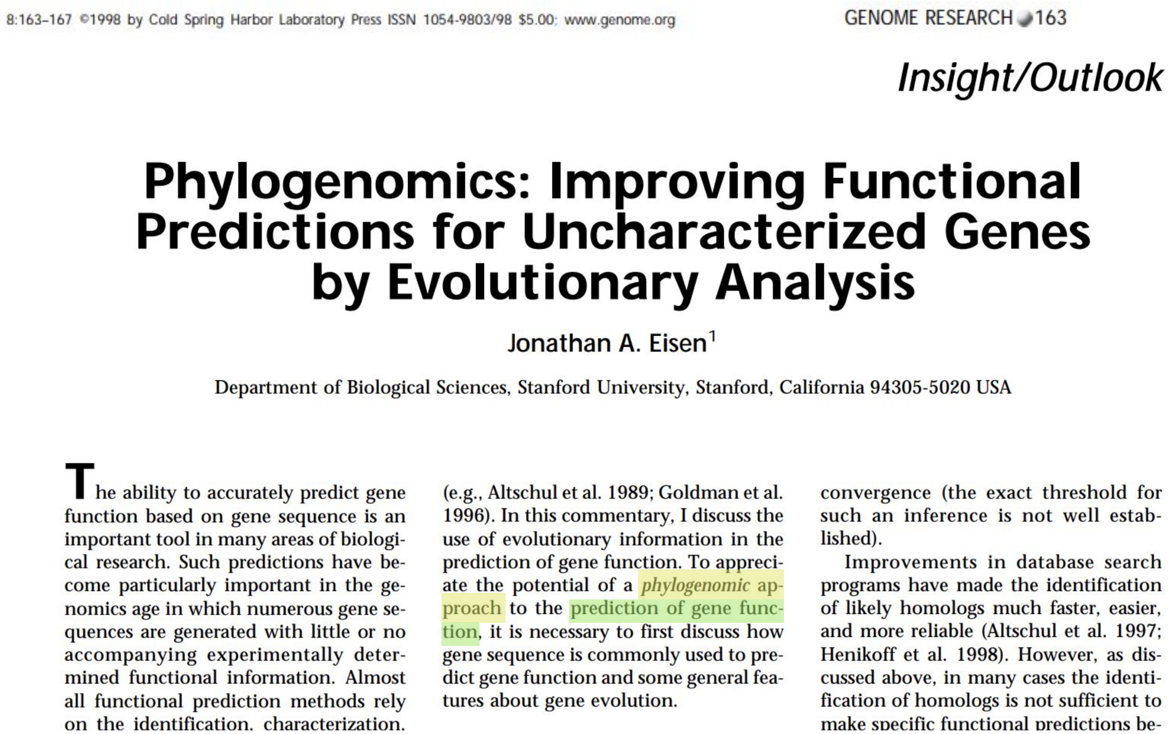
convergence (the exact threshold for such an inference is not well established).

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. Eisen[1]

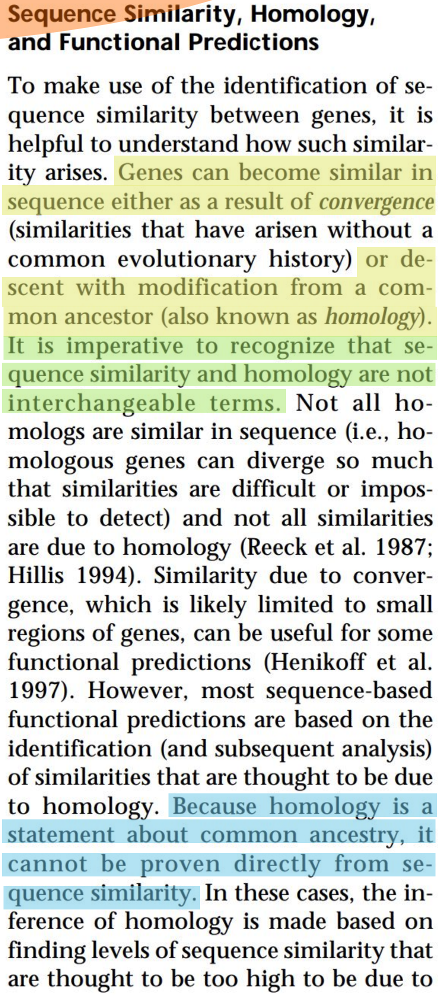
Department of Biological Sciences, Stanford University, Stanford, California 94305-5020 USA

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).

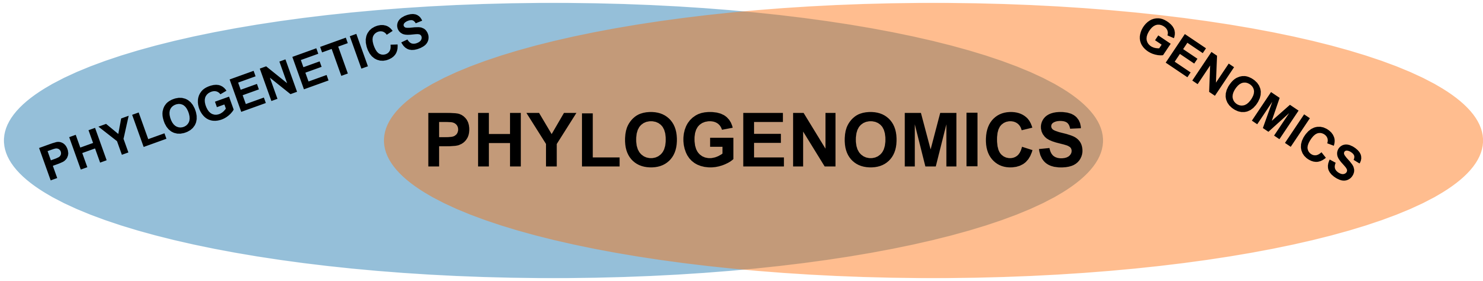
Improvements in database search programs have made the identification of likely homologs much faster, easier, and more reliable (Altschul et al. 1997; Henikoff et al. 1998). However, as discussed above, in many cases the identification of homologs is not sufficient to make specific functional predictions 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. Eisen[1]

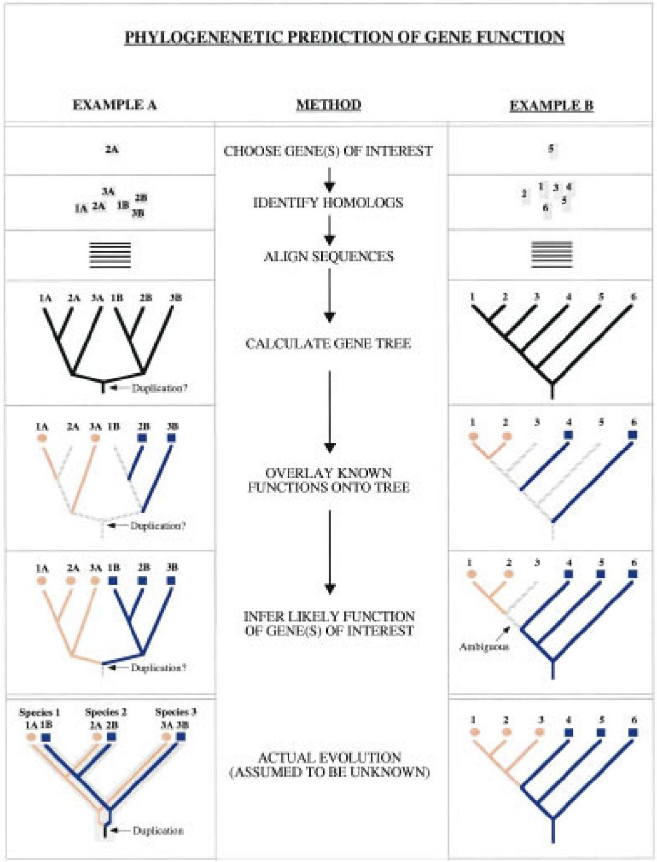
Department of Biological Sciences, Stanford University, Stanford, California 94305-5020 USA

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



PHYLOGENENETIC PREDICTION OF GENE FUNCTION

EXAMPLE A

METHOD

EXAMPLE B

CHOOSE GENE(S) OF INTEREST

IDENTIFY HOMOLOGS

ALIGN SEQUENCES

CALCULATE GENE TREE

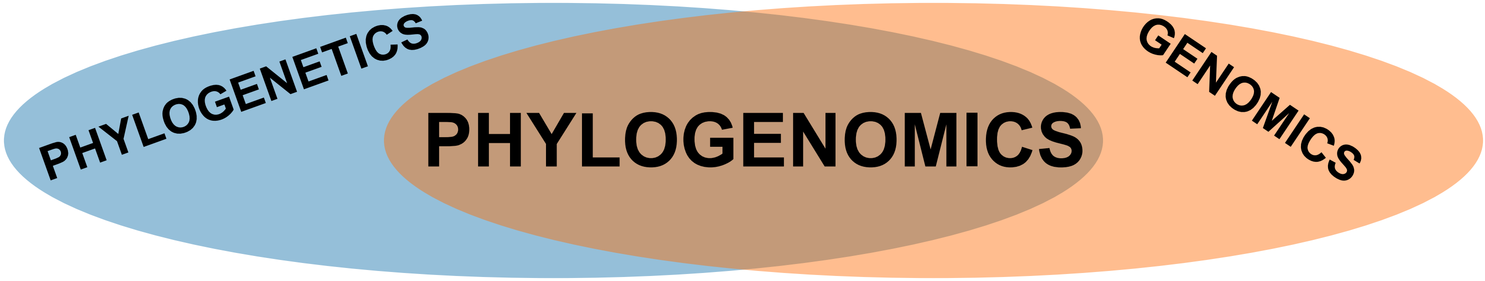
OVERLAY KNOWN FUNCTIONS ONTO TREE

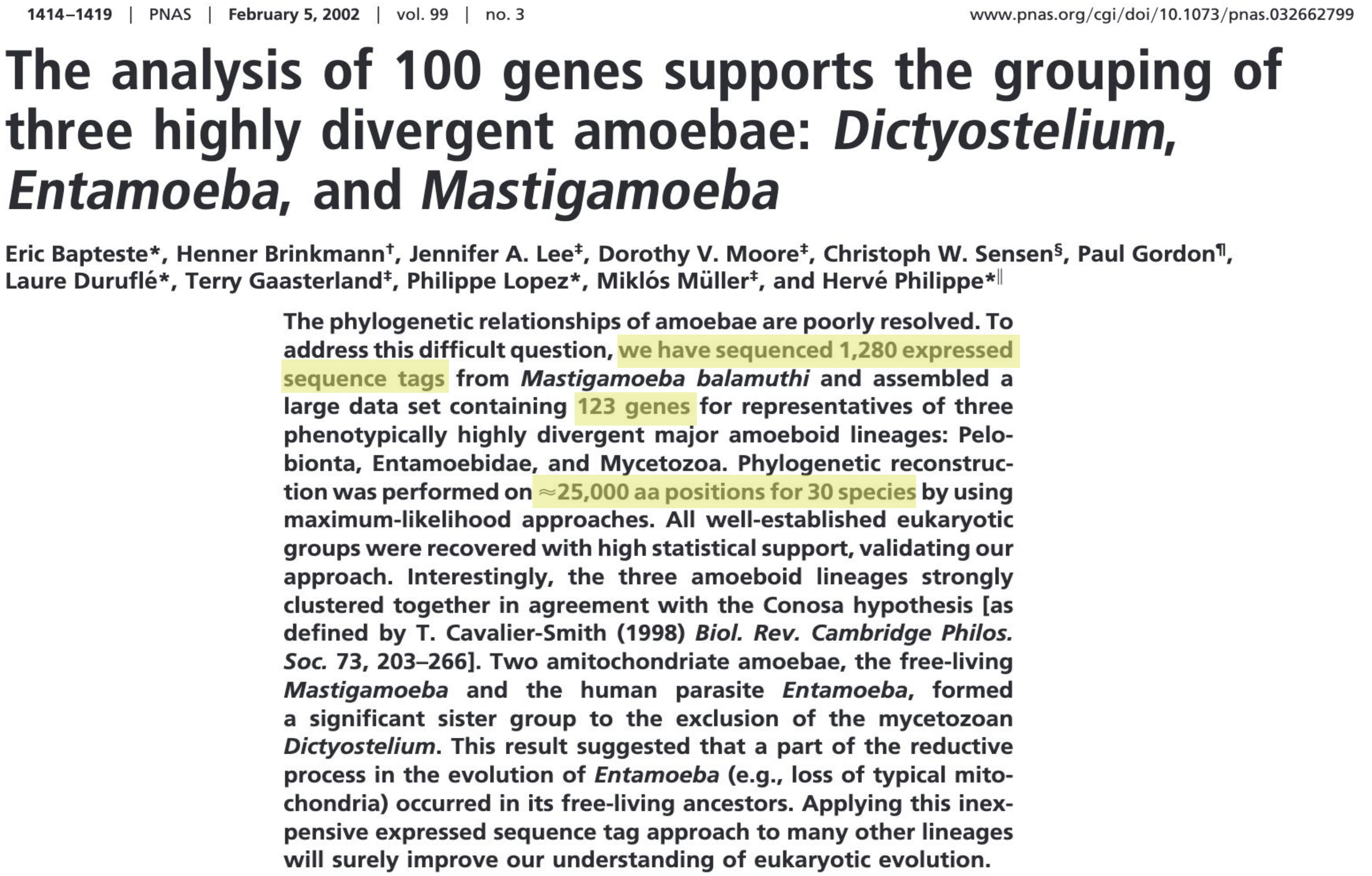
INFER LIKELY FUNCTION OF GENE(S) OF INTEREST

ACTUAL EVOLUTION (ASSUMED TO BE UNKNOWN)

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

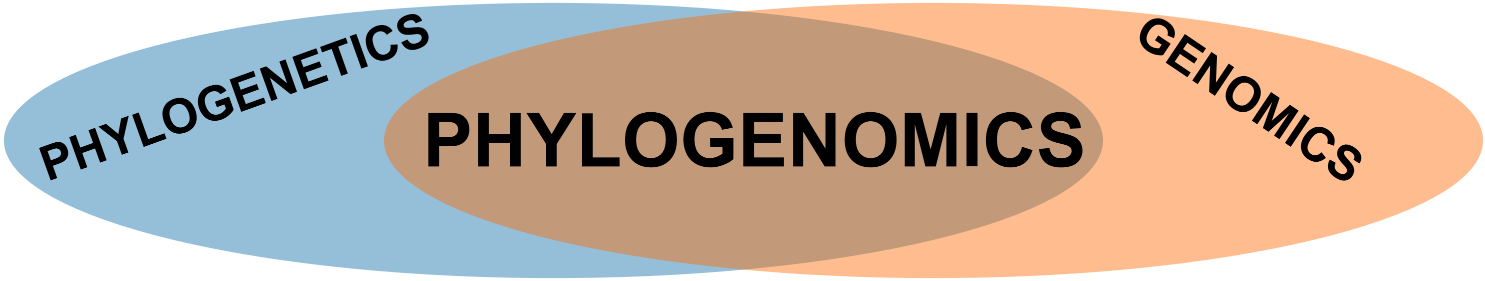
The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*

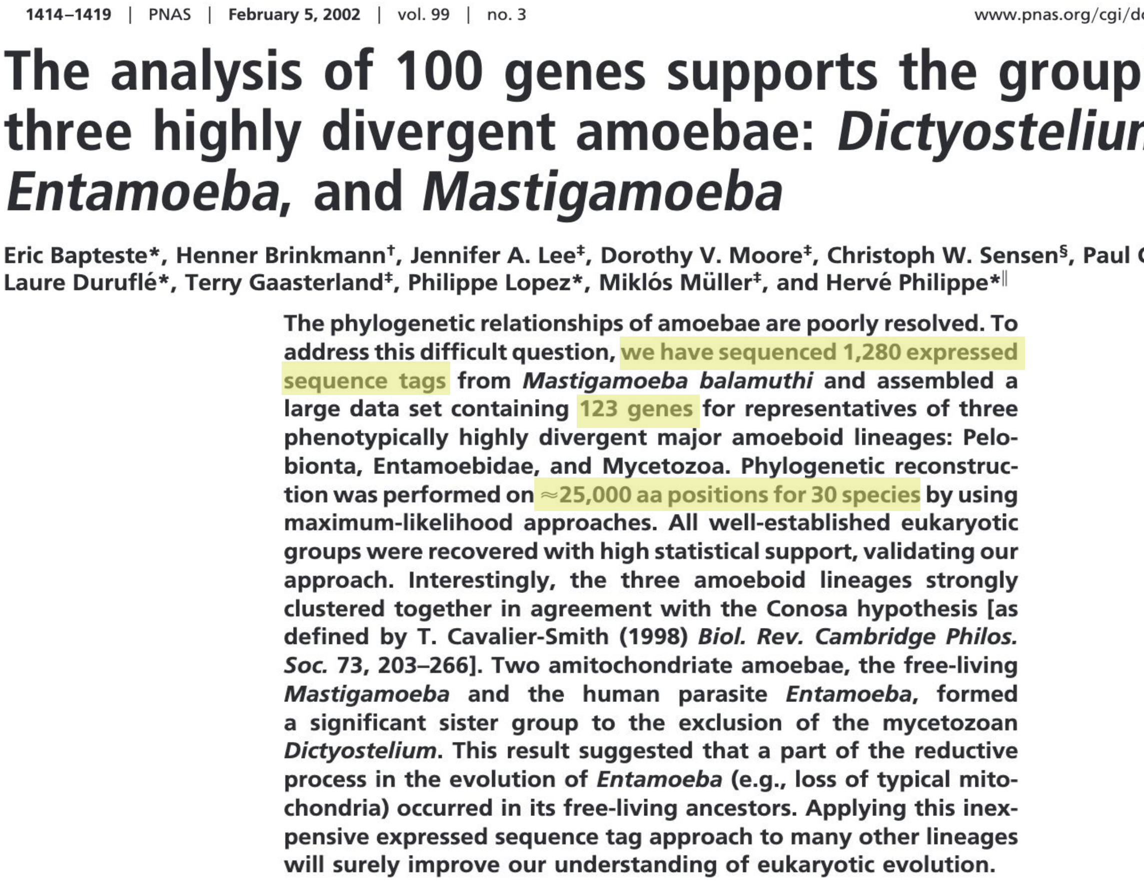
Eric Bapteste*, Henner Brinkmann†, Jennifer A. Lee‡, Dorothy V. Moore‡, Christoph W. Sensen§, Paul Gordon¶, Laure Duruflé*, Terry Gaasterland‡, Philippe Lopez*, Miklós Müller‡, and Hervé Philippe*||

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

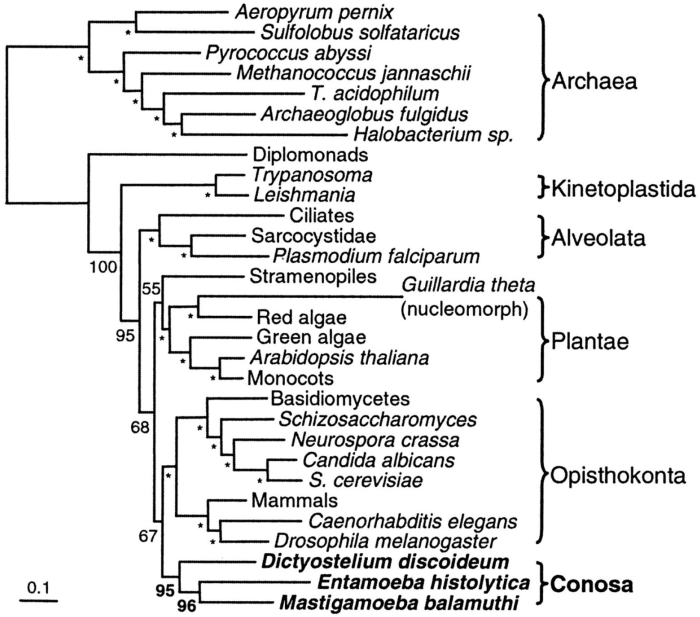



1414–1419 | PNAS | February 5, 2002 | vol. 99 | no. 3 www.pnas.org/cgi/doi/10.1073/pnas.032662799

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

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

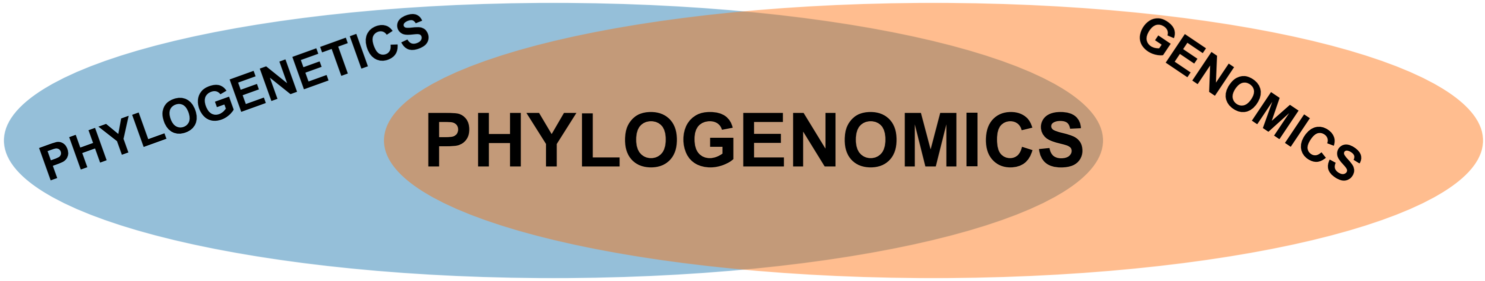
The phylogenetic relationships of amoebae are poorly resolved. To address this difficult question, we have sequenced 1,280 expressed sequence tags from *Mastigamoeba balamuthi* and assembled a large data set containing 123 genes for representatives of three phenotypically highly divergent major amoeboid 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

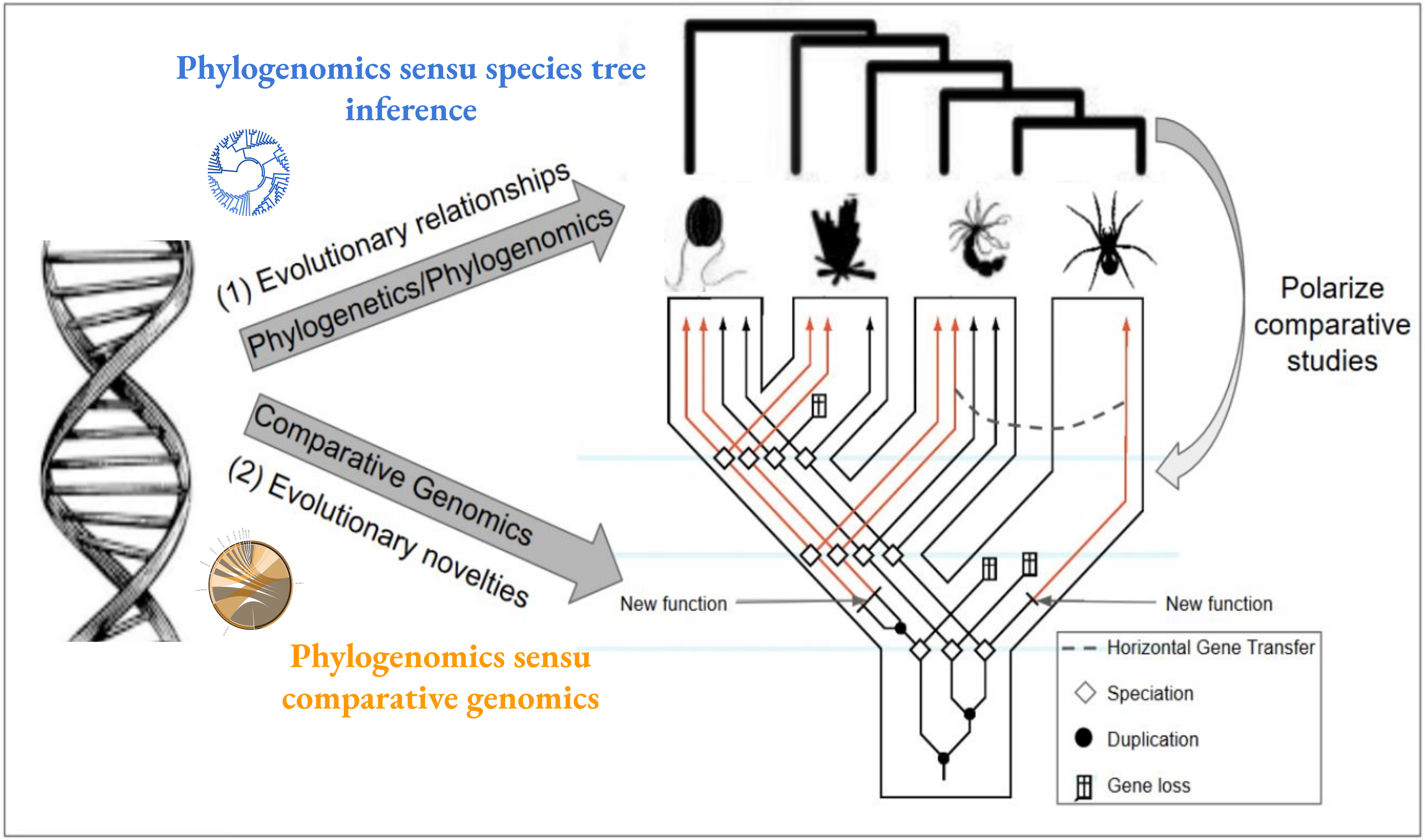
ML tree based on 25,032 aa positions. * indicates a constrained node. We used the JTT model, without taking into account among-sites rate variation. The branch lengths have been computed on the concatenated sequences. BVs were obtained by bootstrapping the 123 genes.

Intro to Phylogenomics


PHYLOGENETICS

PHYLOGENOMICS

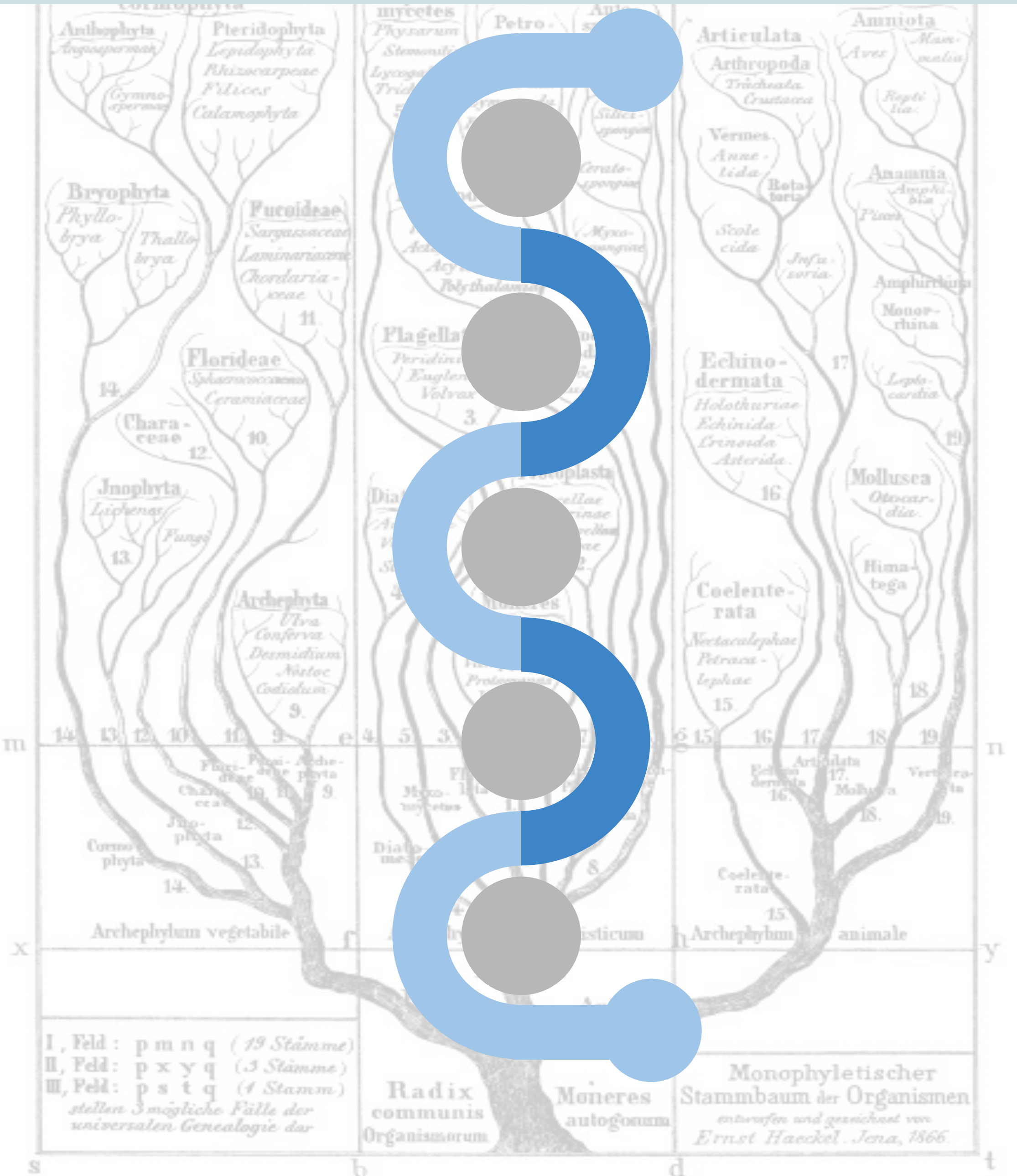
GENOMICS


Phylogenomics sensu species tree inference

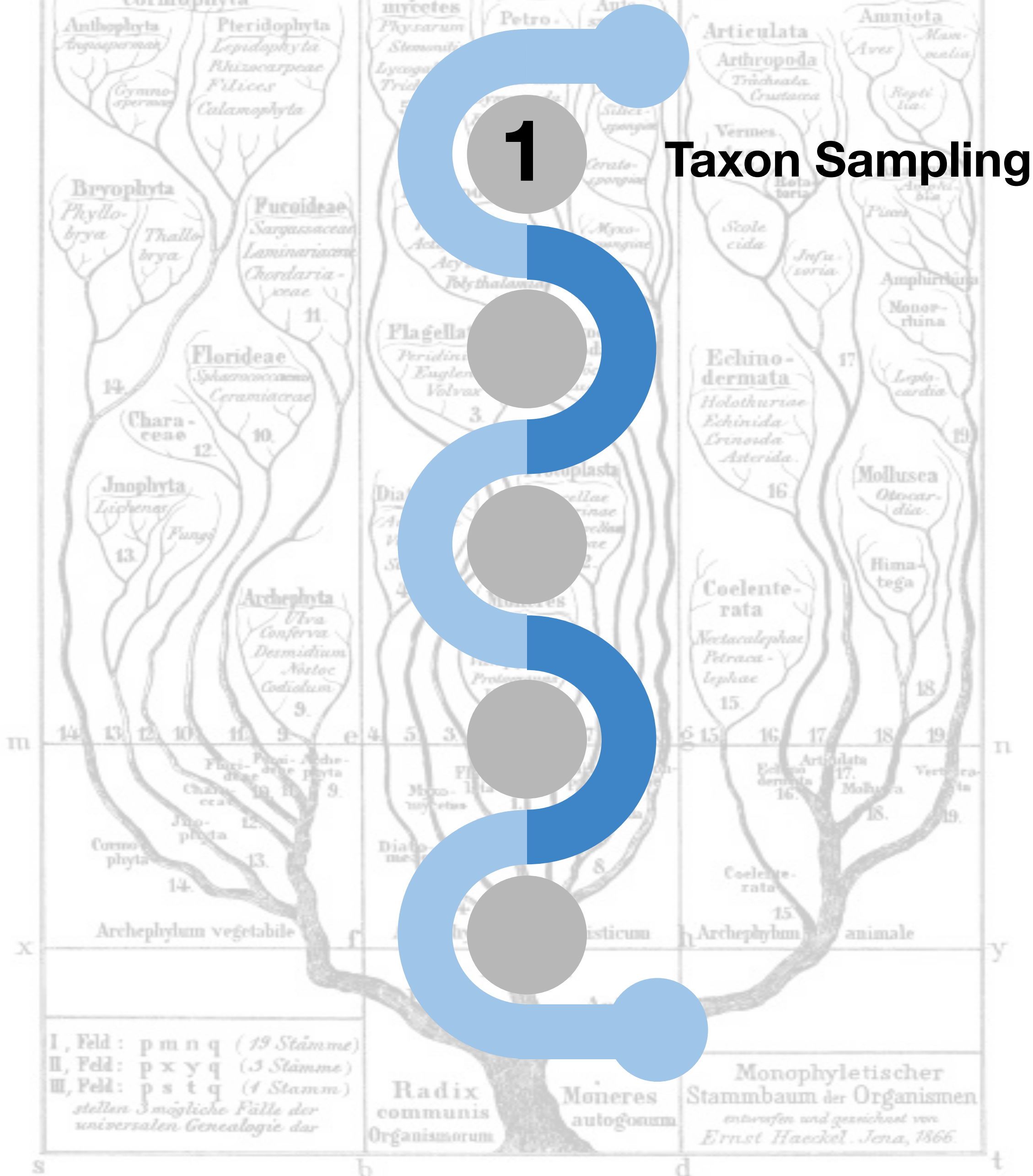
Phylogenomics sensu comparative genomics

(1) Evolutionary relationships
Phylogenetics/Phylogenomics
Comparative Genomics
(2) Evolutionary novelties
Polarize comparative studies
New function
New function
Horizontal Gene Transfer
Speciation
Duplication
Gene loss

How to infer a species tree

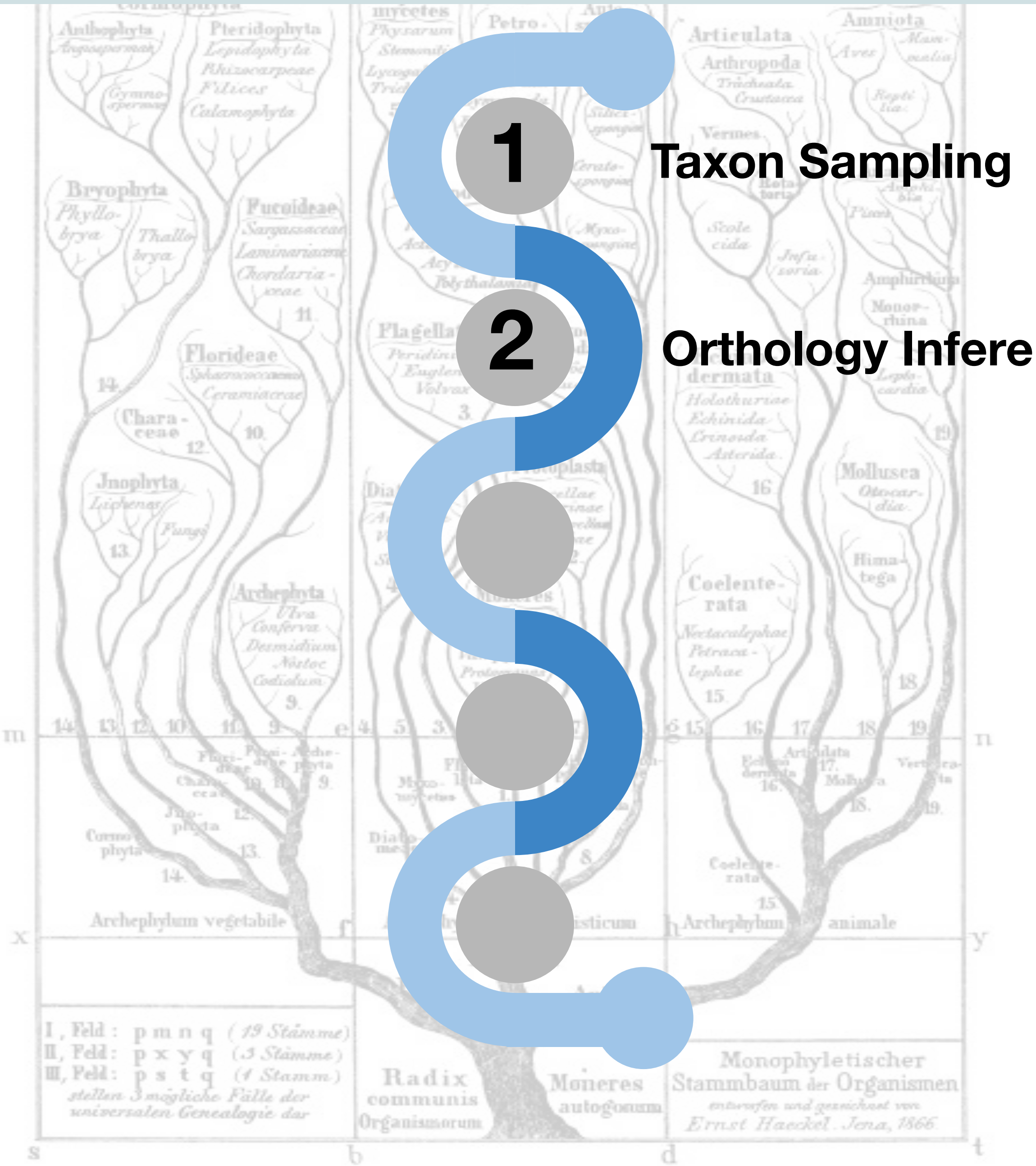


How to infer a species tree


1
Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum
Archephylum vegetabile
Archephylum animale

Taxon Sampling

How to infer a species tree

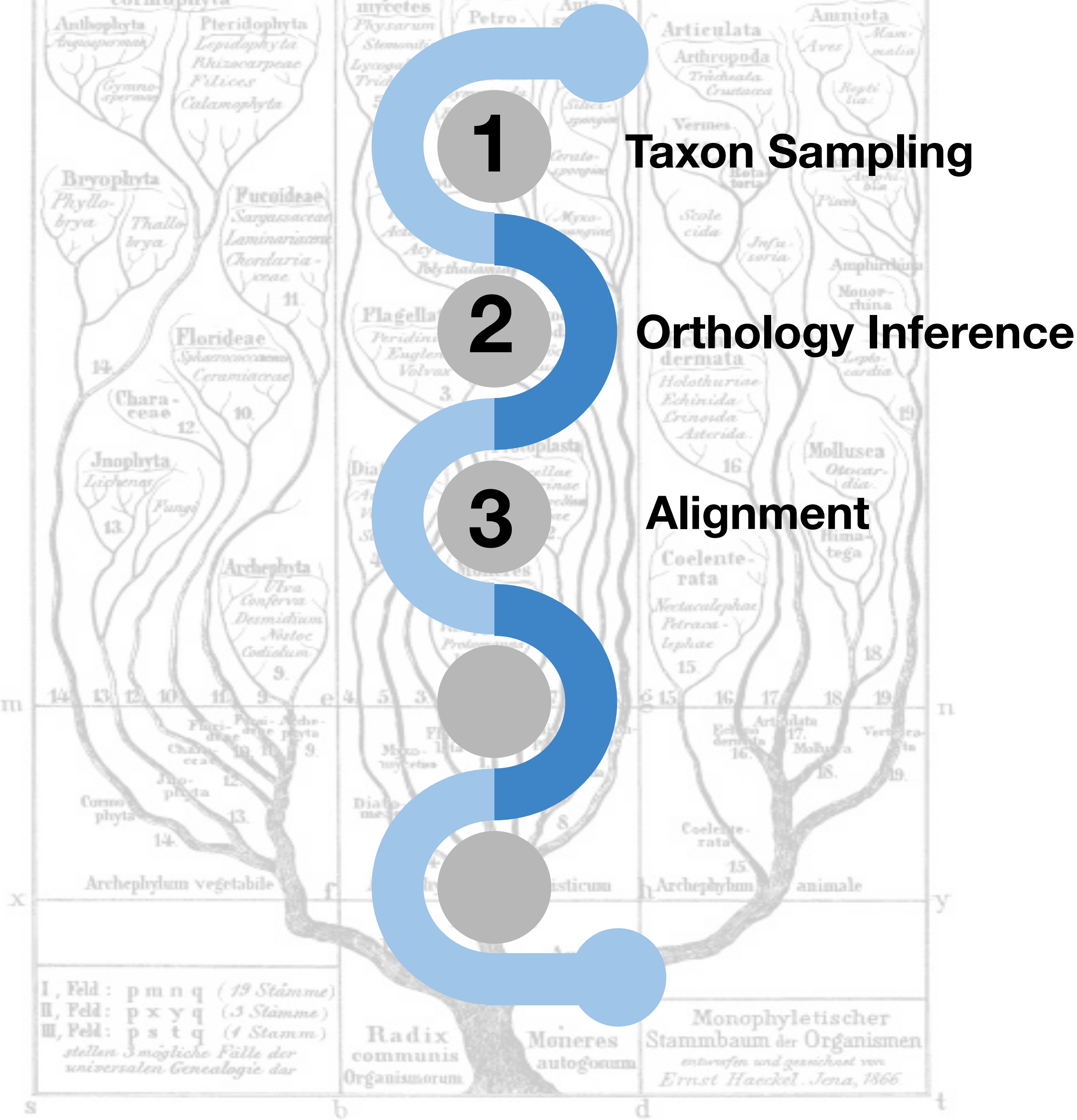


Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866.
Radix communis Organismorum
Moneres autogonum
Archephylum vegetabile
Archephylum animale

1 **Taxon Sampling**

2 **Orthology Inference**

How to infer a species tree

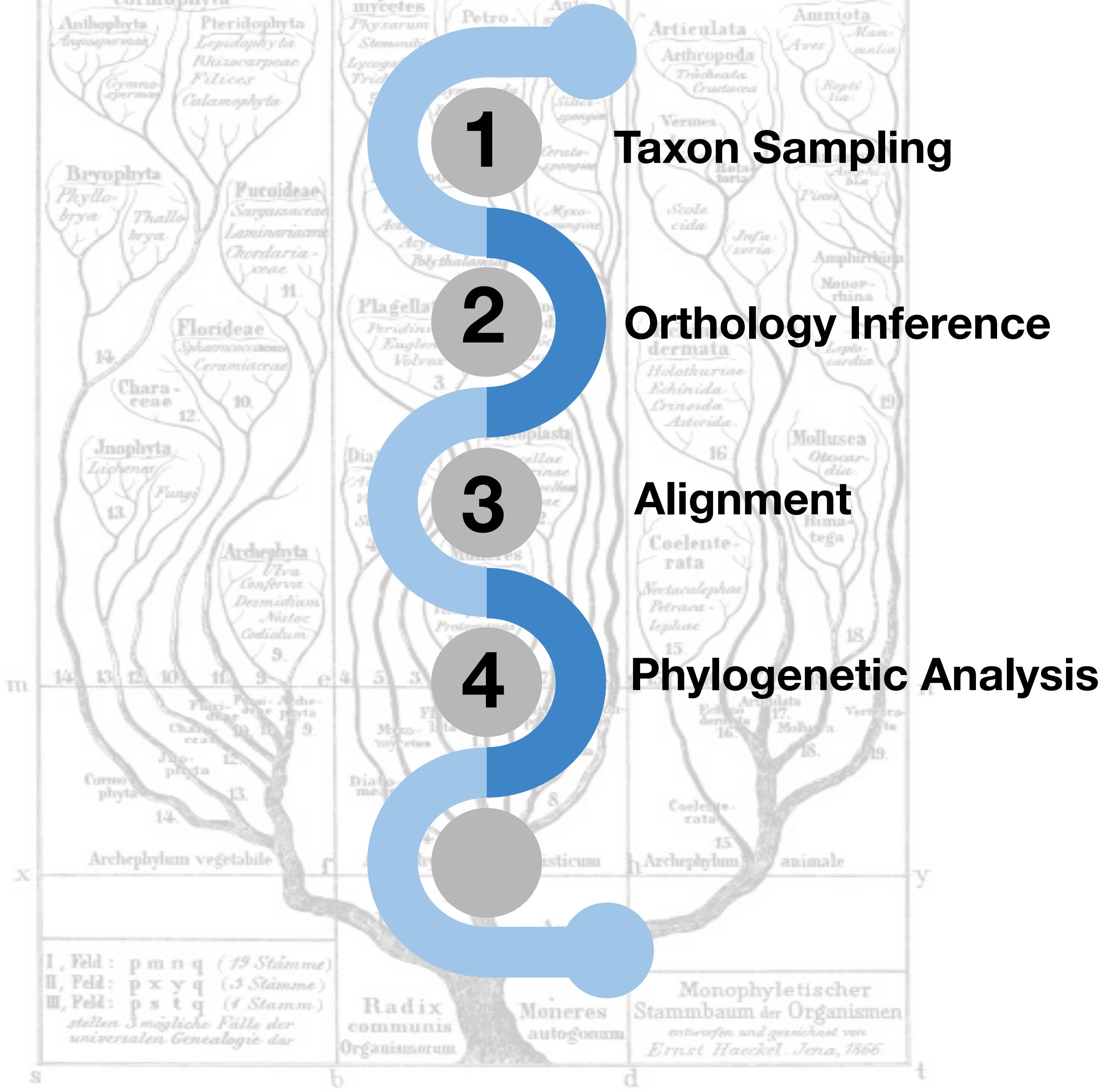

Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum
Archephylum vegetabile
Archephylum animale

1 **Taxon Sampling**

2 **Orthology Inference**

3 **Alignment**

How to infer a species tree



1 Taxon Sampling

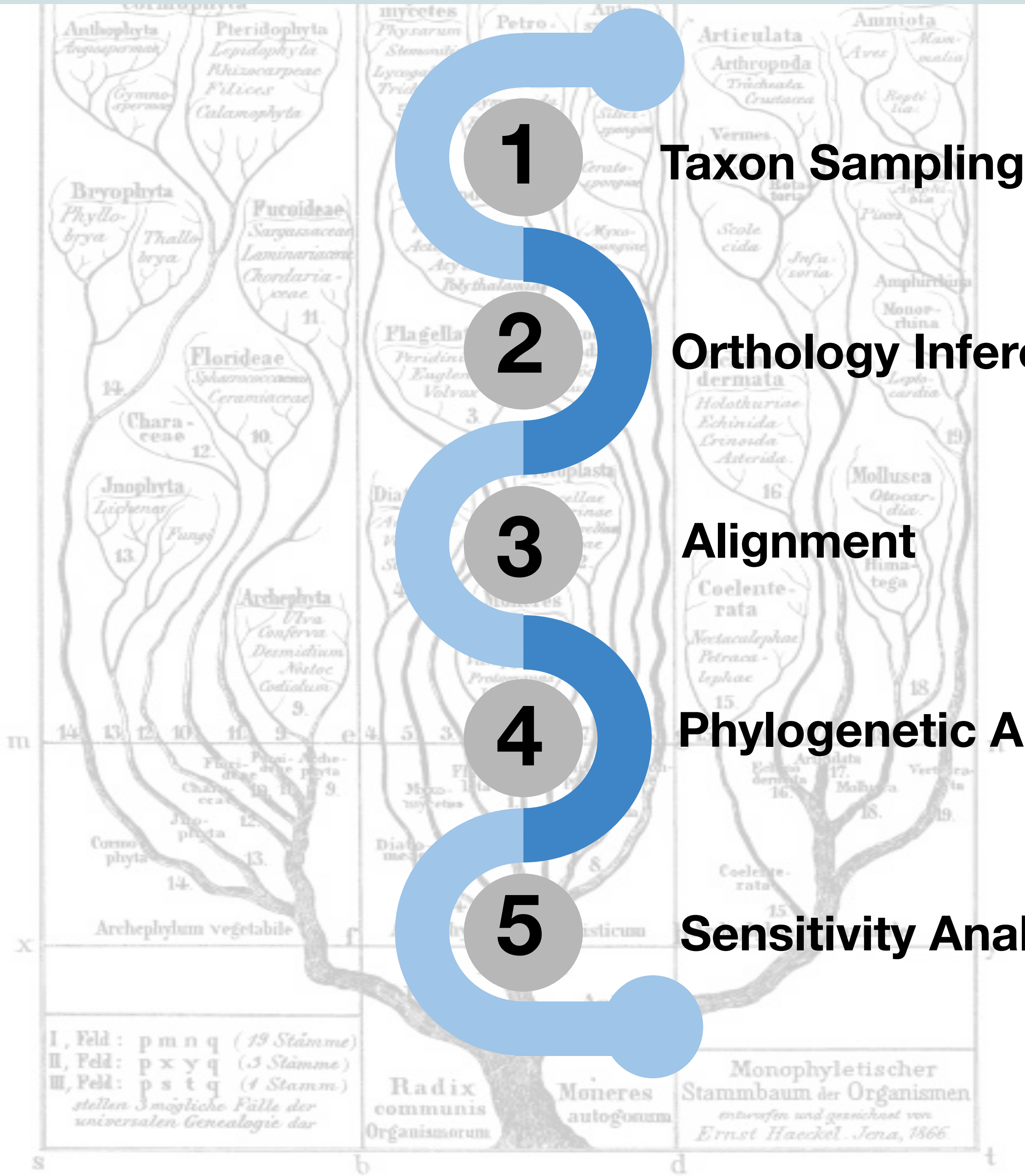
2 Orthology Inference

3 Alignment

4 Phylogenetic Analysis

Plantae
Protista
Animalia
Archephylum vegetabile
Archephylum animale
Radix communis Organismorum
Moneres autogonum
Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866

How to infer a species tree



1 Taxon Sampling

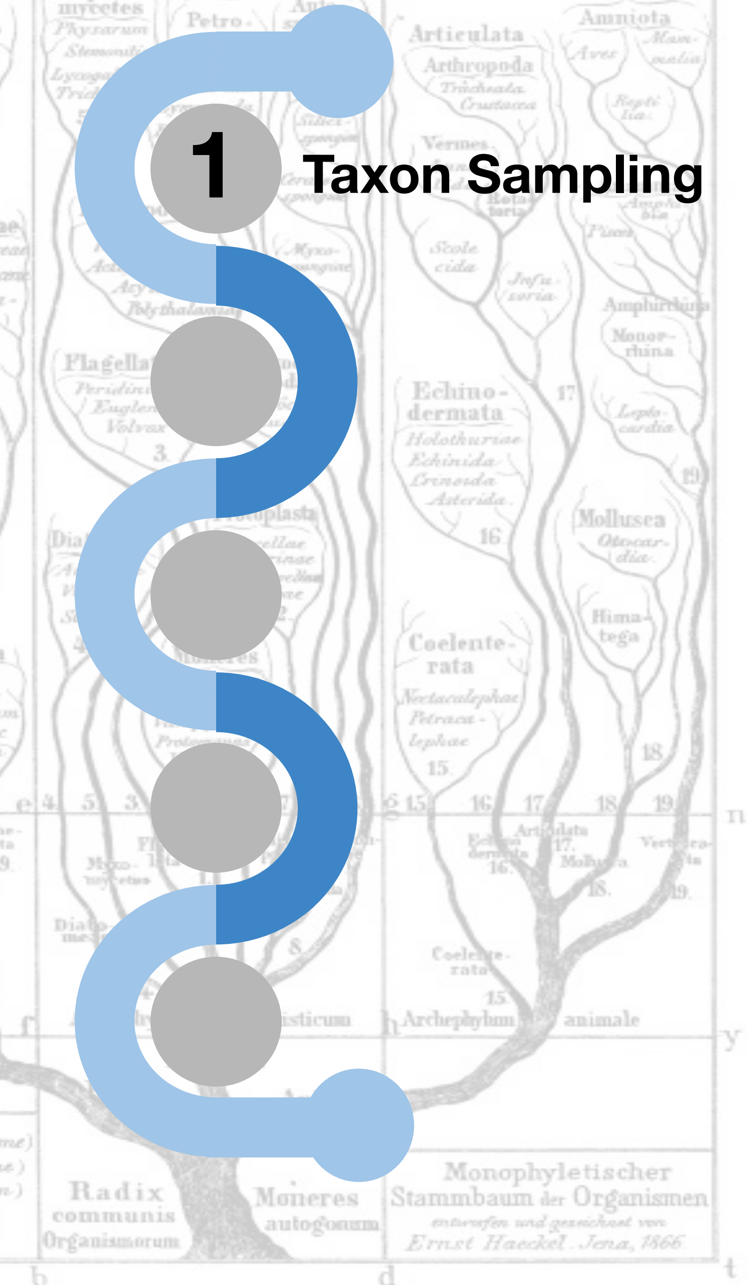
2 Orthology Inference

3 Alignment

4 Phylogenetic Analysis

5 Sensitivity Analysis

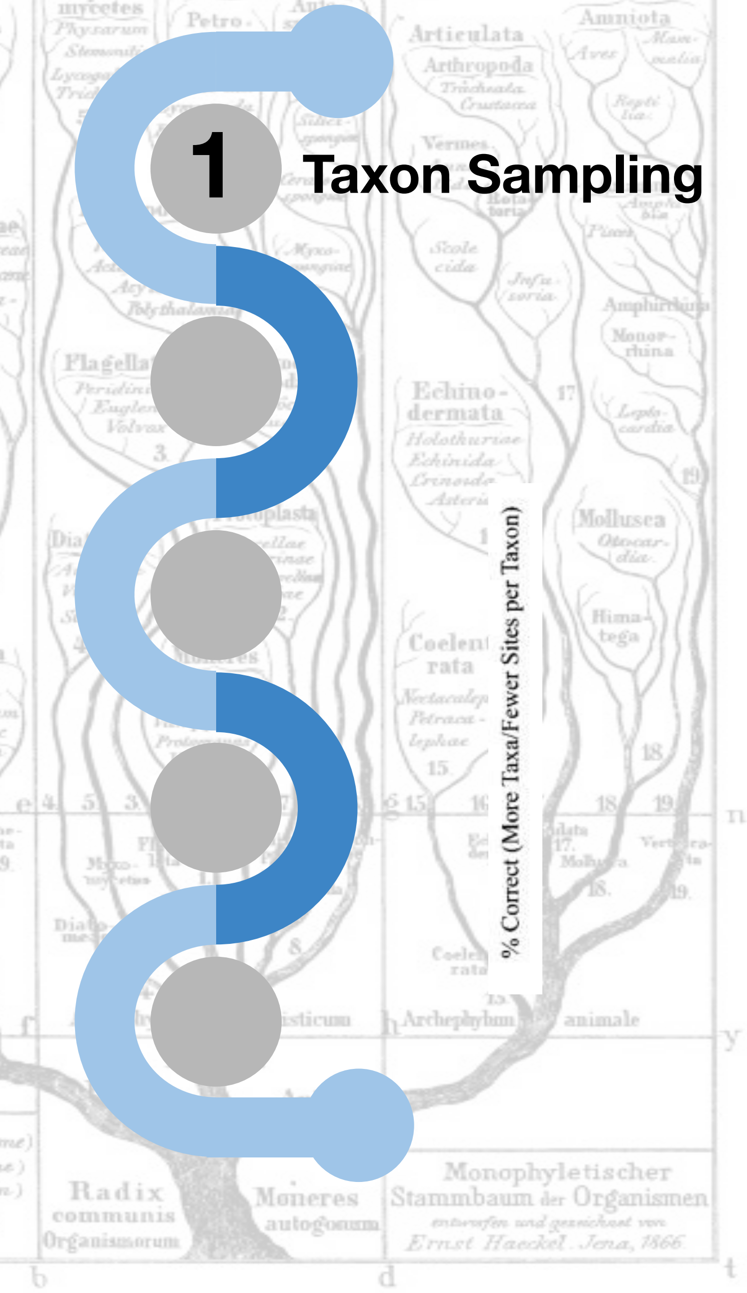
How to infer a species tree



1 Taxon Sampling

Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum

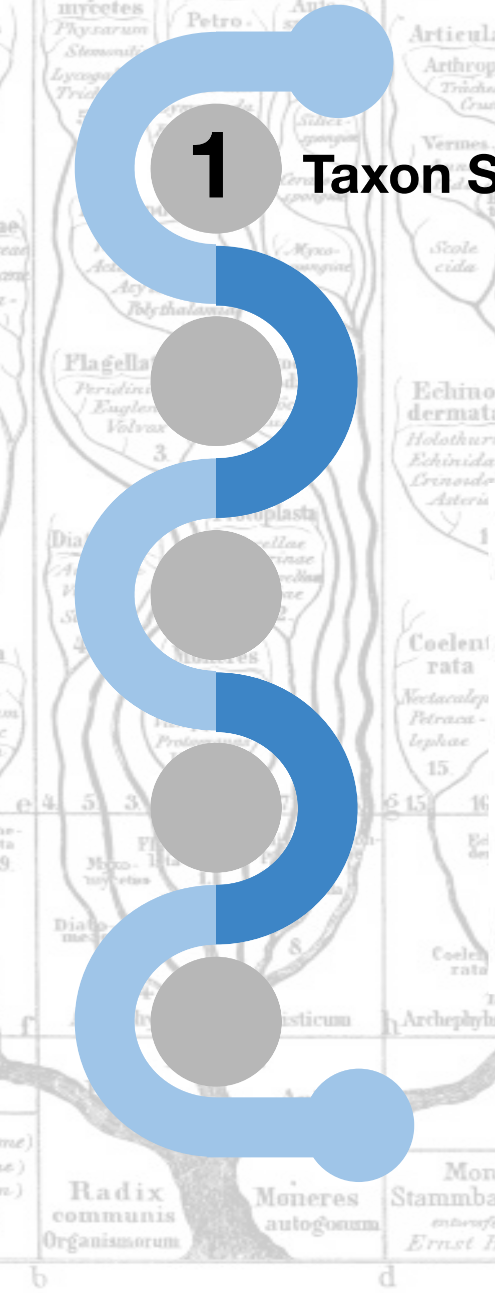
How to infer a species tree


1
Taxon Sampling
% Correct (More Taxa/Fewer Sites per Taxon)
Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum

Key message: taxon sampling matters <u>a lot</u>

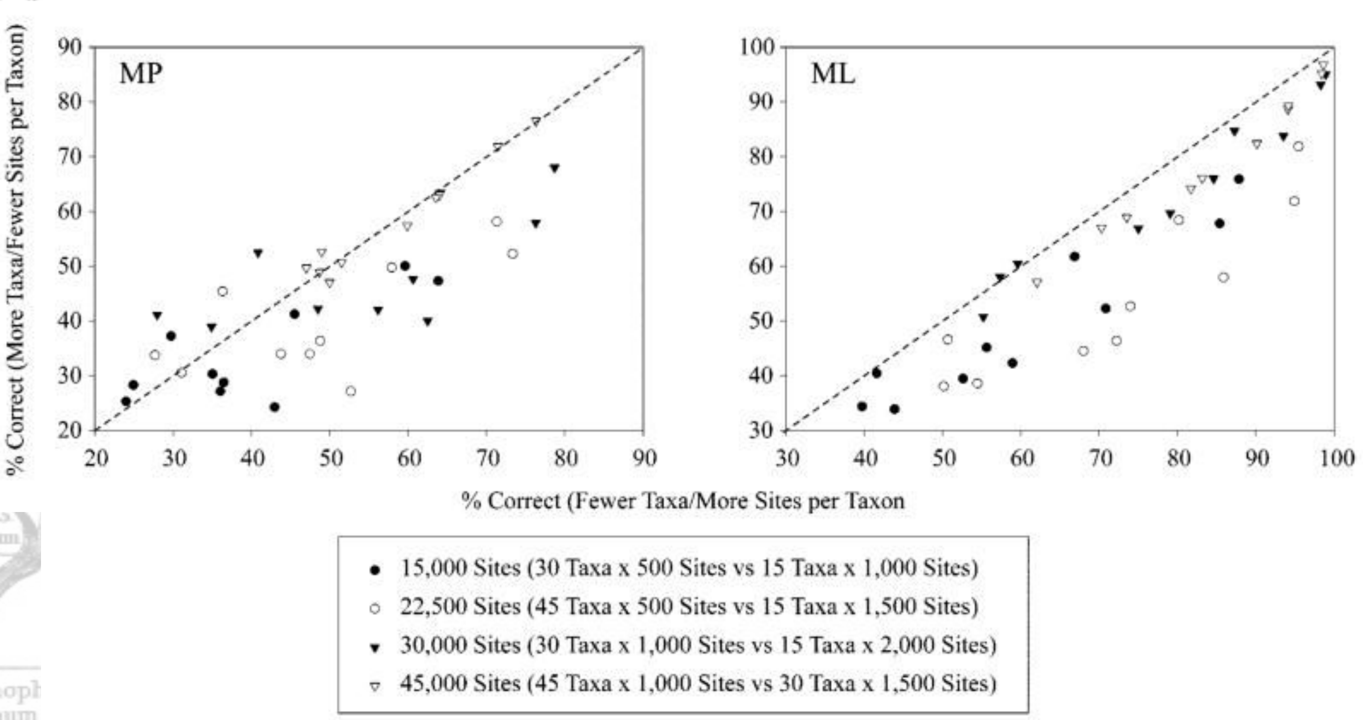
Incomplete, biased, or improper taxon sampling can lead to misleading results in reconstructing evolutionary relationships.

How to infer a species tree



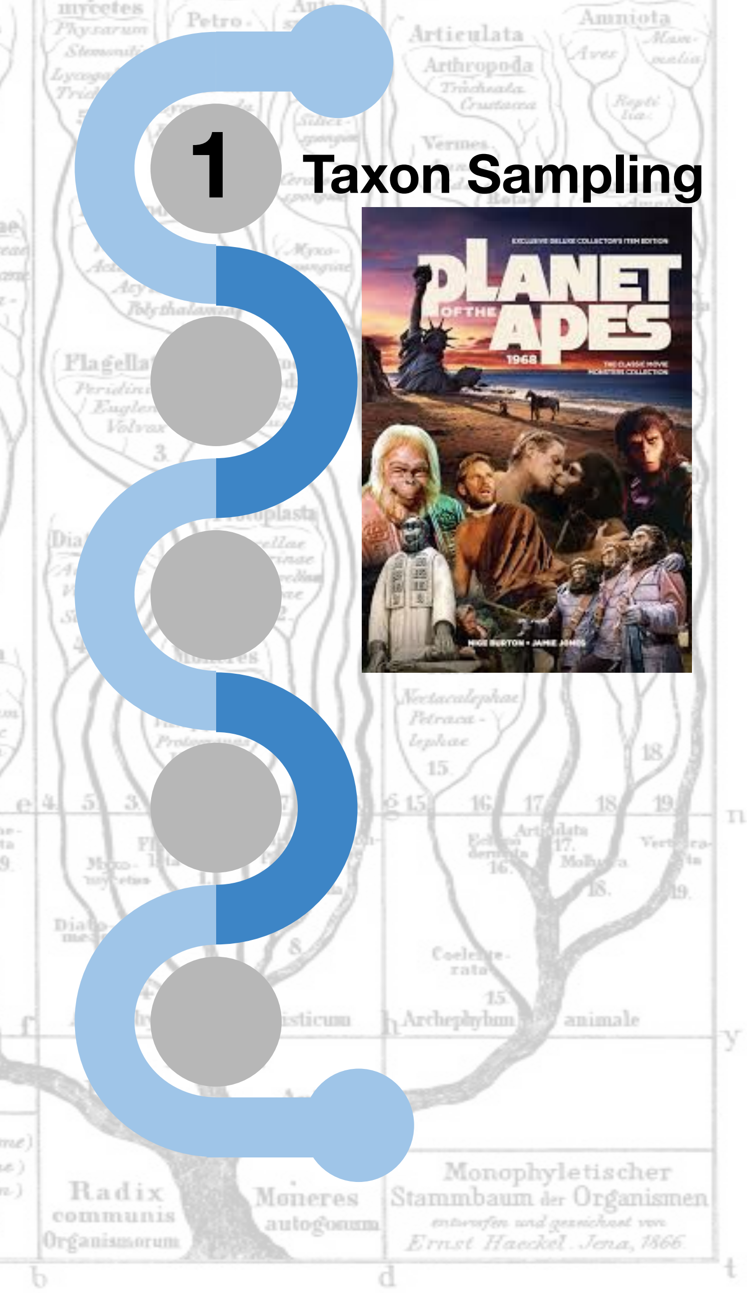
Key message: taxon sampling matters a lot

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



Rosenberg & Kumar 2003

How to infer a species tree

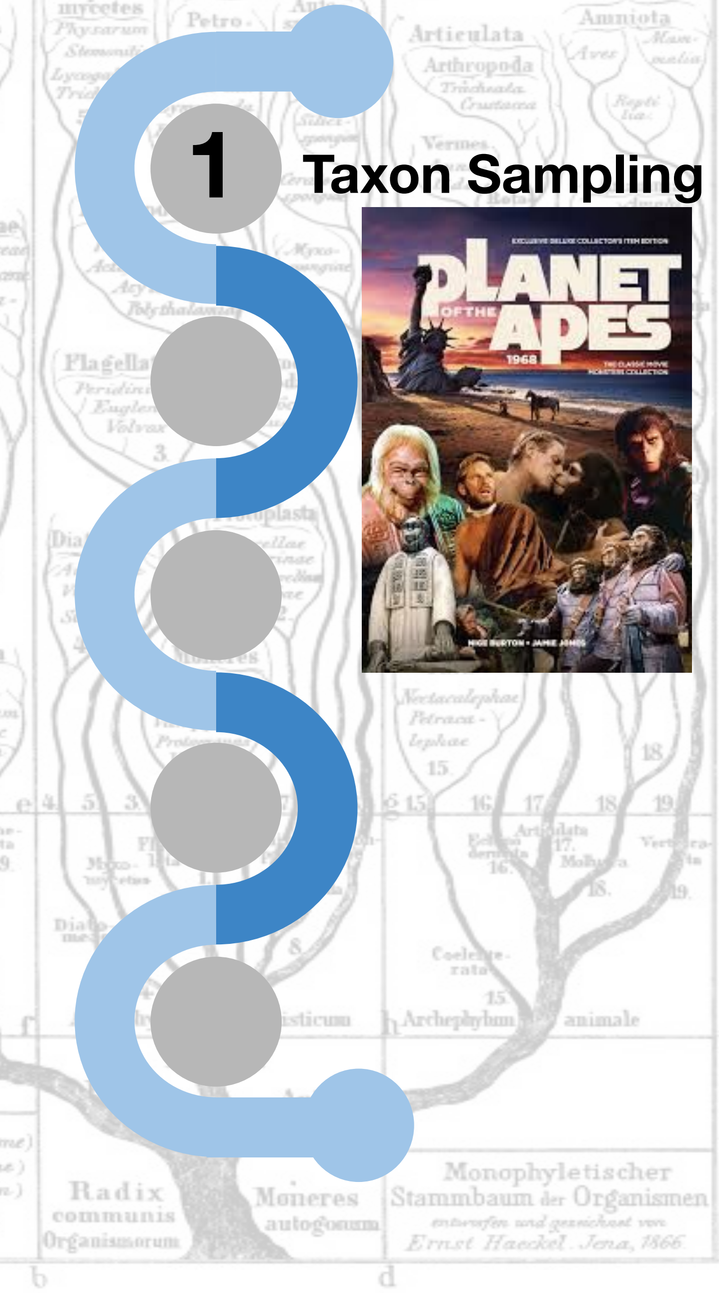


1 Taxon Sampling

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

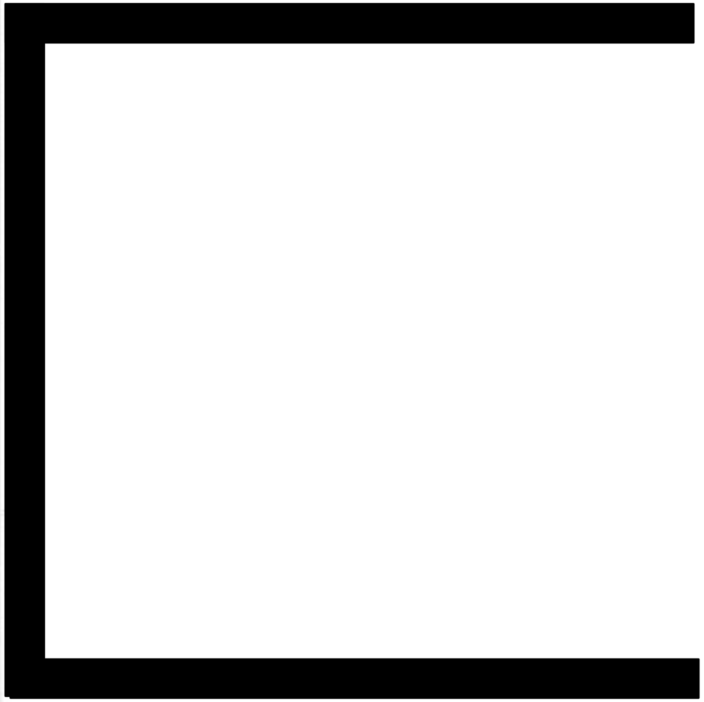


1 **Taxon Sampling**

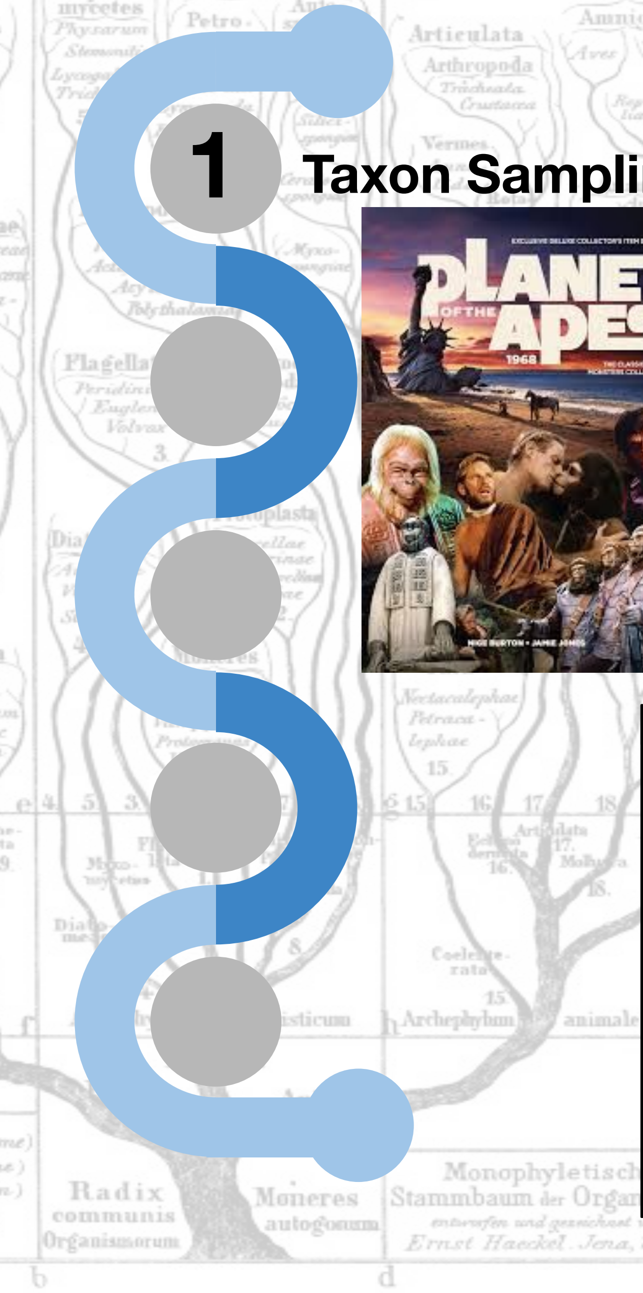
PLANET OF THE APES
1968
Articulata
Arthropoda
Amniota
Vermes
Flagellata
Archephylum
animale
Radix communis Organismorum
Moneres autogonum
Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866

Key message: taxon sampling matters <u>a lot</u>

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



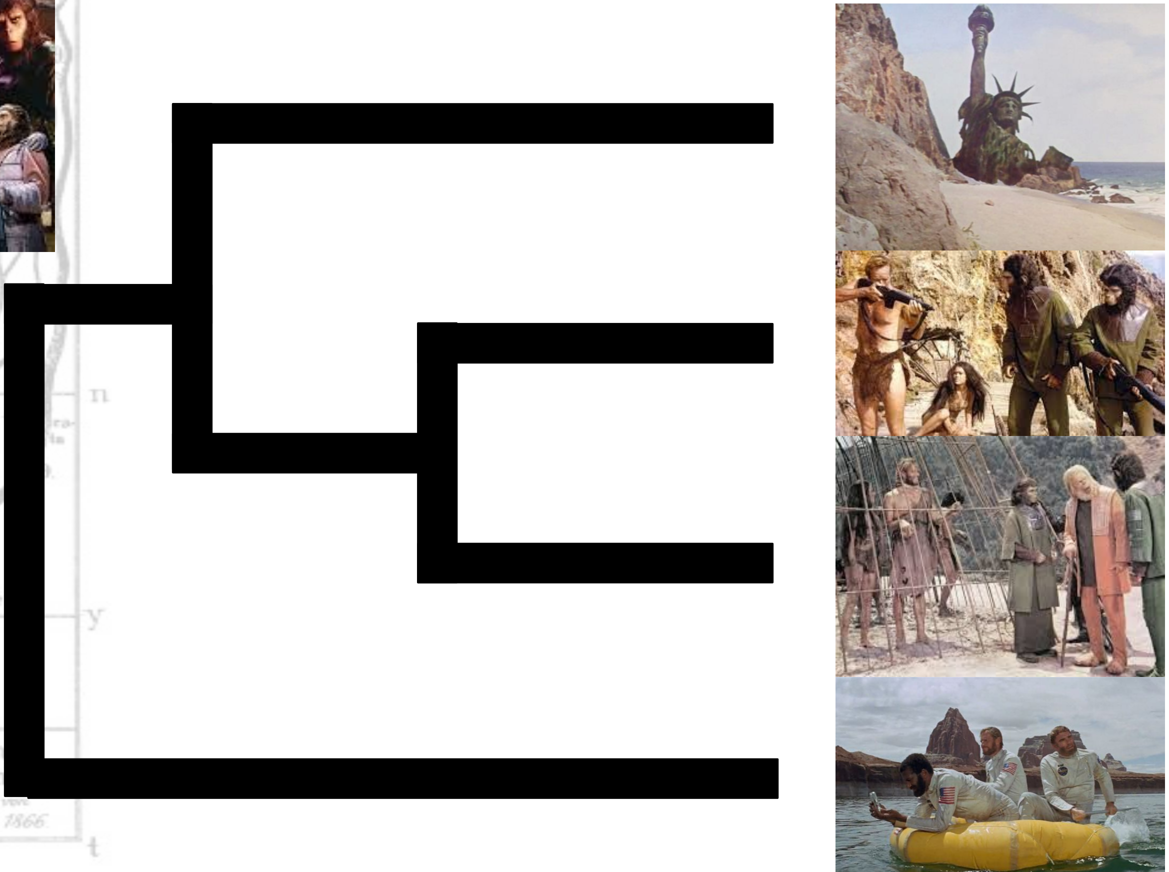
How to infer a species tree



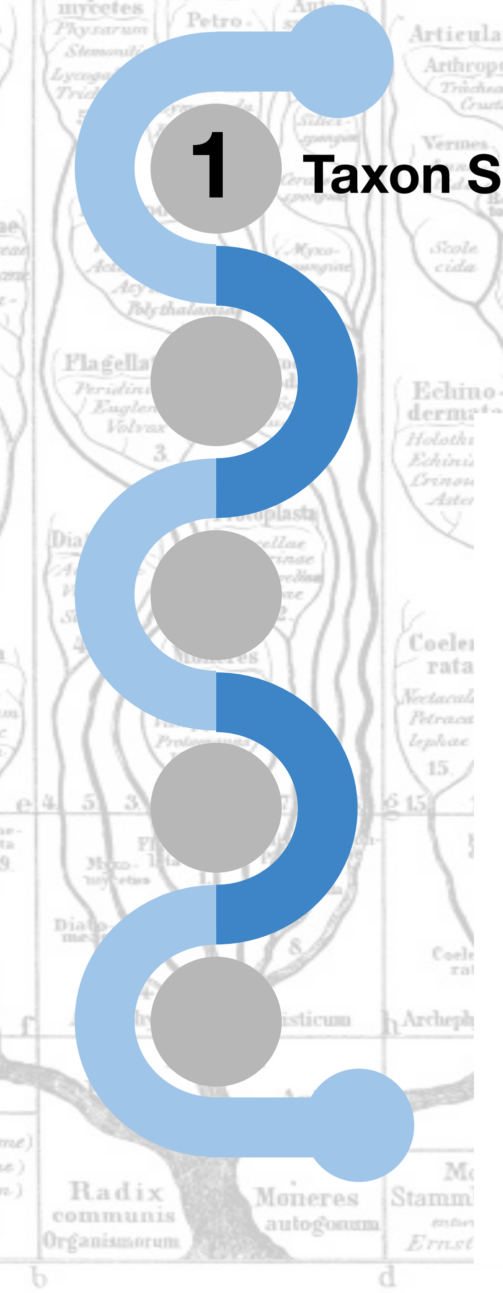
1 Taxon Sampling

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

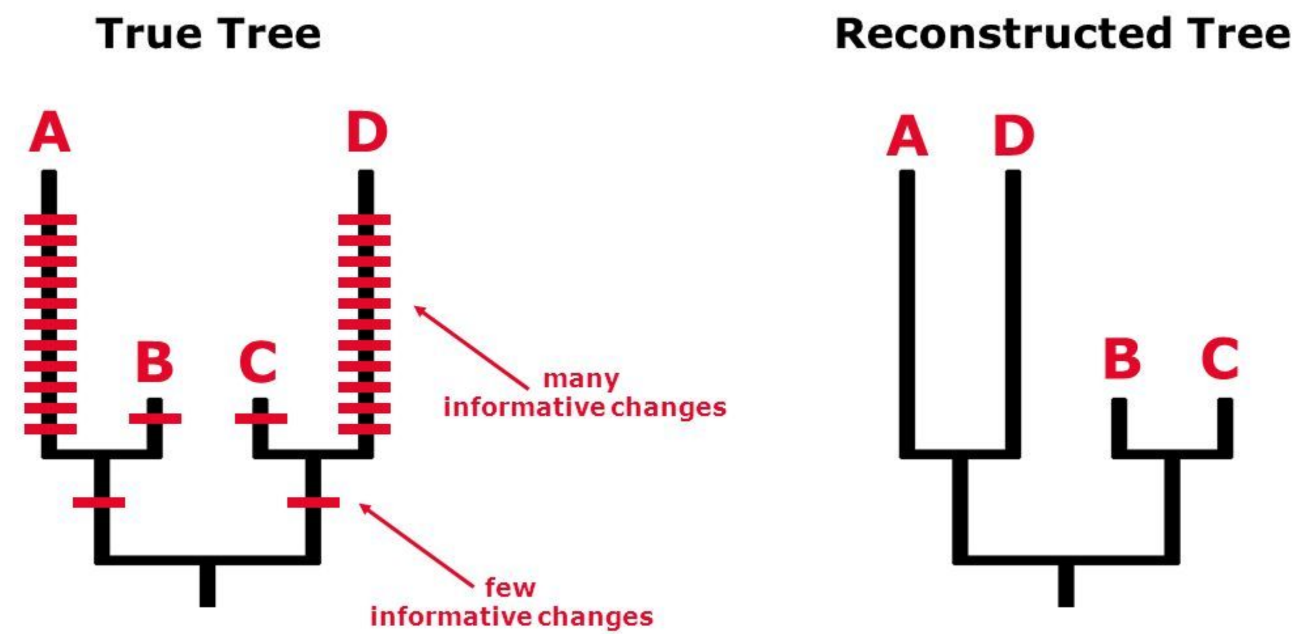


1 Taxon Sampling

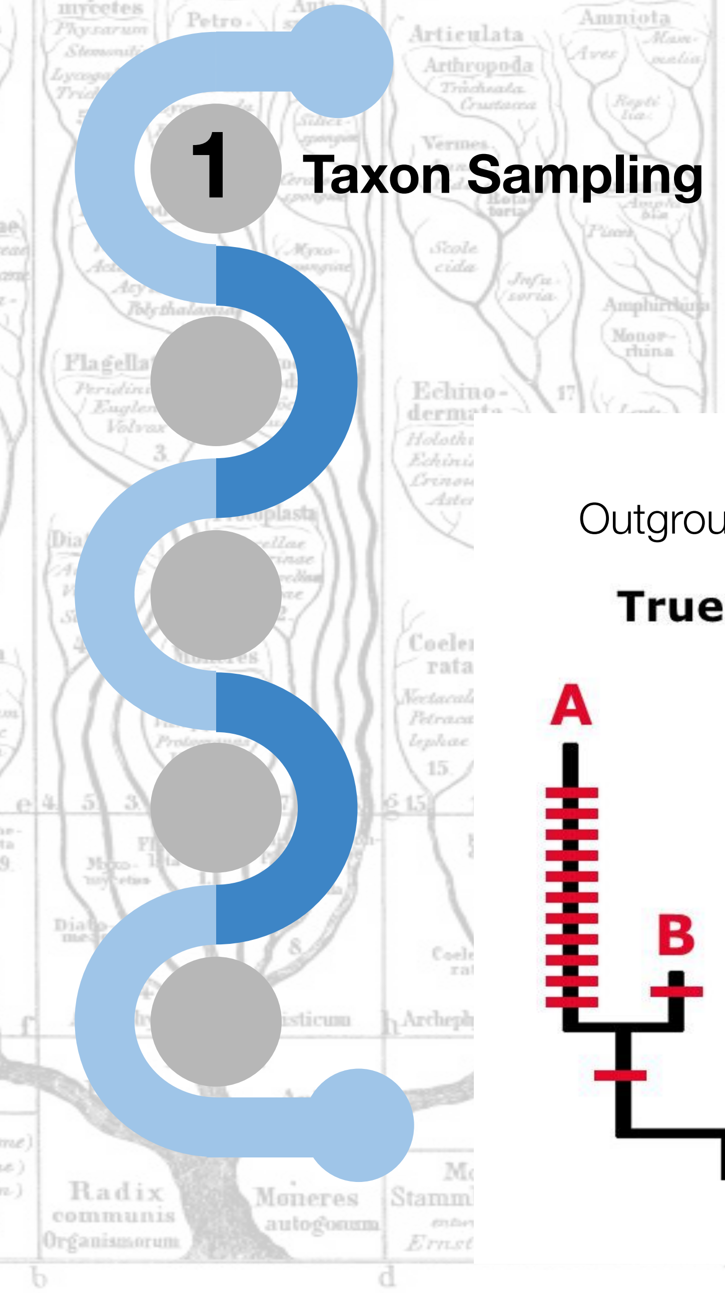
Key message: taxon sampling matters a lot

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

Long Branch Attraction



How to infer a species tree

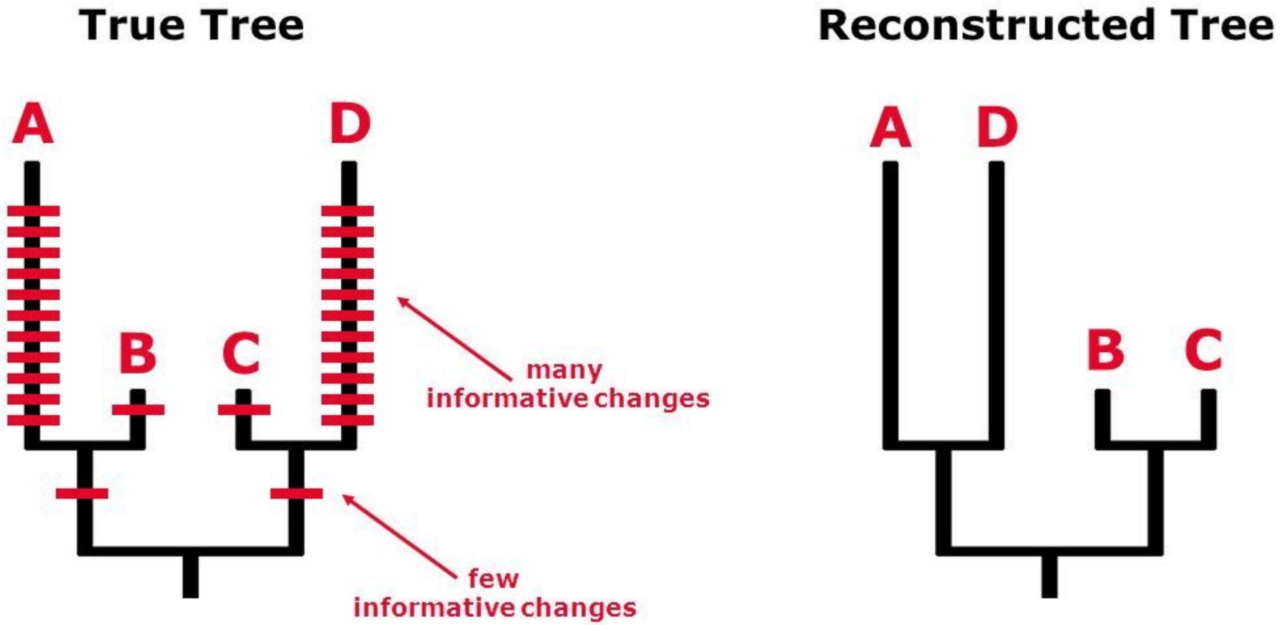


Key message: taxon sampling matters a lot

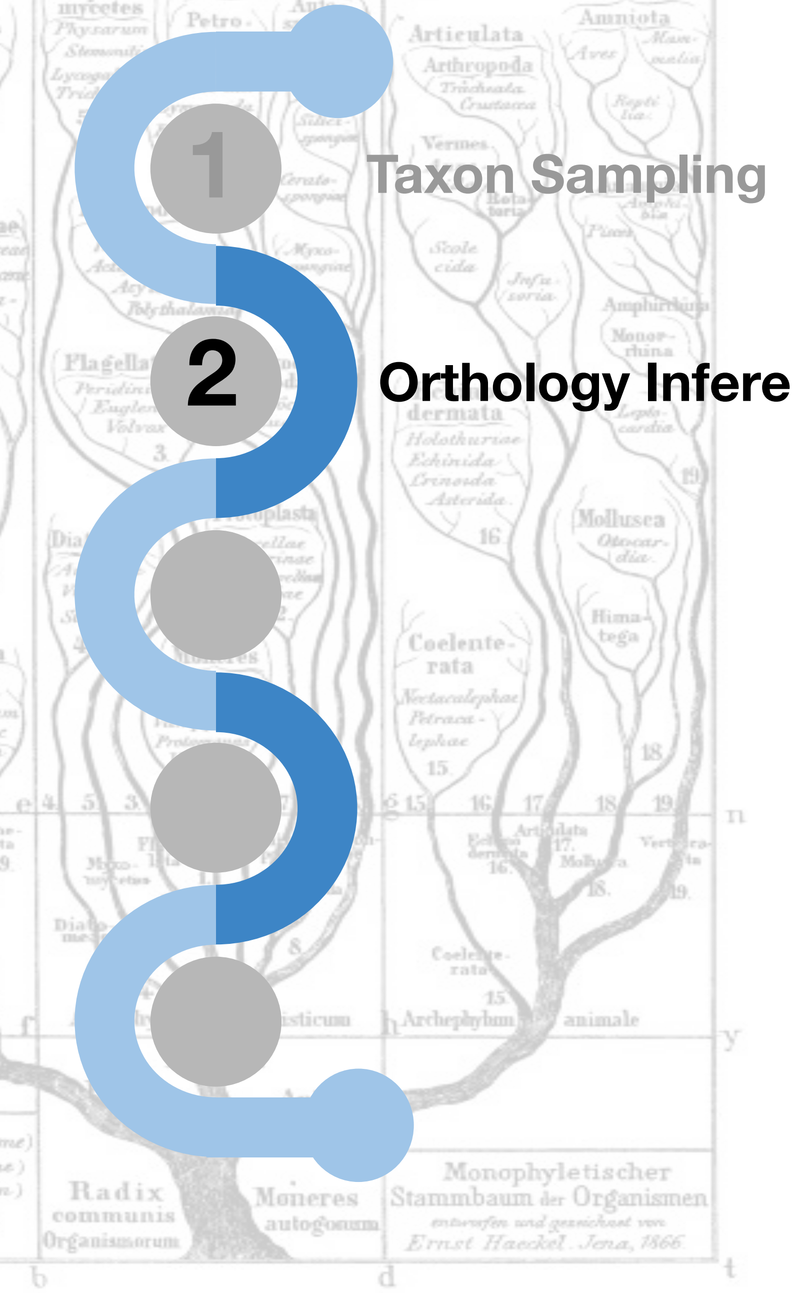
Incomplete, biased, or improper taxon sampling can lead to misleading results in reconstructing evolutionary relationships.

Long Branch Attraction

Outgroups / Fast-evolving lineages / Compositional heterogeneity

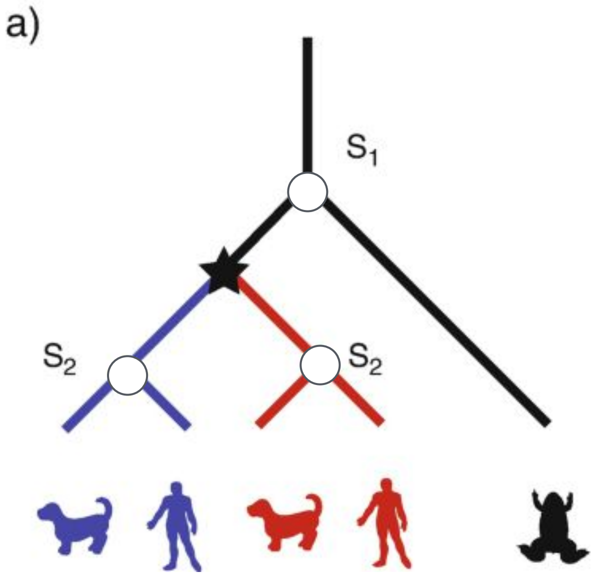


How to infer a species tree



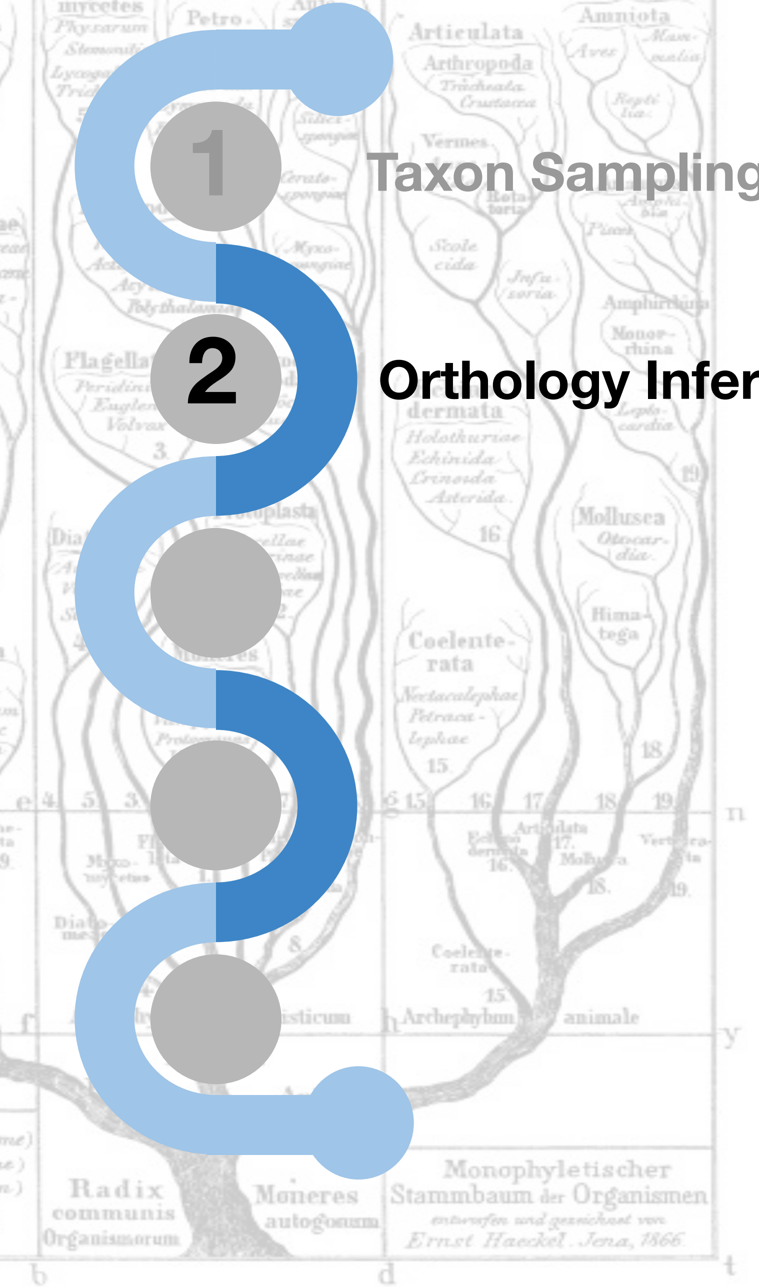
1 Taxon Sampling

2 **Orthology Inference**



Altenhoff, Glover & Dessimoz 2019

How to infer a species tree

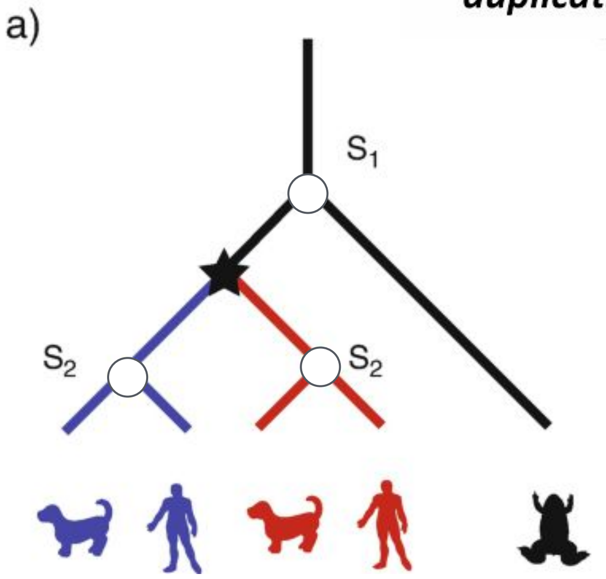

1
2
Monophyletischer Stammbaum der Organismen
Ernst Haeckel. Jena, 1866

1 Taxon Sampling

2 **Orthology Inference**

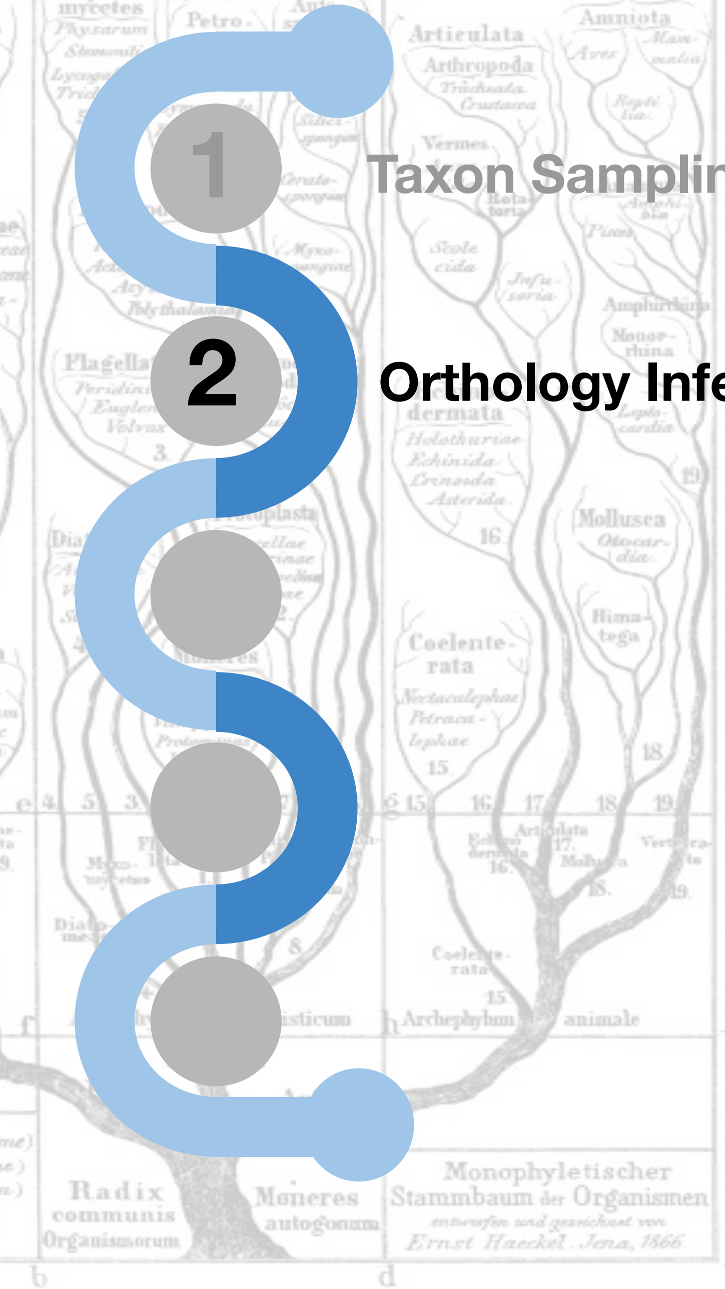
Definitions

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



Altenhoff, Glover & Dessimoz 2019

How to infer a species tree

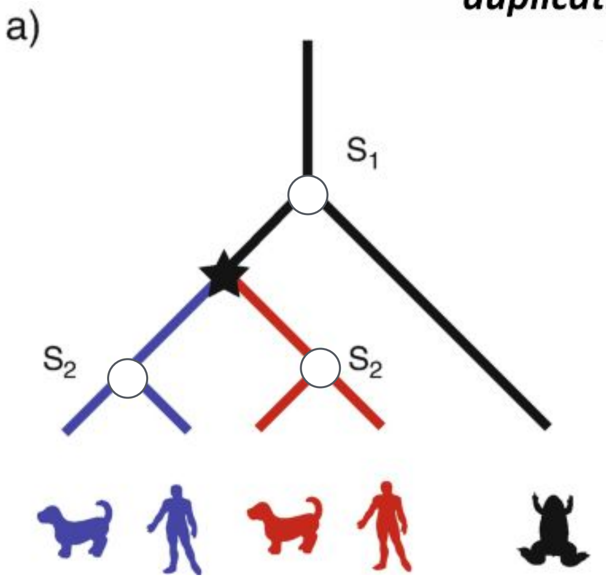

Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum

1 Taxon Sampling

2 **Orthology Inference**

Definitions

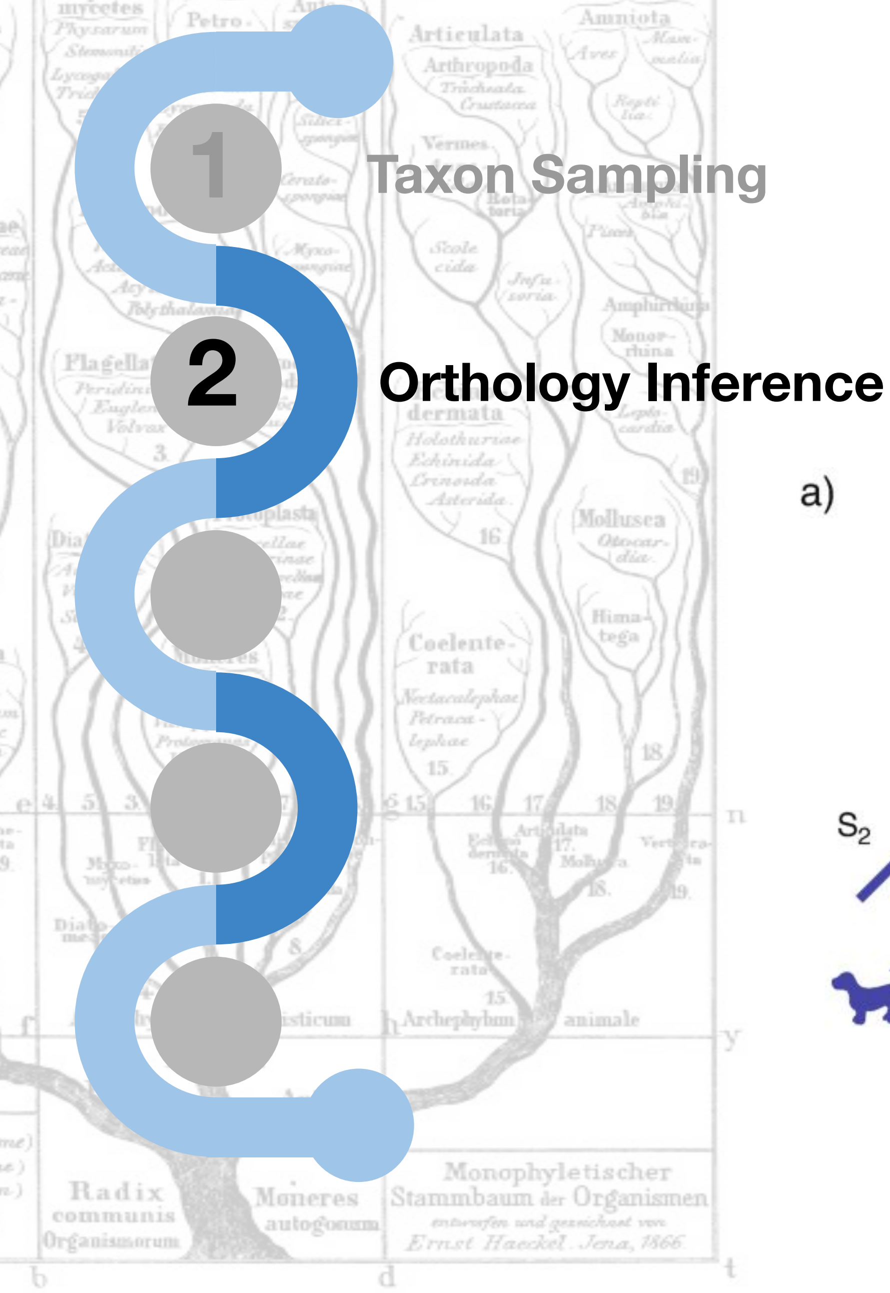
- Two genes are **orthologs** if their MRCA is a ***speciation***: ○
- Two genes are **paralogs** if their MRCA is a ***duplication***: ★



Orthology relationships are inferred pairwise

Altenhoff, Glover & Dessimoz 2019

How to infer a species tree



1 Taxon Sampling

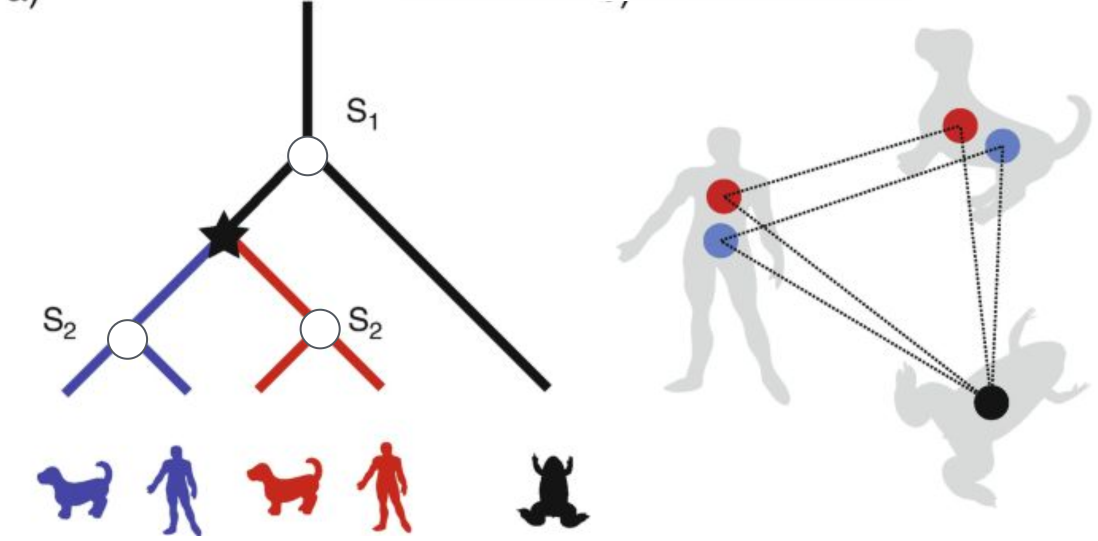
2 **Orthology Inference**

Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum

Definitions

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

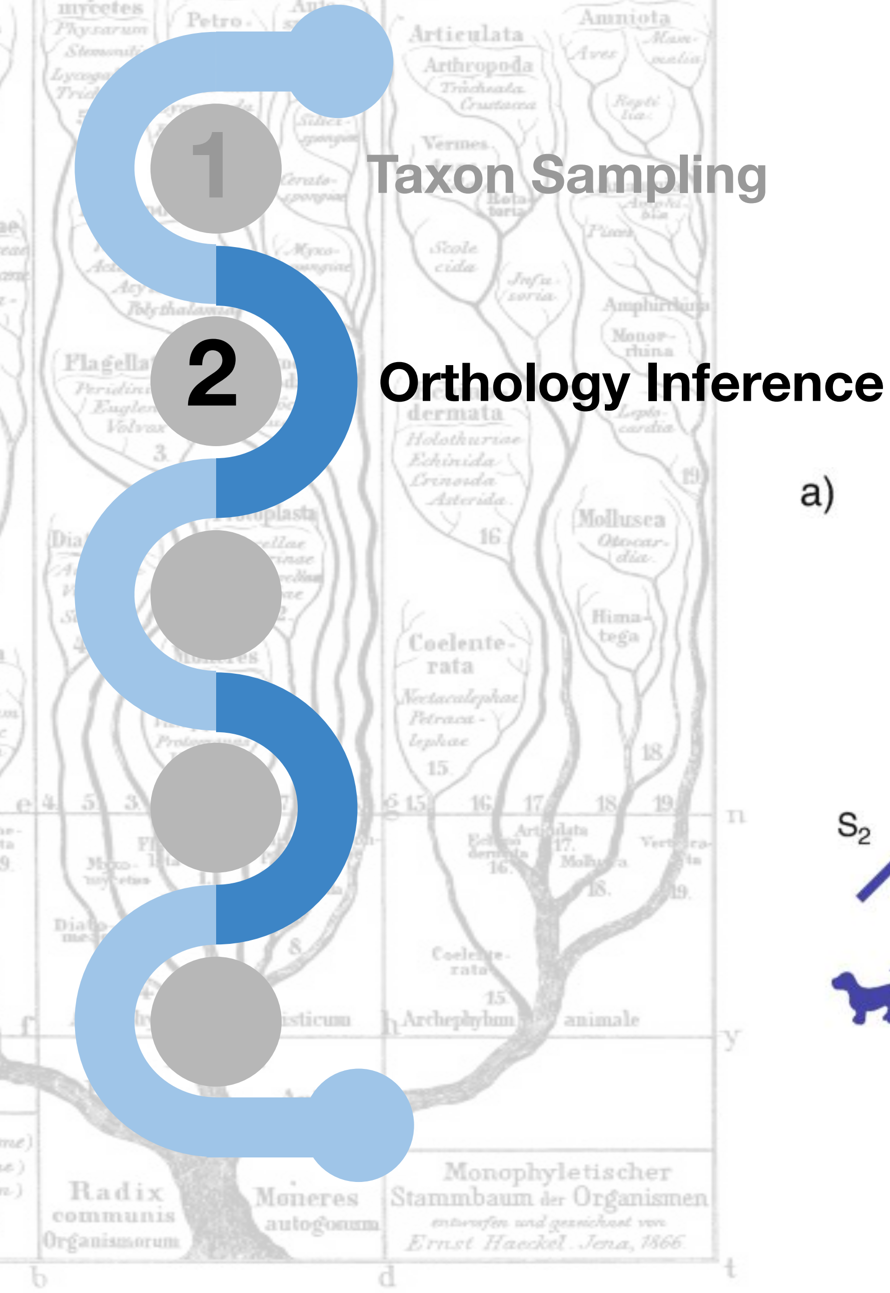
a)



Orthology relationships are inferred pairwise

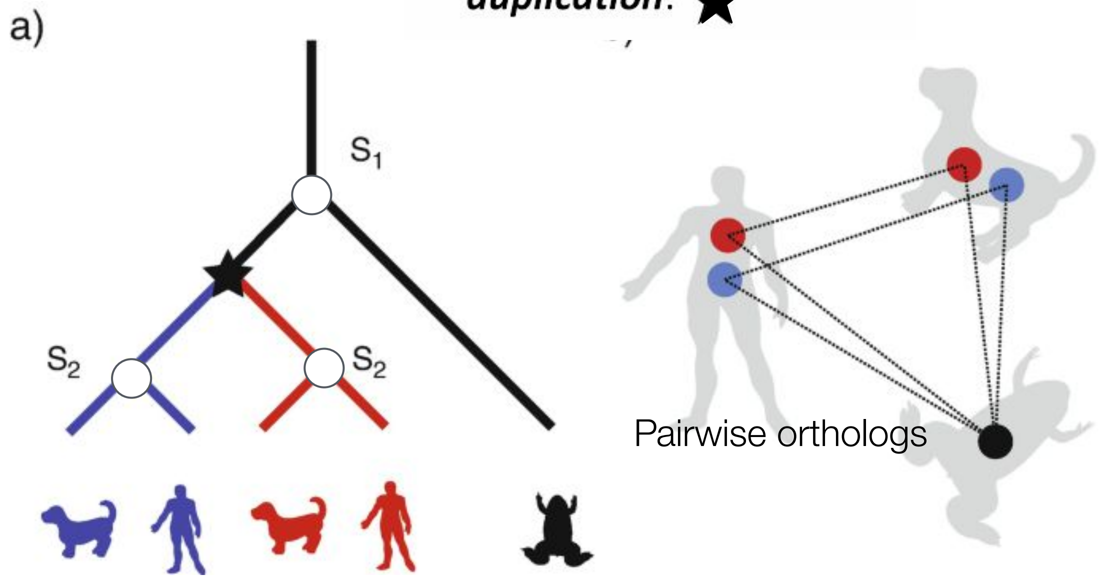
Altenhoff, Glover & Dessimoz 2019

How to infer a species tree


1
Taxon Sampling
2
Orthology Inference
Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866

Definitions

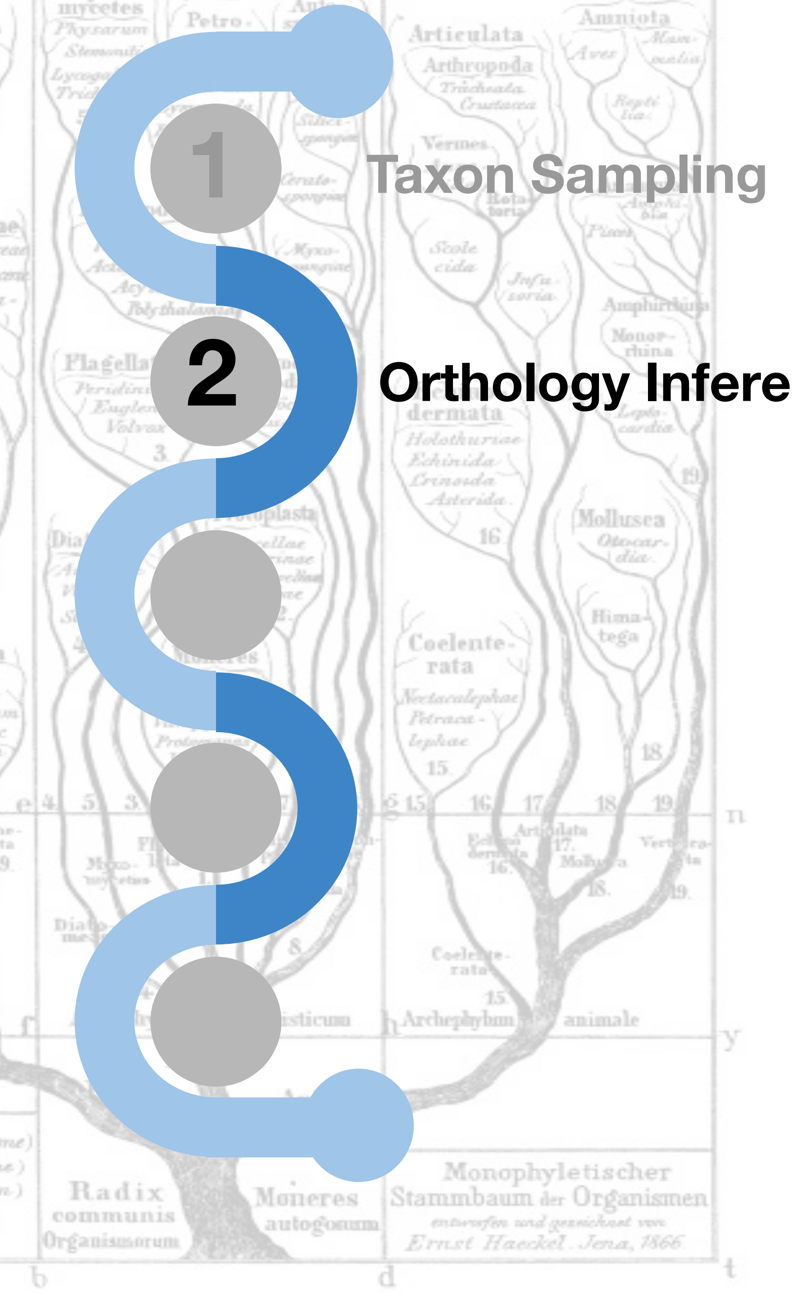
- Two genes are **orthologs** if their MRCA is a ***speciation***: ○
- Two genes are **paralogs** if their MRCA is a ***duplication***: ★



Orthology relationships are inferred pairwise

Altenhoff, Glover & Dessimoz 2019

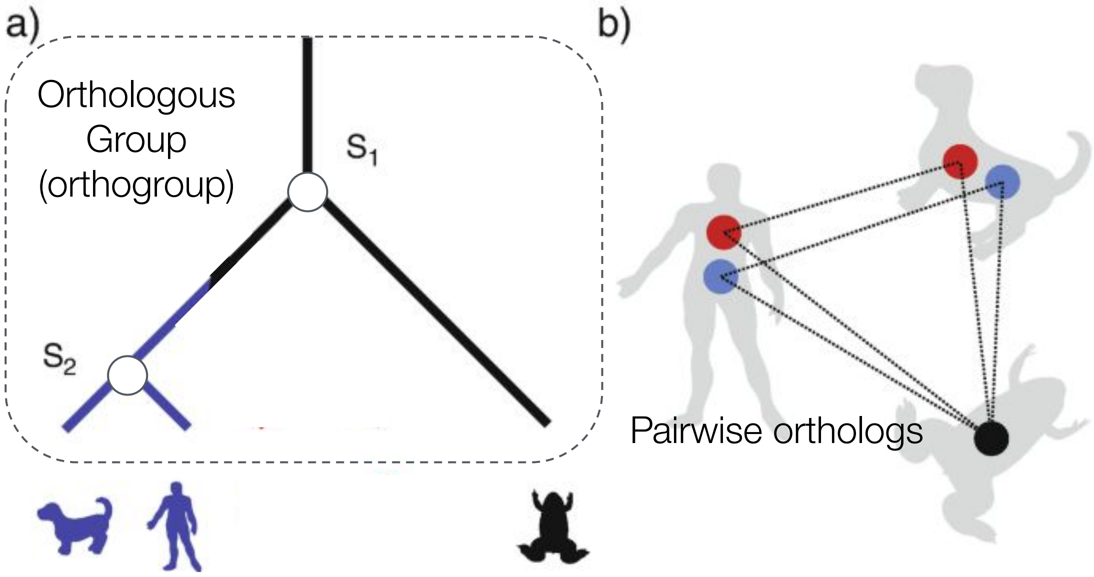
How to infer a species tree



Taxon Sampling

Orthology Inference

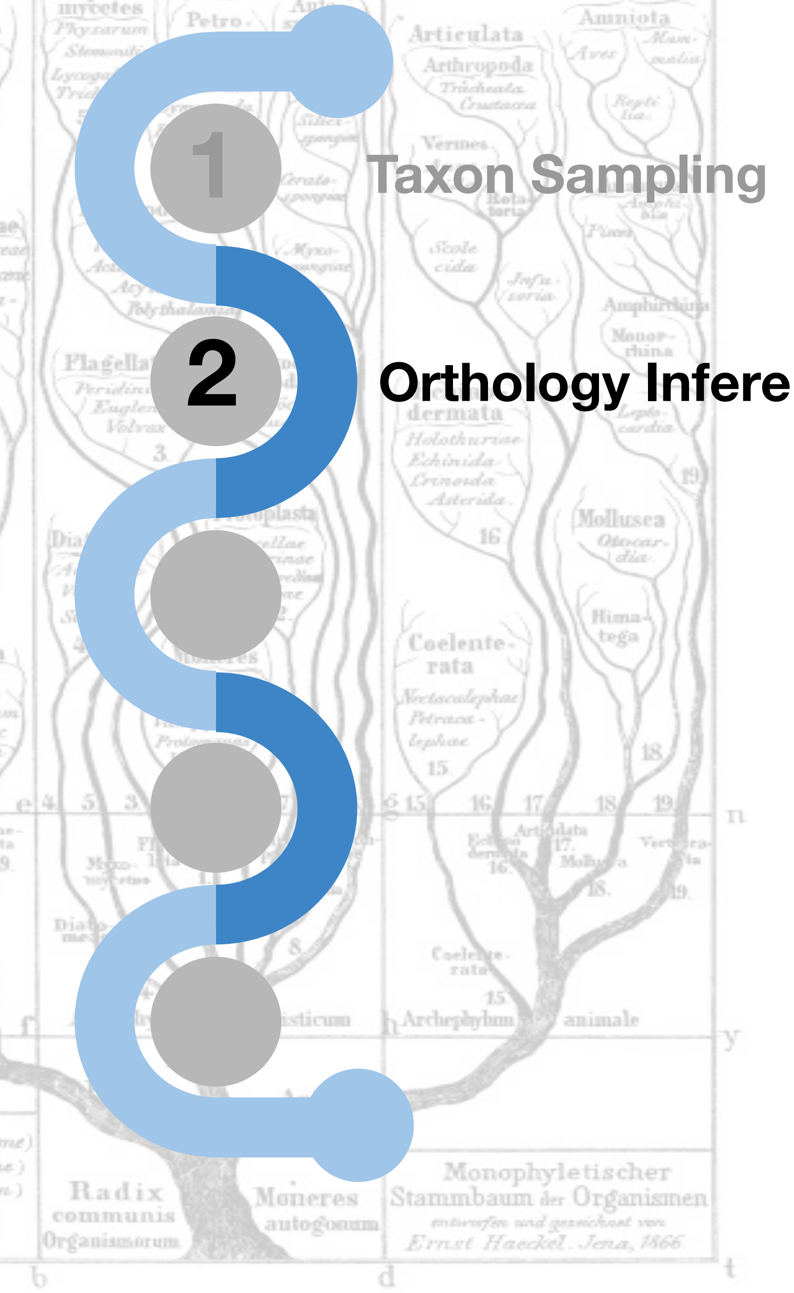
For phylogenetic inference, our starting point are the Orthologous Groups, (OGs) since they are derived from speciation events



Orthology relationships are inferred pairwise

Altenhoff, Glover & Dessimoz 2019

How to infer a species tree

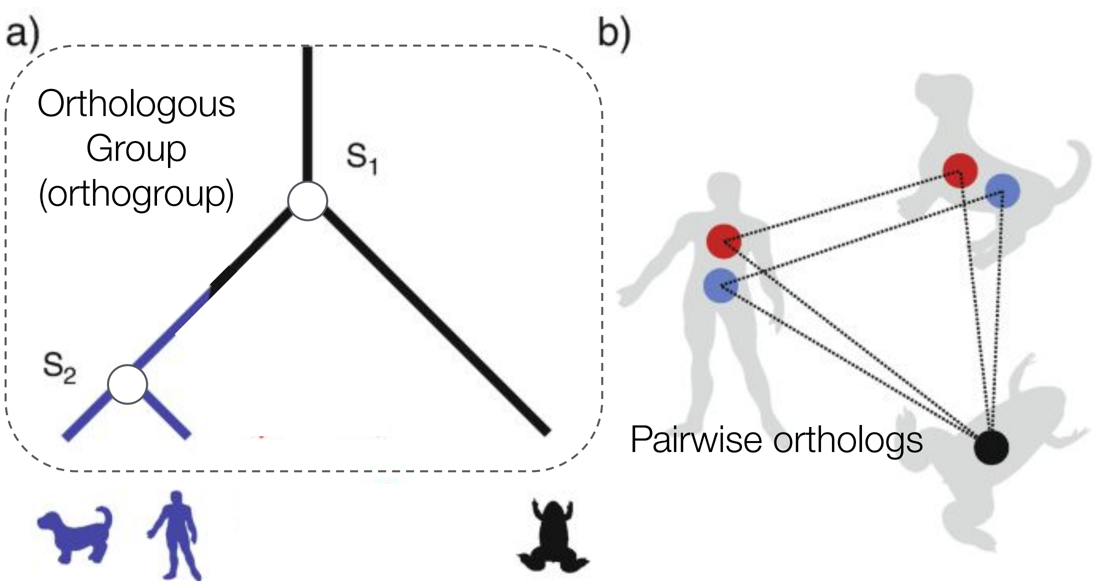


Taxon Sampling

Orthology Inference

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

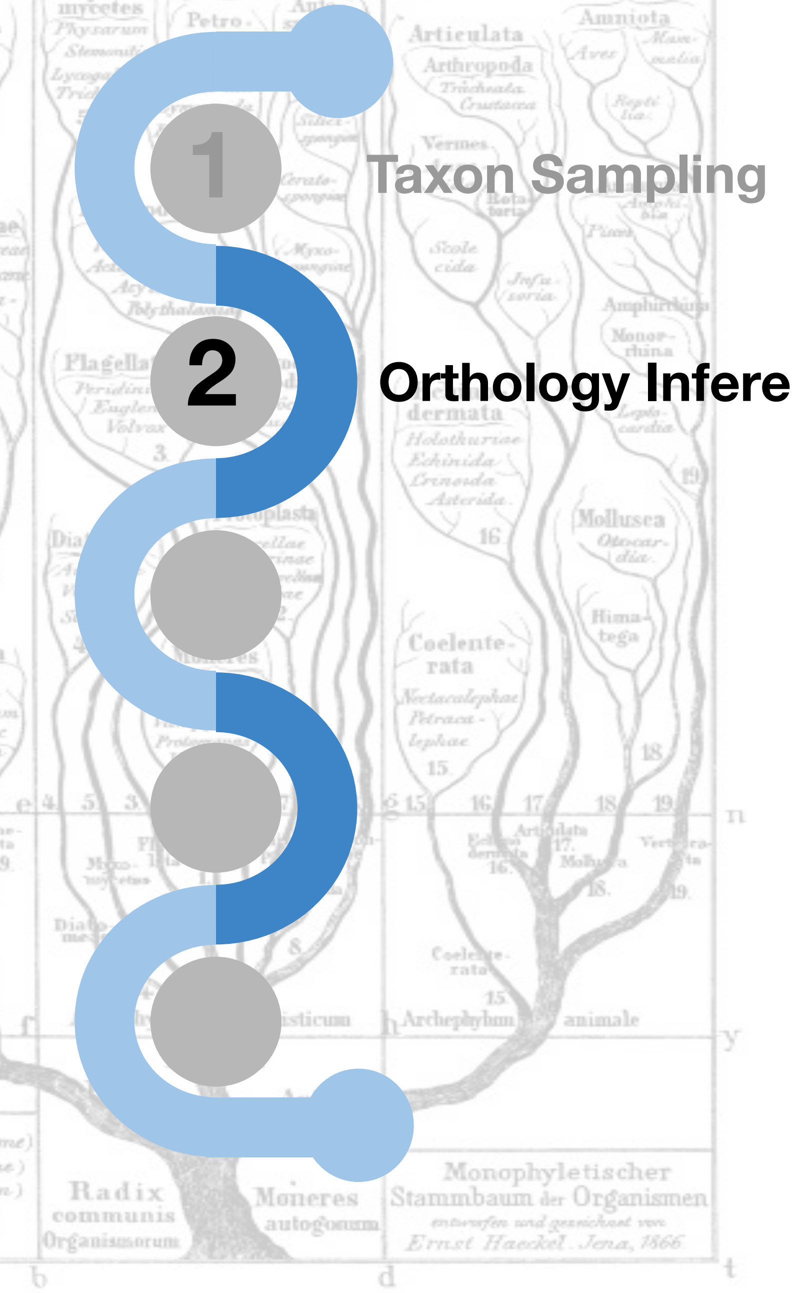
- **single copy OGs** (1:1 OGs: only one gene per species. Ready to continue with the next step!



Orthology relationships are inferred pairwise

Altenhoff, Glover & Dessimoz 2019

How to infer a species tree

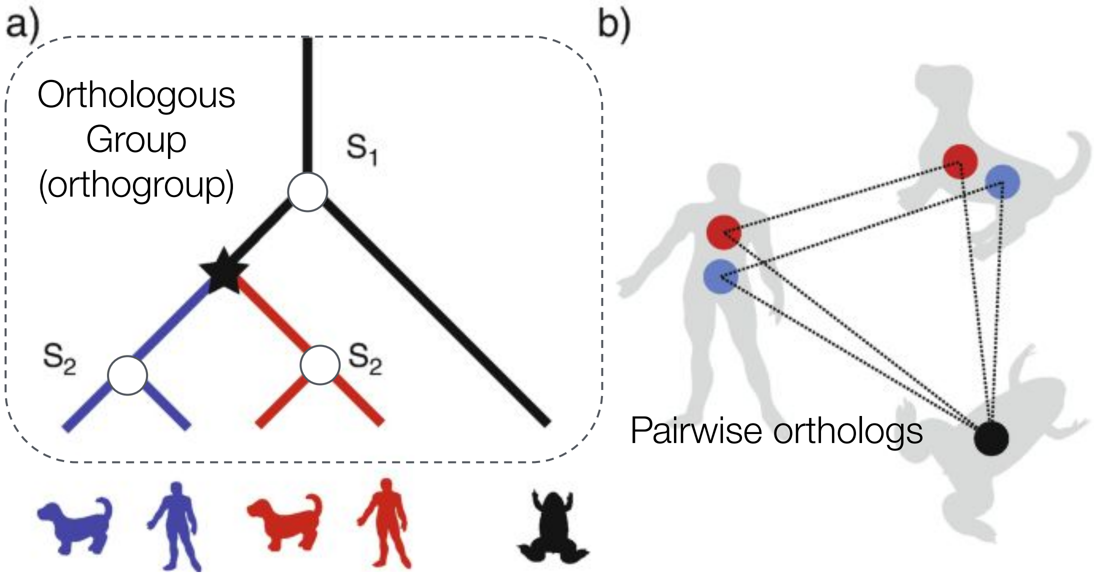

1
2
Monophyletischer Stammbaum der Organismen
Ernst Haeckel. Jena, 1866
Radix communis Organismorum

Taxon Sampling

Orthology Inference

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

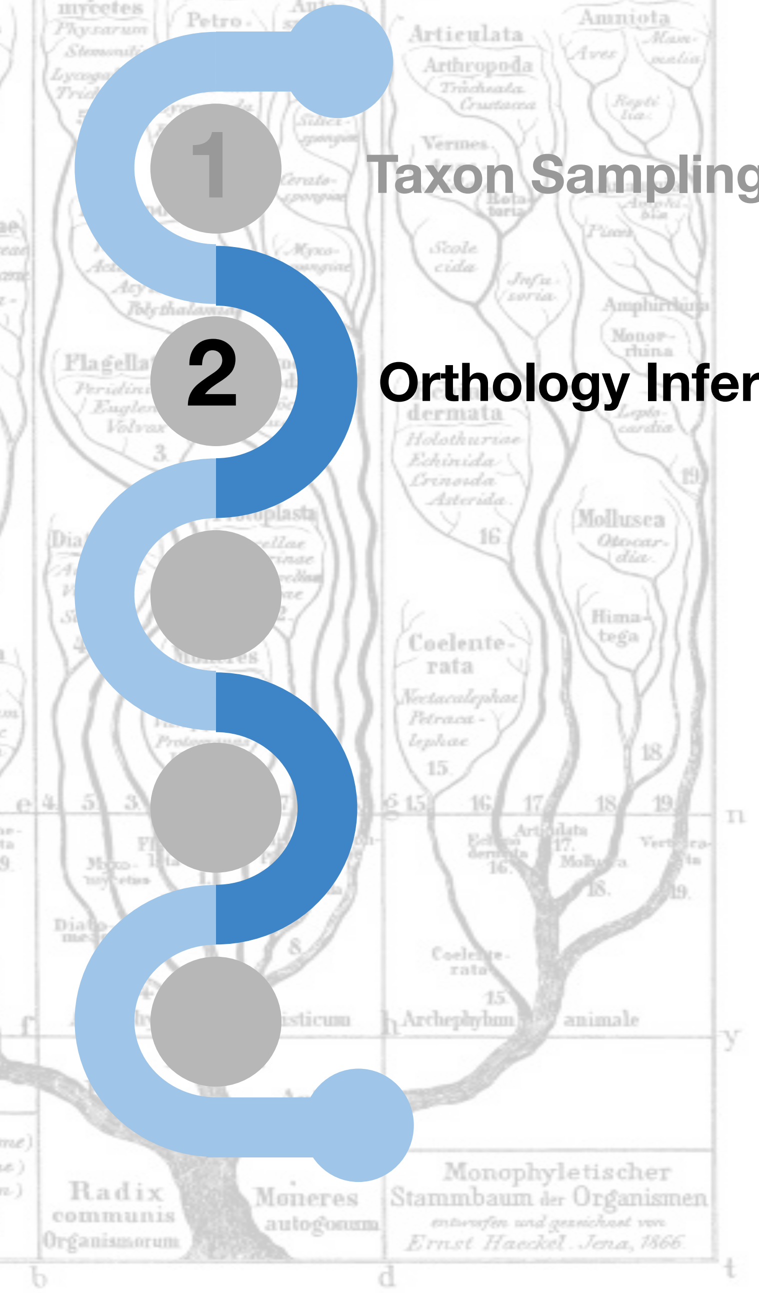
- single copy OGs (1:1 OGs: only one gene per species. Ready to continue with the next step!
- **OGs with duplicates** (1:many, many:many OGs): pruning is necessary to create your matrix



Orthology relationships are inferred pairwise

Altenhoff, Glover & Dessimoz 2019

How to infer a species tree

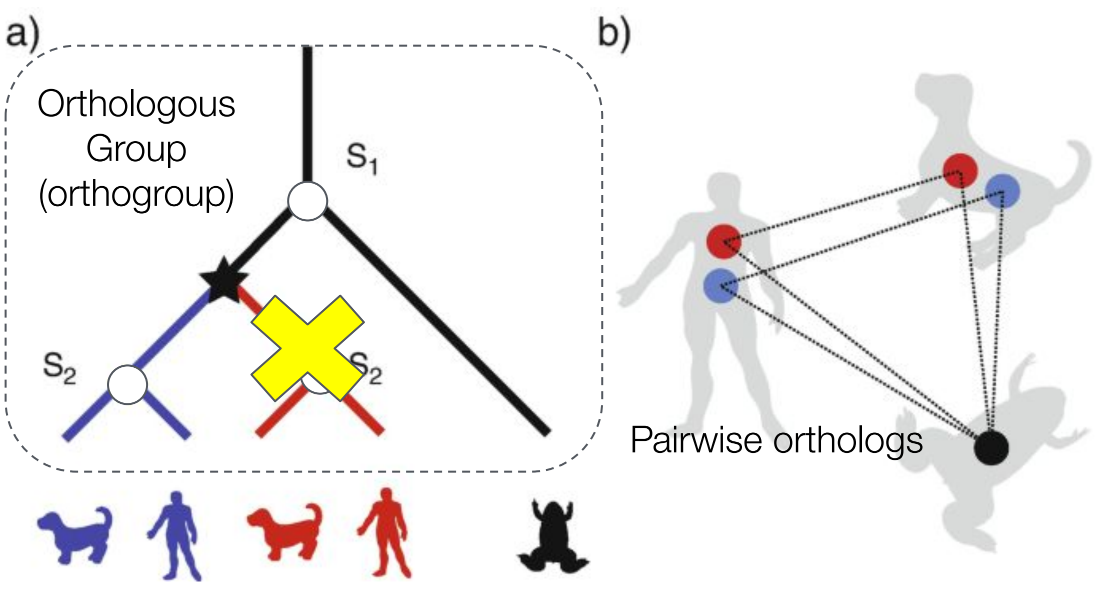


1 Taxon Sampling

2 **Orthology Inference**

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

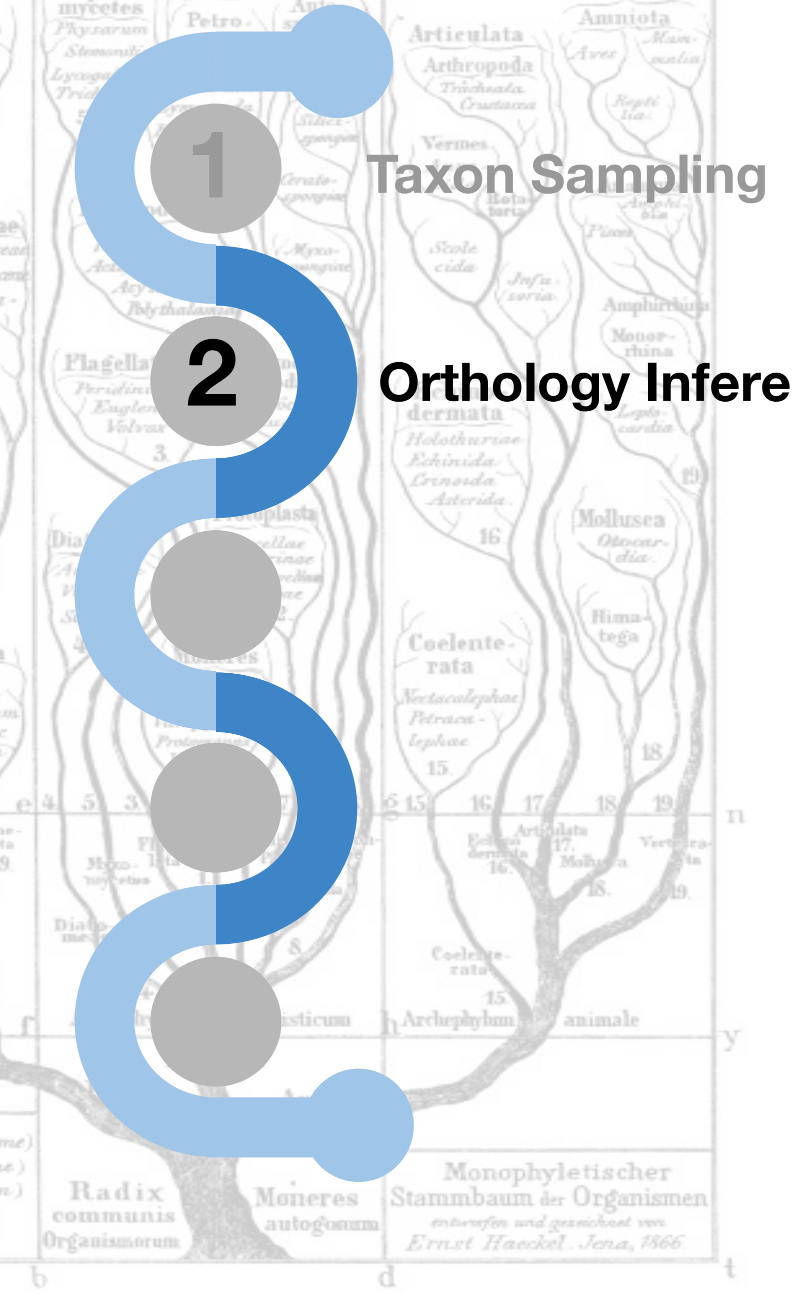
- single copy OGs (1:1 OGs: only one gene per species. Ready to continue with the next step!
- **OGs with duplicates** (1:many, many:many OGs): pruning is necessary to create your matrix



Orthology relationships are inferred pairwise

Altenhoff, Glover & Dessimoz 2019

How to infer a species tree

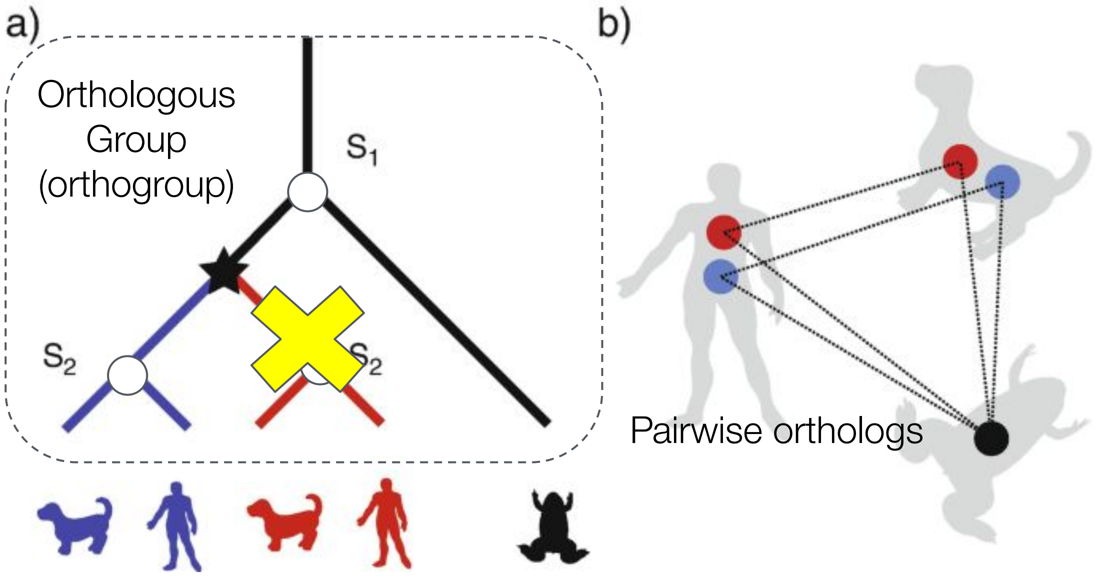


1 Taxon Sampling

2 **Orthology Inference**

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

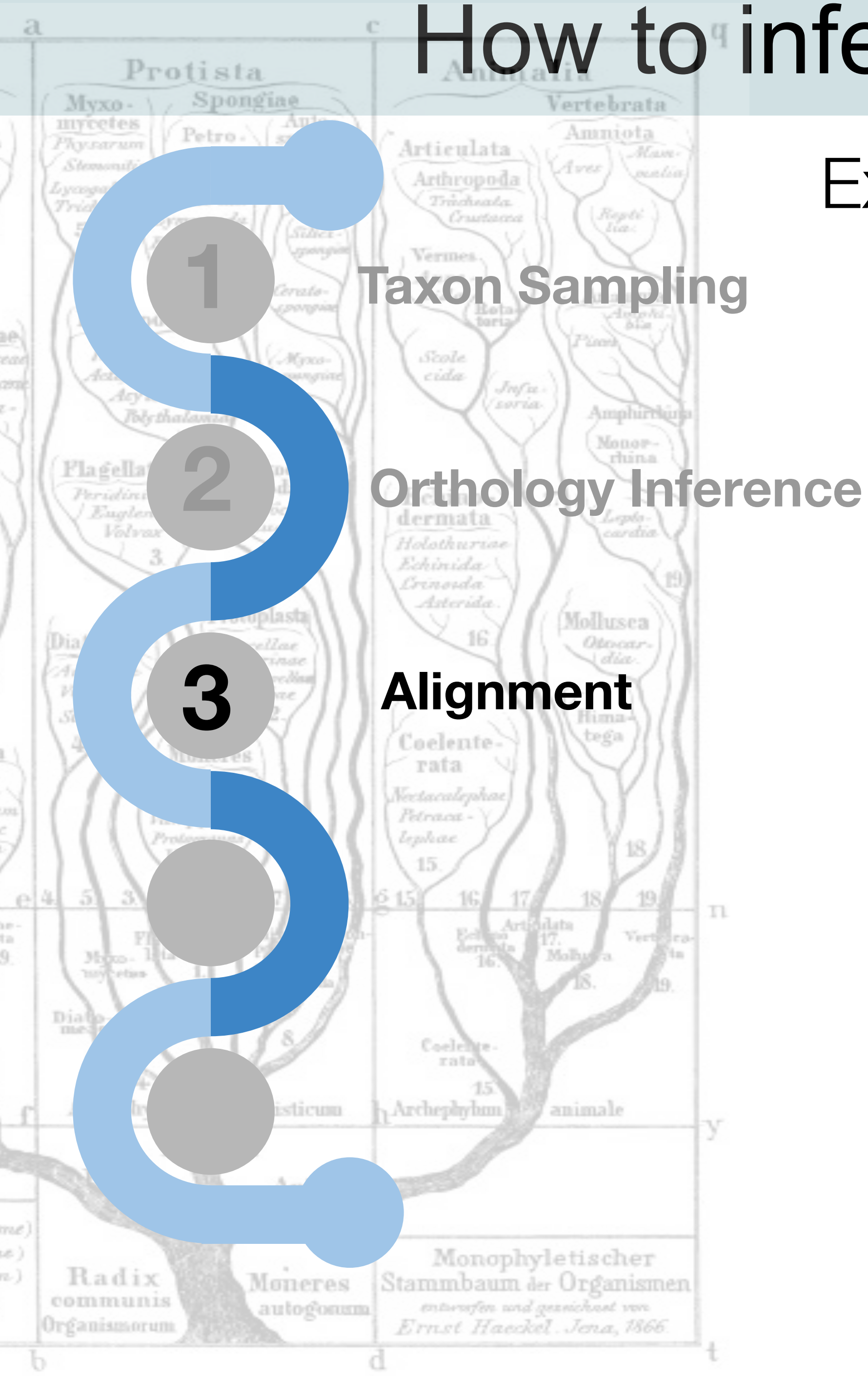
- single copy OGs (1:1 OGs: only one gene per species. Ready to continue with the next step!
- **OGs with duplicates** (1:many, many:many OGs): pruning is necessary to create your matrix



Key message: the selection of OGs for further analysis matters <u>a lot</u>

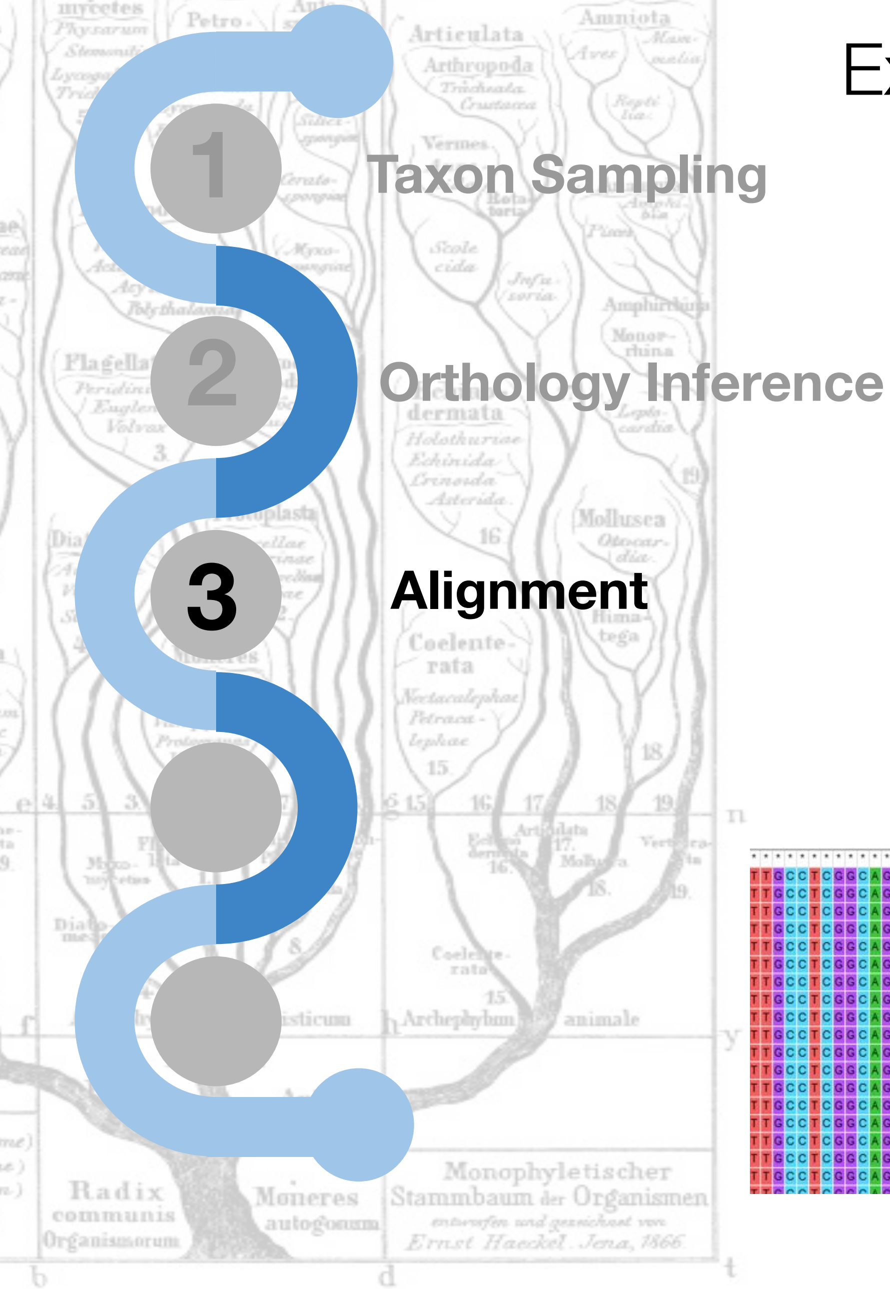
Altenhoff, Glover & Dessimoz 2019

How to infer a species tree



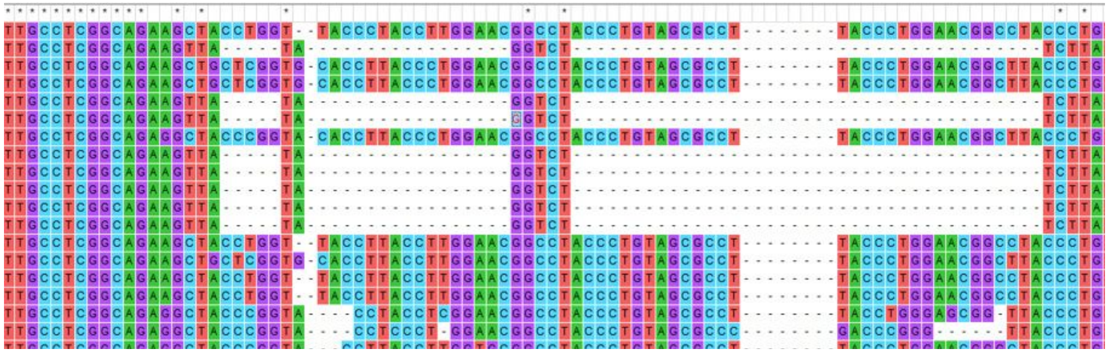
Extensively covered last week
by Mike, Rob & Francesco

How to infer a species tree


1
Taxon Sampling
2
Orthology Inference
3
Alignment
Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866

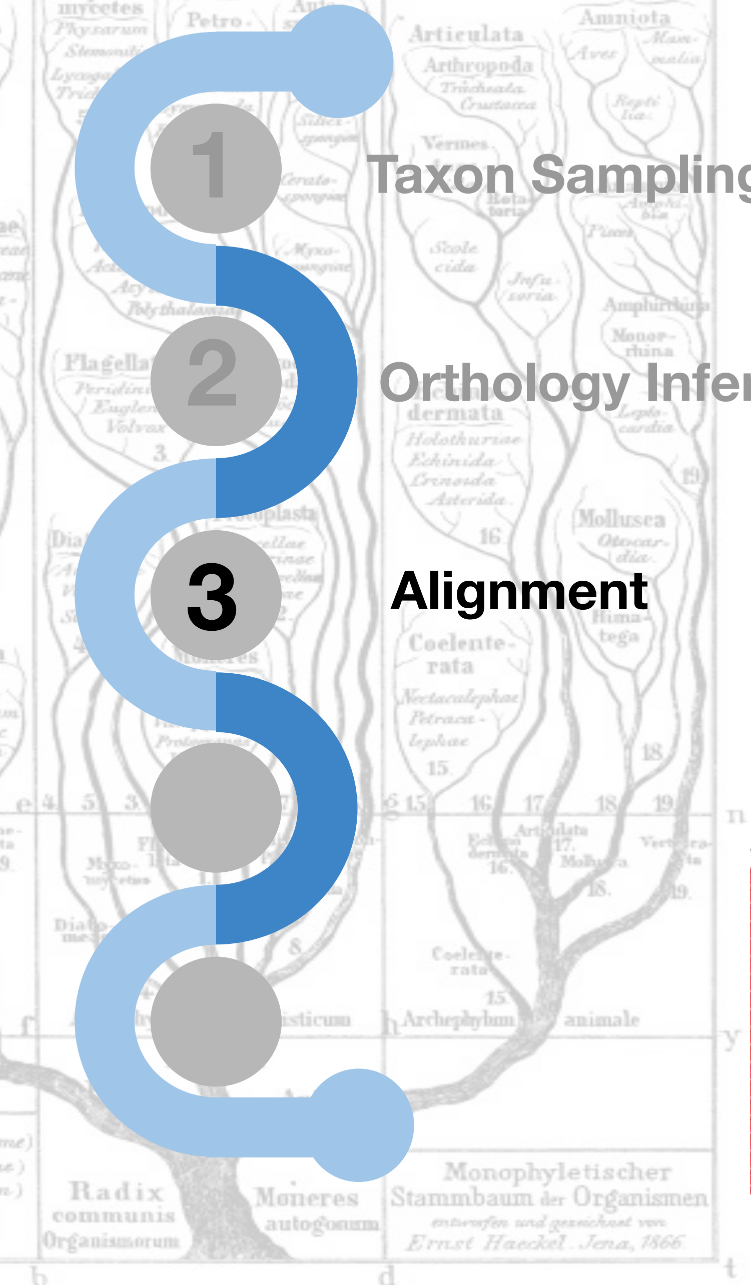
Extensively covered last week by Mike, Rob & Francesco

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



```
************  * *      *                   *  *                                        * * 
TTGCCTCGGCAGAAGCTACCTGGT--TACCCTACCTTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCCTACCCTG
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAAGCTGCTCGGTG-CACCTTACCCTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCTTACCCTG
TTGCCTCGGCAGAAGCTGCTCGGTG-CACCTTACCCTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCTTACCCTG
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAGGCTACCCGGTA-CACCTTACCCTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCTTACCCTG
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAAGCTACCTGGT--TACCTTACCTTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCCTACCCTG
TTGCCTCGGCAGAAGCTGCTCGGTG-CACCTTACCTTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCTTACCCTG
TTGCCTCGGCAGAAGCTACCTGGT--TACCTTACCTTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCCTACCCTG
TTGCCTCGGCAGAAGCTACCTGGT--TACCTTACCTTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCCTACCCTG
TTGCCTCGGCAGAGGCTACCCGGTA----CCTACCTCGGAACGGCCTACCCTGTAGCGCCT--------TACCTGGGAGCGG-TTACCCTG
TTGCCTCGGCAGAGGCTACCCGGTA----CCTCCCT-GGAACGGCCTACCCTGTAGCGCCC--------GACCCGGG------TTACCCTG
```

How to infer a species tree



1 Taxon Sampling

2 Orthology Inference

3 Alignment

Extensively covered last week by Mike, Rob & Francesco

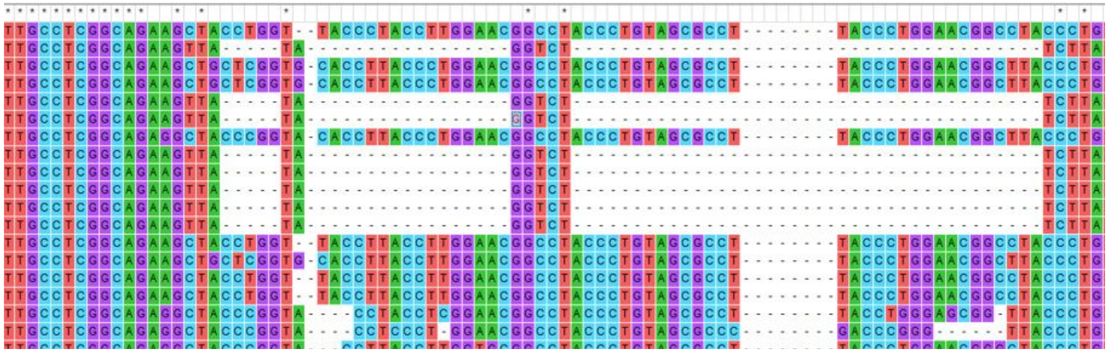
If the sequences are poorly aligned, you may want to consider trimming the poorly aligned areas. There are several tools for it:

trimAl

A tool for automated alignment trimming

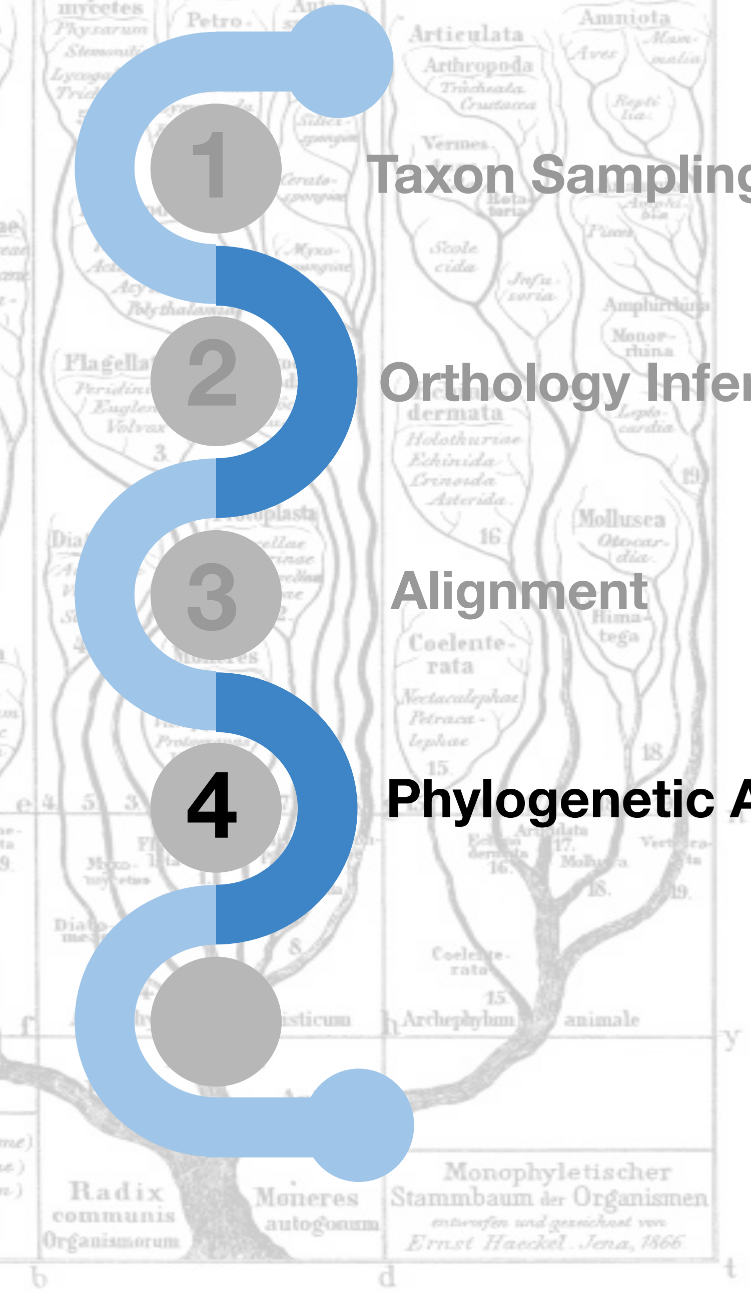
PREQUAL

(prealignment quality filter)



```
************  * *      *                   *  *                                        * * 
TTGCCTCGGCAGAAGCTACCTGGT--TACCCTACCTTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCCTACCCTG
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAAGCTGCTCGGTG-CACCTTACCCTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCTTACCCTG
TTGCCTCGGCAGAAGCTGCTCGGTG-CACCTTACCCTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCTTACCCTG
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAGGCTACCCGGTA-CACCTTACCCTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCTTACCCTG
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAAGTTA-----TA-----------------GGTCT---------------------------------------TCTTA
TTGCCTCGGCAGAAGCTACCTGGT--TACCTTACCTTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCCTACCCTG
TTGCCTCGGCAGAAGCTGCTCGGTG-CACCTTACCTTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCTTACCCTG
TTGCCTCGGCAGAAGCTACCTGGT--TACCTTACCTTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCCTACCCTG
TTGCCTCGGCAGAAGCTACCTGGT--TACCTTACCTTGGAACGGCCTACCCTGTAGCGCCT--------TACCCTGGAACGGCCTACCCTG
TTGCCTCGGCAGAGGCTACCCGGTA----CCTACCTCGGAACGGCCTACCCTGTAGCGCCT--------TACCTGGGAGCGG-TTACCCTG
TTGCCTCGGCAGAGGCTACCCGGTA----CCTCCCT-GGAACGGCCTACCCTGTAGCGCCC--------GACCCGGG------TTACCCTG
```

How to infer a species tree


Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum
Archephylum animale
Articulata
Arthropoda
Amniota
Mollusca
Coelenterata
1
2
3
4

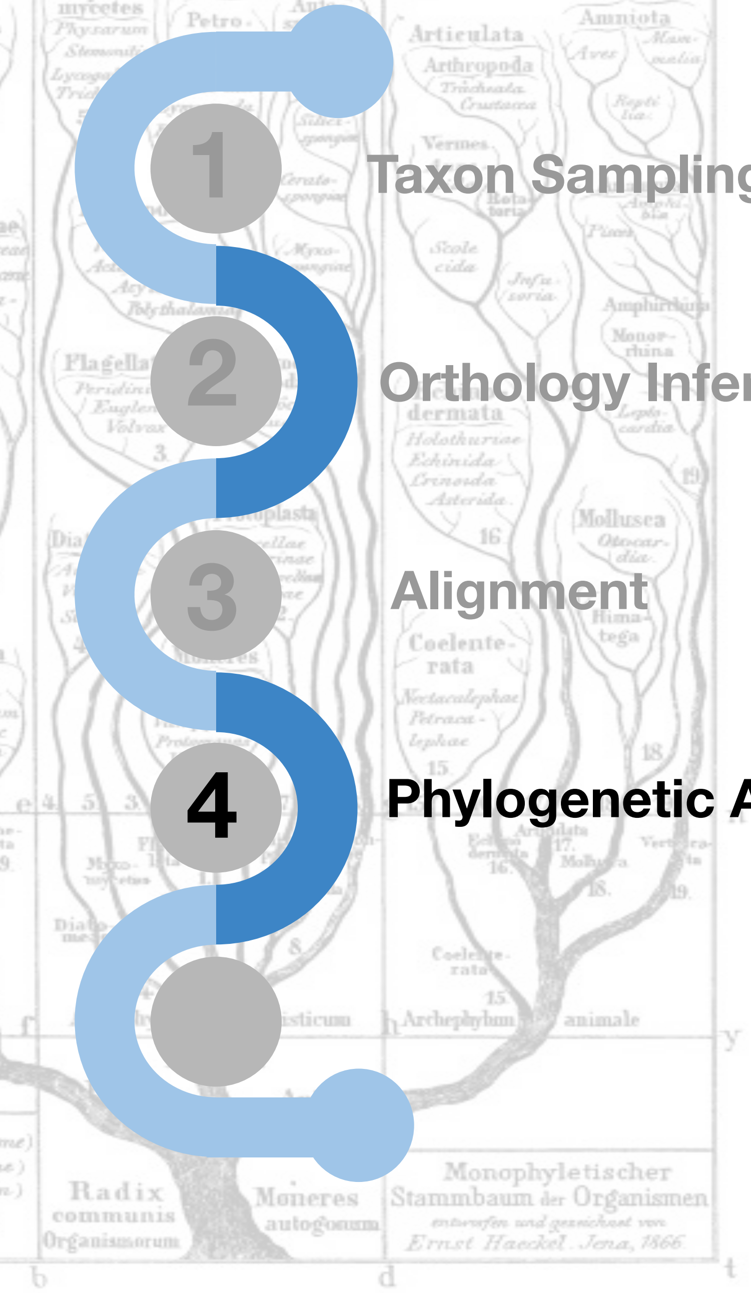
1 Taxon Sampling

2 Orthology Inference

3 Alignment

4 Phylogenetic Analysis

How to infer a species tree



Taxon Sampling

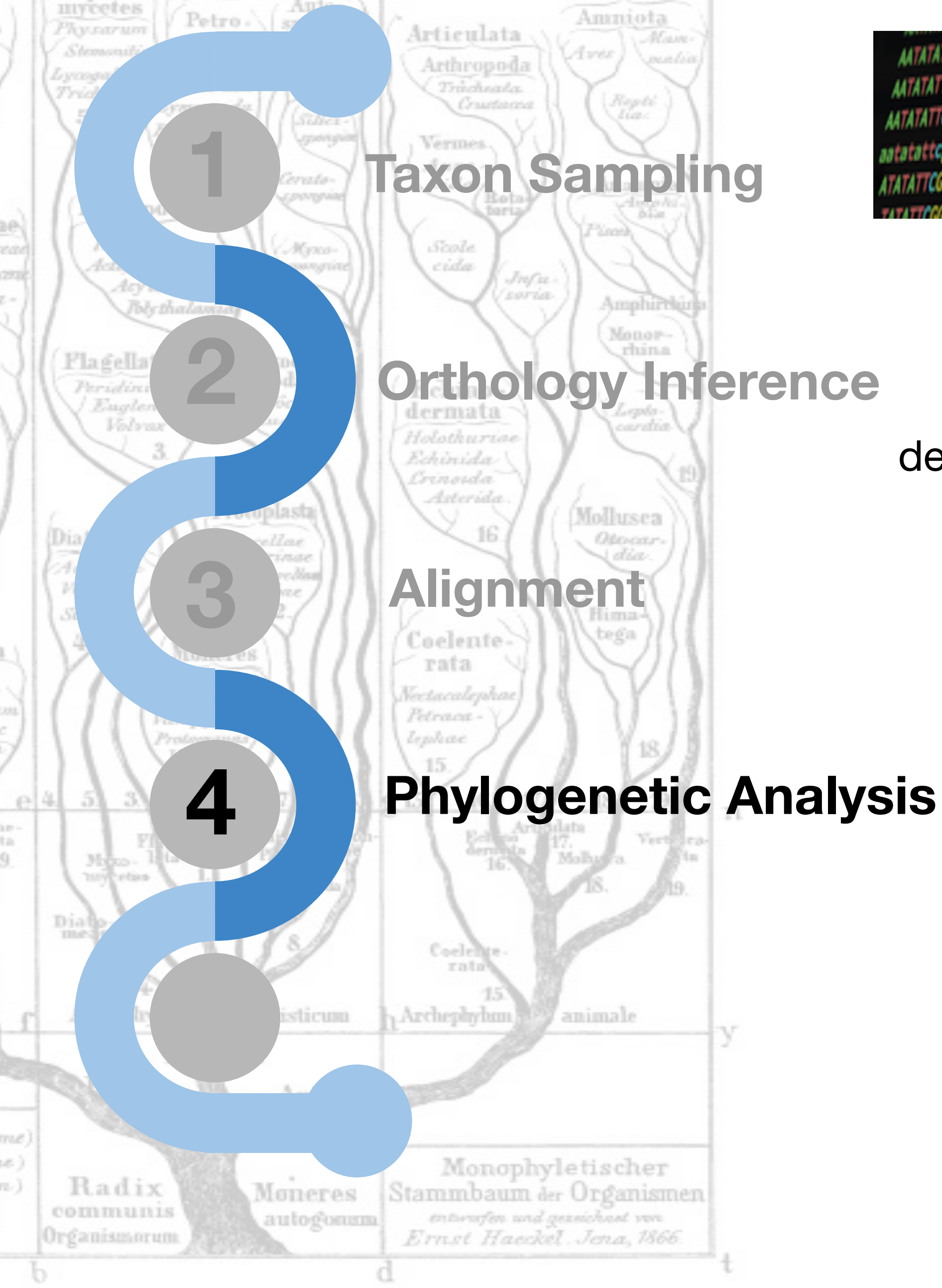
Orthology Inference

Alignment

Phylogenetic Analysis

DATA

How to infer a species tree


Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866

1 Taxon Sampling

2 Orthology Inference

3 Alignment

4 **Phylogenetic Analysis**

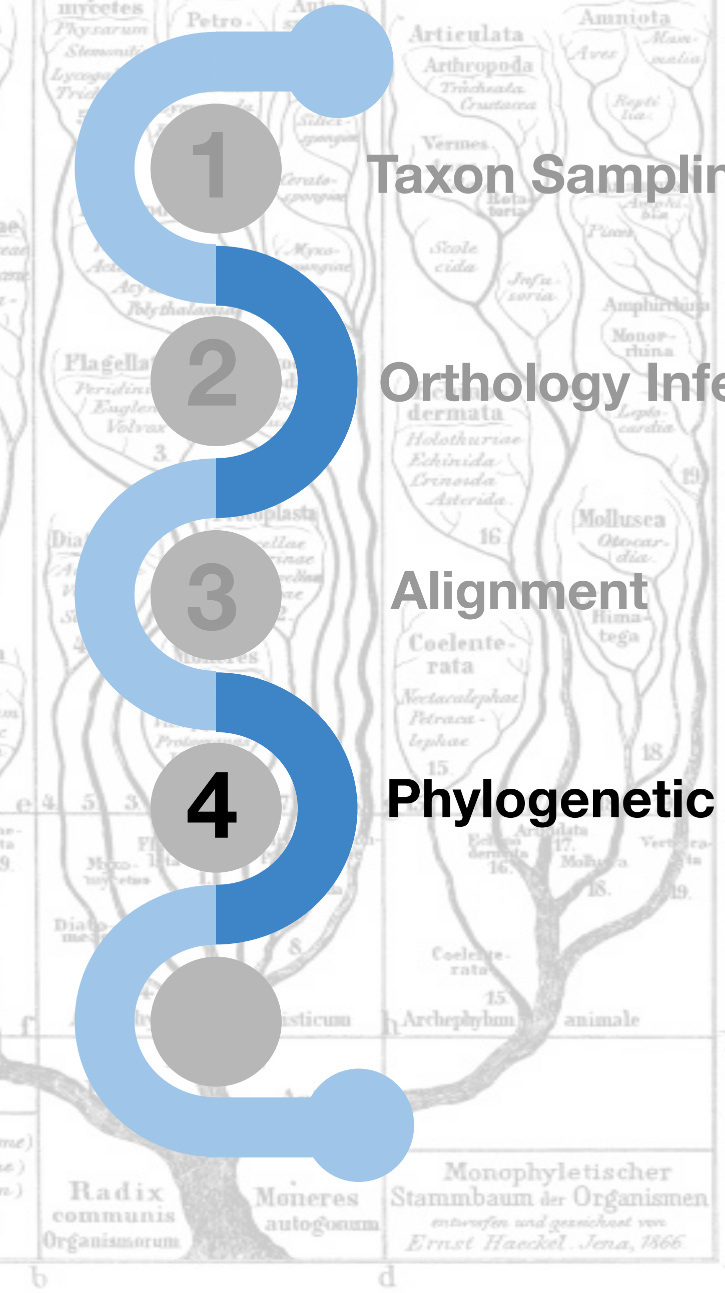
DATA

+

MODEL OF EVOLUTION

describe the relative rates of different changes

How to infer a species tree


1
2
3
4
Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866

Taxon Sampling

Orthology Inference

Alignment

Phylogenetic Analysis

DATA

MODEL OF EVOLUTION

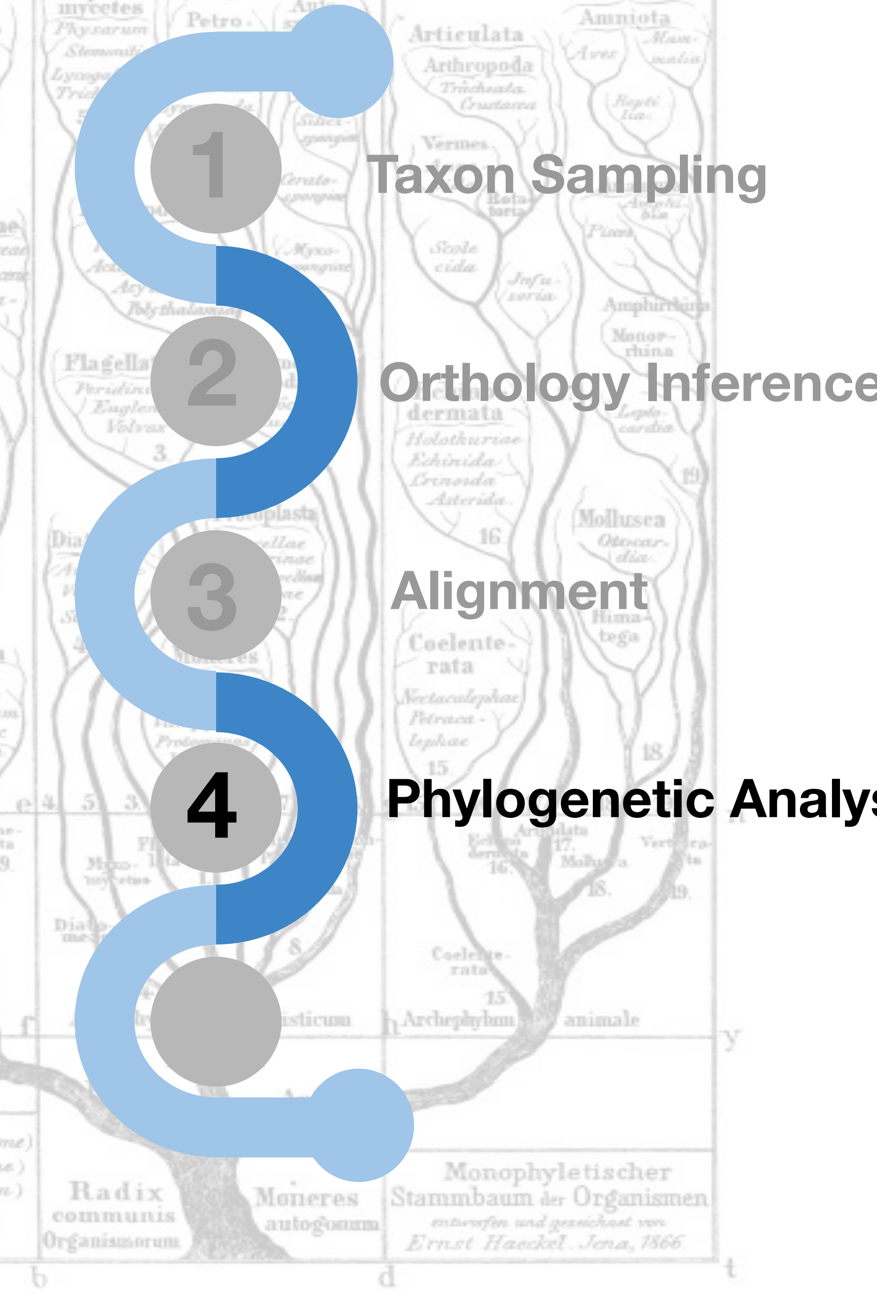
describe the relative rates of different changes

Seq1 ATGGCA

Seq2 ACGCCG

Seq3 AGGGCC

How to infer a species tree



1 Taxon Sampling

2 Orthology Inference

3 Alignment

4 **Phylogenetic Analysis**

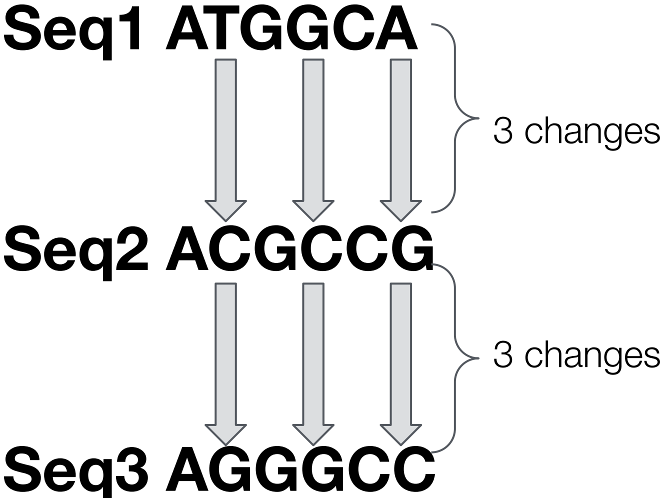
Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum
Archephylum animale
Articulata
Arthropoda
Amniota
Vermes
Mollusca
Coelenterata

DATA

+

MODEL OF EVOLUTION

describe the relative rates of different changes



Seq1 ATGGCA

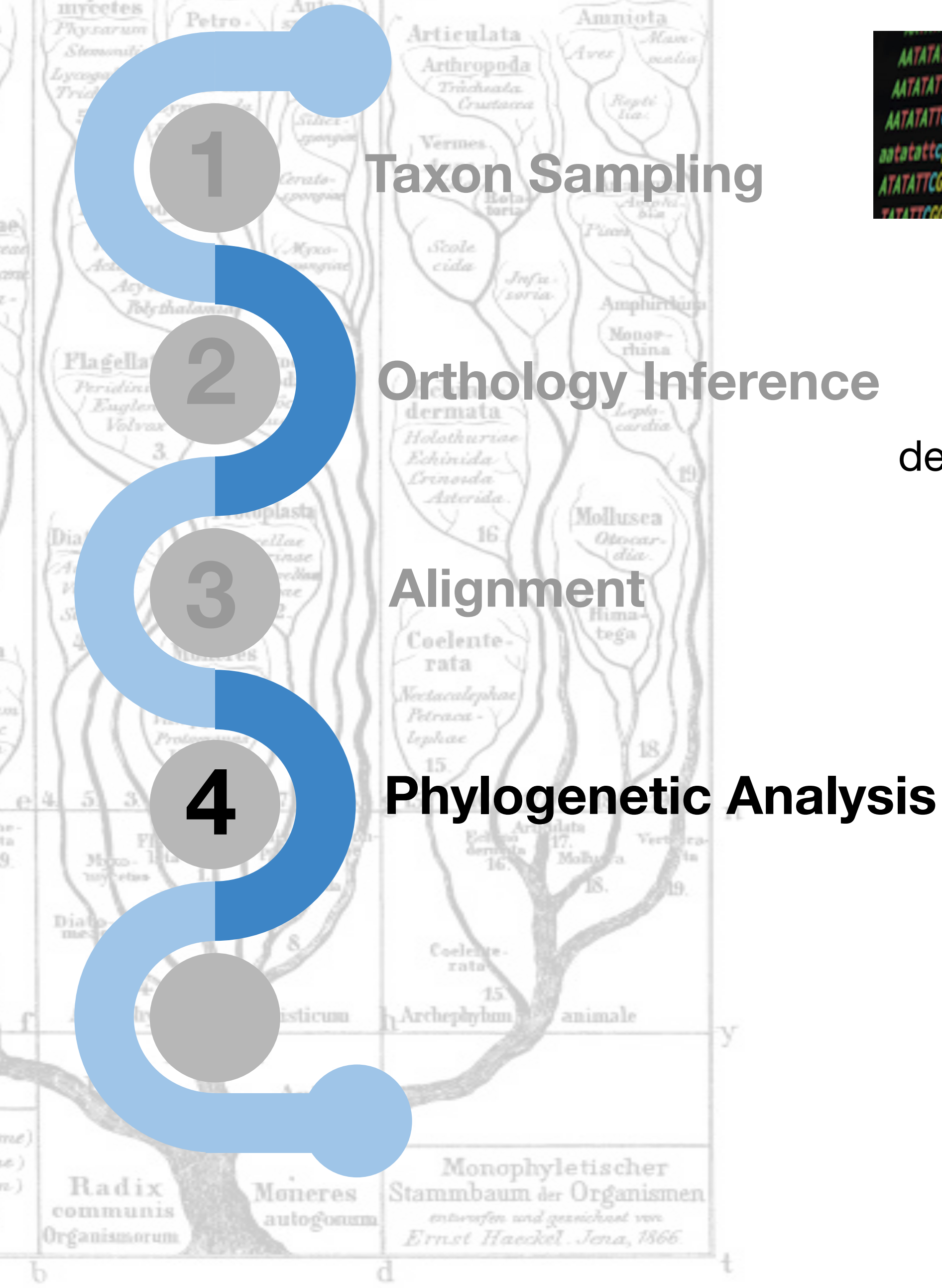
3 changes

Seq2 ACGCCG

3 changes

Seq3 AGGGCC

How to infer a species tree



DATA

+

MODEL OF EVOLUTION

describe the relative rates of different changes

Seq1 ATGGCA

3 changes

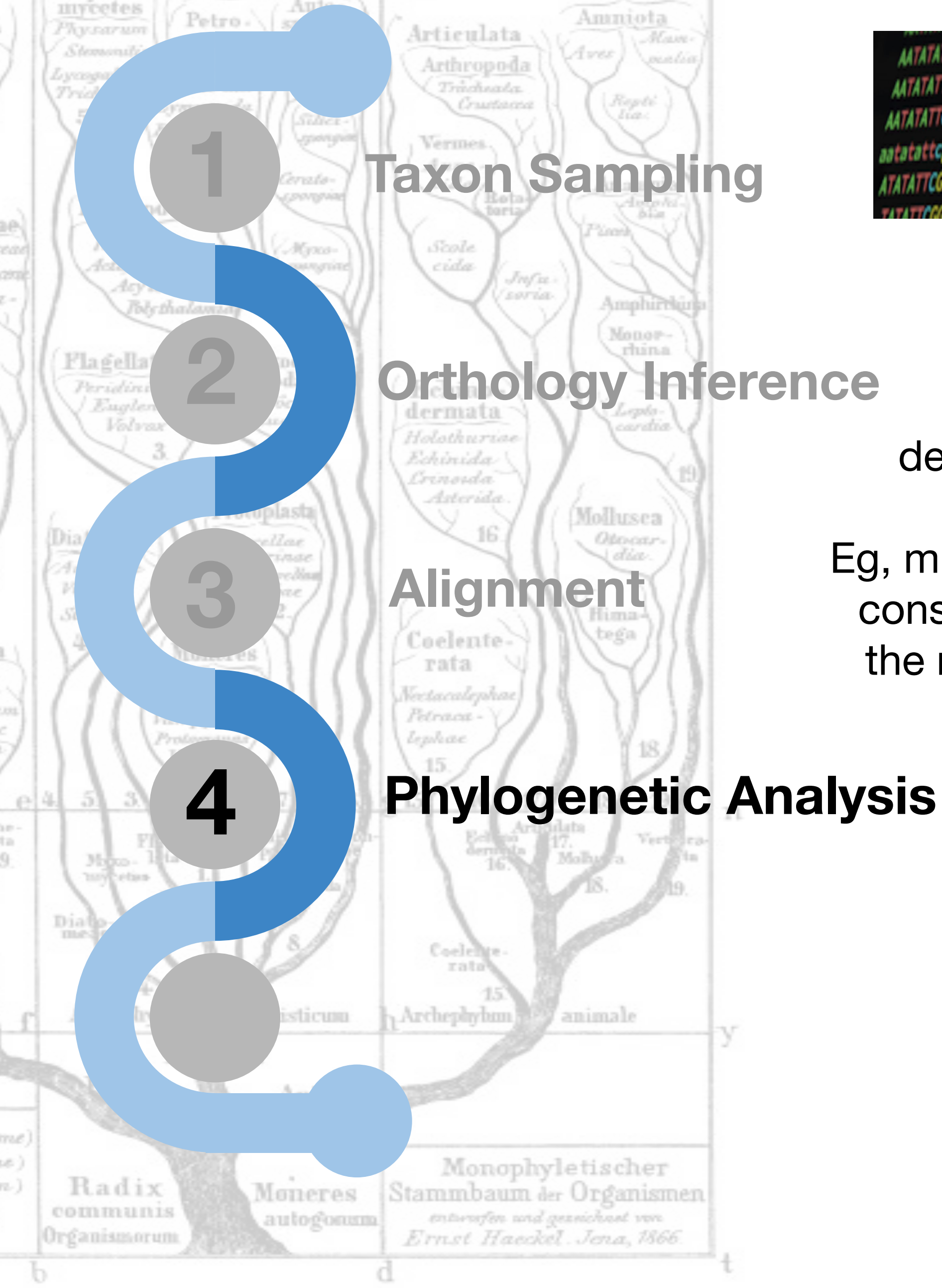
Seq2 ACGCCG

3 changes

Seq3 AGGGCC

Jukes and Cantor 1969 (JC69)

How to infer a species tree



1 Taxon Sampling

2 Orthology Inference

3 Alignment

4 Phylogenetic Analysis

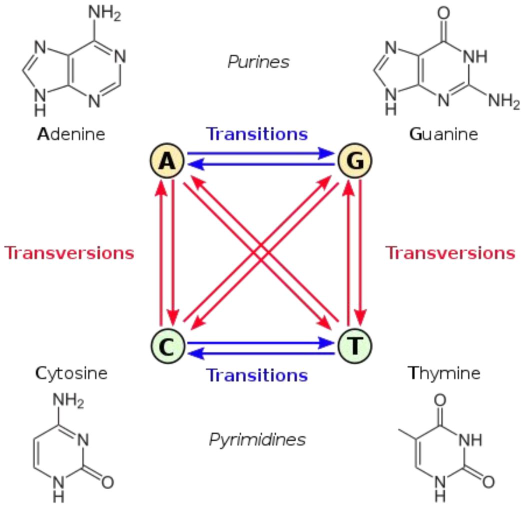
DATA

+

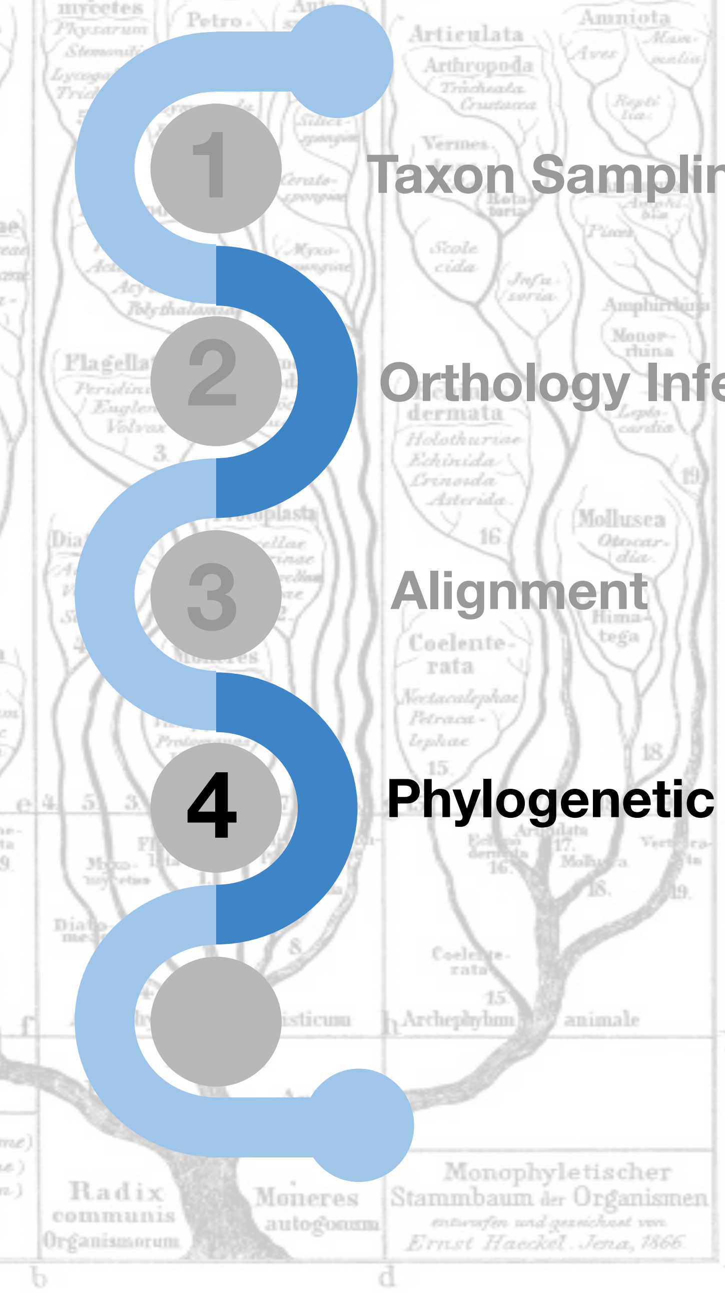
MODEL OF EVOLUTION

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



How to infer a species tree



1
2
3
4
Monophyletischer Stammbaum der Organismen
Ernst Haeckel. Jena, 1866

1 Taxon Sampling

2 Orthology Inference

3 Alignment

4 **Phylogenetic Analysis**

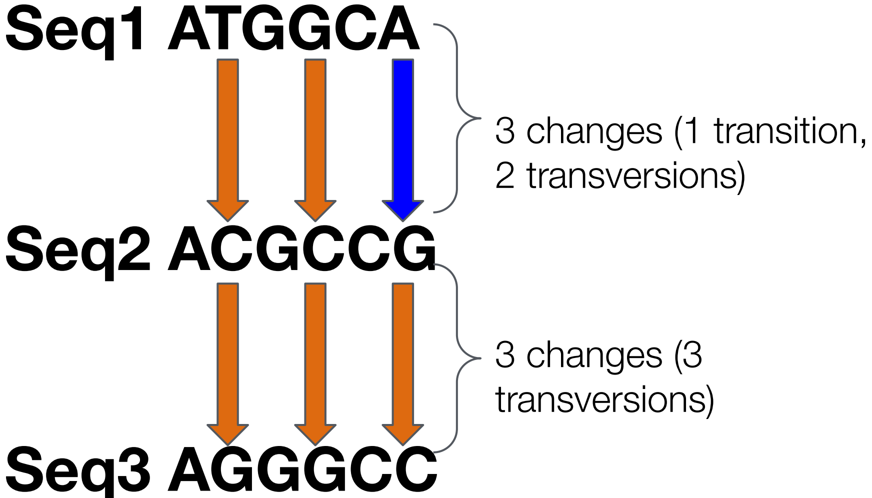
DATA

+

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



Seq1 ATGGCA

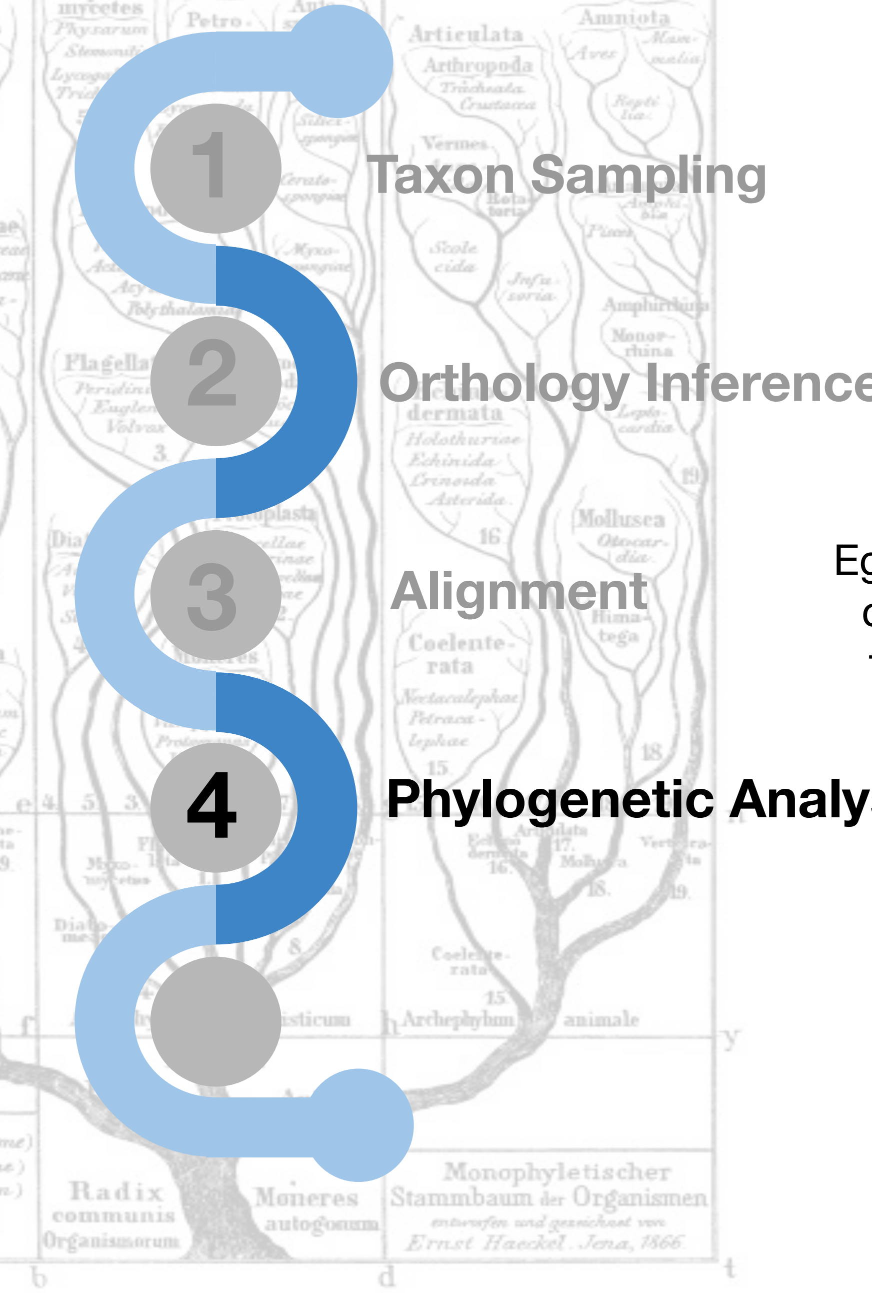
3 changes (1 transition, 2 transversions)

Seq2 ACGCCG

3 changes (3 transversions)

Seq3 AGGGCC

How to infer a species tree



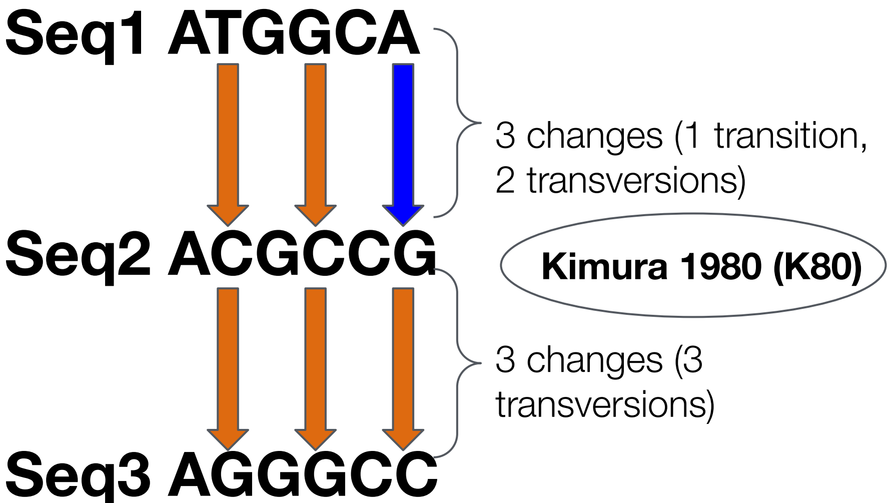
DATA

+

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



Seq1 ATGGCA

3 changes (1 transition, 2 transversions)

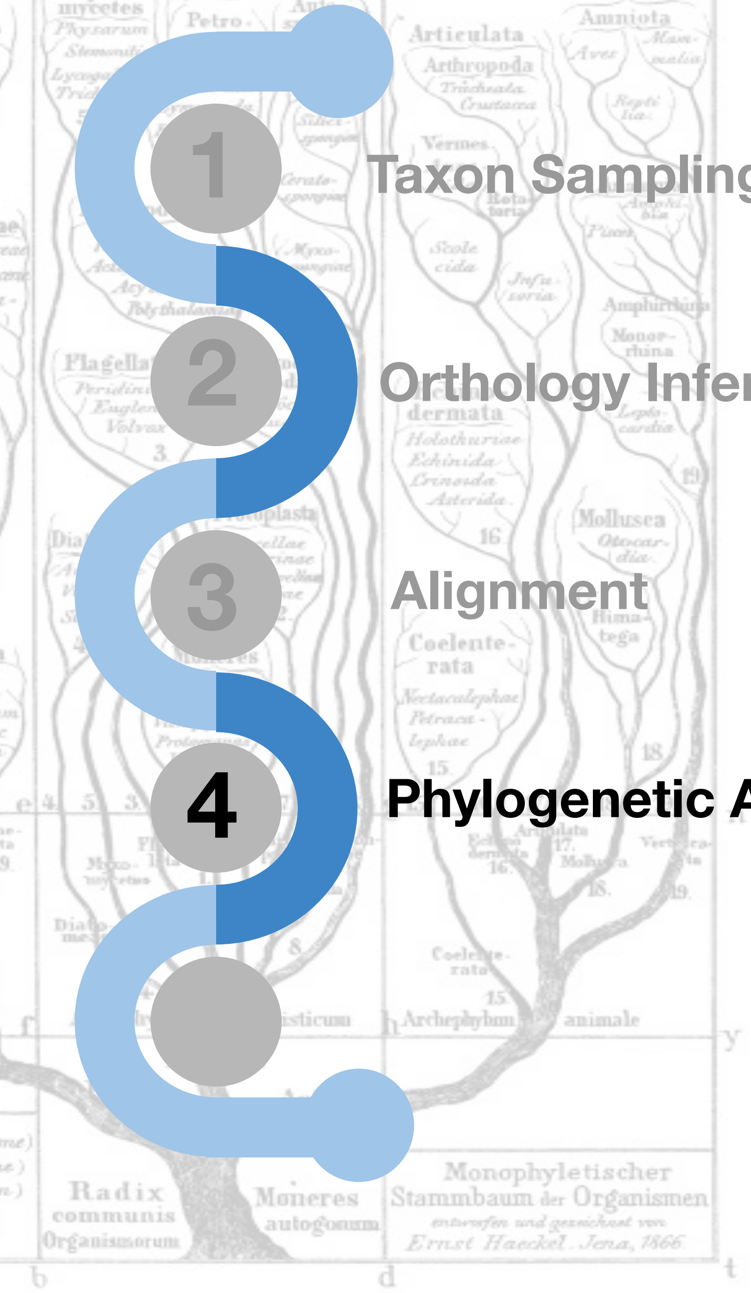
Seq2 ACGCCG

Kimura 1980 (K80)

3 changes (3 transversions)

Seq3 AGGGCC

How to infer a species tree



Taxon Sampling

Orthology Inference

Alignment

Phylogenetic Analysis

DATA

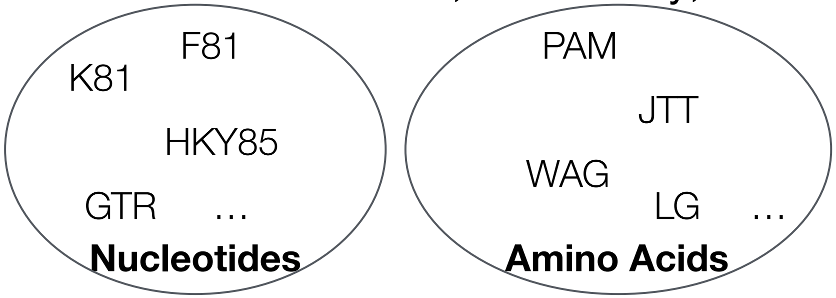
+

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

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



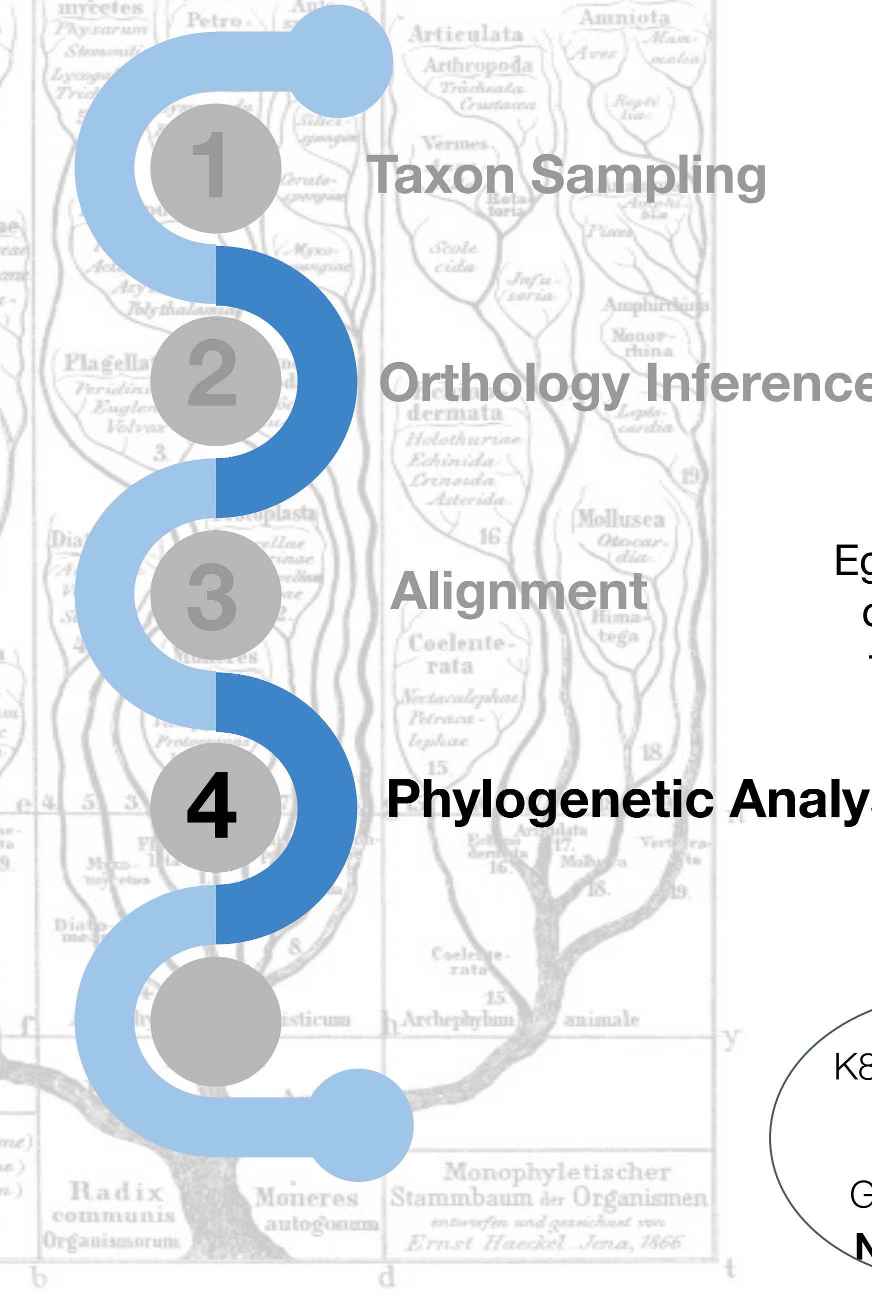
K81 F81 HKY85 GTR …

Nucleotides

PAM JTT WAG LG …

Amino Acids

How to infer a species tree



1 Taxon Sampling

2 Orthology Inference

3 Alignment

4 **Phylogenetic Analysis**

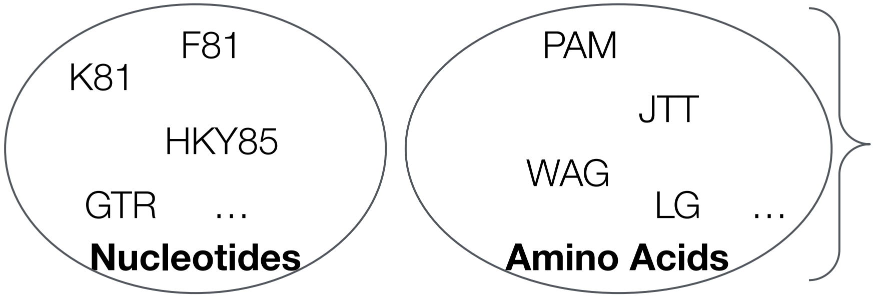
DATA

+

MODEL OF EVOLUTION

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



K81 F81 HKY85 GTR …

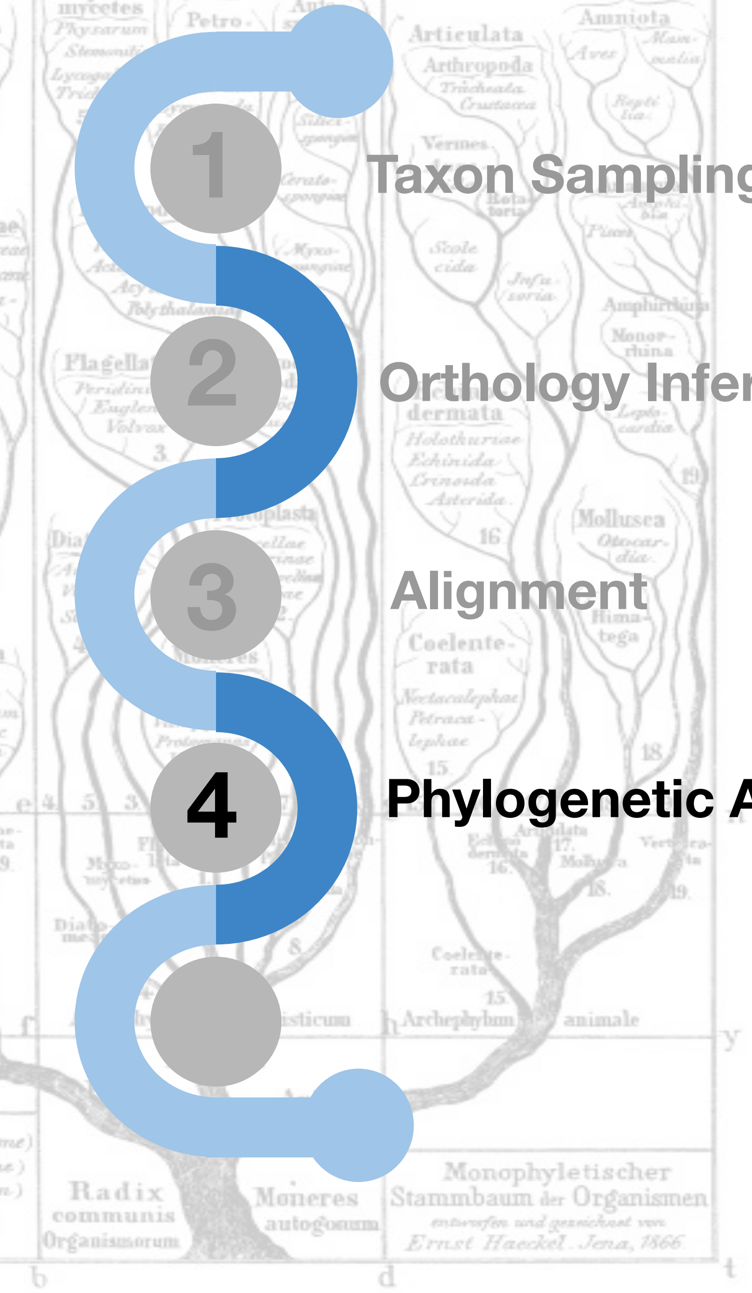
Nucleotides

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

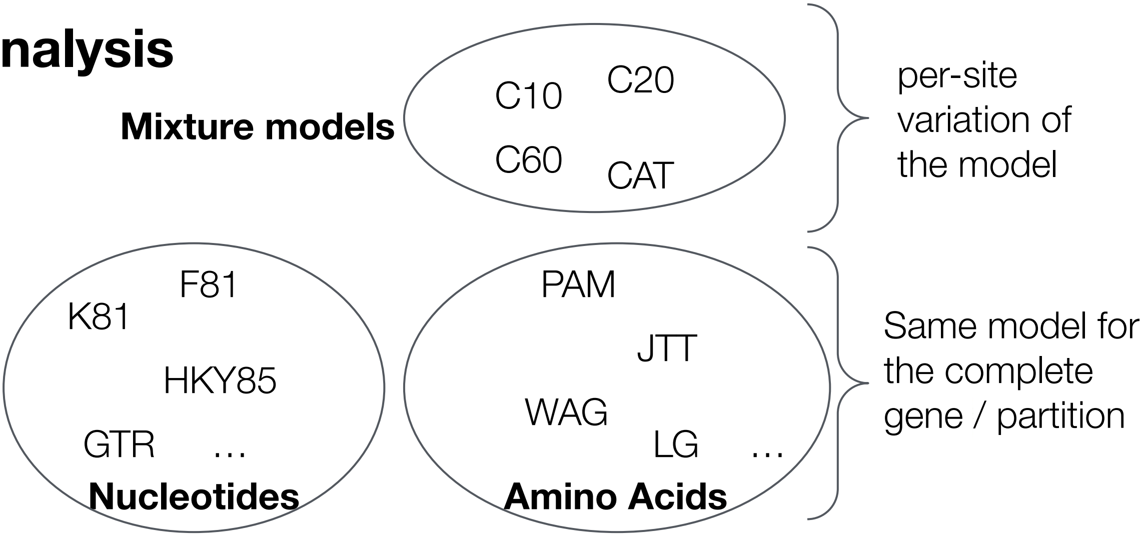
DATA

+

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 K81 HKY85 GTR …

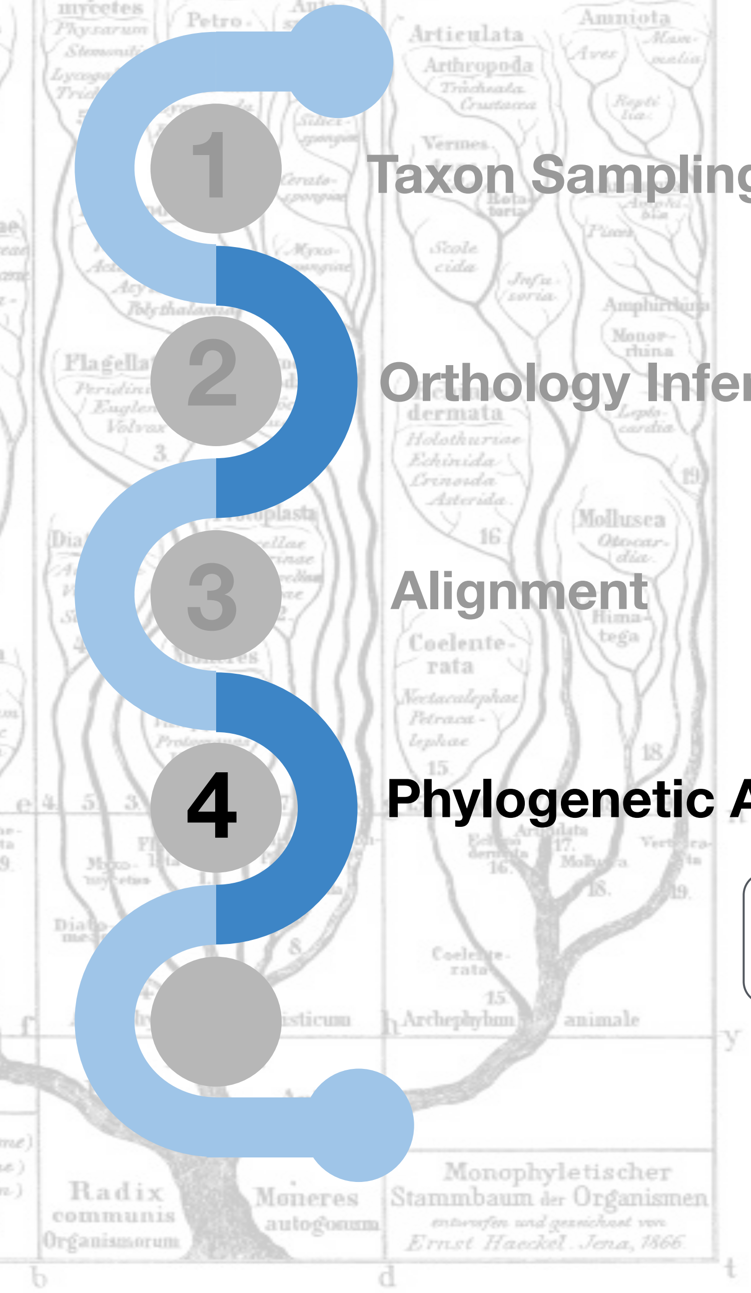
Nucleotides

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

+

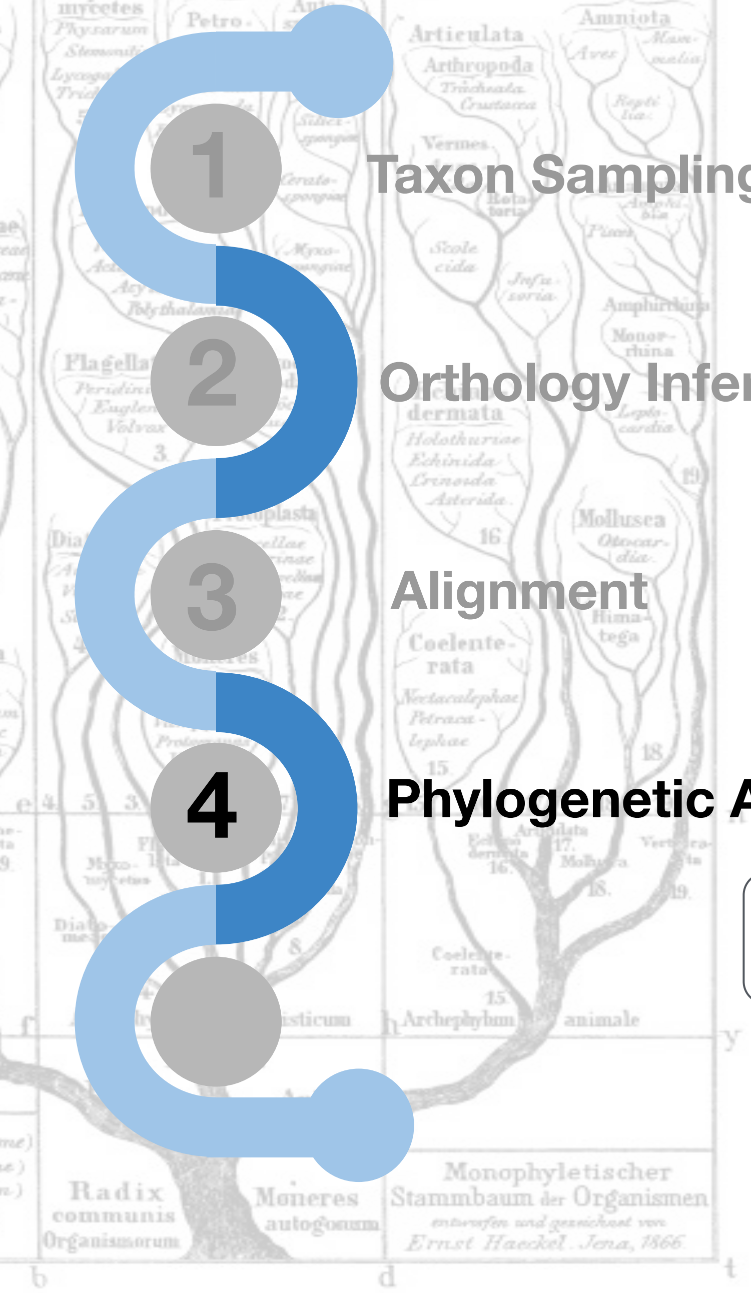
MODEL OF EVOLUTION

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


Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum

1 Taxon Sampling

2 Orthology Inference

3 Alignment

4 **Phylogenetic Analysis**

DATA

➕

MODEL OF EVOLUTION

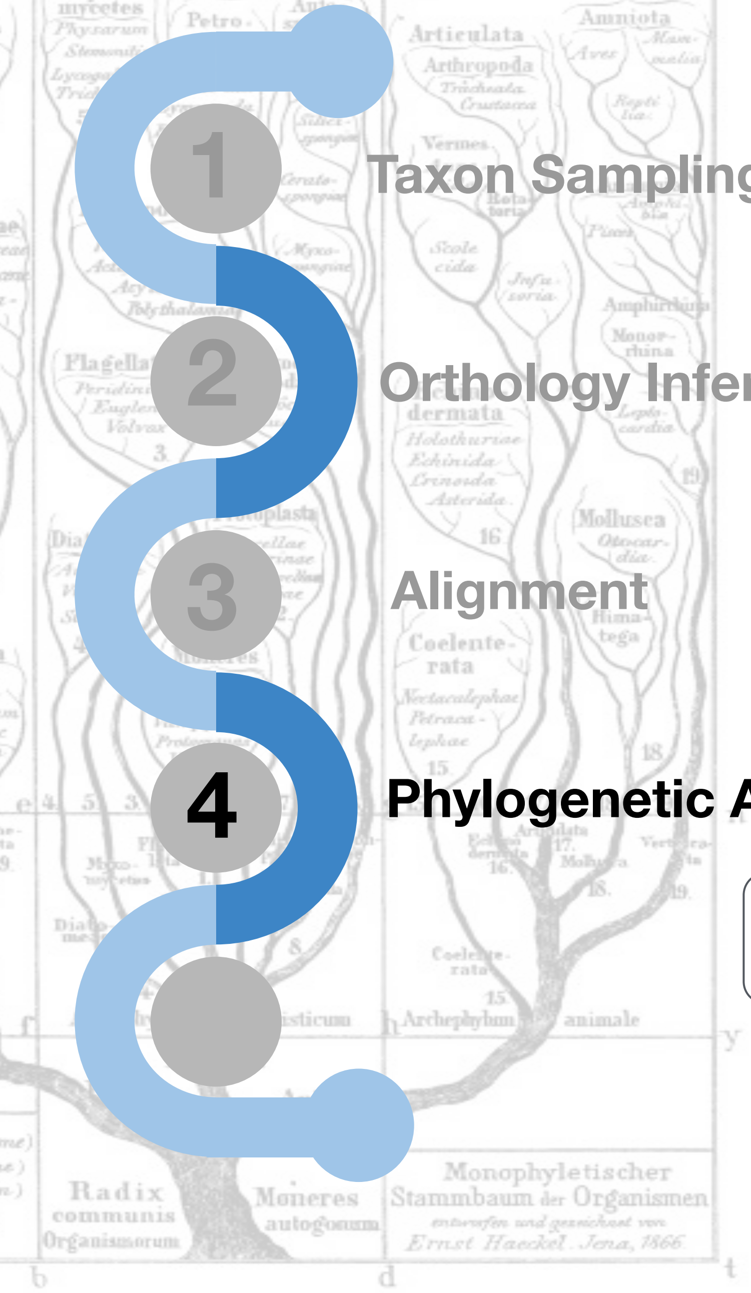
describe the relative rates of different changes

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
2
3
4
Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel Jena, 1866

1 Taxon Sampling

2 Orthology Inference

3 Alignment

4 Phylogenetic Analysis

DATA

+

MODEL OF EVOLUTION

describe the relative rates of different changes

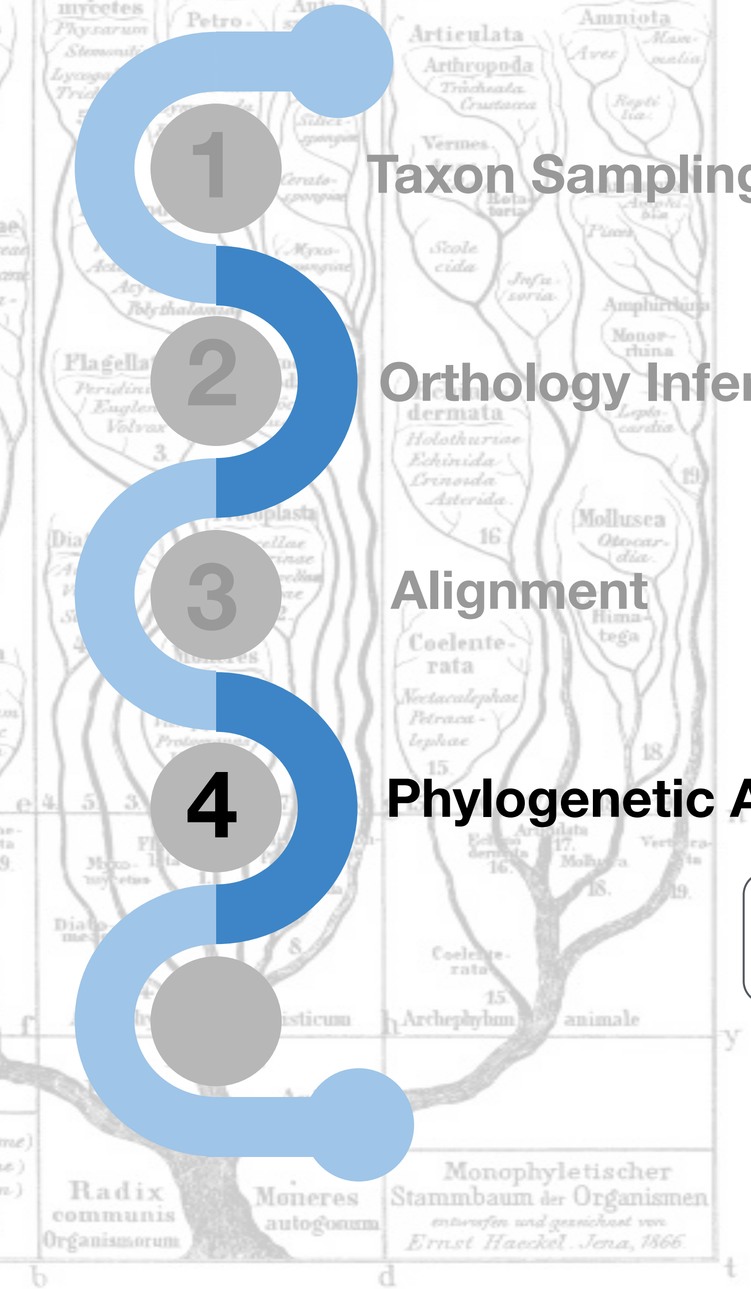
Eg, mutational biases and purifying selection favoring conservative changes are probably responsible for the relatively high rate of **transitions** compared to **transversions** in evolving sequences

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



Taxon Sampling

Orthology Inference

Alignment

Phylogenetic Analysis

AATATATTCGGGTAACCGAAGATGAGAACGATGAGCCCATTGAA
AATATATT.GGGTAACCGAAGATGAGAACGATGAGCCCATTGAA
AATATATTCGGGTAACCGAAGATGAGAACGATGAGCCCATTGAA
aatatattcgggtaaccgaagatgagaacgatgagcccattgaa
ATATATTCGGGTAACCGAAGATGAGAACGATGAGCCCATTGAAT

DATA

+

MODEL OF EVOLUTION

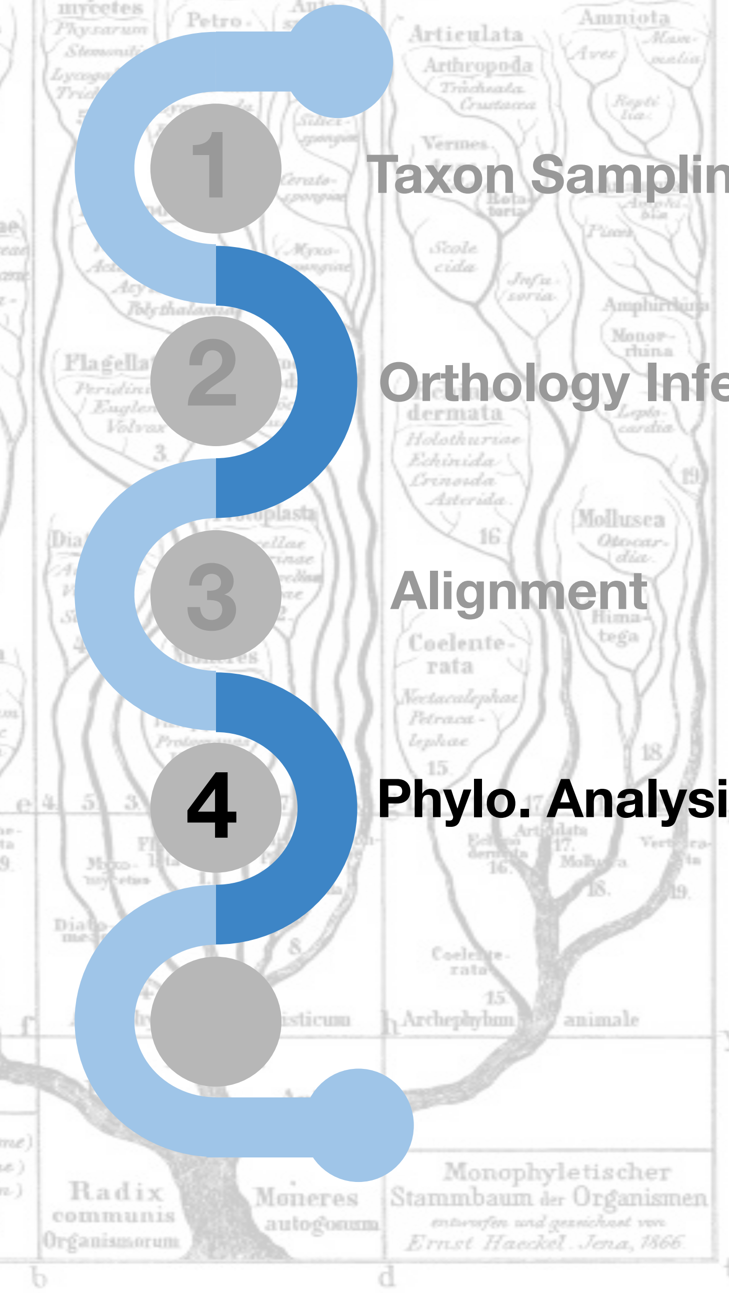
describe the relative rates of different changes

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


Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum

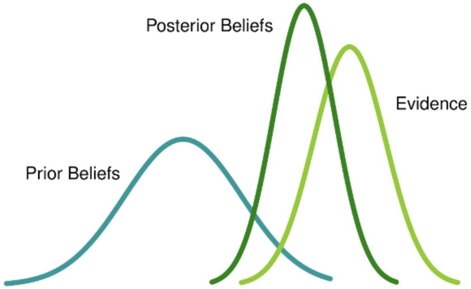
1 Taxon Sampling

2 Orthology Inference

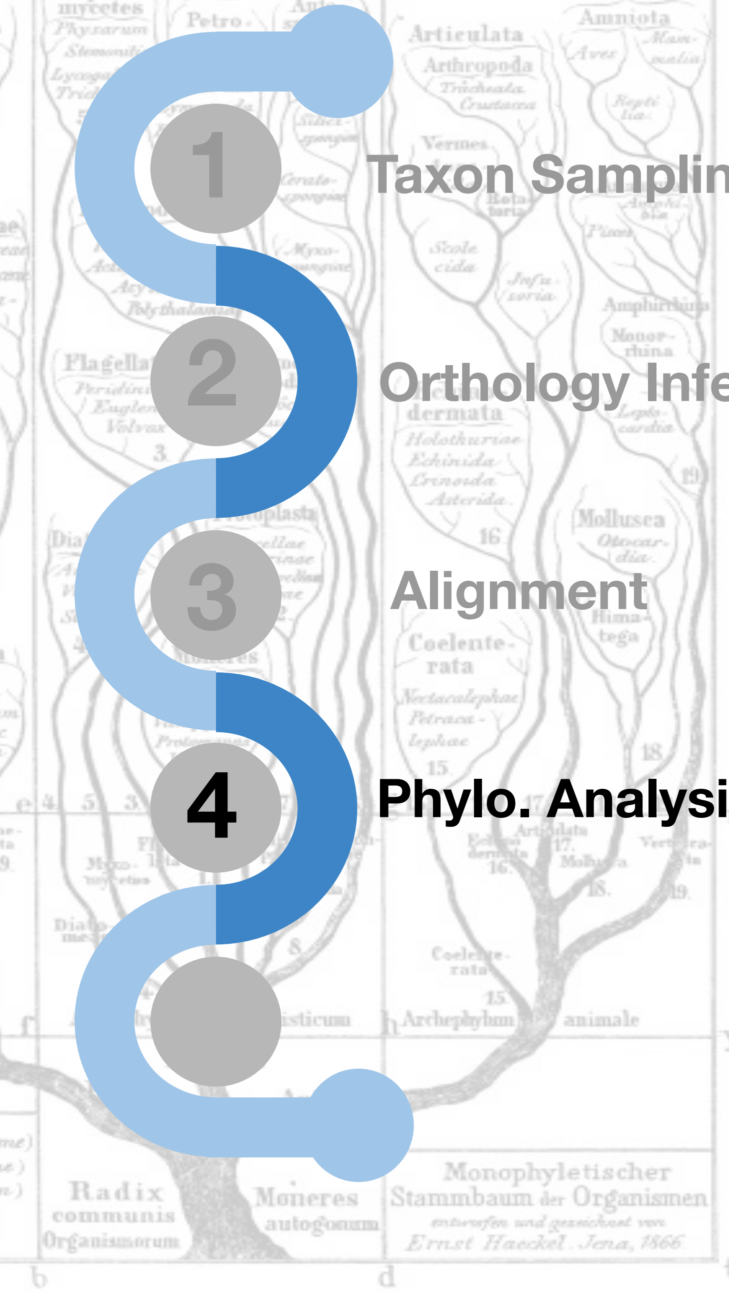
3 Alignment

4 Phylo. Analysis

DATA + **MODEL OF EVOLUTION** + **METHOD**



How to infer a species tree



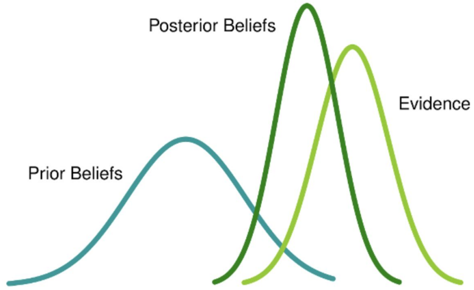
Taxon Sampling

Orthology Inference

Alignment

Phylo. Analysis

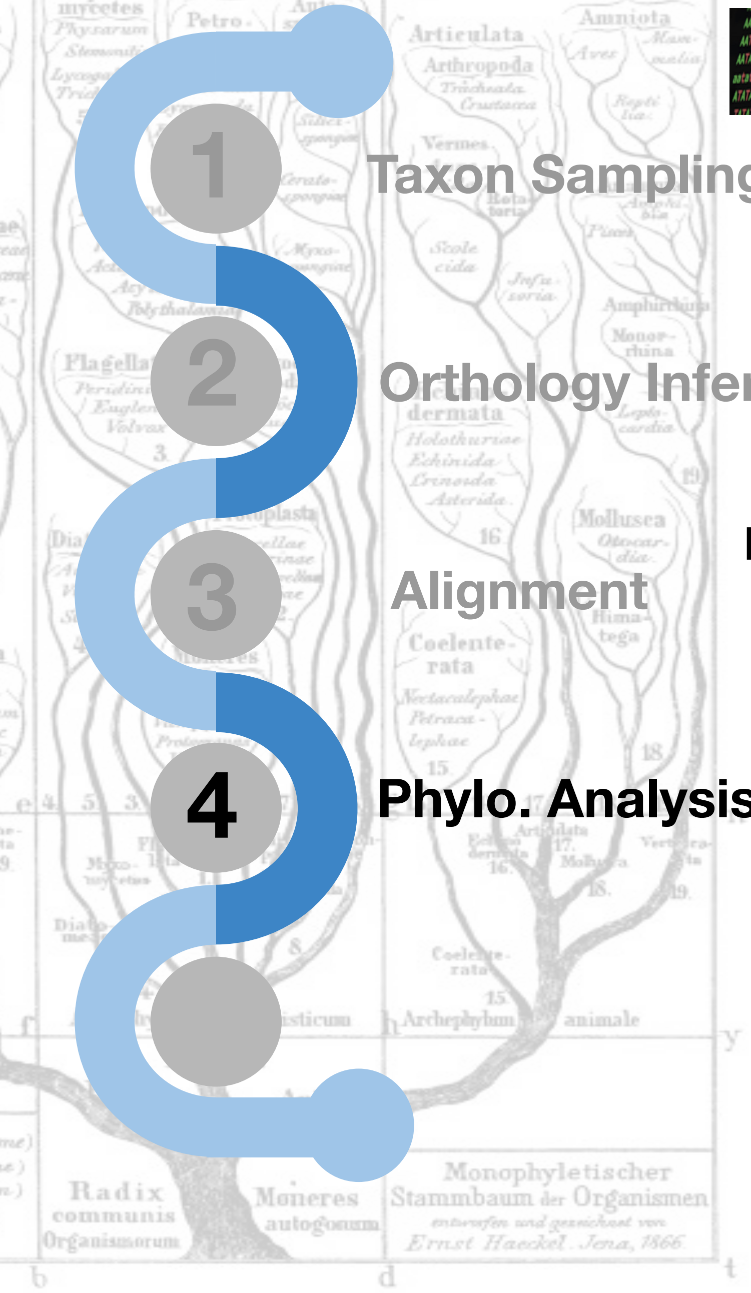
DATA + **MODEL OF EVOLUTION** + **METHOD**



Two main methods:

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

How to infer a species tree



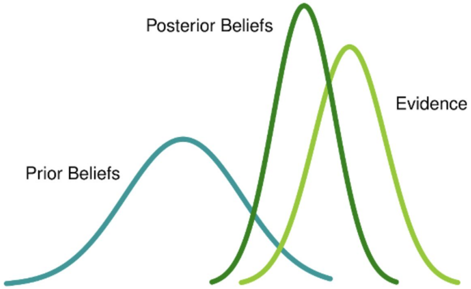
1 Taxon Sampling

2 Orthology Inference

3 Alignment

4 Phylo. Analysis

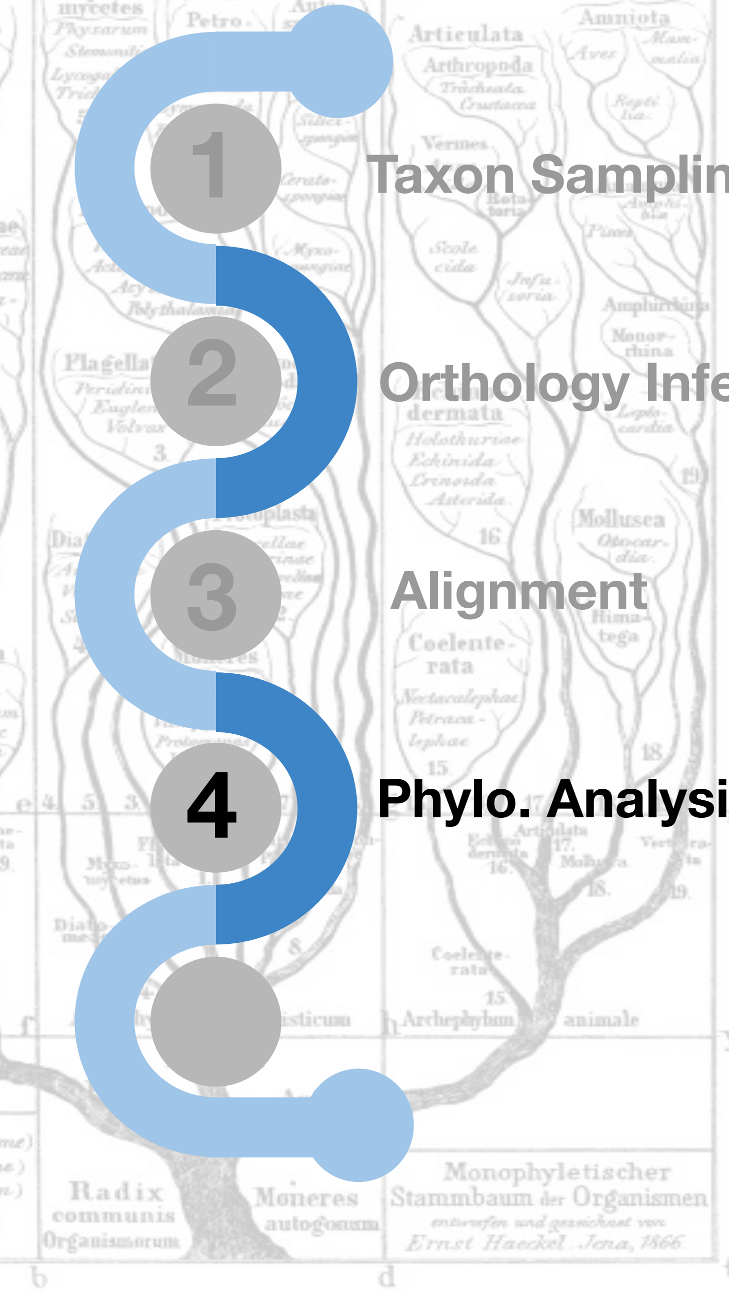
DATA ➕ **MODEL OF EVOLUTION** ➕ **METHOD**



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

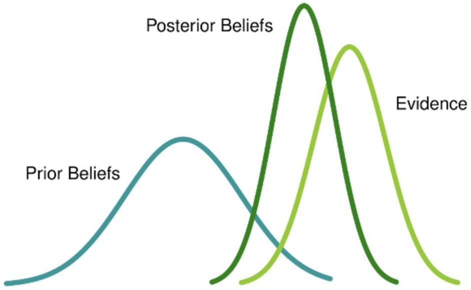
2 Orthology Inference

3 Alignment

4 **Phylo. Analysis**

Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum

DATA + **MODEL OF EVOLUTION** + **METHOD**

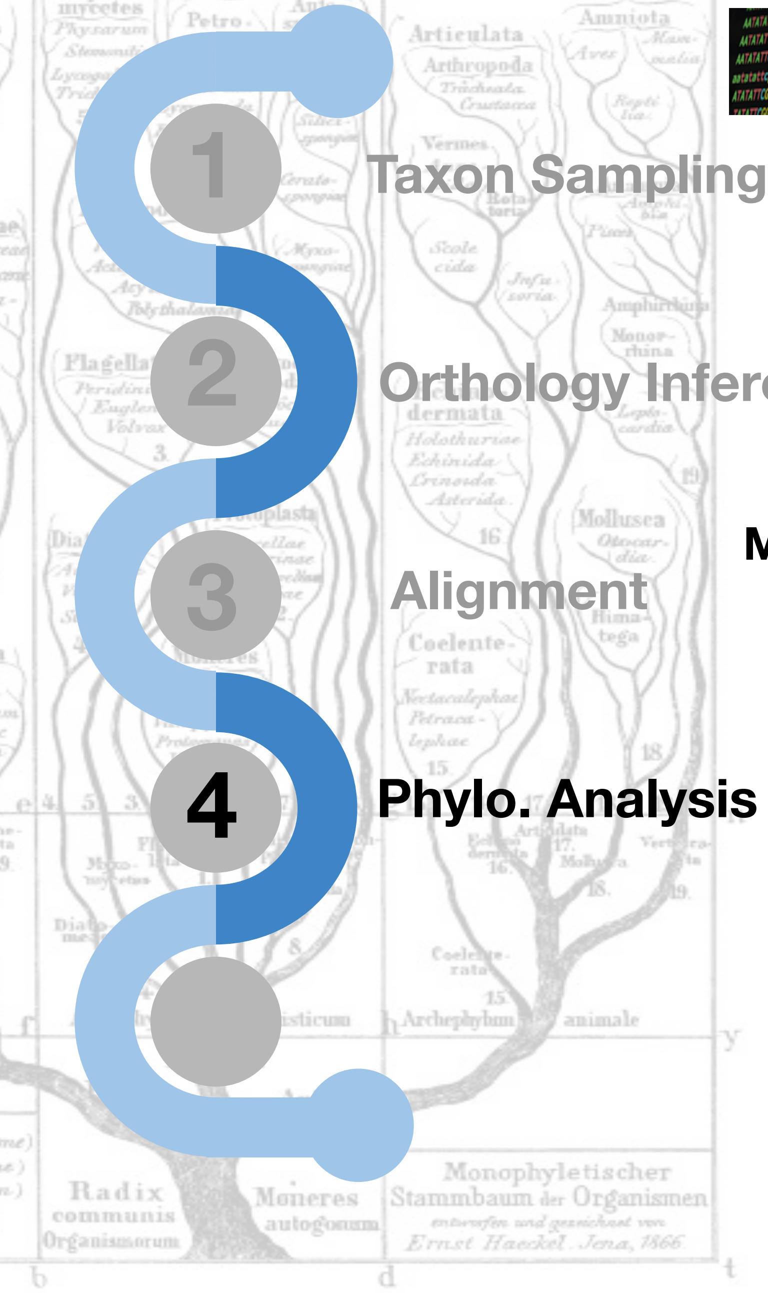


Two main methods:
Maximum Likelihood (ML) and **Bayesian Inference (BI)**

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



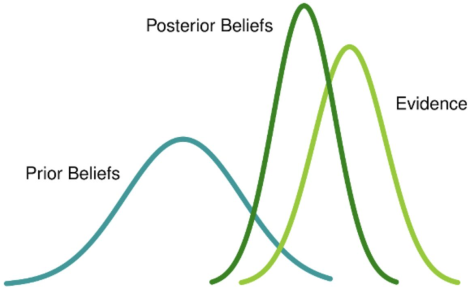
1 Taxon Sampling

2 Orthology Inference

3 Alignment

4 Phylo. Analysis

DATA + **MODEL OF EVOLUTION** + **METHOD**



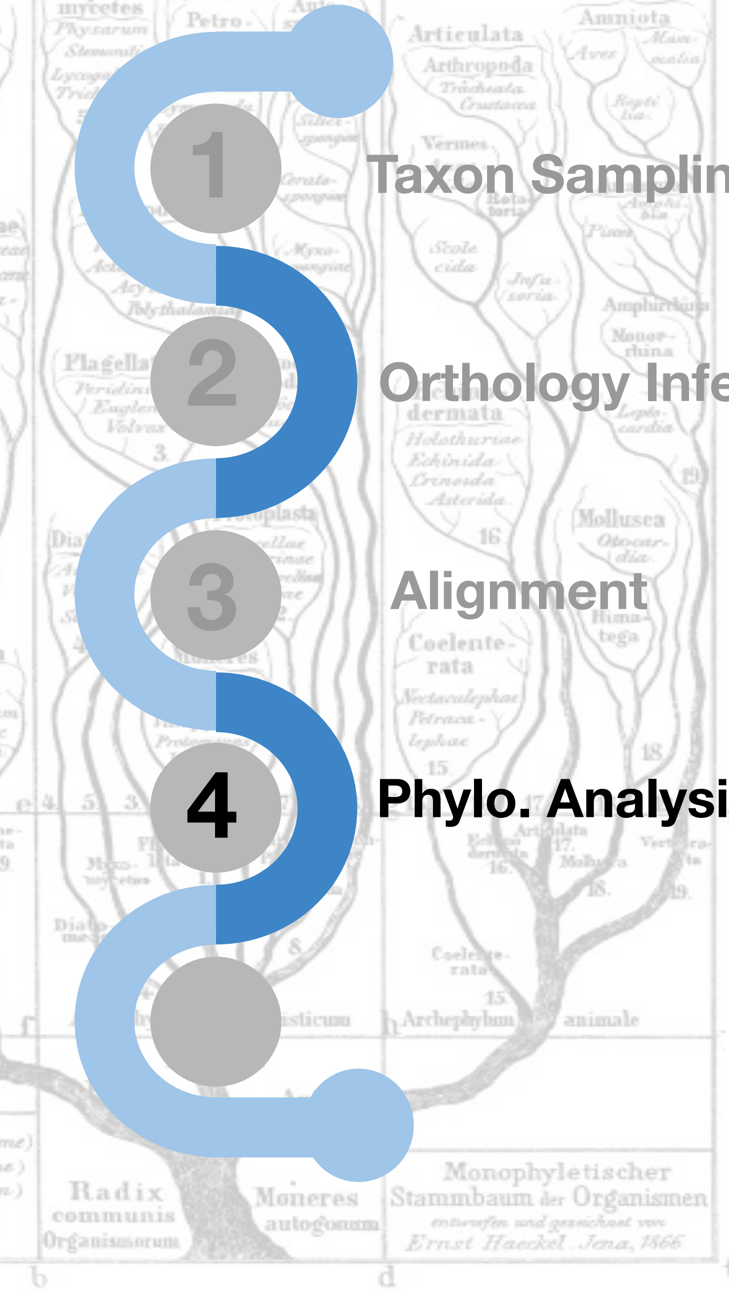
Two main methods:
Maximum Likelihood (ML) and **Bayesian Inference (BI)**

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


Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum

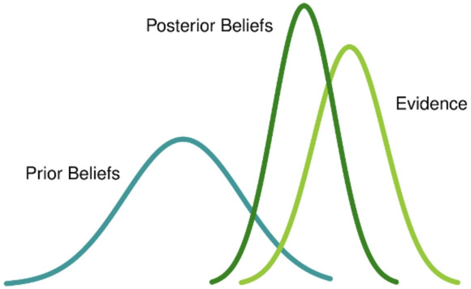
1 Taxon Sampling

2 Orthology Inference

3 Alignment

4 **Phylo. Analysis**

DATA ➕ **MODEL OF EVOLUTION** ➕ **METHOD**



Two main methods:

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

RevBayes

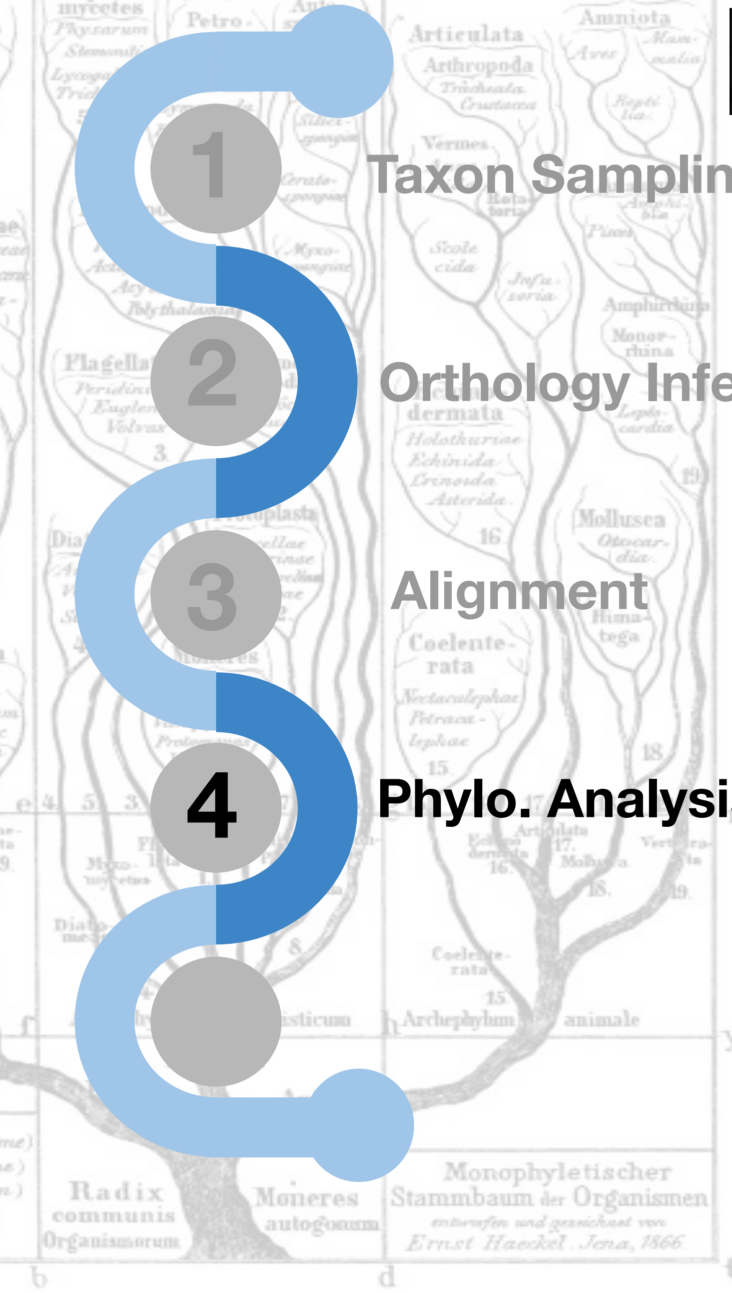
RAxML

IQ-TREE

fasrc/**phylobayes**

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

How to infer a species tree


1
2
3
4
Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum

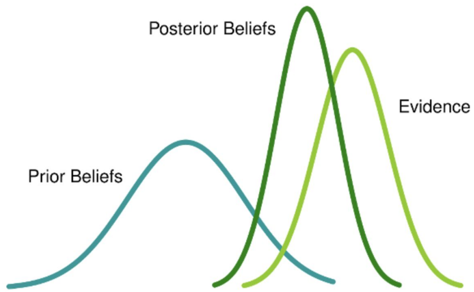
1 Taxon Sampling

2 Orthology Inference

3 Alignment

4 **Phylo. Analysis**

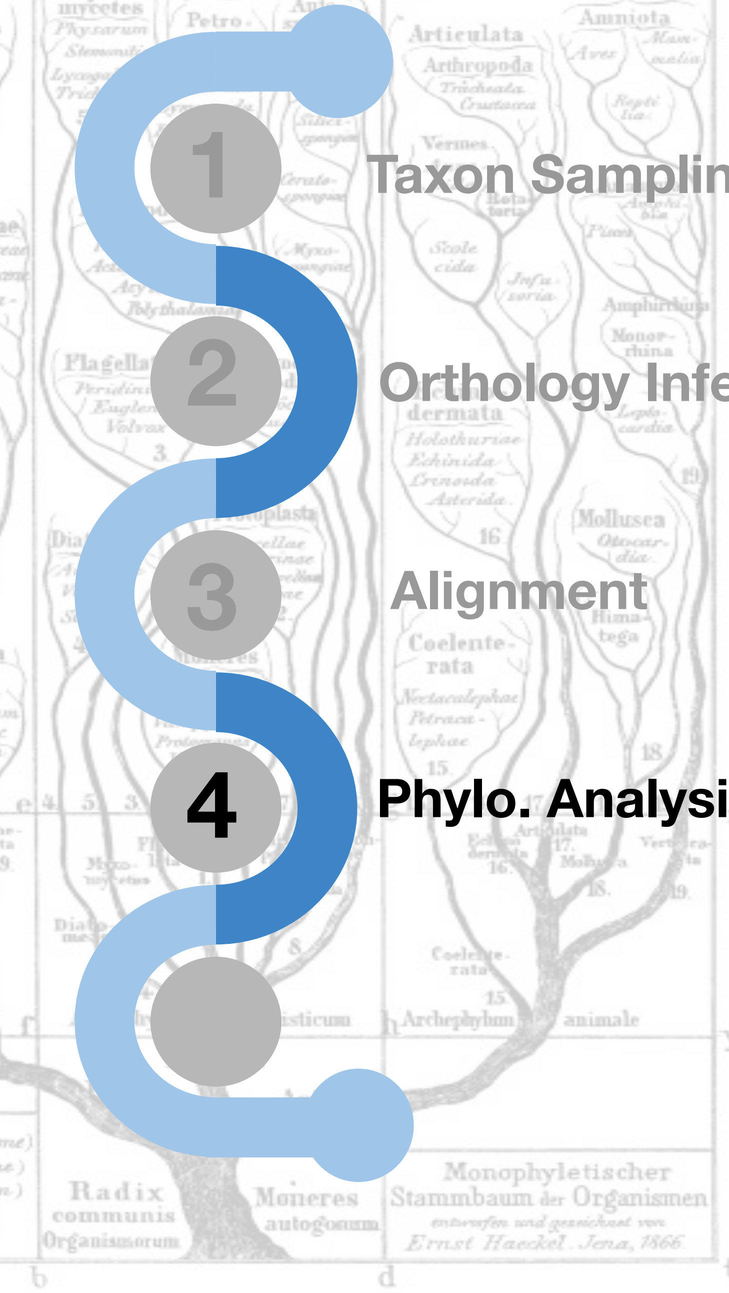
DATA + **MODEL OF EVOLUTION** + **METHOD**



Two main methods:
Maximum Likelihood (ML) and **Bayesian Inference (BI)**

Which one should I choose?

How to infer a species tree



1 Taxon Sampling

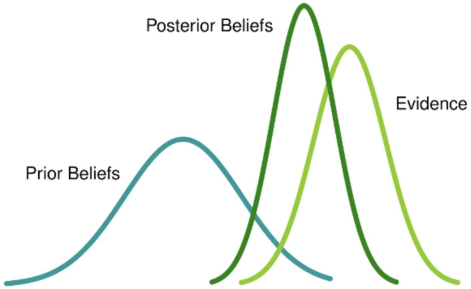
2 Orthology Inference

3 Alignment

4 Phylo. Analysis

Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum

DATA + **MODEL OF EVOLUTION** + **METHOD**

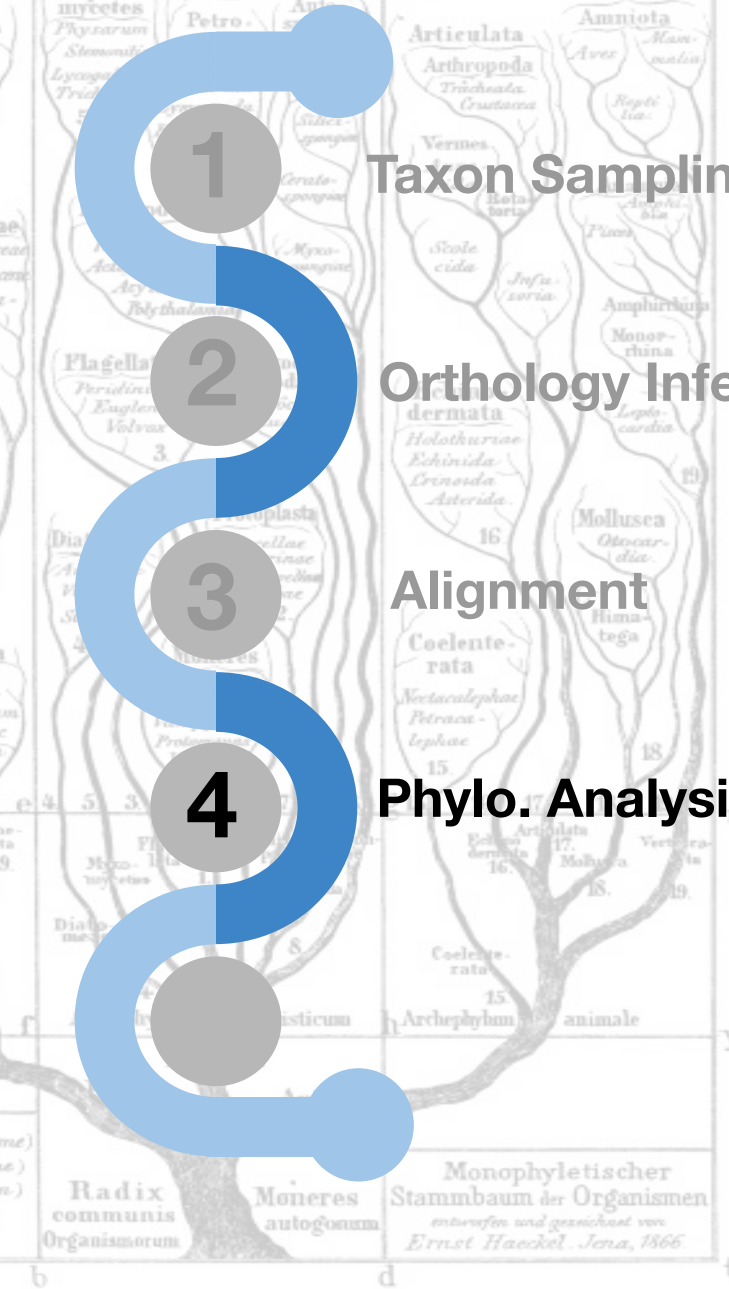


Two main methods:

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

Which one should I choose?

How to infer a species tree


1 Taxon Sampling

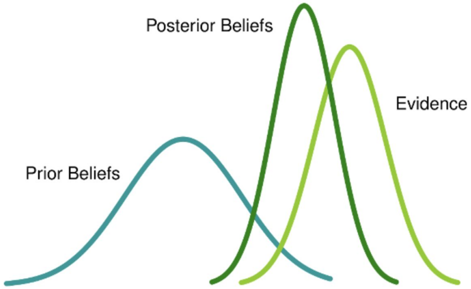
2 Orthology Inference

3 Alignment

4 **Phylo. Analysis**

Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum

DATA + **MODEL OF EVOLUTION** + **METHOD**

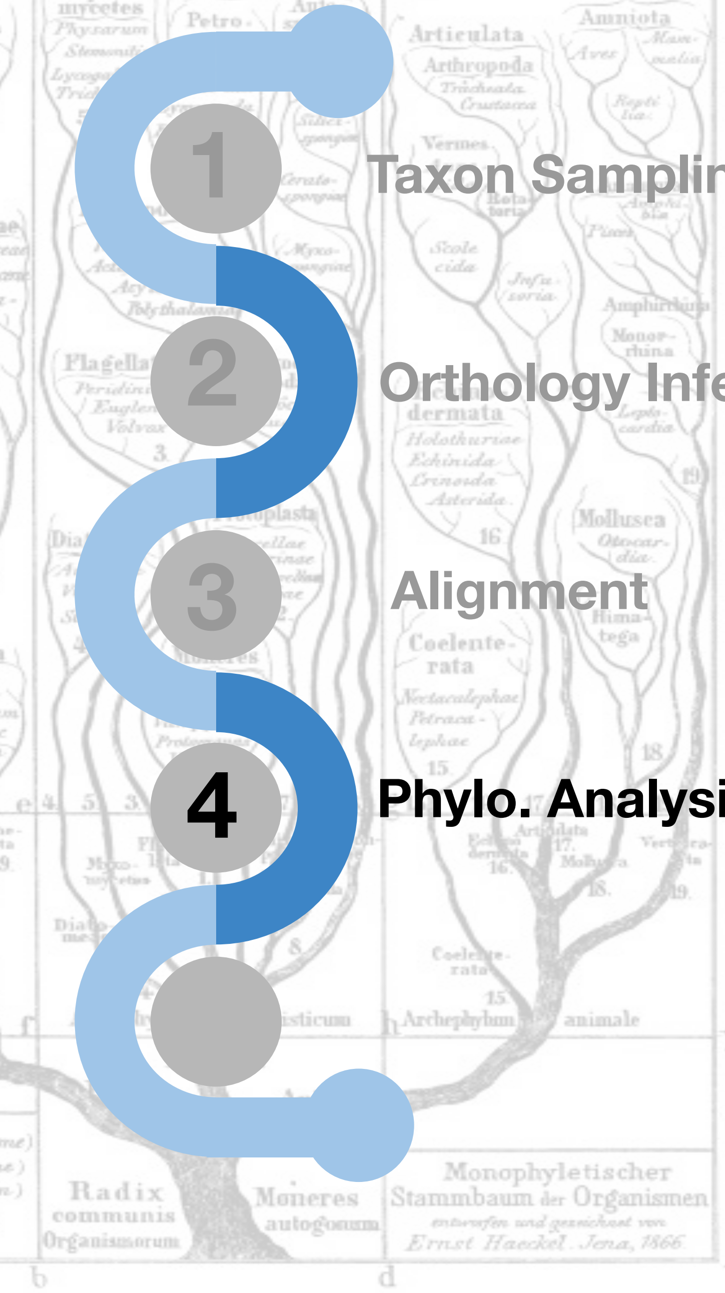


Two main methods:
Maximum Likelihood (ML) and **Bayesian Inference (BI)**

Which one should I choose?

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

How to infer a species tree



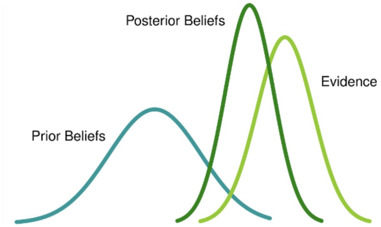
1 Taxon Sampling

2 Orthology Inference

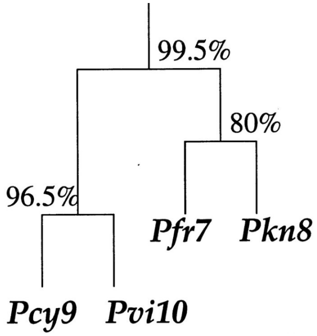
3 Alignment

4 Phylo. Analysis

DATA + **MODEL OF EVOLUTION** +

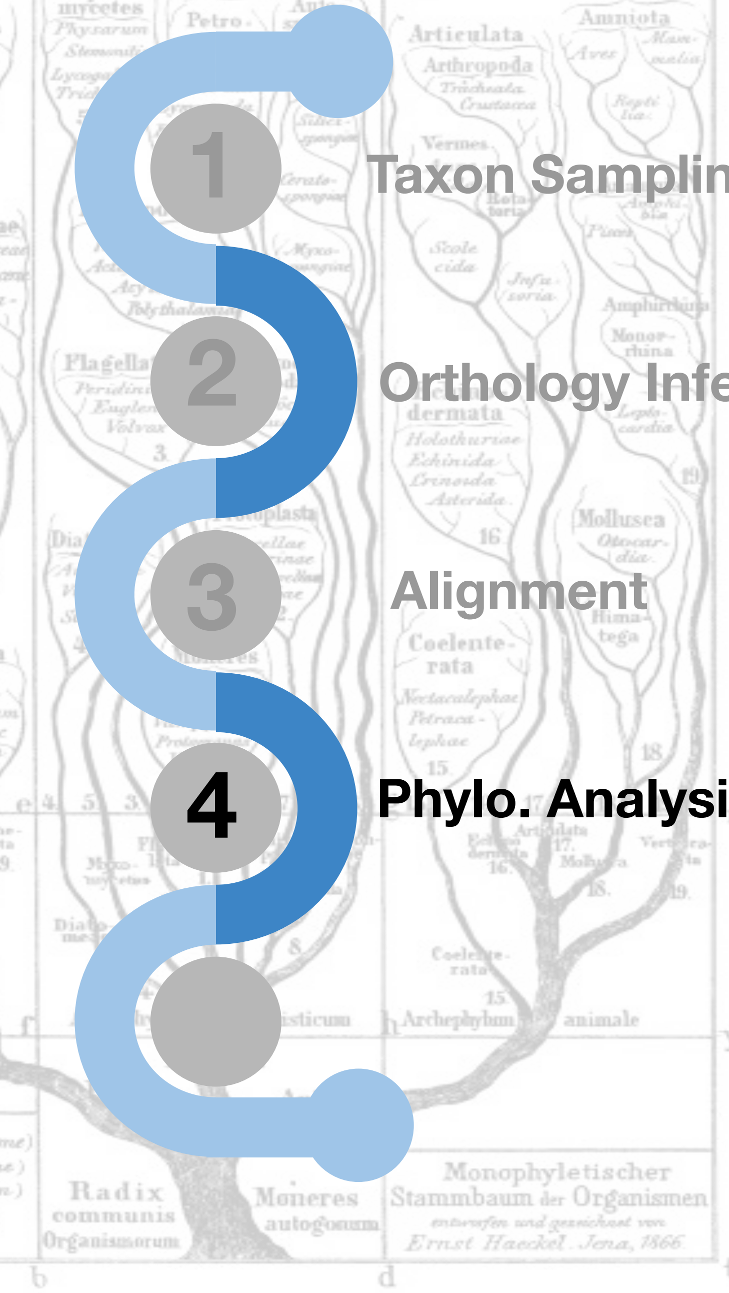


METHOD +



A WAY TO ASSESS HOW GOOD OUR HYPOTHESIS IS

How to infer a species tree



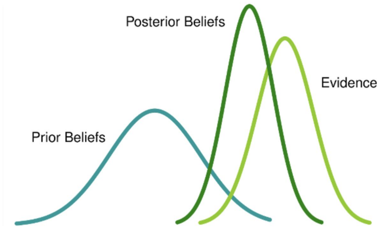
1 Taxon Sampling

2 Orthology Inference

3 Alignment

4 **Phylo. Analysis**

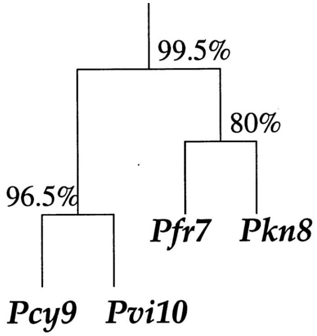
DATA + **MODEL OF EVOLUTION**



+

METHOD

+

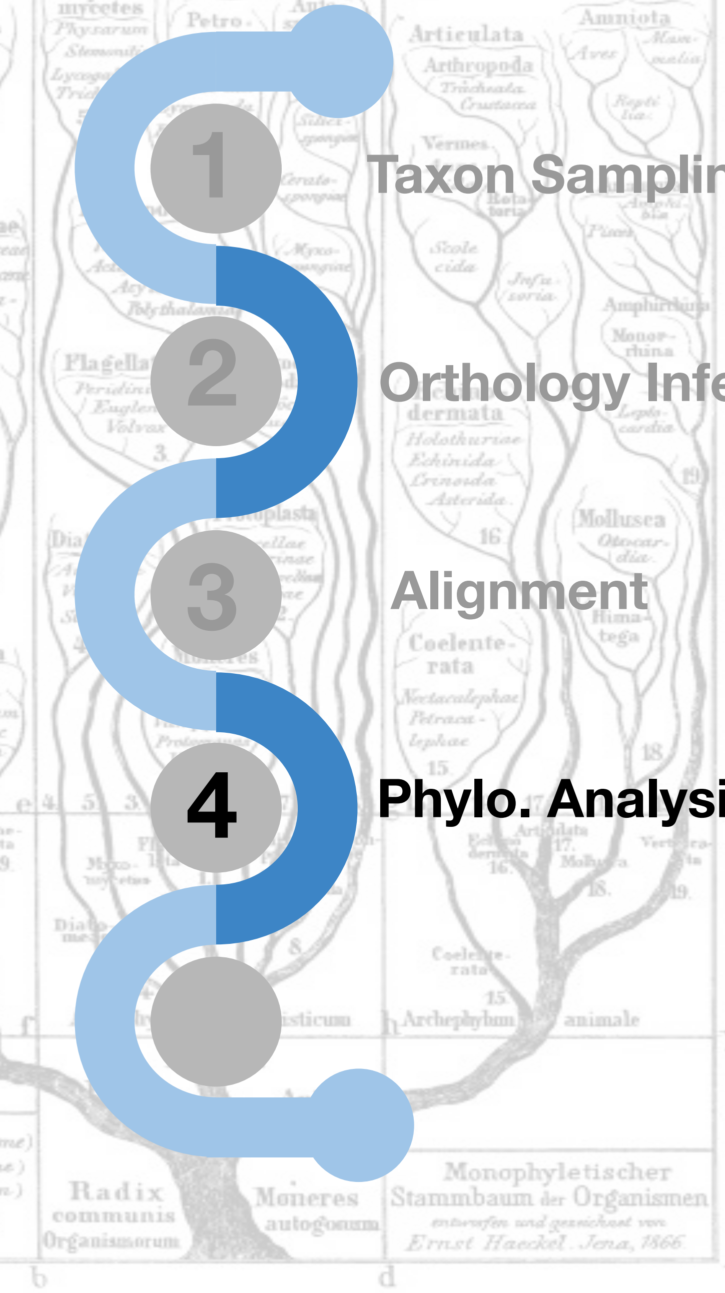


A WAY TO ASSESS HOW GOOD OUR HYPOTHESIS IS

- 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% bootstrap support and/or ultrafast bootstrap > 95% and/or approx. LRT > 80%.

- BI: posterior probability (between 0 and 1)
  - you ‘believe’ in a clade with > 0.9 pp

How to infer a species tree



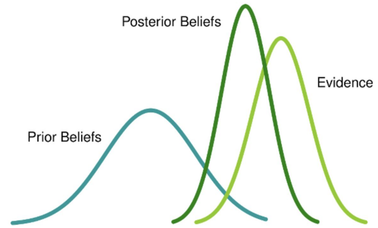
1 Taxon Sampling

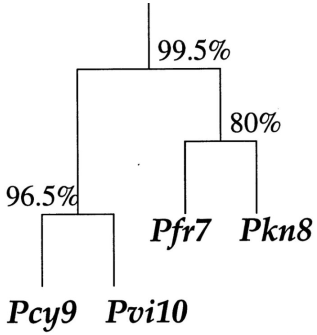
2 Orthology Inference

3 Alignment

4 **Phylo. Analysis**

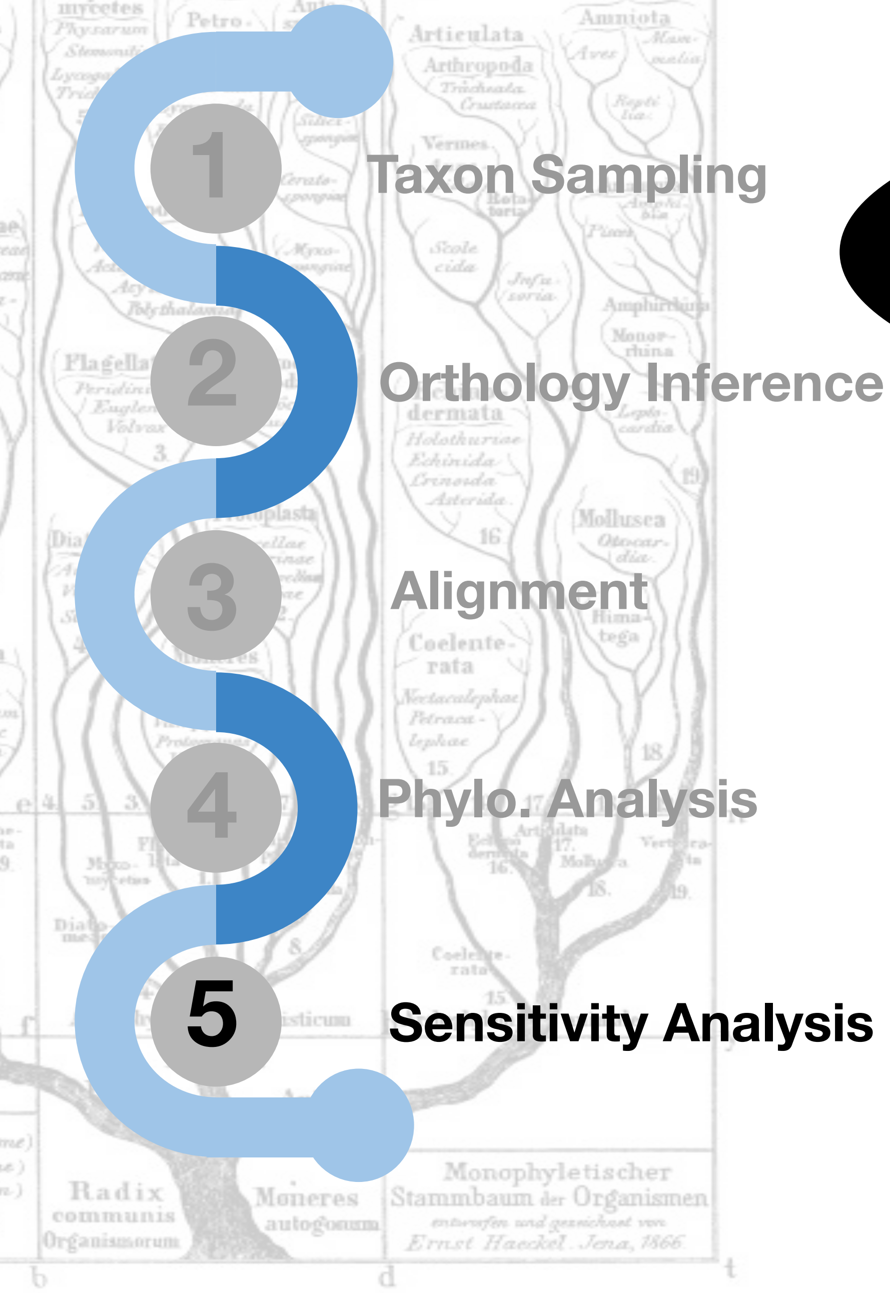
DATA + **MODEL OF EVOLUTION** + **METHOD** + **A WAY TO ASSESS HOW GOOD OUR HYPOTHESIS IS**





- New metrics (highly recommended in large matrices):
  - concordance factor (IQTREE): 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

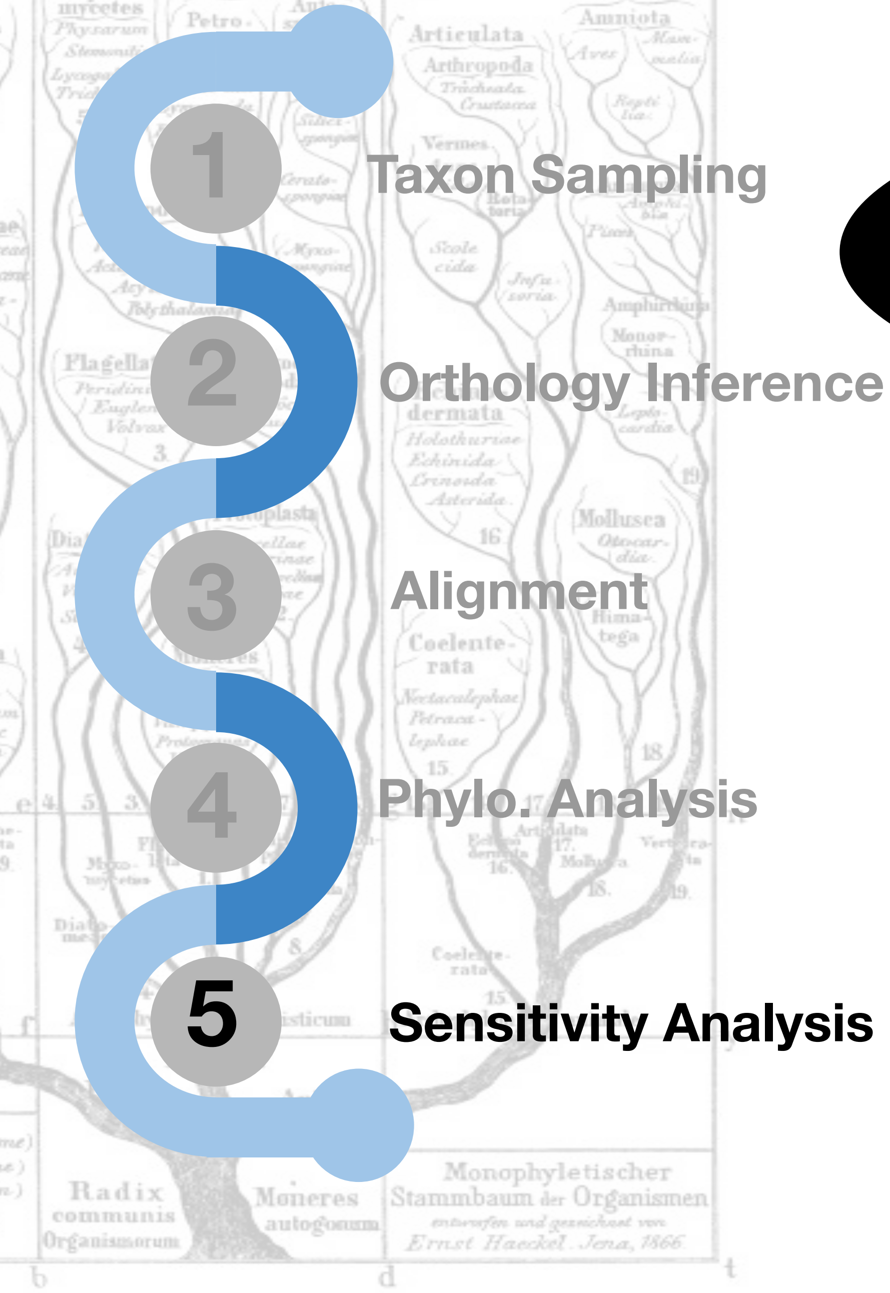
3 Alignment

4 Phylo. Analysis

5 Sensitivity Analysis

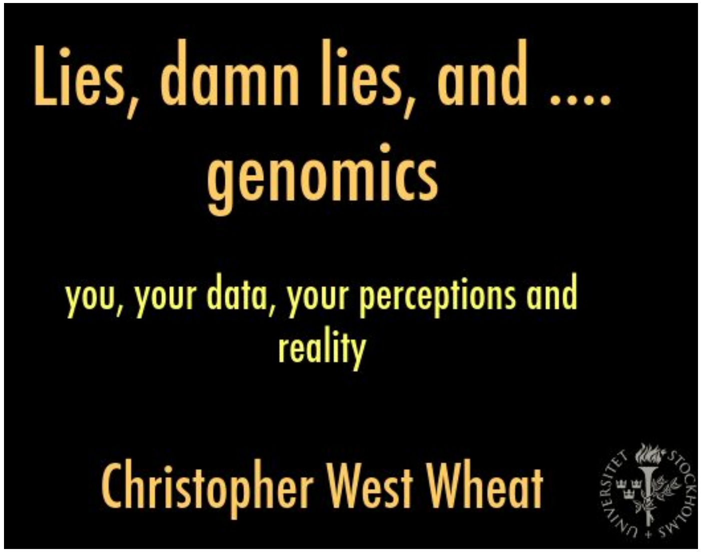
Lies, damn lies and phylogenomics

How to infer a species tree


1
Taxon Sampling
2
Orthology Inference
3
Alignment
4
Phylo. Analysis
5
Sensitivity Analysis
Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena, 1866
Radix communis Organismorum
Moneres autogonum

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)

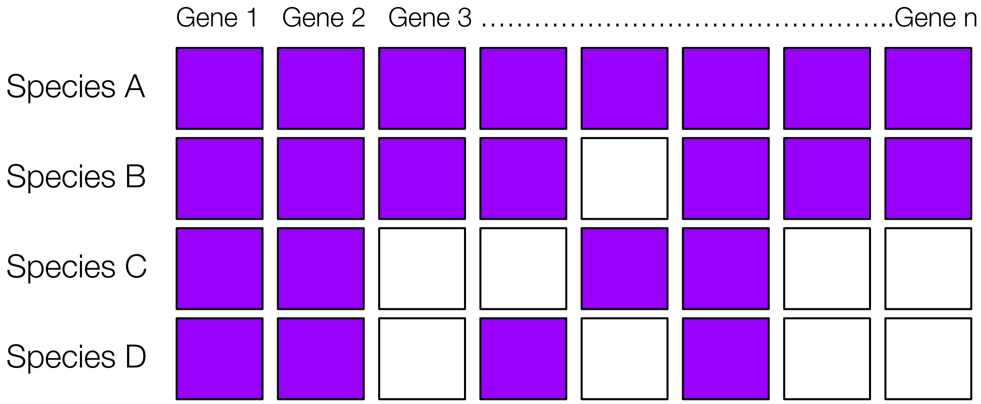
(1) *Intrinsic* properties

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

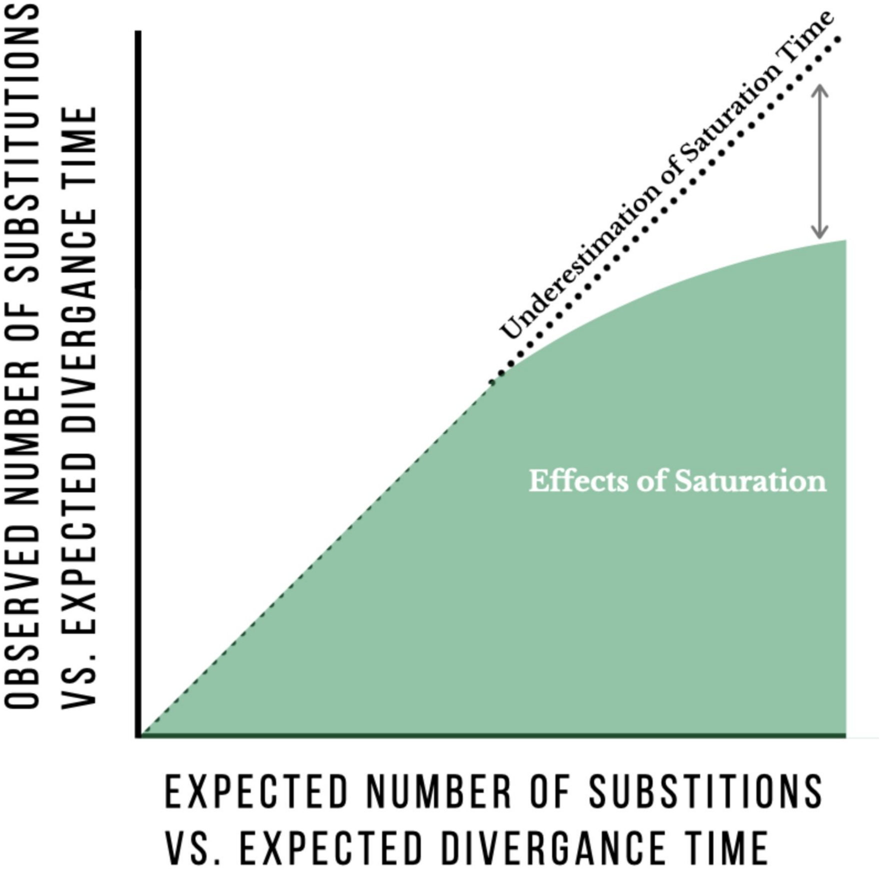
Saturation

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

(1) *Intrinsic* properties

Missing data

Saturation



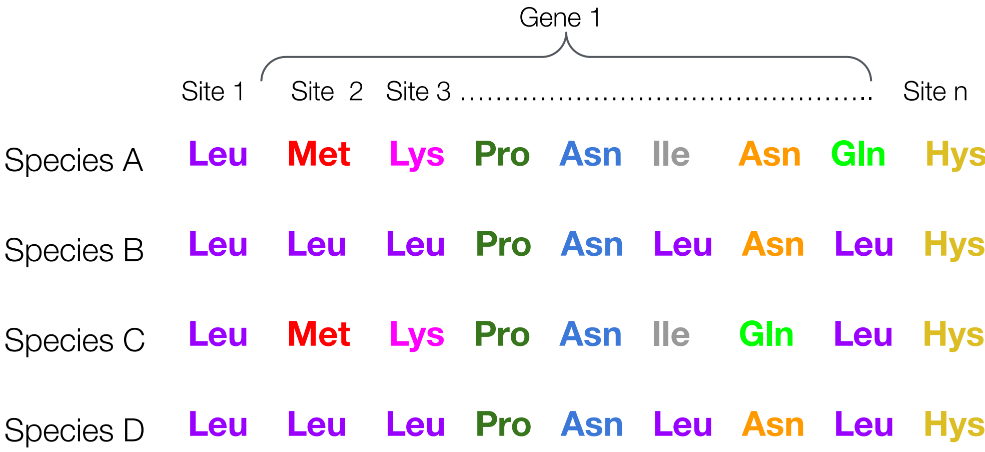
Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity



Gene 1

	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

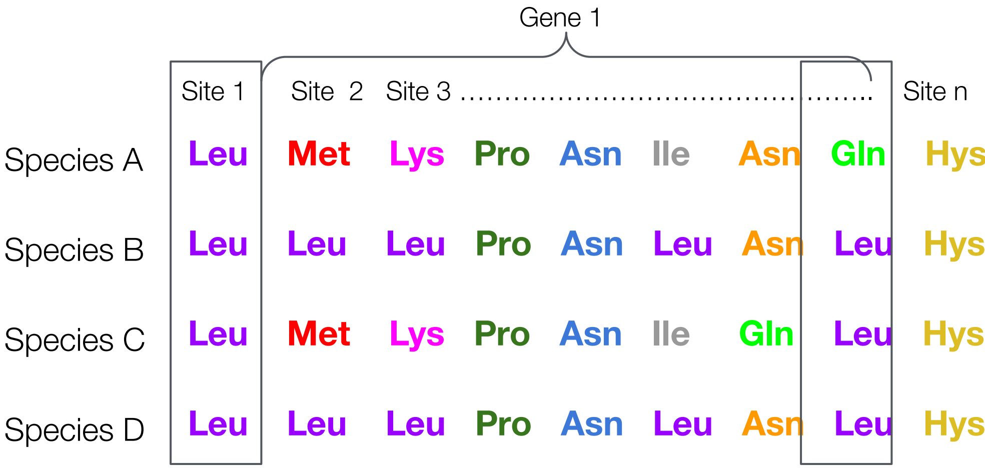
Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity



Gene 1

	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

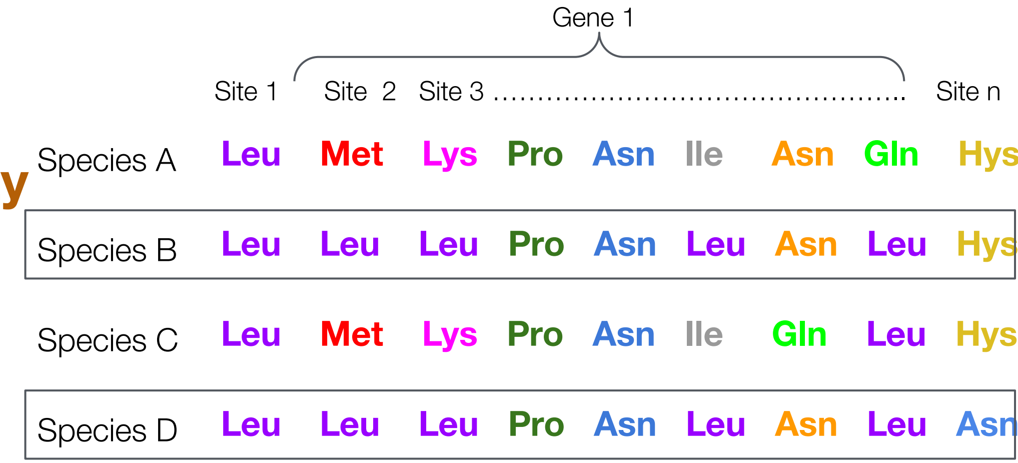
Understanding your data (and the errors it may trigger in downstream analyses)

(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity



Gene 1

	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	Asn

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

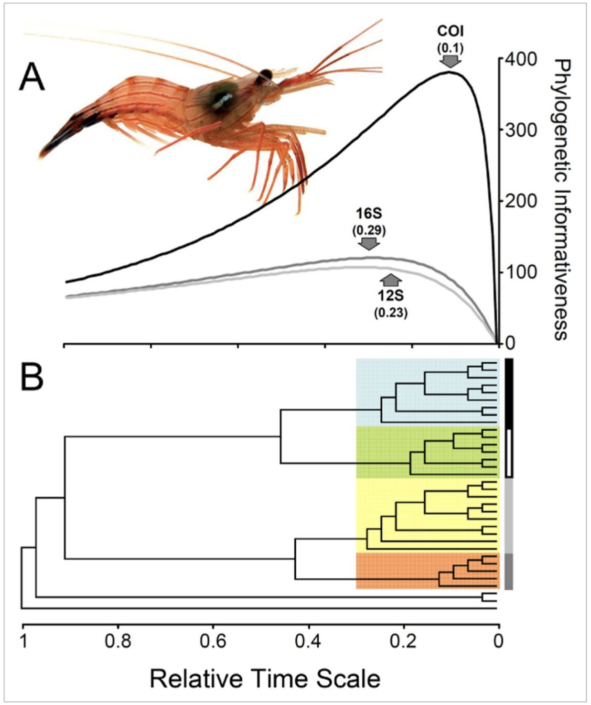
(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest



Baeza & Fuentes 2013

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

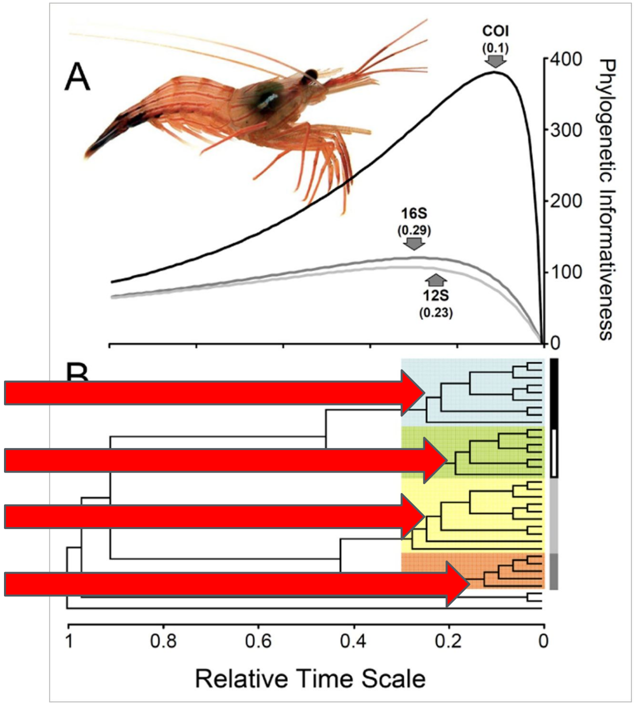
(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest



Good information to resolve these nodes

Baeza & Fuentes 2013

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

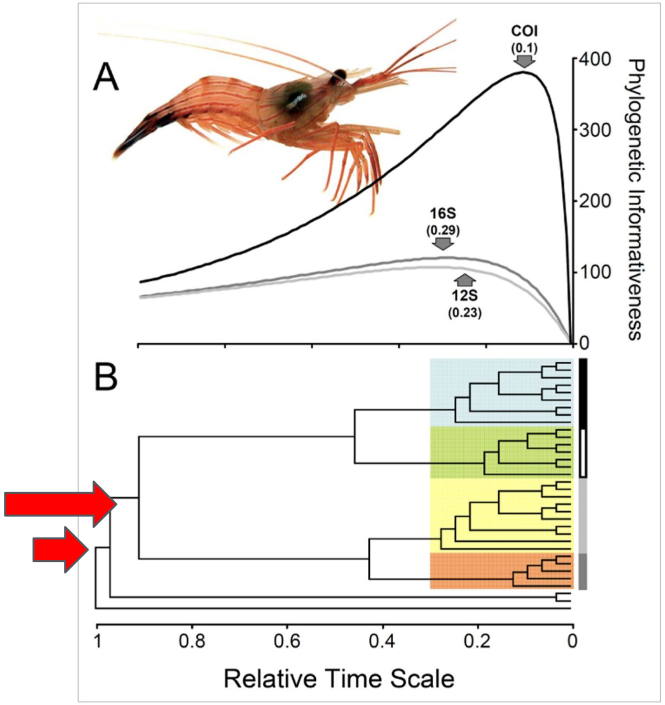
(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest



Not enough information to resolve these nodes

Baeza & Fuentes 2013

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

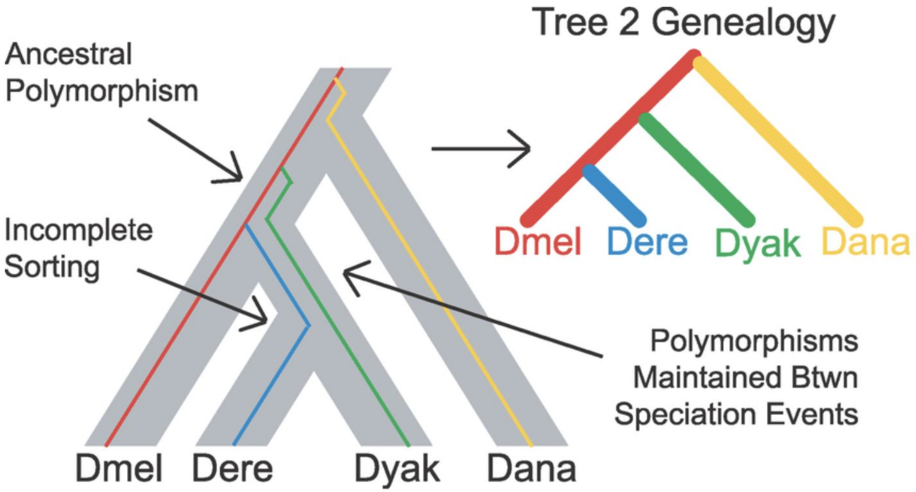
Lack of phylogenetic signal for your node of interest

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

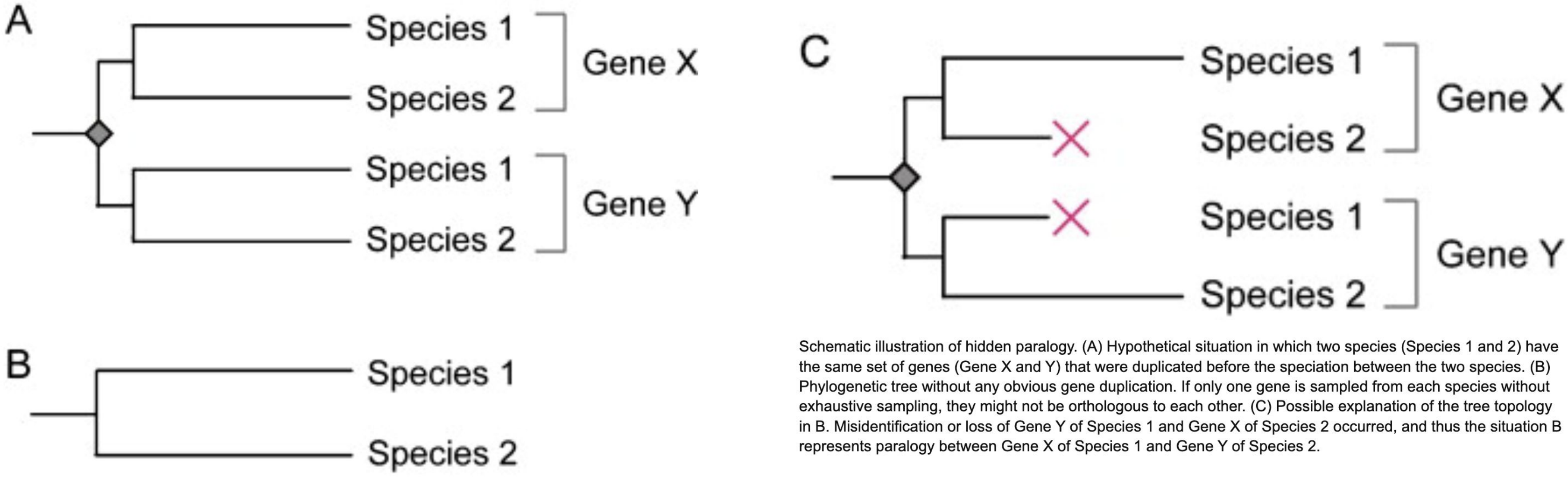
etc.

(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)



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 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)

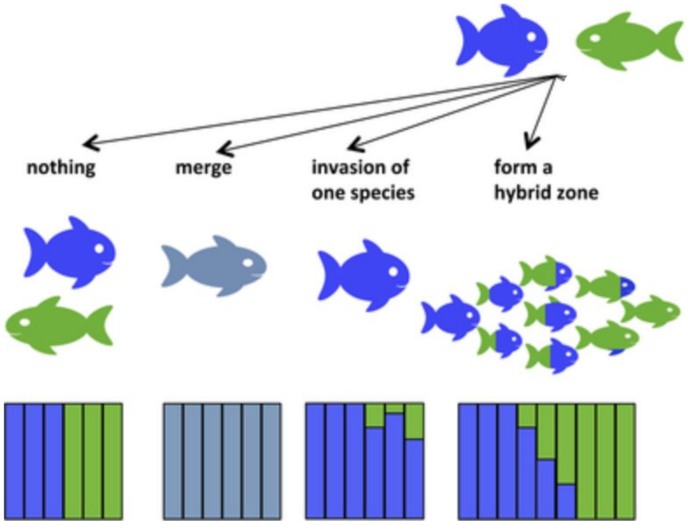


Fig 2.

Schematic representation of homoploid and allopolyploid hybrid speciation.

Runemark et al. 2019

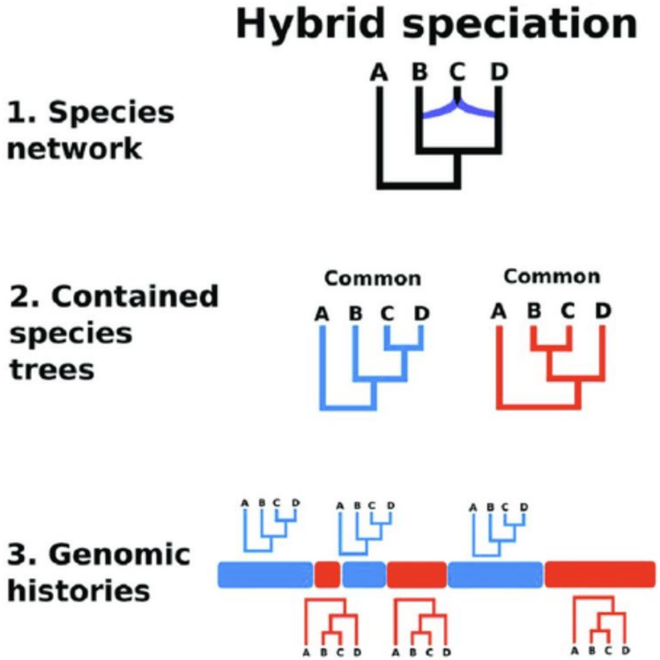
(2) Conflict between individual gene trees and the species tree

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

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



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 ‘real’ species tree

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

Introgression

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

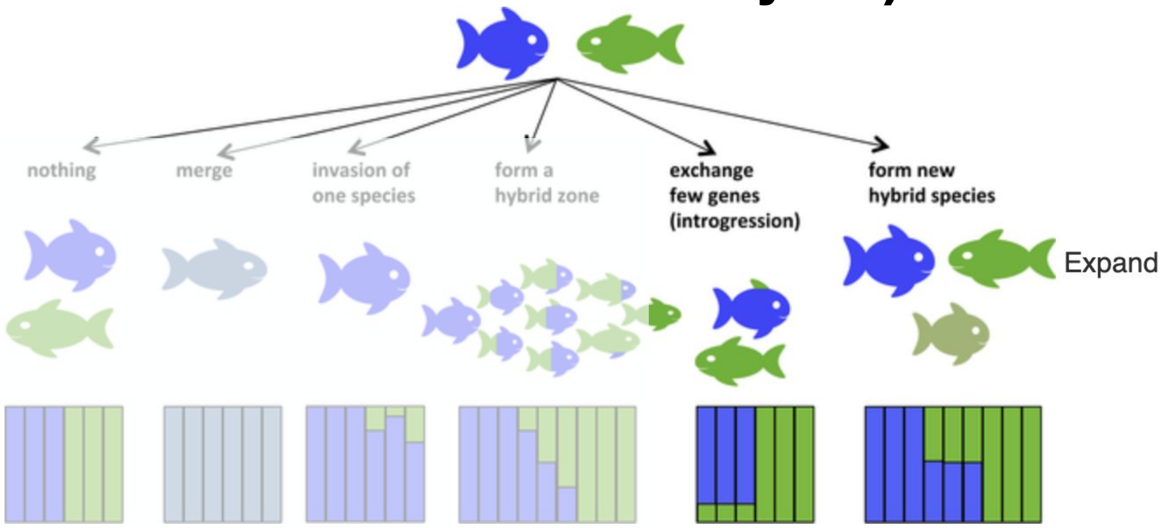


Fig 2.

Schematic representation of homoploid and allopolyploid hybrid speciation.

Runemark et al. 2019

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

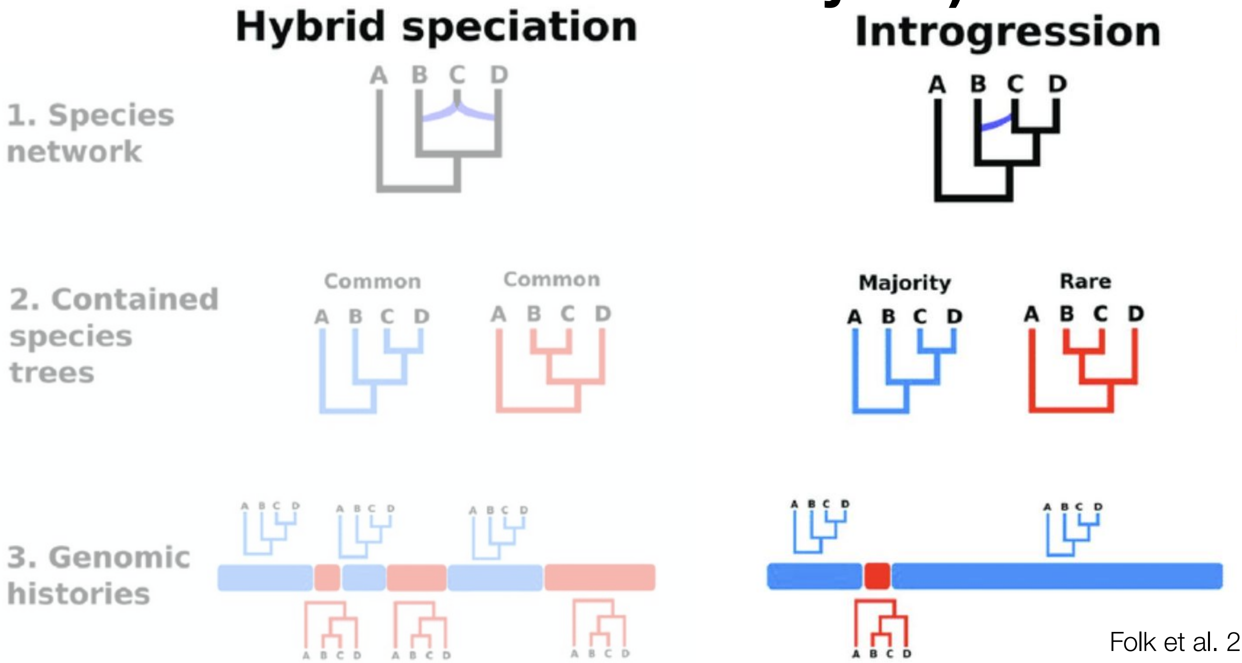
Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

Introgression

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



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

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 ‘real’ species tree

Incomplete lineage sorting

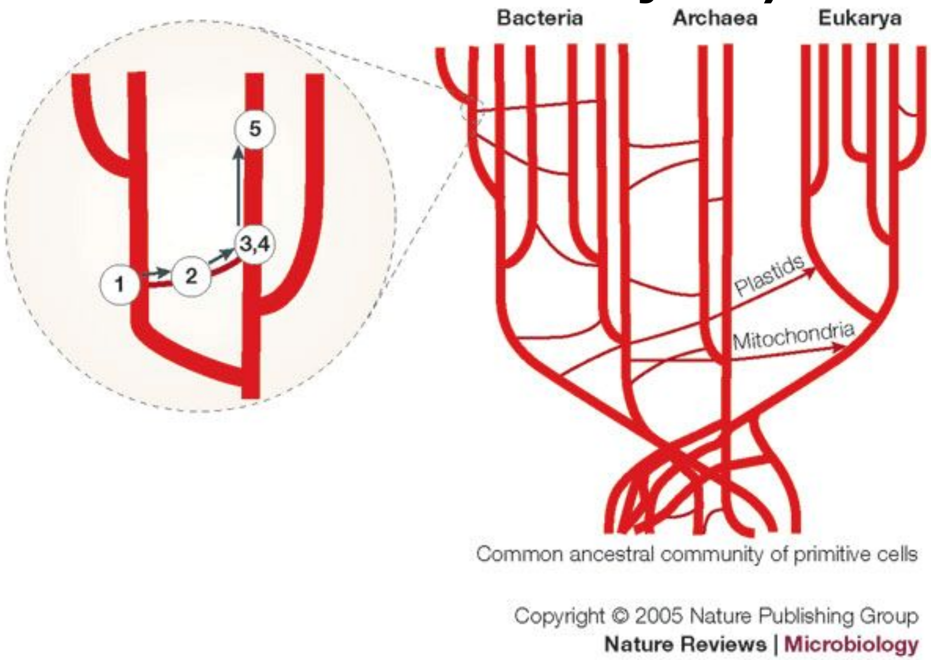
Gene loss (eg, hidden paralogy)

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 'real' 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?

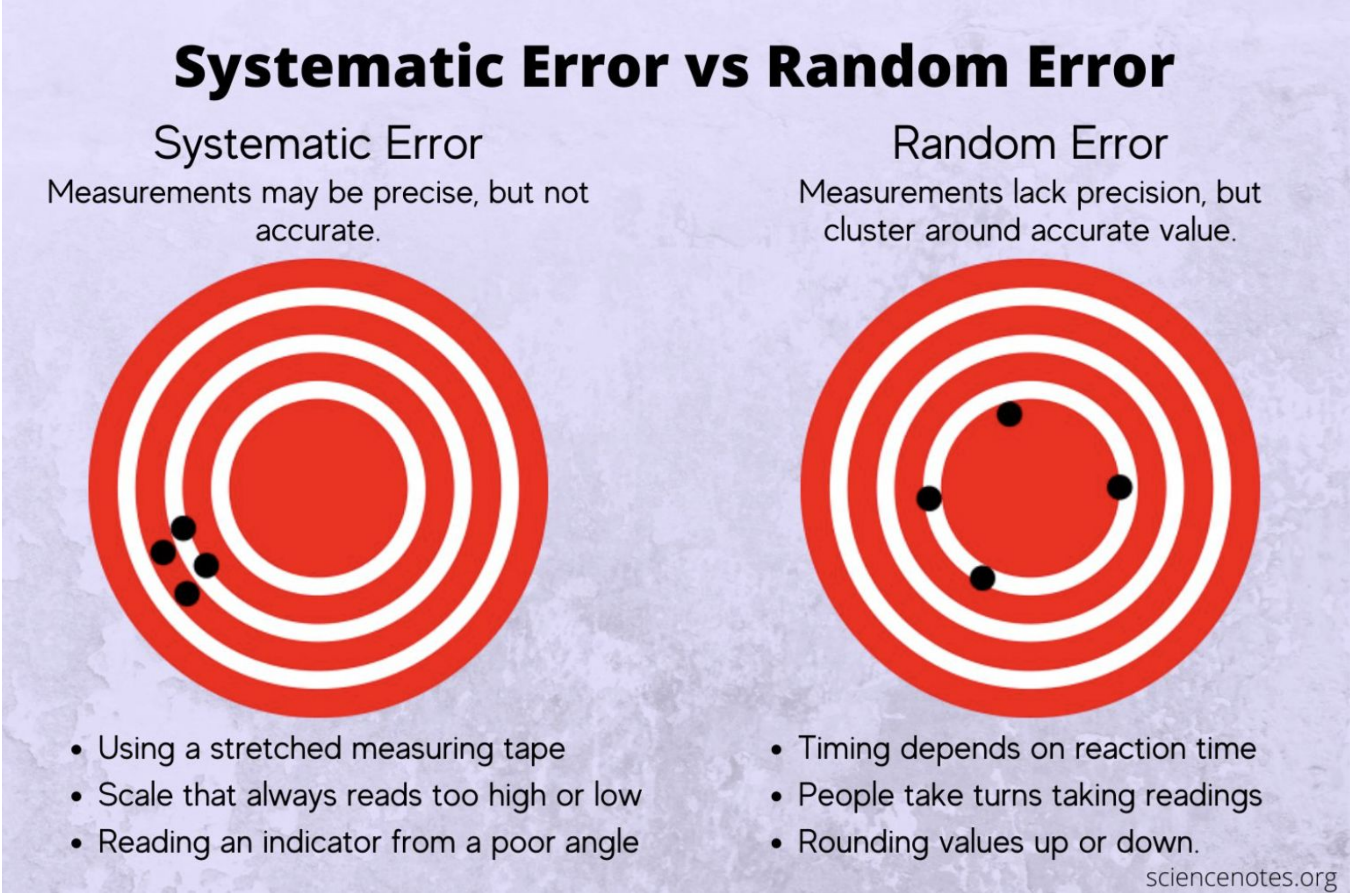
Because of:

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.

- 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.

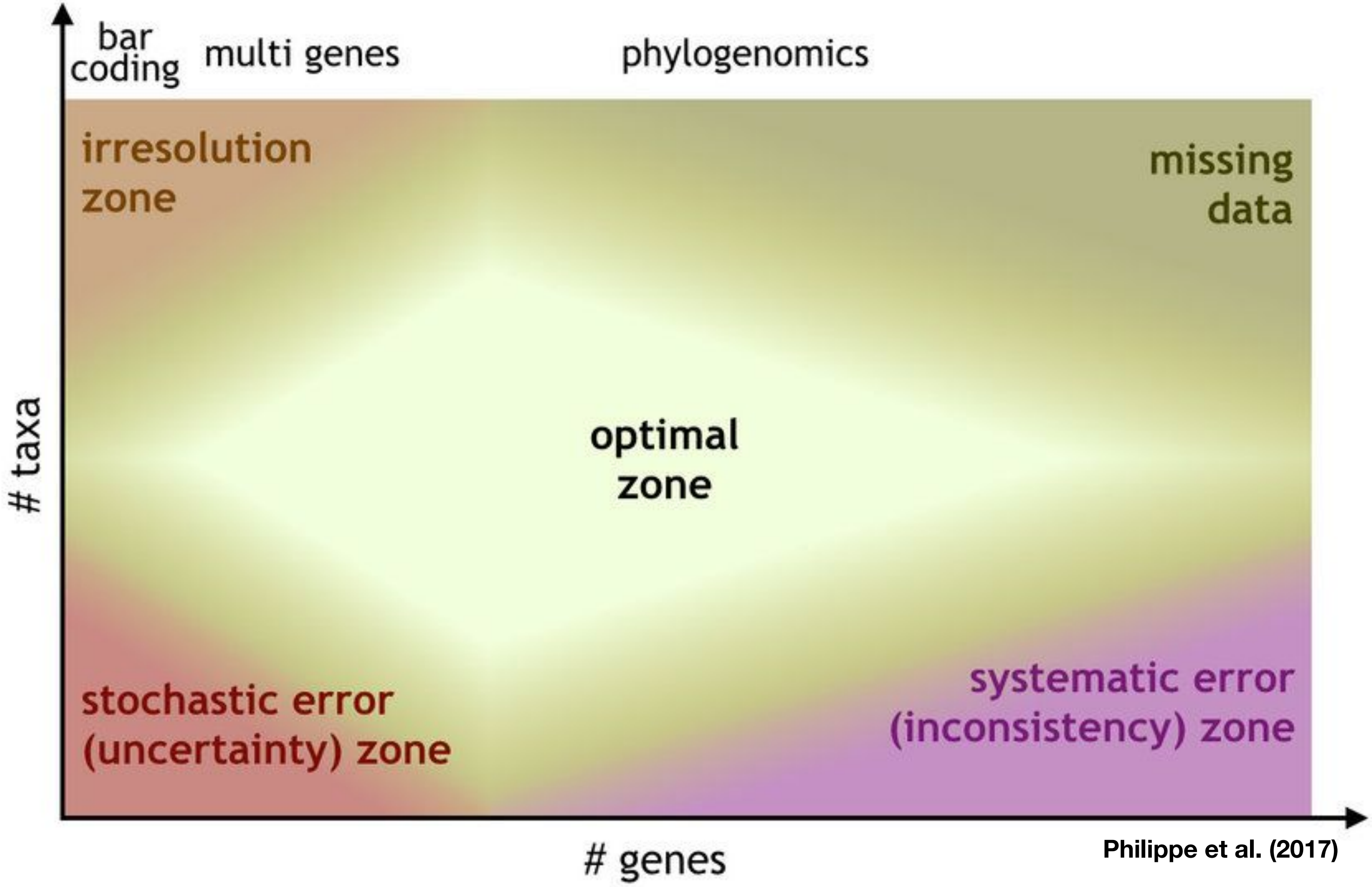
- Timing depends on reaction time
- People take turns taking readings
- Rounding values up or down.

sciencenotes.org

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

Because of:

1) Systematic error



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 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

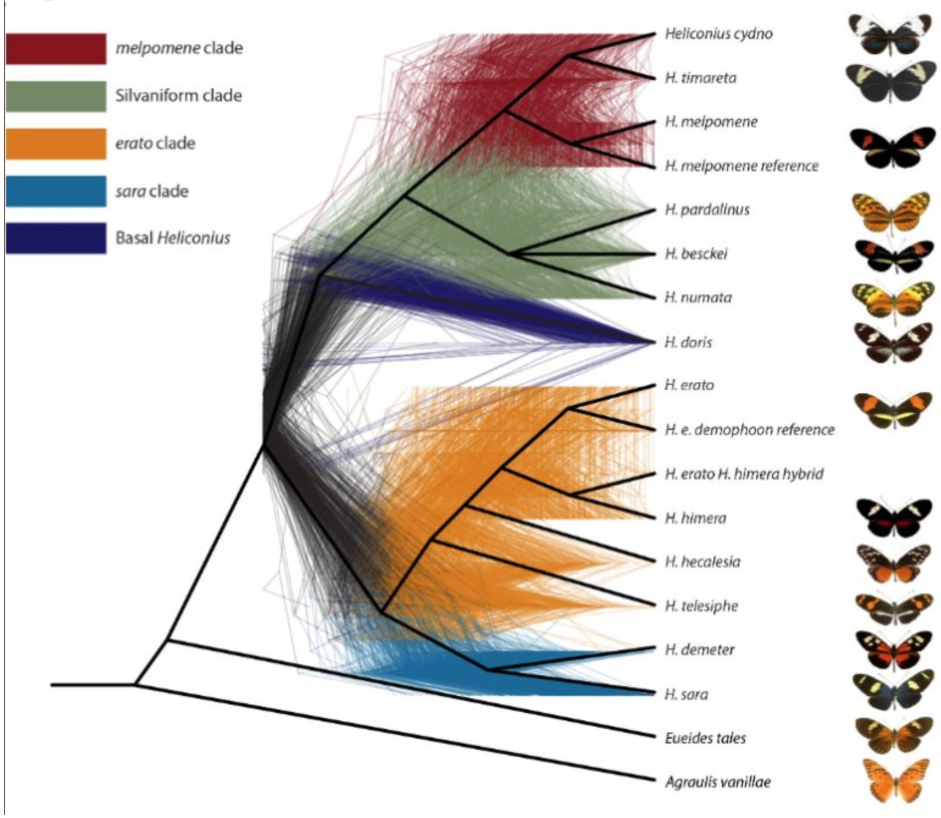
Eg 3, genes evolved through duplication and not through speciation

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**



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 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

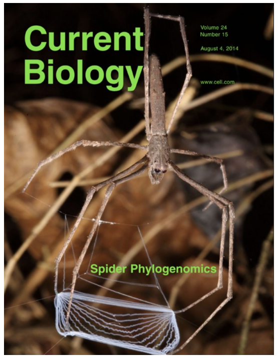
3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*

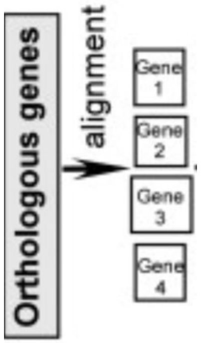
So… what do we do to test the robustness of our tree?

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*





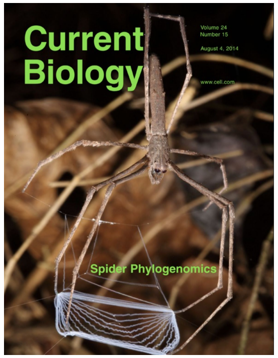
Fernández, Hormiga and Giribet 2014

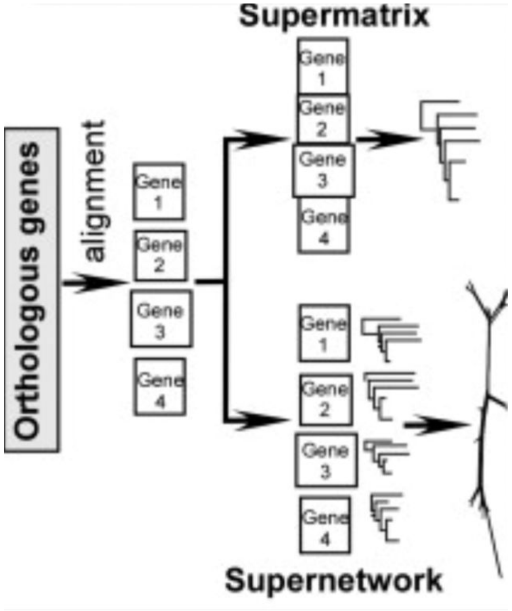
So… what do we do to test the robustness of our tree?

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 and/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)

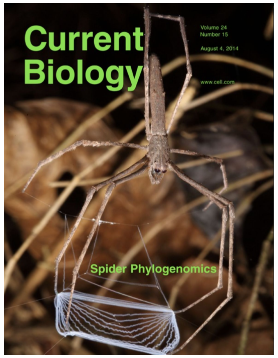
Fernández, Hormiga and Giribet 2014

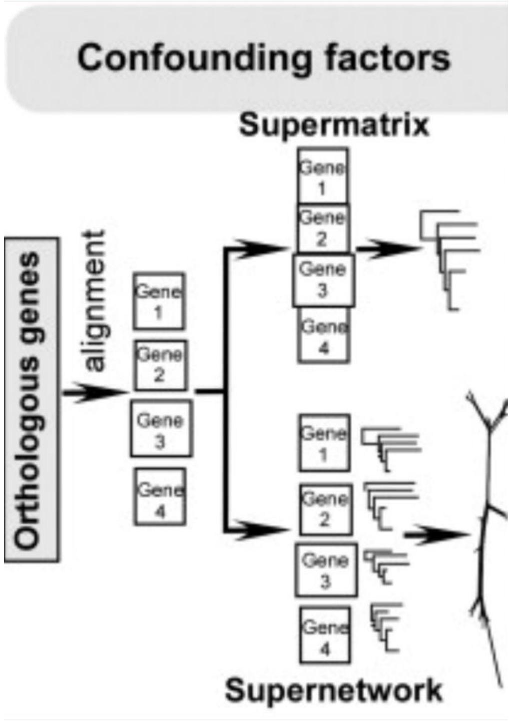
So… what do we do to test the robustness of our tree?

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 and/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)

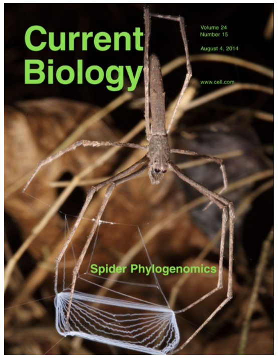
Fernández, Hormiga and Giribet 2014

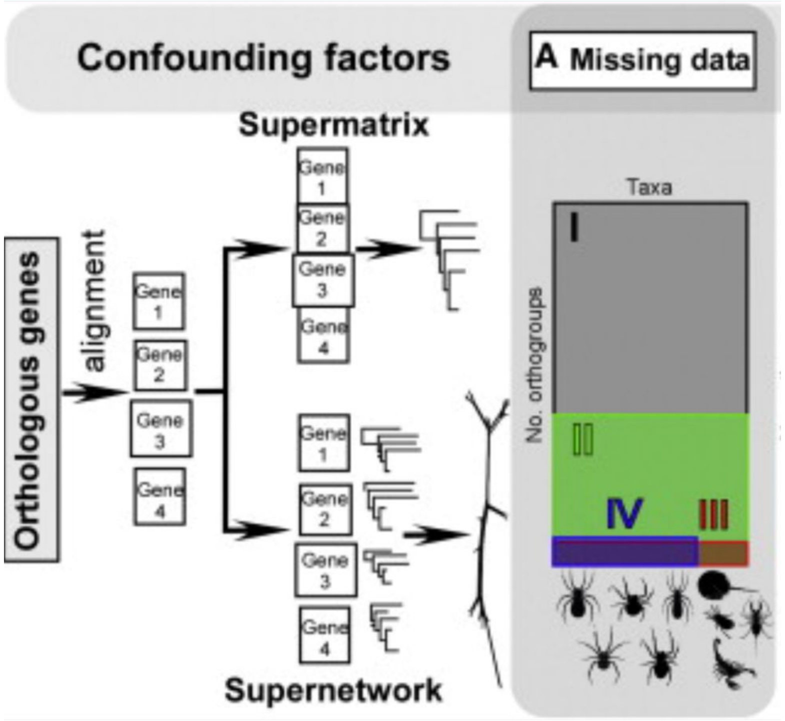
So… what do we do to test the robustness of our tree?

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 and/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)

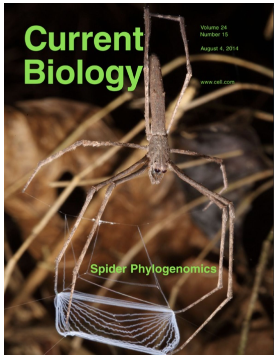
Fernández, Hormiga and Giribet 2014

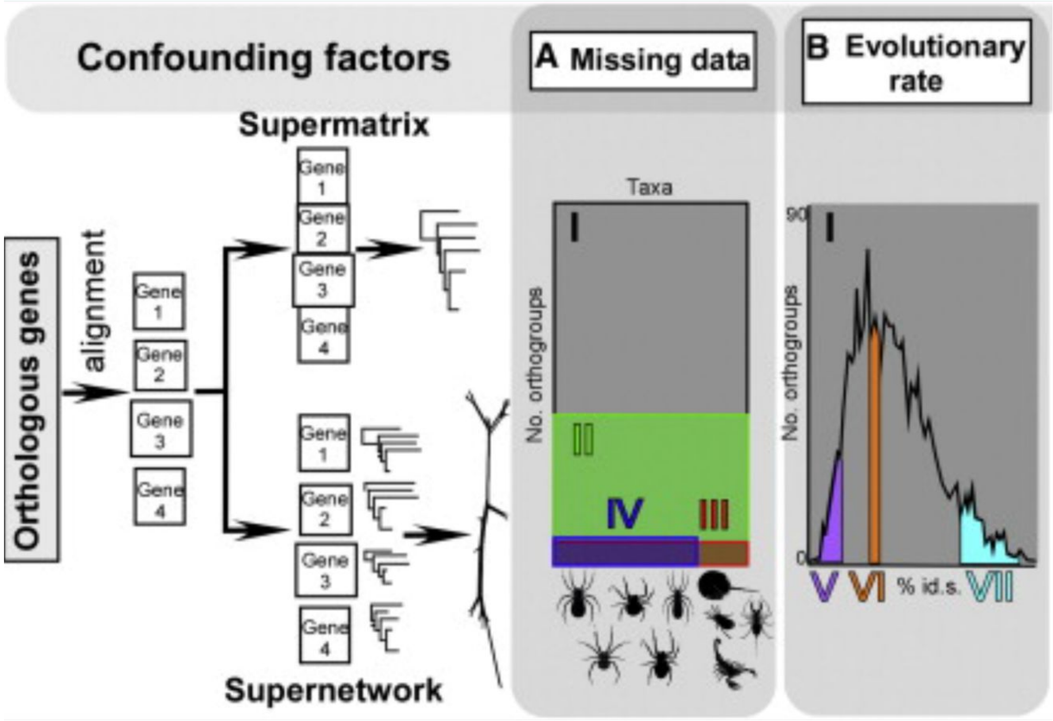
So… what do we do to test the robustness of our tree?

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 and/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)

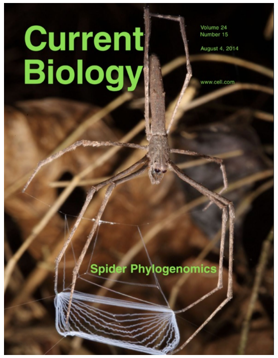
Fernández, Hormiga and Giribet 2014

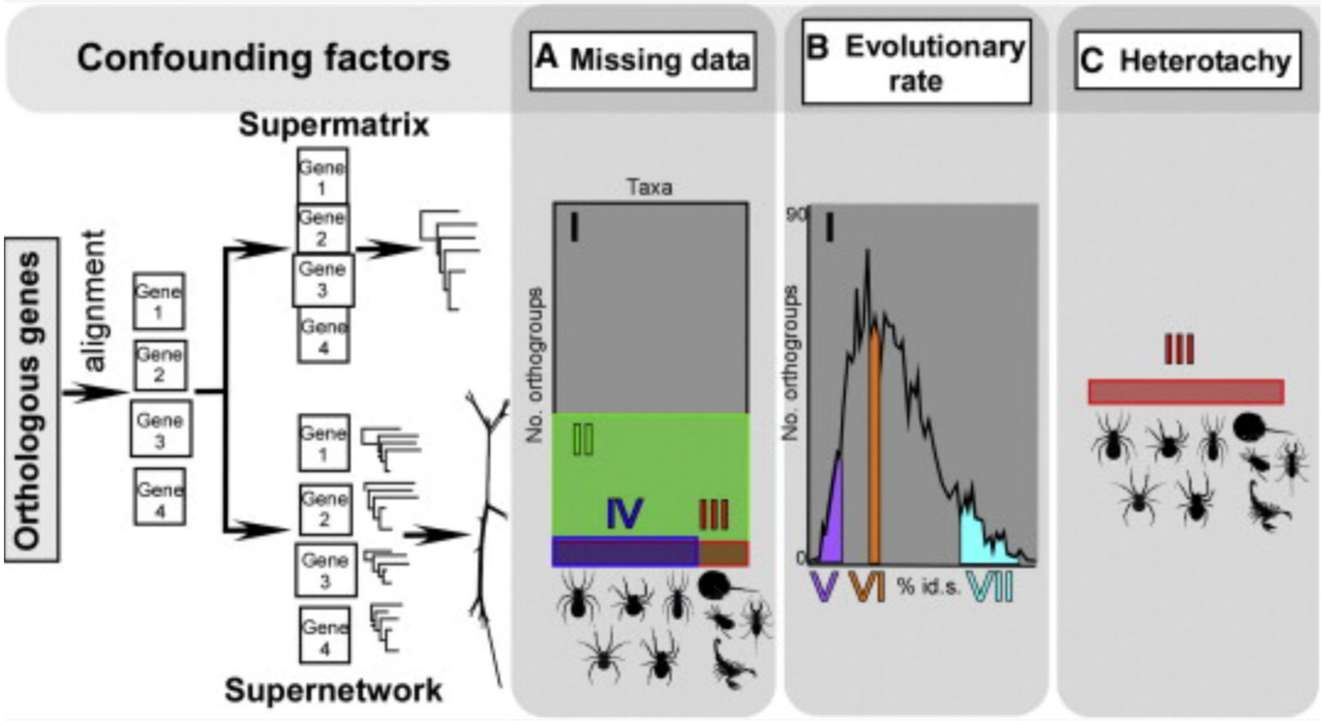
So… what do we do to test the robustness of our tree?

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 and/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

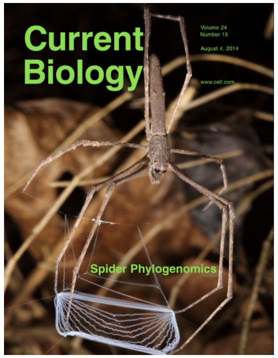
So… what do we do to test the robustness of our tree?

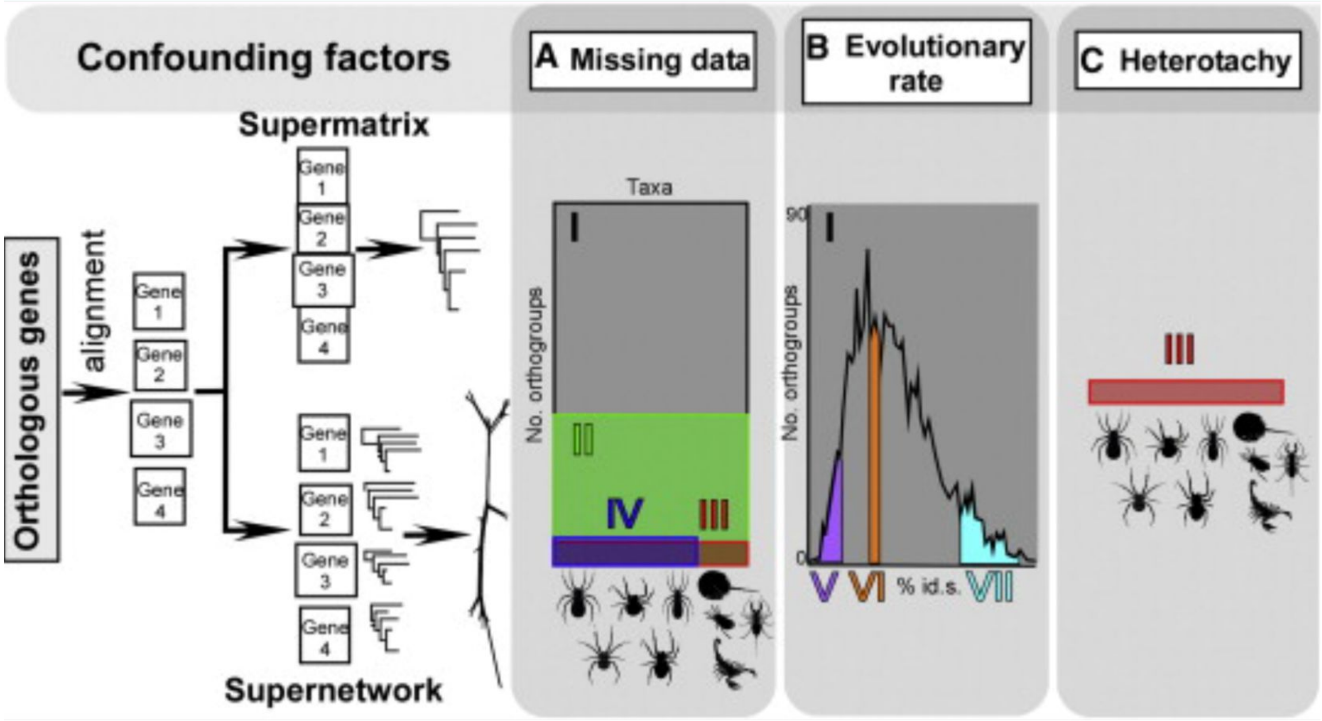
1) Build different subsets of your data through a subs properties

2) Run different analyses that rely on different assum

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)

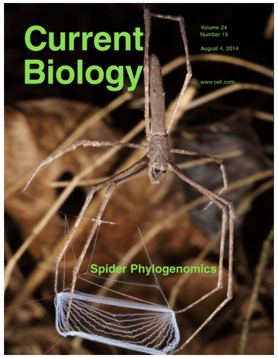
Fernández, Hormiga and Giribet 2014

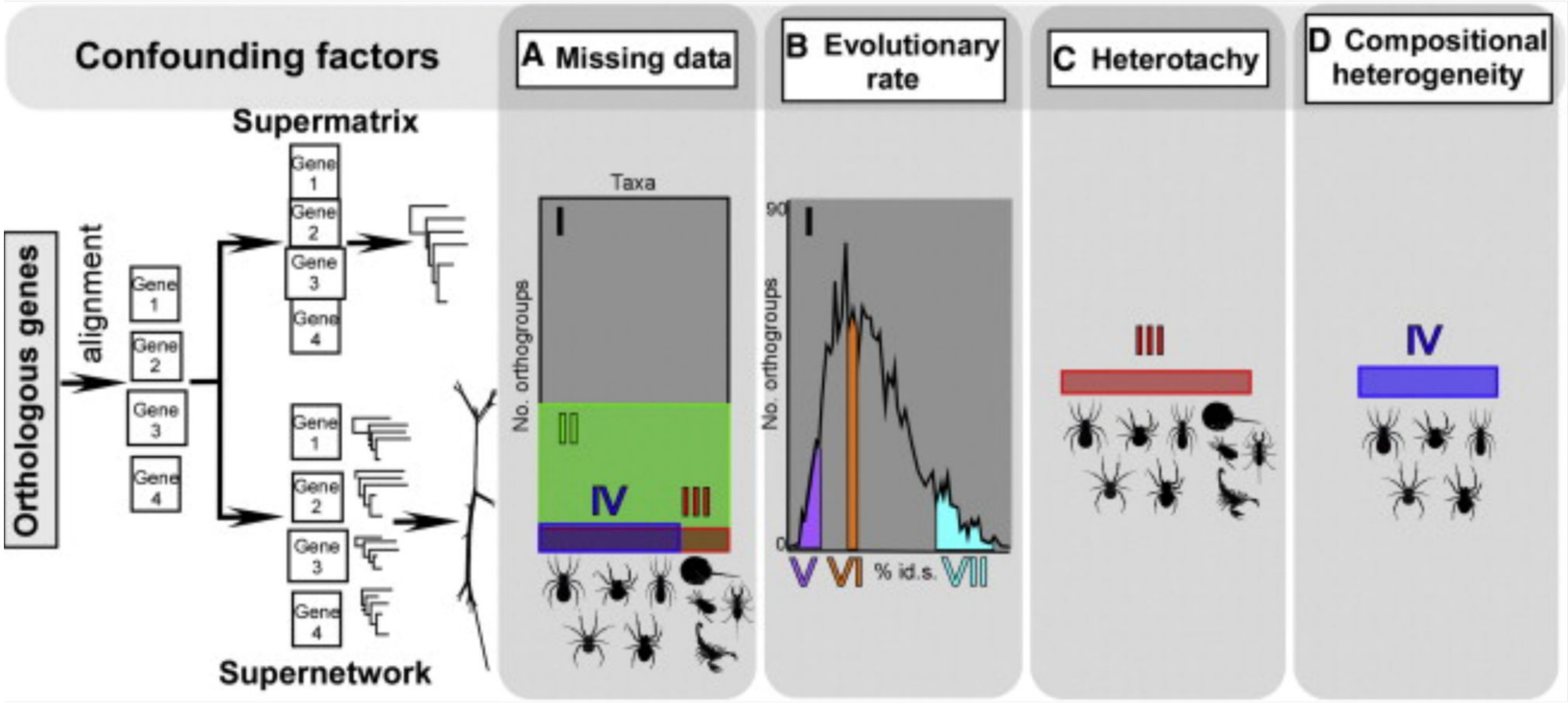
So… what do we do to test the robustness of our tree?

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 and/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***

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 our hands-on session

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

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

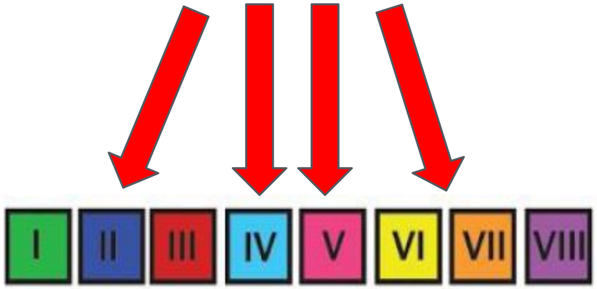
Many.
As many as you can!

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

Fernández, Edgecombe & Giribet (2016) Syst Biol

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

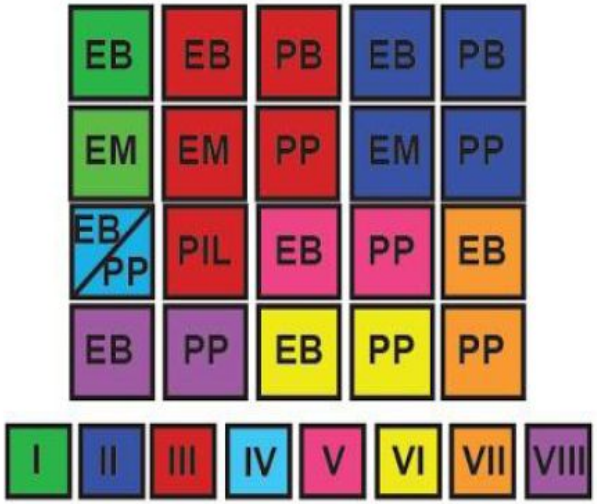
These are **matrices/subsets** of individual gene trees



Fernández, Edgecombe & Giribet (2016) Syst Biol

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

These are **analyses**

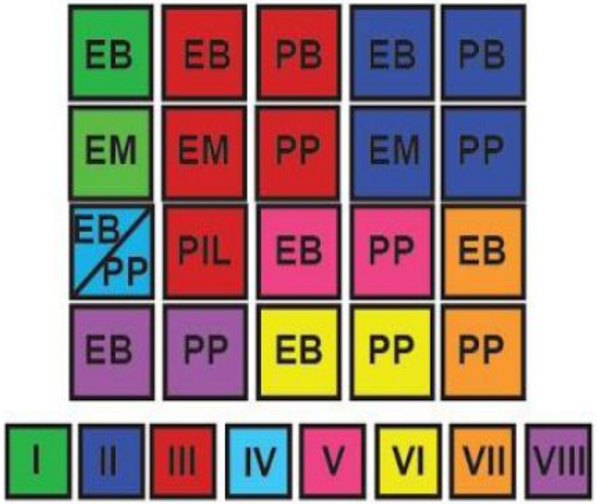


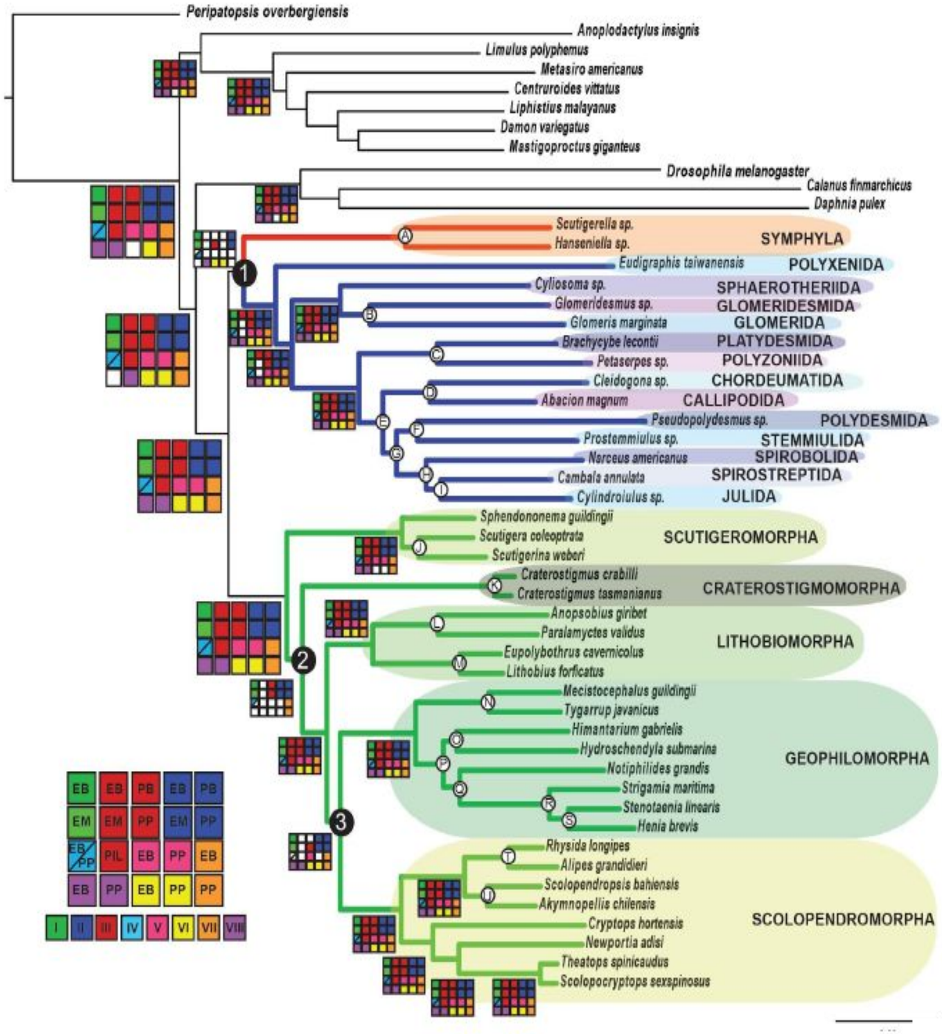
EB	EB	PB	EB	PB
EM	EM	PP	EM	PP
EB/PP	PIL	EB	PP	EB
EB	PP	EB	PP	PP

I II III IV V VI VII VIII

Fernández, Edgecombe & Giribet (2016) Syst Biol

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





Fernández, Edgecombe & Giribet (2016) Syst Biol