INTRODUCTION TO PHYLOGENOMICS

Rosa Fernández
Institute of Evolutionary Biology (CSIC-UPF)

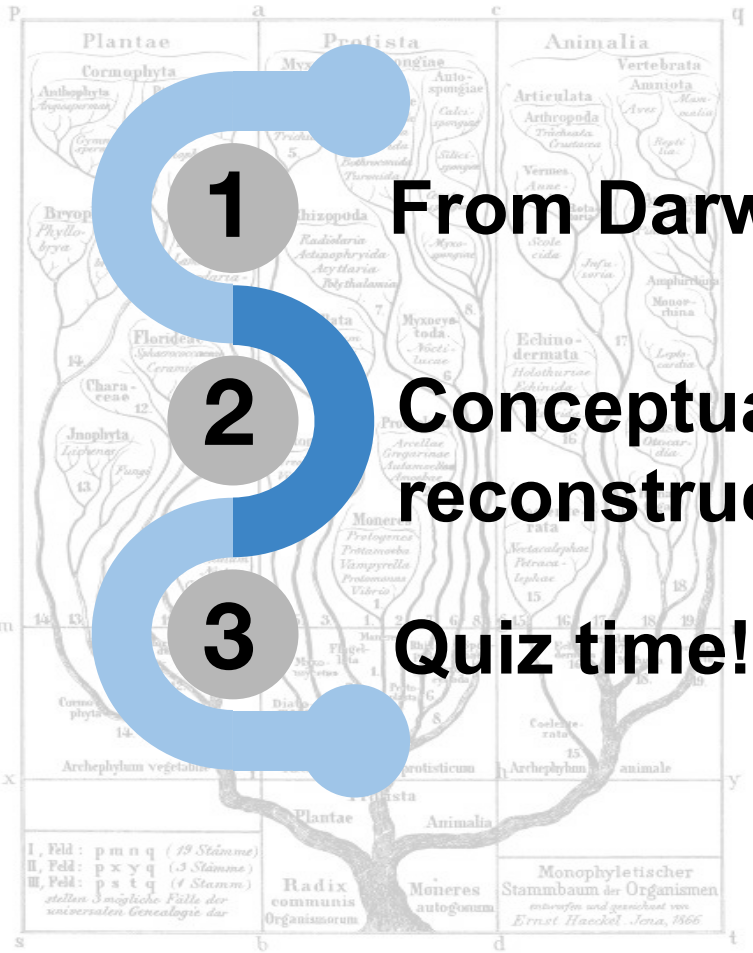
rosa.fernandez@ibe.upf-csic.es

www.metazomics.com

INSTITUT de BIOLOGIA EVOLUTIVA iBe CSIC upf.

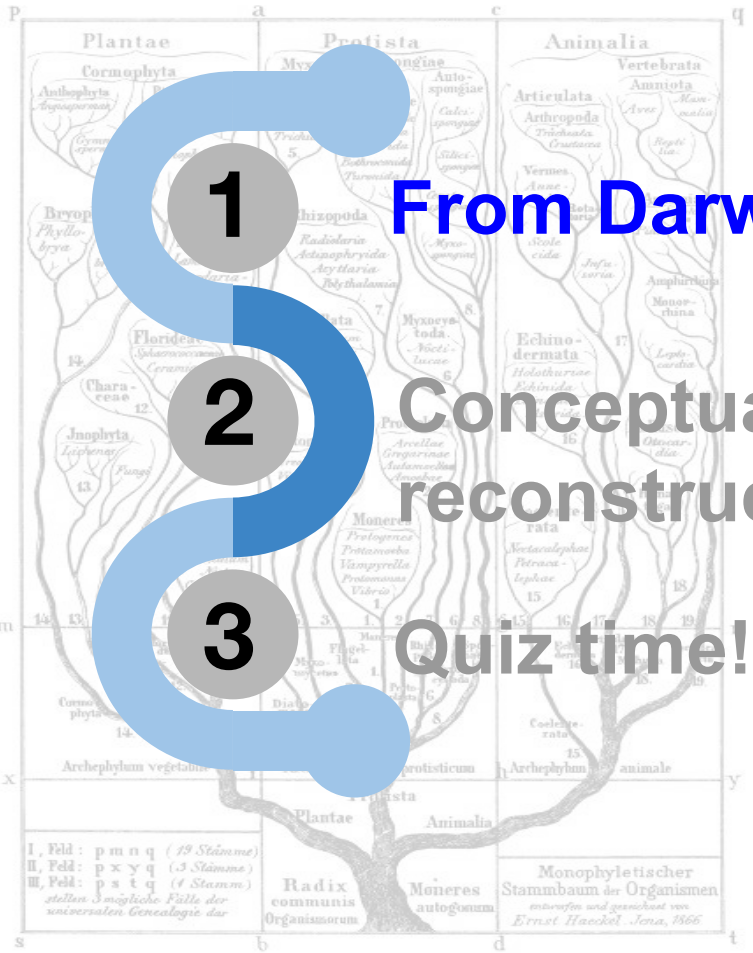
Metazoa Phylogenomics Lab

Content of the lecture



1. From Darwin to phylogenomics
2. Conceptual framework for phylogenomic reconstruction
3. Quiz time!

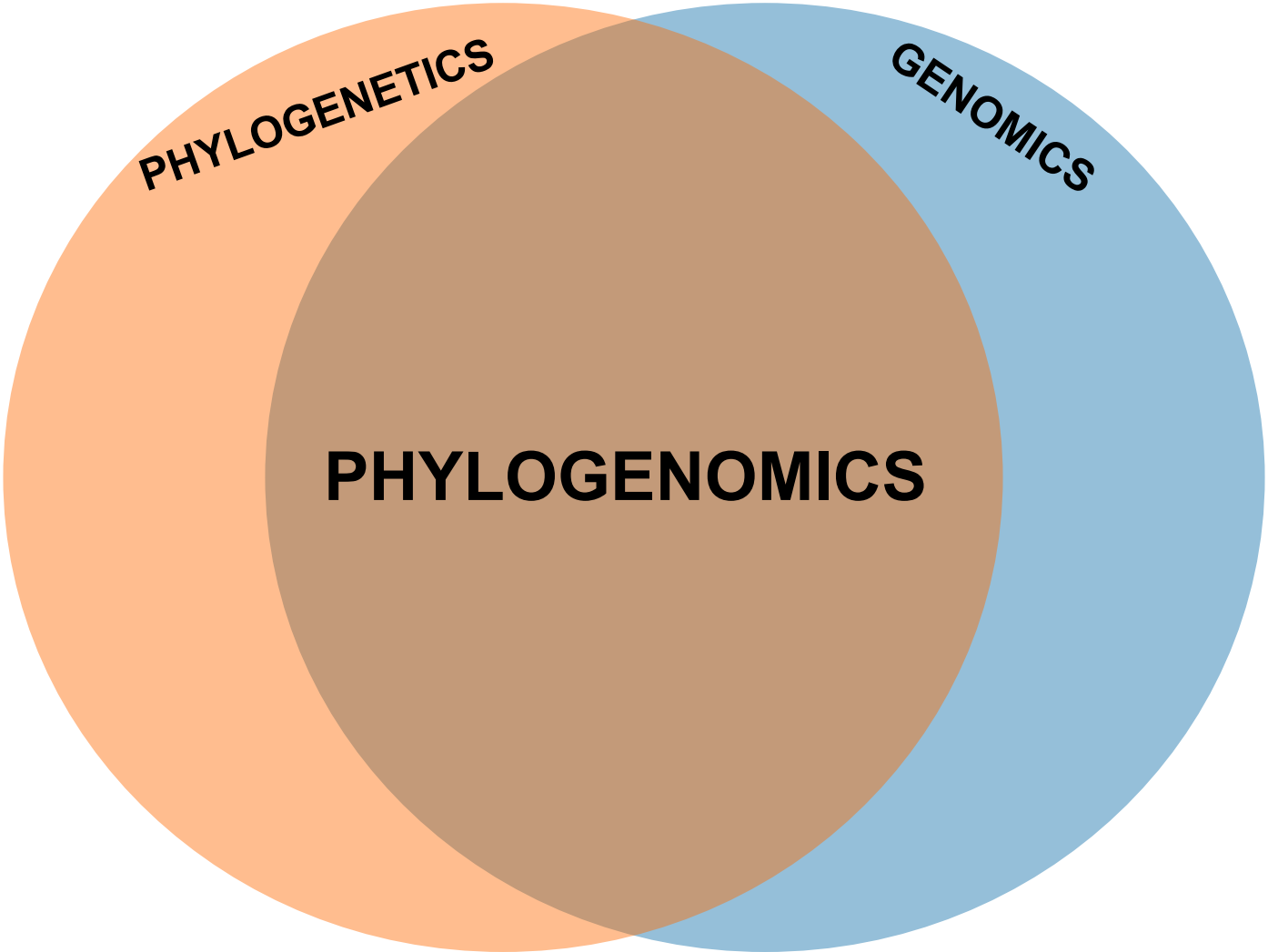
Content of the lecture

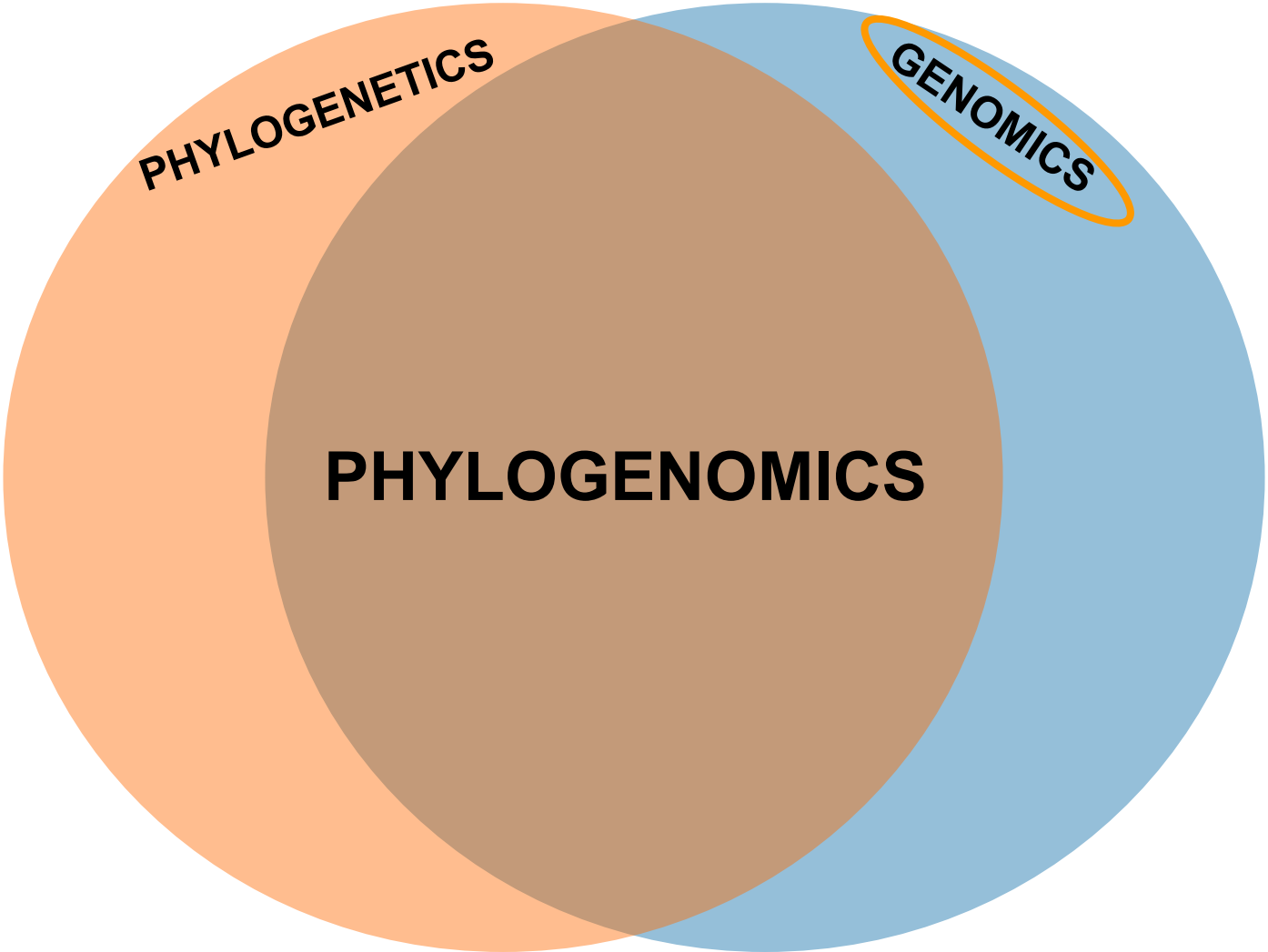


1 **From Darwin to phylogenomics**

2 Conceptual framework for phylogenomic reconstruction

3 Quiz time!


PHYLOGENETICS
GENOMICS
PHYLOGENOMICS


PHYLOGENETICS
GENOMICS
PHYLOGENOMICS

Genomics

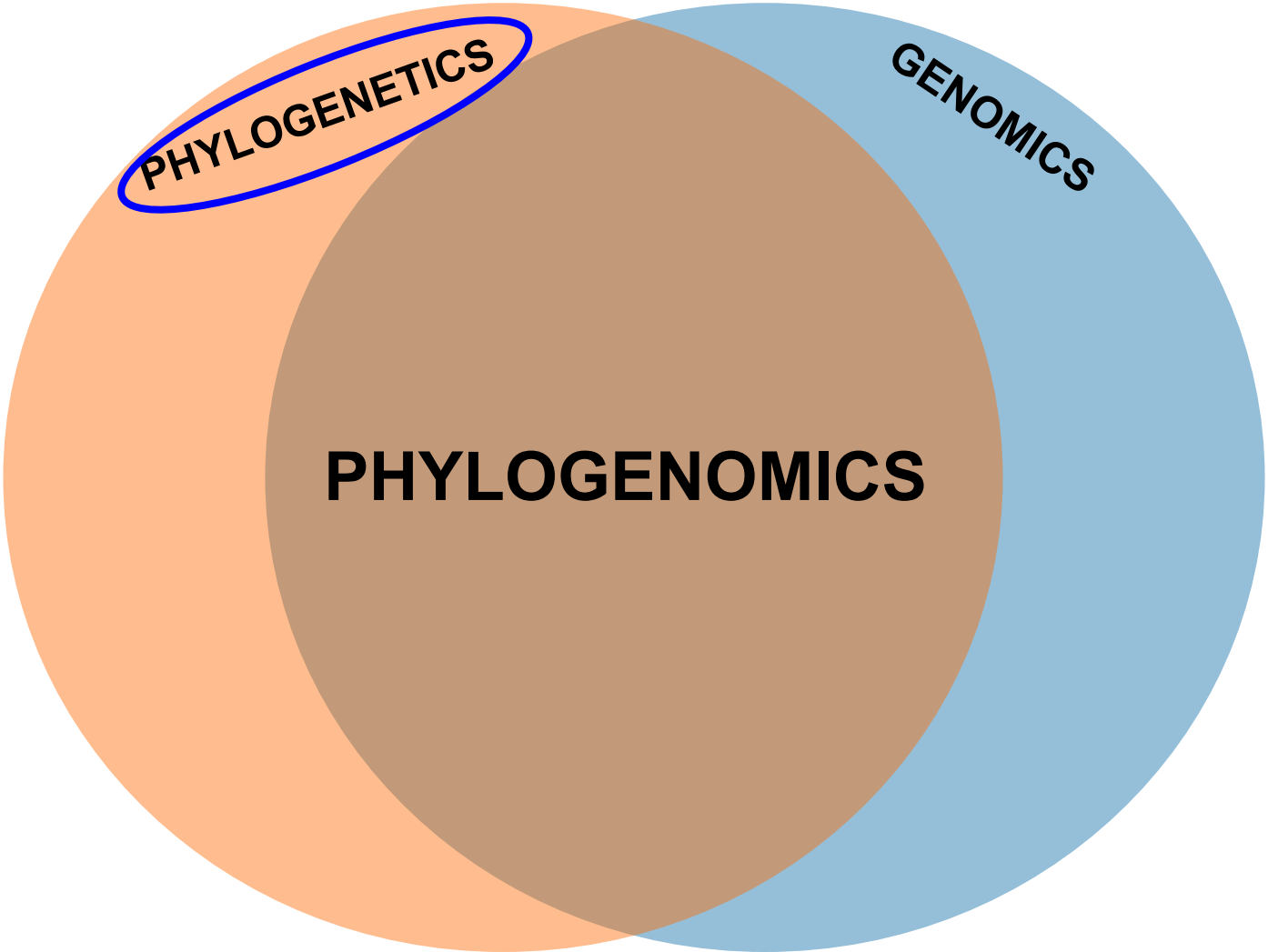
- The study of an organism's complete set of genetic information.
- The genome includes both genes (coding) and non-coding DNA.
- 'Genome': the complete genetic information of an organism.

vs

Genetics

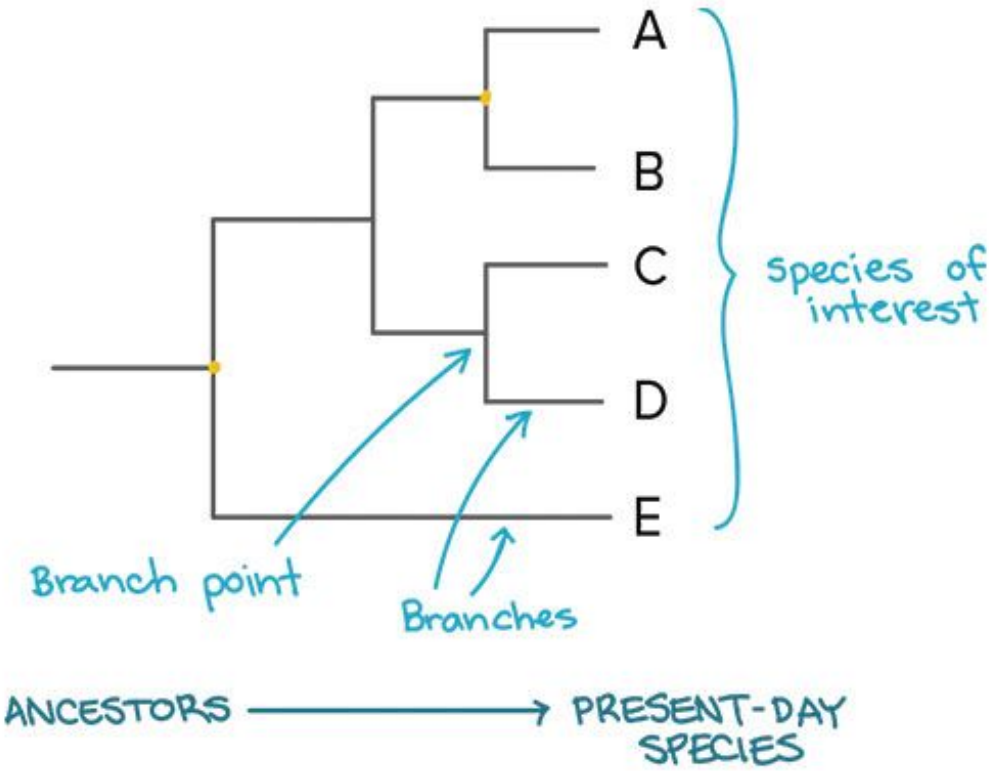
- The study of heredity
- The study of the function and composition of single genes.
- 'Gene': specific sequence of DNA that codes for a functional molecule.

https://www.genomicseducation.hee.nhs.uk/education/core-concepts/what-is-genomics/


PHYLOGENETICS
GENOMICS
PHYLOGENOMICS

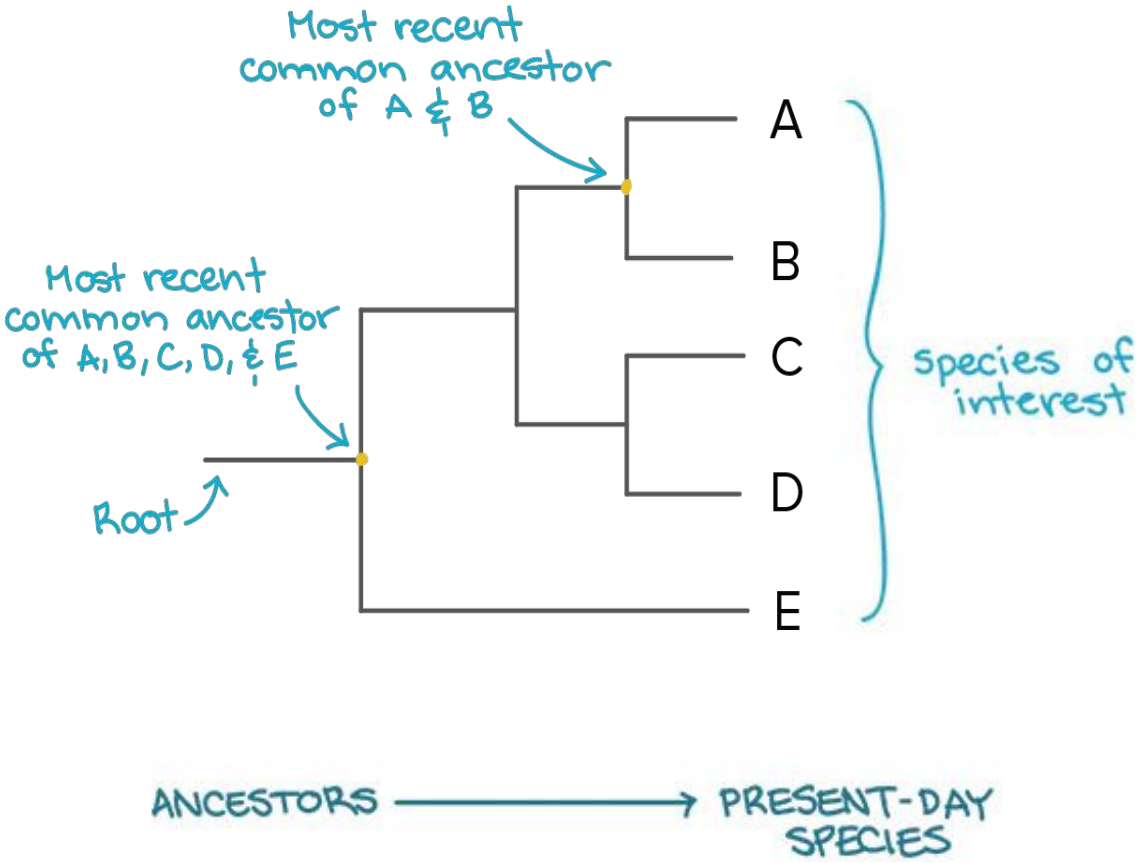
What is a phylogeny…?

What is a phylogeny…?



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

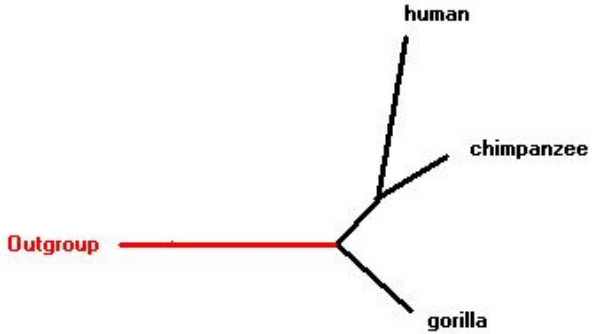
What is a phylogeny…?



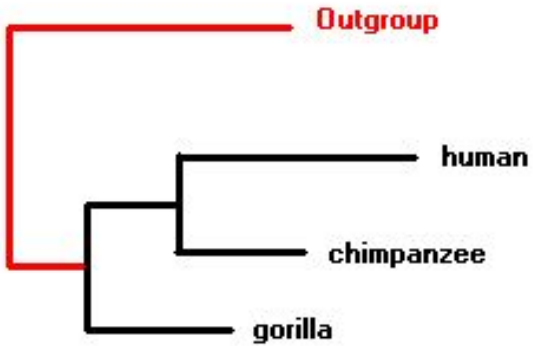
A **phylogenetic tree** is a hypothesis of how species or genes are related through evolution

What is a phylogeny…?

Unrooted tree

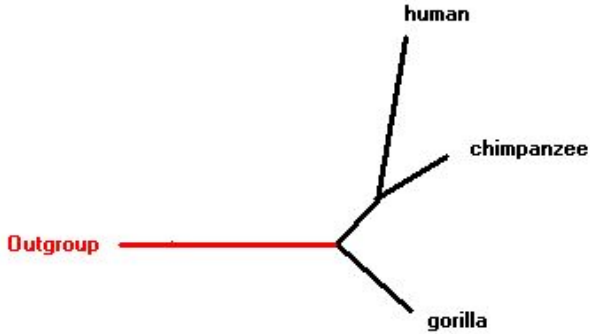


Rooted tree

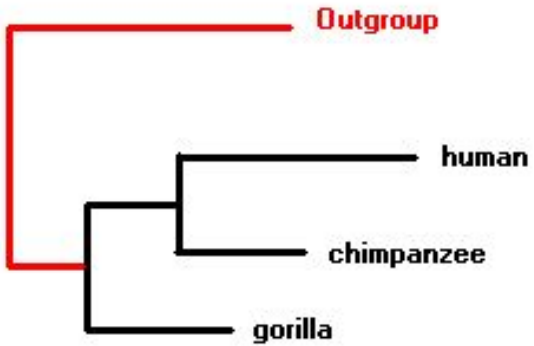


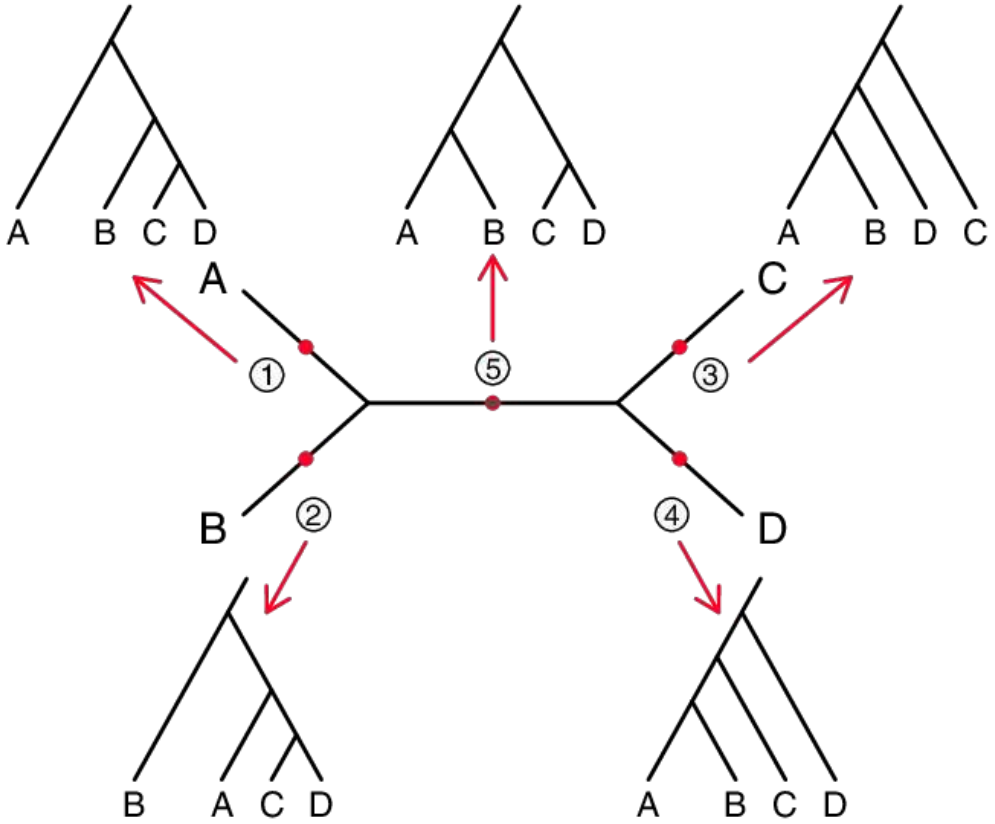
What is a phylogeny…?

Unrooted tree



Rooted tree

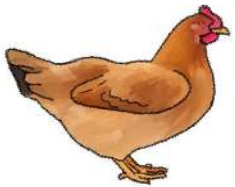




What is a phylogeny, why is it important…?

What is a phylogeny, why is it important…?

Which came first, the chicken or the egg?



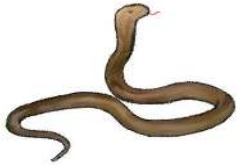
Birds
(Chickens)

What is a phylogeny, why is it important…?

Which came first, the chicken or the egg?

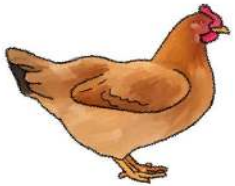
Turtles

Lizards



Snakes

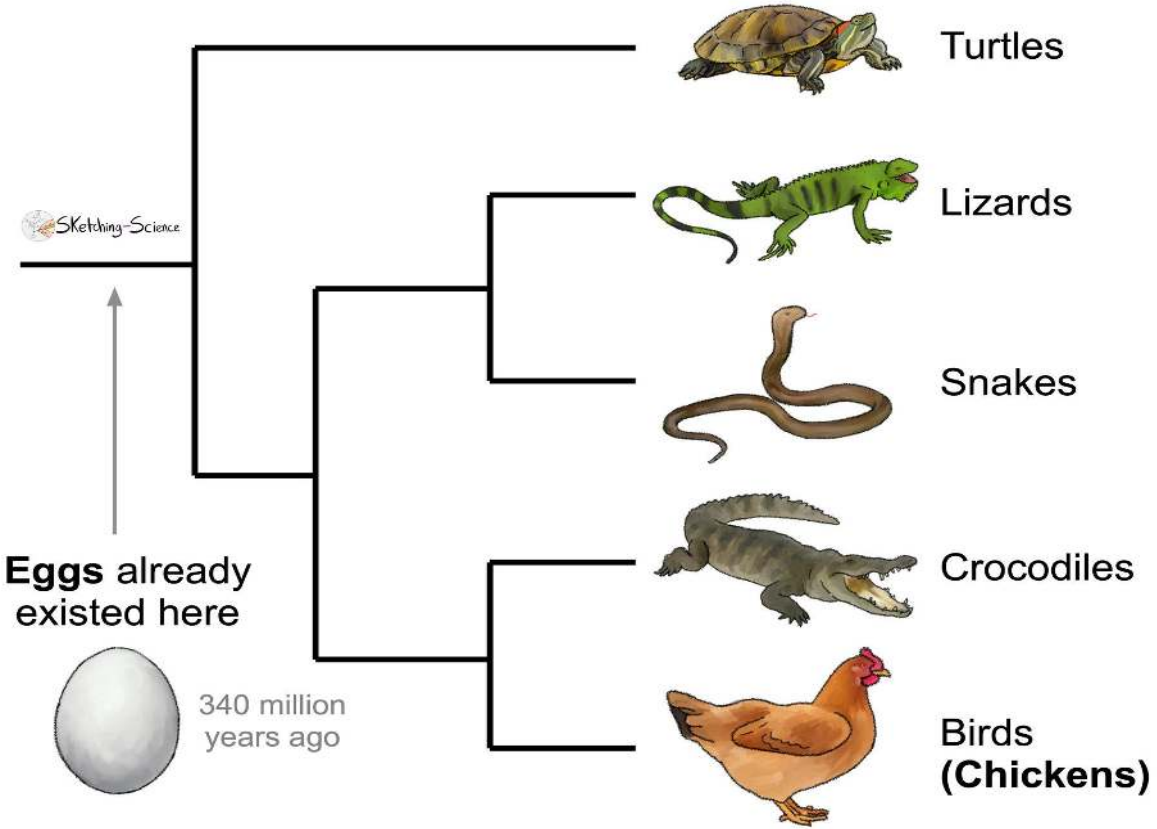
Crocodiles



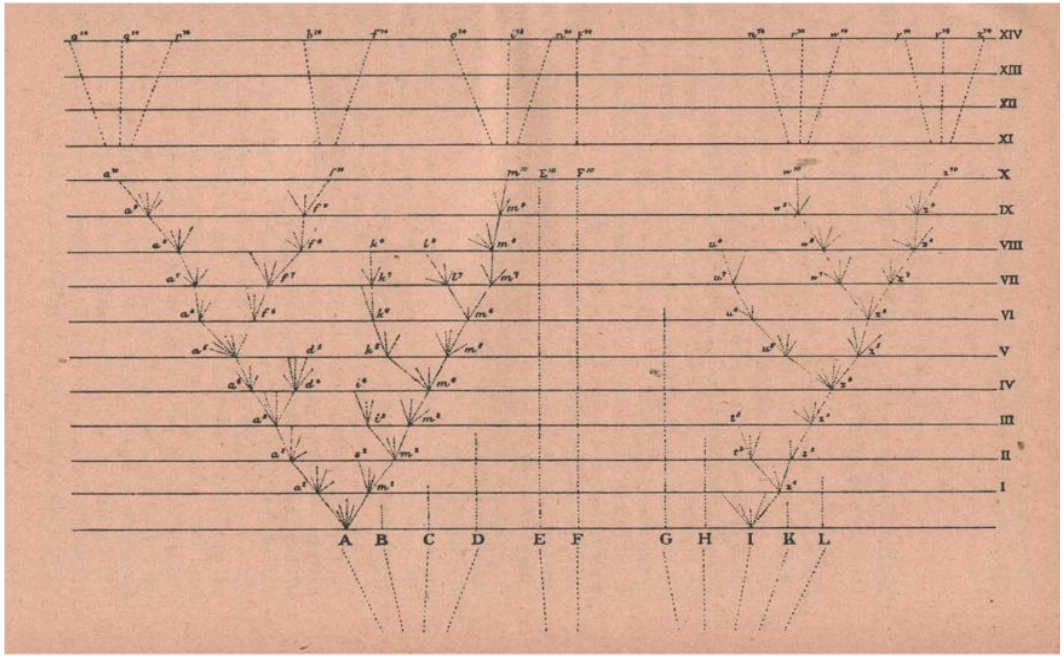
Birds
(Chickens)

What is a phylogeny, why is it important…?

Which came first, the chicken or the egg?



The first phylogenies



(Darwin 1859)

"As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications"

The first phylogenies

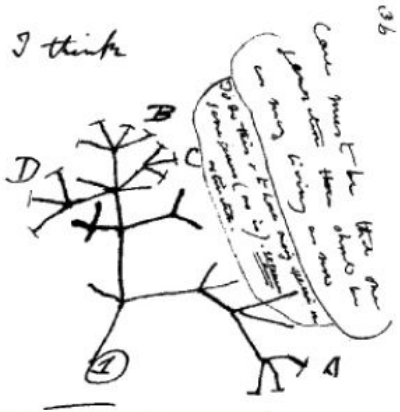
and instinct as the summing up of many contrivances, each useful to the possessor, nearly in the same way as when we look at any great mechanical invention as the summing up of the labour, the experience, the reason, and even the blunders of numerous workmen; when we thus view each organic being, how far more interesting, I speak from experience, will the study of natural history become!

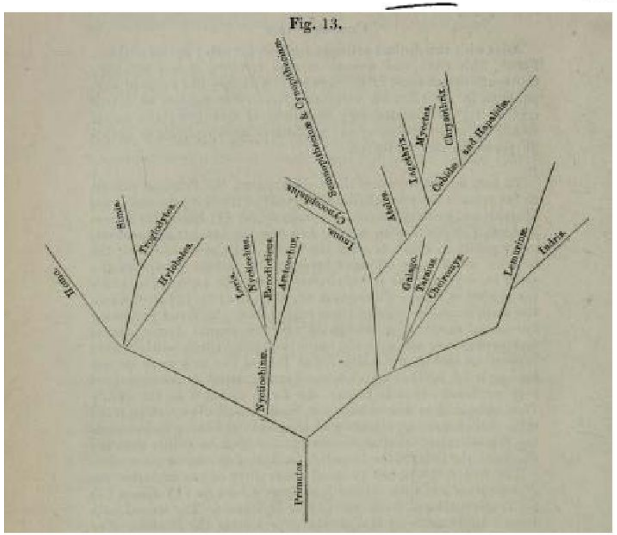
A grand and almost untrodden field of inquiry will be opened, on the causes and laws of variation, on correlation of growth, on the effects of use and disuse, on the direct action of external conditions, and so forth. The study of domestic productions will rise immensely in value. A new variety raised by man will be a far more important and interesting subject for study than one more species added to the infinitude of already recorded species. Our classifications will come to be, as far as they can be so made, genealogies; and will then truly give what may be called the plan of creation. The rules for classifying will no doubt become simpler when we have a definite object in view. We possess no pedigrees or armorial bearings; and we have to discover and trace the many diverging lines of descent in our natural genealogies, by characters of any kind which have long been inherited. Rudimentary organs will speak infallibly with respect to the nature of long-lost structures. Species and groups of species, which are called aberrant, and which may fancifully be called living fossils, will aid us in forming a picture of the ancient forms of life. Embryology will reveal to us the structure, in some degree obscured, of the prototypes of each great class.

When we can feel assured that all the individuals of the same species, and all the closely allied species of most genera, have within a not very remote period de-

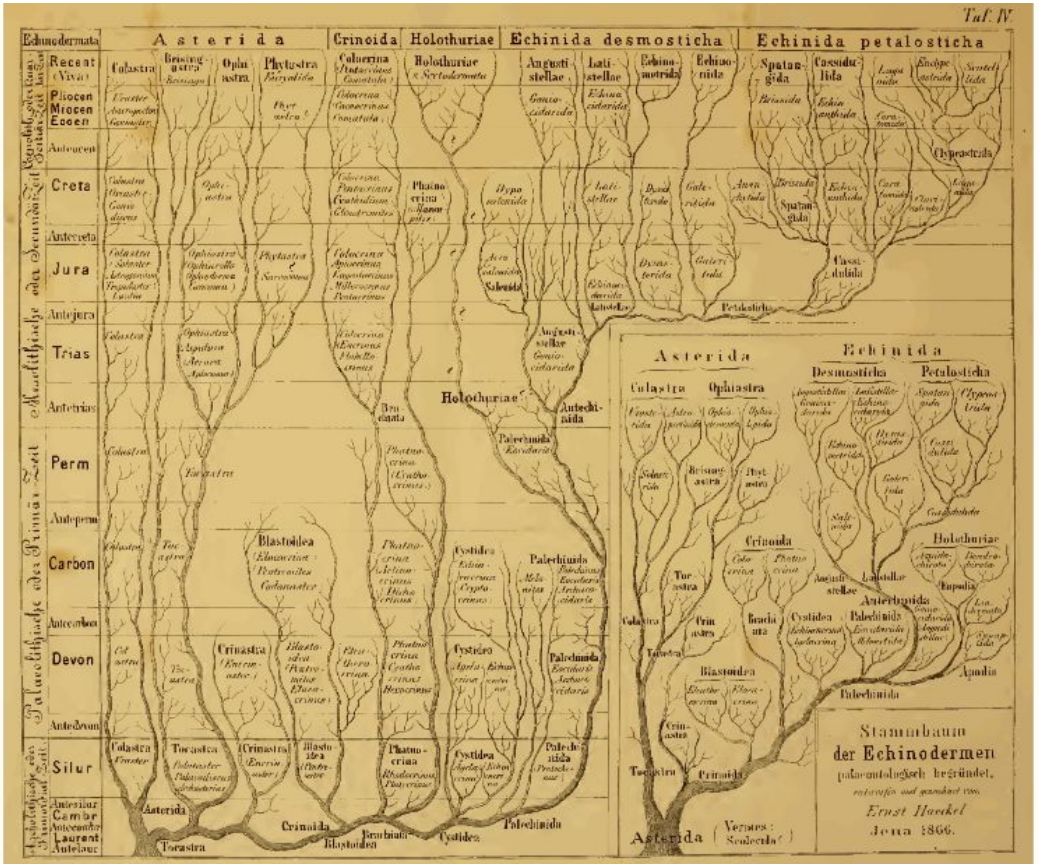
The first phylogenies

The concept:
Darwin's 'I think'
(1837)





Mivart (1865) Proc. Zool. Soc. London

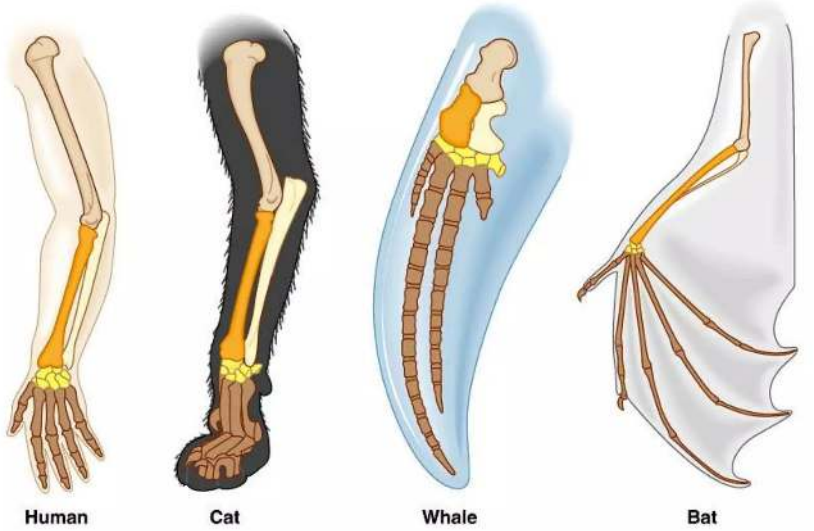


Haeckel (1866)

What is a phylogeny, why is it important… and how do you build one?

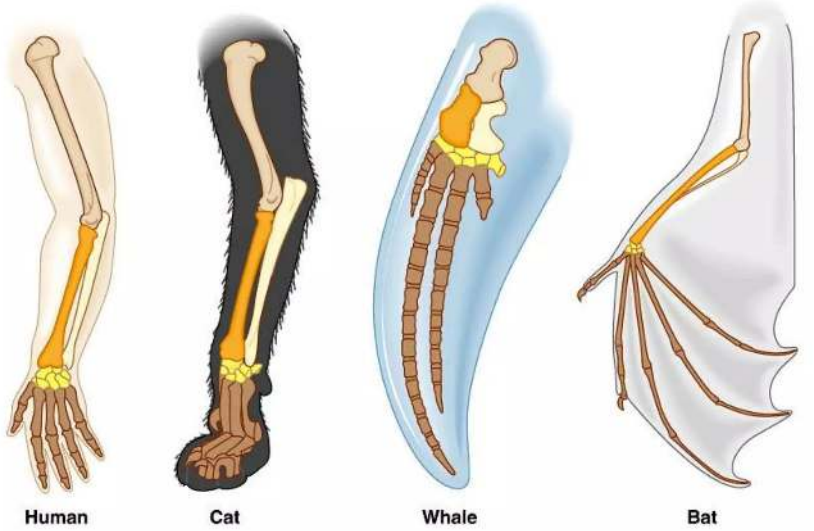
What is a phylogeny, why is it important… and how do you build one?

Homologous Structures



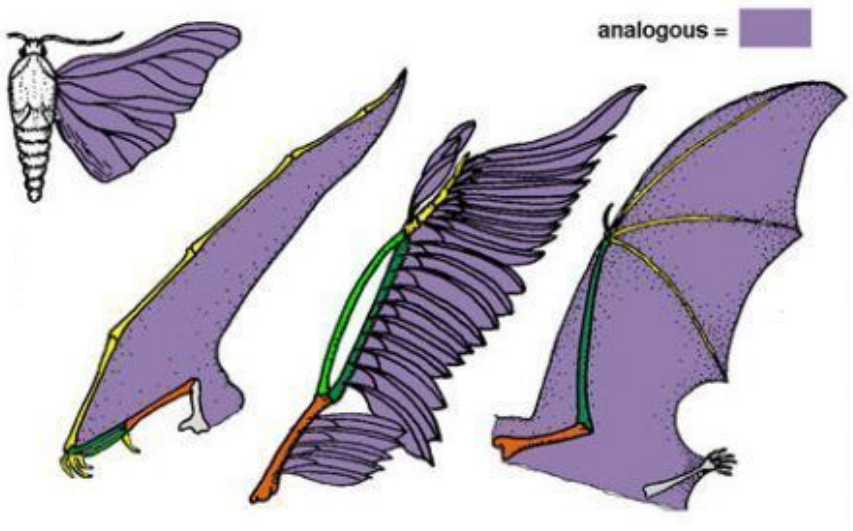
What is a phylogeny, why is it important… and how do you build one?

Homologous Structures



VS

Analogous Structures

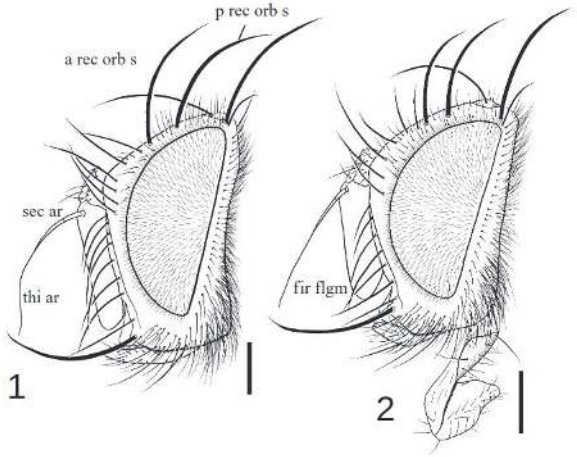


What is a phylogeny, why is it important… and how do you build one?

Systematic study of the genus *Phorinia* Robineau-Desvoidy of the Palearctic, Oriental and Oceanian regions (Diptera : Tachinidae)

Takuji Tachi[A,C] *and Hiroshi Shima*[B]

Invertebrate Systematics, 2006, **20**, 255–287



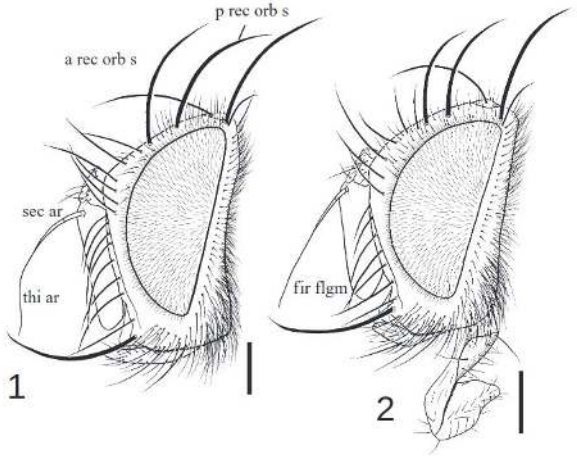
Figs 1–2. Male heads in profile: *1, Phorinia spinulosa*, sp. nov.; *2, P. breviata*, sp. nov. (Abbreviations: fir flgm, first flagellomere; sec ar, second aristomere; thi ar, third aristomere; a rec orb s, anterior reclinate orbital seta; p rec orb s, posterior reclinate orbital seta). Scale bars = 0.5 mm.

What is a phylogeny, why is it important… and how do you build one?

Systematic study of the genus *Phorinia* Robineau-Desvoidy of the Palearctic, Oriental and Oceanian regions (Diptera : Tachinidae)

Takuji Tachi[A,C] *and Hiroshi Shima*[B]

Invertebrate Systematics, 2006, **20**, 255–287



Figs 1–2. Male heads in profile: *1, Phorinia spinulosa*, sp. nov.; *2, P. breviata*, sp. nov. (Abbreviations: fir flgm, first flagellomere; sec ar, second aristomere; thi ar, third aristomere; a rec orb s, anterior reclinate orbital seta; p rec orb s, posterior reclinate orbital seta). Scale bars = 0.5 mm.

Table 2. Characters used for phylogenetic analysis

Lengths (L), consistency indices (*CI*) and retention indices (*RI*) are described from the unweighted analysis.

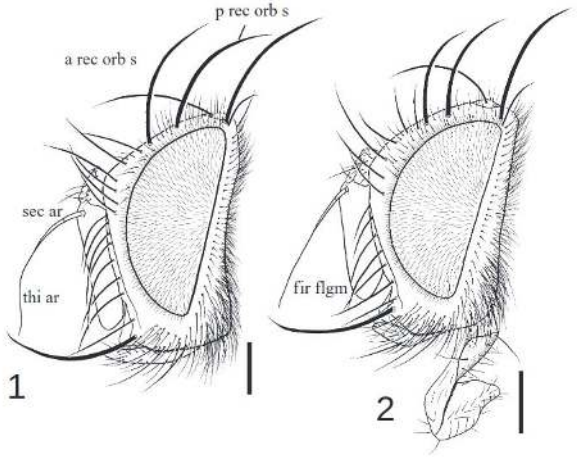
(*1*) *Eye:* 0, setulose (Figs 1–4); 1, bare or sparsely haired. L = 4; *CI* = 0.25; *RI* = 0.73.
(*2*) *Ocellar setae:* 0, present and strong (Figs 1–4); 1, absent or short and weak. L = 2; *CI* = 0.50; *RI* = 0.50.
(*3*) *Facial ridge:* 0, bare; 1, with short setae; 2, with strong setae (Figs 1–4). L = 3; *CI* = 0.67; *RI* = 0.94.
(*4*) *Occiput:* 0, without black setulae behind postocular row; 1, with black setulae behind postocular row. L = 2; *CI* = 0.50; *RI* = 0.86.
(*5*) *First supra-alar setae (sa):* 0, longer than first intra-alar seta (ia); 1, shorter than first intra-alar seta. L = 1; *CI* = 1; *RI* = 0.
(*6*) *Apical scutellar setae:* 0, horizontal or absent; 1, directed upwards. L = 4; *CI* = 0.25; *RI* = 0.81.
(*7*) *Setae on vein* R_{4+5}*:* 0, only base (at most to halfway to crossvein r-m); 1, from base nearly to crossvein r-m or beyond. L = 3; *CI* = 0.33; *RI* = 0.89.

What is a phylogeny, why is it important… and how do you build one?

Systematic study of the genus *Phorinia* Robineau-Desvoidy of the Palearctic, Oriental and Oceanian regions (Diptera : Tachinidae)

Takuji Tachi[A,C] *and Hiroshi Shima*[B]

Invertebrate Systematics, 2006, **20**, 255–287



Figs 1–2. Male heads in profile: *1, Phorinia spinulosa*, sp. nov.; *2, P. breviata*, sp. nov. (Abbreviations: fir flgm, first flagellomere; sec ar, second aristomere; thi ar, third aristomere; a rec orb s, anterior reclinate orbital seta; p rec orb s, posterior reclinate orbital seta). Scale bars = 0.5 mm.

Table 3. Morphological data matrix used for phylogenetic analysis

Taxa	Characters			
	0000000001	1111111112	2222222223	3
	1234567890	1234567890	1234567890	1
Winthemia venusta	0000000000	0000000000	-000001000	0
Drinomyia hokkaidensis	1000100001	0100000000	-000002000	0
Phorocerosoma vicarium	0000100000	0010000000	-000001000	0
Austrophorocera grandis	0120100000	0101010001	0000003000	1
A. hirsuta	0020100000	0001010001	0000003000	1
Bessa parallela	1021101000	0001010001	1000003000	1

Table 2. Characters used for phylogenetic analysis

Lengths (L), consistency indices (*CI*) and retention indices (*RI*) are described from the unweighted analysis.

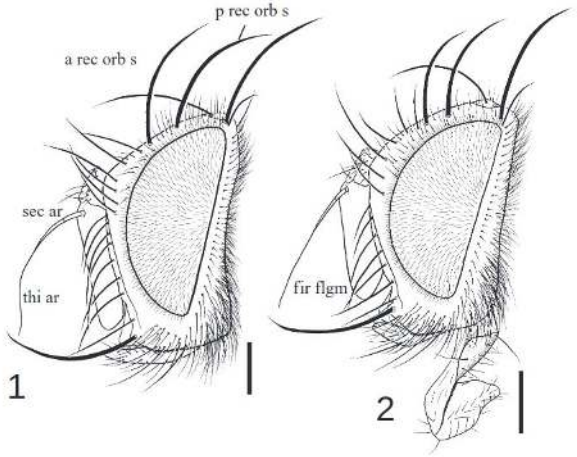
(*1*) *Eye:* 0, setulose (Figs 1–4); 1, bare or sparsely haired. L = 4; *CI* = 0.25; *RI* = 0.73.
(*2*) *Ocellar setae:* 0, present and strong (Figs 1–4); 1, absent or short and weak. L = 2; *CI* = 0.50; *RI* = 0.50.
(*3*) *Facial ridge:* 0, bare; 1, with short setae; 2, with strong setae (Figs 1–4). L = 3; *CI* = 0.67; *RI* = 0.94.
(*4*) *Occiput:* 0, without black setulae behind postocular row; 1, with black setulae behind postocular row. L = 2; *CI* = 0.50; *RI* = 0.86.
(*5*) *First supra-alar setae (sa):* 0, longer than first intra-alar seta (ia); 1, shorter than first intra-alar seta. L = 1; *CI* = 1; *RI* = 0.
(*6*) *Apical scutellar setae:* 0, horizontal or absent; 1, directed upwards. L = 4; *CI* = 0.25; *RI* = 0.81.
(*7*) *Setae on vein* R_{4+5}*:* 0, only base (at most to halfway to crossvein r-m); 1, from base nearly to crossvein r-m or beyond. L = 3; *CI* = 0.33; *RI* = 0.89.

What is a phylogeny, why is it important… and how do you build one?

Systematic study of the genus *Phorinia* Robineau-Desvoidy of the Palearctic, Oriental and Oceanian regions (Diptera : Tachinidae)

Takuji Tachi[A,C] *and Hiroshi Shima*[B]

Invertebrate Systematics, 2006, **20**, 255–287



Figs 1–2. Male heads in profile: *1, Phorinia spinulosa*, sp. nov.; *2, P. breviata*, sp. nov. (Abbreviations: fir flgm, first flagellomere; sec ar, second aristomere; thi ar, third aristomere; a rec orb s, anterior reclinate orbital seta; p rec orb s, posterior reclinate orbital seta). Scale bars = 0.5 mm.

Table 3. Morphological data matrix used for phylogenetic analysis

Taxa	Characters			
	0000000001	1111111112	2222222223	3
	1234567890	1234567890	1234567890	1
Winthemia venusta	0000000000	0000000000	-000001000	0
Drinomyia hokkaidensis	1000100001	0100000000	-000002000	0
Phorocerosoma vicarium	0000100000			
Austrophorocera grandis	0120100000			
A. hirsuta	0020100000			
Bessa parallela	1021101000			

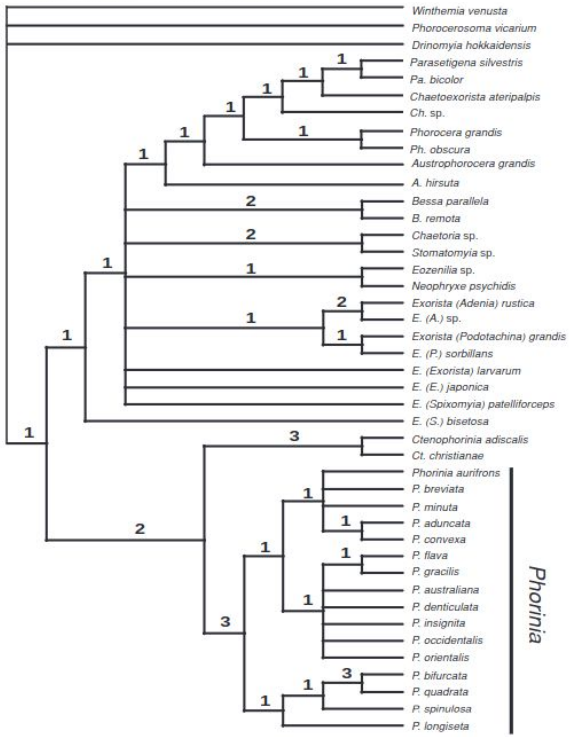


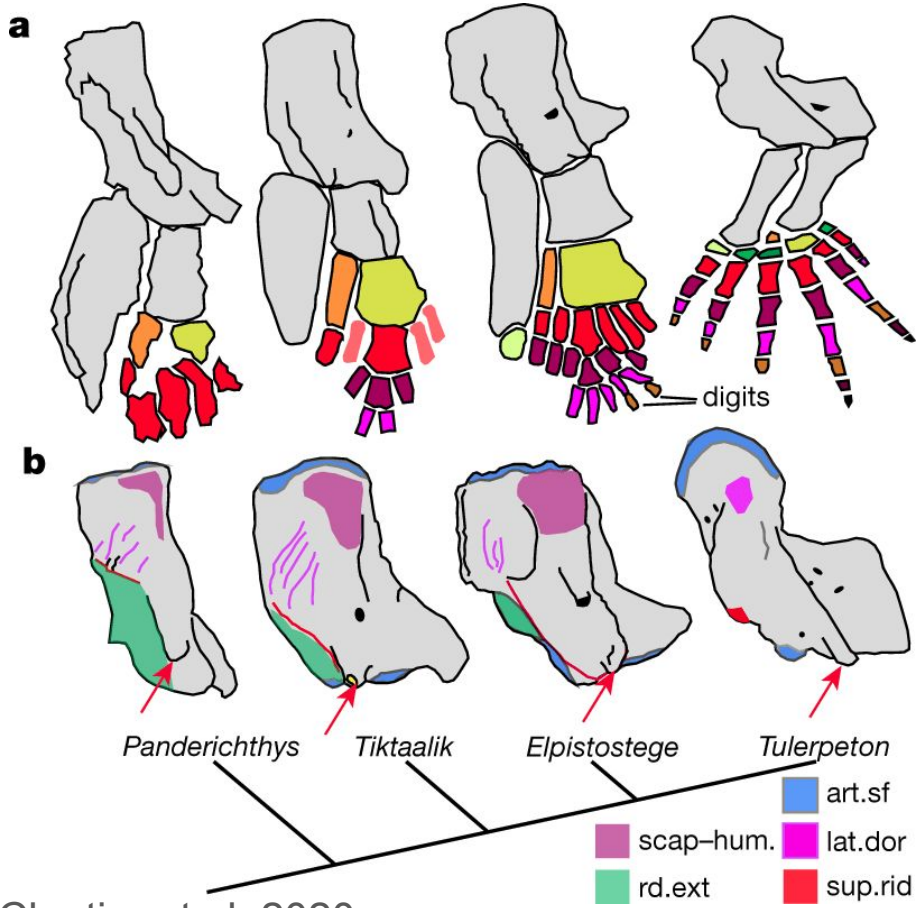
Fig. 79. Strict consensus of 186 equally most parsimonious cladograms (length = 66, consistency index (*CI*) = 0.530, rescaled consistency index (*RC*) = 0.462) generated from an analysis of thirty-one morphological characters. Bremer support values are given on the branches.

Table 2. Characters used for phylogenetic analysis

Lengths (L), consistency indices (*CI*) and retention indices (*RI*) are described from the unweighted analysis.

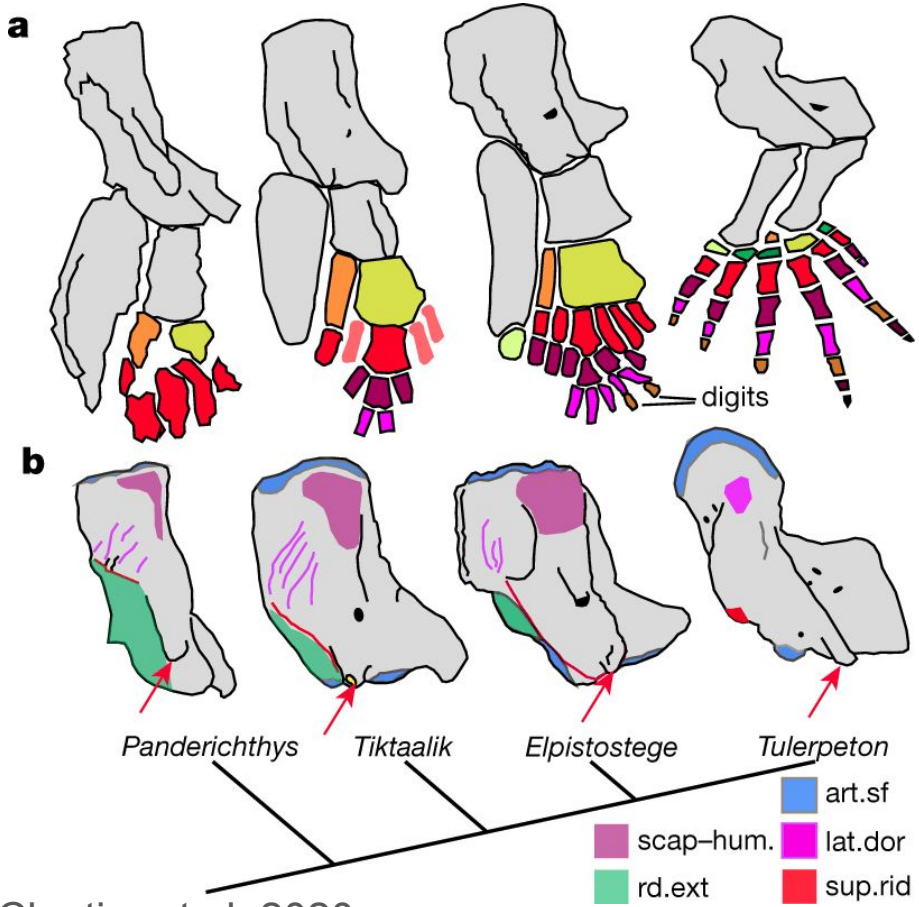
(*1*) *Eye:* 0, setulose (Figs 1–4); 1, bare or sparsely haired. L = 4; *CI* = 0.25; *RI* = 0.73.
(*2*) *Ocellar setae:* 0, present and strong (Figs 1–4); 1, absent or short and weak. L = 2; *CI* = 0.50; *RI* = 0.50.
(*3*) *Facial ridge:* 0, bare; 1, with short setae; 2, with strong setae (Figs 1–4). L = 3; *CI* = 0.67; *RI* = 0.94.
(*4*) *Occiput:* 0, without black setulae behind postocular row; 1, with black setulae behind postocular row. L = 2; *CI* = 0.50; *RI* = 0.86.
(*5*) *First supra-alar setae (sa):* 0, longer than first intra-alar seta (ia); 1, shorter than first intra-alar seta. L = 1; *CI* = 1; *RI* = 0.
(*6*) *Apical scutellar setae:* 0, horizontal or absent; 1, directed upwards. L = 4; *CI* = 0.25; *RI* = 0.81.
(*7*) *Setae on vein* R_{4+5}*:* 0, only base (at most to halfway to crossvein r-m); 1, from base nearly to crossvein r-m or beyond. L = 3; *CI* = 0.33; *RI* = 0.89.

What is a phylogeny, why is it important… and how do you build one?

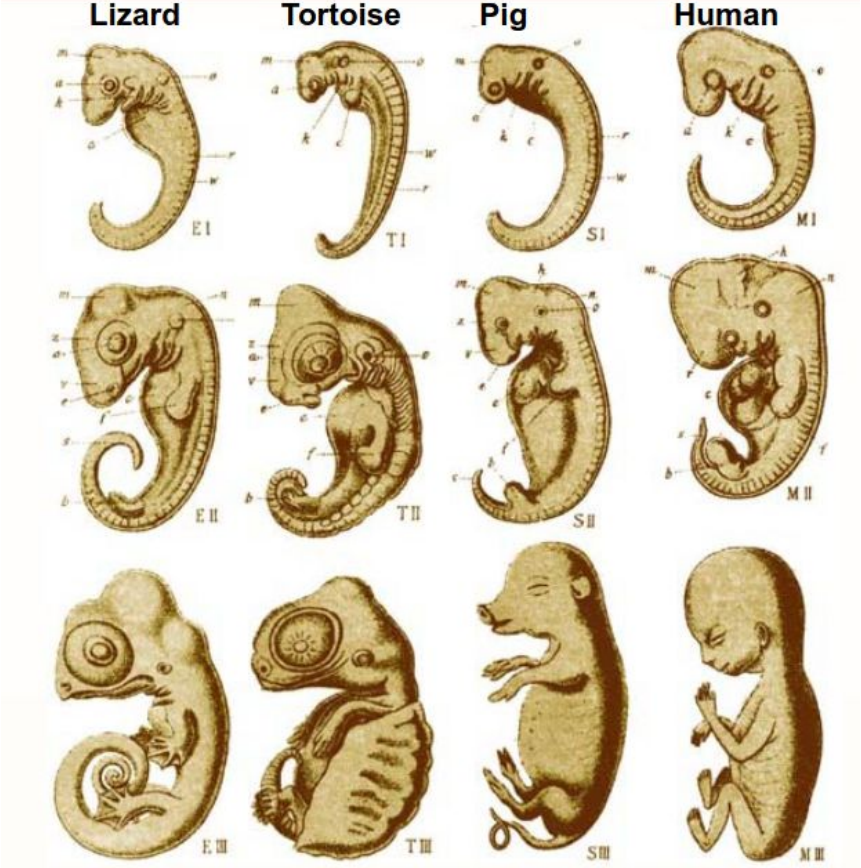


Cloutier et al. 2020

What is a phylogeny, why is it important… and how do you build one?



Cloutier et al. 2020



http://www.nature.com/nrg/journal/v7/n11/images/nrg1918-f2.jpg

The origin of molecular phylogenetics

The origin of molecular phylogenetics

BLOOD IMMUNITY

AND

BLOOD RELATIONSHIP

A DEMONSTRATION OF CERTAIN BLOOD-RELATIONSHIPS AMONGST ANIMALS BY MEANS OF

THE PRECIPITIN TEST FOR BLOOD

by

GEORGE H. F. NUTTALL, M.A., M.D., PH.D.

University Lecturer in Bacteriology and Preventive Medicine, Cambridge.

Including

Original Researches by

G. S. GRAHAM-SMITH, M.A., M.B., D.P.H. (Camb.)

and

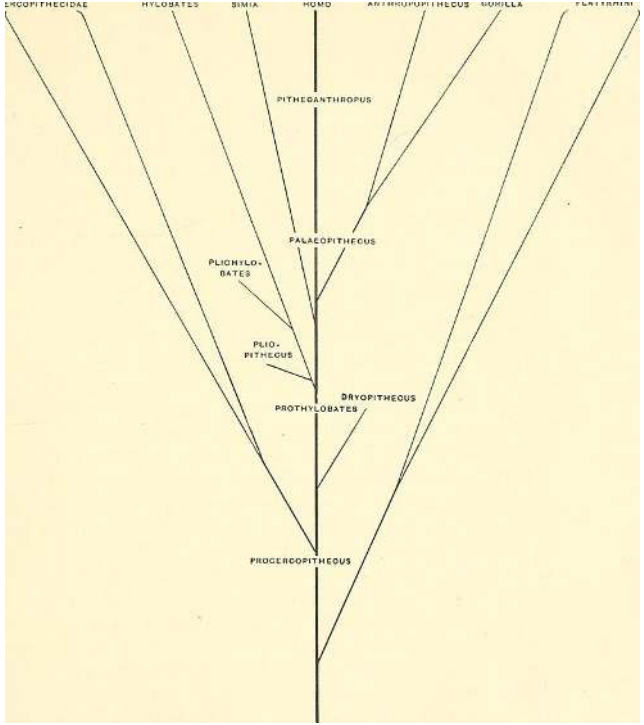
T. S. P. STRANGEWAYS, M.A., M.R.C.S.

CAMBRIDGE:

at the University Press

1904

Nuttal (1904) - serological cross-reactions were stronger for more closely related organisms -> phylogeny of apes



The origin of molecular phylogenetics

BLOOD IMMUNITY
AND
BLOOD RELATIONSHIP

Nuttal (1904) - serological cross-reactions were stronger for more closely related organisms -> phylogeny of apes

Dobzhansky & Sturtevant (1938) - genomic rearrangements in *Drosophila* as phylogenetic markers

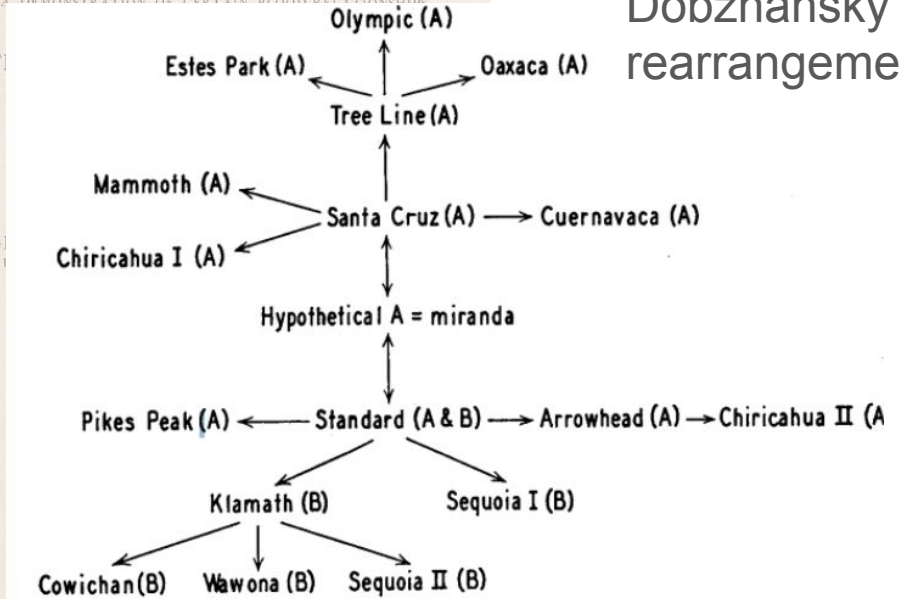
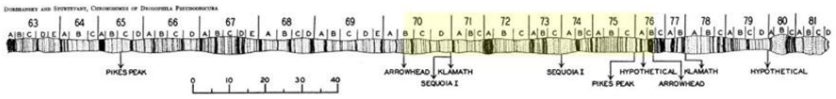
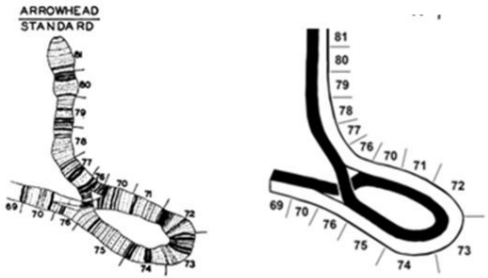


FIGURE 3.—Phylogeny of the gene arrangements in the third chromosome of *Drosophila pse doobscura*. Any two arrangements connected by an arrow in the diagram differ by a single i version. Further explanation in text.

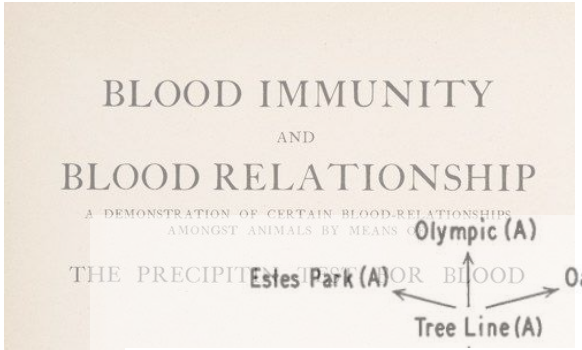


Chromosome 3 of *Drosophila pseudoobscura*



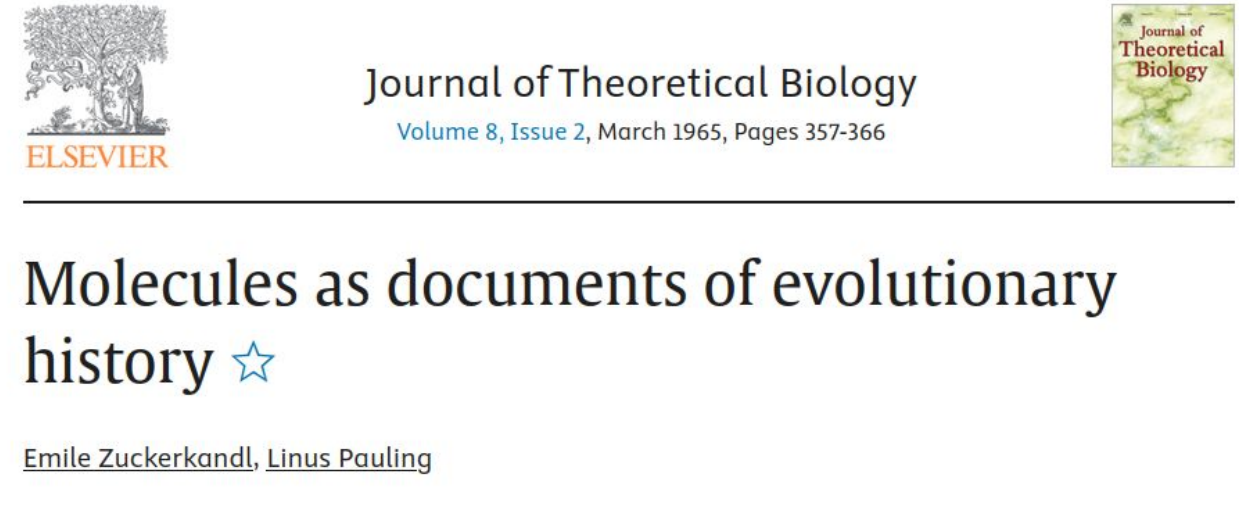
Standard and Arrowhead arrangements differ by an inversion from segments 70 to 76

The origin of molecular phylogenetics


BLOOD IMMUNITY
AND
BLOOD RELATIONSHIP
A DEMONSTRATION OF CERTAIN BLOOD-RELATIONSHIPS
AMONGST ANIMALS BY MEANS
THE PRECIPITIN
Olympic (A)
Estes Park (A)
Oaxaca (A)
Tree Line (A)
version. Further explanation in text.

Nuttal (1904) - serological cross-reactions were stronger for more closely related organisms -> phylogeny of apes

Dobzhansky & Sturtevant (1938) - genomic rearrangements in Drosophila as phylogenetic markers


ELSEVIER

Journal of Theoretical Biology

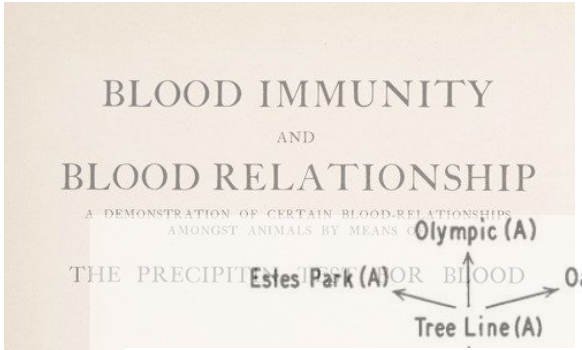
Volume 8, Issue 2, March 1965, Pages 357-366

Molecules as documents of evolutionary history ☆

Emile Zuckerkandl, Linus Pauling

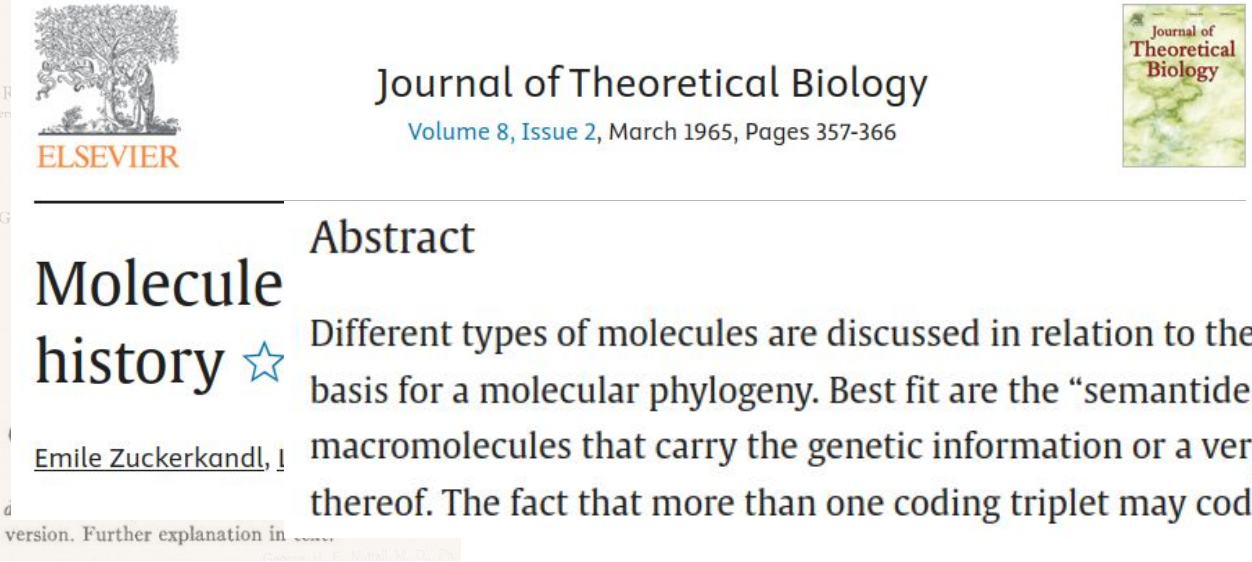
Zuckerkandl & Pauling (1965) -

The origin of molecular phylogenetics


BLOOD IMMUNITY
AND
BLOOD RELATIONSHIP
A DEMONSTRATION OF CERTAIN BLOOD-RELATIONSHIPS
AMONGST ANIMALS BY MEANS
THE PRECIPIT OR BLOOD
Olympic (A)
Estes Park (A)
Oaxaca (A)
Tree Line (A)

Nuttal (1904) - serological cross-reactions were stronger for more closely related organisms -> phylogeny of apes

Dobzhansky & Sturtevant (1938) - genomic rearrangements in Drosophila as phylogenetic markers


ELSEVIER

Journal of Theoretical Biology

Volume 8, Issue 2, March 1965, Pages 357-366

Zuckerkandl & Pauling (1965) -

Molecule history ☆

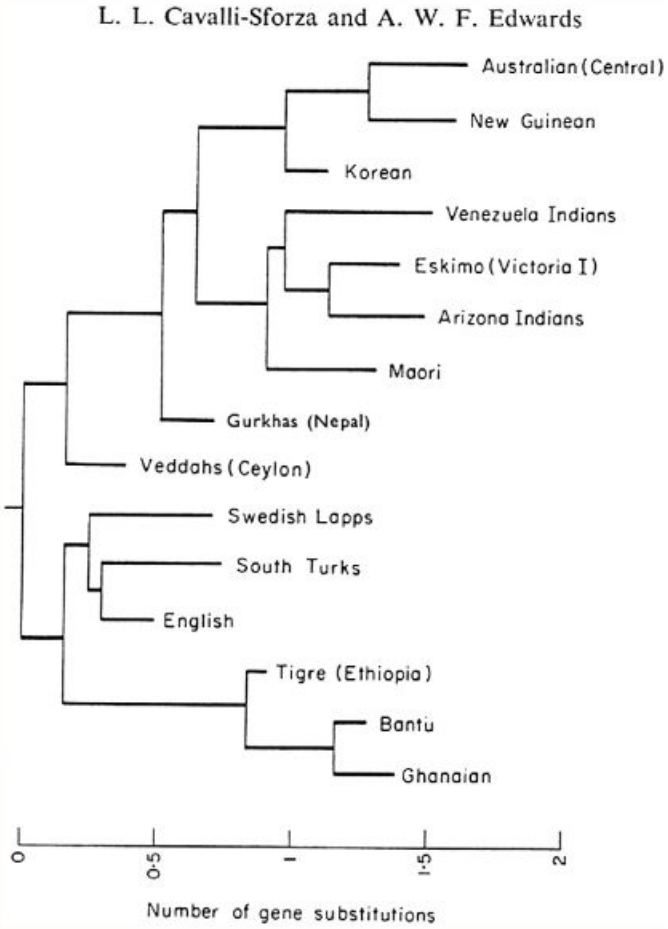
Emile Zuckerkandl, I

version. Further explanation in

Abstract

Different types of molecules are discussed in relation to their fitness for providing the basis for a molecular phylogeny. Best fit are the “semantides”, i.e. the different types of macromolecules that carry the genetic information or a very extensive translation thereof. The fact that more than one coding triplet may code for a given amino acid

Molecular phylogenetics: the new wave

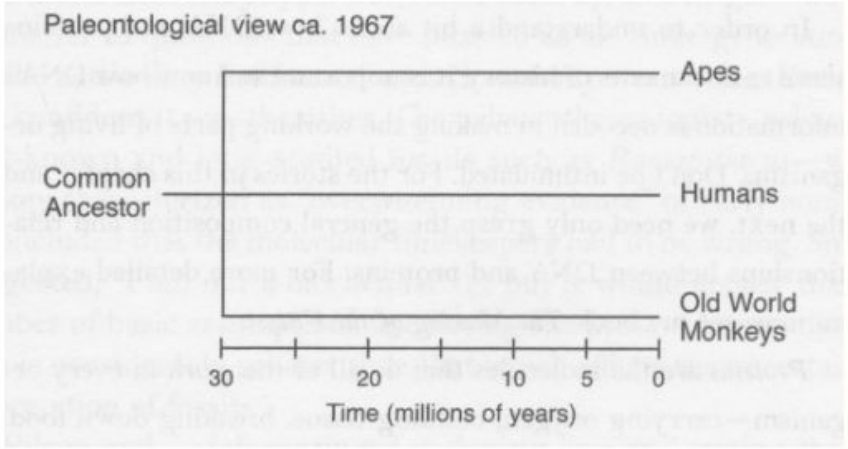


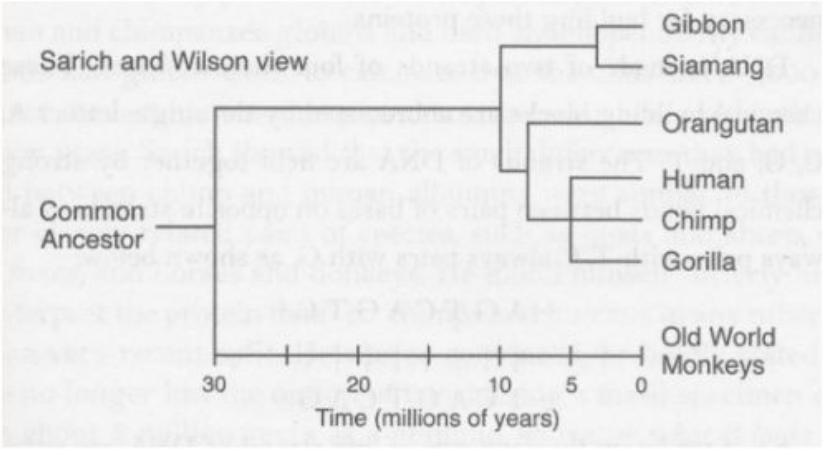
Phylogeny inferred from blood group allele frequencies from 15 populations

Cavalli-Sforza & Edwards (1965) in Genetics Today

Molecular phylogenetics: the new wave

Divergence times were estimated by measuring the immunological cross-reaction of blood serum albumin between pairs of primates





"no fuss, no muss, no dishpan hands. Just throw some proteins into a laboratory apparatus, shake them up, and bingo! – we have an answer to questions that have puzzled us for three generations."

Sarich & Wilson (1967) Science

Molecular phylogenetics: the new wave

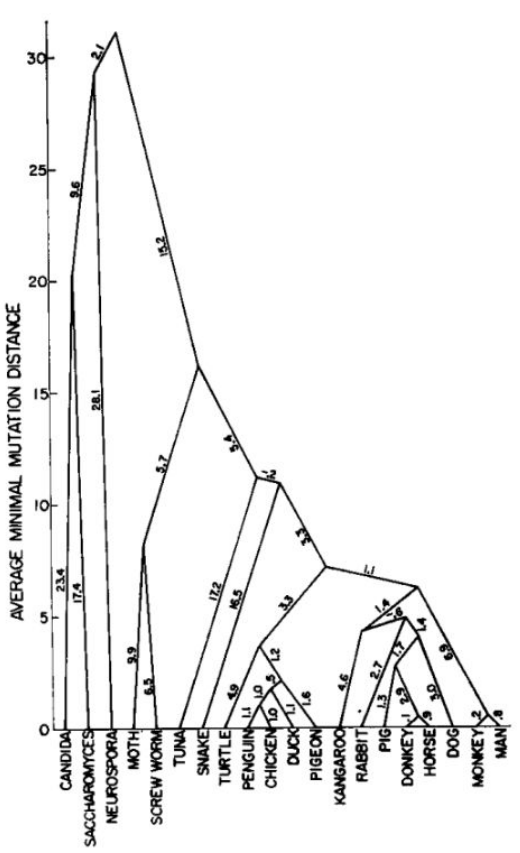
Construction of Phylogenetic Trees

A method based on mutation distances as estimated from cytochrome *c* sequences is of general applicability.

Walter M. Fitch and Emanuel Margoliash

Biochemists have attempted to use quantitative estimates of variance between substances obtained from different species to construct phylogenetic trees. Examples of this approach include studies of the degree of interspecific hybridization of DNA (*1*), the degree of cross reactivity of antisera to purified proteins (*2*), the number of differences in the peptides from enzymic digests of purified homologous proteins, both as estimated by paper electrophoresis-chromatography or column chromatography and as estimated from the amino acid compositions of the proteins (*3*), and the number of amino acid replacements between homologous proteins whose complete primary structures had been determined (*4*). These methods have not been completely satisfactory because (i) the portion of the genome examined

20 JANUARY 1967



Molecular phylogenetics: the new wave

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 11, pp. 5088–5090, November 1977
Evolution

Phylogenetic structure of the prokaryotic domain: The primary kingdoms

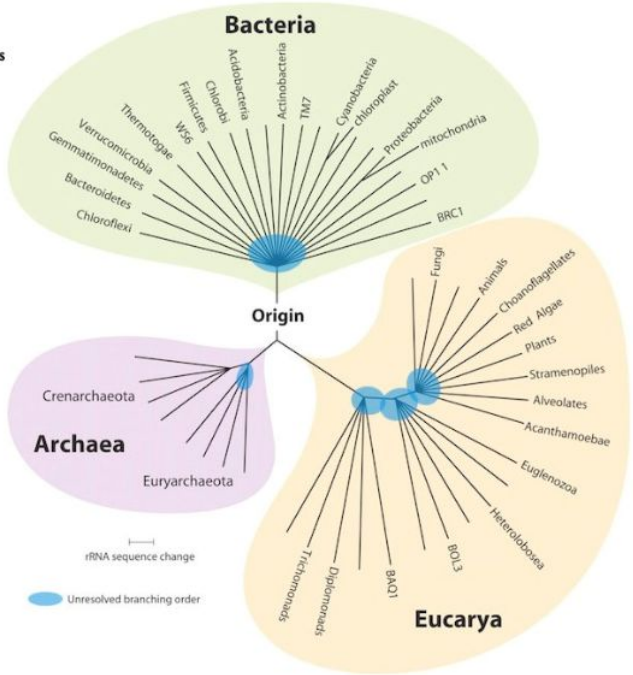
(archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

CARL R. WOESE AND GEORGE E. FOX*

Department of Genetics and Development, University of Illinois, Urbana, Illinois

Communicated by T. M. Sonneborn, August 18, 1977

ABSTRACT **A phylogenetic analysis based upon ribosomal RNA sequence characterization reveals that living systems represent one of three aboriginal lines of descent: (*i*) the eubacteria, comprising all typical bacteria; (*ii*) the archaebacteria, containing methanogenic bacteria; and (*iii*) the urkaryotes, now represented in the cytoplasmic component of eukaryotic cells.**



The dawn of phylogenomics

The dawn of phylogenomics

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

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

The dawn of phylogenomics

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

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

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

The dawn of phylogenomics

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

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.

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

Phylogenomics: prediction of gene function and gene family evolution

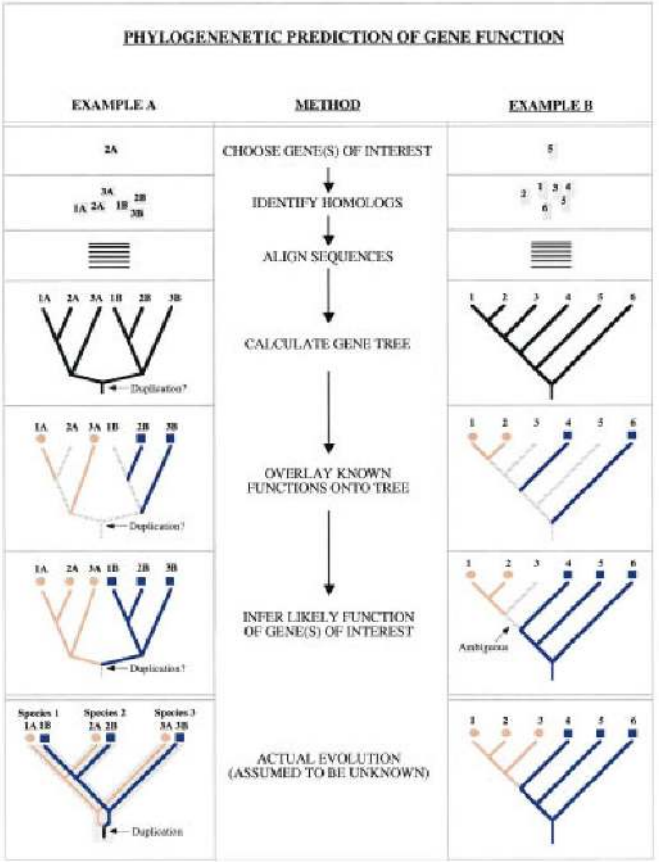


Figure 1 Outline of a phylogenomic methodology. In this method, information about the evolutionary relationships among genes is used to predict the functions of uncharacterized genes (see text for details). Two hypothetical scenarios are presented and the path of trying to

The dawn of 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

The dawn of phylogenomics

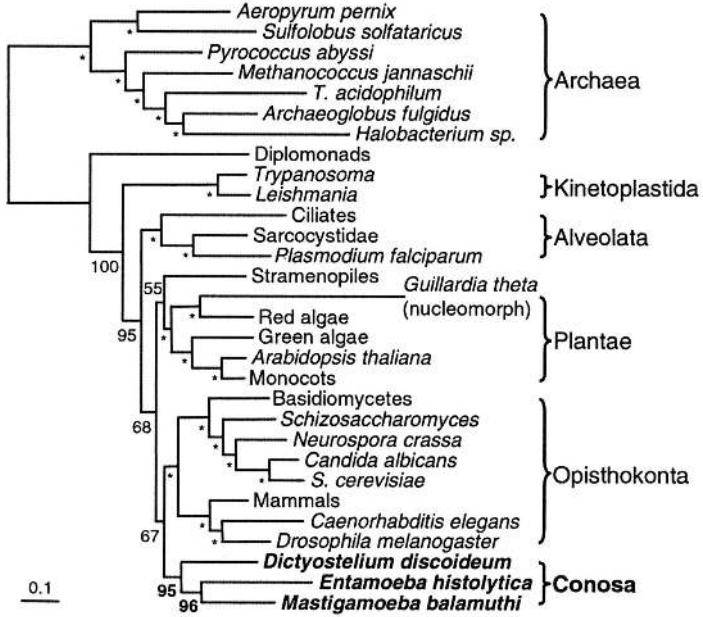
1414–1419 | PNAS | February 5, 2002 | vol. 99 | no. 3 www.pnas.org/cgi/doi/10.1073/pnas.032662799

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*‖

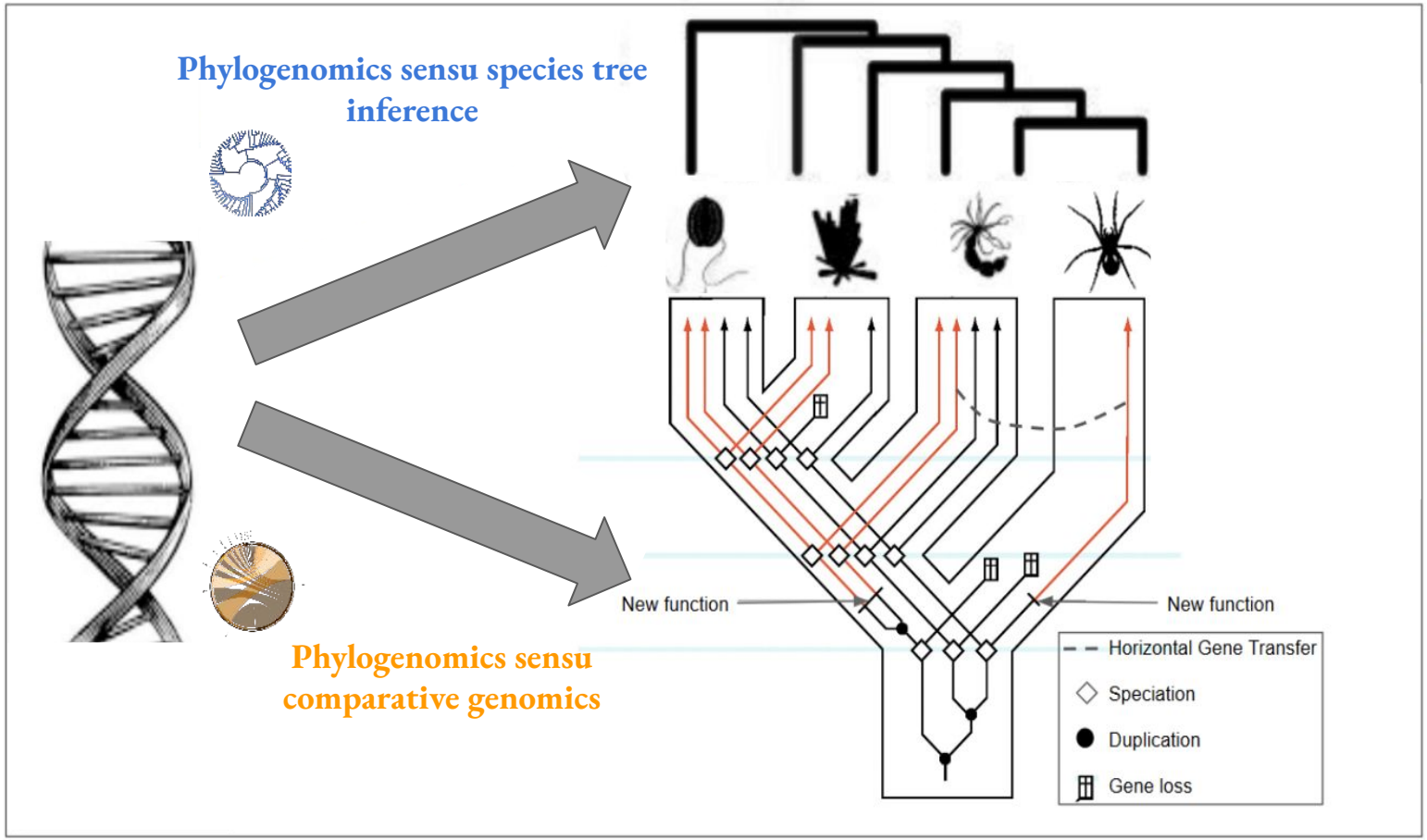
The phylogenetic relationships of amoebae are poorly resolved. To address this difficult question, we have sequenced 1,280 expressed sequence tags from *Mastigamoeba balamuthi* and assembled a large data set containing 123 genes for representatives of three phenotypically highly divergent major amoeboid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

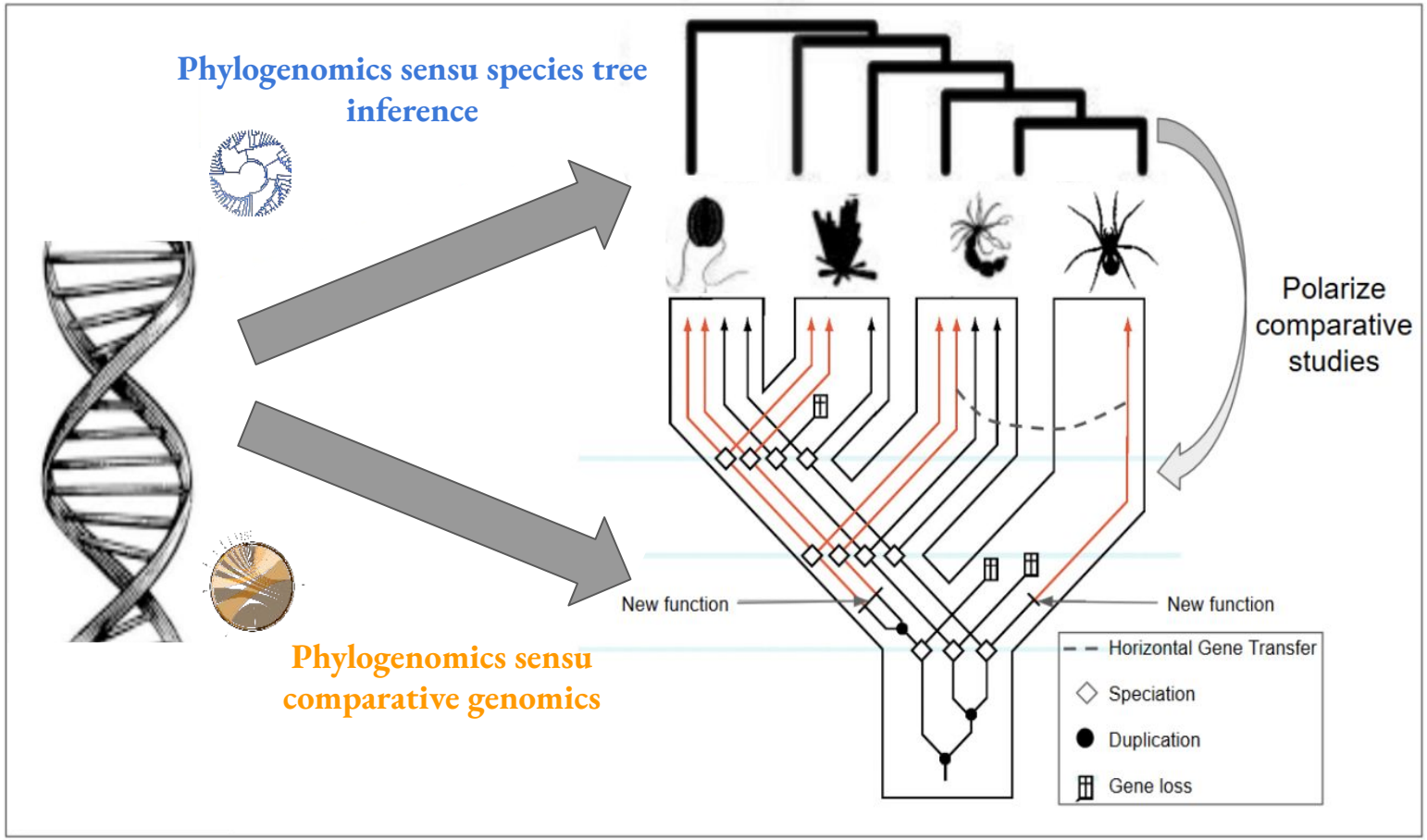


ML tree based on 25,032 aa positions. * indicates a constrained node. We used the JTT model, without taking into account among-sites rate variation. The branch lengths have been computed on the concatenated sequences. BVs were obtained by bootstrapping the 123 genes.

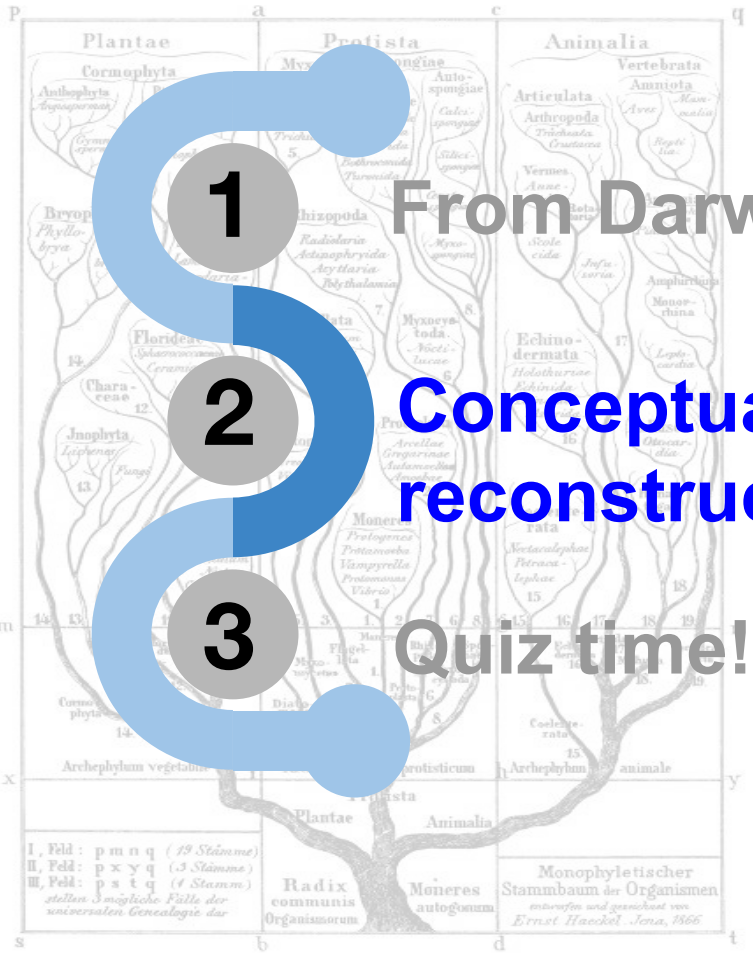
The dawn of phylogenomics



The dawn of phylogenomics



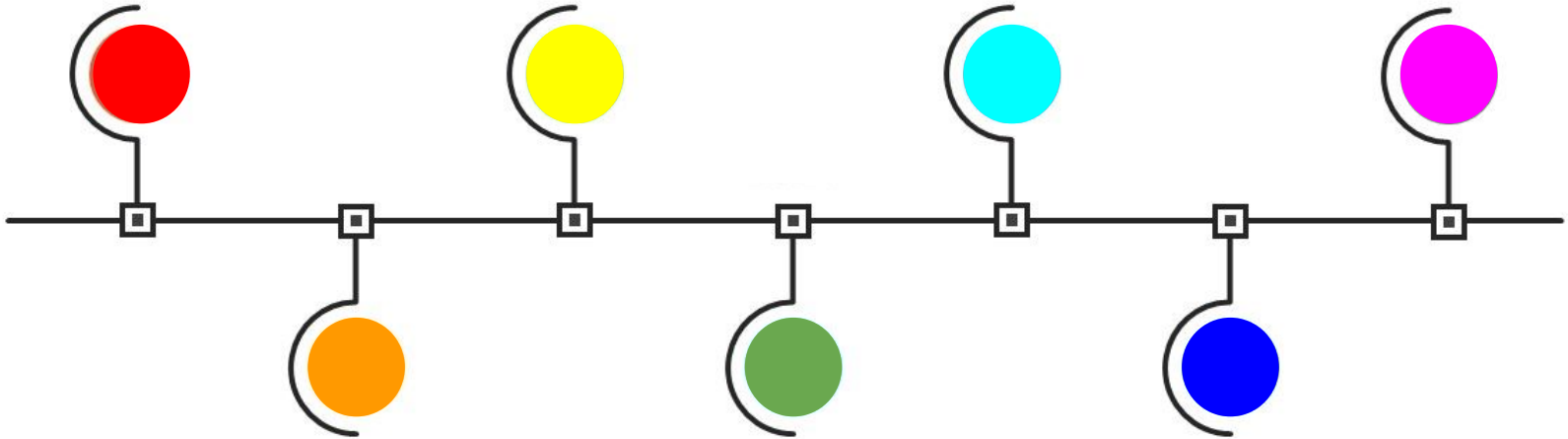
Content of the lecture



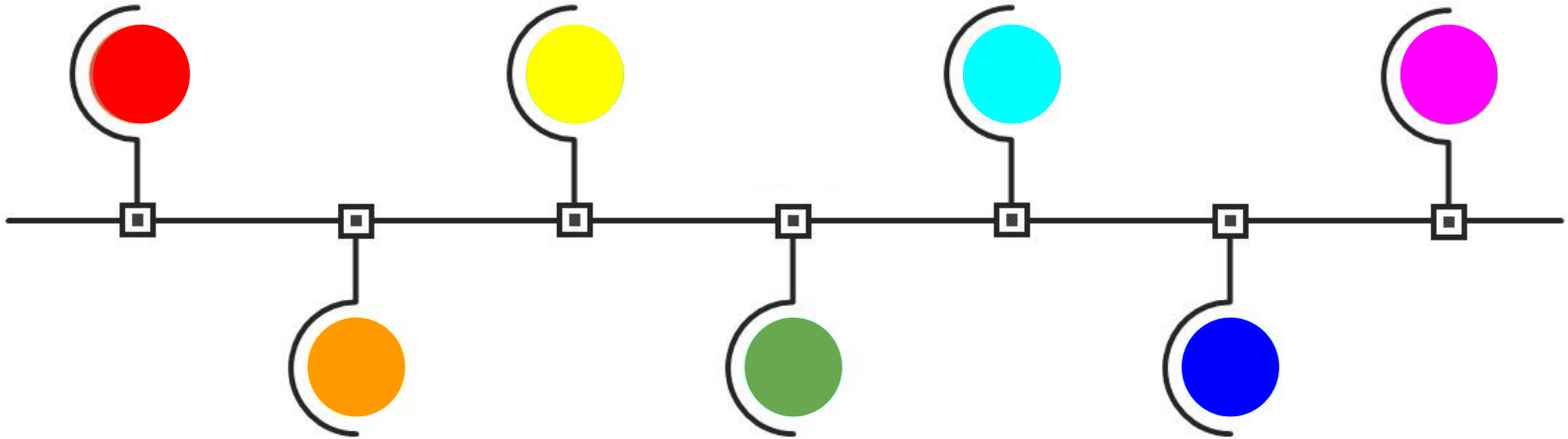
1 From Darwin to phylogenomics

2 **Conceptual framework for phylogenomic reconstruction**

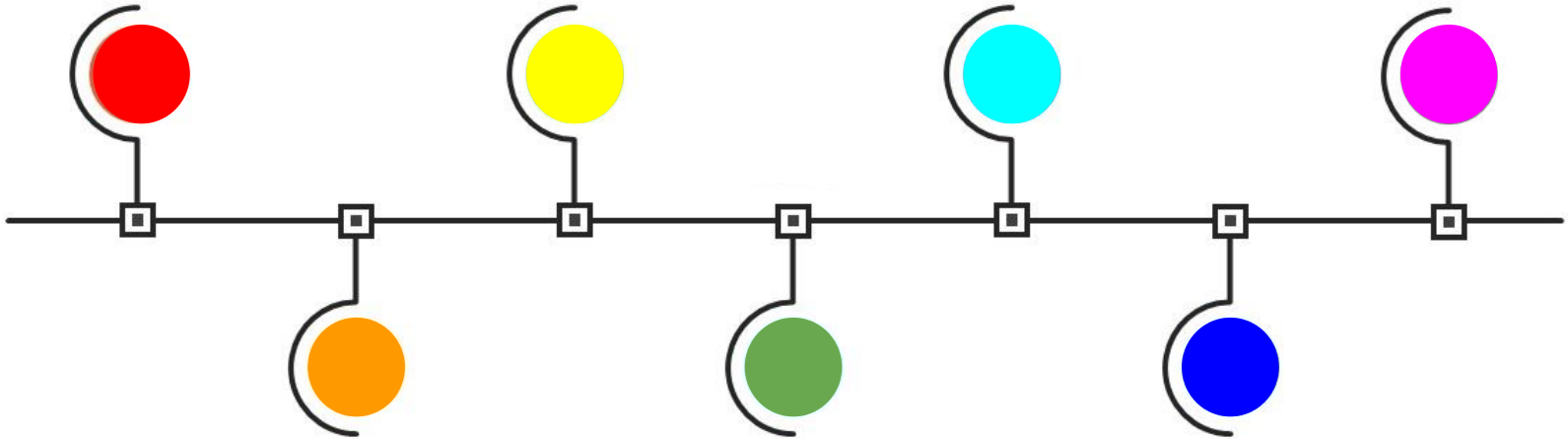
3 Quiz time!



01 DATA



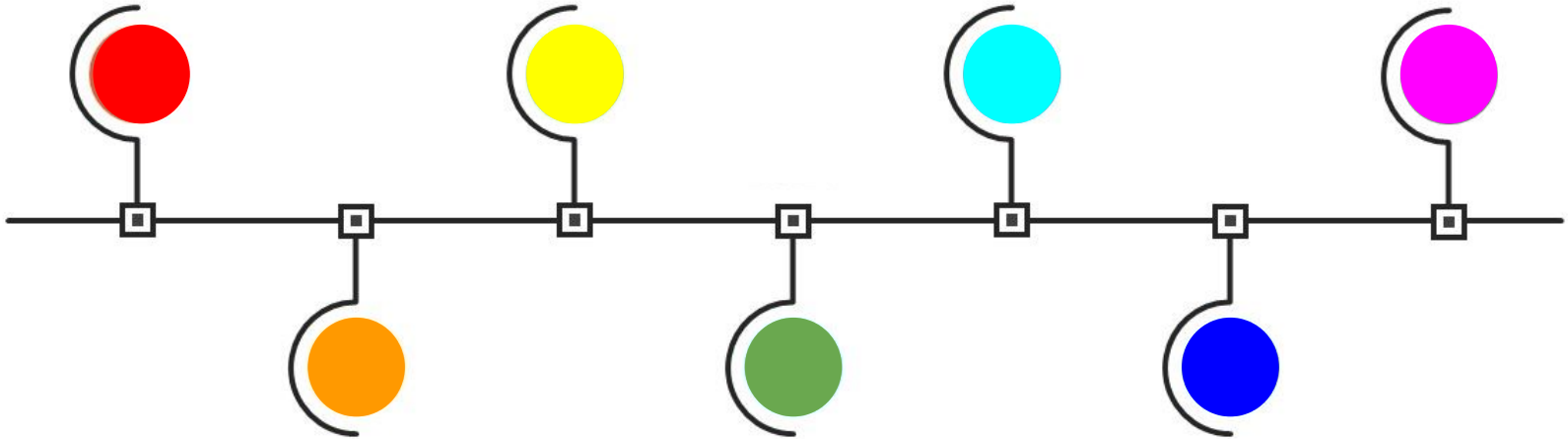
01 DATA



02 ORTHOLOGY INFERENCE

01 DATA

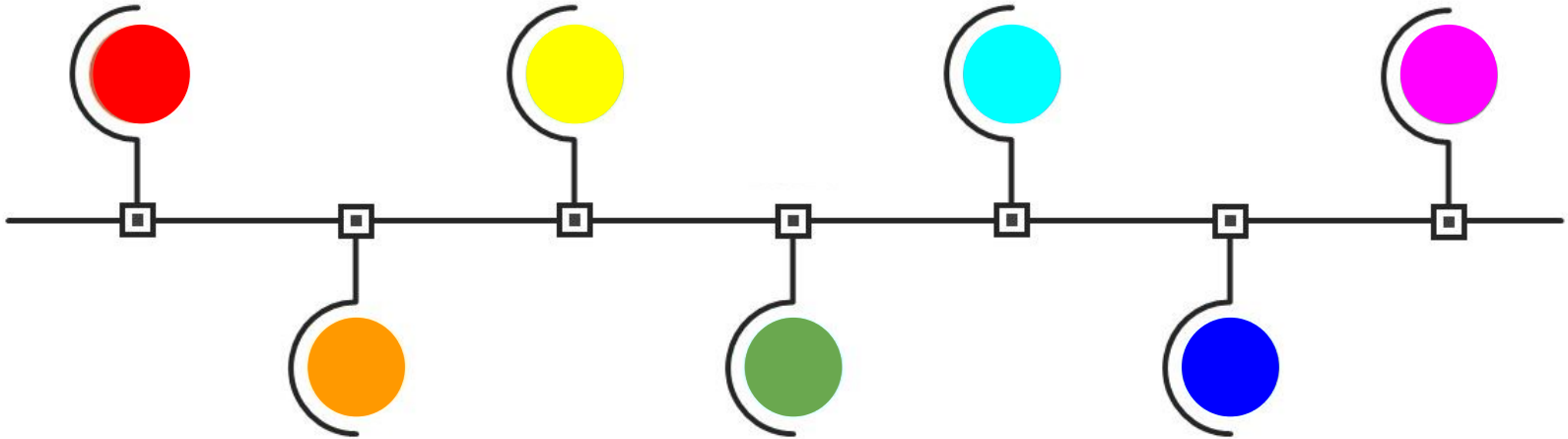
03 ALIGNMENT & TRIMMING



02 ORTHOLOGY INFERENCE

01 DATA

03 ALIGNMENT & TRIMMING



02 ORTHOLOGY INFERENCE

04 PHYLOGENOMIC SUBSAMPLING

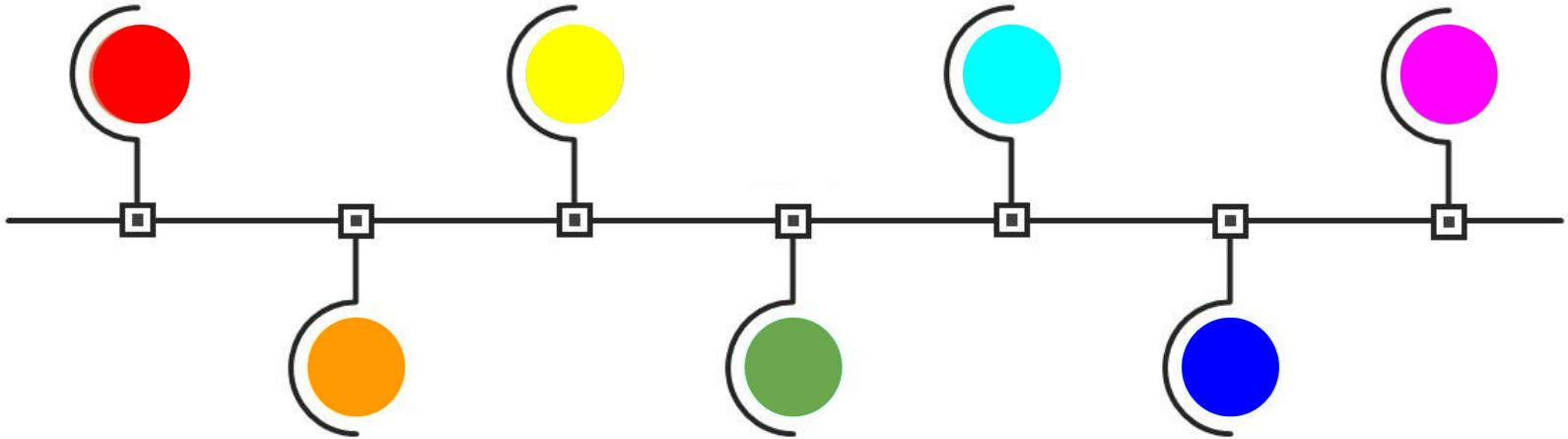
01 DATA

02 ORTHOLOGY INFERENCE

03 ALIGNMENT & TRIMMING

04 PHYLOGENOMIC SUBSAMPLING

05 SUPERMATRIX VS INDIVIDUAL GENES



01 DATA

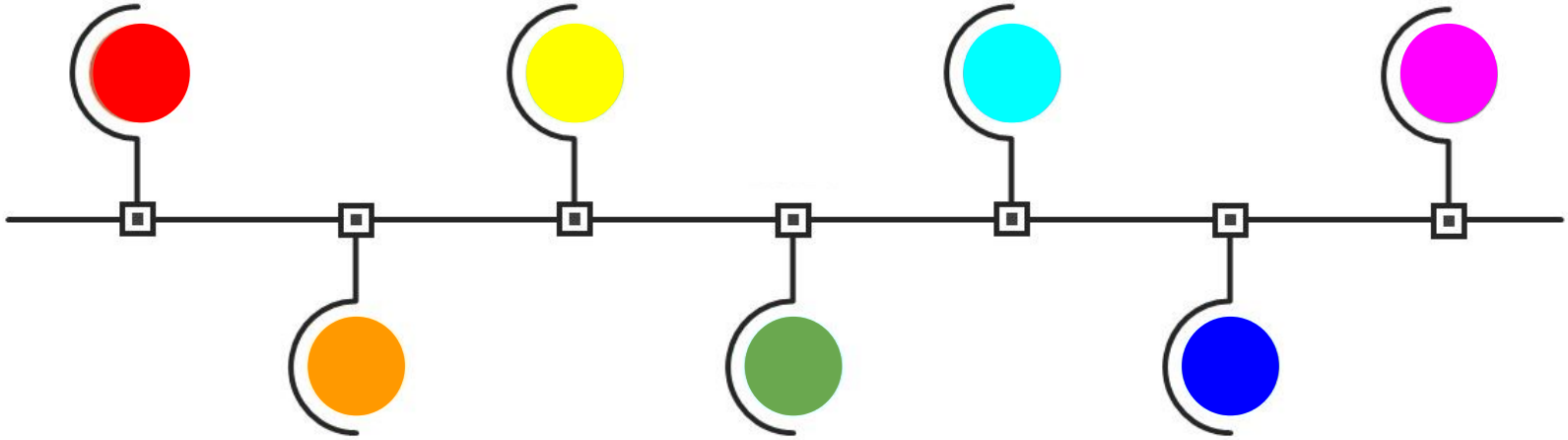
02 ORTHOLOGY INFERENCE

03 ALIGNMENT & TRIMMING

04 PHYLOGENOMIC SUBSAMPLING

05 SUPERMATRIX VS INDIVIDUAL GENES

06 MODEL SELECTION & PHYLOGENETIC INFERENCE



01 DATA

02 ORTHOLOGY INFERENCE

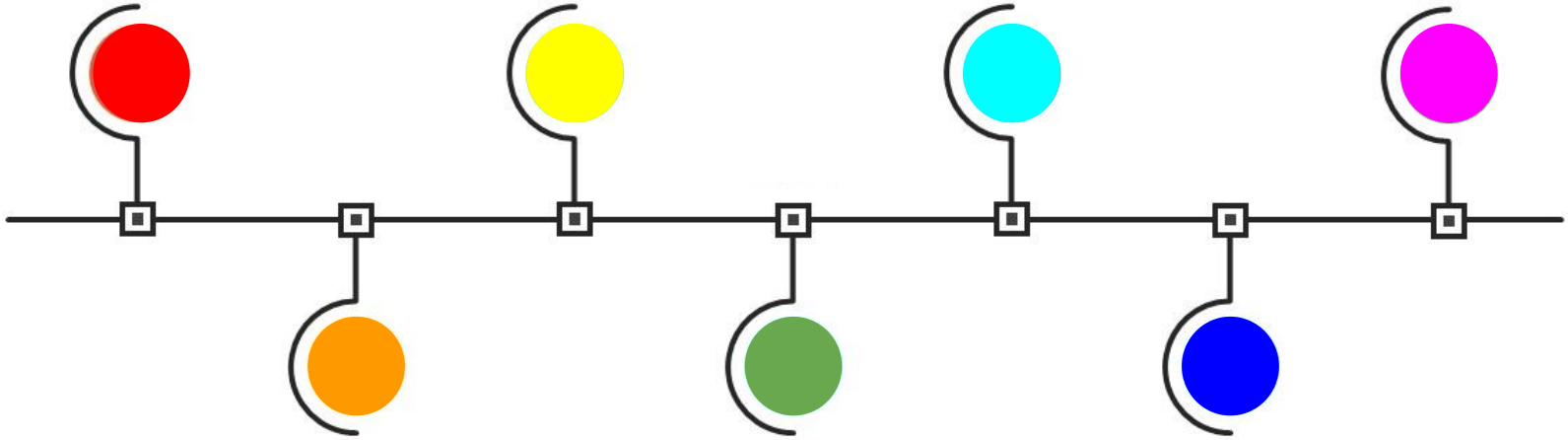
03 ALIGNMENT & TRIMMING

04 PHYLOGENOMIC SUBSAMPLING

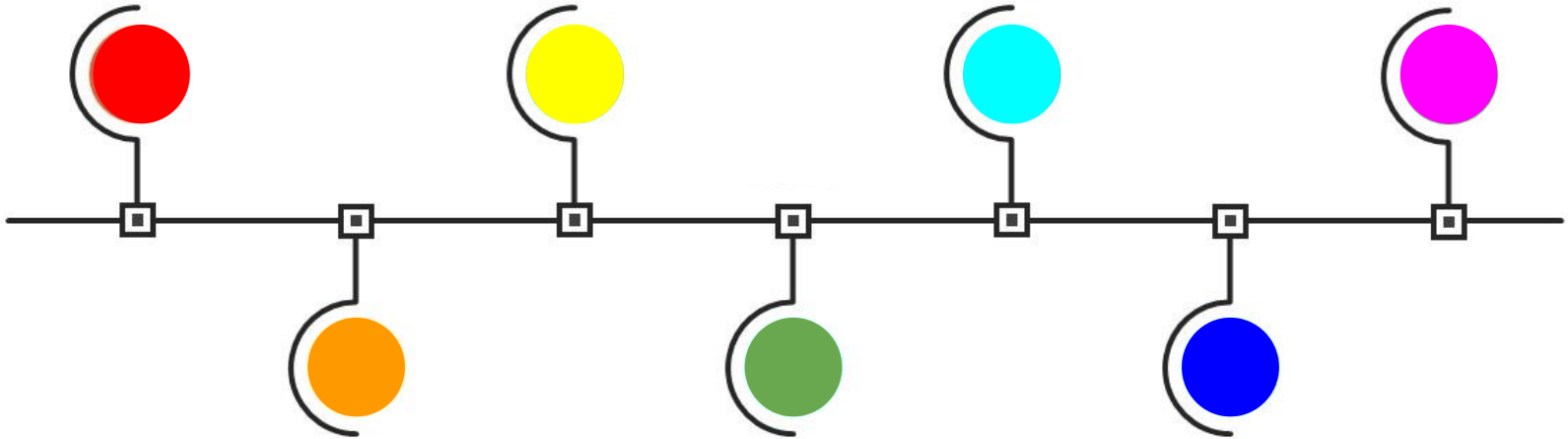
05 SUPERMATRIX VS INDIVIDUAL GENES

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

07 TESTING THE ROBUSTNESS OF YOUR TREE



01 DATA

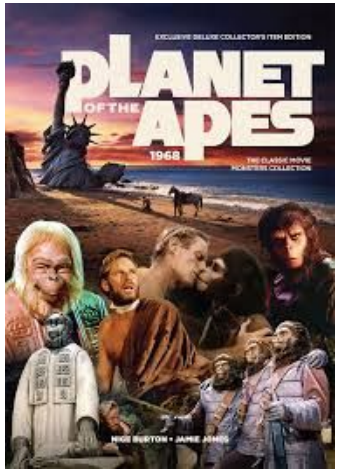


01 DATA

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

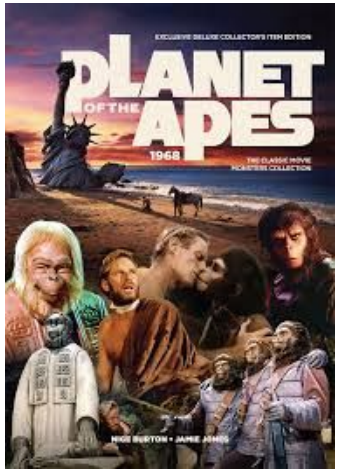
01 DATA

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



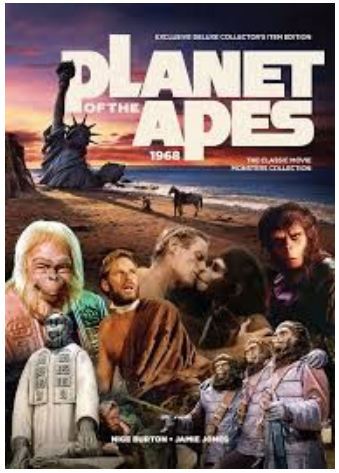
01 DATA

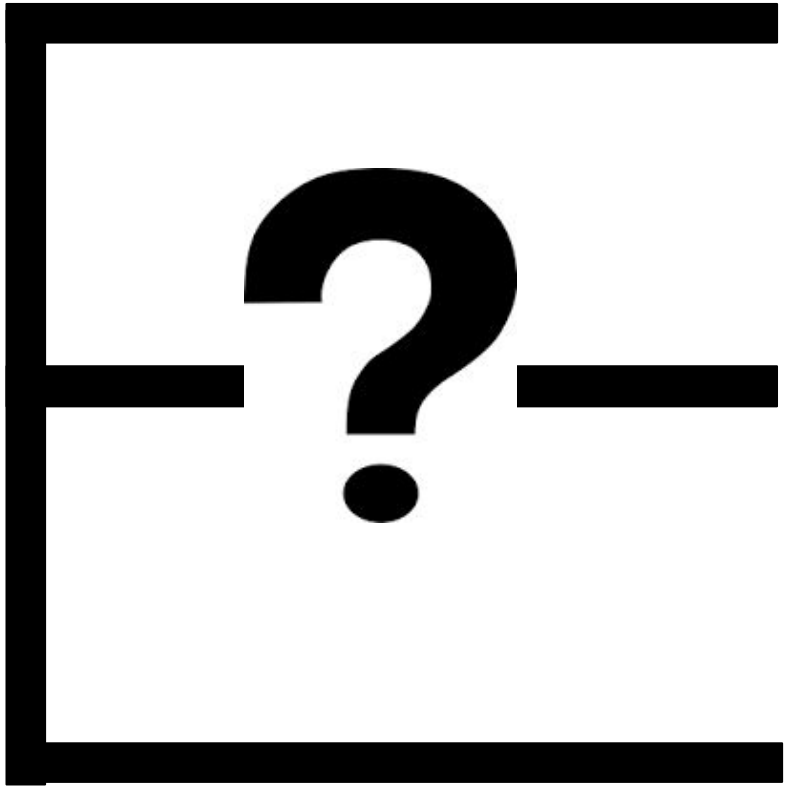
Incomplete, biased, or improper **taxon sampling** can lead to misleading results in reconstructing evolutionary relationships.



01 DATA

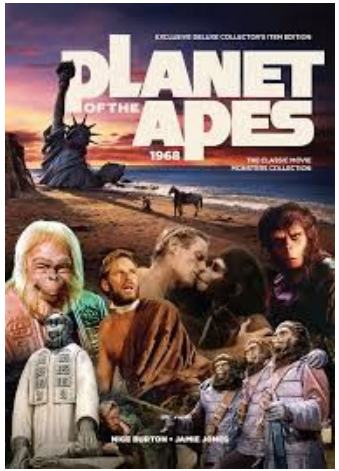
Incomplete, biased, or improper **taxon sampling** can lead to misleading results in reconstructing evolutionary relationships.





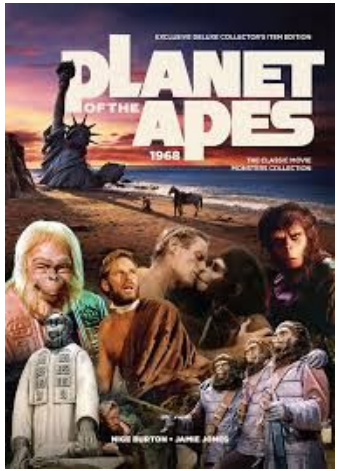
01 DATA

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



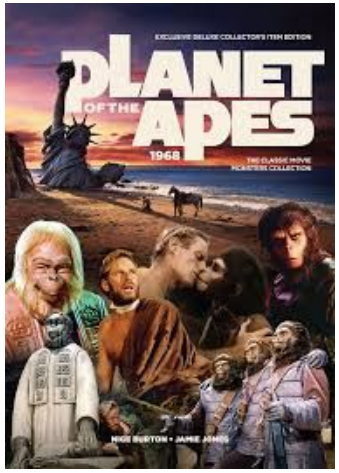
01 DATA

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



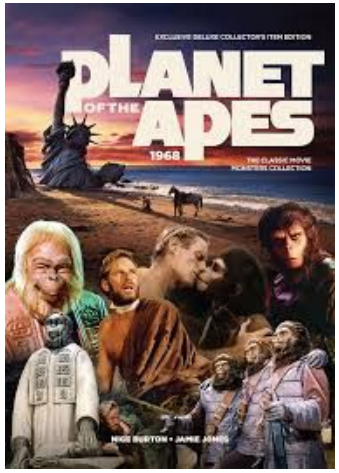
01 DATA

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



01 DATA

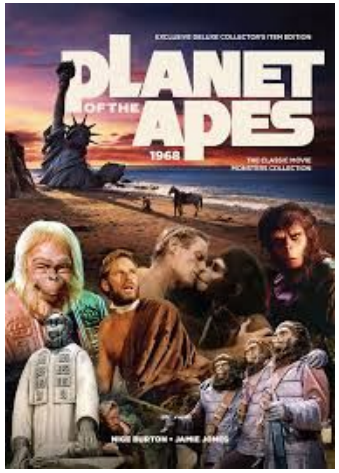
Incomplete, biased, or improper **taxon sampling** can lead to misleading results in reconstructing evolutionary relationships.



[redacted]

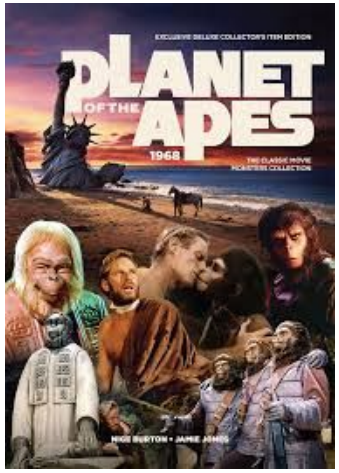
01 DATA

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



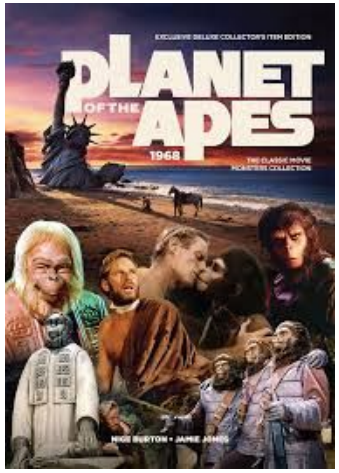
01 DATA

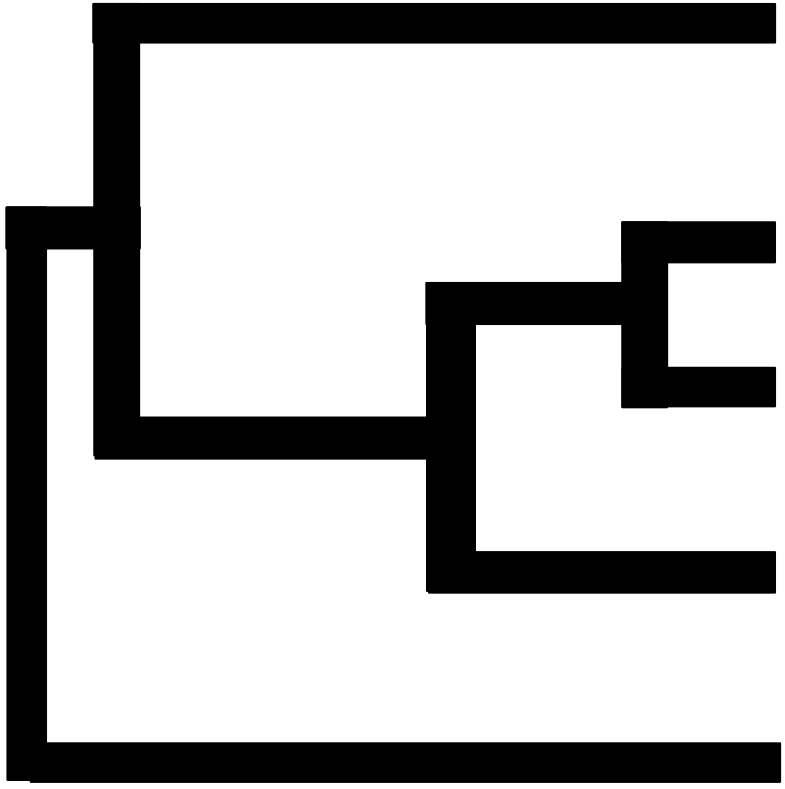
Incomplete, biased, or improper **taxon sampling** can lead to misleading results in reconstructing evolutionary relationships.



01 DATA

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

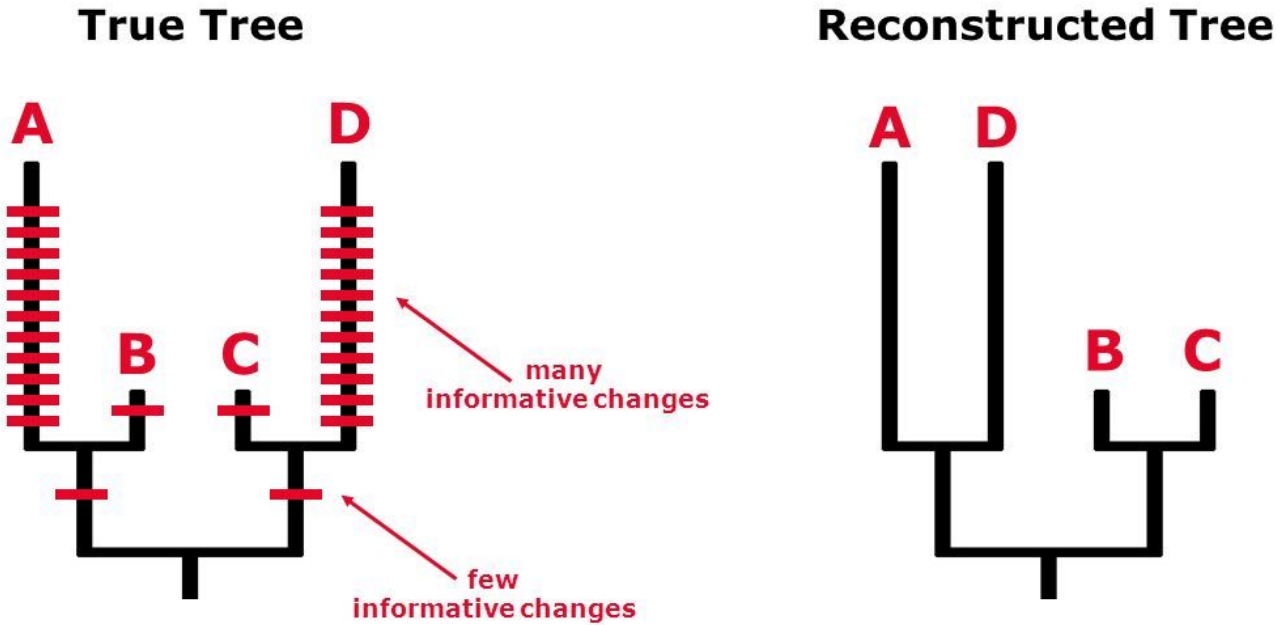




01 DATA

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

Long Branch Attraction

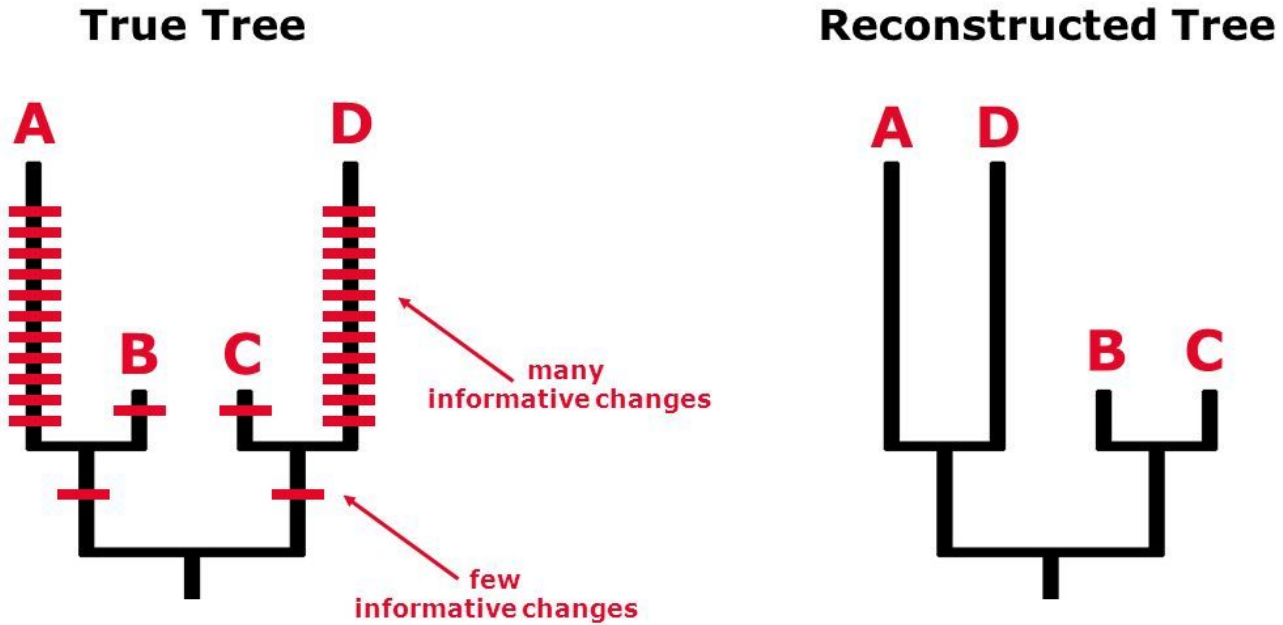


01 DATA

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

Long Branch Attraction

Outgroups / Fast-evolving lineages / Missing data



01 DATA

Source of your data

01 DATA

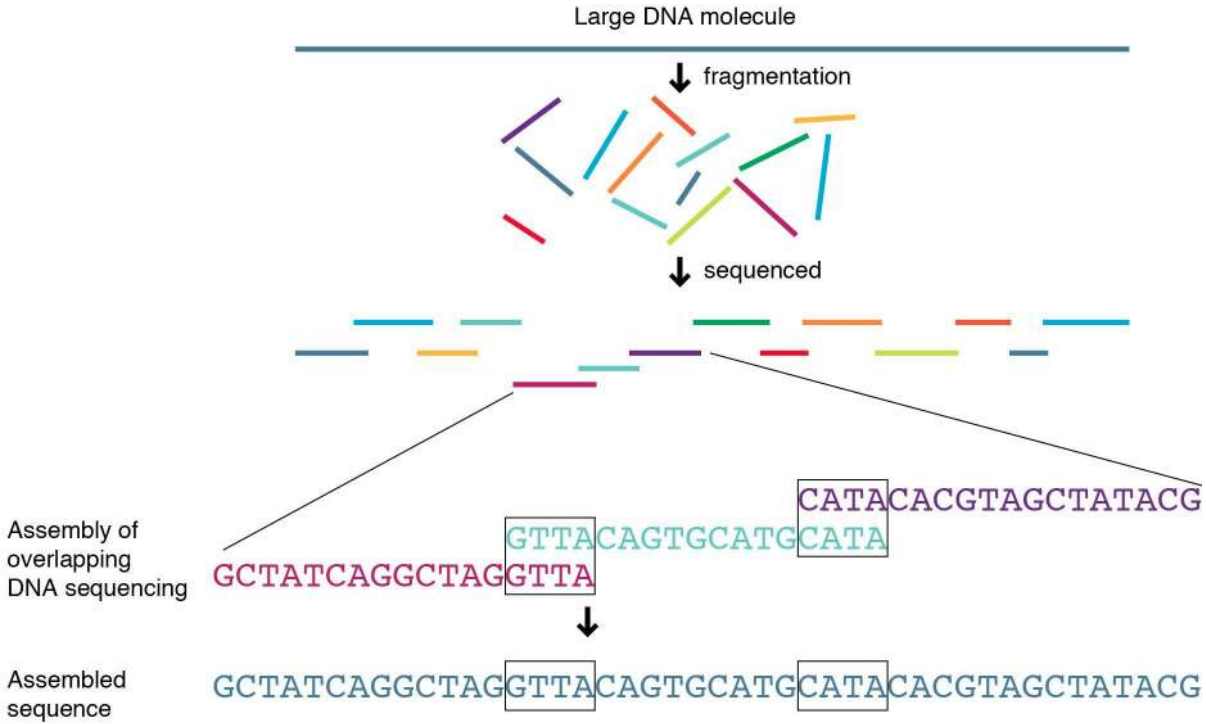
Source of your data

GENOMES

01 DATA

Source of your data

GENOMES



https://knowgenetics.org/whole-genome-sequencing/

- Assembled and annotated.
- Coding genes are retrieved (longest isoform) -> this is your dataset!

01 DATA

Source of your data

GENOMES

Pros:

- Very large set of genetic markers
- Good identification of full-length genes, less chimeras (if the assembly and annotation are of good quality)
- Good for shallow and deep evolutionary distances
- Ethanol-fixed tissue OK (for draft genomes)

01 DATA

Source of your data

GENOMES

Pros:

- Very large set of genetic markers
- Good identification of full-length genes, less chimeras (if the assembly and annotation are of good quality)
- Good for shallow and deep evolutionary distances
- Ethanol-fixed tissue OK (for draft genomes)

Cons:

- Annotation may vary quite a lot between species (source, software, etc), may not be comparable.
- Expensive (money and computing time)
- More difficult to have a high number of species
- Fresh tissue needed (for chromosome-level genomes)

01 DATA

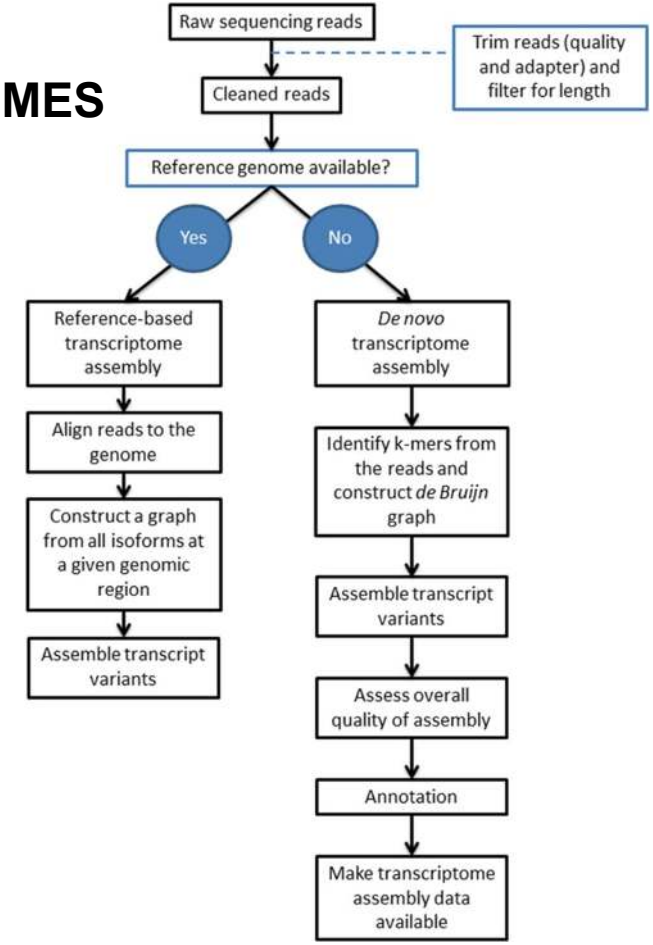
Source of your data

TRANSCRIPTOMES

01 DATA

Source of your data

TRANSCRIPTOMES

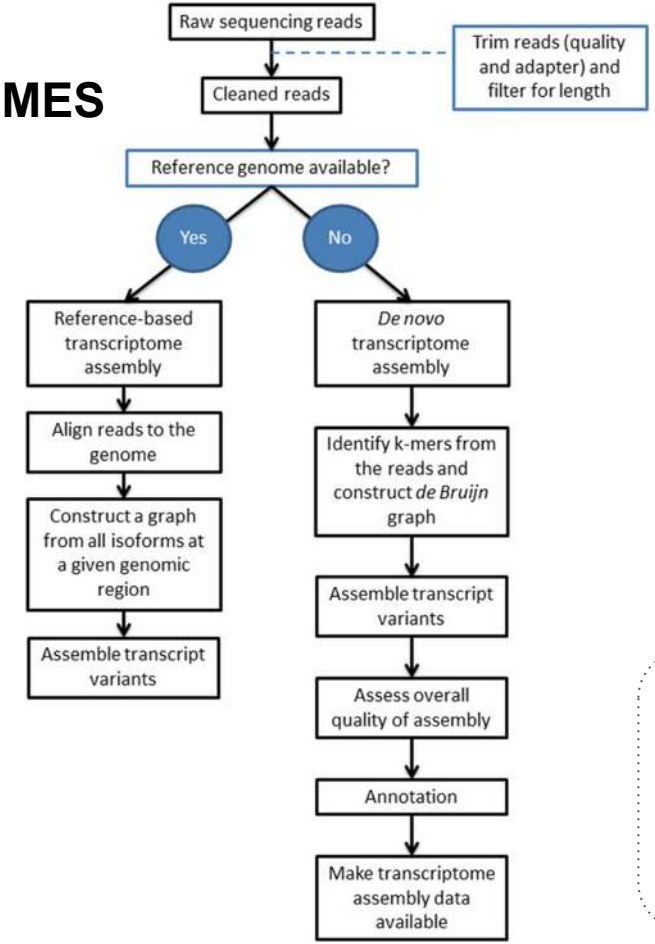


Moreton et al. 2016

01 DATA

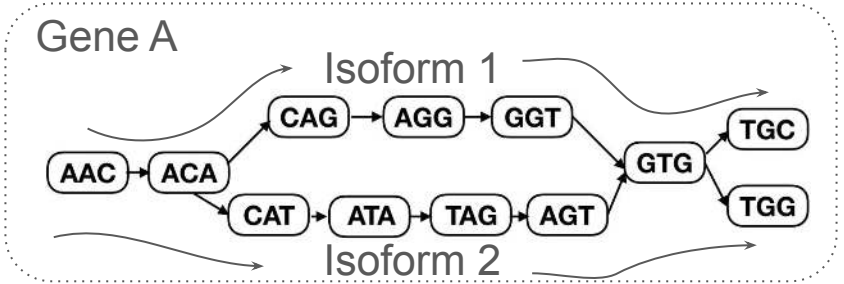
Source of your data

TRANSCRIPTOMES



- Assembled de novo

De Bruijn Graph

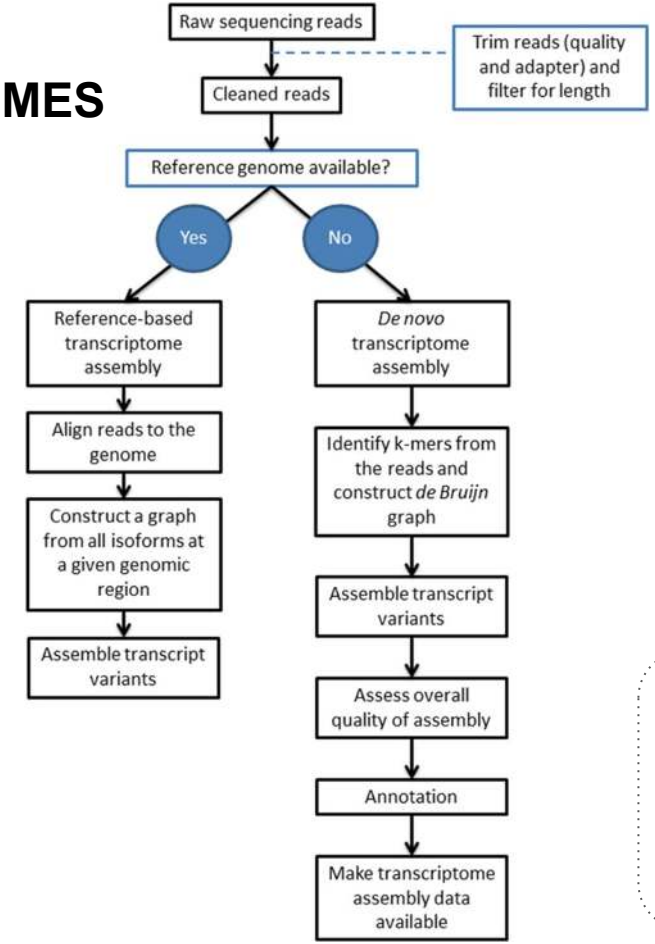


Moreton et al. 2016

01 DATA

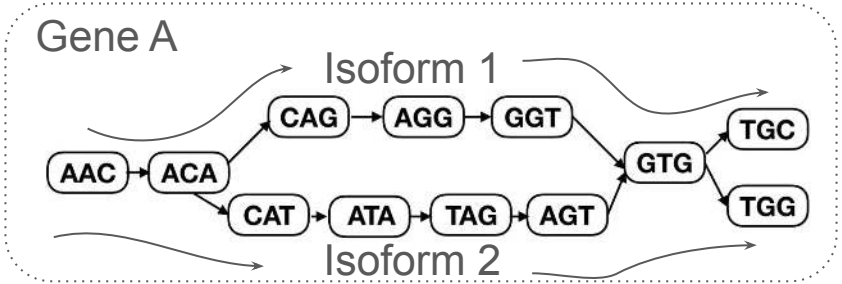
Source of your data

TRANSCRIPTOMES



- Assembled de novo
- Coding genes are retrieved (after inferring ORFs; longest isoform) -> this is your dataset!

De Bruijn Graph

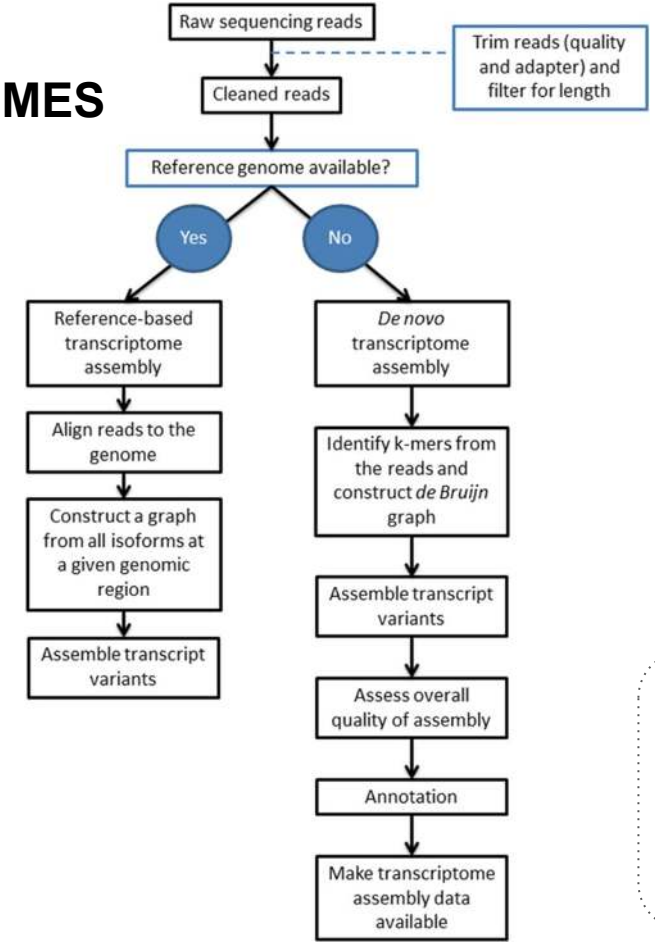


Moreton et al. 2016

01 DATA

Source of your data

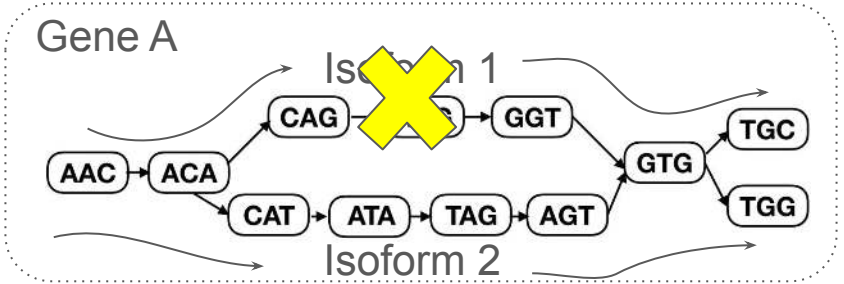
TRANSCRIPTOMES



Moreton et al. 2016

- Assembled de novo
- Coding genes are retrieved (after inferring ORFs; longest isoform) -> this is your dataset!

De Bruijn Graph



01 DATA

Source of your data

TRANSCRIPTOMES

Pros:

- Very large set of genetic markers
- Much cheaper than sequencing genomes -> easier to have a high number of species
- Not dependent upon a reference genome
- Good for shallow and deep evolutionary distances

01 DATA

Source of your data

TRANSCRIPTOMES

Pros:

- Very large set of genetic markers
- Much cheaper than sequencing genomes -> easier to have a high number of species
- Not dependent upon a reference genome
- Good for shallow and deep evolutionary distances

Cons:

- Incomplete identification of full-length genes and single-copy transcripts.
- Potential misassembly of transcripts (especially when duplicates are present)
- Missing data as a product of the transcriptome representing a snapshot of expression (but this could also affect genome annotation)
- Fresh tissue needed

01 DATA

Source of your data

ULTRACONSERVED ELEMENTS (UCEs)

01 DATA

Source of your data

ULTRACONSERVED ELEMENTS (UCEs)

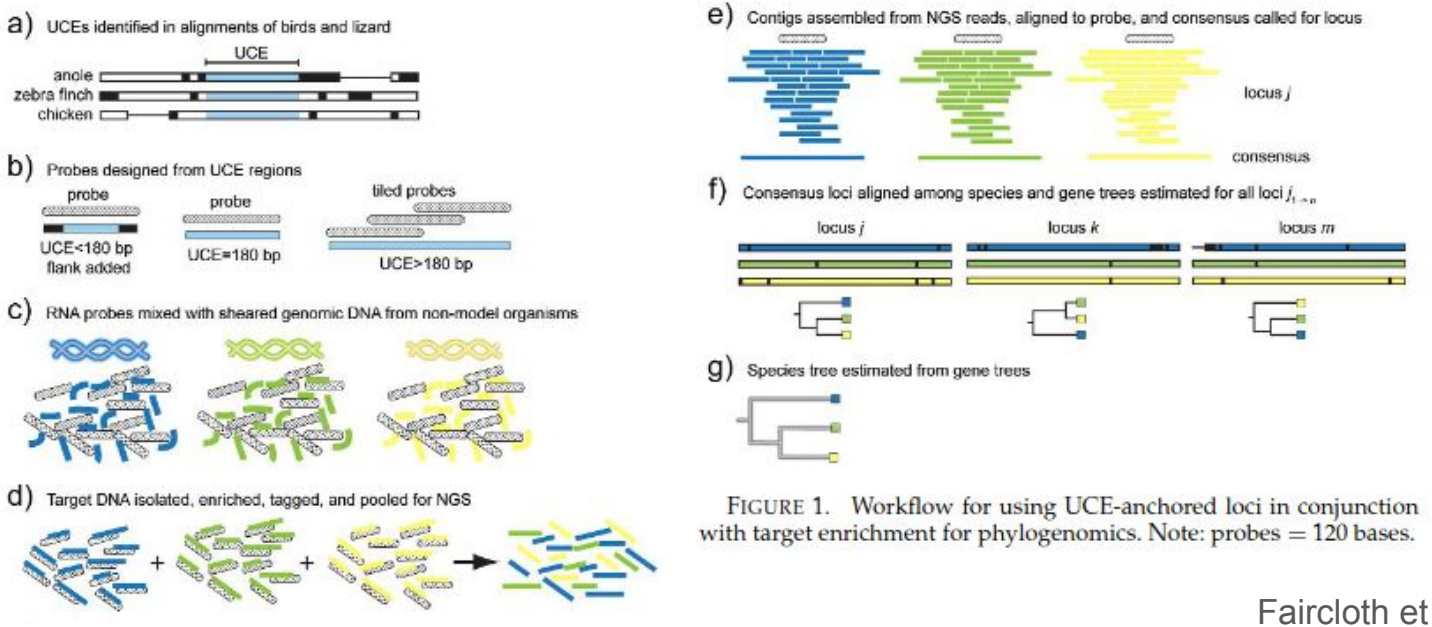


FIGURE 1. Workflow for using UCE-anchored loci in conjunction with target enrichment for phylogenomics. Note: probes = 120 bases.

Faircloth et al. 2012

The UCEs are designed a priori -> after hybridization, sequencing, assembly and mapping, this is your data!

01 DATA

Source of your data

ULTRACONSERVED ELEMENTS (UCEs)

Pros:

- Medium-large set of genetic markers
- Much cheaper than sequencing genomes -> easier to have a high number of species
- Not dependent upon a reference genome
- Tissues fixed in EtOH or museum specimens are OK

01 DATA

Source of your data

(Lisa Pokorny's talk on 31st Jan)

ULTRACONSERVED ELEMENTS (UCEs)

Pros:

- Medium-large set of genetic markers
- Much cheaper than sequencing genomes -> easier to have a high number of species
- Not dependent upon a reference genome
- Tissues fixed in EtOH or museum specimens are OK

Cons:

- Limited availability of markes outside the designed ones.
- Potential misassembly (if probes are designed with a limited amount of species)
- Retrieval success dependent on DNA quality
- Usefulness of markers known a posteriori
- No proper orthology inference

01 DATA

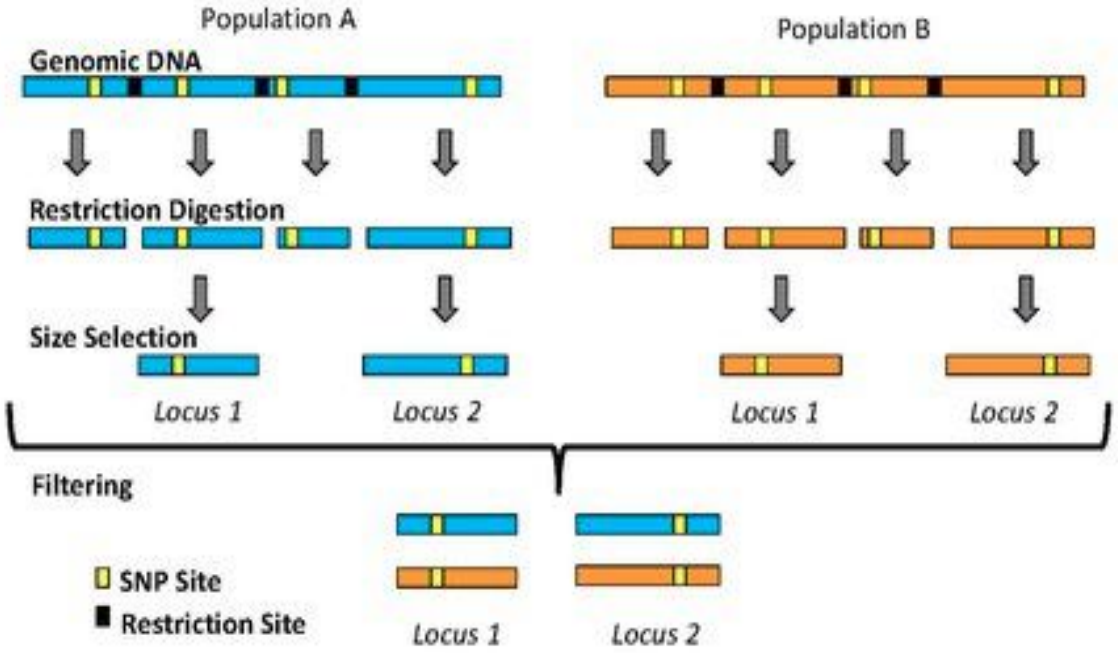
Source of your data

REDUCED REPRESENTATION (RADseq, GBS)

01 DATA

Source of your data

REDUCED REPRESENTATION (RADseq, GBS)



After digestion, sequencing and mapping, this is your data!

01 DATA

Source of your data

REDUCED REPRESENTATION (RADseq, GBS)

Pros:

- The cheapest of the methods
- Not dependent upon a reference genome
- Samples fixed in ethanol OK
- Markers distributed evenly across the genome

01 DATA

Source of your data

REDUCED REPRESENTATION (RADseq, GBS)

Pros:

- The cheapest of the methods
- Not dependent upon a reference genome
- Samples fixed in ethanol OK
- Markers distributed evenly across the genome

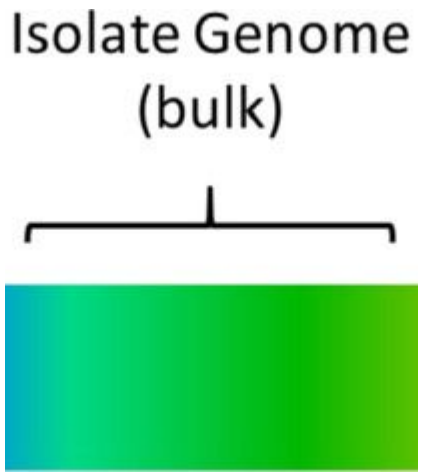
Cons:

- No full genes, only SNPs
- Only for population genomics or phylogeny including closely-related species
- Missing data as a product of the transcriptome representing a snapshot of expression (but this could also affect genome annotation)
- No proper orthology inference

01 DATA

Source of your data

METAGENOMICS/METATRANSCRIPTOMICS

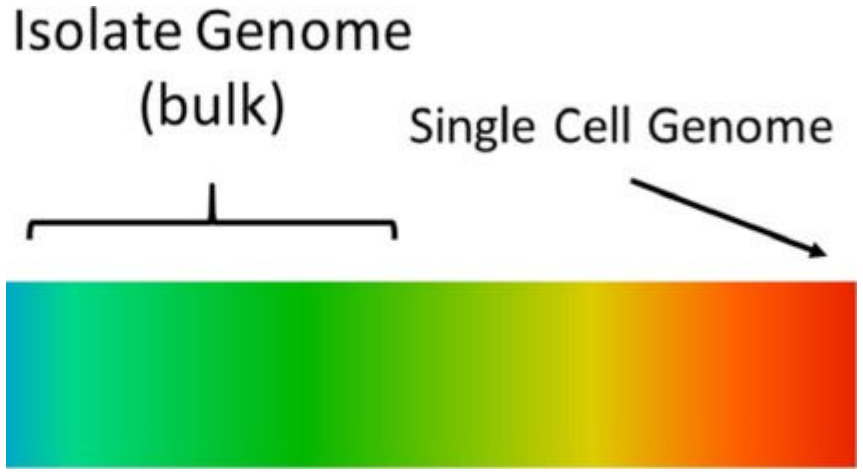


One individual, multiple cells

01 DATA

Source of your data

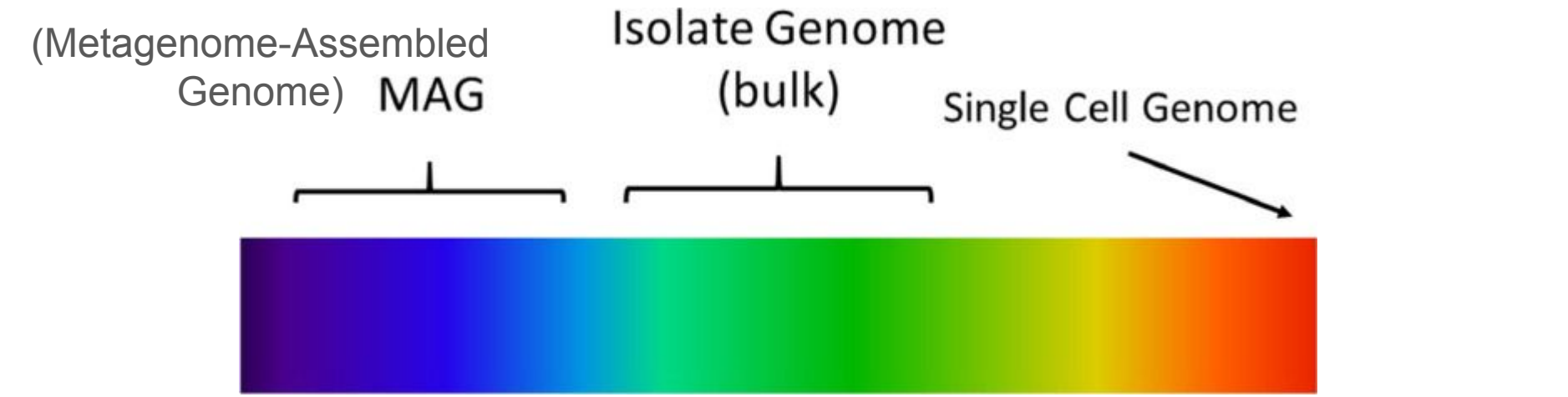
METAGENOMICS/METATRANSCRIPTOMICS



01 DATA

Source of your data

METAGENOMICS/METATRANSCRIPTOMICS



01 DATA

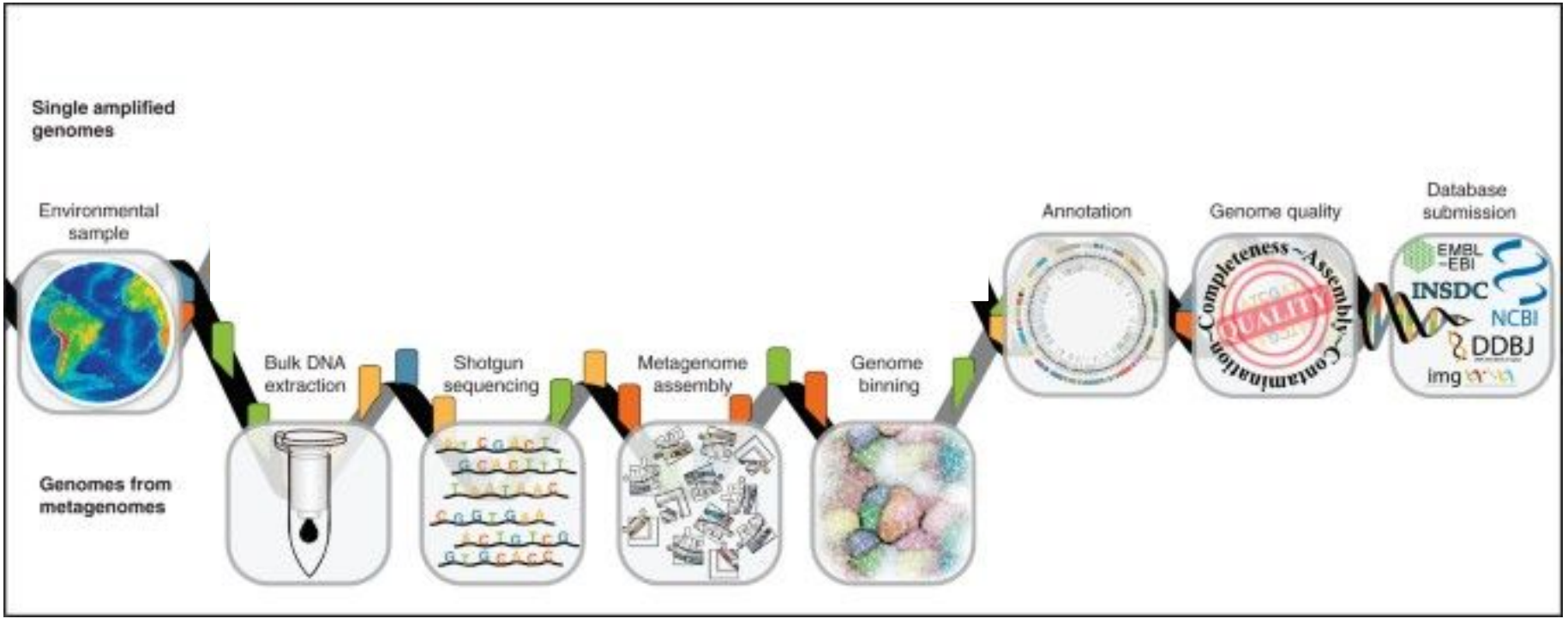
Source of your data

METAGENOMICS - single cell

01 DATA

Source of your data

METAGENOMICS - single cell

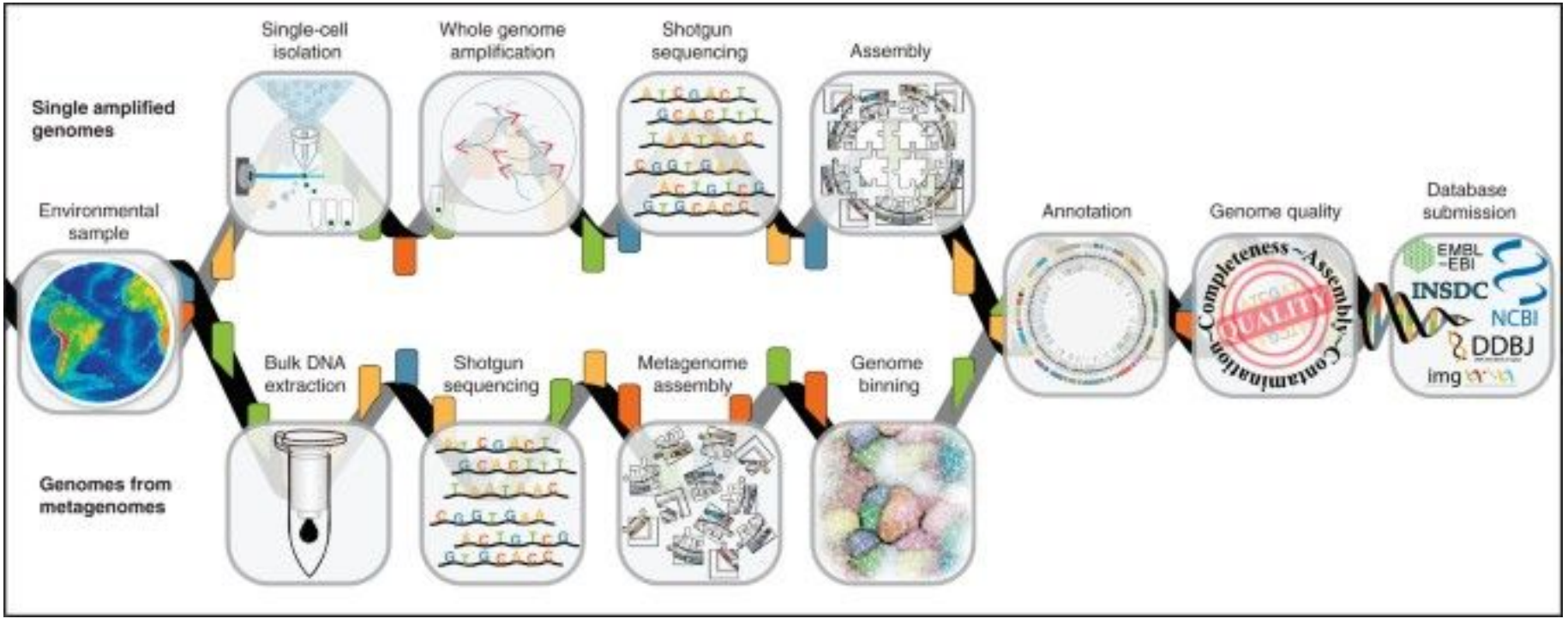


Bowers et al. 2017

01 DATA

Source of your data

METAGENOMICS - single cell vs MAGs

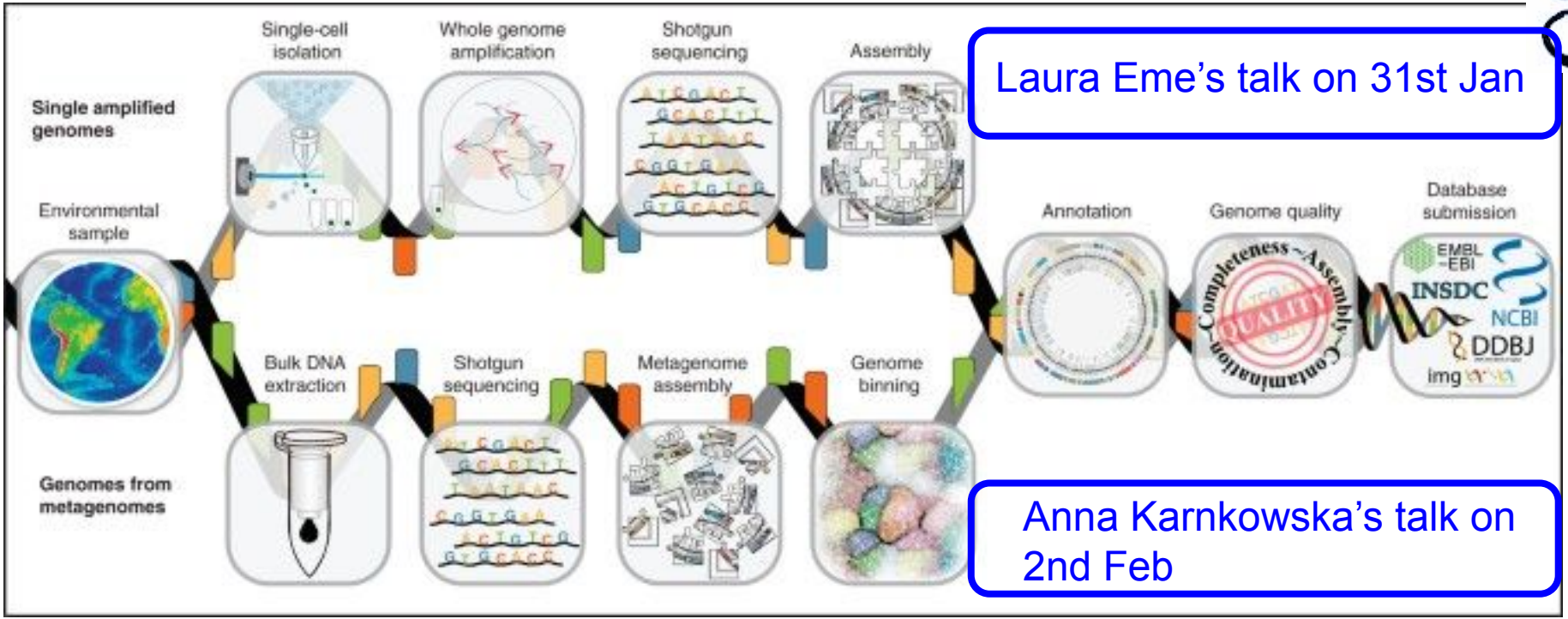


Bowers et al. 2017

01 DATA

Source of your data

METAGENOMICS - single cell vs MAGs

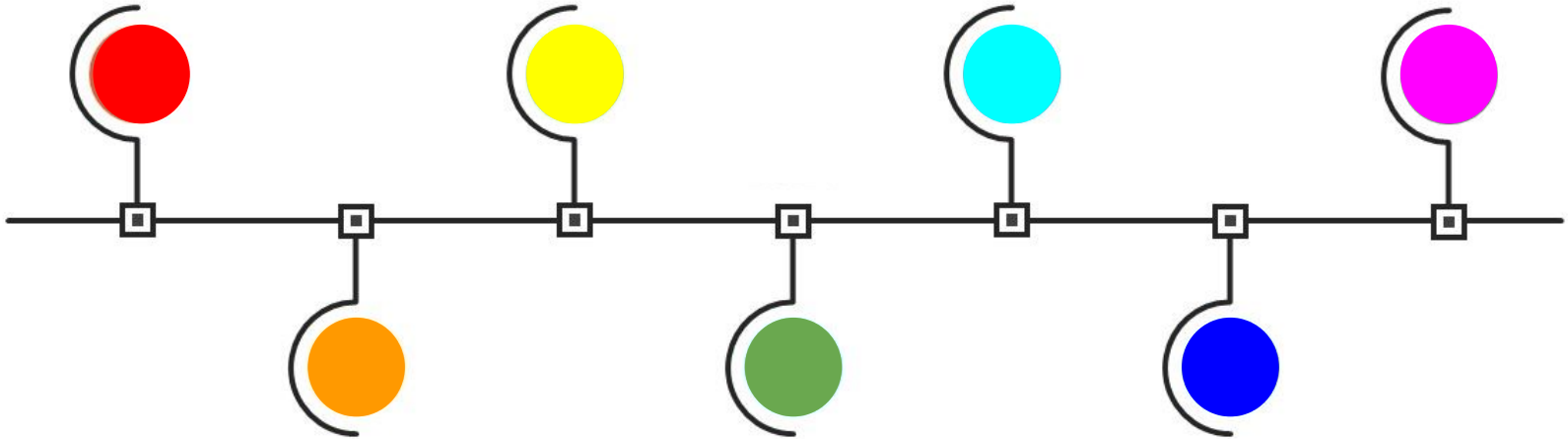


Laura Eme's talk on 31st Jan

Anna Karnkowska's talk on 2nd Feb

Bowers et al. 2017

01 DATA



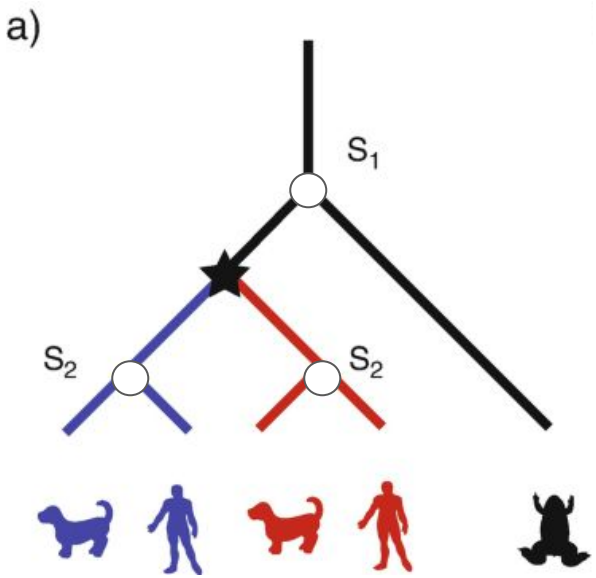
02 ORTHOLOGY INFERENCE

02 ORTHOLOGY INFERENCE

02 ORTHOLOGY INFERENCE

Definitions

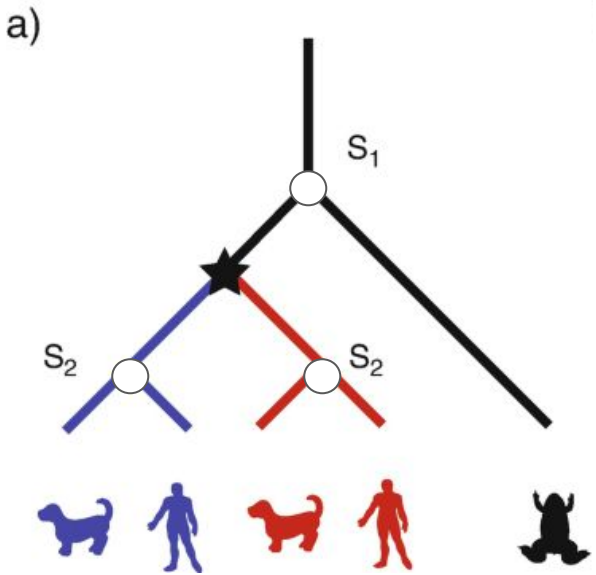
- Two genes are **orthologs** if their MRCA is a ***speciation***:○



02 ORTHOLOGY INFERENCE

Definitions

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

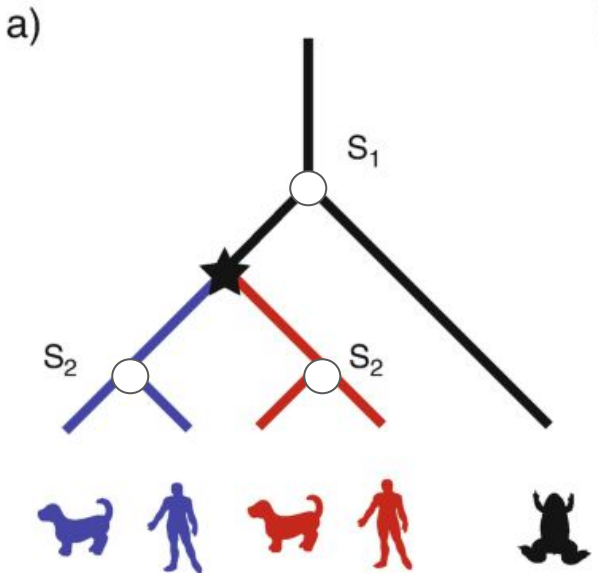


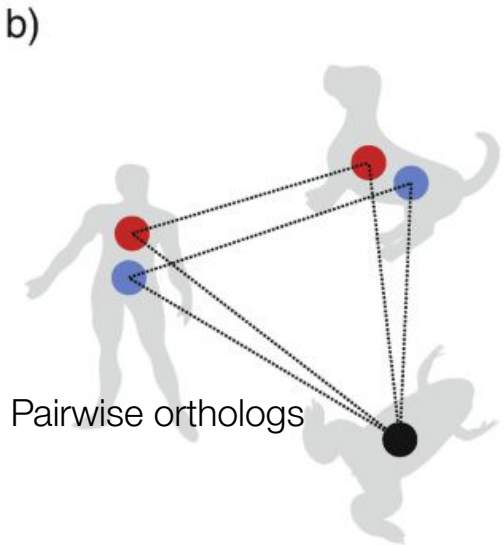
02 ORTHOLOGY INFERENCE

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*





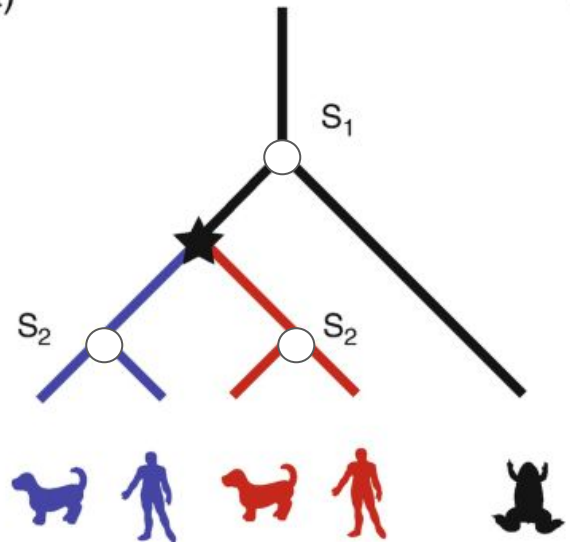
Altenhoff, Glover & Dessimoz 2019

02 ORTHOLOGY INFERENCE

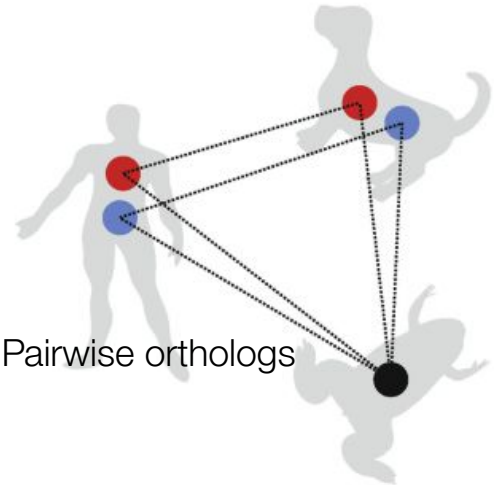
Definitions

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

a)



b)



Pairwise orthologs

Orthology relationships are inferred *pairwise*

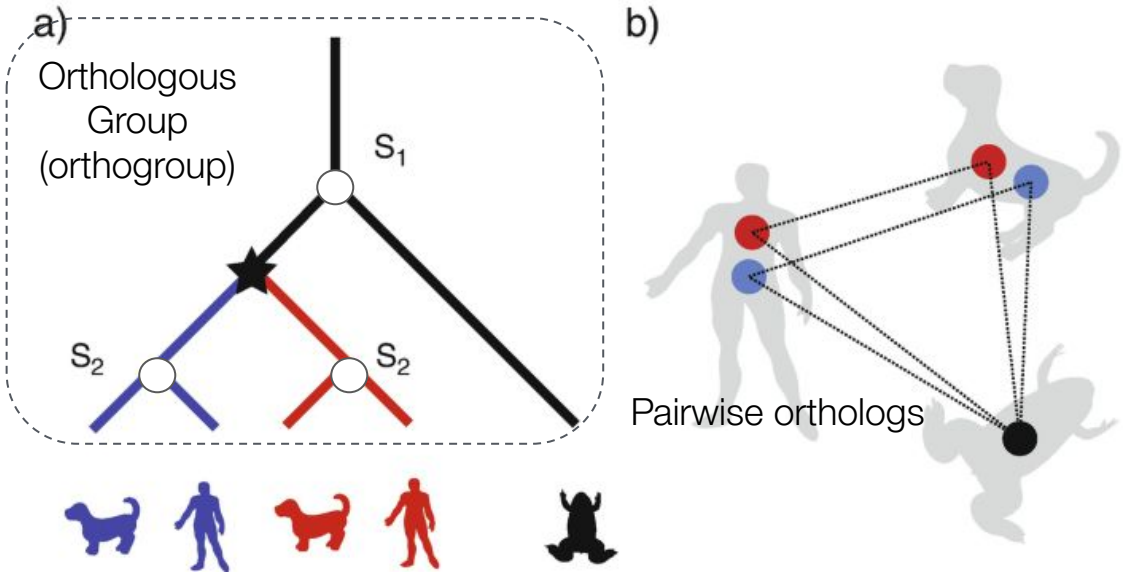
When we have multiple species, we should consider the concept of *orthogroup*

Altenhoff, Glover & Dessimoz 2019

02 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*

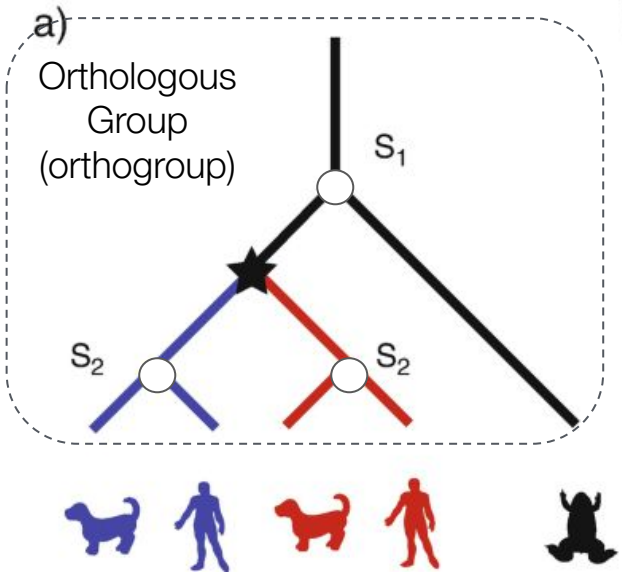
When we have multiple species, we should consider the concept of *orthogroup*

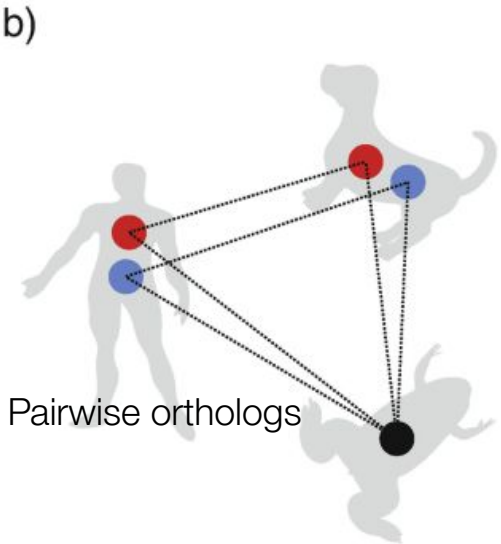
Altenhoff, Glover & Dessimoz 2019

02 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*

When we have multiple species, we should consider the concept of *orthogroup*

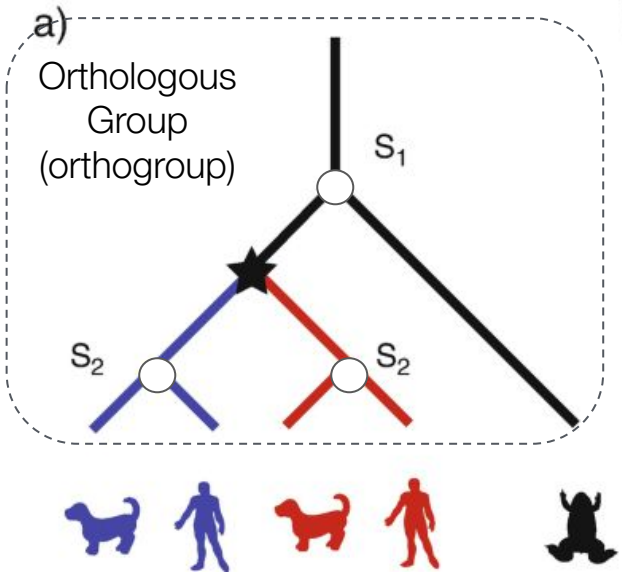
Orthology inference is essential for phylogenomics, as you want to consider only genes that arouse through speciation events

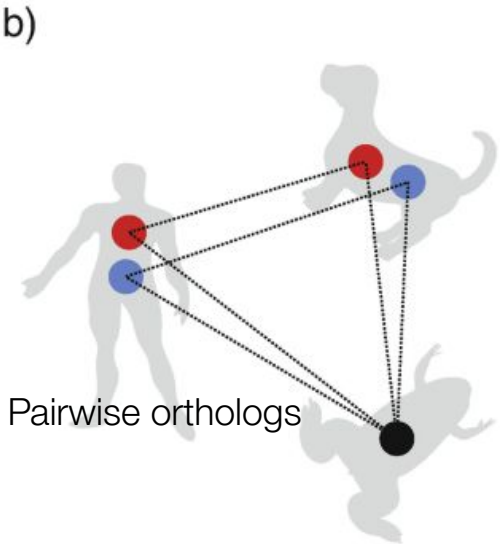
Altenhoff, Glover & Dessimoz 2019

02 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*

When we have multiple species, we should consider the concept of *orthogroup*

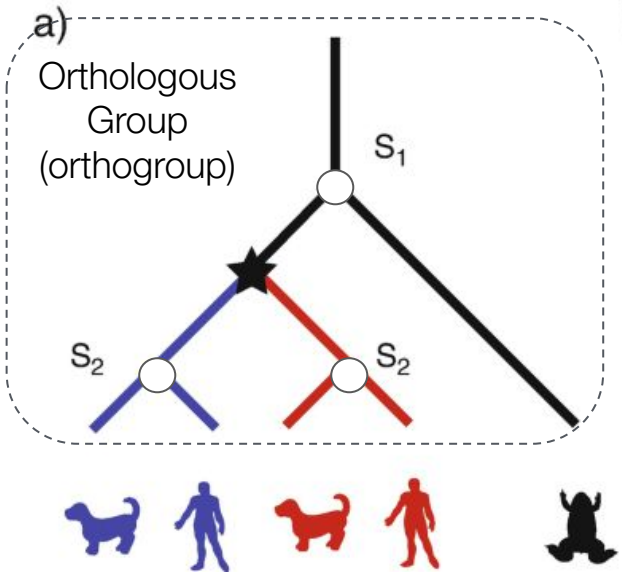
For phylogenomic inference, we want either:

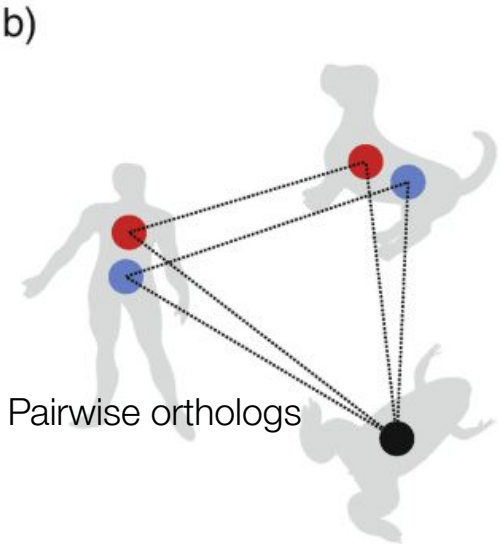
Altenhoff, Glover & Dessimoz 2019

02 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*

When we have multiple species, we should consider the concept of *orthogroup*

For phylogenomic inference, we want either:

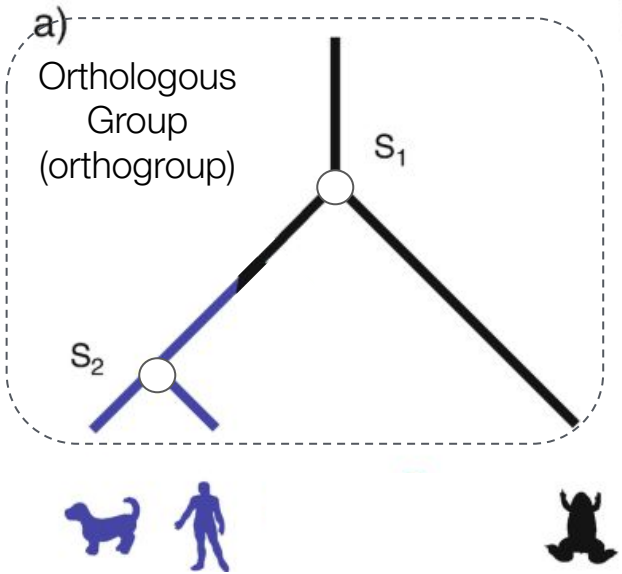
- Single-copy orthogroups (ie, one gene per species)

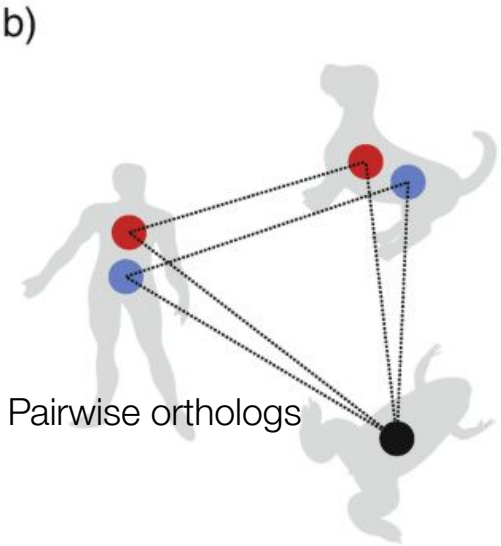
Altenhoff, Glover & Dessimoz 2019

02 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*

When we have multiple species, we should consider the concept of *orthogroup*

For phylogenomic inference, we want either:

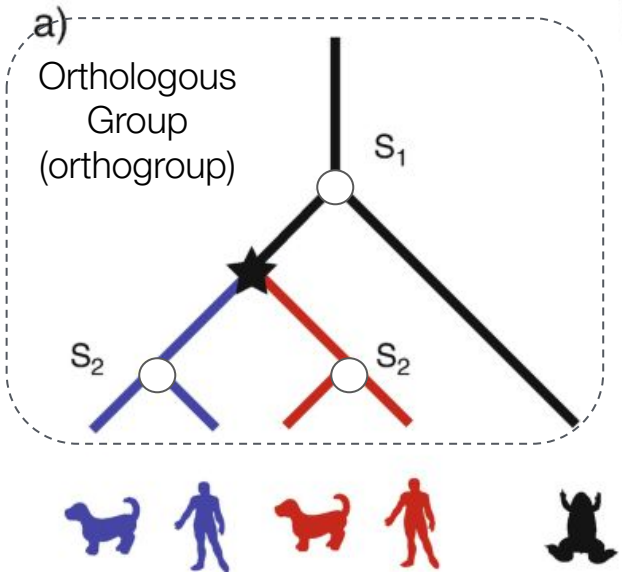
- Single-copy orthogroups (ie, one gene per species)

Altenhoff, Glover & Dessimoz 2019

02 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*

When we have multiple species, we should consider the concept of *orthogroup*

For phylogenomic inference, we want either:

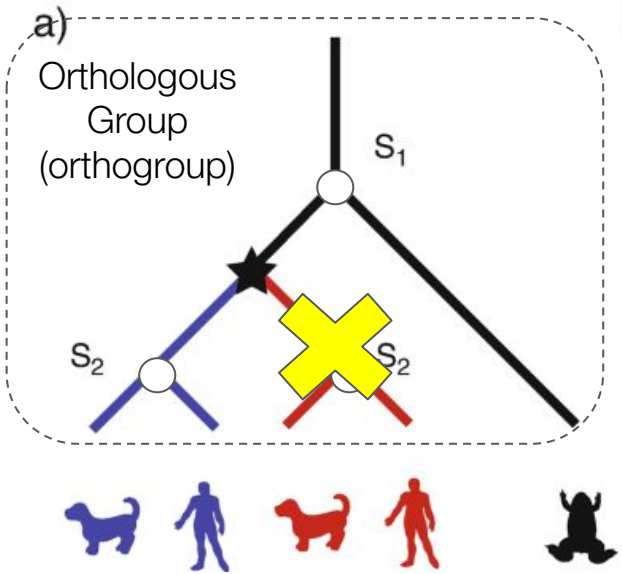
- Single-copy orthogroups (ie, one gene per species)
- Trimmed orthogroups (ie, removing genes from duplication events)

Altenhoff, Glover & Dessimoz 2019

02 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*

When we have multiple species, we should consider the concept of *orthogroup*

For phylogenomic inference, we want either:

- Single-copy orthogroups (ie, one gene per species)
- Trimmed orthogroups (ie, removing genes from duplication events)

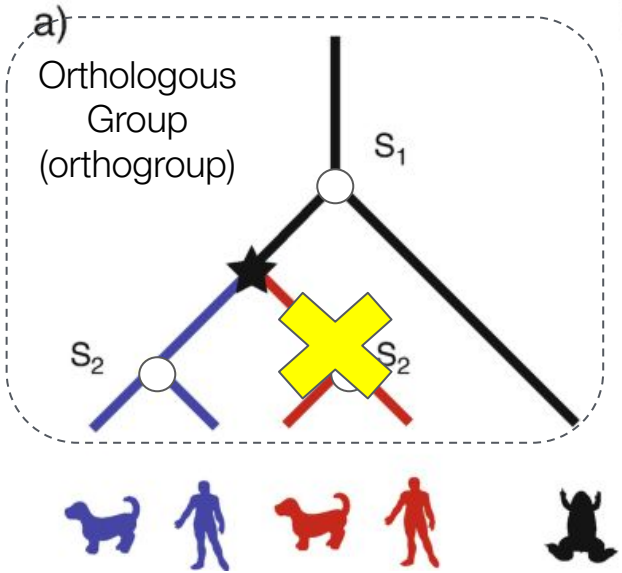
Altenhoff, Glover & Dessimoz 2019

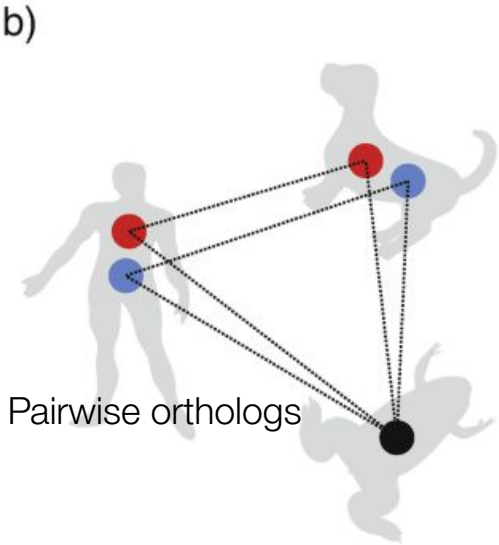
02 ORTHOLOGY INFERENCE

Marina Marcet-Houben tomorrow

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*

When we have multiple species, we should consider the concept of *orthogroup*

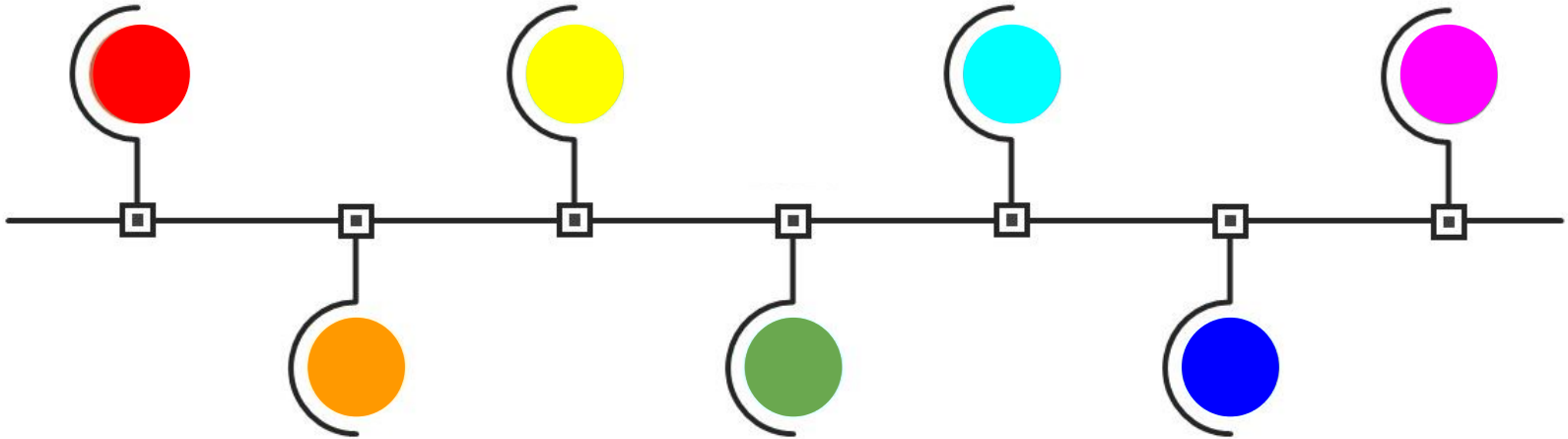
For phylogenomic inference, we want either:

- Single-copy orthogroups (ie, one gene per species)
- Trimmed orthogroups (ie, removing genes from duplication events)

Altenhoff, Glover & Dessimoz 2019

01 DATA

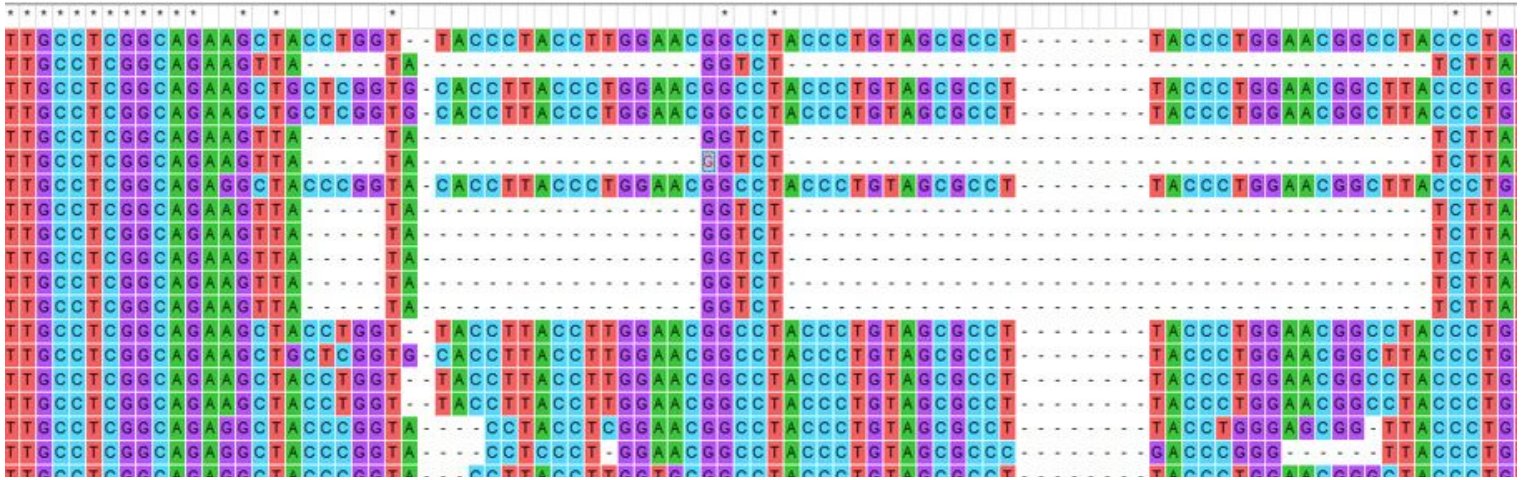
03 ALIGNMENT & TRIMMING



02 ORTHOLOGY INFERENCE

03 ALIGNMENT AND TRIMMING

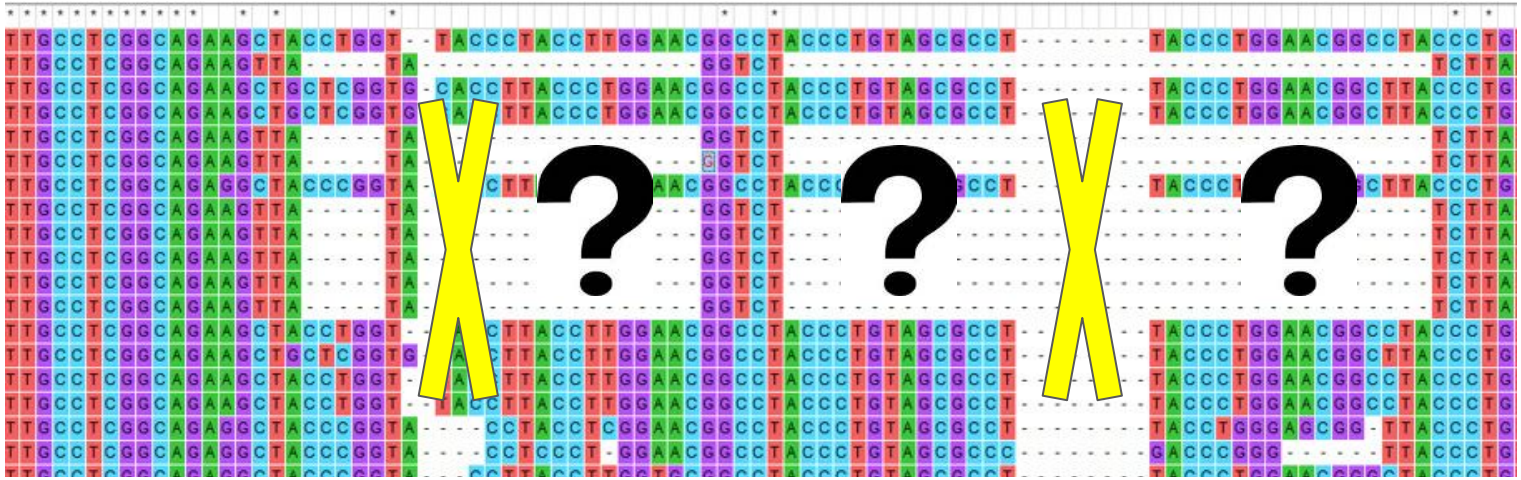
The goal of the alignment procedure should be to identify the events associated with the homologies, so that the aligned sequences accurately reflect those events.



03 ALIGNMENT AND TRIMMING

The goal of the alignment procedure should be to identify the events associated with the homologies, so that the aligned sequences accurately reflect those events.

If the sequences are poorly aligned, you may want to consider trimming the poorly aligned areas.

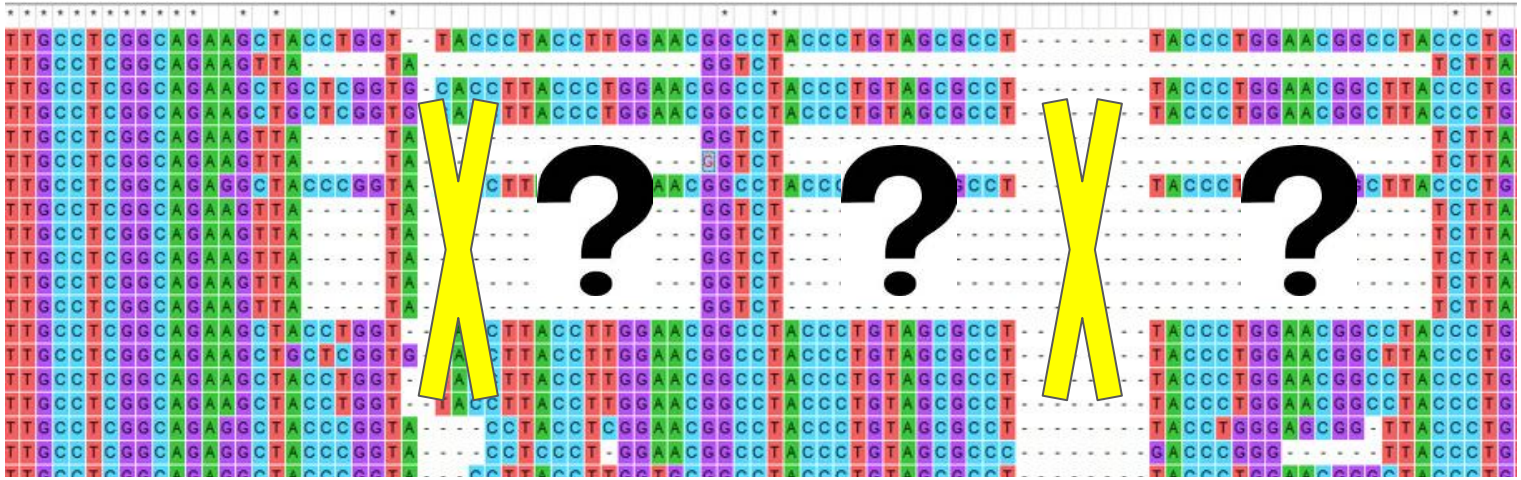


03 ALIGNMENT AND TRIMMING

Jacob and Marina today

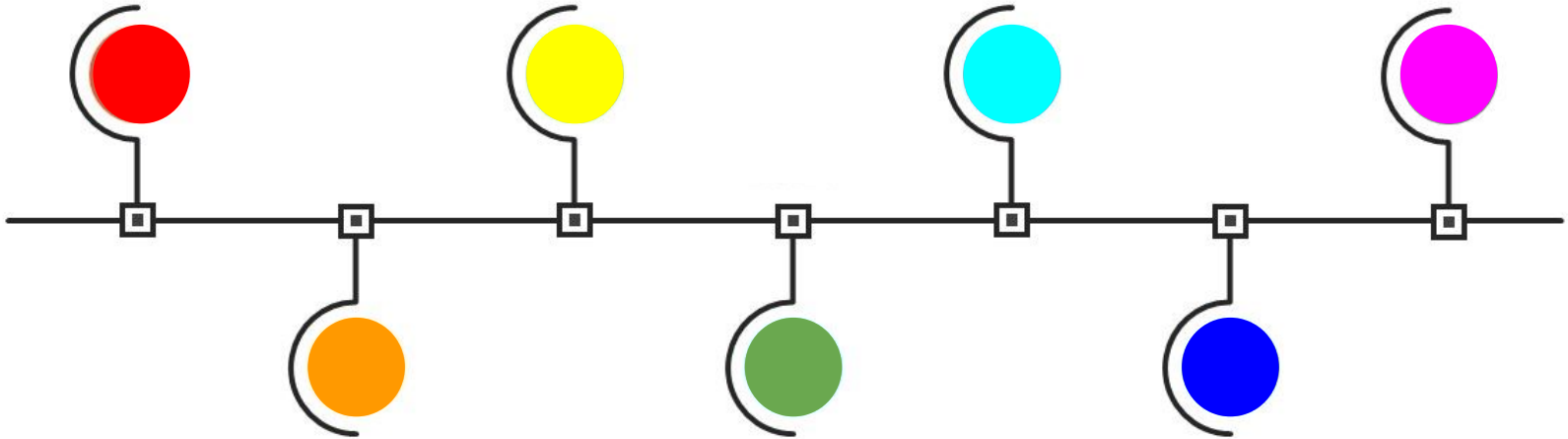
The goal of the alignment procedure should be to identify the events associated with the homologies, so that the aligned sequences accurately reflect those events.

If the sequences are poorly aligned, you may want to consider trimming the poorly aligned areas.



01 DATA

03 ALIGNMENT & TRIMMING



02 ORTHOLOGY INFERENCE

04 PHYLOGENOMIC SUBSAMPLING

04 PHYLOGENOMIC SUBSAMPLING

04 PHYLOGENOMIC SUBSAMPLING

What? Sets of loci are selected from large genome-scale data sets and used for phylogenetic inference.

04 PHYLOGENOMIC SUBSAMPLING

What? Sets of loci are selected from large genome-scale data sets and used for phylogenetic inference.

Why? To avoid an accumulation of nonphylogenetic signals as a product of heterogeneities in evolutionary processes, reduce computing time and improve model fit.

This step can be used to *explore phylogenetic conflicts*, *test specific hypotheses* of relationships, measure the impact of *different sources of bias*, and allow for a *better modeling* of evolutionary processes.

04 PHYLOGENOMIC SUBSAMPLING

What? Sets of loci are selected from large genome-scale data sets and used for phylogenetic inference.

Why? To avoid an accumulation of nonphylogenetic signals as a product of heterogeneities in evolutionary processes, reduce computing time and improve model fit.

This step can be used to *explore phylogenetic conflicts*, *test specific hypotheses* of relationships, measure the impact of *different sources of bias*, and allow for a *better modeling* of evolutionary processes.

How? By checking the properties of genes or sites and selecting the ones that minimize bias.

04 PHYLOGENOMIC SUBSAMPLING

Which properties?

04 PHYLOGENOMIC SUBSAMPLING

Which properties?

Information content

-> length of alignment
-> missing data
-> level of occupancy

04 PHYLOGENOMIC SUBSAMPLING

Which properties?

Information content

-> length of alignment
-> missing data
-> level of occupancy

Phylogenetic signal

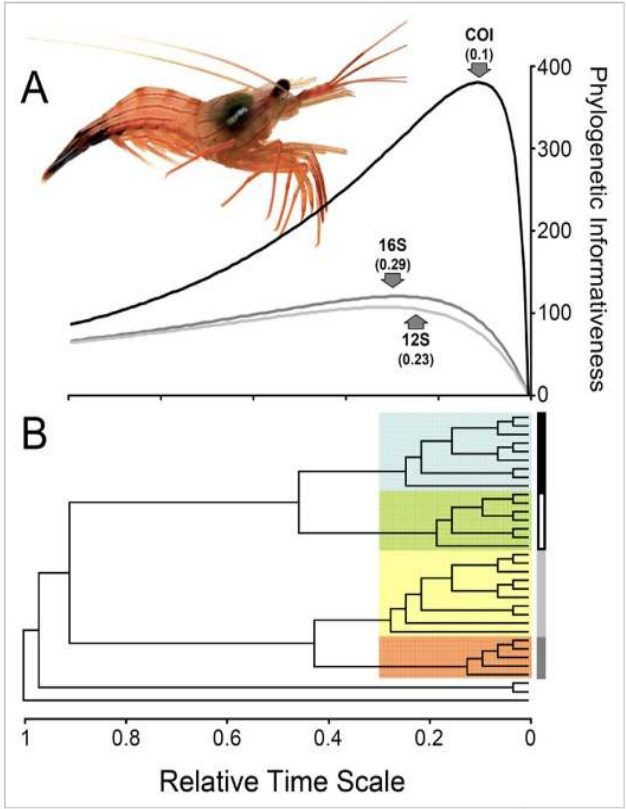
04 PHYLOGENOMIC SUBSAMPLING

Which properties?

Information content

-> length of alignment
-> missing data
-> level of occupancy

Phylogenetic signal



Baeza & Fuentes 2013

04 PHYLOGENOMIC SUBSAMPLING

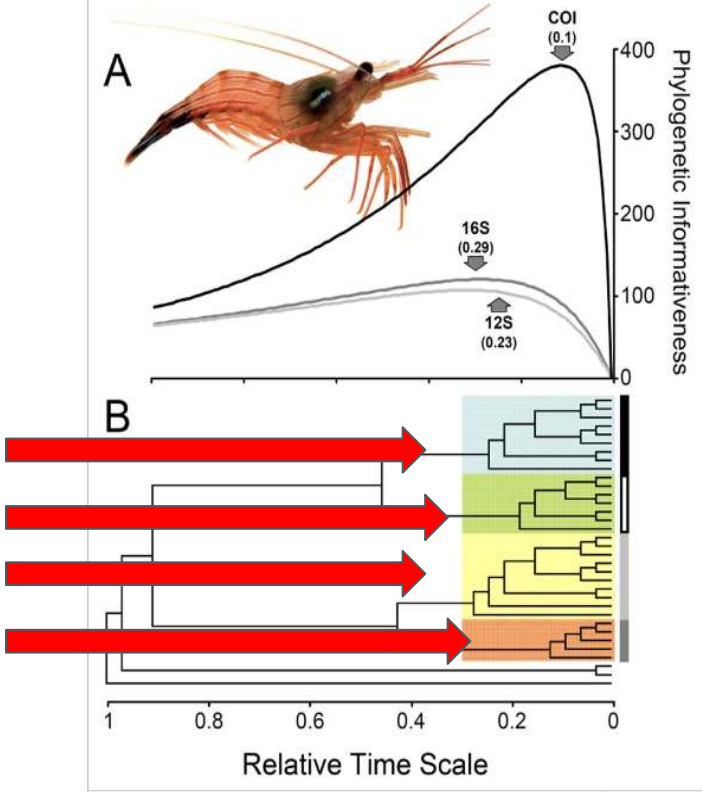
Which properties?

Information content

-> length of alignment
-> missing data
-> level of occupancy

Phylogenetic signal

Good information to infer these nodes



Baeza & Fuentes 2013

04 PHYLOGENOMIC SUBSAMPLING

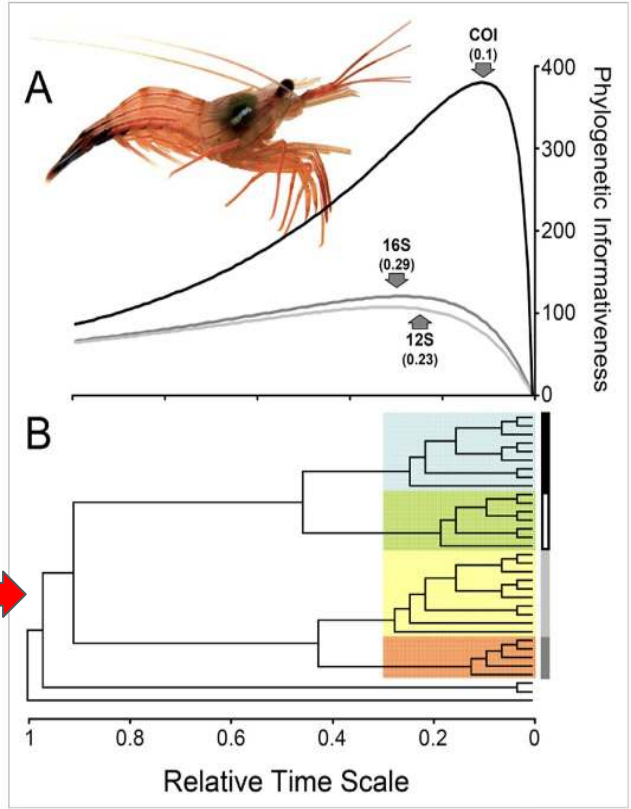
Which properties?

Information content

-> length of alignment
-> missing data
-> level of occupancy

Phylogenetic signal

Not enough information to infer these nodes



Baeza & Fuentes 2013

04 PHYLOGENOMIC SUBSAMPLING

Which properties?

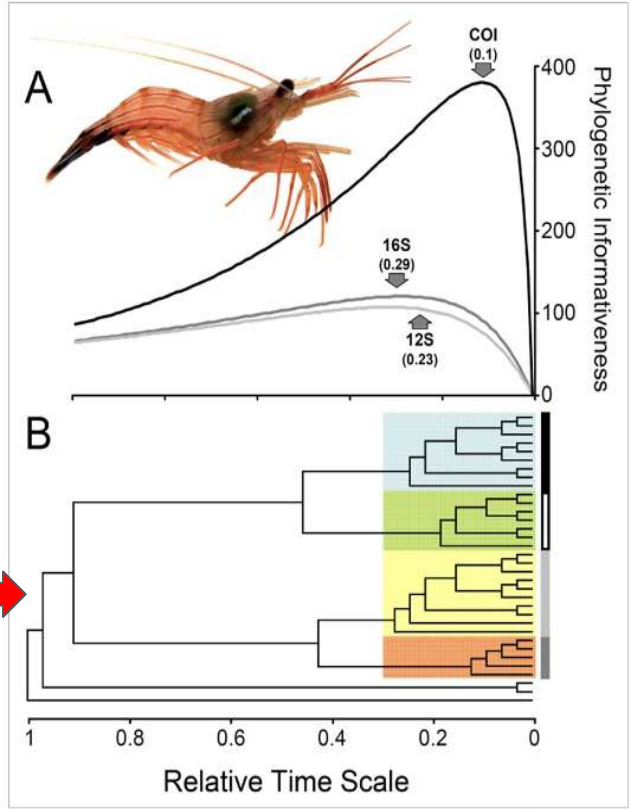
Information content

-> length of alignment
-> missing data
-> level of occupancy

Phylogenetic signal

-> average support
-> Robinson-Foulds distance

Not enough information to infer these nodes



Baeza & Fuentes 2013

04 PHYLOGENOMIC SUBSAMPLING

Which properties?

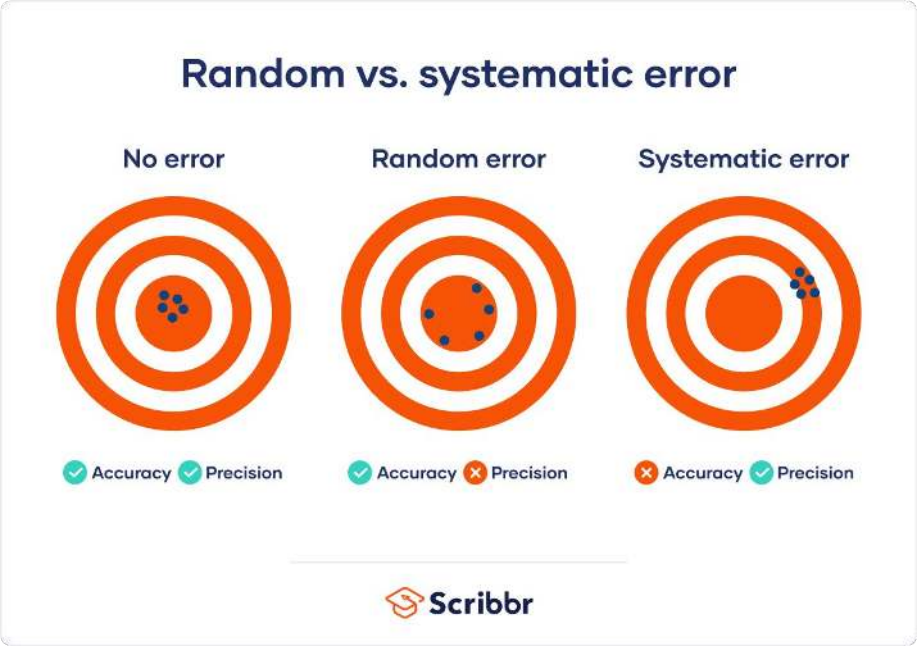
Information content

-> length of alignment
-> missing data
-> level of occupancy

Phylogenetic signal

-> average support
-> Robinson-Foulds distance

Systematic error: when a calculated value deviates from the true value in a consistent way.



04 PHYLOGENOMIC SUBSAMPLING

Which properties?

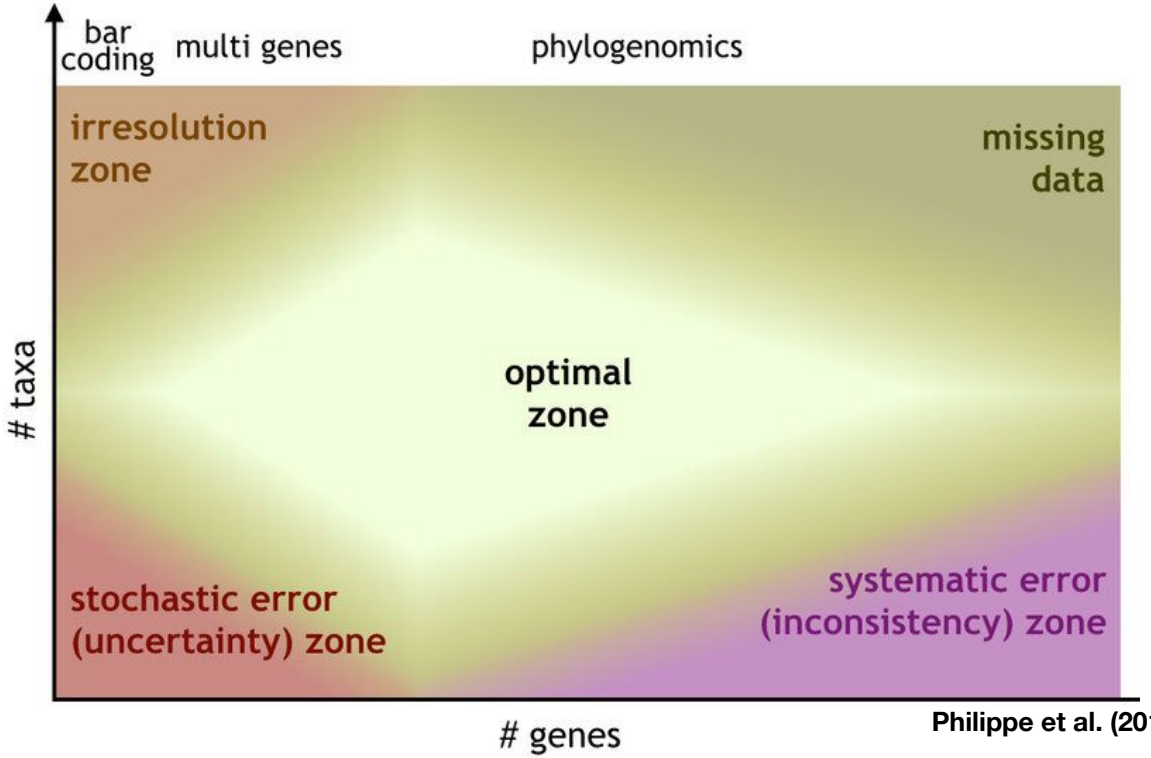
Information content

-> length of alignment
-> missing data
-> level of occupancy

Phylogenetic signal

-> average support
-> Robinson-Foulds dista

Systematic error:



Philippe et al. (2017)

04 PHYLOGENOMIC SUBSAMPLING

Which properties?

Information content

-> length of alignment
-> missing data
-> level of occupancy

Phylogenetic signal

-> average support
-> Robinson-Foulds distance

Systematic error

-> root-to-tip distance (ie, the degree of deviation from a strict clock-like behavior)

04 PHYLOGENOMIC SUBSAMPLING

Which properties?

Information content

-> length of alignment
-> missing data
-> level of occupancy

Phylogenetic signal

-> average support
-> Robinson-Foulds distance

Systematic error

-> root-to-tip distance (ie, the degree of deviation from a strict clock-like behavior)
-> average pair-wise patristic distance between terminals (indicative of susceptibility to long-branch attraction)

04 PHYLOGENOMIC SUBSAMPLING

Which properties?

Information content

-> length of alignment
-> missing data
-> level of occupancy

Phylogenetic signal

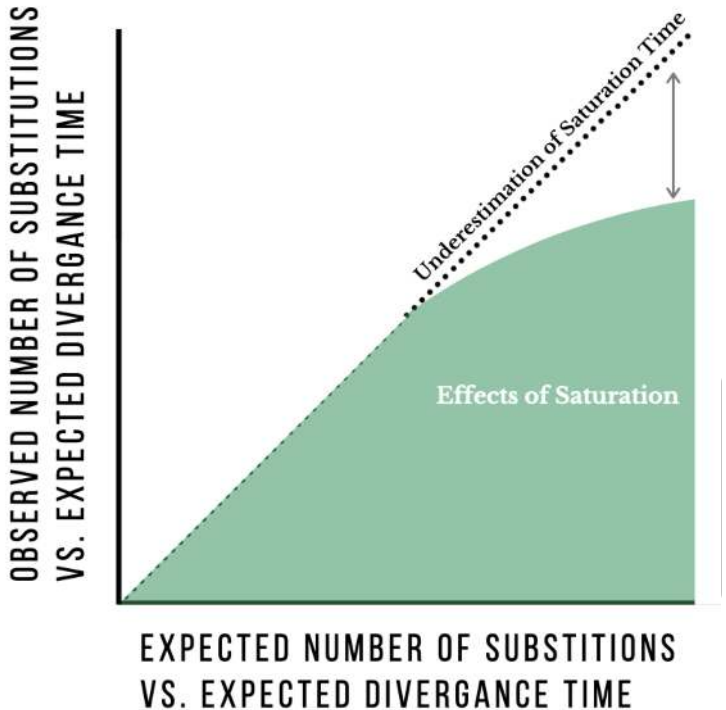
-> average support
-> Robinson-Foulds distance

Systematic error

-> root-to-tip distance (ie, the degree of deviation from a strict clock-like behavior)
-> average pair-wise patristic distance between terminals (indicative of susceptibility to long-branch attraction)
-> level of saturation

04 PHYLOGENOMIC SUBSAMPLING

Which properties?



Systematic error

-> root-to-tip distance (ie, the degree of deviation from a strict clock-like behavior)

-> average pair-wise patristic distance between terminals (indicative of susceptibility to long-branch attraction)

-> level of saturation

04 PHYLOGENOMIC SUBSAMPLING

Which properties?

Information content

-> length of alignment
-> missing data
-> level of occupancy

Phylogenetic signal

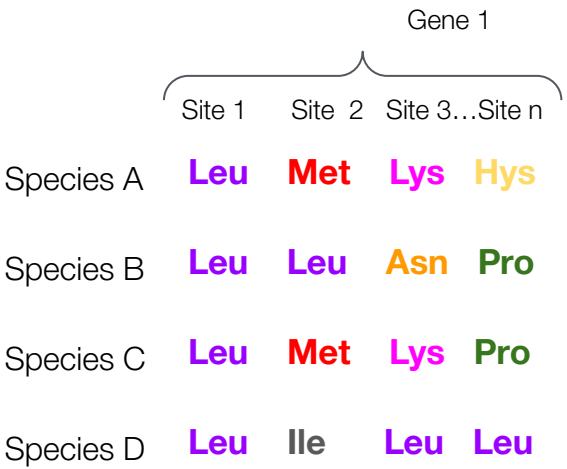
-> average support
-> Robinson-Foulds distance

Systematic error

-> root-to-tip distance (ie, the degree of deviation from a strict clock-like behavior)
-> average pair-wise patristic distance between terminals (indicative of susceptibility to long-branch attraction)
-> level of saturation
-> compositional heterogeneity

04 PHYLOGENOMIC SUBSAMPLING

Which properties?



Systematic error

-> root-to-tip distance (ie, the degree of deviation from a strict clock-like behavior)

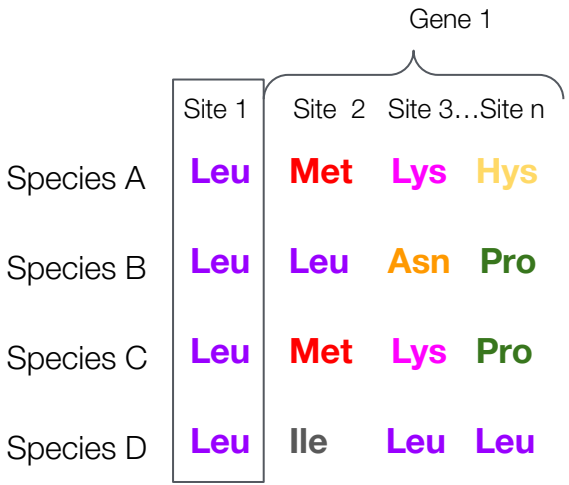
-> average pair-wise patristic distance between terminals (indicative of susceptibility to long-branch attraction)

-> level of saturation

-> compositional heterogeneity

04 PHYLOGENOMIC SUBSAMPLING

Which properties?



Systematic error

-> root-to-tip distance (ie, the degree of deviation from a strict clock-like behavior)

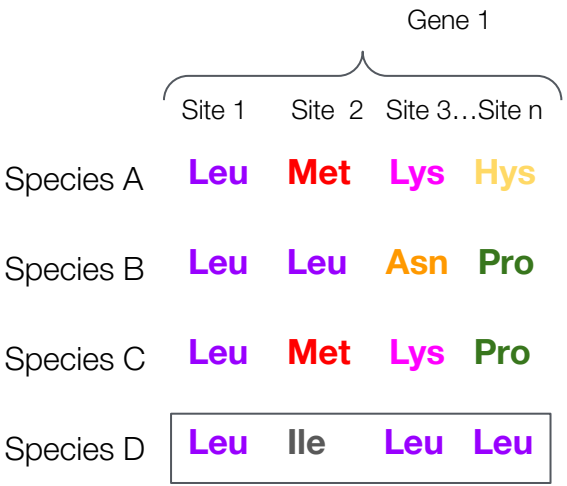
-> average pair-wise patristic distance between terminals (indicative of susceptibility to long-branch attraction)

-> level of saturation

-> compositional heterogeneity

04 PHYLOGENOMIC SUBSAMPLING

Which properties?



Systematic error

-> root-to-tip distance (ie, the degree of deviation from a strict clock-like behavior)

-> average pair-wise patristic distance between terminals (indicative of susceptibility to long-branch attraction)

-> level of saturation

-> compositional heterogeneity

04 PHYLOGENOMIC SUBSAMPLING

Which properties?

Information content

-> length of alignment
-> missing data
-> level of occupancy

Phylogenetic signal

-> average support
-> Robinson-Foulds distance

Systematic error

-> root-to-tip distance (ie, the degree of deviation from a strict clock-like behavior)
-> average pair-wise patristic distance between terminals (indicative of susceptibility to long-branch attraction)
-> level of saturation
-> compositional heterogeneity

Jacob and Marina today

Antonis Rokas and Jacob on 29th Jan

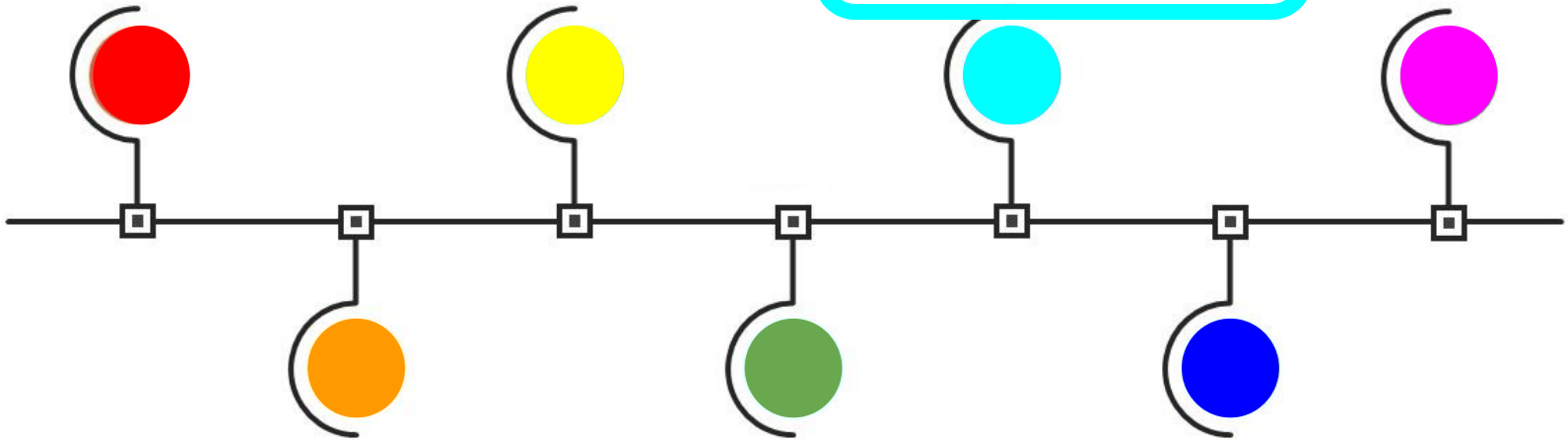
01 DATA

02 ORTHOLOGY INFERENCE

03 ALIGNMENT & TRIMMING

04 PHYLOGENOMIC SUBSAMPLING

05 SUPERMATRIX VS INDIVIDUAL GENES



05 SUPERMATRIX VS INDIV. GENE TREES

05 SUPERMATRIX VS INDIV. GENE TREES

~~Gene tree ≈ Species phylogeny~~

Gene tree ≠ Species phylogeny

05 SUPERMATRIX VS INDIV. GENE TREES

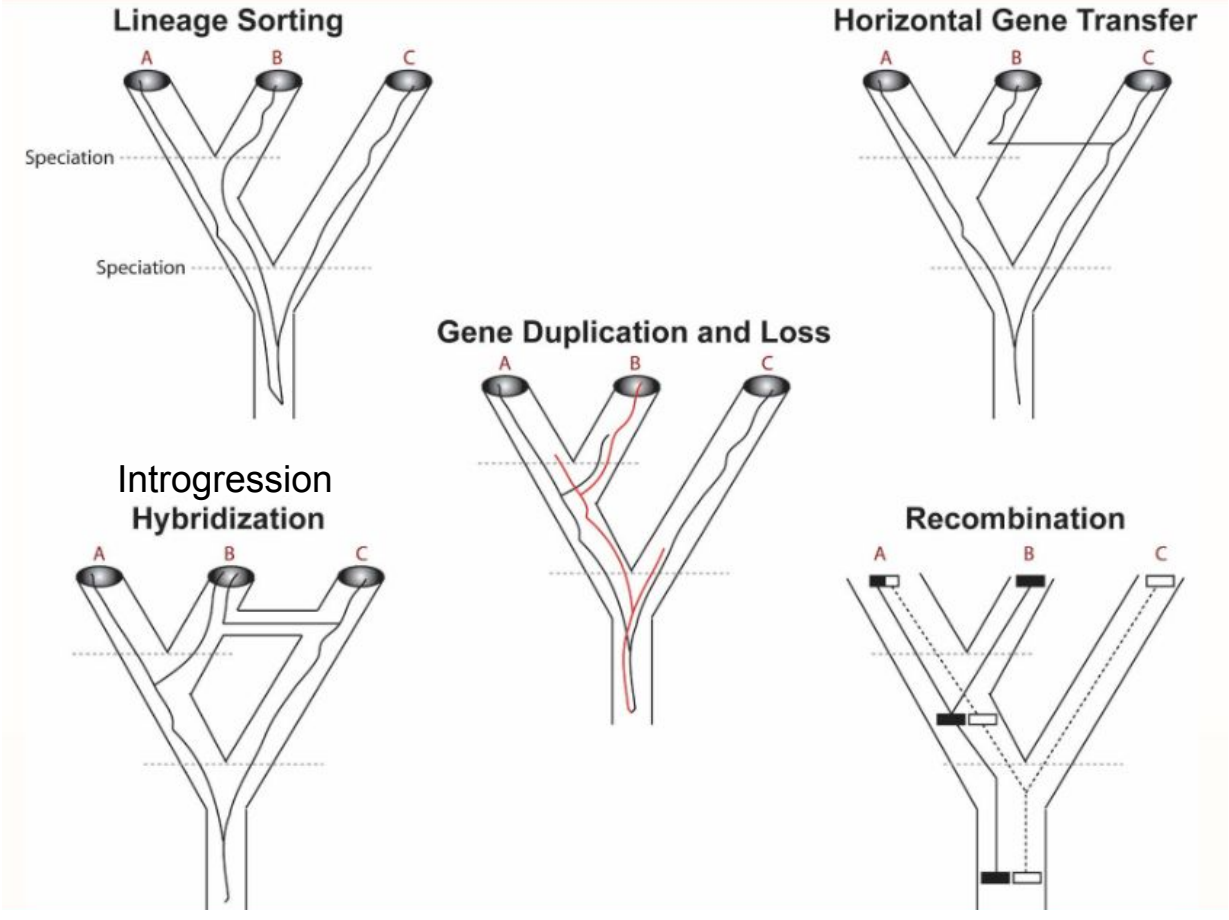
Analytical factors

They lead to failure in accurately inferring a gene tree; these can be either due to **stochastic error** (e.g., insufficient sequence length or taxon samples) or due to **systematic error** (e.g., observed data far depart from model assumptions)

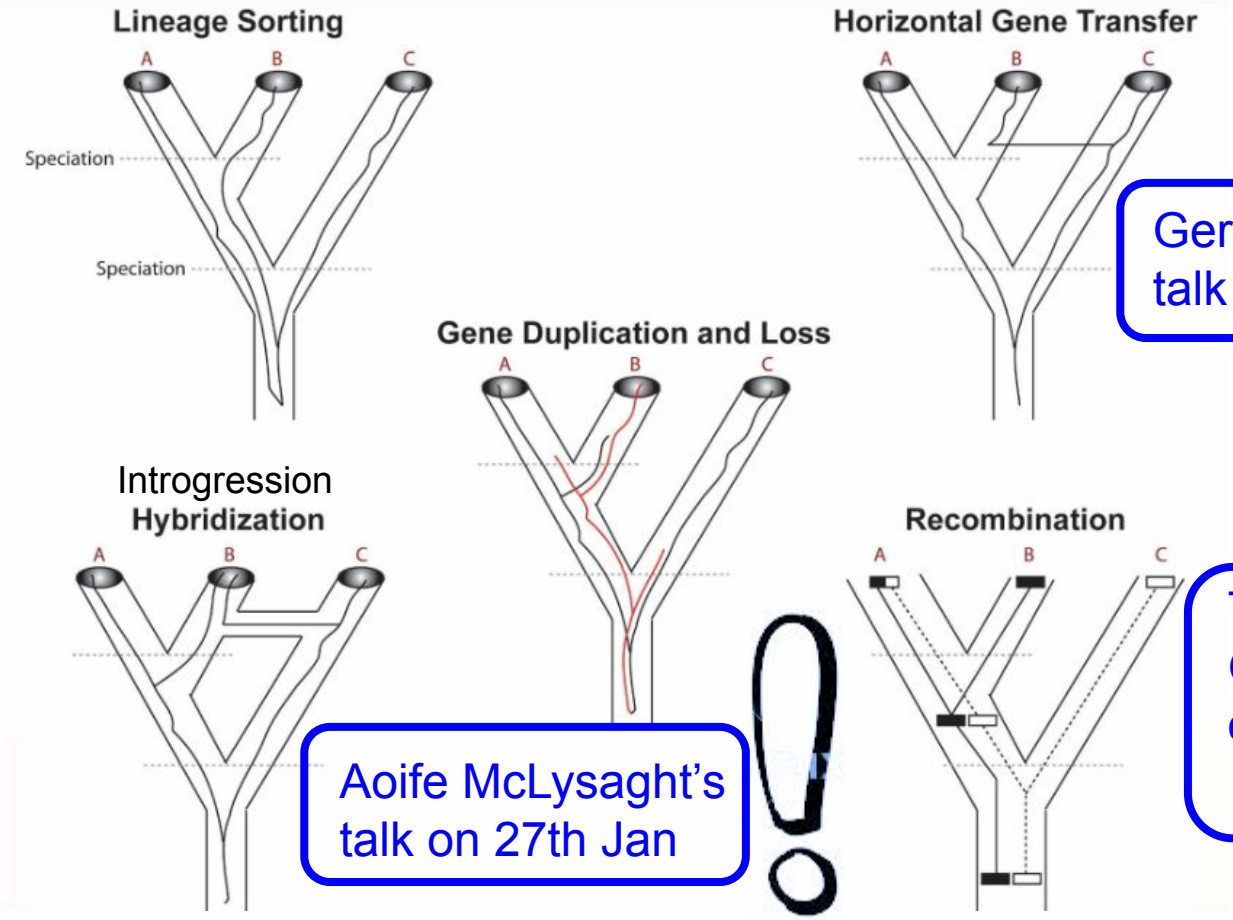
Biological factors

They lead to gene trees that are topologically distinct from each other and from the species tree. Known factors include **stochastic lineage sorting**, **hidden paralogy**, **horizontal gene transfer**, **recombination** and **natural selection**

05 SUPERMATRIX VS INDIV. GENE TREES



05 SUPERMATRIX VS INDIV. GENE TREES

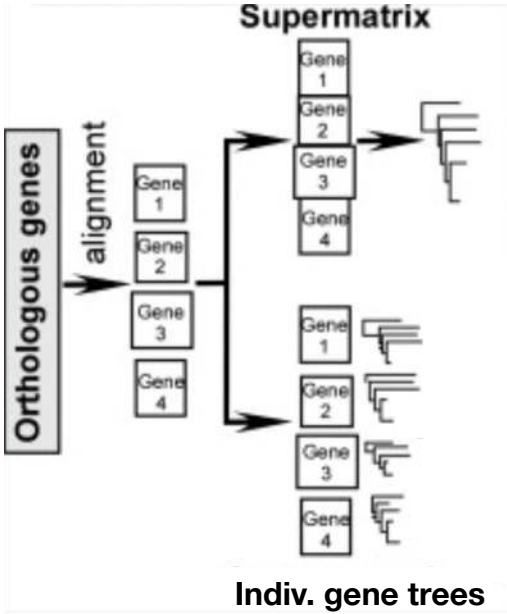


Gergely Szöllősi's talk on 1st Feb

Aoife McLysaght's talk on 27th Jan

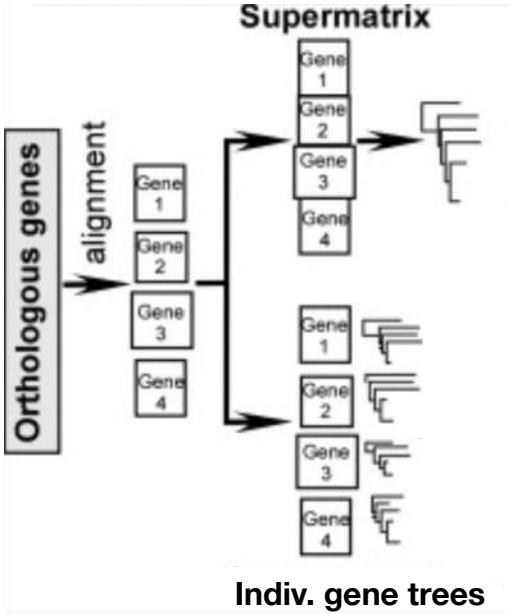
Toni Gabaldón's tal on 30th Jan

05 SUPERMATRIX VS INDIV. GENE TREES



Adapted from Fernández, Hormiga & Giribet (2014)

05 SUPERMATRIX VS INDIV. GENE TREES

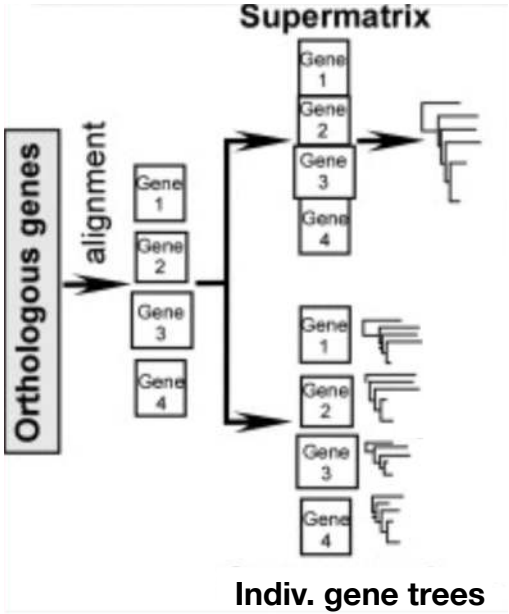


Phylogenetic analysis
(one tree)

Phylogenetic analysis
(multiple trees)

Adapted from Fernández, Hormiga & Giribet (2014)

05 SUPERMATRIX VS INDIV. GENE TREES



Phylogenetic analysis (one tree)

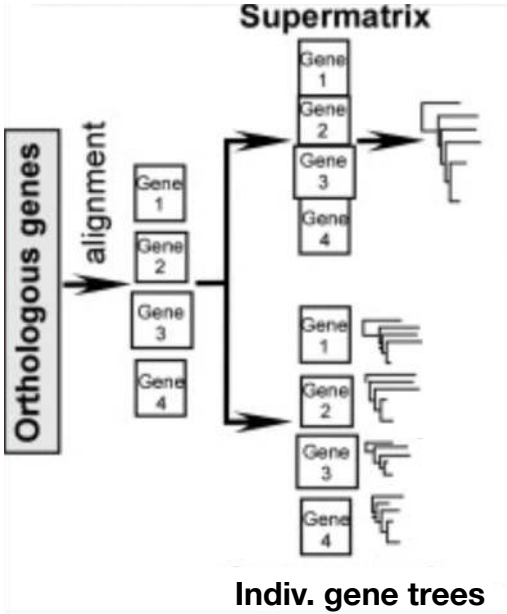
Phylogenetic analysis (multiple trees)

Estimation of a species tree given a set of gene trees

Multispecies coalescent

Adapted from Fernández, Hormiga & Giribet (2014)

05 SUPERMATRIX VS INDIV. GENE TREES



Phylogenetic analysis (one tree)

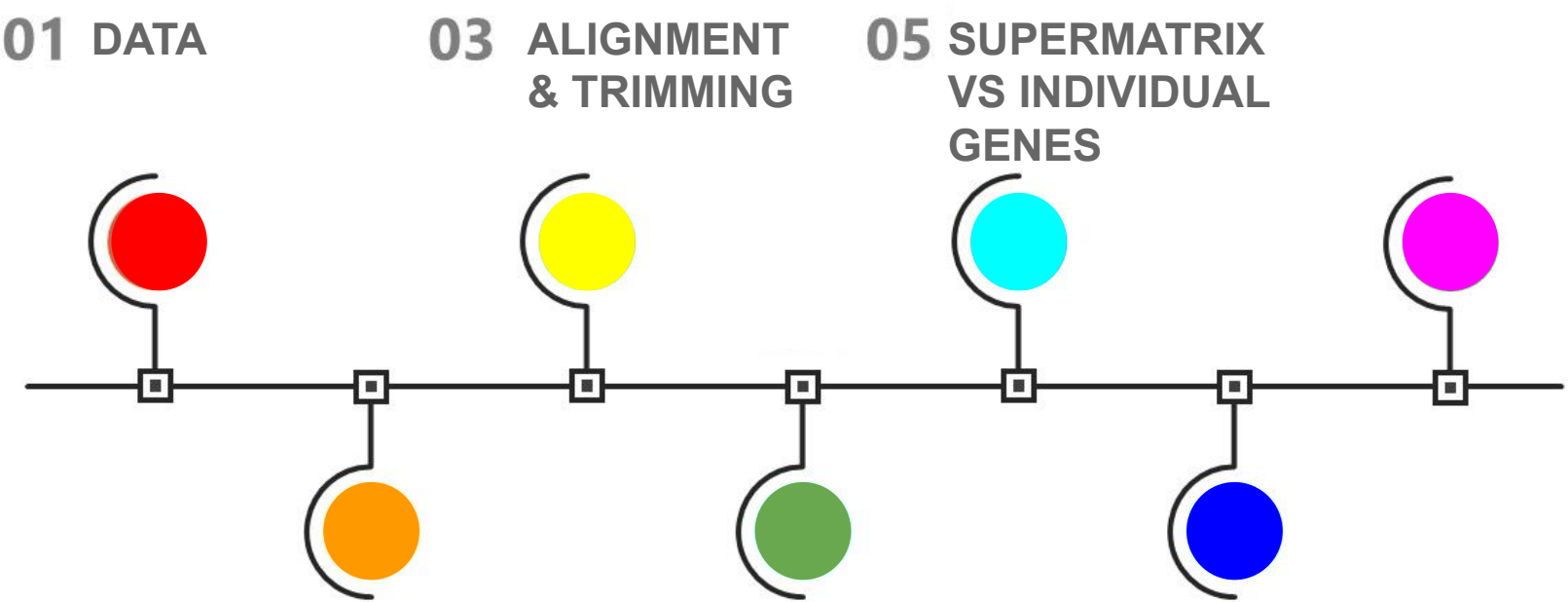
Phylogenetic analysis (multiple trees)

Estimation of a species tree given a set of gene trees

Multispecies coalescent

Erin Molloy's talks on 25th and 26th Jan

Adapted from Fernández, Hormiga & Giribet (2014)


01 DATA
02 ORTHOLOGY INFERENCE
03 ALIGNMENT & TRIMMING
04 PHYLOGENOMI SUBSAMPLING
05 SUPERMATRIX VS INDIVIDUAL GENES
06 MODEL SELECTION & PHYLOGENETIC INFERENCE

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + MODEL OF EVOLUTION

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + MODEL OF EVOLUTION + METHOD

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + MODEL OF EVOLUTION

+ METHOD

+ A WAY TO ASSESS HOW GOOD YOUR HYPOTHESIS IS

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + MODEL OF EVOLUTION

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

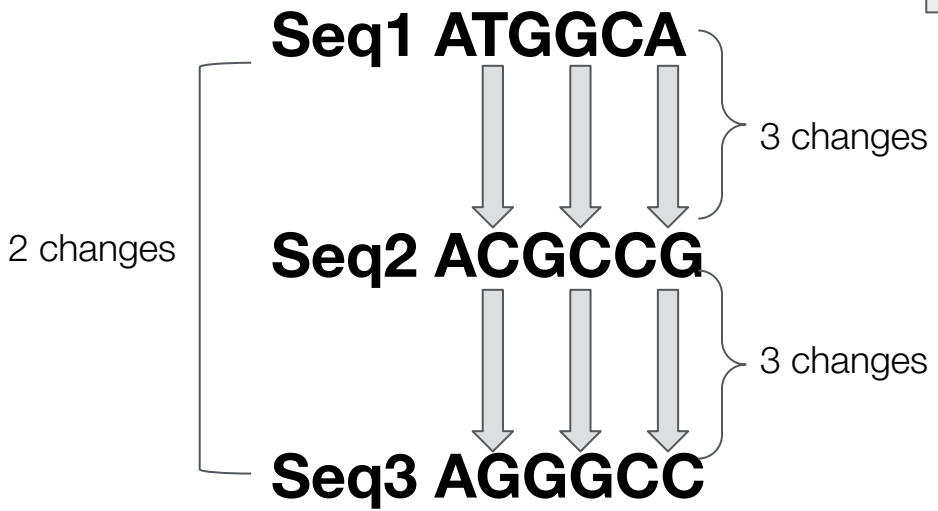
Observed number of changes + **Equation** = **Evolutionary distance**

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

Observed number of changes + **Equation** = **Evolutionary distance**



06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

Observed number of changes + **Equation** = **Evolutionary distance**

Seq1 ATGGCA

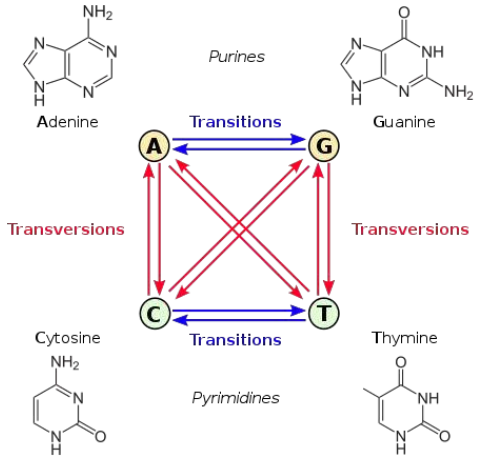
3 changes

2 changes

Seq2 ACGCCG

3 changes

Seq3 AGGGCC



06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

Observed number of changes + **Equation** = **Evolutionary distance**

Seq1 ATGGCA

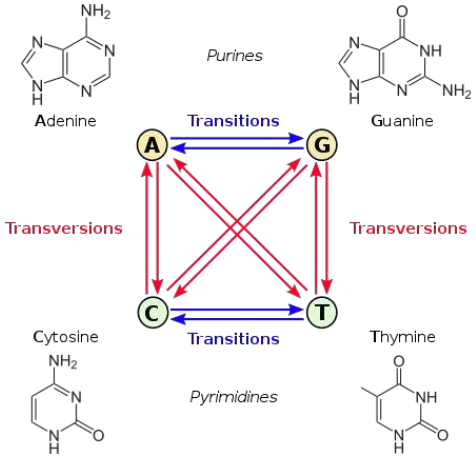
3 changes
(1 transition, 2 transversions)

2 changes

Seq2 ACGCCG

3 changes (3 transvesions)

Seq3 AGGGCC

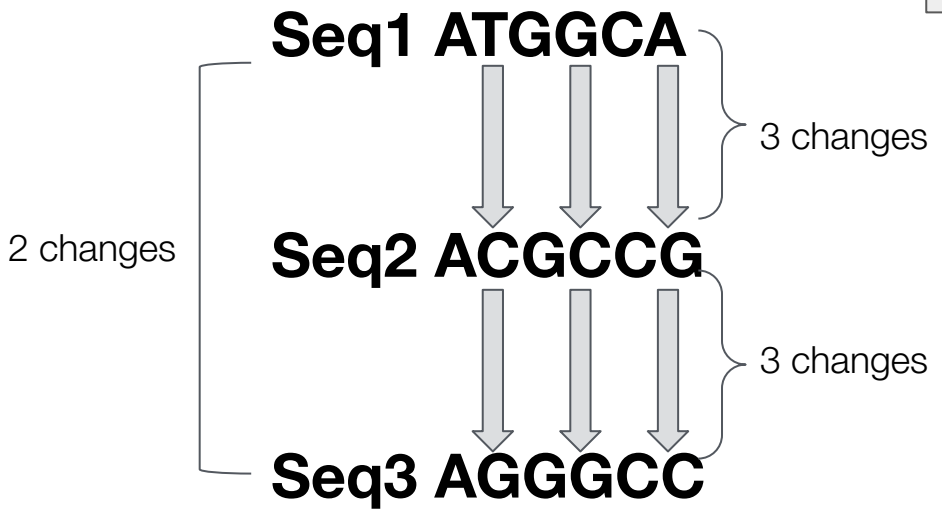


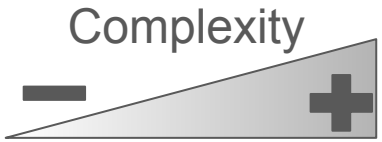
06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

Observed number of changes + **Equation** = **Evolutionary distance**





06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

Observed number of changes + **Equation** = **Evolutionary distance**

Seq1 ATGGCA

3 changes

Seq2 ACGCCG

3 changes

Seq3 AGGGCC

2 changes

Complexity

Jukes & Cantor

nucleotides

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

Observed number of changes + **Equation** = **Evolutionary distance**

Seq1 ATGGCA

3 changes

Seq2 ACGCCG

2 changes

3 changes

Seq3 AGGGCC

Complexity

Jukes & Cantor

Kimura 2P

nucleotides

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

Observed number of changes + **Equation** = **Evolutionary distance**

Seq1 ATGGCA

3 changes

Seq2 ACGCCG

2 changes

3 changes

Seq3 AGGGCC

Complexity

Jukes & Cantor

Kimura 2P

Felsenstein 81

nucleotides

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

Observed number of changes + **Equation** = **Evolutionary distance**

Seq1 ATGGCA

3 changes

2 changes

Seq2 ACGCCG

3 changes

Seq3 AGGGCC

Complexity

Jukes & Cantor

Kimura 2P

Felsenstein 81

GTR…

nucleotides

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

Observed number of changes + **Equation** = **Evolutionary distance**

Seq1 ATGGCA

3 changes

Seq2 ACGCCG

2 changes

3 changes

Seq3 AGGGCC

Complexity

Jukes & Cantor

Kimura 2P

Felsenstein 81

GTR…

PAM

nucleotides

amino acids

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

Observed number of changes + **Equation** = **Evolutionary distance**

Seq1 ATGGCA

3 changes

Seq2 ACGCCG

2 changes

3 changes

Seq3 AGGGCC

Complexity

Jukes & Cantor

Kimura 2P

Felsenstein 81

GTR…

PAM

BLOSUM

nucleotides

amino acids

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

Observed number of changes + **Equation** = **Evolutionary distance**

Seq1 ATGGCA

3 changes

Seq2 ACGCCG

2 changes

3 changes

Seq3 AGGGCC

Complexity

Jukes & Cantor

Kimura 2P

Felsenstein 81

GTR…

PAM

BLOSUM

JTT

nucleotides

amino acids

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

Observed number of changes + **Equation** = **Evolutionary distance**

Seq1 ATGGCA

3 changes

Seq2 ACGCCG

2 changes

3 changes

Seq3 AGGGCC

Complexity

Jukes & Cantor

Kimura 2P

Felsenstein 81

GTR…

nucleotides

PAM

BLOSUM

JTT

LG…

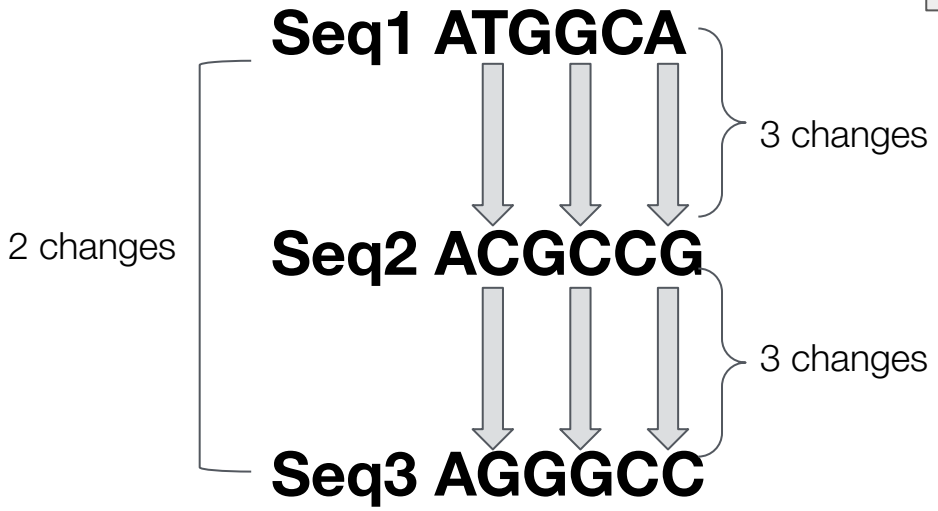
amino acids

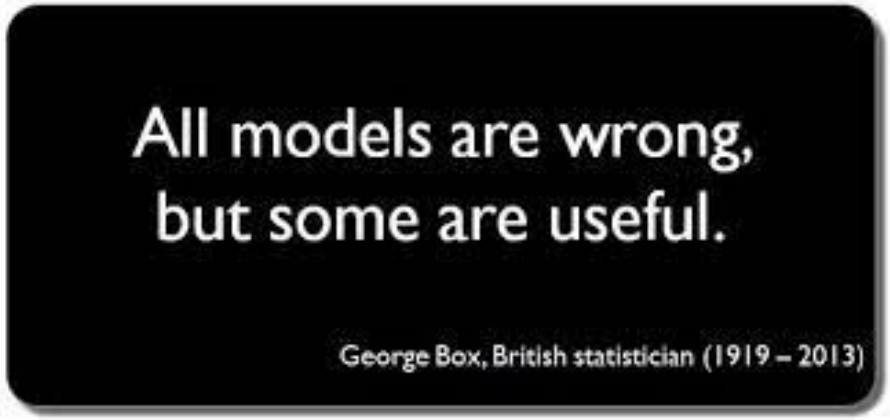
06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

Observed number of changes + **Equation** = **Evolutionary distance**



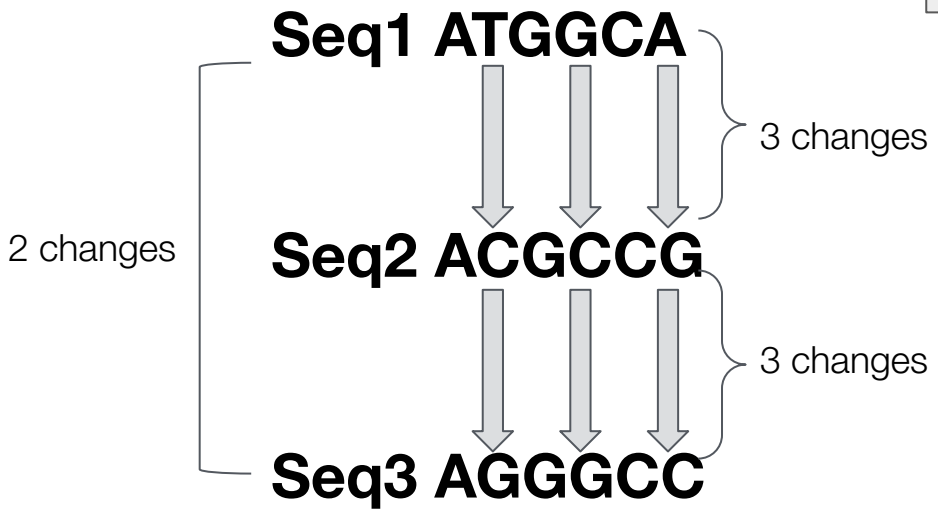


06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION** (= substitution model)

A model that describes changes in sequences over evolutionary time and transforms the number of changes in an evolutionary distance

Observed number of changes + **Equation** = **Evolutionary distance**



More about models:

Olivier Gascuel's talk on 25th Jan

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + MODEL OF EVOLUTION + METHOD

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION**

+ **METHOD**

Two main methods:

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

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

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?'

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

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?'

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

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)

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + MODEL OF EVOLUTION

+ METHOD

+ A WAY TO ASSESS HOW GOOD YOUR HYPOTHESIS IS

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION**

+ **METHOD**

+ **A WAY TO ASSESS HOW GOOD YOUR HYPOTHESIS IS**

Traditional metrics:

- ML: standard nonparametric bootstrap (100 reps), approximate likelihood ratio test (1,000 reps), ultrafast bootstrap (1,000 reps)(between 1 and 100)
- BI: posterior probability (between 0 and 1)

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION**

+ **METHOD**

+ **A WAY TO ASSESS HOW GOOD YOUR HYPOTHESIS IS**

Traditional metrics:

- ML: standard nonparametric bootstrap (100 reps), approximate likelihood ratio test (1,000 reps), ultrafast bootstrap (1,000 reps)(between 1 and 100)
- BI: posterior probability (between 0 and 1)

Novel metrics:

- concordance factor: for every branch of a reference tree, the percentage of “decisive” gene trees containing that branch.
- internode certainty/tree certainty: 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.
- Felsenstein’s bootstrap proportion (FBP)
- Transfer bootstrap expectation (TBE)

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

DATA + **MODEL OF EVOLUTION**

+ **METHOD**

+ **A WAY TO ASSESS HOW GOOD YOUR HYPOTHESIS IS**

Traditional metrics:

- ML: standard nonparametric bootstrap (100 reps), approximate likelihood ratio test (1,000 reps), ultrafast bootstrap (1,000 reps)(between 1 and 100)
- BI: posterior probability (between 0 and 1)

Olivier Gascuel and Oleksyi Kozlov's talks on 25th Jan

Novel metrics:

- concordance factor: for every branch of a reference tree, the percentage of "decisive" gene trees containing that branch.
- internode certainty/tree certainty: 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.
- Felsenstein's bootstrap proportion (FBP)
- Transfer bootstrap expectation (TBE)

01 DATA

02 ORTHOLOGY INFERENCE

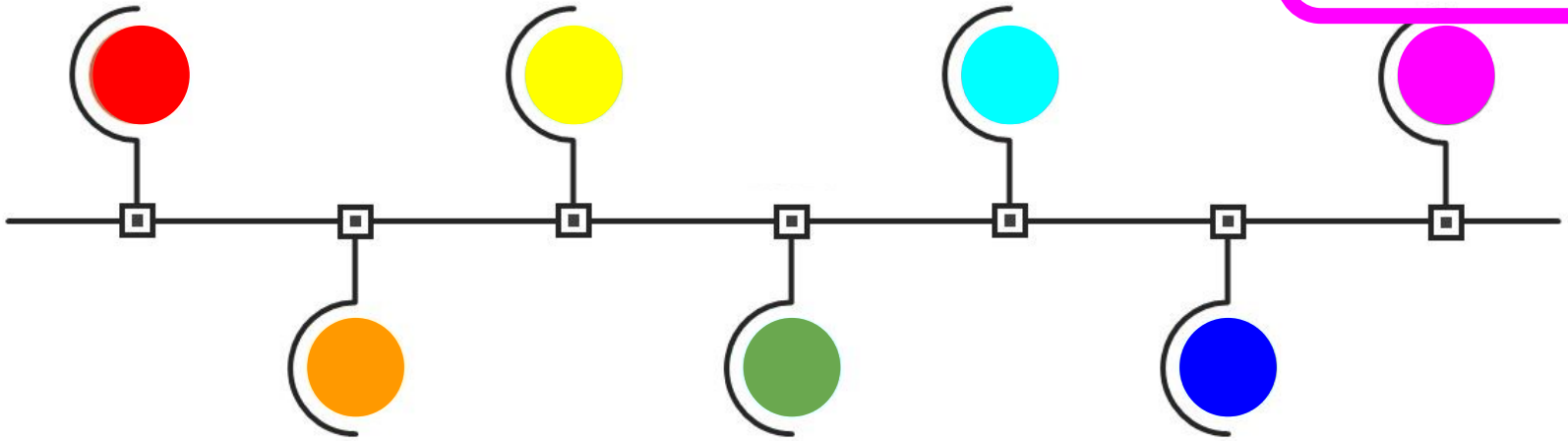
03 ALIGNMENT & TRIMMING

04 PHYLOGENOMIC SUBSAMPLING

05 SUPERMATRIX VS INDIVIDUAL GENES

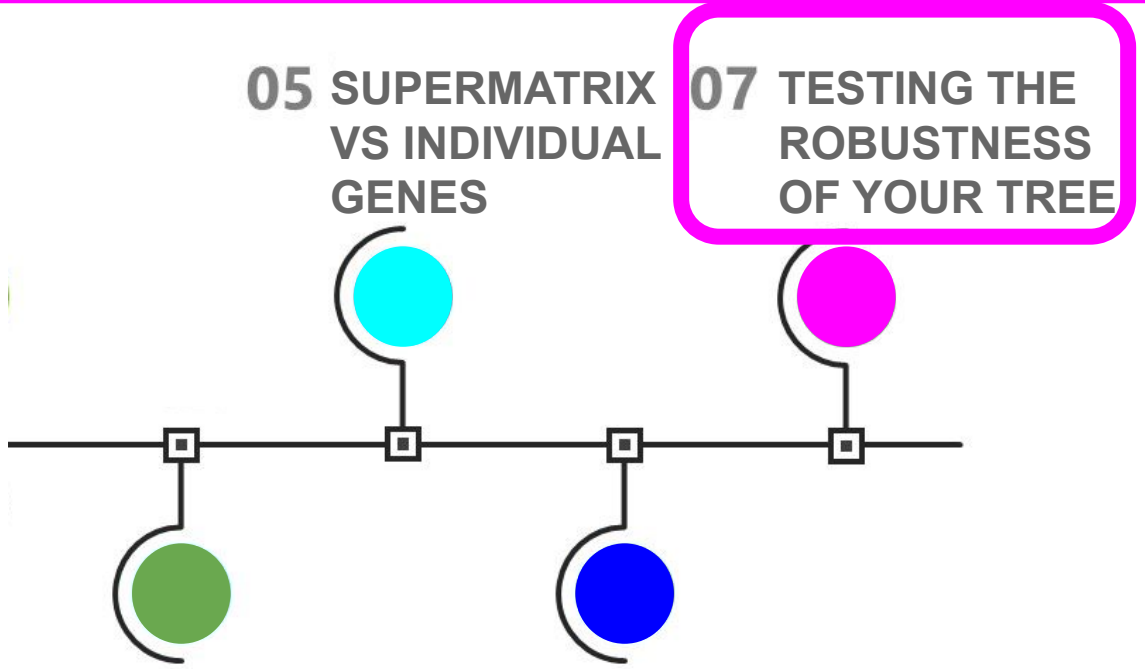
06 MODEL SELECTION & PHYLOGENETIC INFERENCE

07 TESTING THE ROBUSTNESS OF YOUR TREE

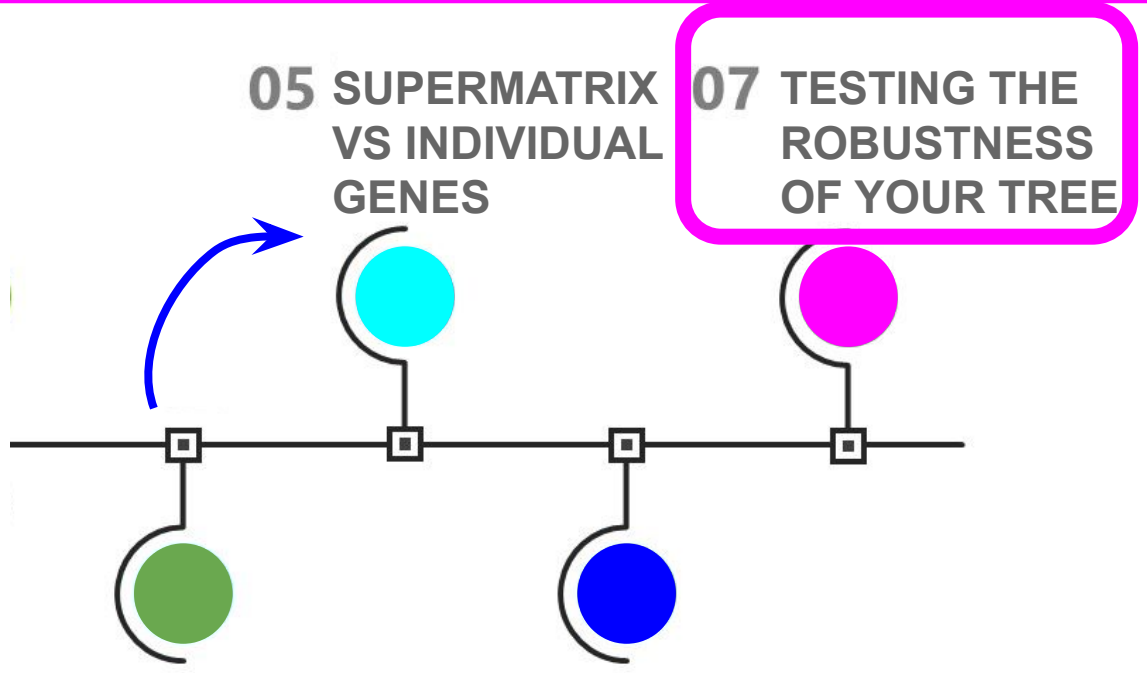


07 TESTING THE ROBUSTNESS OF YOUR TREE

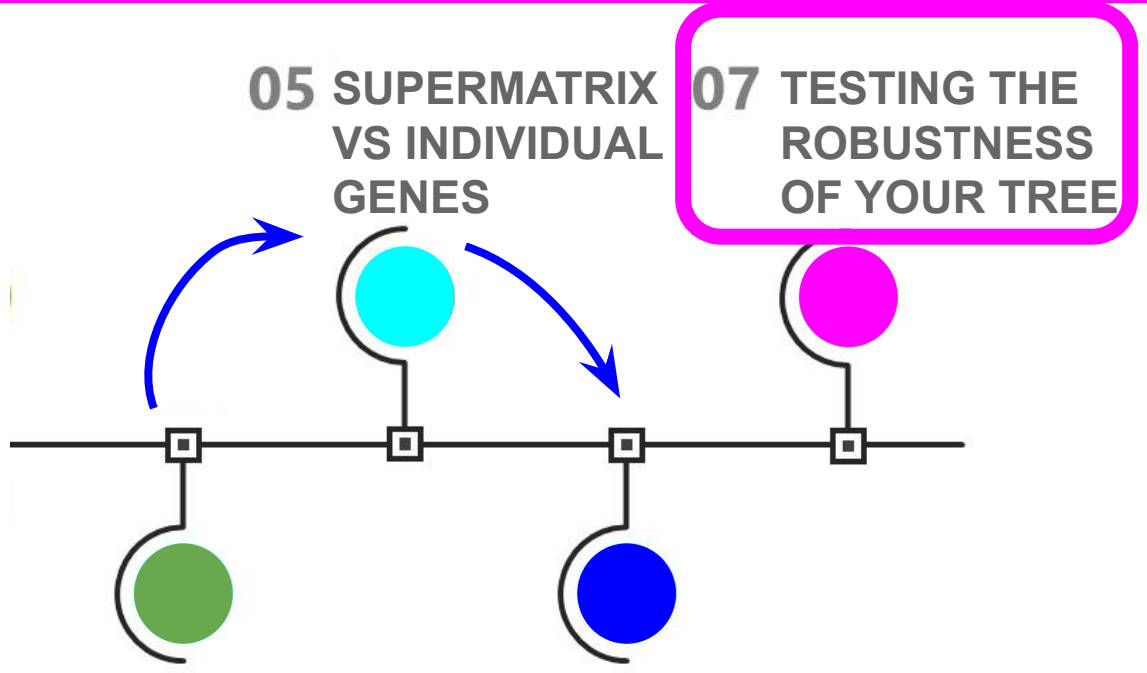
07 TESTING THE ROBUSTNESS OF YOUR TREE



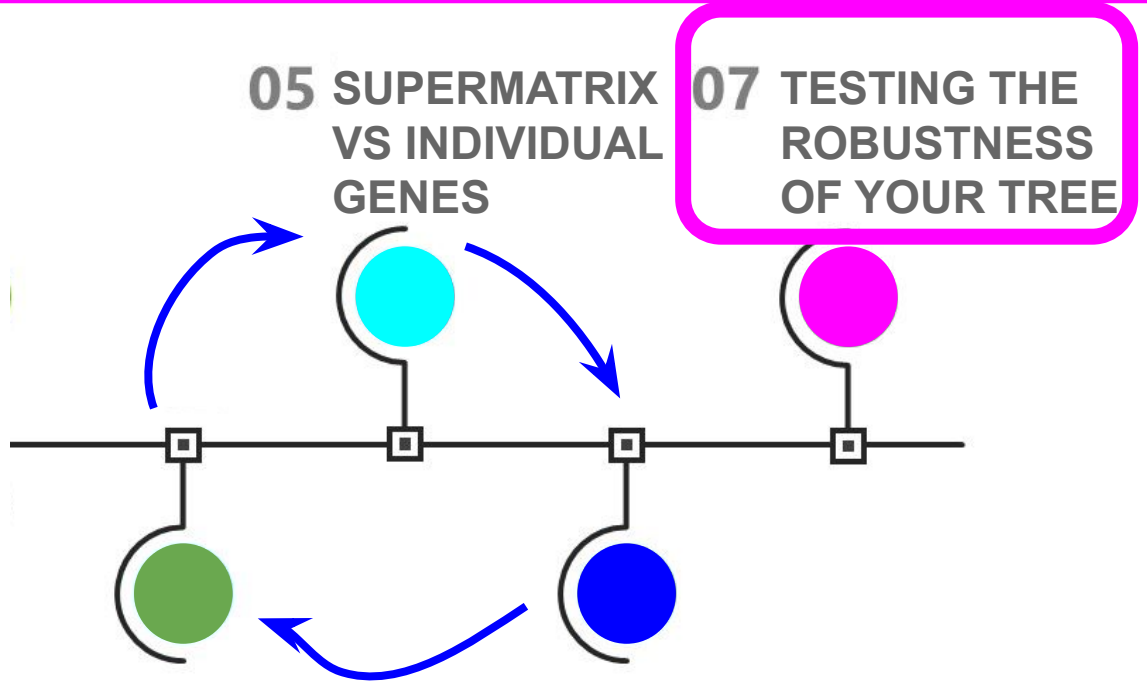
07 TESTING THE ROBUSTNESS OF YOUR TREE



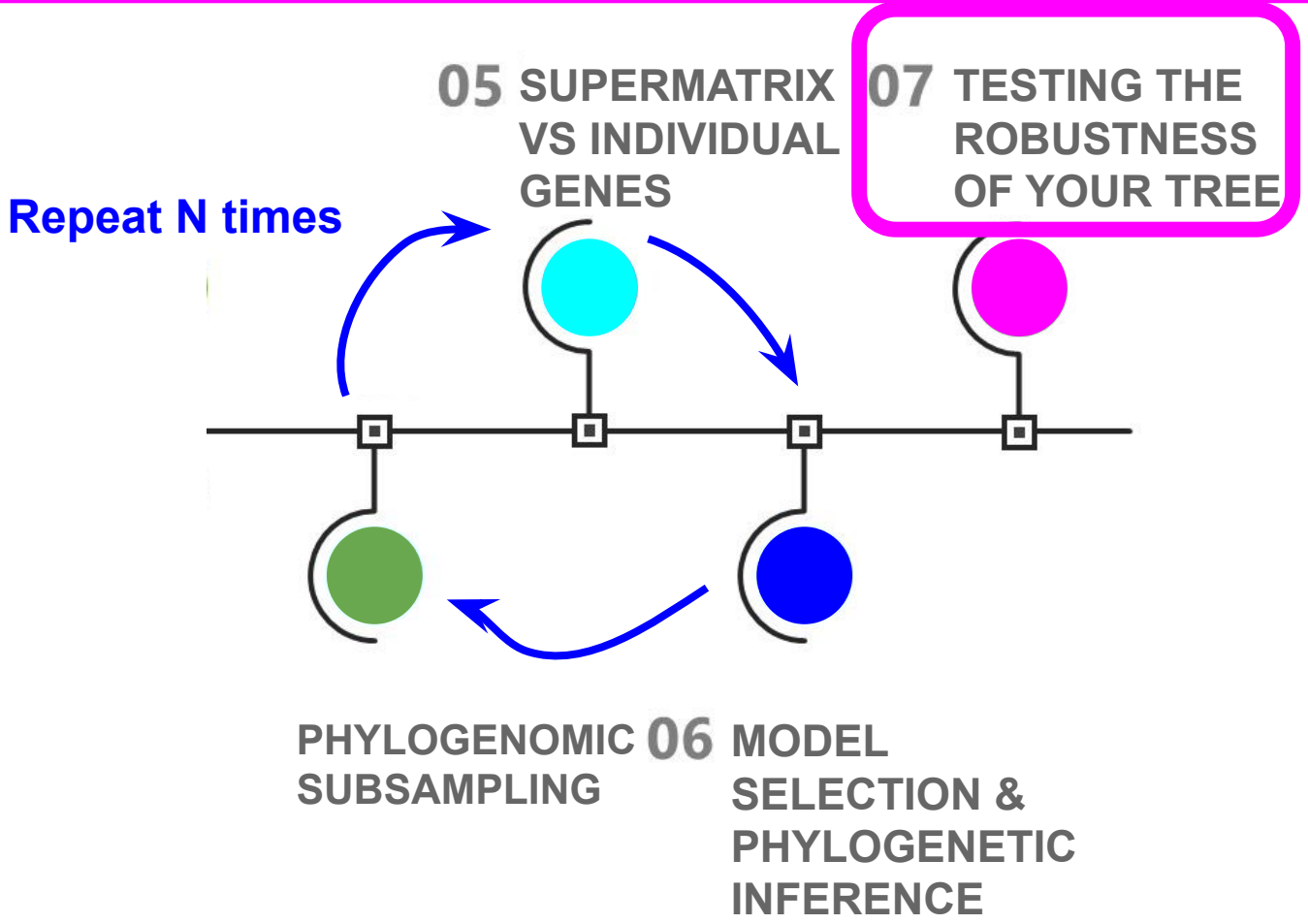
07 TESTING THE ROBUSTNESS OF YOUR TREE



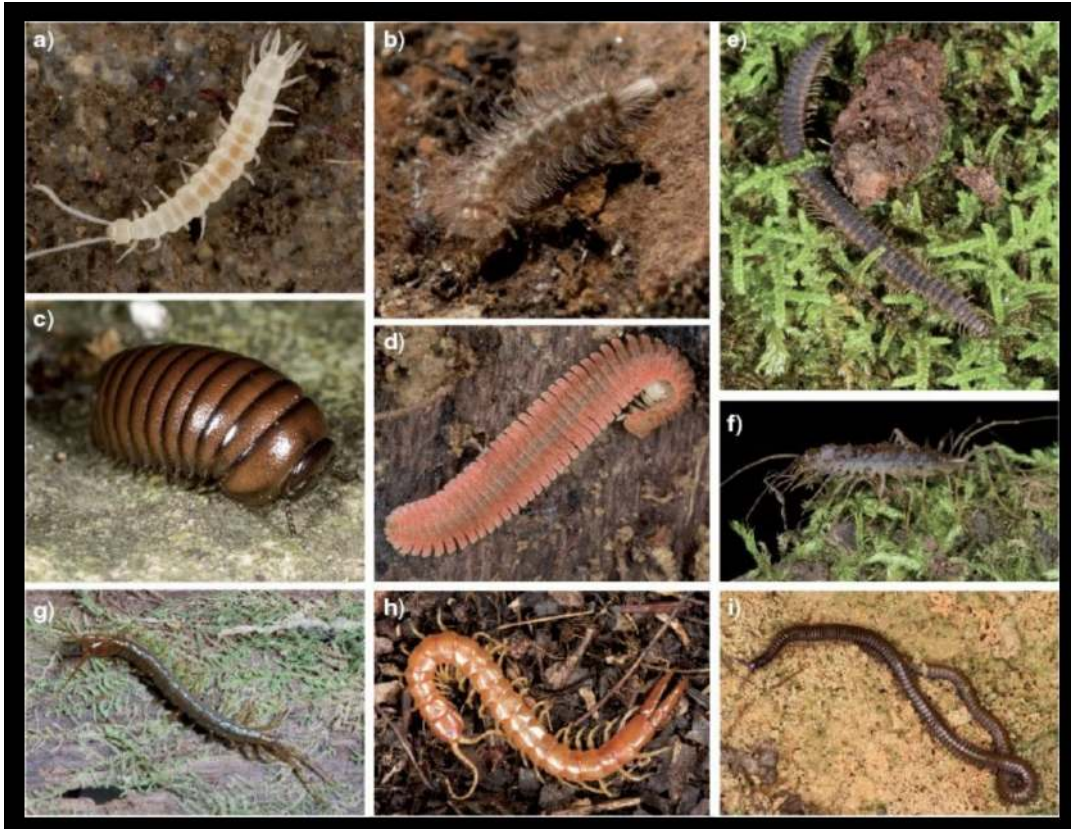
07 TESTING THE ROBUSTNESS OF YOUR TREE



07 TESTING THE ROBUSTNESS OF YOUR TREE



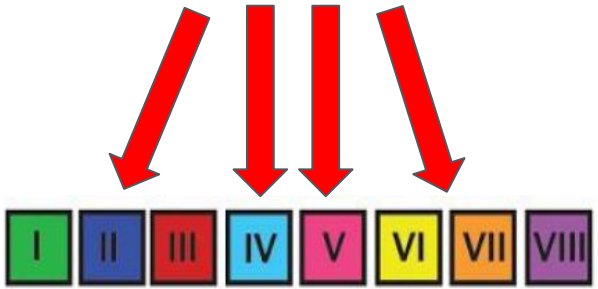
07 TESTING THE ROBUSTNESS OF YOUR TREE

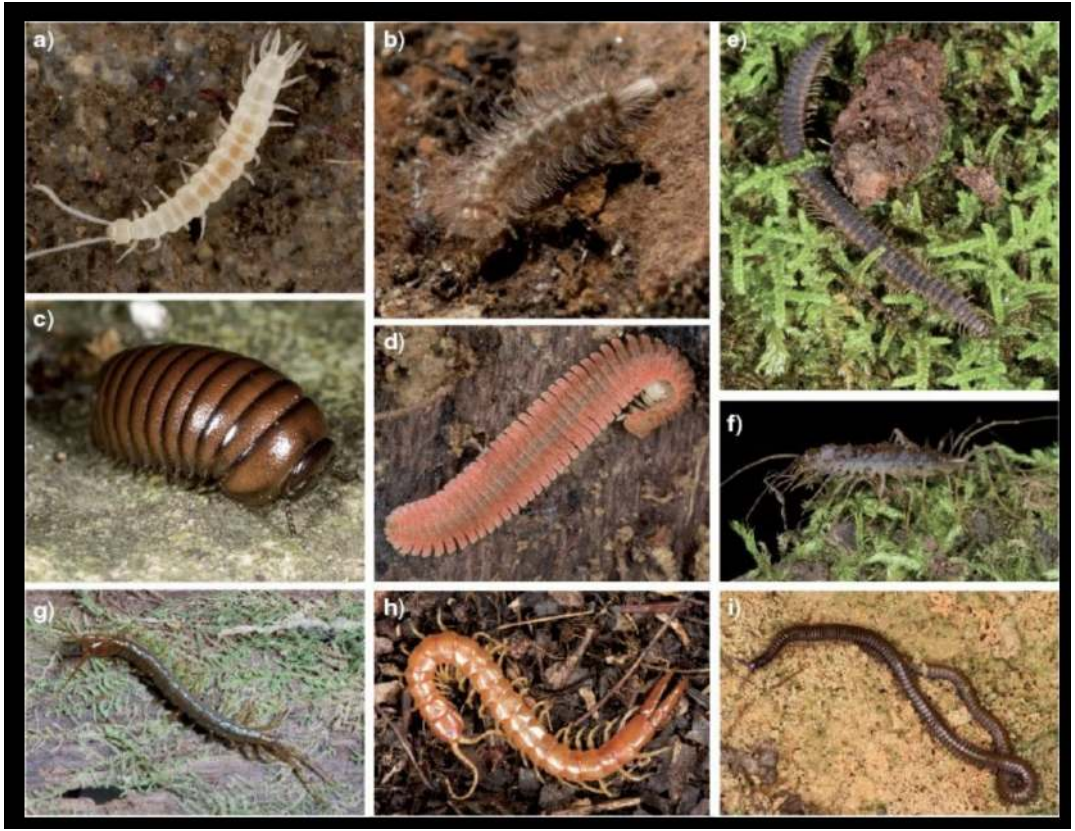


Fernández, Edgecombe & Giribet (2016) Syst Biol

07 TESTING THE ROBUSTNESS OF YOUR TREE

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

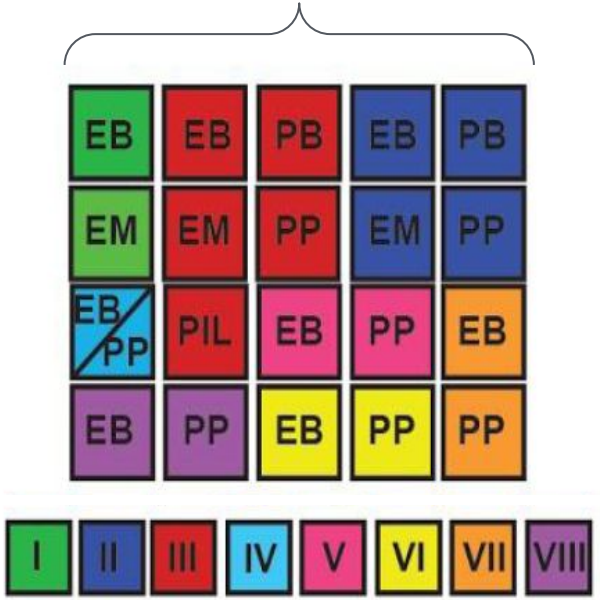


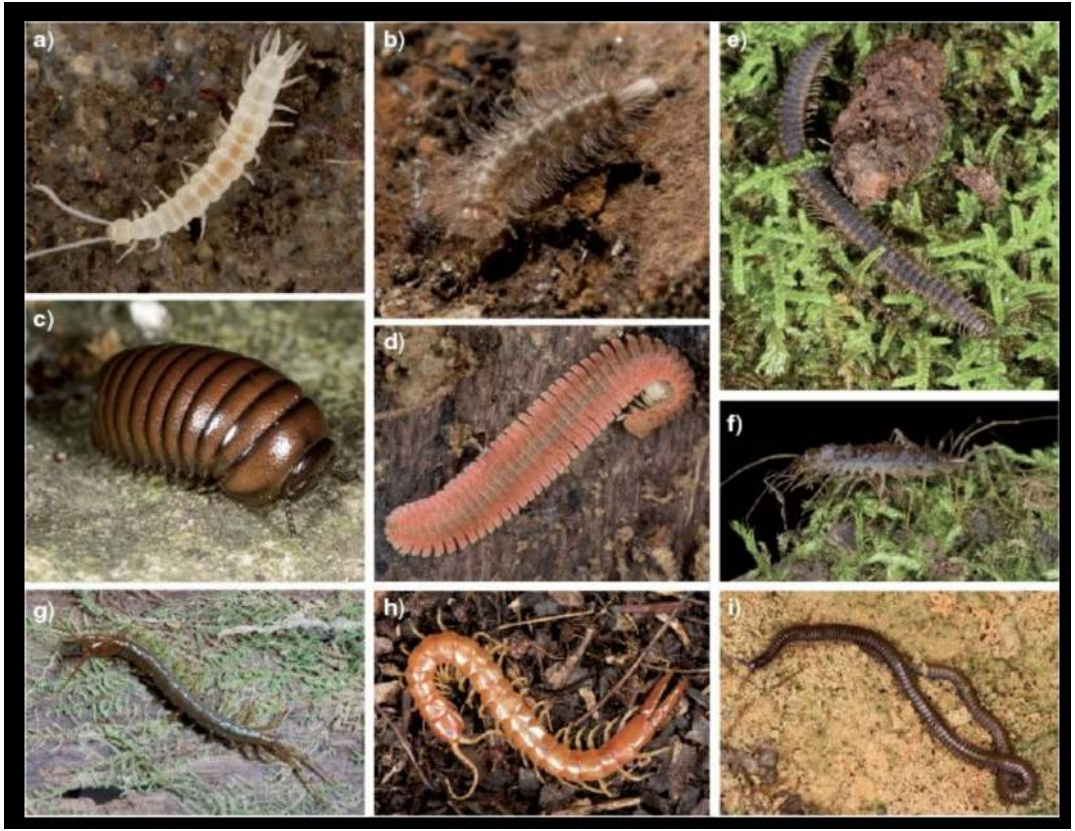


Fernández, Edgecombe & Giribet (2016) Syst Biol

07 TESTING THE ROBUSTNESS OF YOUR TREE

These are **analyses**

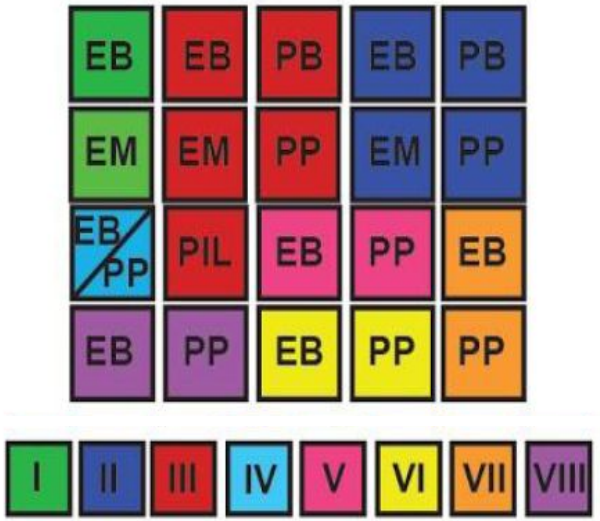


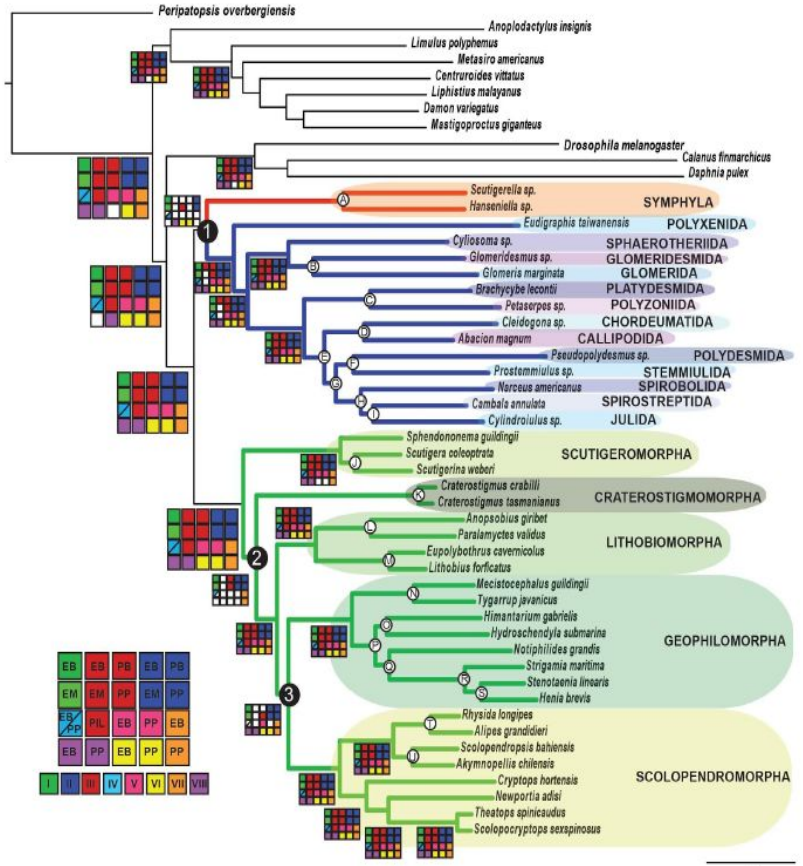


Fernández, Edgecombe & Giribet (2016) Syst Biol

07 TESTING THE ROBUSTNESS OF YOUR TREE

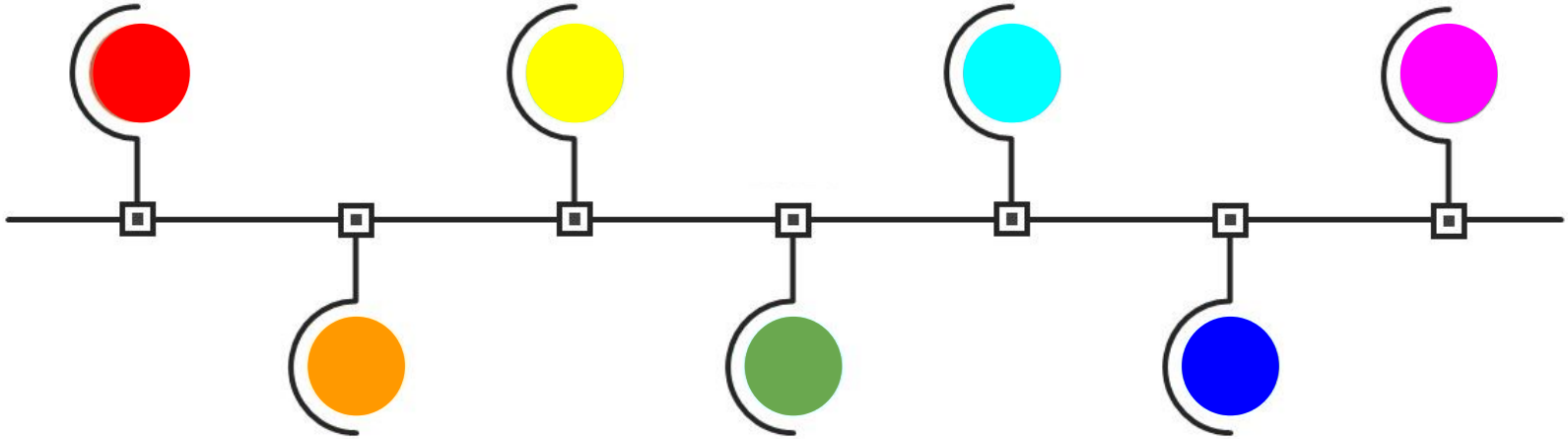
These are **analyses**





Fernández, Edgecombe & Giribet (2016) Syst Biol

AND YOU, HOW IS **YOUR** PROJECT?



01 DATA

02 ORTHOLOGY INFERENCE

03 ALIGNMENT & TRIMMING

04 PHYLOGENOMIC SUBSAMPLING

05 SUPERMATRIX VS INDIVIDUAL GENES

06 MODEL SELECTION & PHYLOGENETIC INFERENCE

07 TESTING THE ROBUSTNESS OF YOUR TREE


QUIZ TIME