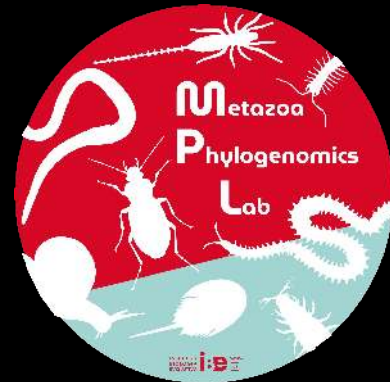


INTRODUCTION TO PHYLOGENOMICS

Rosa Fernández
Institute of Evolutionary Biology (CSIC-UPF)

rosa.fernandez@ibe.upf-csic.es

www.metazomics.com



A little bit about myself

Madrid (PhD)



Boston (1st postdoc)



Barcelona (2nd postdoc & my lab)



www.metazomics.com

@rosafernandez.bsky.social

Fun Facts:

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



I did my PhD on earthworms

Main lines of research:



MACROevolution :-)



Mother of two amazing girls :-)

ME AND THE WORKSHOP(S)



2017

Workshop on
Phylogenomics
(1st edition), TA

2019



COVID

2023

Workshop on Genomics
(Faculty & Scientific
Advisory Board)



2024

Workshop on Genomics
(Faculty)
Workshop on Phylogenomics
(3rd Edition), Co-Director

2026

Today :-)

Today's menu

1

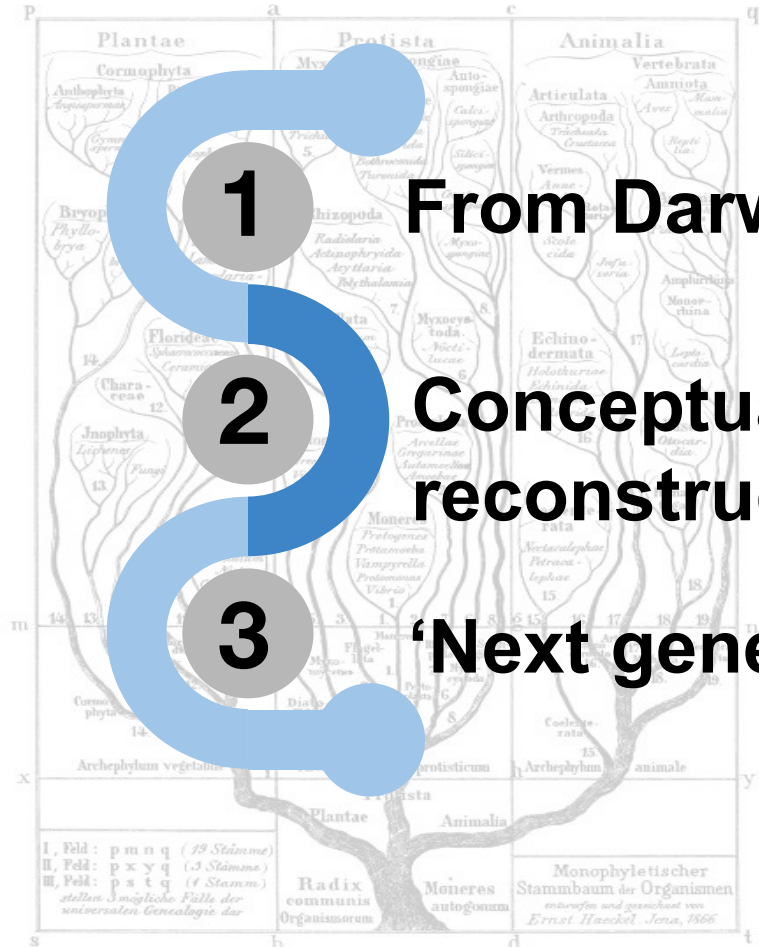
From Darwin to phylogenomics

2

Conceptual framework for phylogenomic reconstruction

3

'Next generation' phylogenomics



Today's menu

1

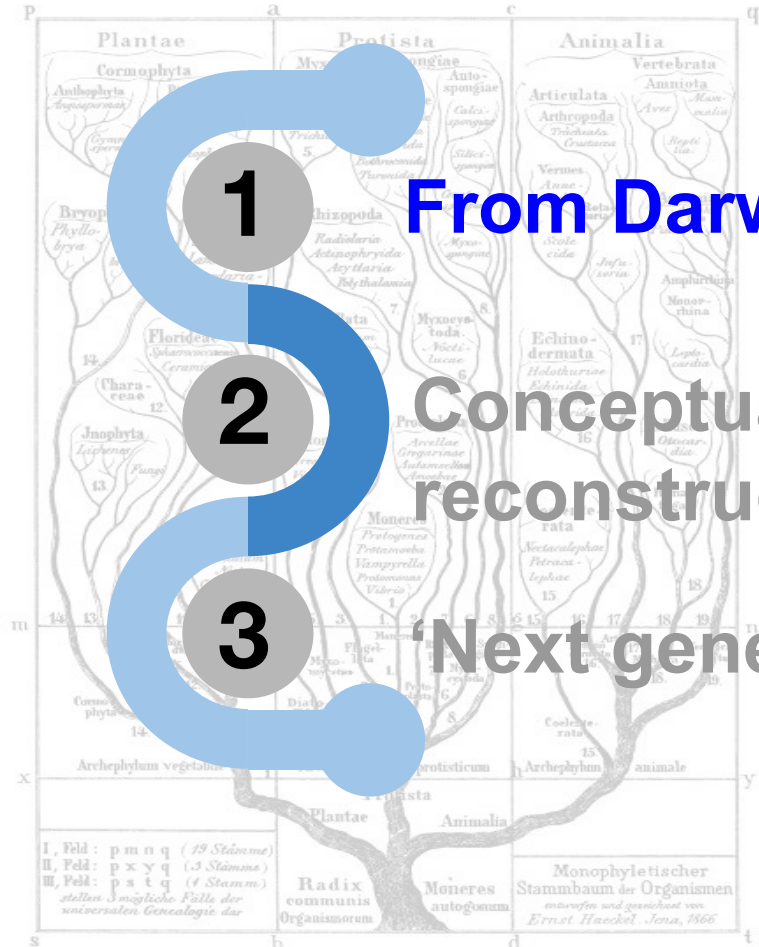
From Darwin to phylogenomics

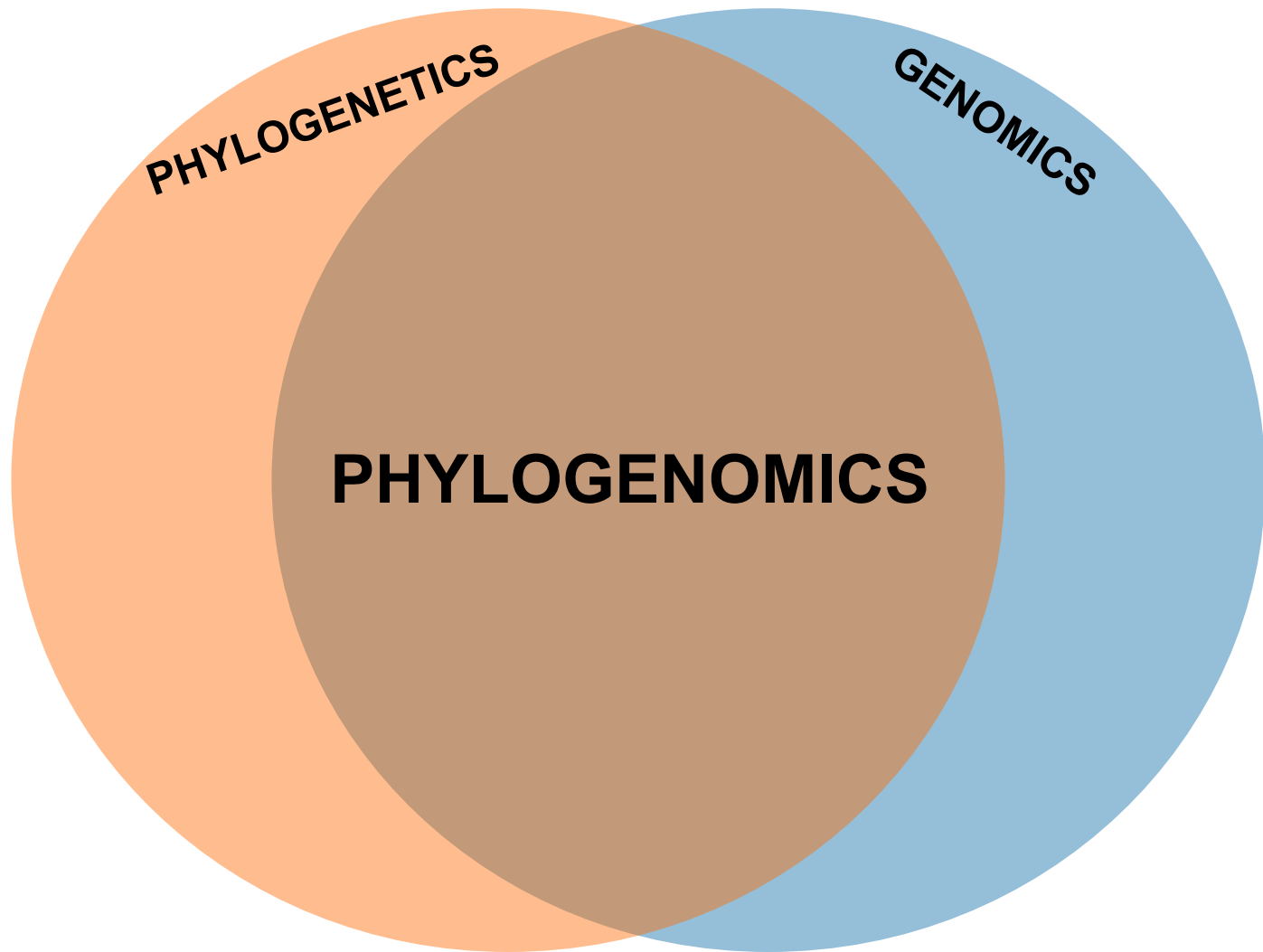
2

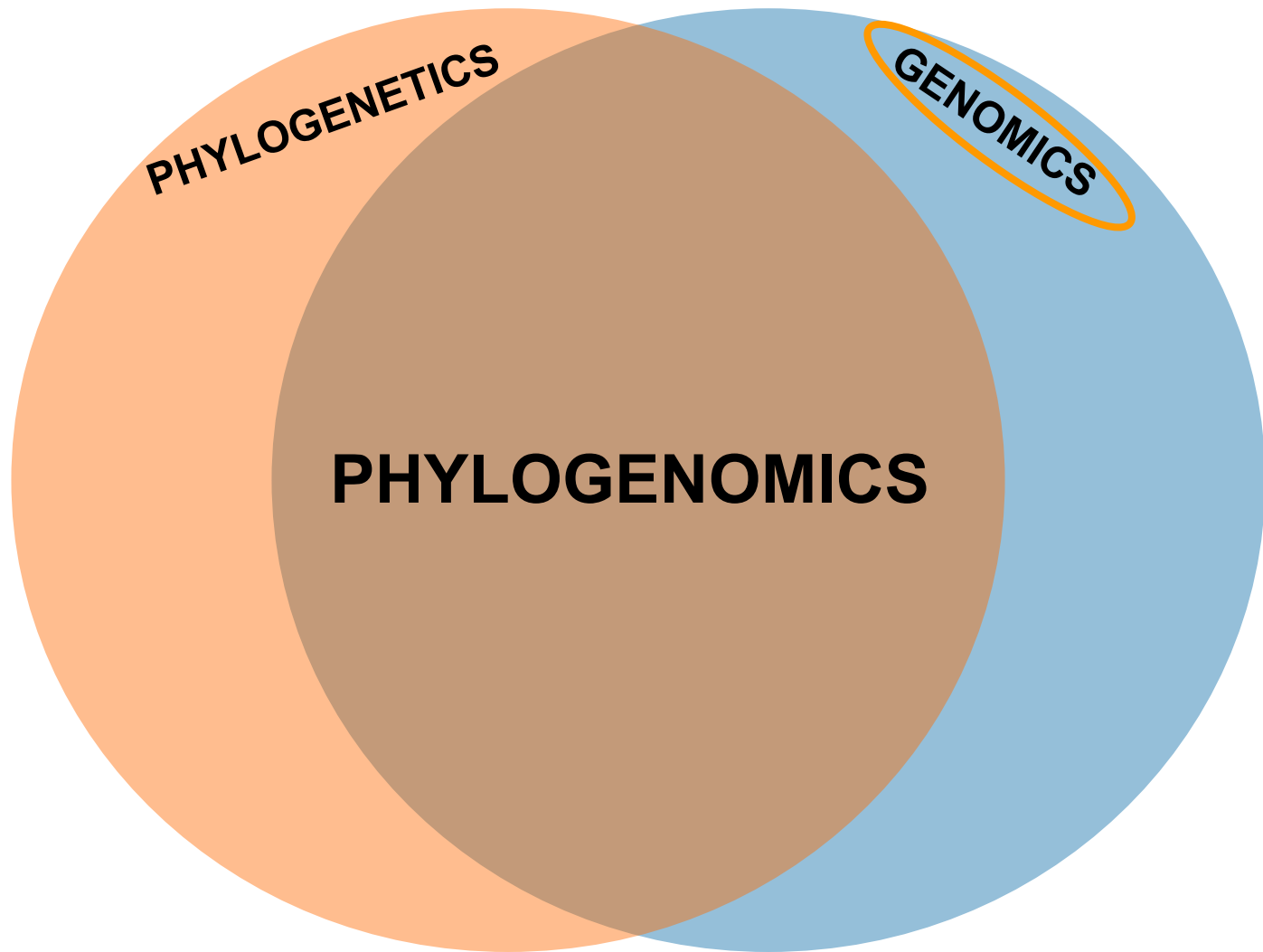
Conceptual framework for phylogenomic reconstruction

3

'Next generation' phylogenomics









Genomics

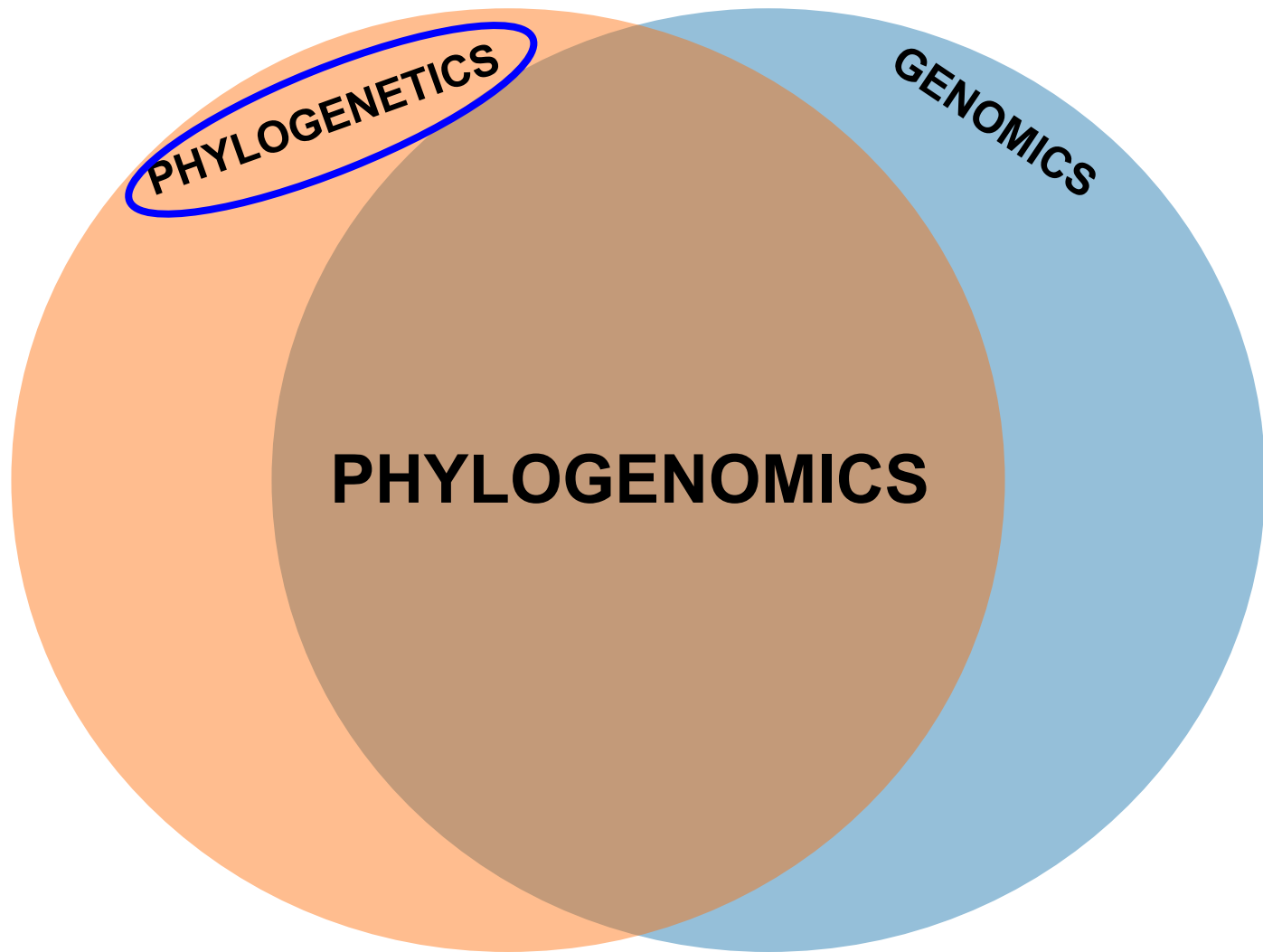
- The study of an organism's complete set of genetic information.
- The genome includes both genes (coding) and non-coding DNA.
- 'Genome': the complete genetic information of an organism.

VS



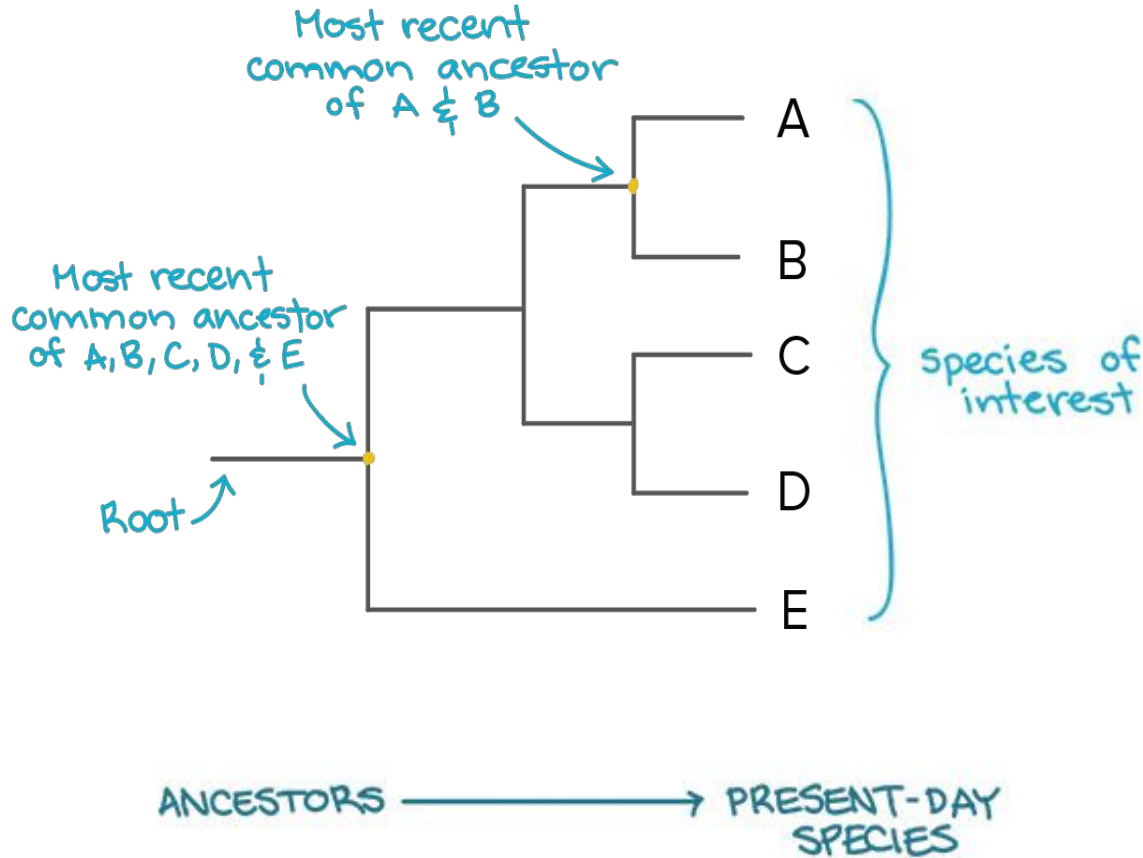
Genetics

- 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.



What is a phylogeny...?

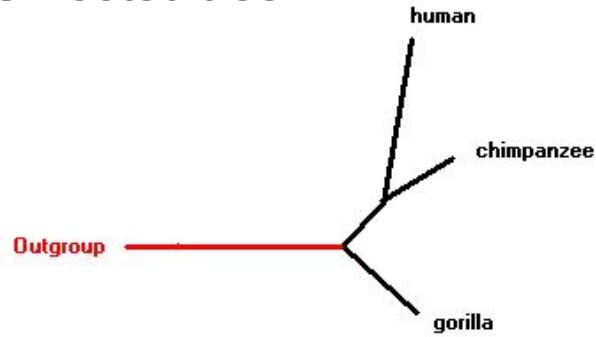
What is a phylogeny...?



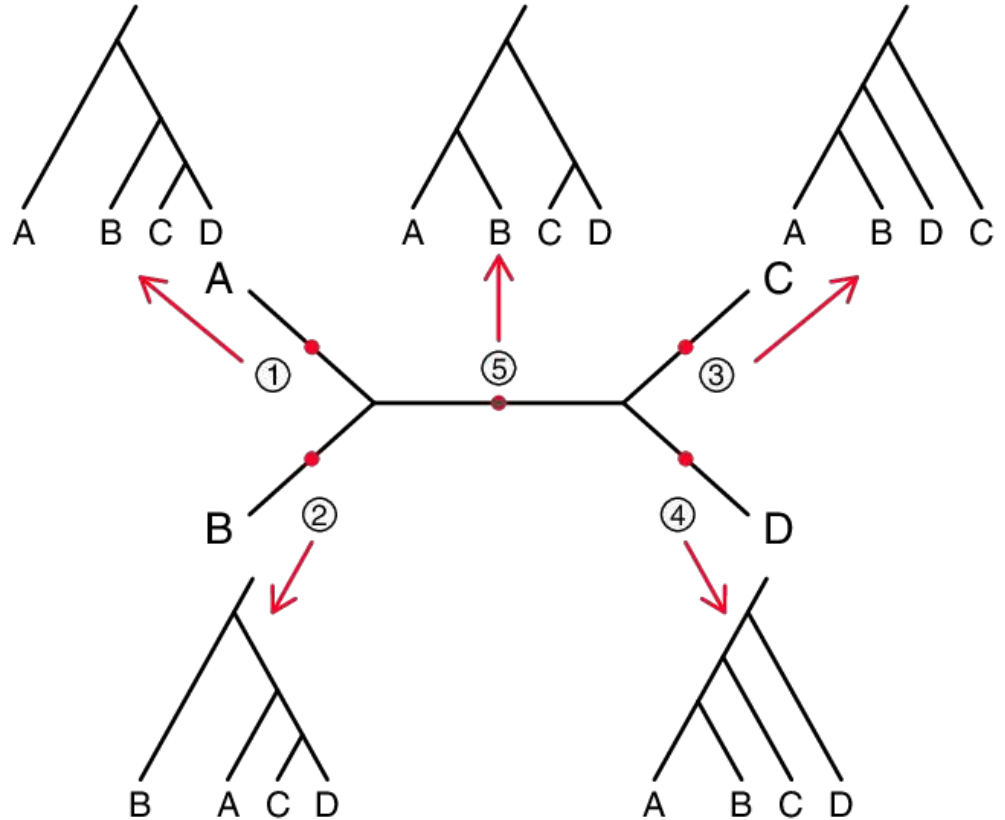
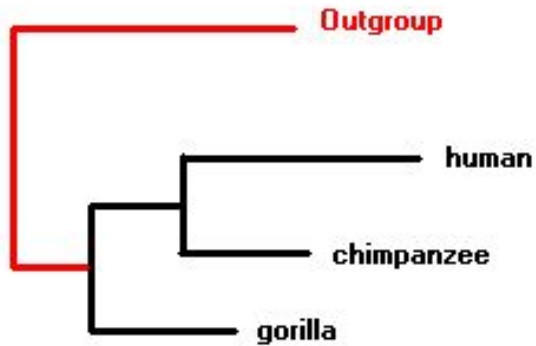
A **phylogenetic tree** is a hypothesis of how species or genes are related through evolution

What is a phylogeny...?

Unrooted tree

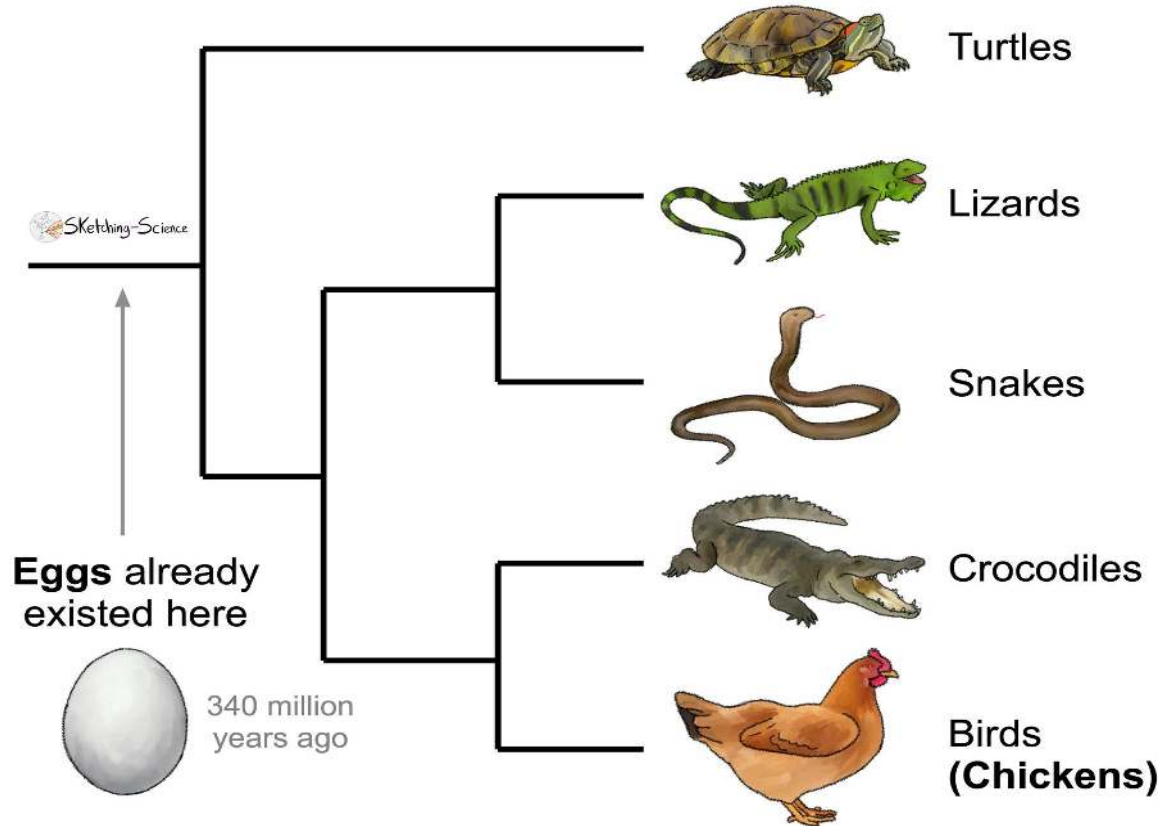


Rooted tree

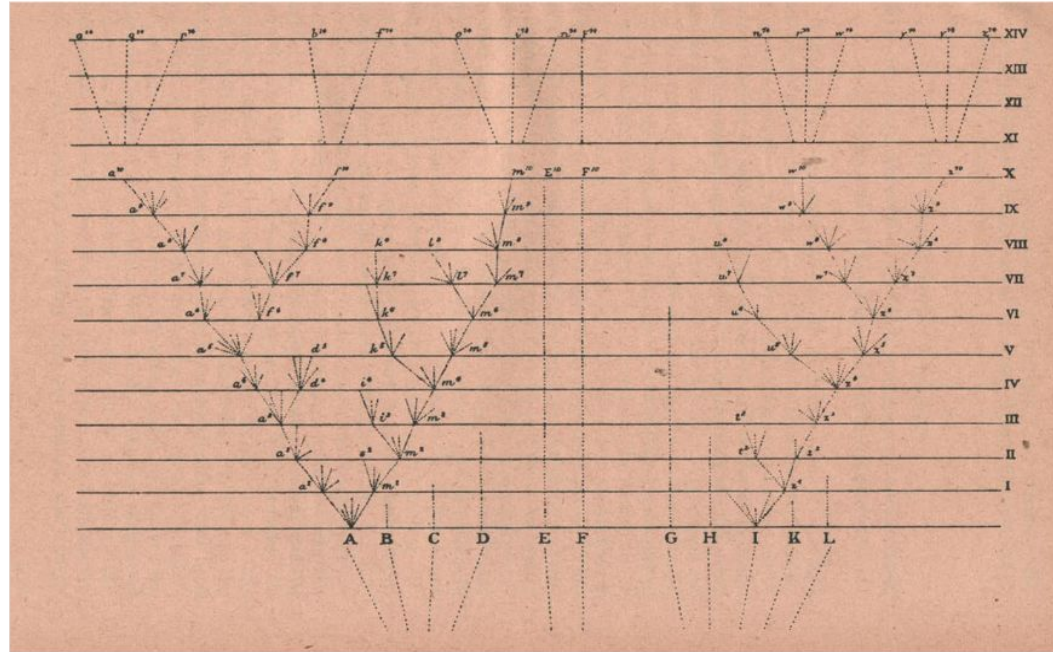


What is a phylogeny, why is it important...?

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

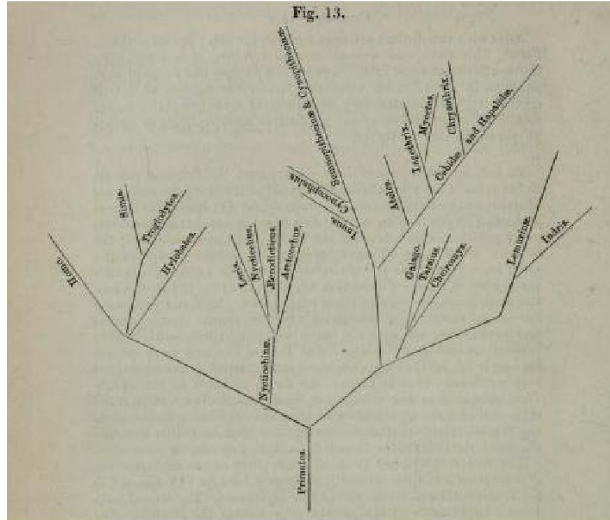
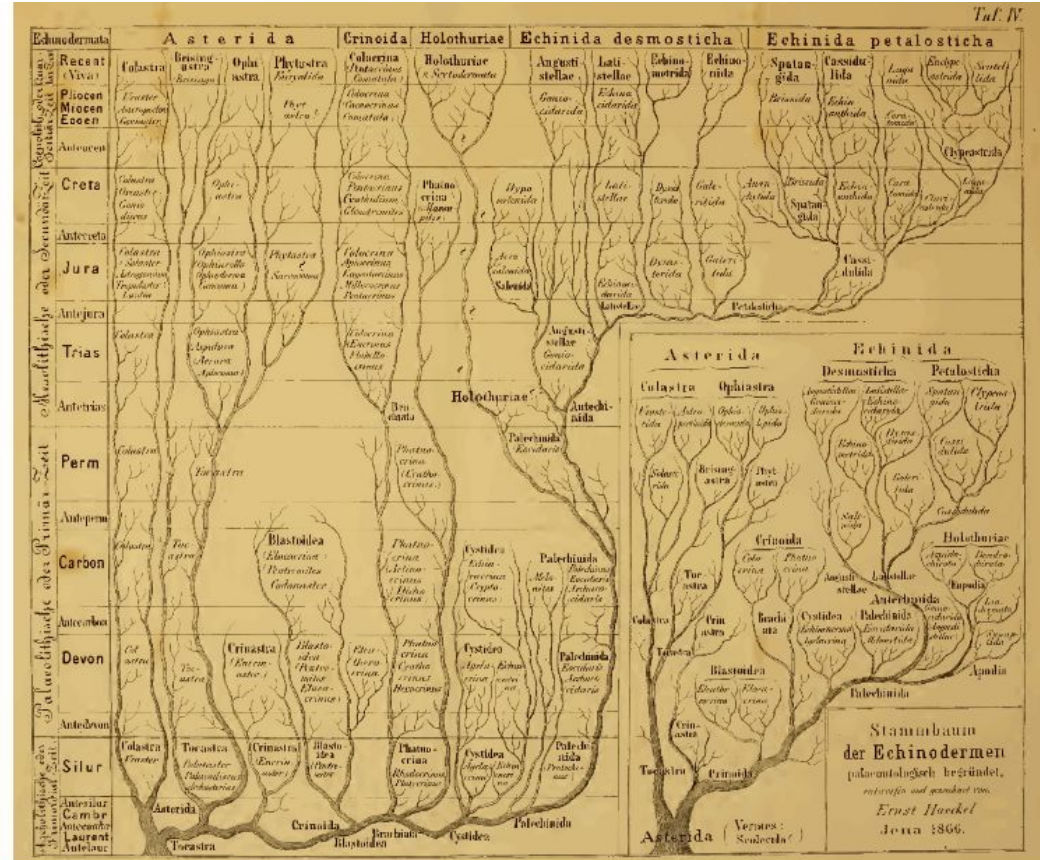
and instinct as the summing up of many contrivances, each useful to the possessor, nearly in the same way as when we look at any great mechanical invention as the summing up of the labour, the experience, the reason, and even the blunders of numerous workmen; when we thus view each organic being, how 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)

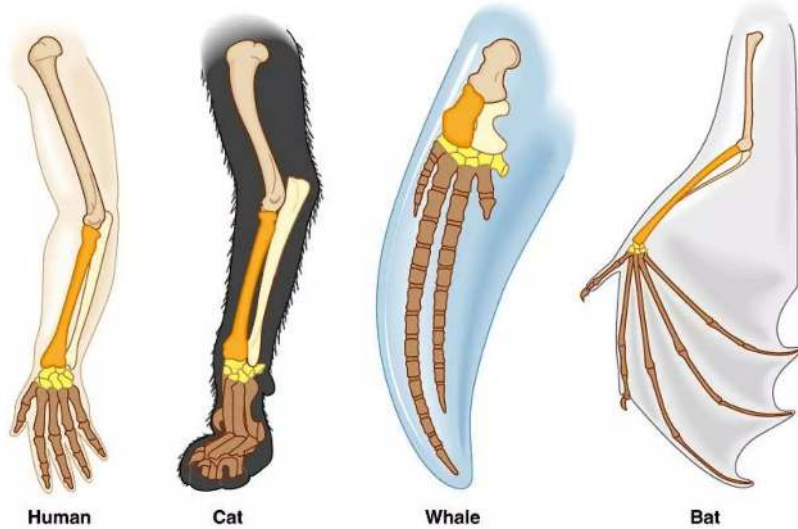


Mivart (1865) Proc. Zool. Soc. London

Haeckel (1866)

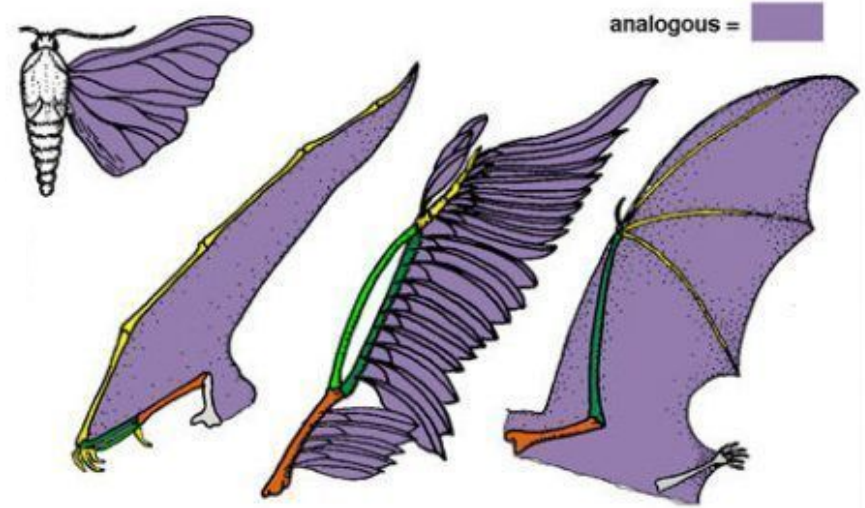
What is a phylogeny, why is it important... and how do you build one?

Homologous Structures



VS

Analogous Structures

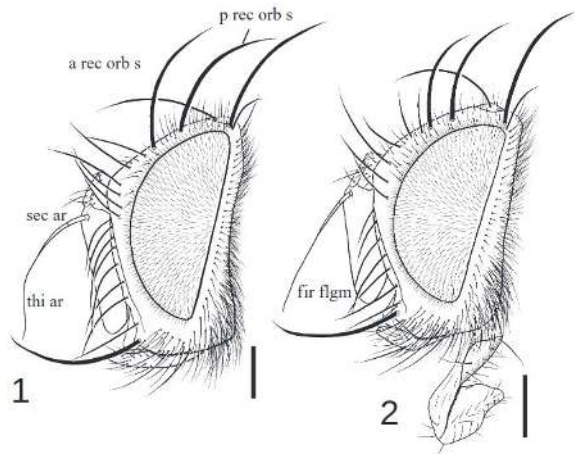


What is a phylogeny, why is it important... and how do you build one?

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.

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 R₄₊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.

Table 3. Morphological data matrix used for phylogenetic analysis

Taxa	Characters		
	0000000001	1111111112	2222222223
	1234567890	1234567890	1234567890
<i>Winthemia venusta</i>	0000000000	0000000000	~000001000
<i>Drinomyia hokkaidensis</i>	1000100001	0100000000	~000002000
<i>Phorocerosoma vicarium</i>	0000100000		
<i>Austrophorocera grandis</i>	0120100000		
<i>A. hirsuta</i>	0020100000		
<i>Bessa parallela</i>	1021101000		
<i>B. remota</i>	1021101000		

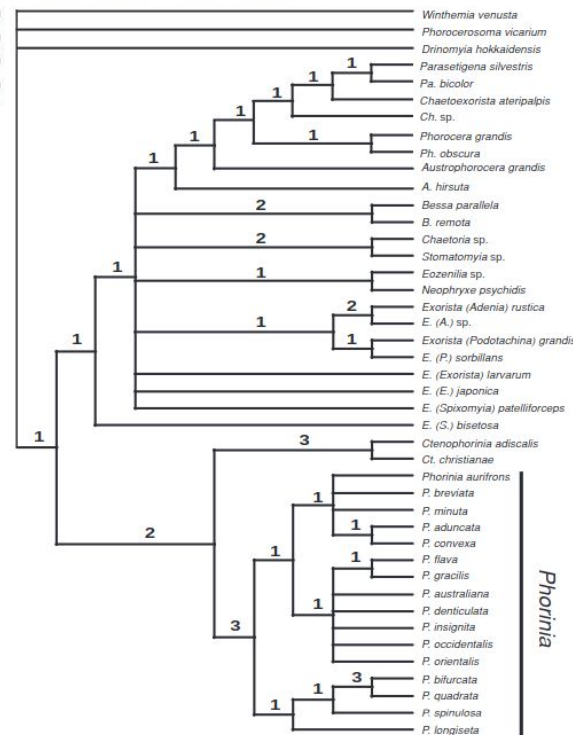
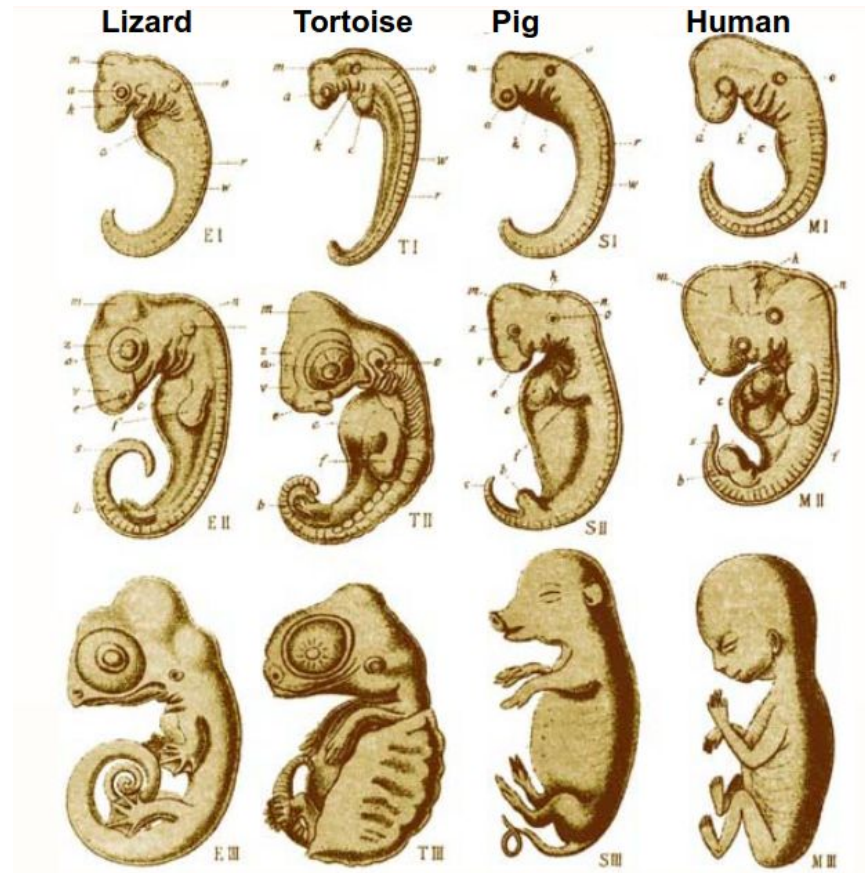
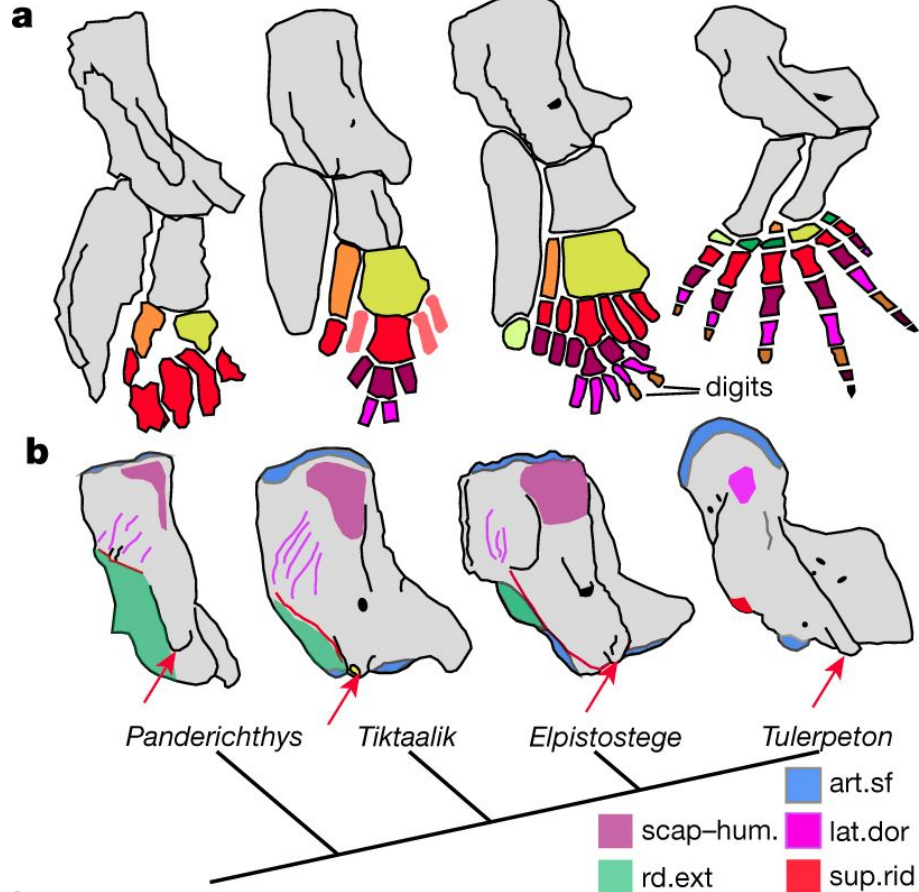


Fig. 79. Strict consensus of 186 equally most parsimonious cladograms (length = 66, consistency index (CI) = 0.530, rescaled consistency index (RC) = 0.462) generated from an analysis of thirty-one morphological characters. Bremer support values are given on the branches.

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



The origin of molecular phylogenetics

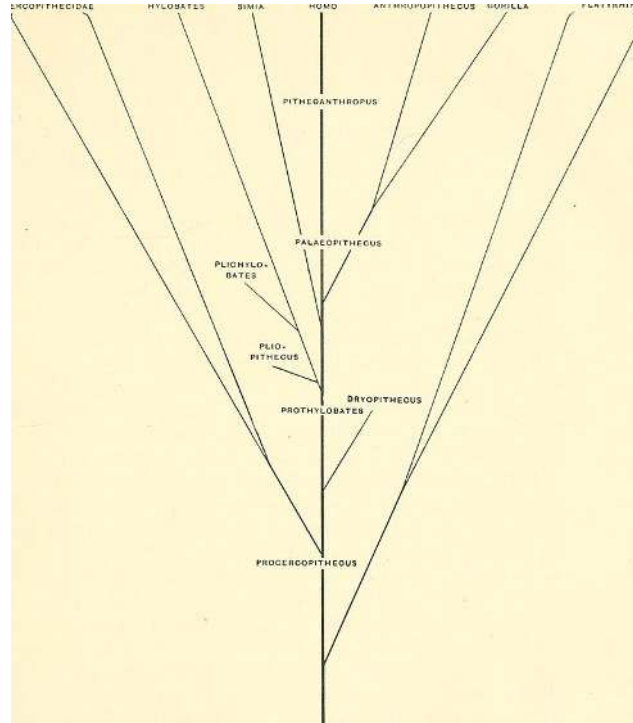
Nuttall (1904) - serological cross-reactions were stronger for more closely related organisms -> phylogeny of apes

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



The origin of molecular phylogenetics

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

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

BLOOD IMMUNITY AND BLOOD RELATIONSHIP

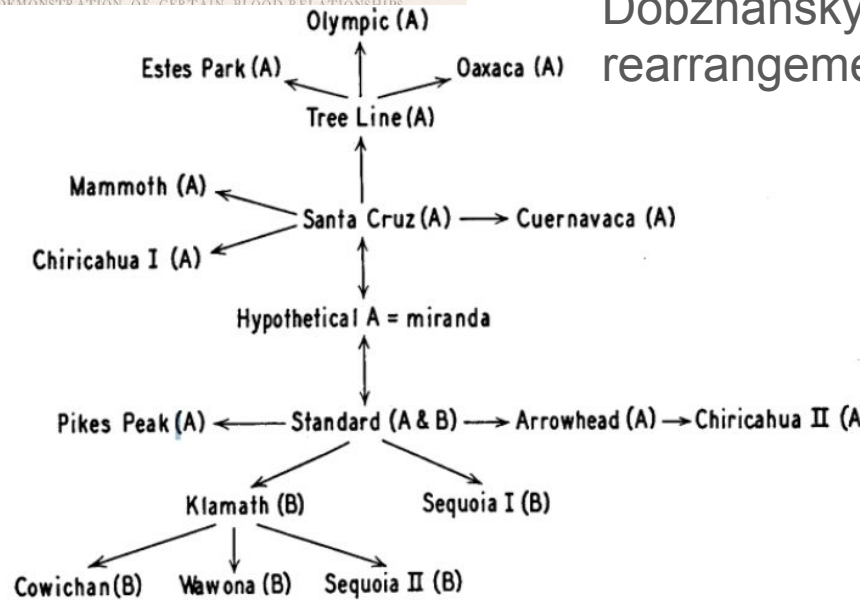
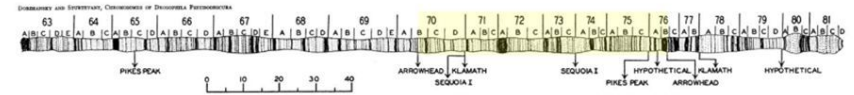
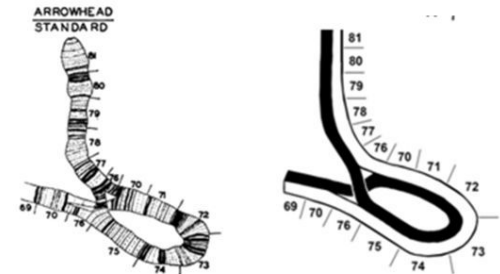


FIGURE 3.—Phylogeny of the gene arrangements in the third chromosome of *Drosophila pseudoobscura*. Any two arrangements connected by an arrow in the diagram differ by a single inversion. Further explanation in text.



Chromosome 3 of *Drosophila pseudoobscura*



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

The origin of molecular phylogenetics

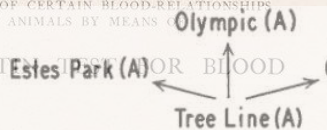
Nuttall (1904) - serological cross-reactions were stronger for more closely related organisms -> phylogeny of apes

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

BLOOD IMMUNITY AND BLOOD RELATIONSHIP

A DEMONSTRATION OF CERTAIN BLOOD RELATIONSHIPS
AMONGST ANIMALS BY MEANS OF

THE PRECIPITATION TEST FOR BLOOD



ELSEVIER

Journal of Theoretical Biology

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



Zuckerlandl &
Pauling (1965) -

Abstract

Molecule history ☆

Emile Zuckerlandl, I

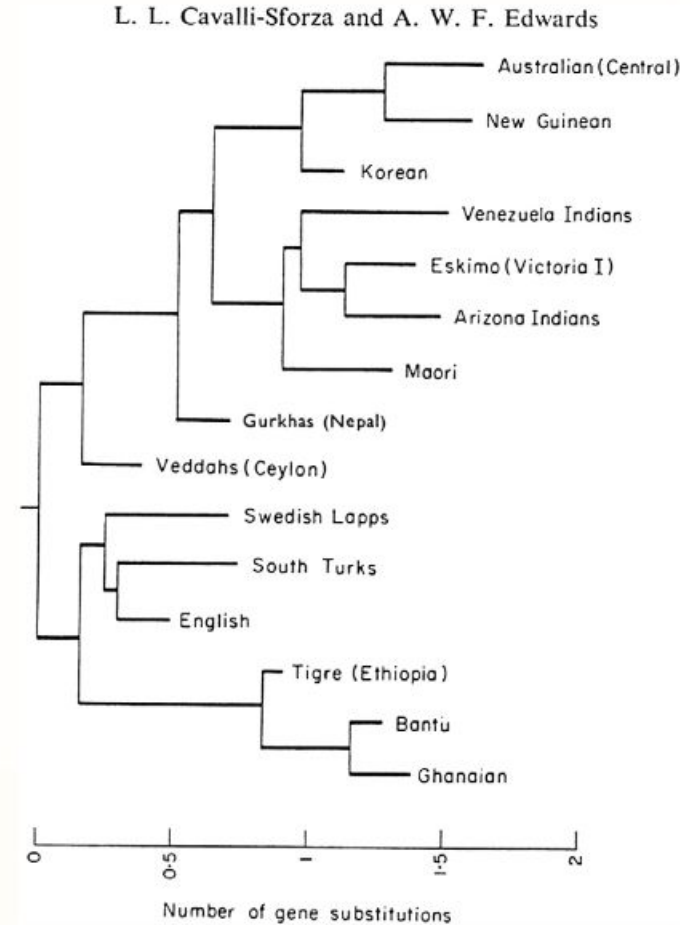
Different types of molecules are discussed in relation to their fitness for providing the basis for a molecular phylogeny. Best fit are the “semantides”, i.e. the different types of macromolecules that carry the genetic information or a very extensive translation thereof. The fact that more than one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

version. Further explanation in

Molecular phylogenetics: the new wave



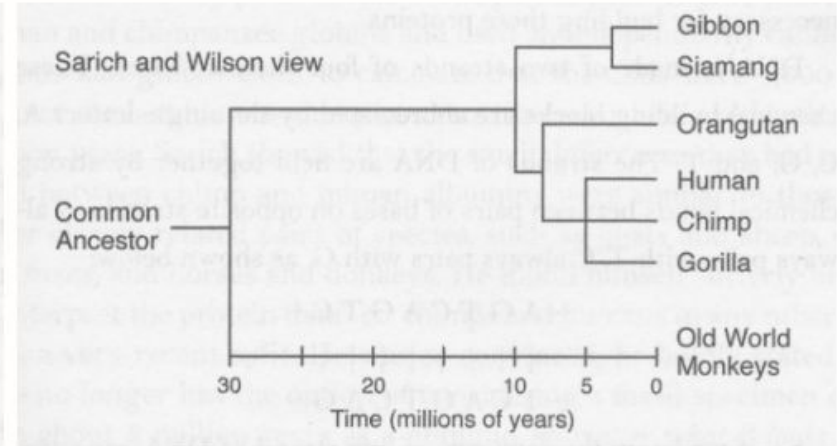
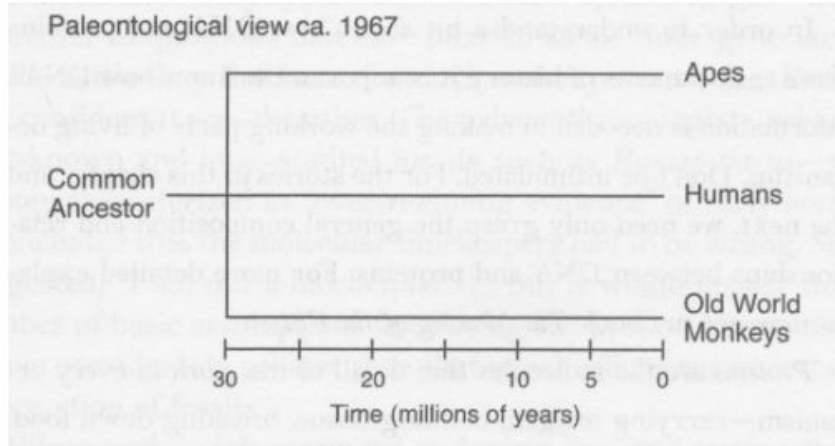
Phylogeny inferred from blood group allele frequencies from 15 populations



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

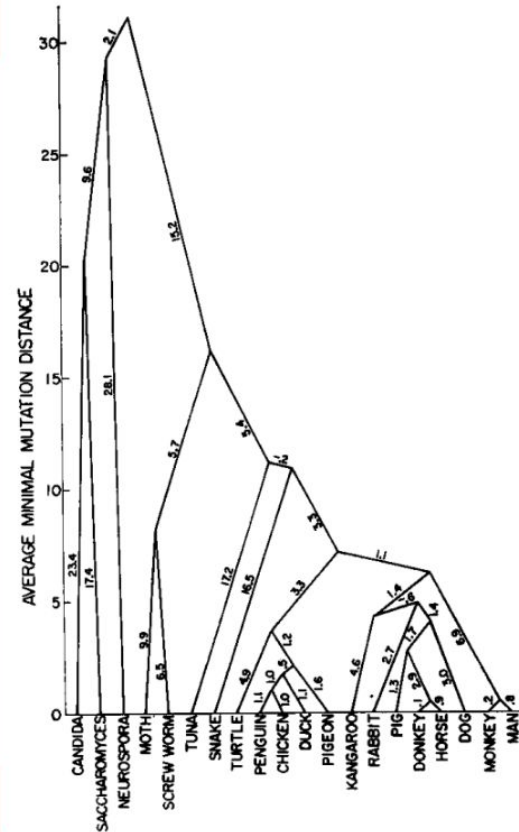
Construction of Phylogenetic Trees

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 homol-

ogous 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



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

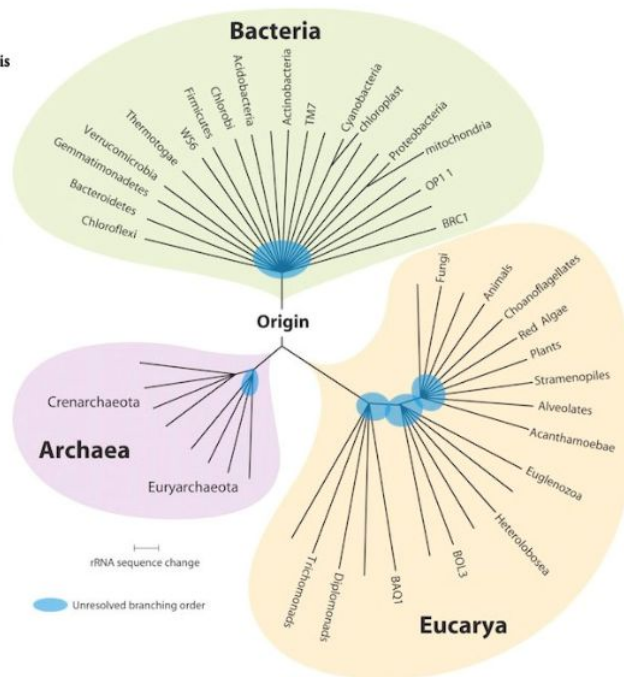
(archaeobacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

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 archaeobacteria, containing methanogenic bacteria; and (iii) the urkaryotes, now represented in the cytoplasmic component of eukaryotic cells.



The dawn of phylogenomics



Insight/Outlook

Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis

Jonathan A. Eisen¹

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

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

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

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

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

Sequence Similarity, Homology, and Functional Predictions

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

Phylogenomics: prediction of gene function and gene family evolution

The dawn of phylogenomics

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

GENOME RESEARCH 163

Insight/Outlook

Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis

Jonathan A. Eisen¹

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

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

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

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

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

Phylogenomics: prediction of gene function and gene family evolution

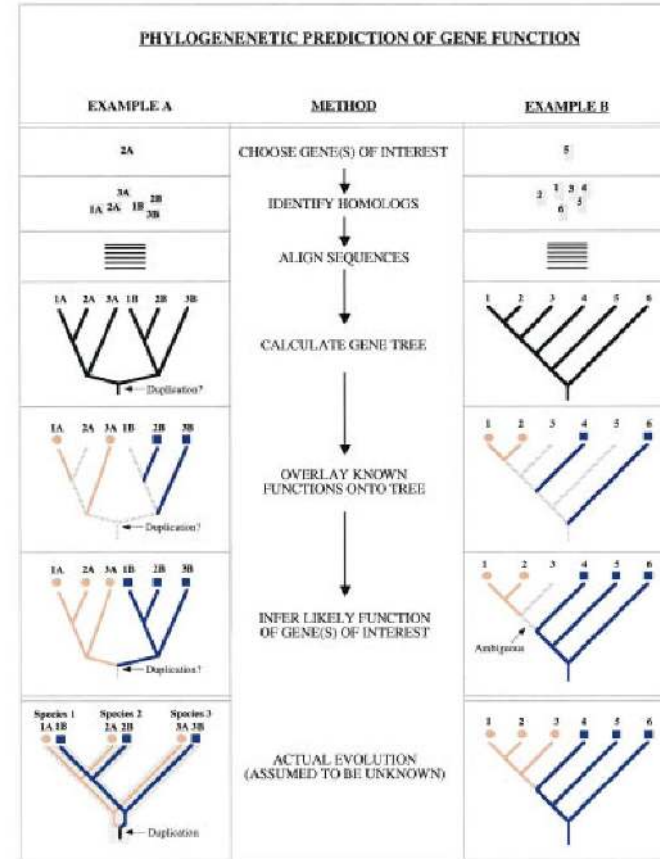


Figure 1 Outline of a phylogenomic methodology. In this method, information about the evolutionary relationships among genes is used to predict the functions of uncharacterized genes (see text for details). Two hypothetical scenarios are presented and the path of trying to

The dawn of phylogenomics

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

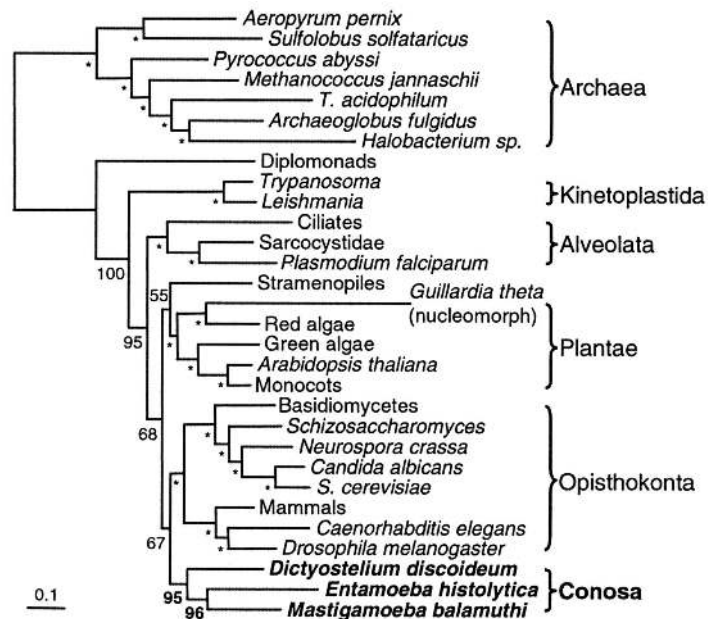
www.pnas.org/cgi/doi/10.1073/pnas.032662799

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

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

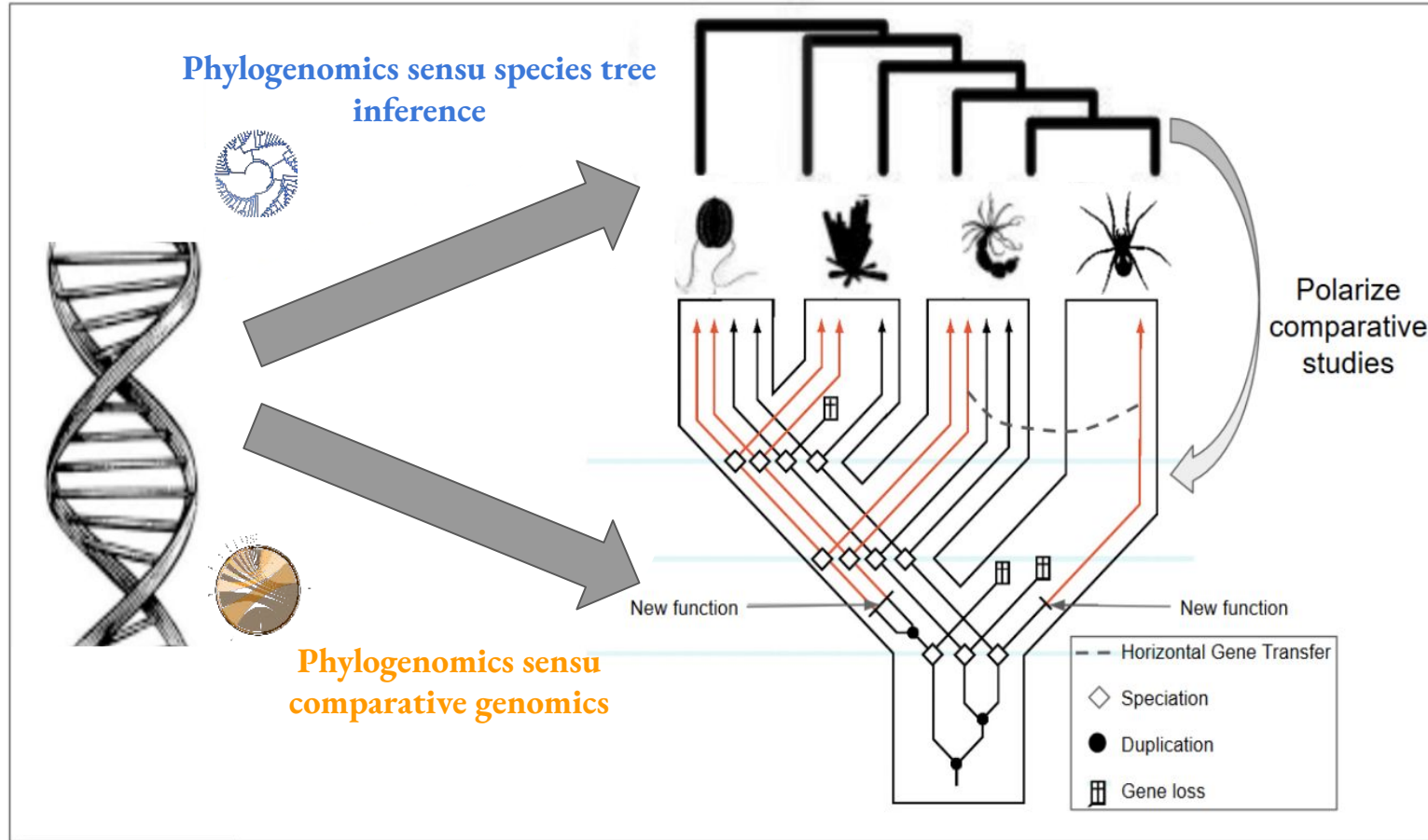
The phylogenetic relationships of amoebae are poorly resolved. To address this difficult question, we have sequenced 1,280 expressed sequence tags from *Mastigamoeba balamuthi* and assembled a large data set containing 123 genes for representatives of three phenotypically highly divergent major amoeboid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

Phylogenomics: species tree inference



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



Content of the lecture

1

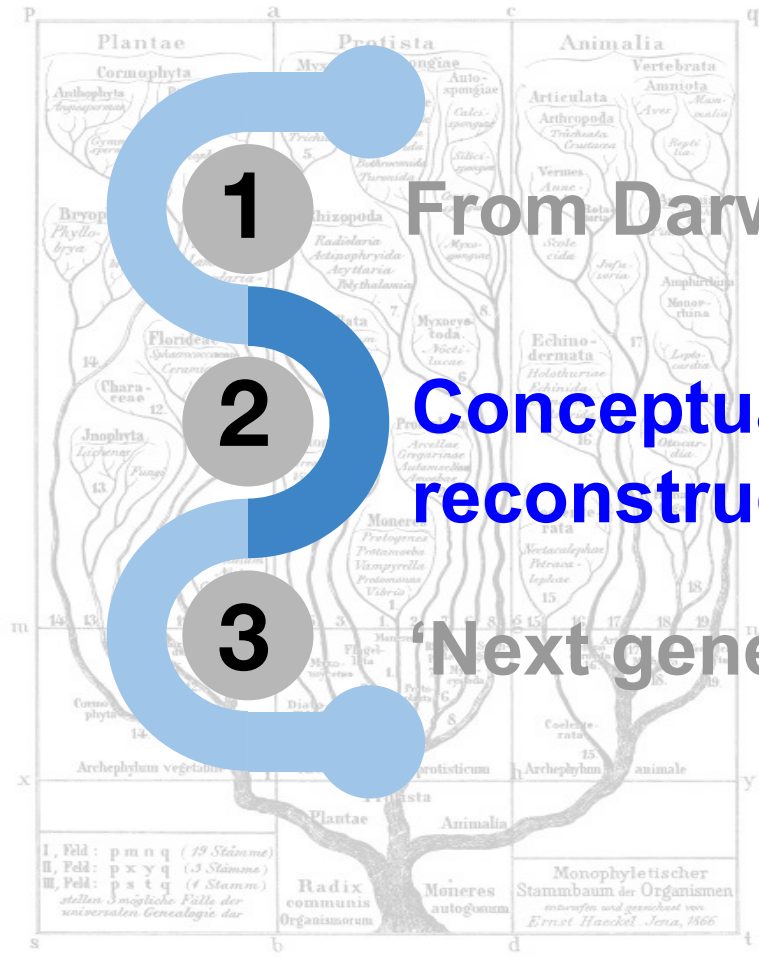
From Darwin to phylogenomics

2

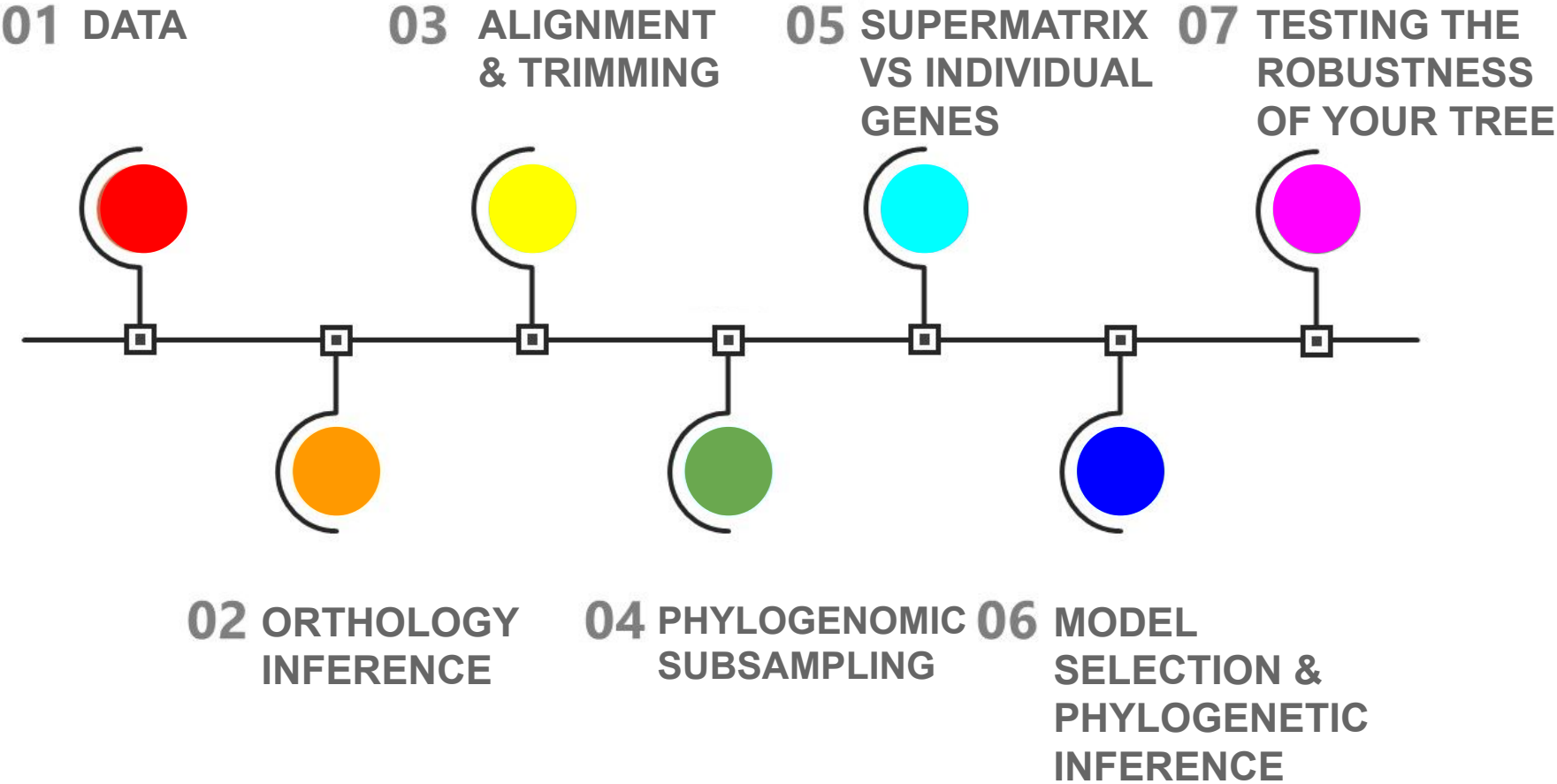
Conceptual framework for phylogenomic reconstruction

3

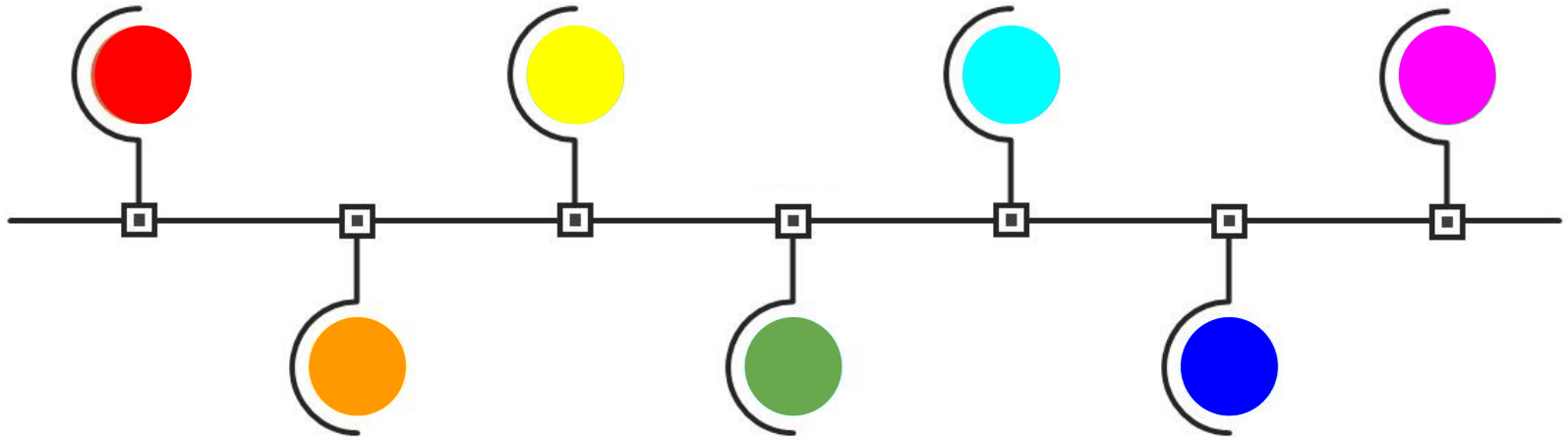
‘Next generation’ phylogenomics



Tutorials and hands-on sessions available at
<https://evomics.org/2024-workshop-on-phylogenomics-cesky-krumlov/>

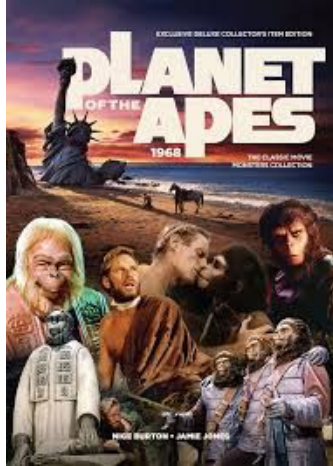


01 DATA



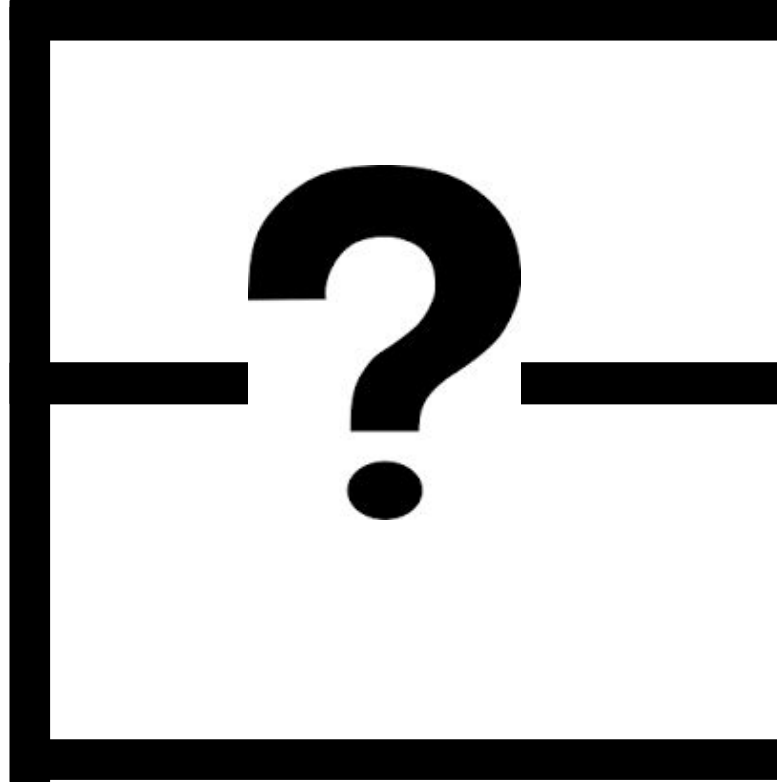
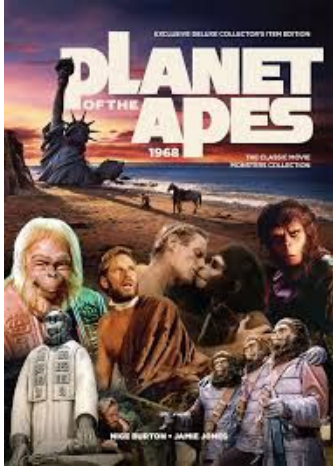
01 DATA

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



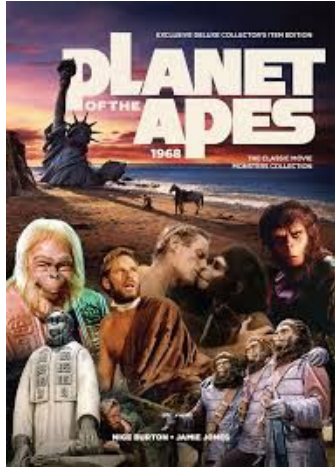
01 DATA

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



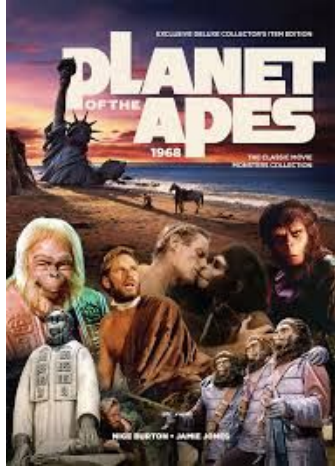
01 DATA

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



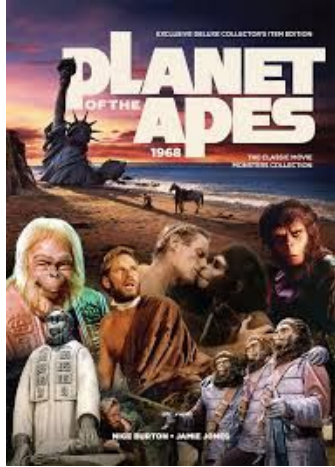
01 DATA

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



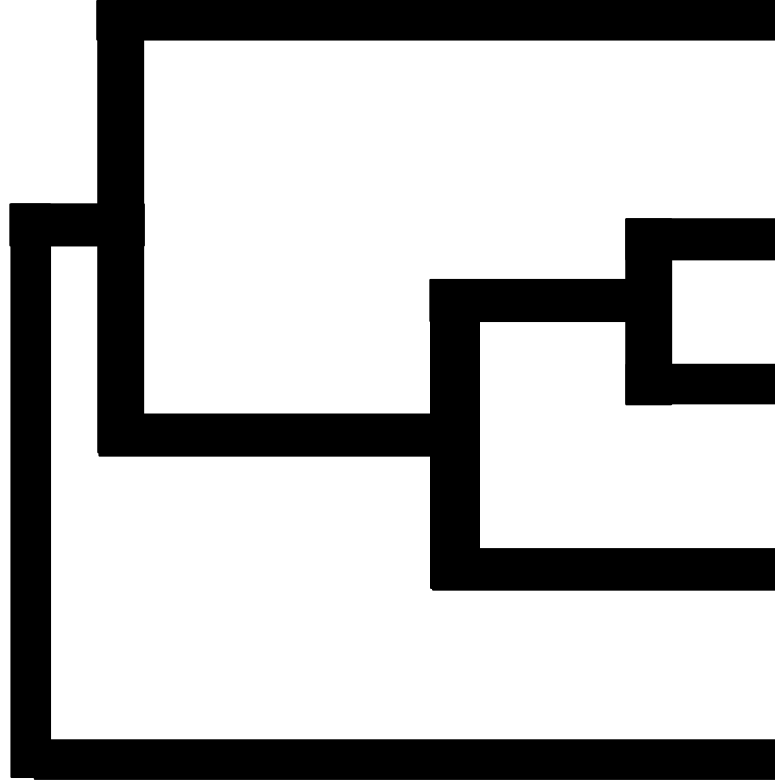
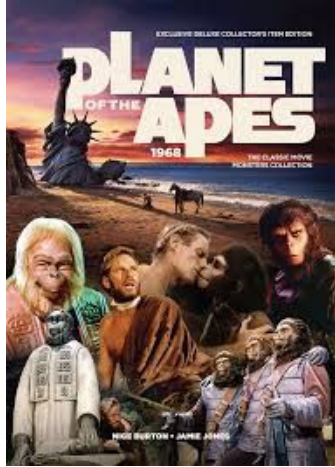
01 DATA

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.



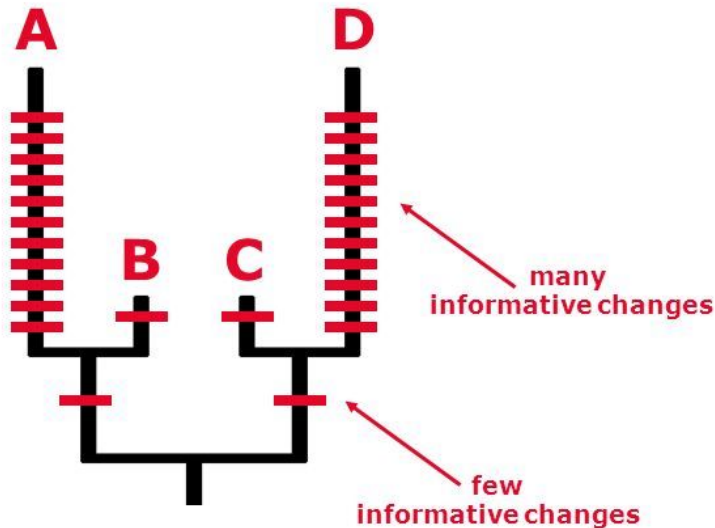
01 DATA

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

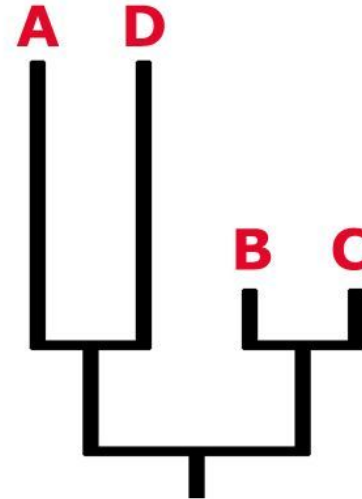
Long Branch Attraction

Outgroups / Fast-evolving lineages / Missing data

True Tree



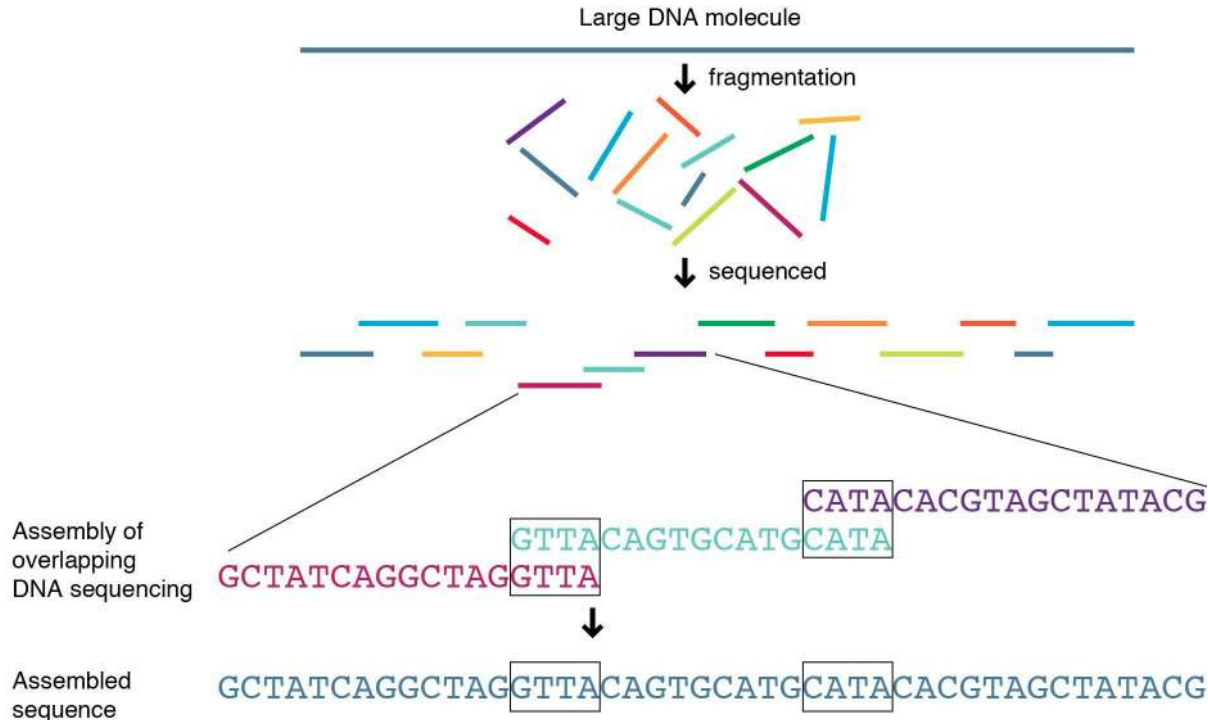
Reconstructed Tree



01 DATA

Source of your data

GENOMES



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

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)

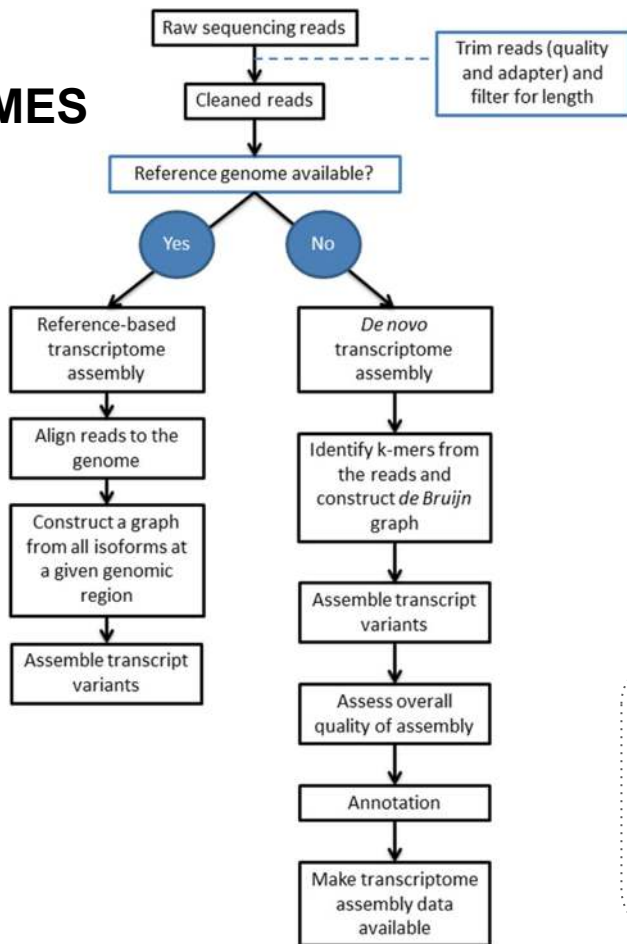
Cons:

- 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

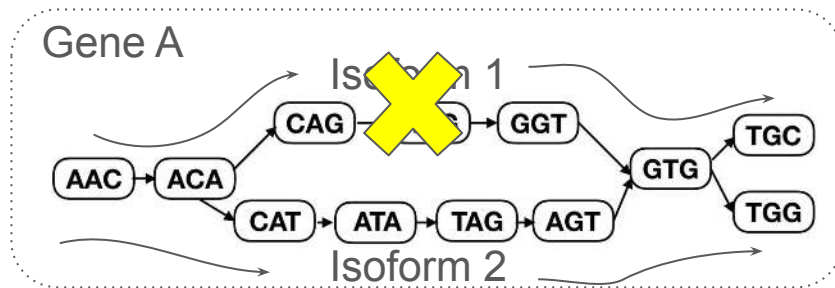
Source of your data

TRANSCRIPTOMES



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

De Bruijn Graph



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

ULTRA-CONSERVED ELEMENTS (UCEs)

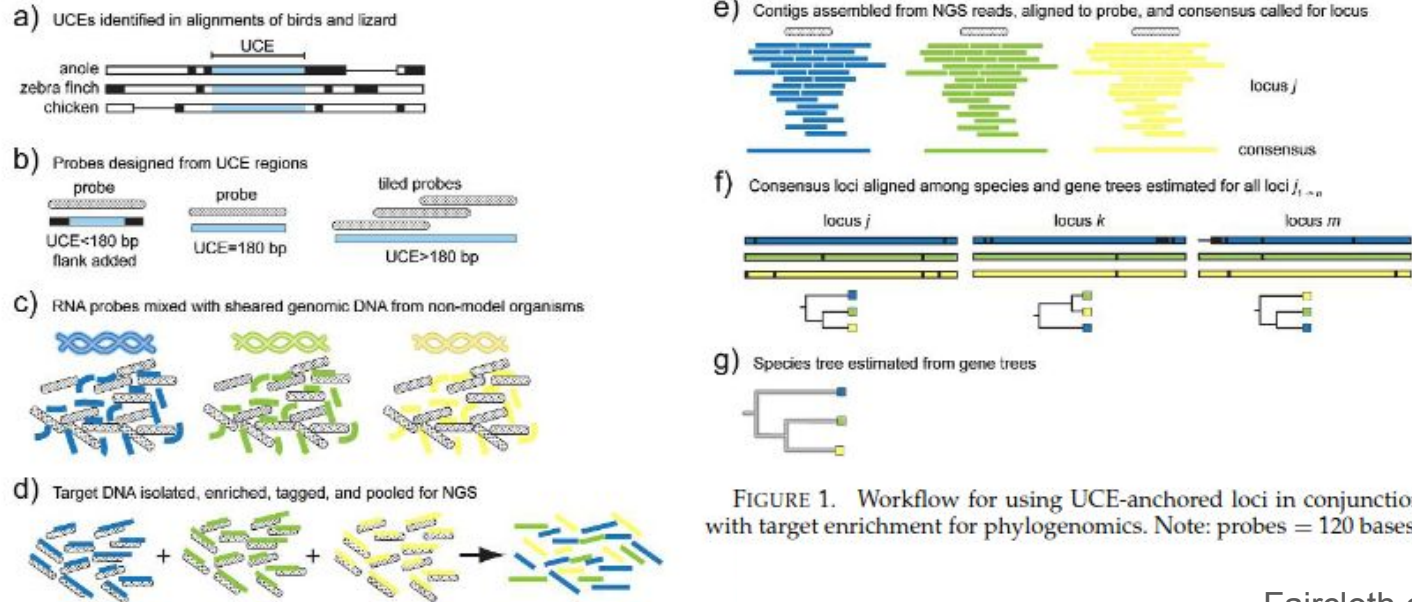


FIGURE 1. Workflow for using UCE-anchored loci in conjunction with target enrichment for phylogenomics. Note: probes = 120 bases.

Faircloth et al. 2012

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

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

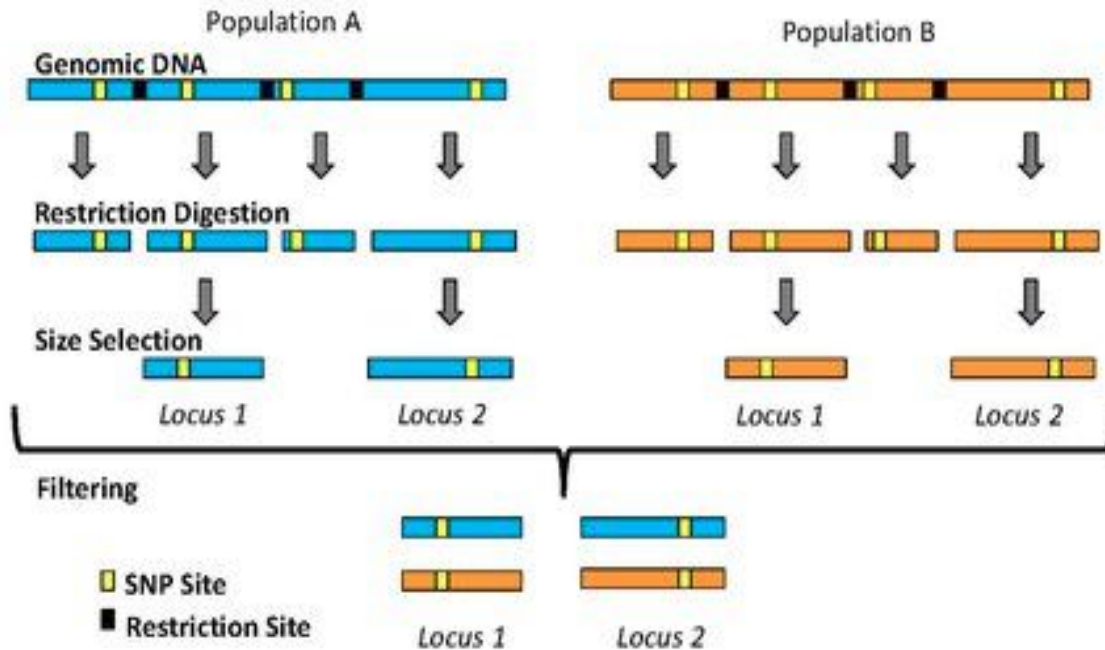
Cons:

- Limited availability of marks 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)



After digestion, sequencing and mapping, this is 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

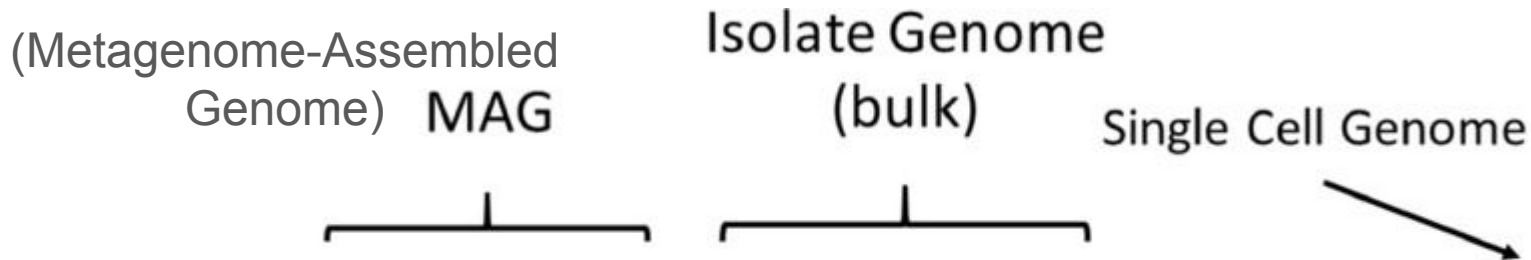
Cons:

- 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 cell, multiple organisms

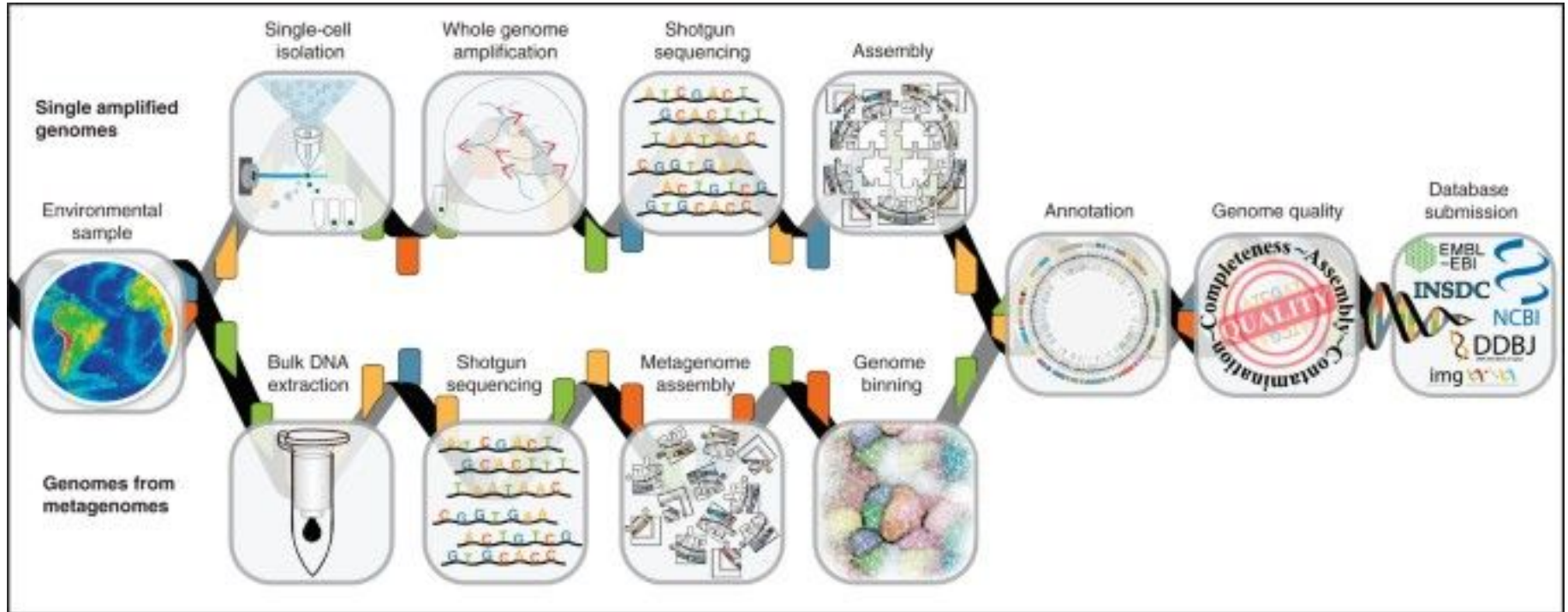
One individual, multiple cells

One cell, one organism

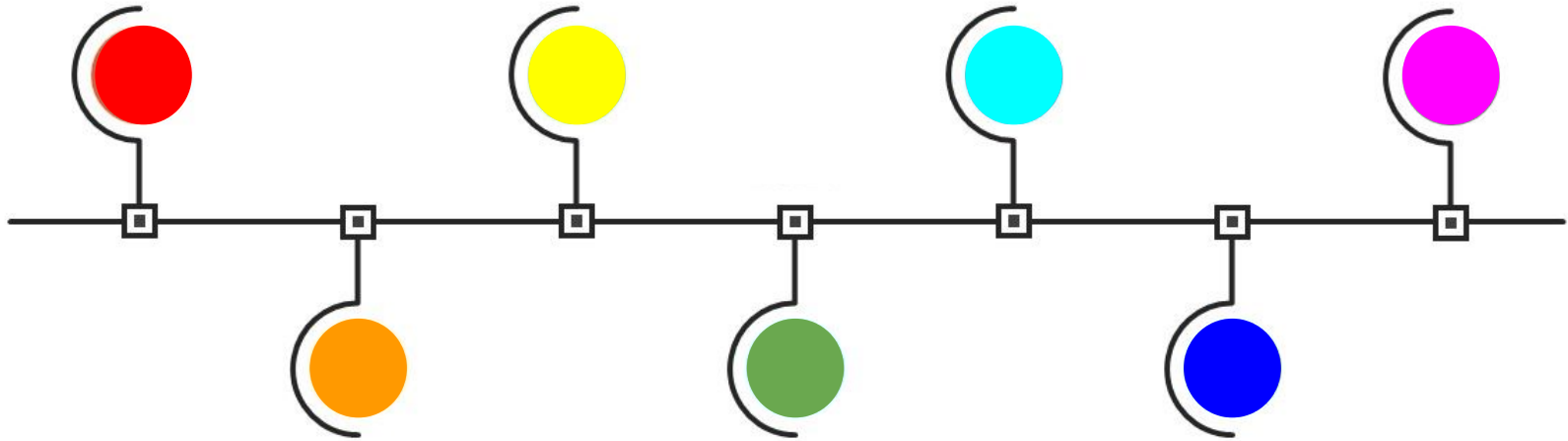
01 DATA

Source of your data

METAGENOMICS - single cell vs MAGs



01 DATA

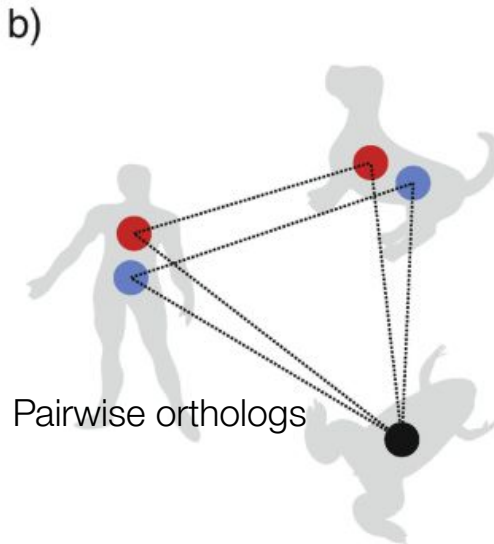
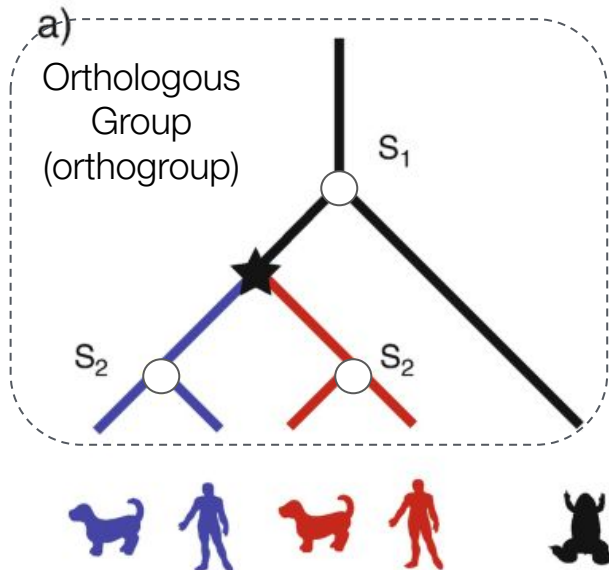


**02 ORTHOLOGY
INFERENCE**

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*

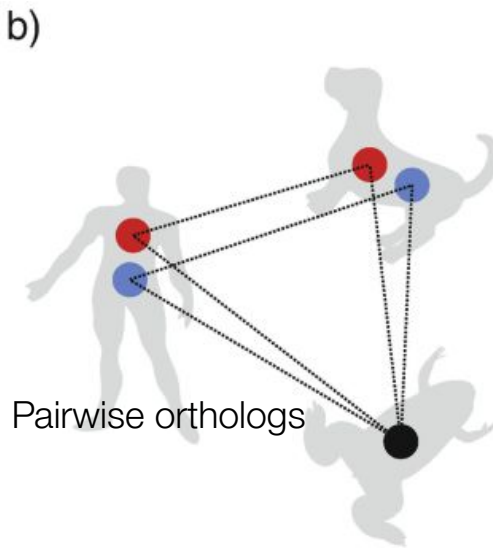
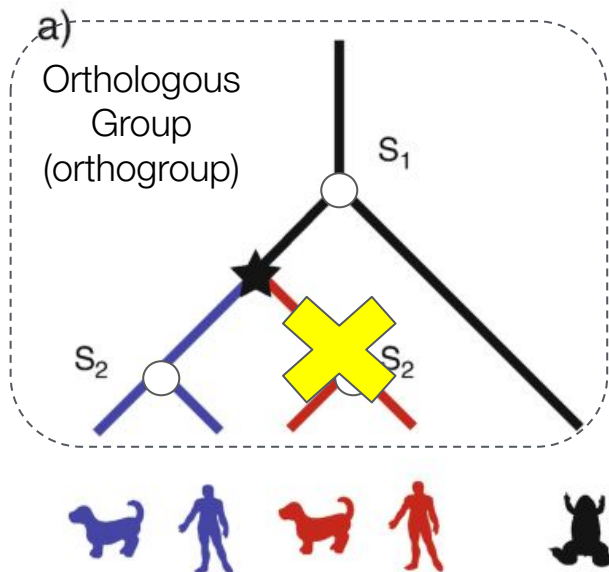
When we have multiple species, we should consider the concept of *orthogroup*

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

02 ORTHOLOGY INFERENCE

Definitions

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



Software:

- OrthoFinder
- OMA
- TOGA (synteny; vertebrates)

Orthology relationships are inferred *pairwise*

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

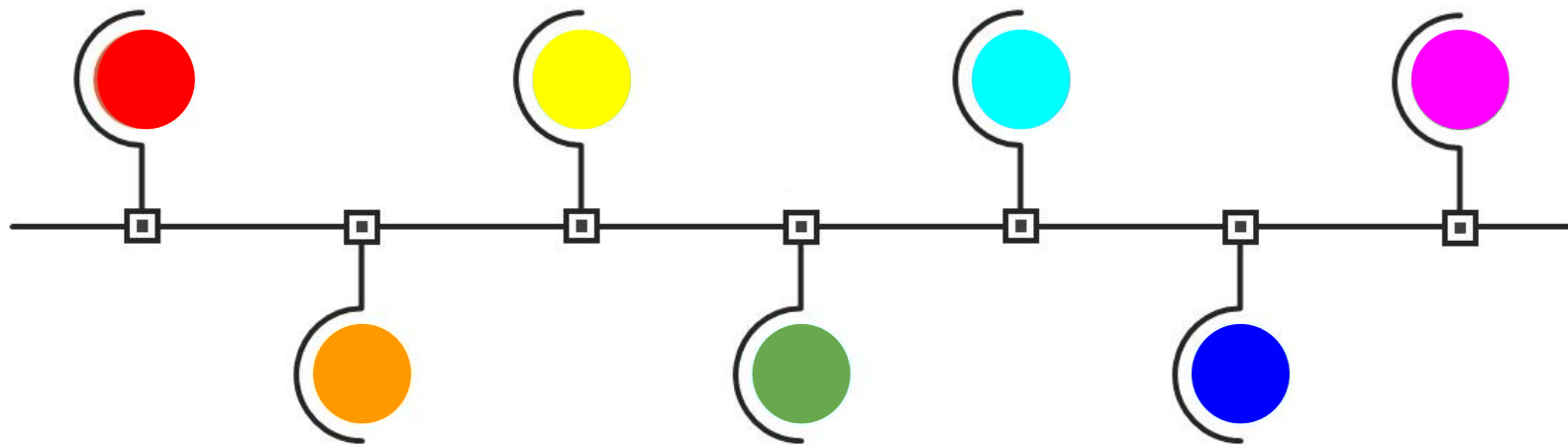
For phylogenomic inference, we want either:

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

01 DATA

**03 ALIGNMENT
& TRIMMING**

**02 ORTHOLOGY
INFERENCE**



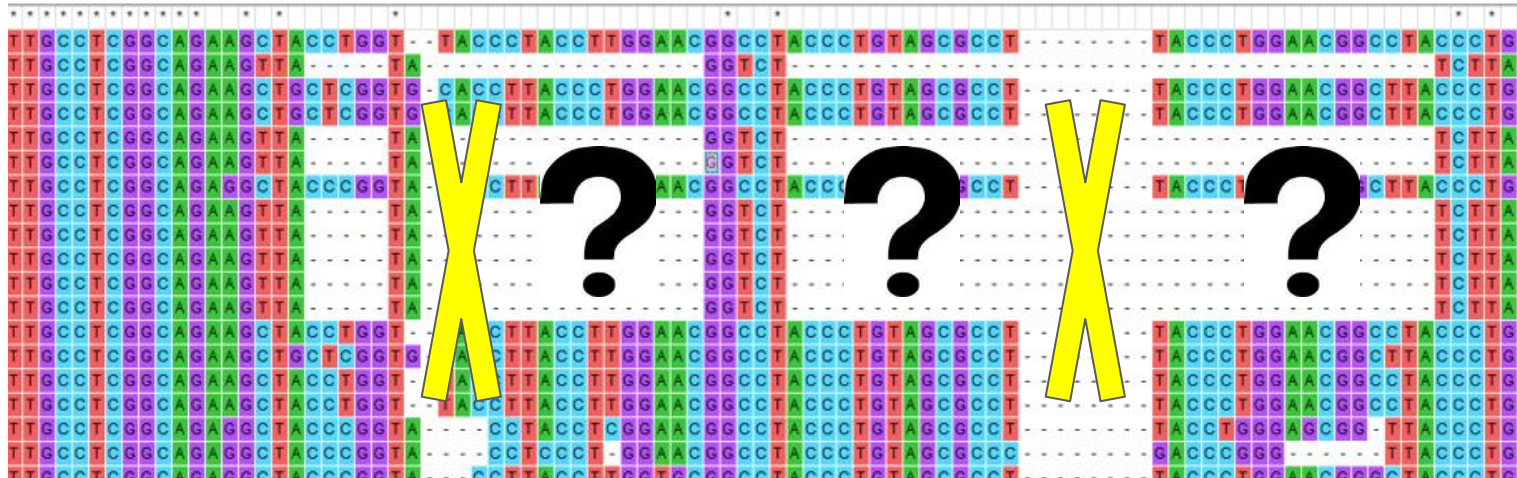
03 ALIGNMENT AND TRIMMING

Software:

- Muscle5, MAFFT
- PhyKIT, trimAL

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

Software:

- Muscle5, MAFFT
- PhyKIT, trimAL

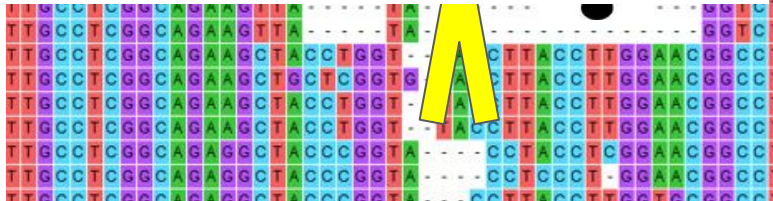
The goal of the alignment procedure should be to identify the events associated with the homologies, so that the aligned sequences accurately reflect those events.

Article | [Open access](#) | Published: 15 November 2022

Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny

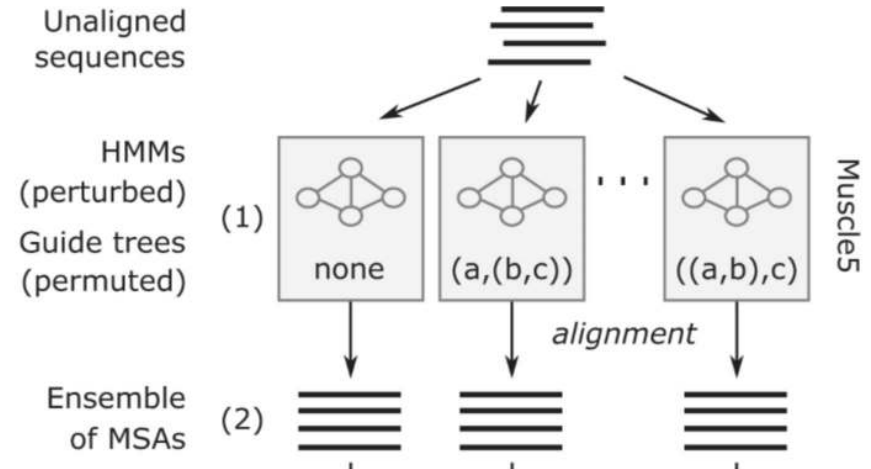
[Robert C. Edgar](#) 

[Nature Communications](#) **13**, Article number: 6968 (2022) | [Cite this article](#)



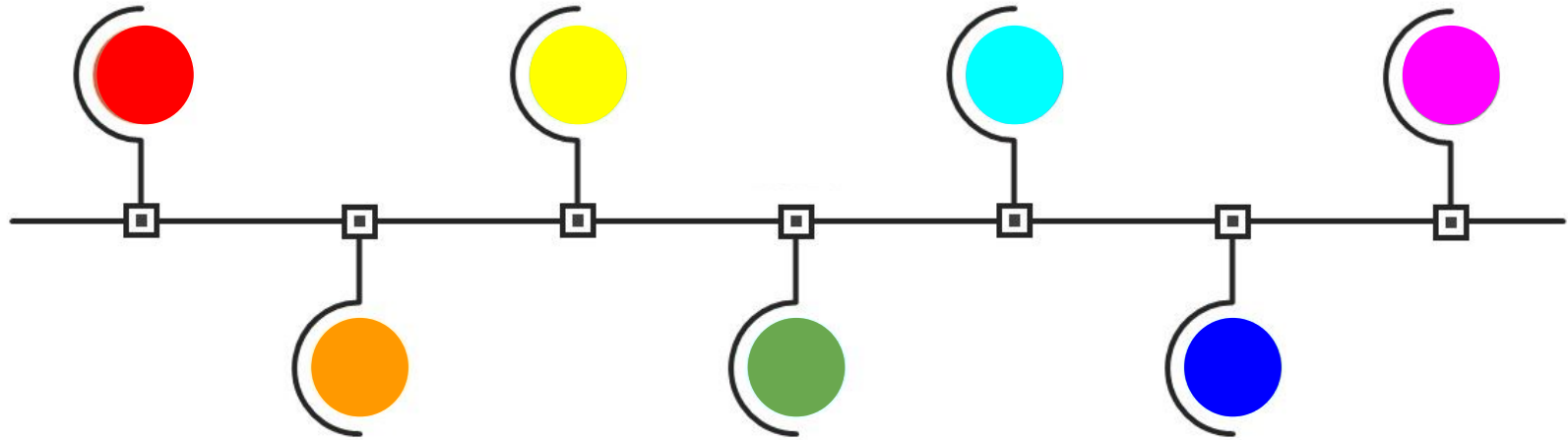
If the sequences are poorly aligned, you may want to consider trimming the poorly aligned areas.

Fig. 1: Typical ensemble workflow for alignment and phylogeny assessment.



01 DATA

**03 ALIGNMENT
& TRIMMING**



**02 ORTHOLOGY
INFERENCE**

**04 PHYLOGENOMIC
SUBSAMPLING**

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

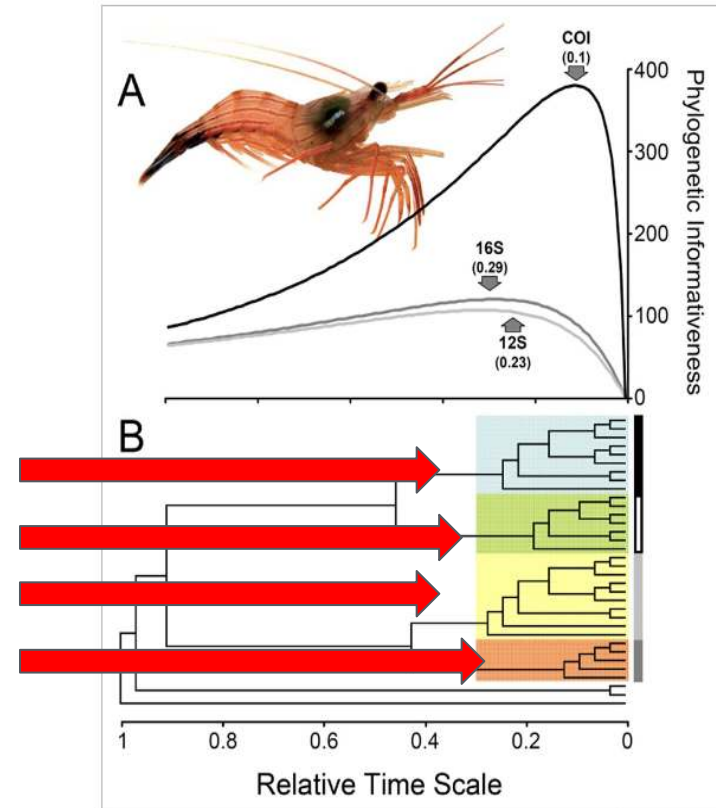
Which properties?

Information content

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

Phylogenetic signal

Good information
to infer these
nodes



04 PHYLOGENOMIC SUBSAMPLING

Which properties?

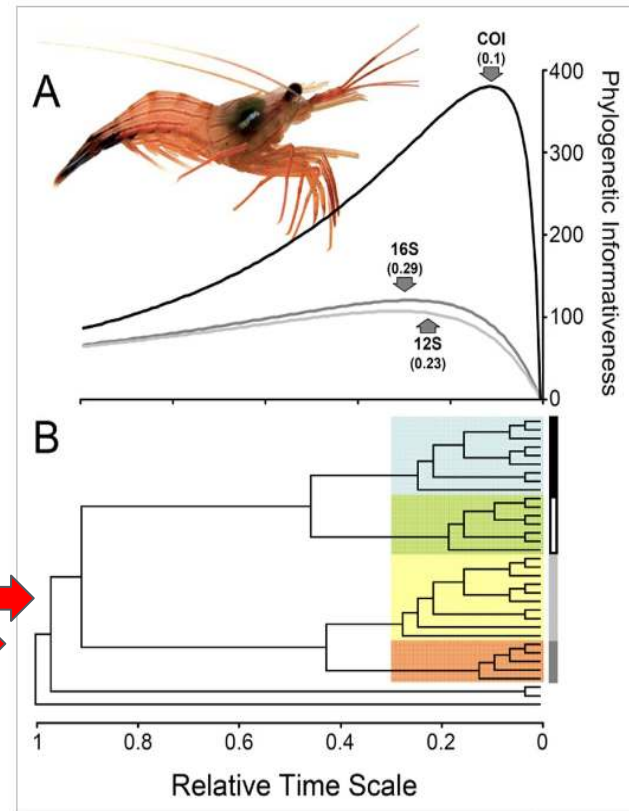
Information content

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

Phylogenetic signal

- > average support
- > Robinson-Foulds distance

Not enough
information to infer
these nodes



04 PHYLOGENOMIC SUBSAMPLING

Which properties?

Information content

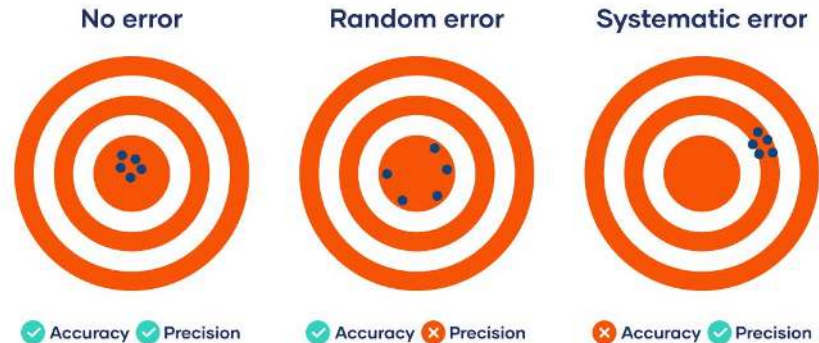
- > length of alignment
- > missing data
- > level of occupancy

Phylogenetic signal

- > average support
- > Robinson-Foulds distance

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

Random vs. systematic error



04 PHYLOGENOMIC SUBSAMPLING

Which properties?

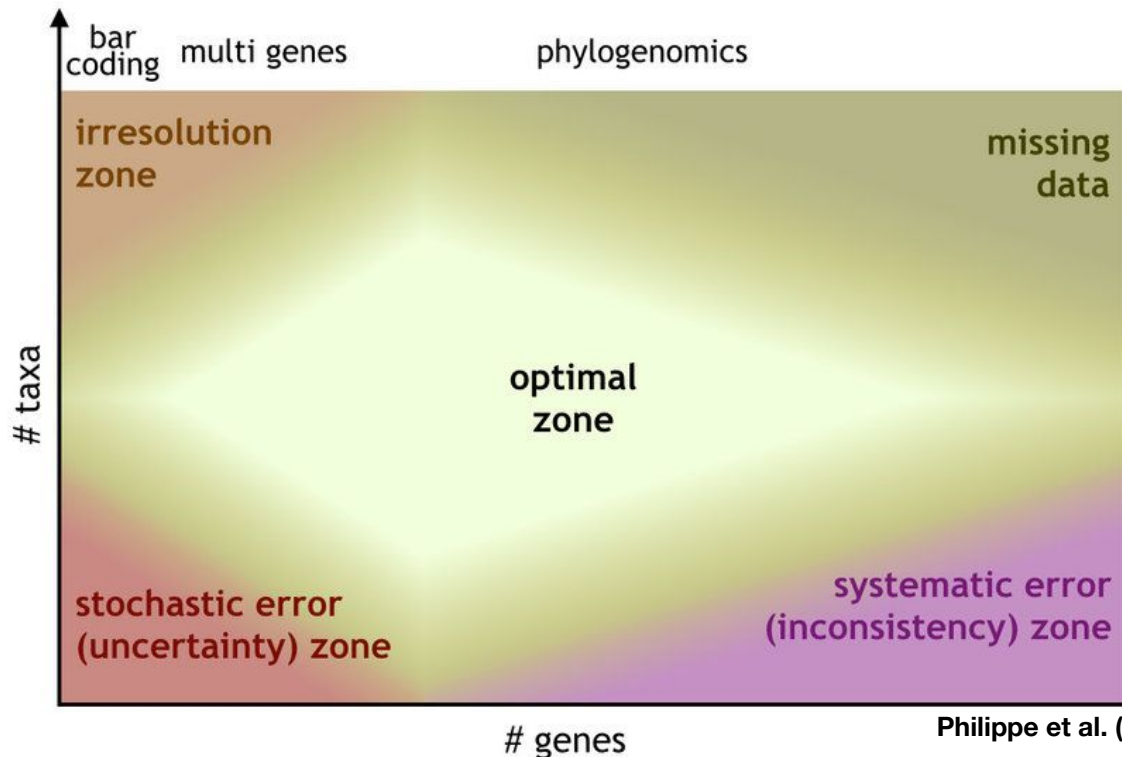
Information content

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

Phylogenetic signal

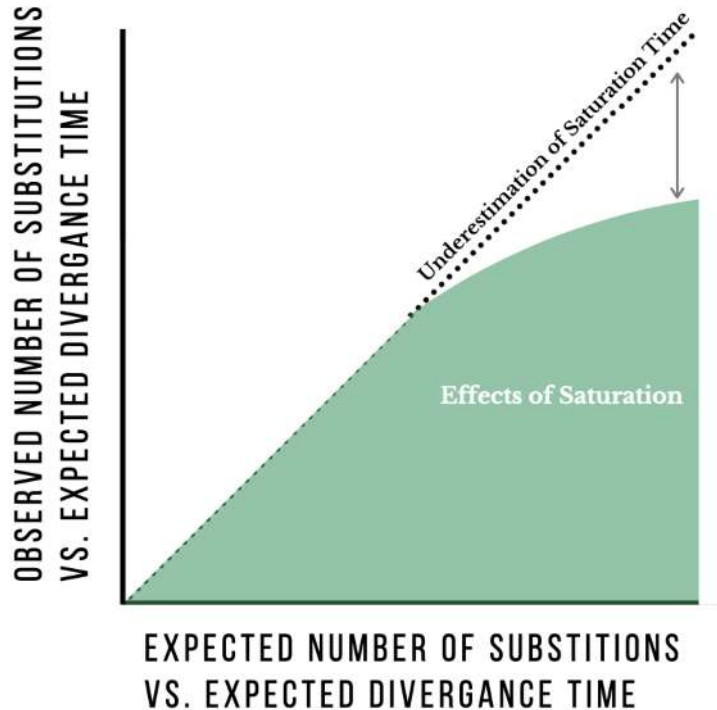
- > average support
- > Robinson-Foulds distance

Systematic error:



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?

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

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?

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

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

Software:

- PhyKIT
- genesortR

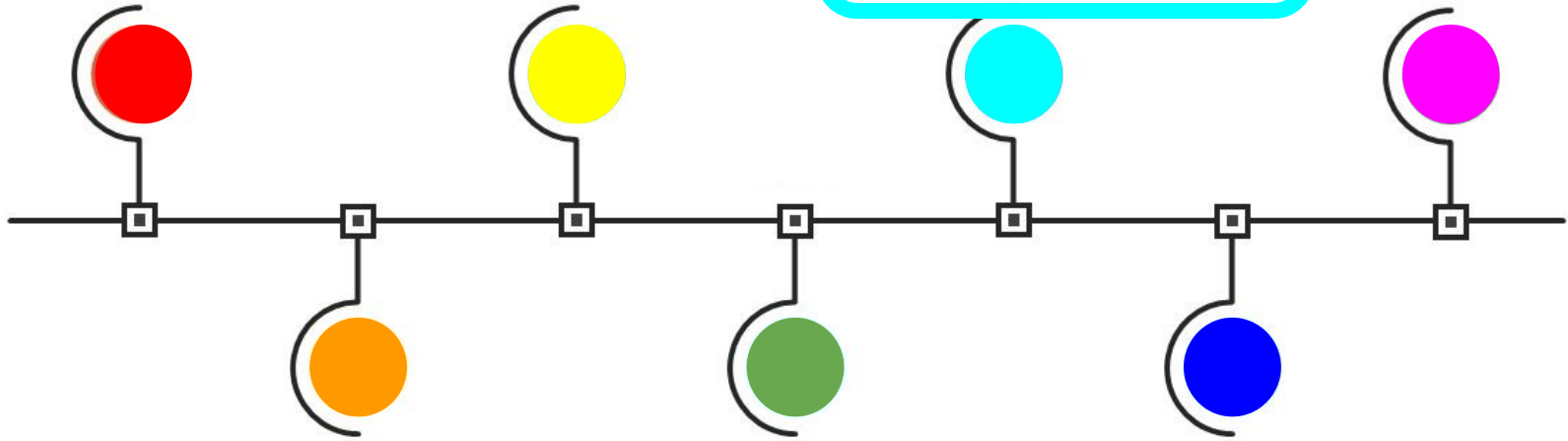
01 DATA

**03 ALIGNMENT
& TRIMMING**

**05 SUPERMATRIX
VS INDIVIDUAL
GENES**

**02 ORTHOLOGY
INFERENCE**

**04 PHYLOGENOMIC
SUBSAMPLING**



05 SUPERMATRIX VS INDIV. GENE TREES

~~Gene tree $\approx$ Species phylogeny~~

Gene tree $\neq$ Species phylogeny

05 SUPERMATRIX VS INDIV. GENE TREES

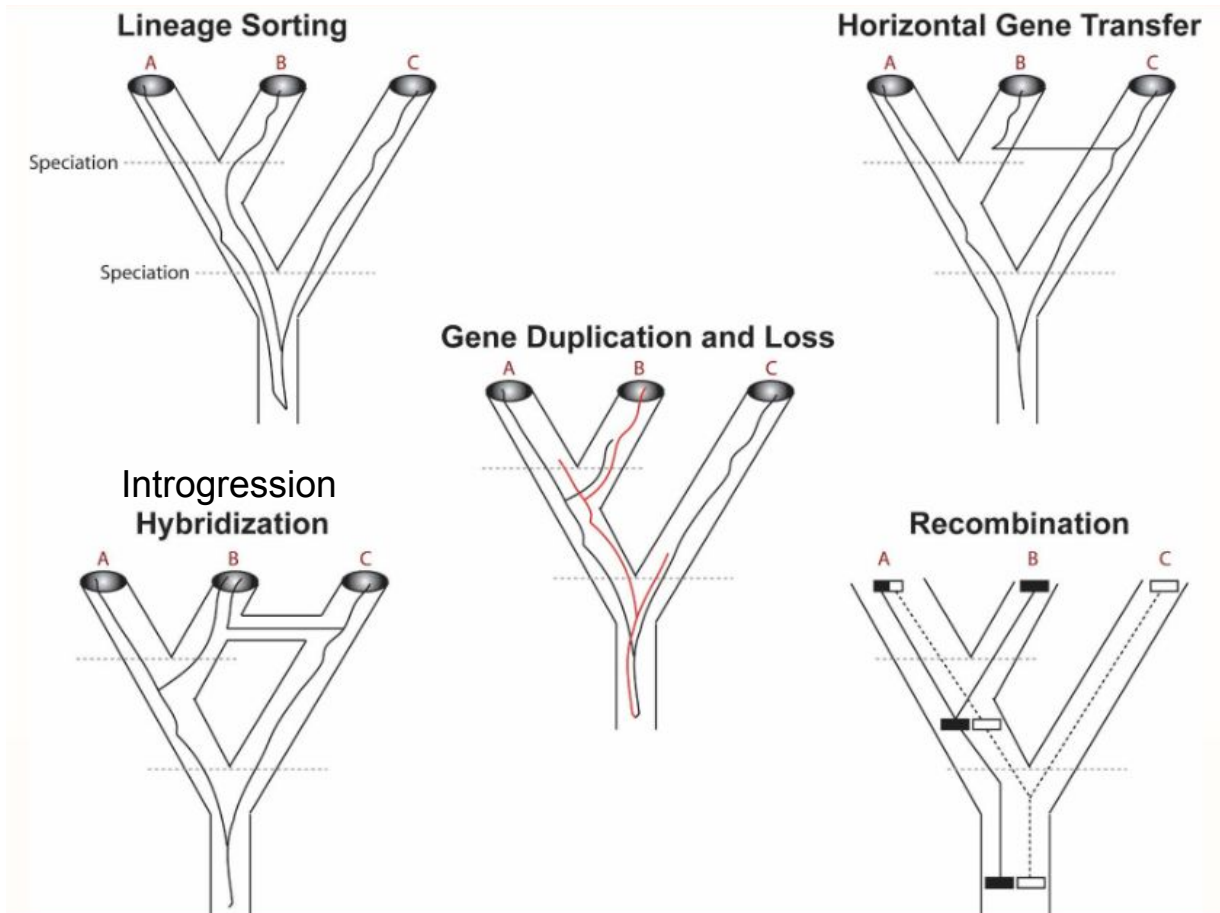
Analytical factors

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)

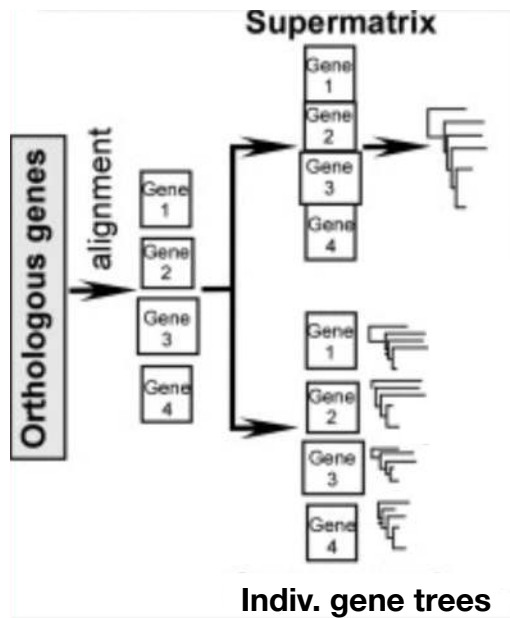
Biological factors

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



Phylogenetic analysis
(one tree)

Phylogenetic analysis
(multiple trees)

Software:

- ASTRAL
- TREE-QMC/TOB-QMC
- StarBeast3

Estimation of a species
tree given a set of gene
trees

Multispecies coalescent

01 DATA

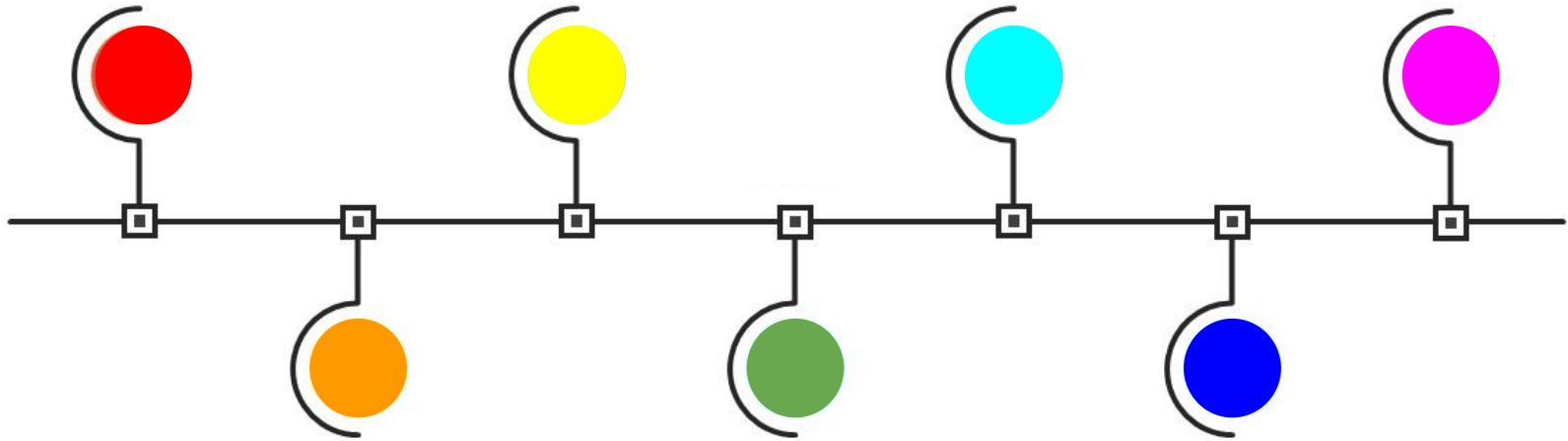
**03 ALIGNMENT
& TRIMMING**

**05 SUPERMATRIX
VS INDIVIDUAL
GENES**

**02 ORTHOLOGY
INFERENCE**

**04 PHYLOGENOMIC
SUBSAMPLING**

**06 MODEL
SELECTION &
PHYLOGENETIC
INFERENCE**



06 MODEL SELECTION & PHYLOGENETIC INFERENCE



DATA + MODEL OF EVOLUTION

+ METHOD

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

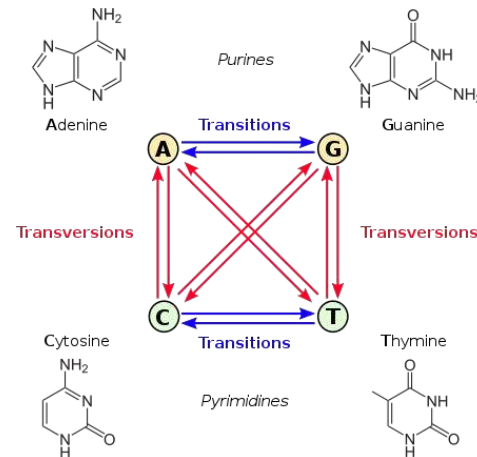
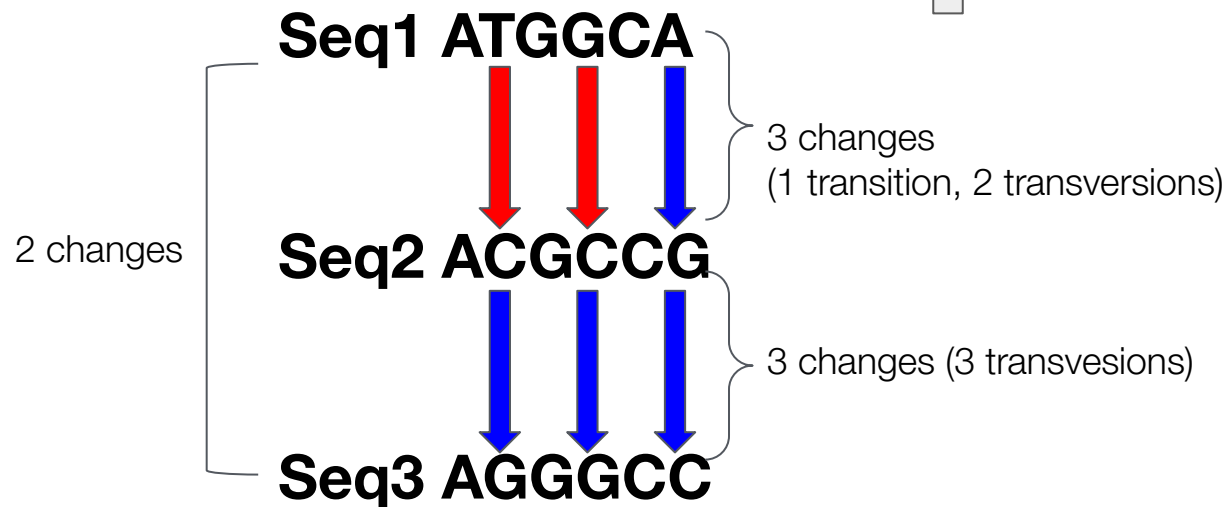
06 MODEL SELECTION & PHYLOGENETIC INFERENCE



DATA + **MODEL OF EVOLUTION** (= 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**

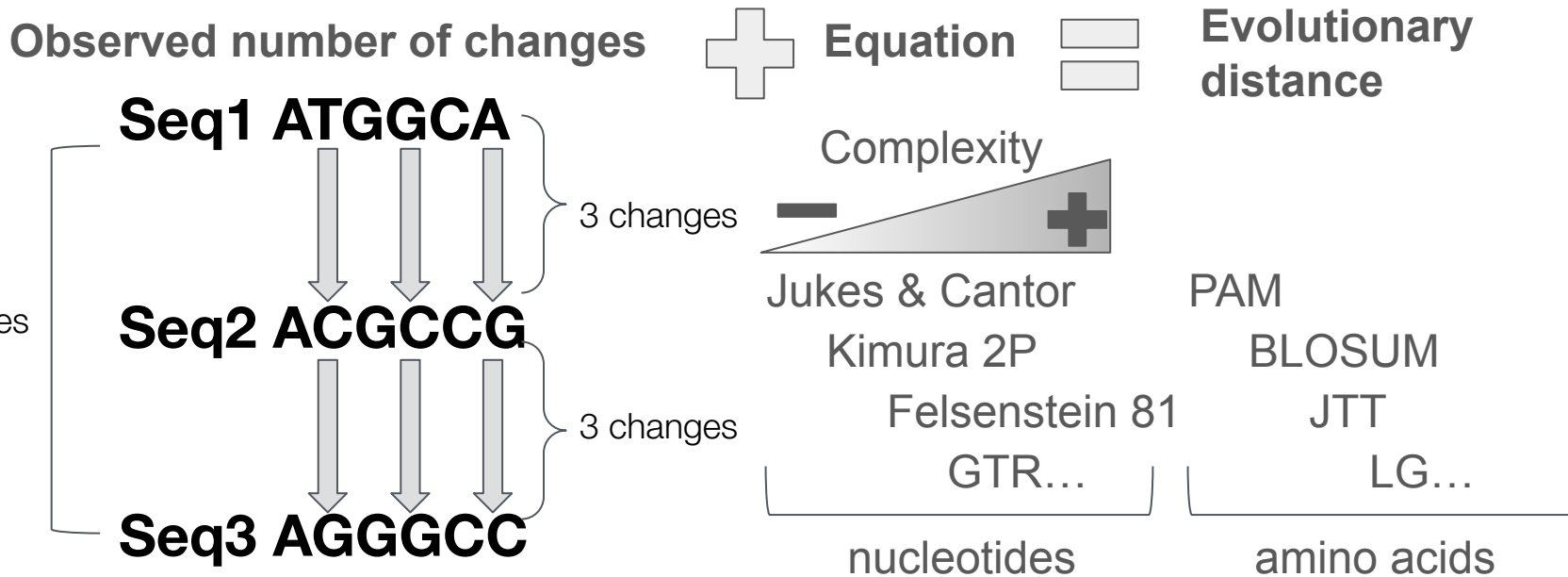


06 MODEL SELECTION & PHYLOGENETIC INFERENCE



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



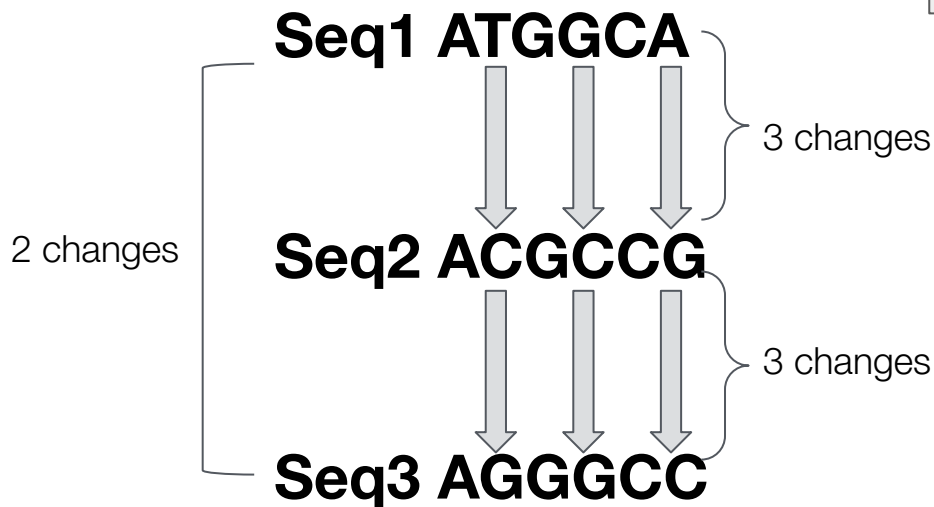
06 MODEL SELECTION & PHYLOGENETIC INFERENCE



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



All models are wrong,
but some are useful.

George Box, British statistician (1919 – 2013)

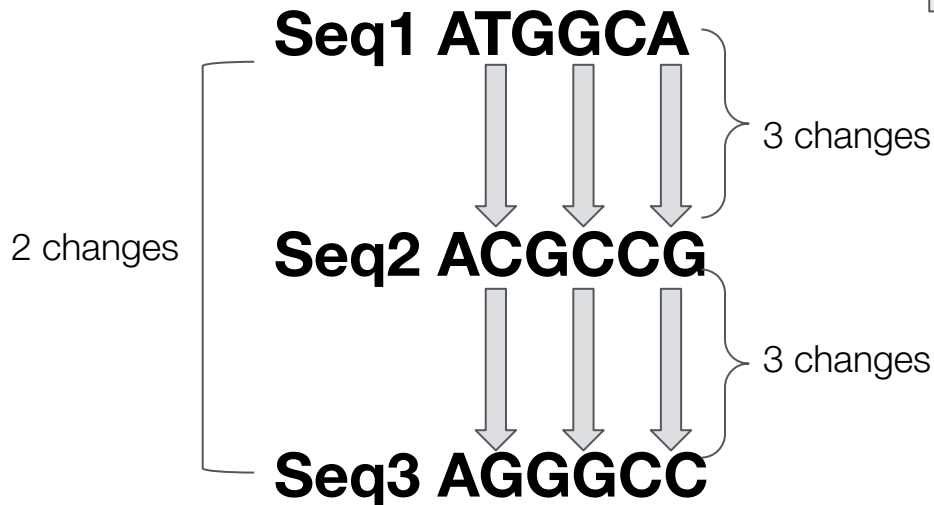
06 MODEL SELECTION & PHYLOGENETIC INFERENCE



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**



Software:

- ModelFinder (IQ-TREE3)
- ModelTest

06 MODEL SELECTION & PHYLOGENETIC INFERENCE



DATA + MODEL OF EVOLUTION
+ METHOD

Two main methods:

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

Software:

RevBayes
BEAST2
ExaBayes

Basic question in BI:

'What is the probability that this model (M) is correct, given the data (D) that we have observed?'

IQ-TREE3
RAxML-ng
ExaML

Basic question in ML:

'What is the probability of seeing the observed data (D) given that a certain model (M) is true?'

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

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)

01 DATA

**03 ALIGNMENT
& TRIMMING**

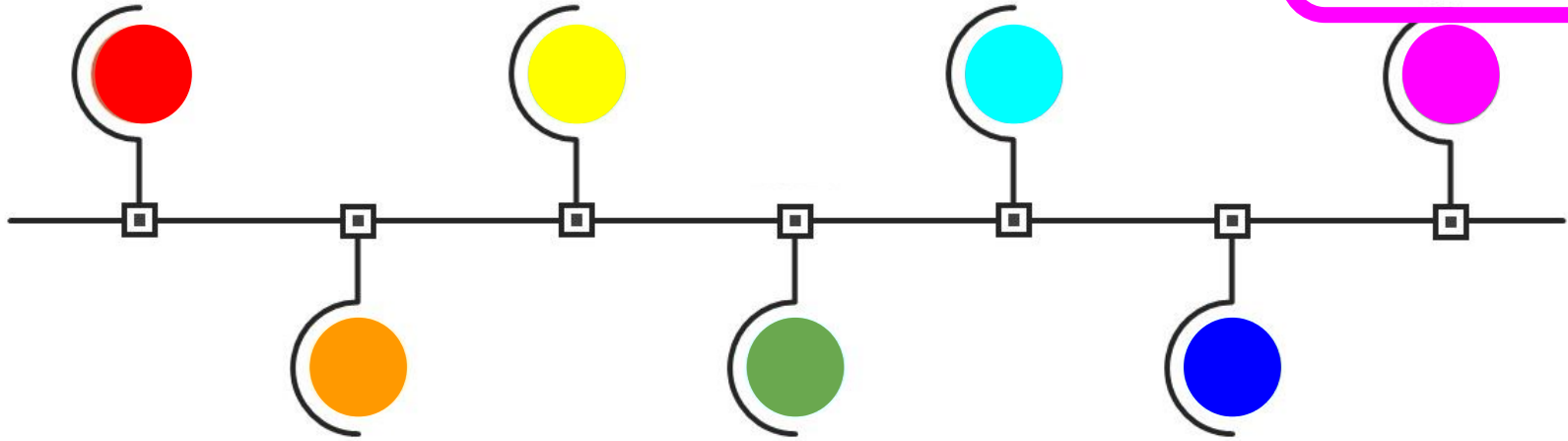
**05 SUPERMATRIX
VS INDIVIDUAL
GENES**

**07 TESTING THE
ROBUSTNESS
OF YOUR TREE**

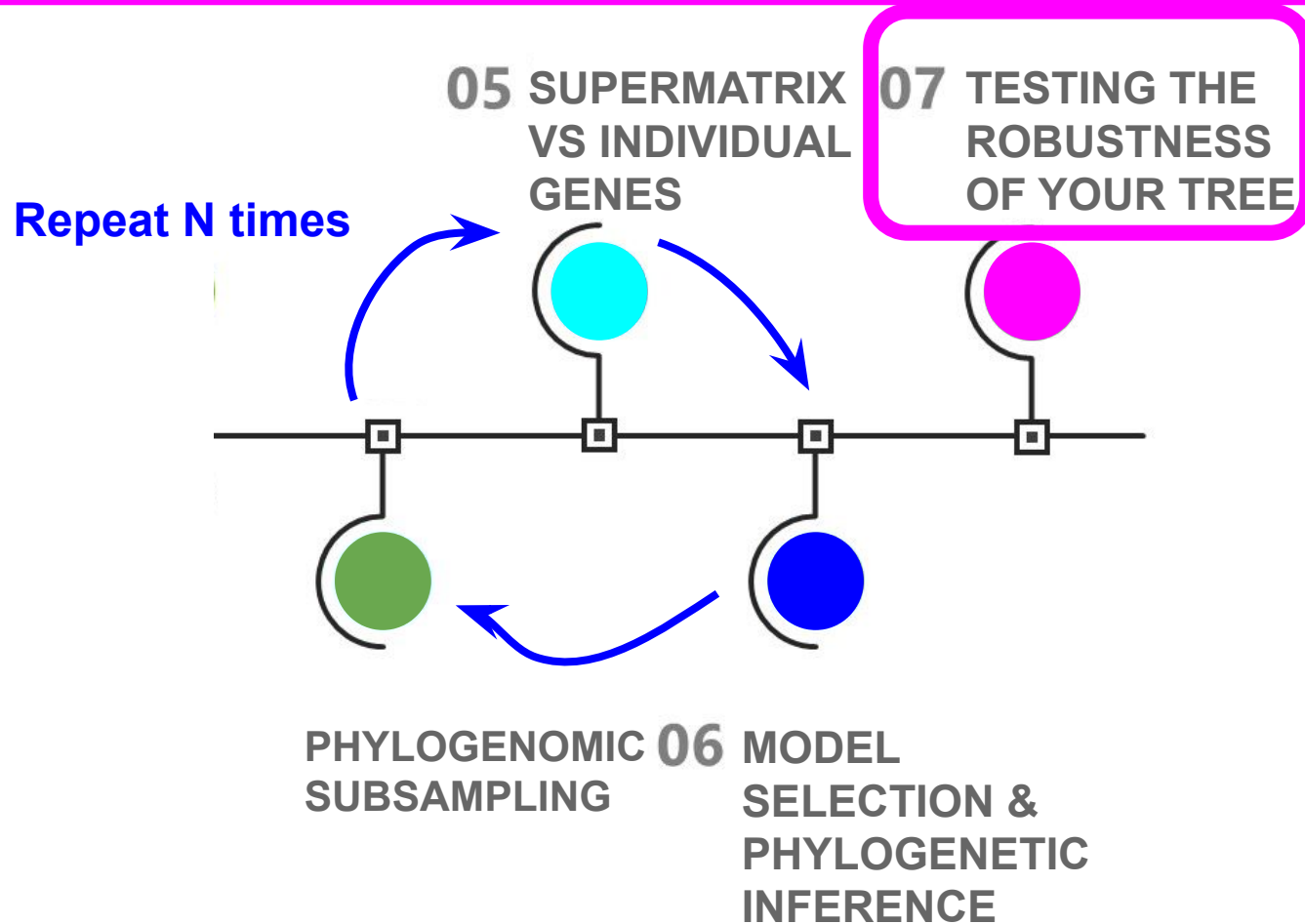
**02 ORTHOLOGY
INFERENCE**

**04 PHYLOGENOMIC
SUBSAMPLING**

**06 MODEL
SELECTION &
PHYLOGENETIC
INFERENCE**

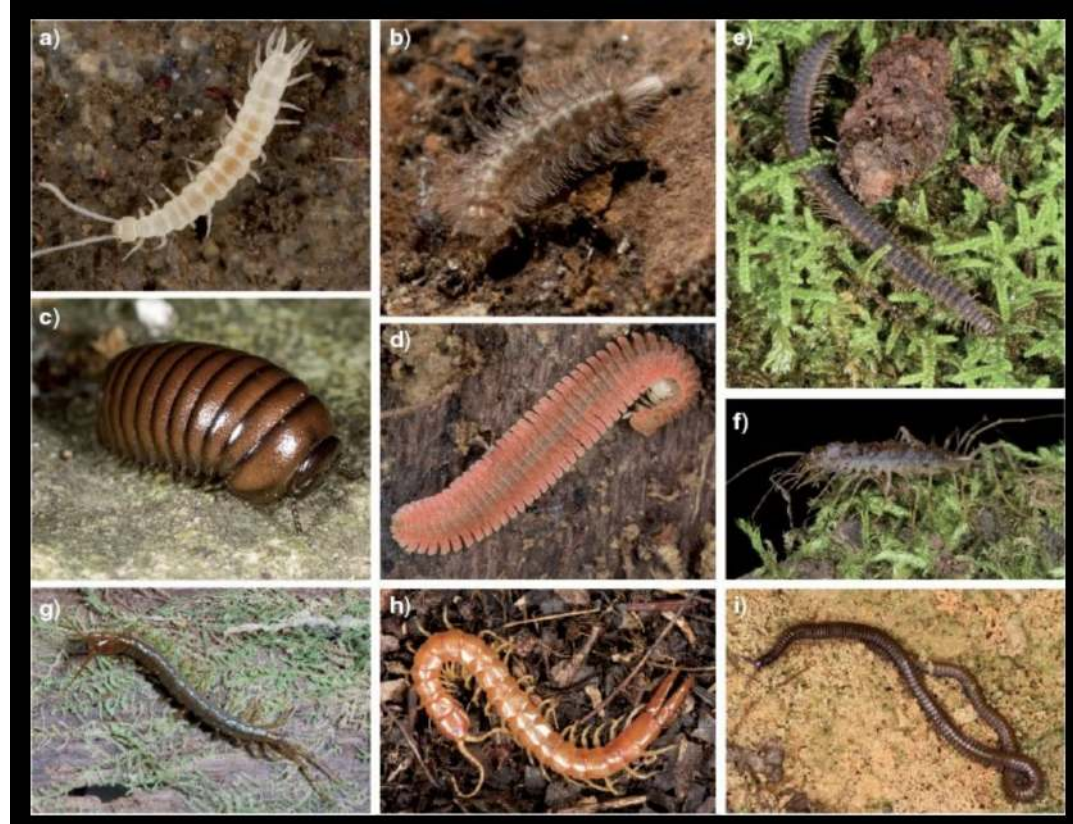
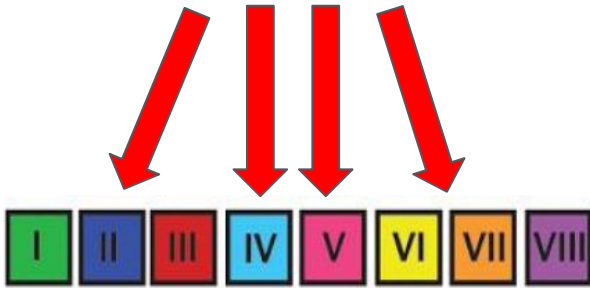


07 TESTING THE ROBUSTNESS OF YOUR TREE



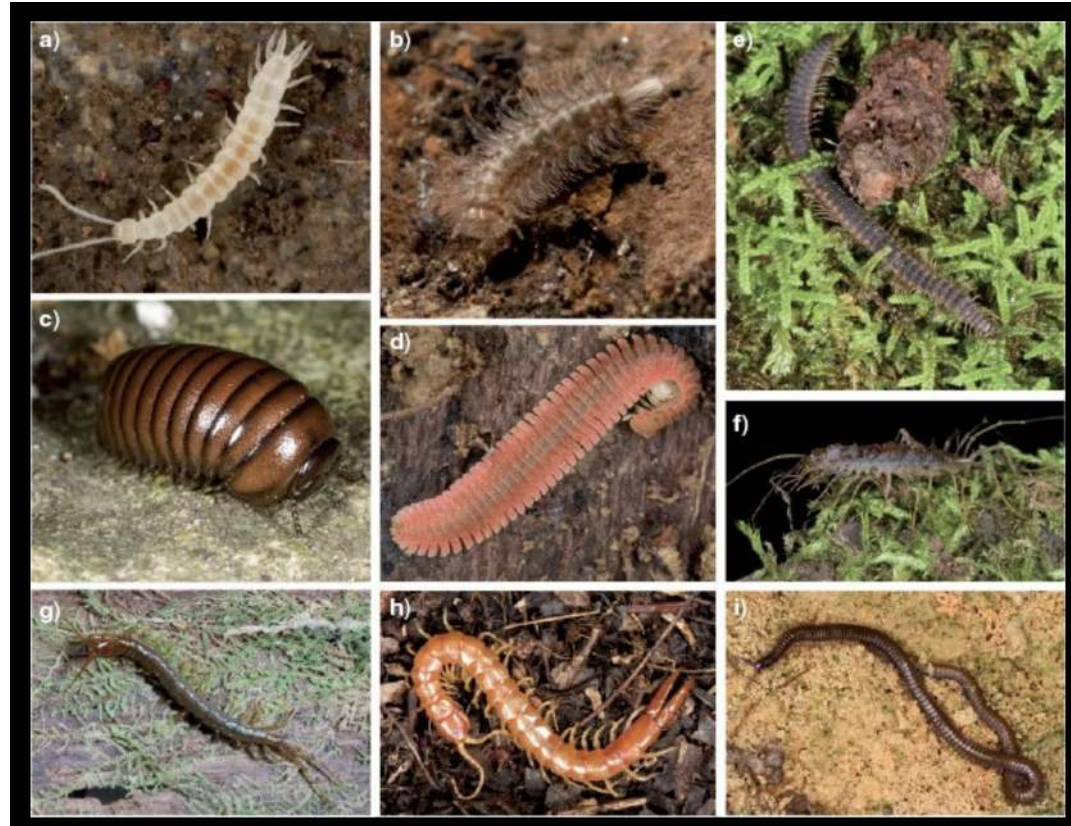
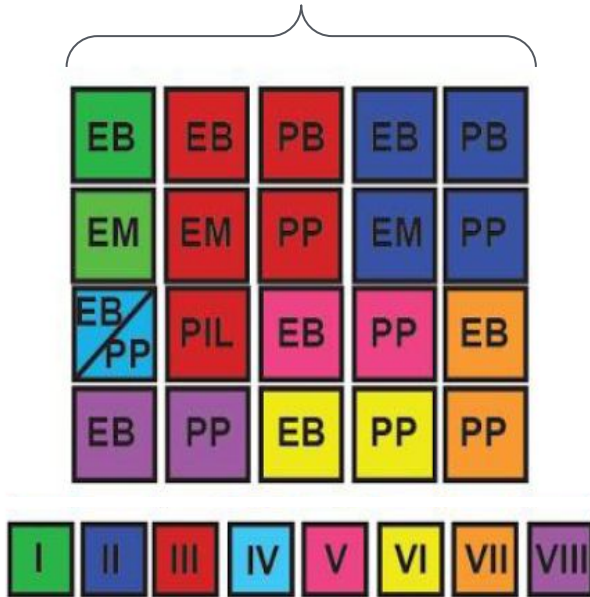
07 TESTING THE ROBUSTNESS OF YOUR TREE

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



07 TESTING THE ROBUSTNESS OF YOUR TREE

These are **analyses**



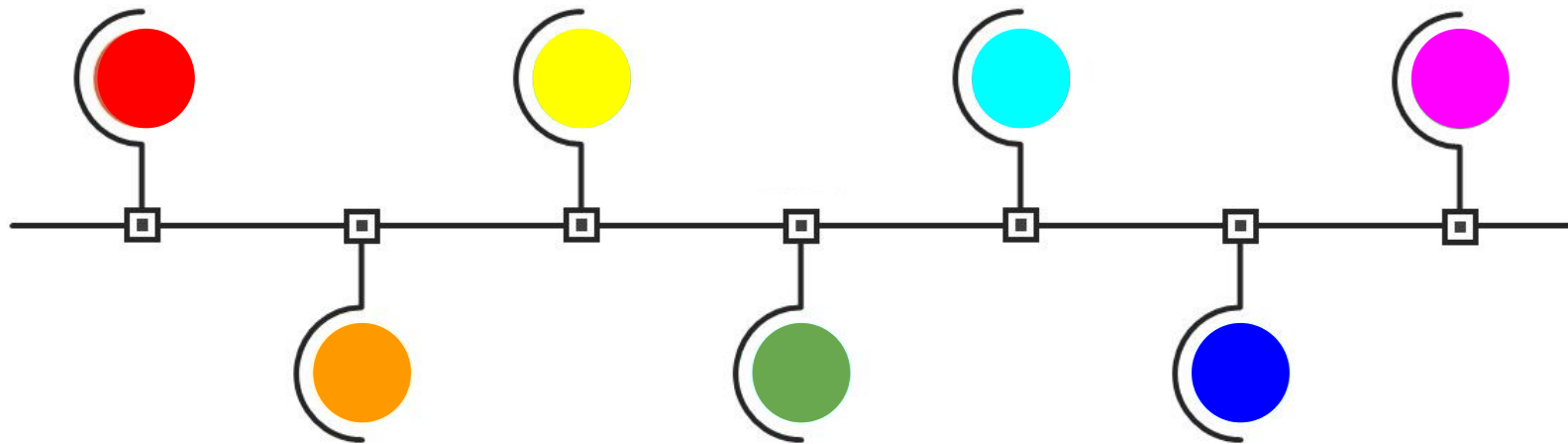
AND YOU, HOW IS **YOUR** PROJECT?

01 DATA

**03 ALIGNMENT
& TRIMMING**

**05 SUPERMATRIX
VS INDIVIDUAL
GENES**

**07 TESTING THE
ROBUSTNESS
OF YOUR TREE**



**02 ORTHOLOGY
INFERENCE**

**04 PHYLOGENOMIC
SUBSAMPLING**

**06 MODEL
SELECTION &
PHYLOGENETIC
INFERENCE**

1

2

3

From Darwin to Conceptual reconstruction of the 'Next generation' of phylogenetic trees

I, Feld: p m n q (19 Stämme)
 II, Feld: p x y q (3 Stämme)
 III, Feld: p s t q (4 Stämme)
 stellen Smögliche Fälle der konvergenen realogic dar

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

1

2

3



“Here be dragons”. This phrase refers to the practice of medieval map makers of drawing dragons and sea serpents in the uncharted areas at the edge of the map.

WARNING

**THIS PLAY AREA IS USED
AT YOUR OWN RISK**

'Next generation' phylogenomics: Why rethink phylogenomics?

Thousands of loci $\neq$ resolved trees

- Deep divergences, rapid radiations, short internodes
- Sequence signal saturates faster than we like
- Genomes contain **more information than alignments**

The Limits of the 'Bag of Genes' Model

Sequence signal saturates faster than structural signal.

The Status Quo

The Problem:

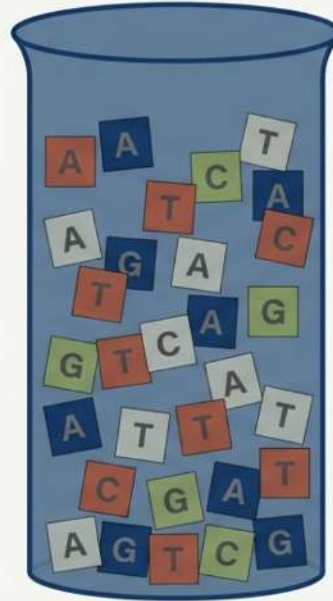
Classic phylogenomics treats genomes as disordered collections of independent loci.

The Result:

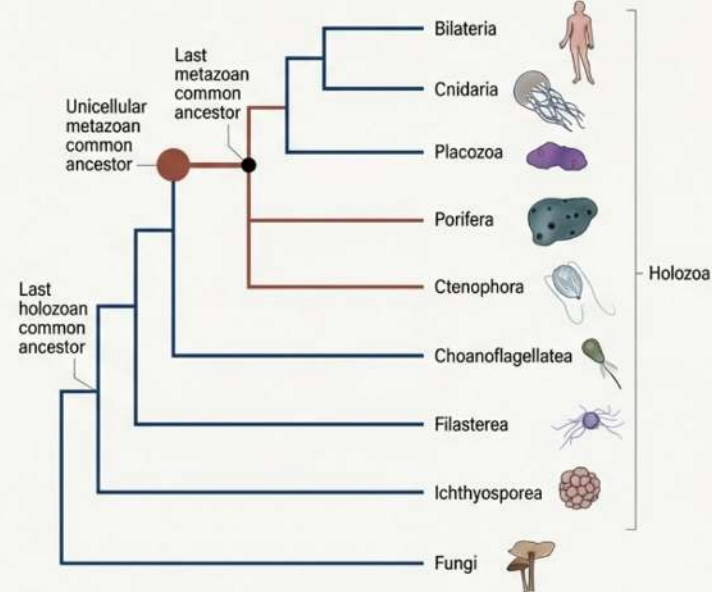
Despite using thousands of genes, deep divergences (like the base of Metazoa) and rapid radiations remain unresolved.

Key Question:

If sequence signal saturates, what other signals remain?



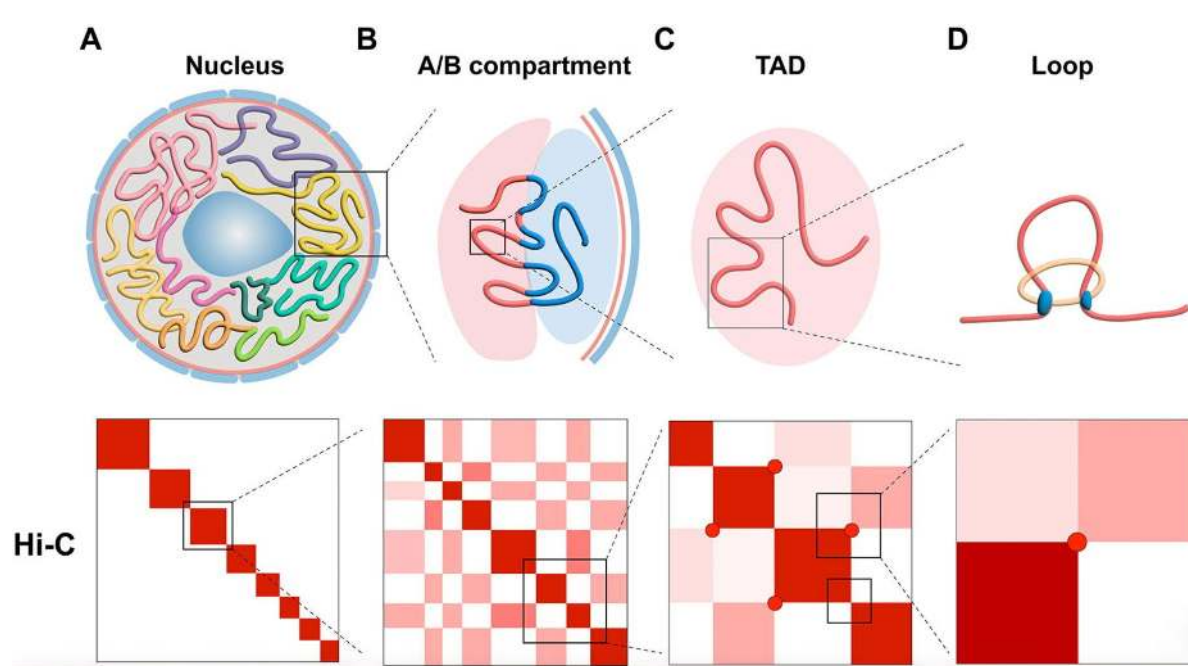
Phylogeny



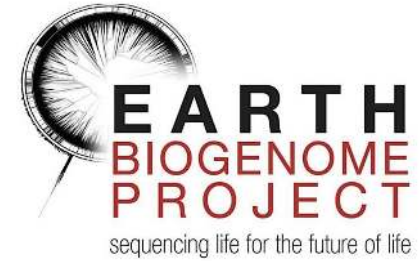
Two new sources of phylogenetic signal

- **Genome architecture**

- Gene order, chromosomes, 3D folding (chromosome-level genomes galore!)

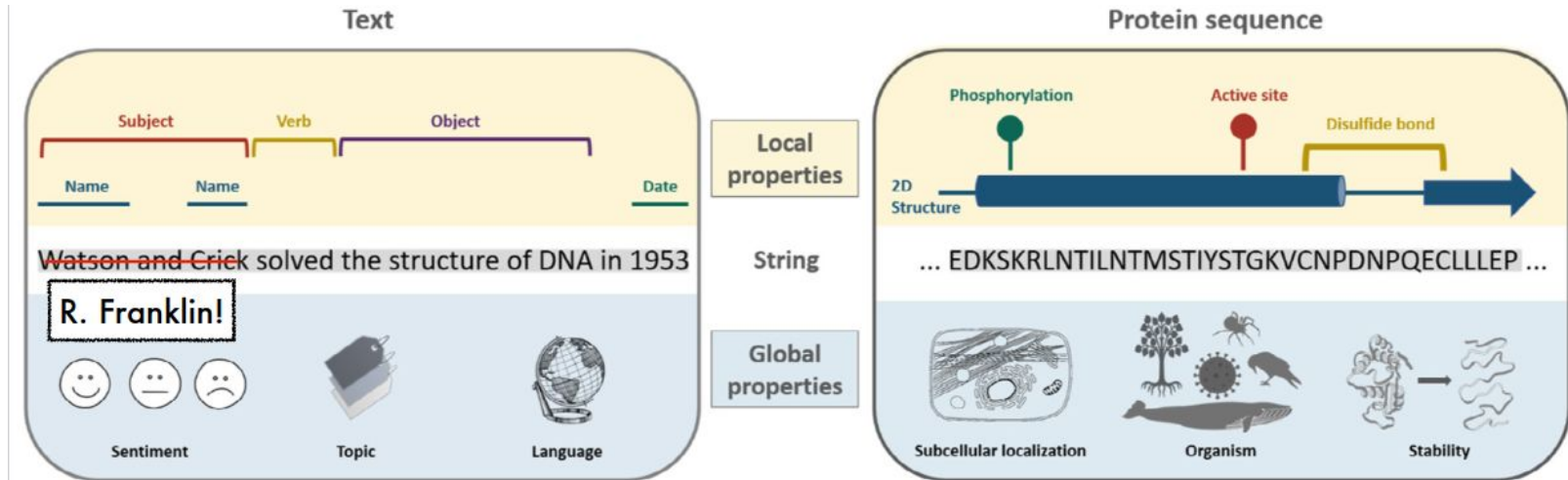


Yang & Ma 2022

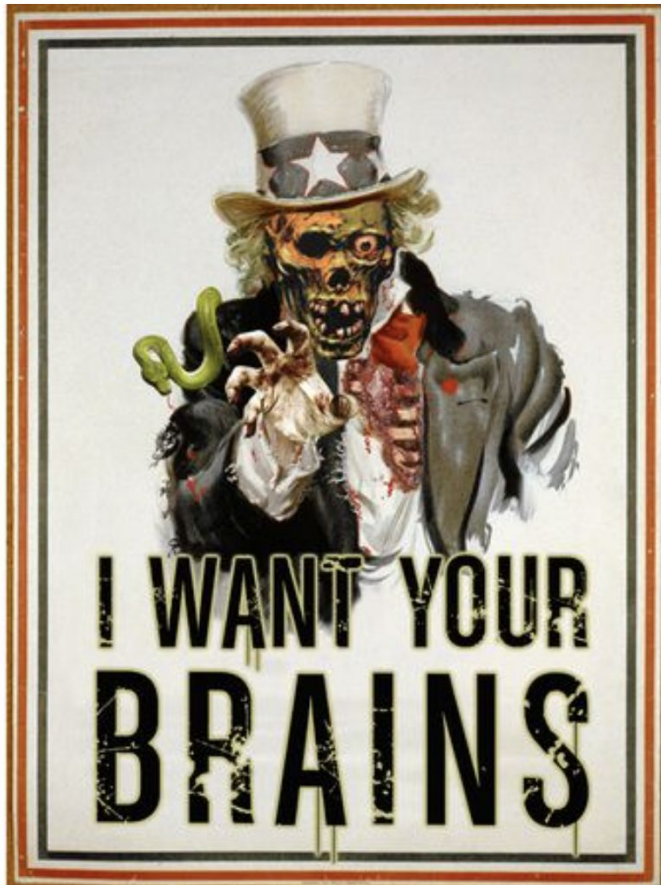


Two new sources of phylogenetic signal

- **Genome architecture**
 - Gene order, chromosomes, 3D folding (chromosome-level genomes galore!)
- **AI-based methods applied to phylogenomics/comparative genomics**
 - Encoding sequences as '*something else*', based on AI learning



WARNING (AGAIN!!): THIS IS ALL EXPLORATORY



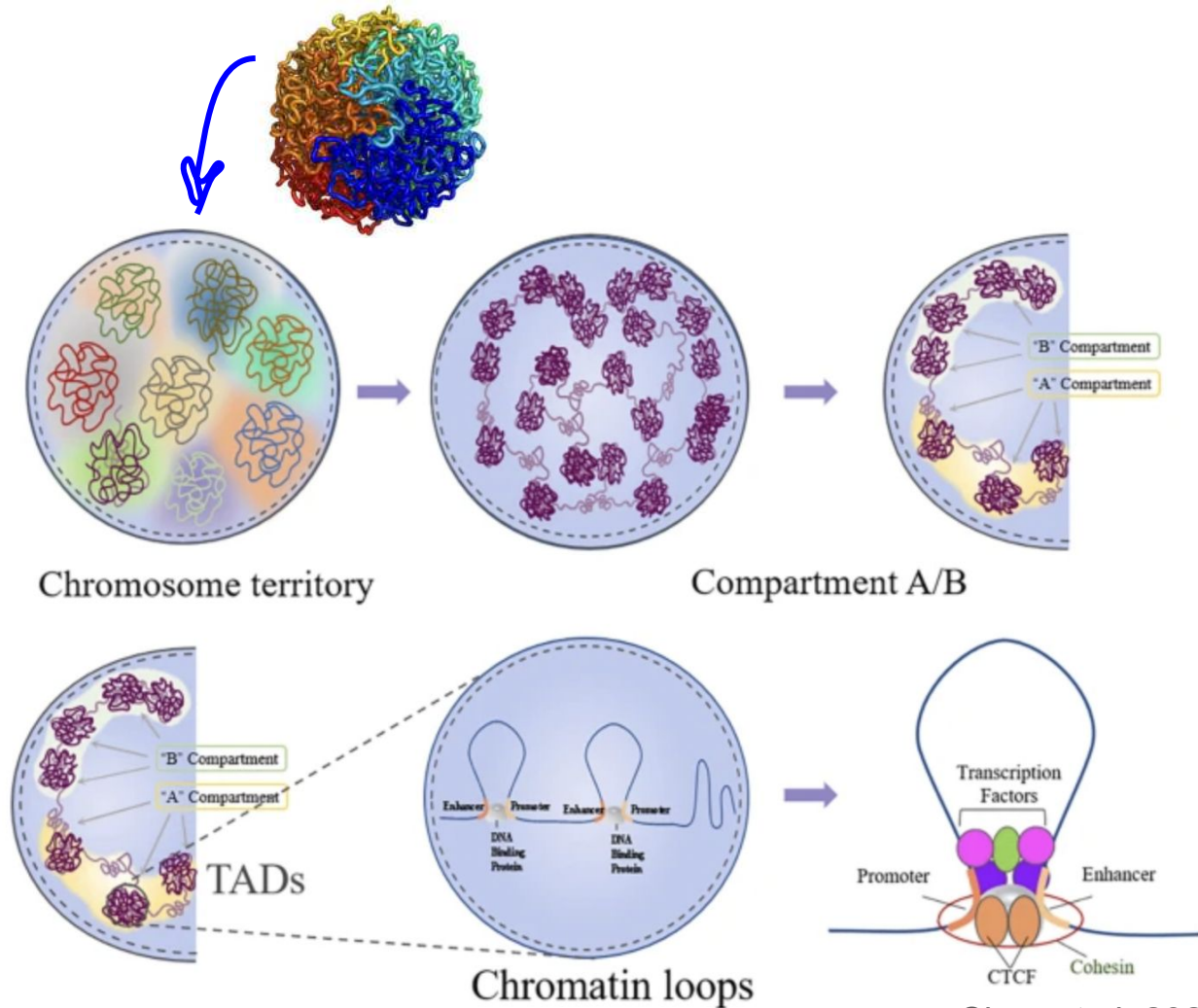
- Uncharted territory, emerging concepts that still need to be properly defined and tested.
 - still exploring: we need your brains!!
- Fields expanding exponentially, great potential, great investment (i.e. chromosome-level genomes, AI in China*)
 - we need to build literacy and critical thinking
- Results may be GREAT... or may be bullshit

(*China investment in AI surpasses by far that in Europe & USA)

PART I — Genome architecture—aware phylogenomics

Genomes are not bags of genes

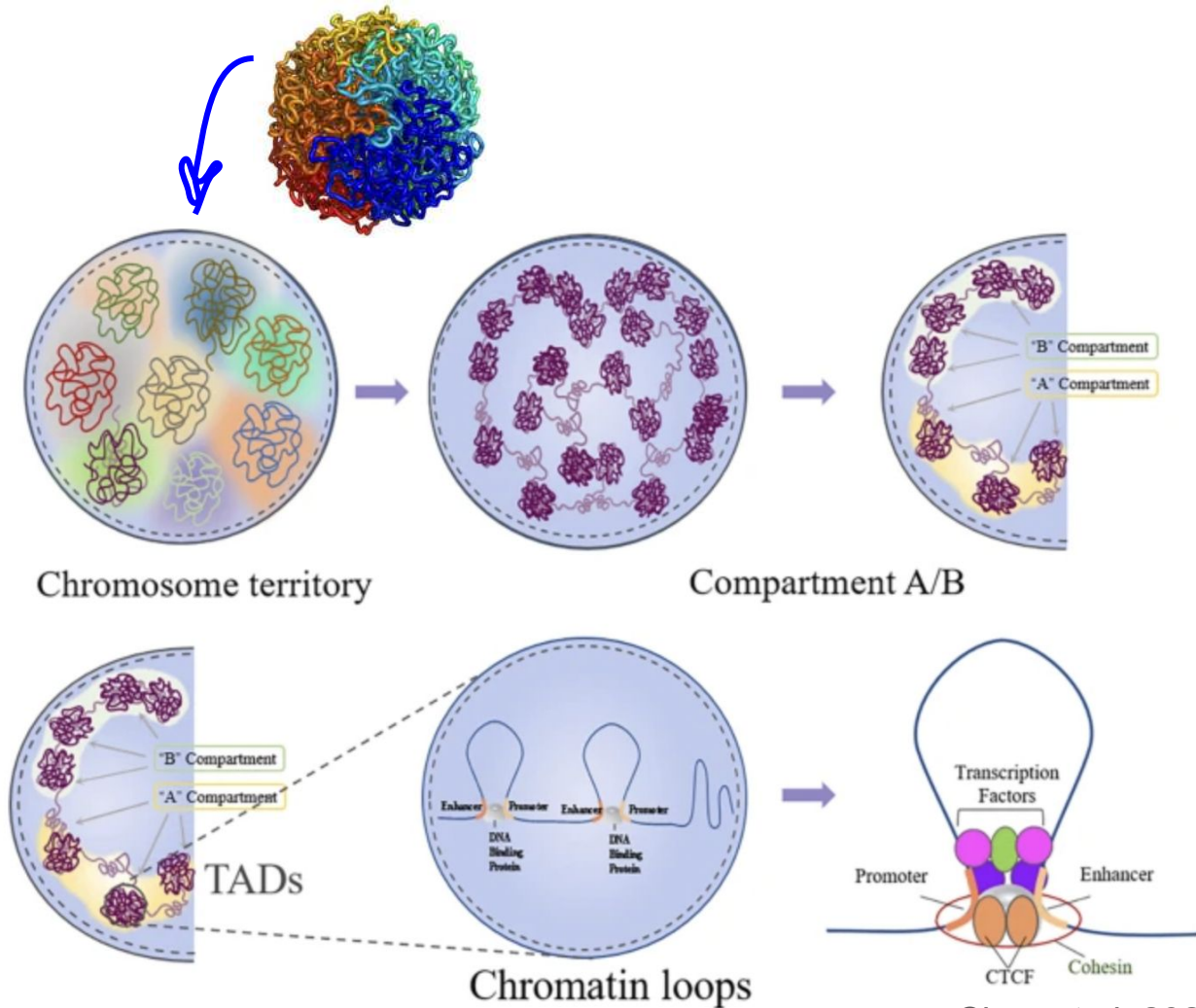
- Genes have **order**, **orientation**, **neighbors**
- Chromosomes evolve via fusions, fissions, inversions
- Structure persists when sequence similarity is gone: **SYNTENY**



PART I — Genome architecture—aware phylogenomics

Genomes are not bags of genes

- Genes have **order**, **orientation**, **neighbors**
- Chromosomes evolve via fusions, fissions, inversions
- Structure persists when sequence similarity is gone: **SYNTENY** (... or does it??)



PART I — Genome architecture—aware phylogenomics

Genomes as documents of evolutionary history: a probabilistic macrosynteny model for the reconstruction of ancestral genomes

Yoichiro Nakatani* and Aoife McLysaght* (2017)

Ancestral Linkage Groups (ALGs): conserved blocks of genes that remained together on ancestral chromosomes over vast evolutionary periods

Macrosynteny survives deep time

- Ancestral linkage groups conserved across animals
- Detected even after >500 My of divergence
- Provides signal when alignments fail

Examples

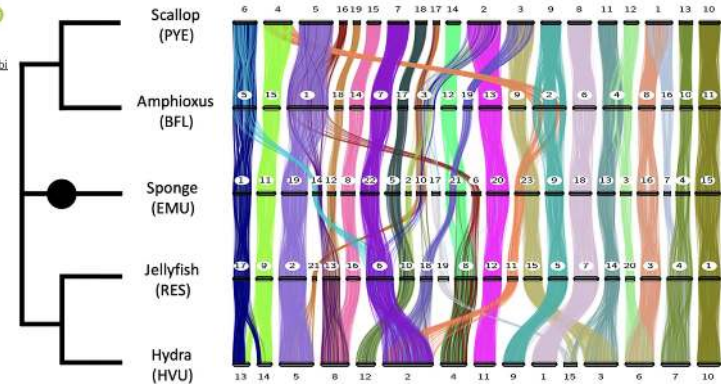
- Amphioxus as proxy for ancestral chordate genome
- Bilaterian chromosomal blocks conserved across phyla



Deeply conserved synteny and the evolution of metazoan chromosomes (2022)

OLEG SIMAKOV , JESSEN BREDESON , KODIAK BERKOFF , FERDINAND MARLETAZ , THERESE MITROS , DARRIN T. SCHULTZ , BRENDAN L. O'CONNELL
, PAUL DEAR, DANIEL E. MARTINEZ , [...] AND DANIEL S. ROKHSAR

SCIENCE ADVANCES • 2 Feb 2022 • Vol 8, Issue 5 • DOI:10.1126/sciadv.abi



PART I — Genome architecture—aware phylogenomics

Article [Open access](#) | Published: 17 May 2023

Ancient gene linkages support ctenophores as sister to other animals

[Darrin T. Schultz](#) , [Steven H. D. Haddock](#), [Jessen V. Bredeson](#), [Richard E. Green](#), [Oleg Simakov](#)  &

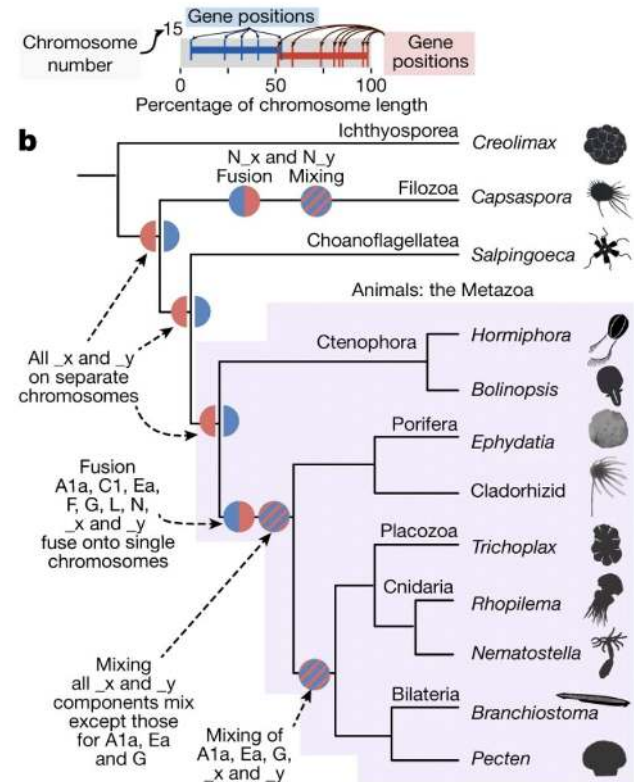
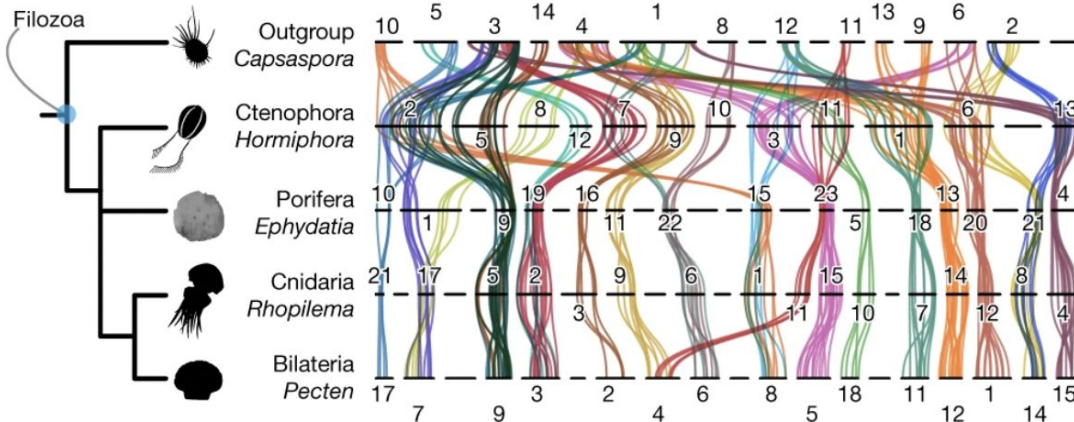
[Daniel S. Rokhsar](#) 

Synteny as a rare genomic change

- Rearrangements = discrete evolutionary events
- Shared fusions/fissions → low homoplasy
- Conceptually similar to indels or retroposons

Key idea

- Fewer characters, but more reliable



PART I — Genome architecture—aware phylogenomics

Synteny as a rare genomic change

- *REALLY??*

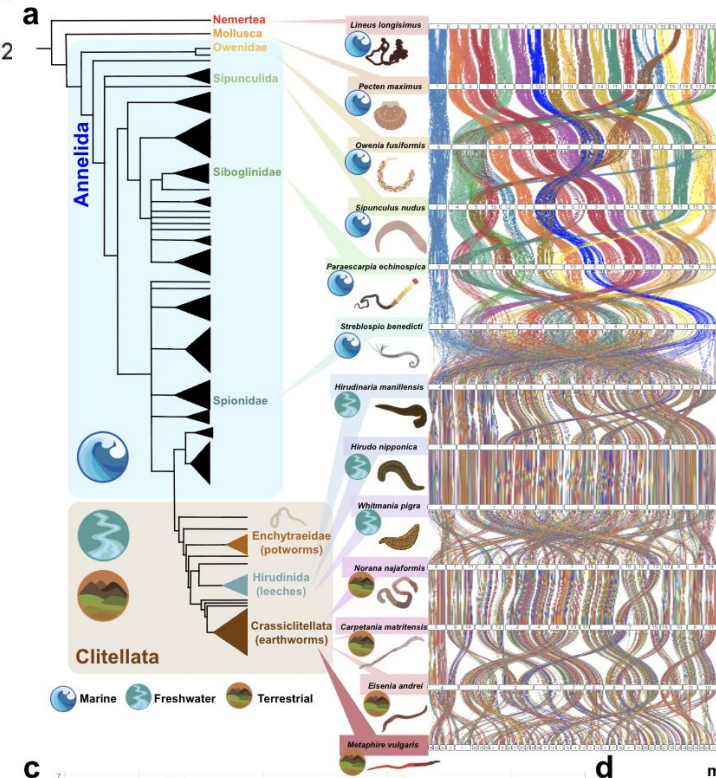


Article | Published: 18 June 2025

An episodic burst of massive genomic rearrangements and the origin of non-marine annelids

Carlos Vargas-Chávez, Lisandra Benítez-Álvarez, Gemma I. Martínez-Redondo, Lucía Álvarez-González, Judit Salces-Ortiz, Klara Eleftheriadi, Nuria Escudero, Nadège Guiguielmoni, Jean-François Flot, Marta Novo, Aurora Ruiz-Herrera, Aoife McLysaght & Rosa Fernández

Nature Ecology & Evolution 9, 12



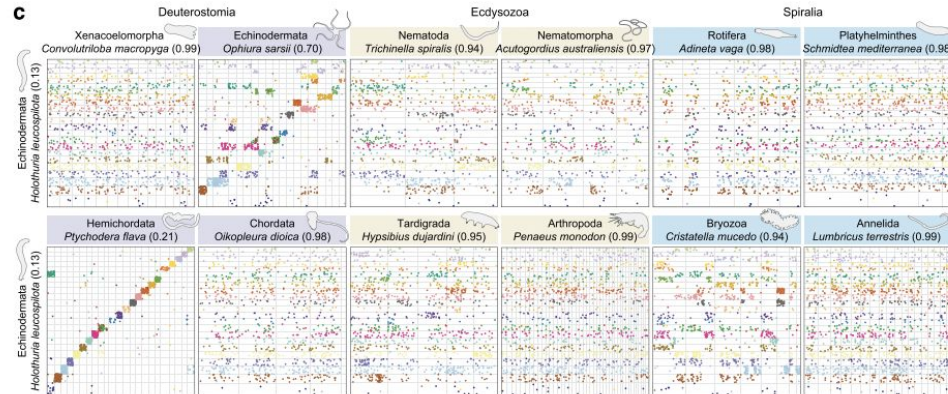
PART I — Genome architecture—aware phylogenomics

Synteny as a rare genomic change

- **REALLY??**

Conservation of bilaterian genome structure is the exception, not the rule

Thomas D. Lewin^{1*}, Isabel Jiah-Yih Liao¹ and Yi-Jyun Luo^{1*}

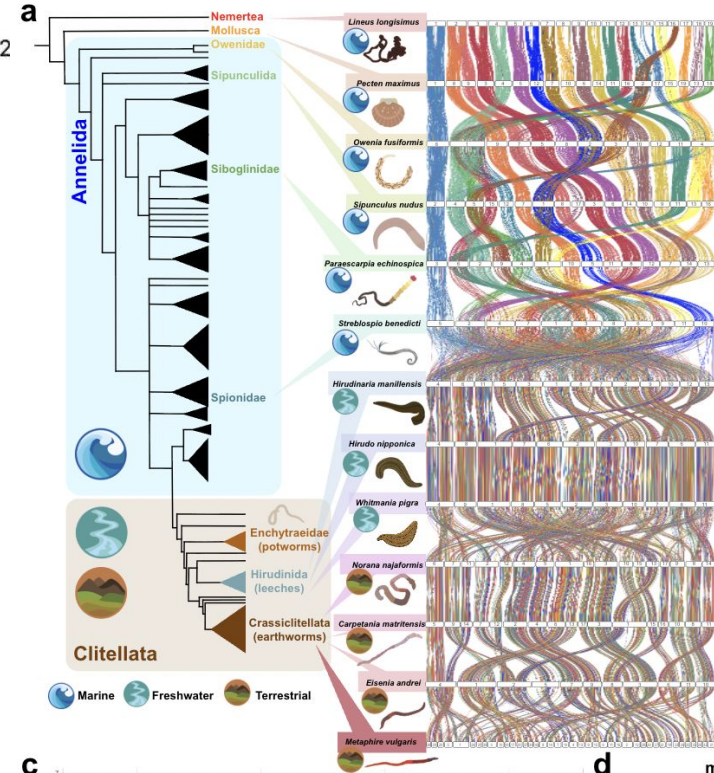


Article | Published: 18 June 2025

An episodic burst of massive genomic rearrangements and the origin of non-marine annelids

Carlos Vargas-Chávez, Lisandra Benítez-Álvarez, Gemma I. Martínez-Redondo, Lucía Álvarez-González, Judit Salces-Ortiz, Klara Eleftheriadi, Nuria Escudero, Nadège Guiguielmoni, Jean-François Flot, Marta Novo, Aurora Ruiz-Herrera, Aoife McLysaght & Rosa Fernández

Nature Ecology & Evolution 9, 12



PART I — Genome architecture—aware phylogenomics

Synteny as a rare genomic change

- **REALLY??**

Rearrangement rate heterogeneity is high

- Some lineages: highly stable genomes
- Others: massive reshuffling (even within phylum/genus!!)
- Rate heterogeneity is lineage-specific: we need models & new tools (e.g. to infer ALGs with more precision, simulations of SV scenarios, etc)

Implications

- Architecture works best when reshuffling is not extreme
- If extreme, be creative :-) (*feel free to reach out for tips!*)
- Not all 3D signal is phylogenetically useful

Article | Published: 18 June 2025

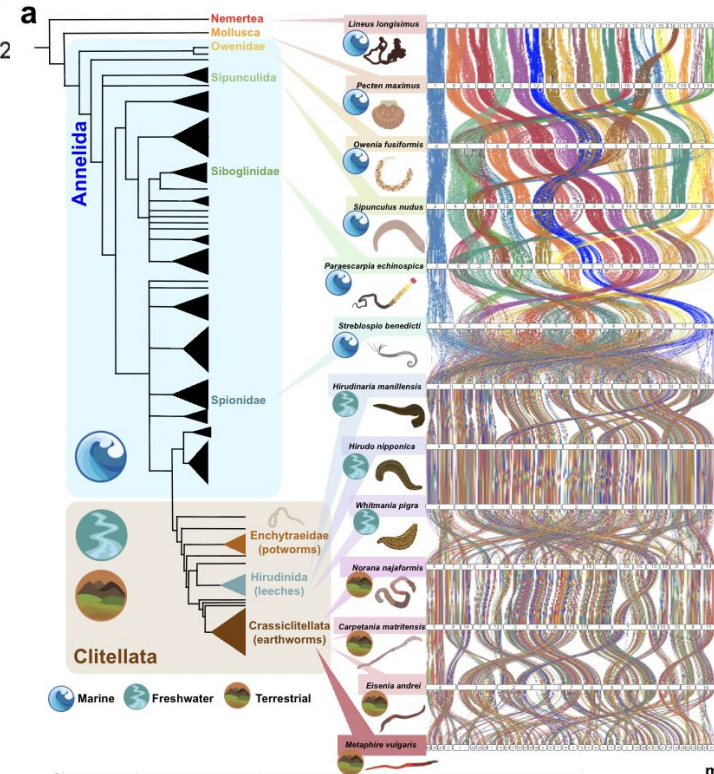
An episodic burst of massive genomic rearrangements and the origin of non-marine annelids

Carlos Vargas-Chávez, Lisandra Benítez-Álvarez, Gemma I. Martínez-Redondo, Lucía Álvarez-González

Judit Salces-Ortiz, Klara Eleftheriadi, Nuria Escudero, Nadège Guiguielmoni, Jean-François Flot, Marta

Novo, Aurora Ruiz-Herrera, Aoife McLysaght & Rosa Fernández

Nature Ecology & Evolution 9, 12



PART I — Genome architecture-aware phylogenomics

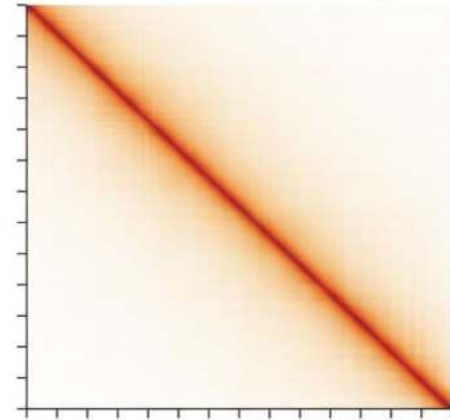
When genome architecture can mislead

- Assembly errors mimic rearrangements
- TE-driven convergence of breakpoints
- Paralogy confounds synteny blocks

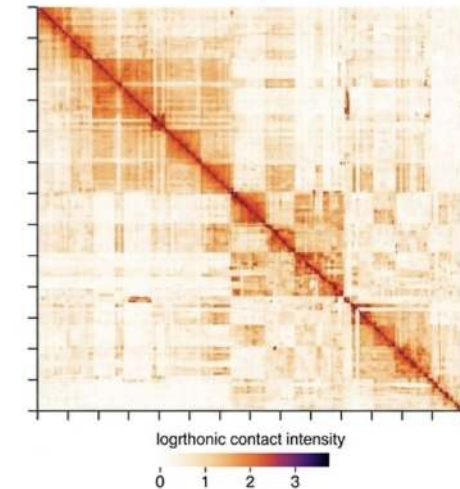
Rules

- Chromosome-level assemblies are mandatory (good quality!!)
- Hi-C data needs to be comparable (same kits/enzymes) & of enough depth

Correct Scaffolding



Assembly Error



PART I — Genome architecture—aware phylogenomics

Review

Breaking bad: when clitellate genomes go rogue

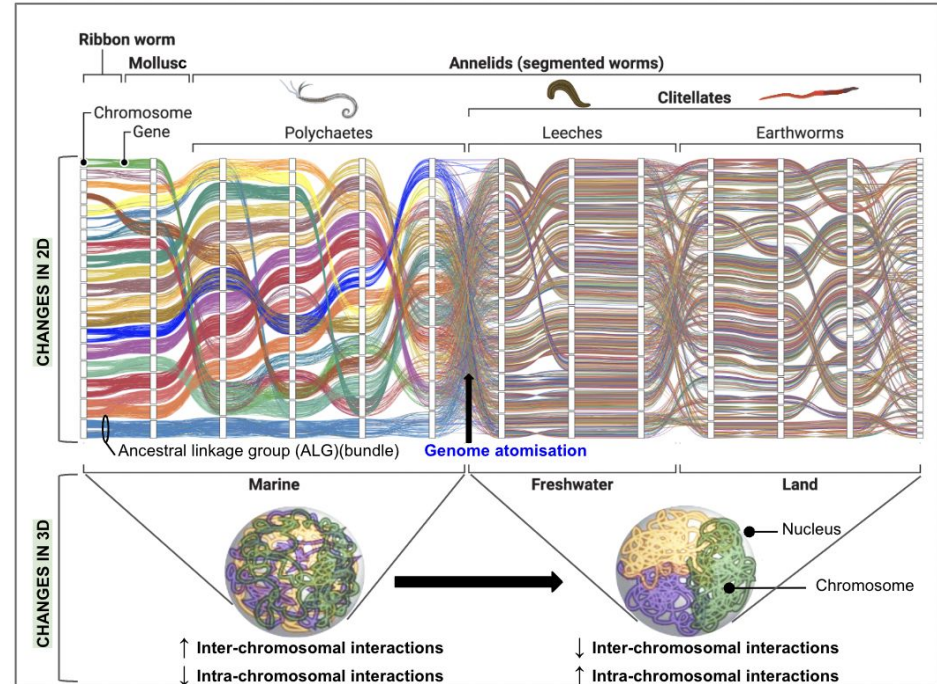
Carlos Vargas-Chávez¹, Aoife McLysaght², and Rosa Fernández ^{1,*}

Can 3D data inform phylogeny?

- Comparative contact decay curves
- Compartment similarity metrics
- Architecture-aware distance measures: ‘3D linkage groups’?

Are we there yet?

- Promising, exploratory, not standardized yet. A lot of fun work to do here!!



PART II — AI-assisted phylogenomics

Two main 'lines' of development of methods

- Complex pattern recognition via Machine learning & Deep learning

JOURNAL ARTICLE

Phylogenetic Methods Meet Deep Learning

Svitlana Braichenko, Rui Borges, Carolin Kosiol ✉ Author Notes

Genome Biology and Evolution, Volume 17, Issue 10, October 2025, evaf177,
<https://doi.org/10.1093/gbe/evaf177>

Published: 19 September 2025 **Article history** ▼

PDF Split View Cite Permissions Share ▼

Abstract

Deep learning (DL) has been widely used in various scientific fields, but its integration into phylogenetics has been slower, primarily due to the complex nature of phylogenetic data. The studies that apply DL to sequencing data often

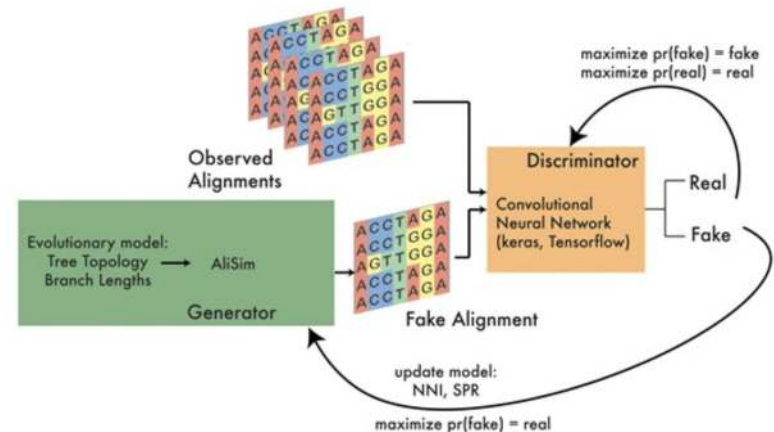
JOURNAL ARTICLE

Phylogenetic inference using generative adversarial networks

Megan L Smith ✉, Matthew W Hahn

Bioinformatics, Volume 39, Issue 9, September 2023, btad543,
<https://doi.org/10.1093/bioinformatics/btad543>

Published: 05 September 2023 **Article history** ▼



PART II — AI-assisted phylogenomics

Two main 'lines' of development of methods

- Complex pattern recognition via Machine learning & Deep learning

JOURNAL ARTICLE

Phylogenetic Methods Meet Deep Learning

Svitlana Braichenko, Rui Borges, Carolin Kosiol ✉ [Author Notes](#)

Genome Biology and Evolution, Volume 17, Issue 10, October 2025, evaf177,
<https://doi.org/10.1093/gbe/evaf177>

Published: 19 September 2025 **Article history** ▼

 PDF  Split View  Cite  Permissions  Share ▼

Abstract

Deep learning (DL) has been widely used in various scientific fields, but its integration into phylogenetics has been slower, primarily due to the complex nature of phylogenetic data. The studies that apply DL to sequencing data often

JOURNAL ARTICLE

Phylogenetic inference using generative adversarial networks

Megan L Smith ✉, Matthew W Hahn

Bioinformatics, Volume 39, Issue 9, September 2023, btad543,
<https://doi.org/10.1093/bioinformatics/btad543>

Published: 05 September 2023 **Article history** ▼

Phyloformer: Fast, Accurate, and Versatile Phylogenetic Reconstruction with Deep Neural Networks

Luca Nesterenko, Luc Blassel, Philippe Veber, Bastien Boussau, Laurent Jacob ✉
[Author Notes](#)

Molecular Biology and Evolution, Volume 42, Issue 4, April 2025, msaf051,

PART II — AI-assisted phylogenomics

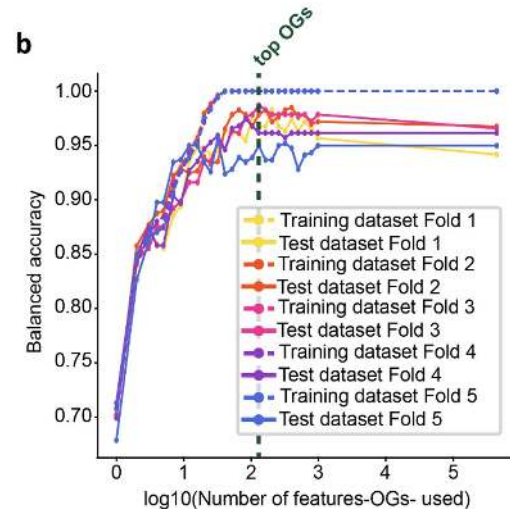
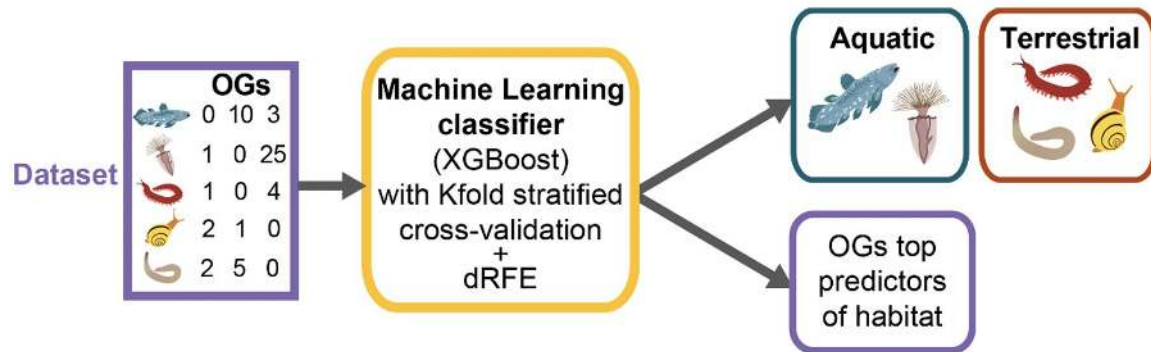
Two main 'lines' of development of methods

- Complex pattern recognition via Machine learning & Deep learning

Independent genomic trajectories shape adaptation to life on land across animal lineages

 Gemma I. Martínez-Redondo,  Klara Eleftheriadi, Judit Salces-Ortiz, Nuria Escudero, Fernando Ángel Fernández-Álvarez, Belén Carbonetto, Carlos Vargas-Chávez, Raquel García-Vernet, Javier Palma-Guerrero, Libe Rentería, Iñaki Rojo, Cristina Chiva, Eduard Sabidó, Aureliano Bombarely,  Rosa Fernández

Ca. 1,000 animal genomes, 24M genes, 520k orthogroups (OGs)



130 OGs are relevant for terrestrial animals (none shared across phyla)

PART II — AI-assisted phylogenomics

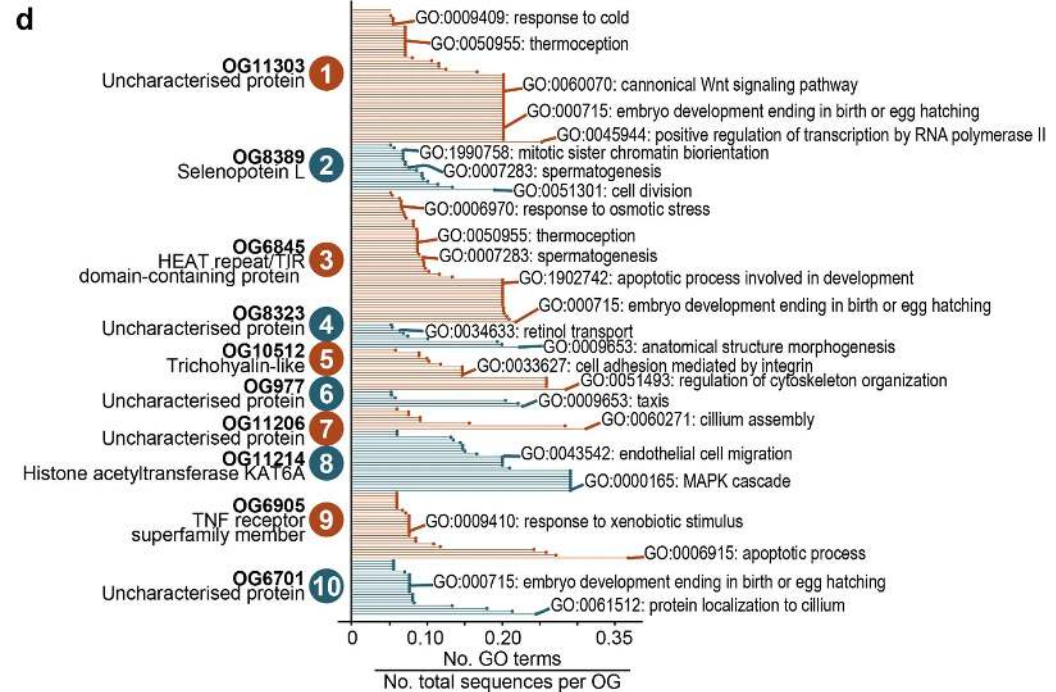
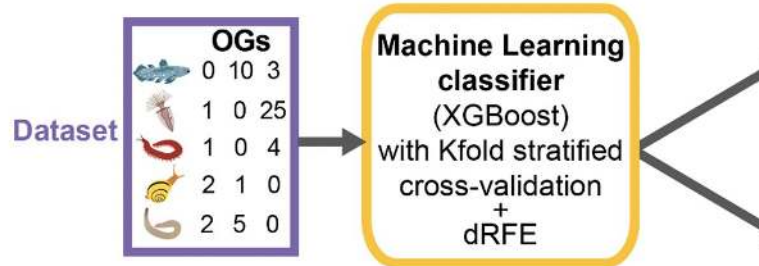
Two main 'lines' of development of methods

- Complex pattern recognition via Machine learning & Deep learning

Independent genomic trajectories shape adaptation to life on land across animal lineages

 Gemma I. Martínez-Redondo,  Klara Eleftheriadi, Judit Salces-Olivera, Fernando Ángel Fernández-Álvarez, Belén Carbonetto, Carlos Vargas, Javier Palma-Guerrero, Libe Rentería, Iñaki Rojo, Cristina Chiva, Edua Rosa Fernández

Ca. 1,000 animal genomes, 24M genes, 5

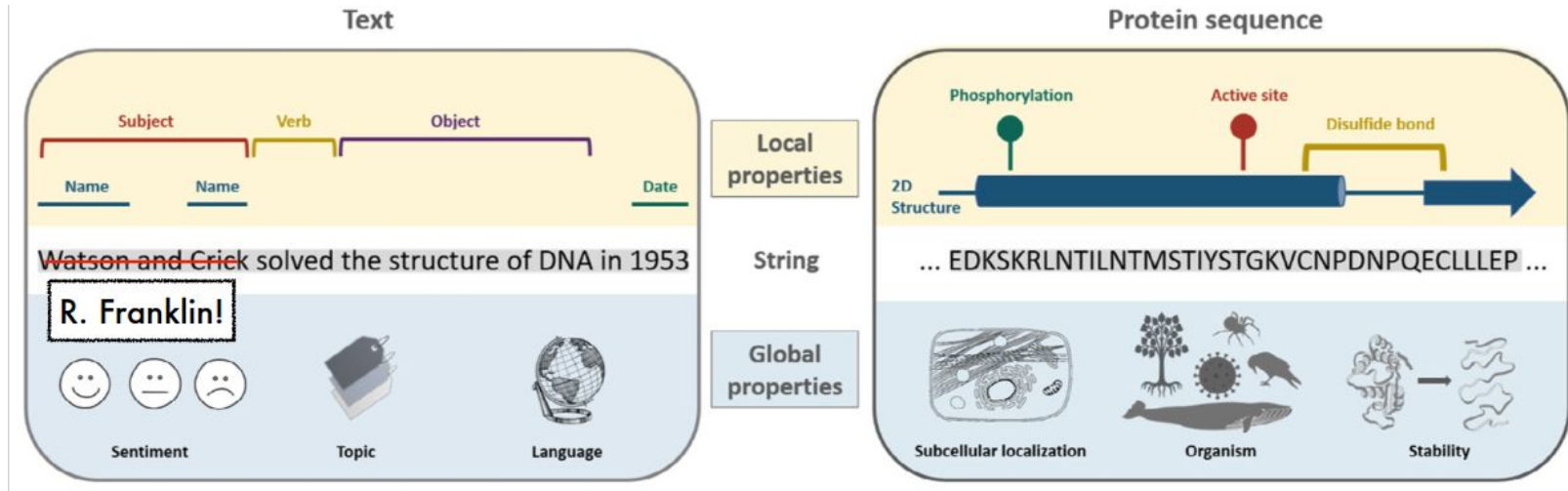


PART II — AI-assisted phylogenomics

Two main 'lines' of development of methods

- Genome/Protein Language Models to recode sequences and 'learn' the *grammar* of genomes

Encoding proteins as numerical vectors ('*embeddings*')

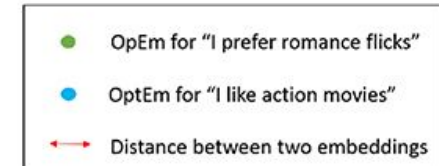
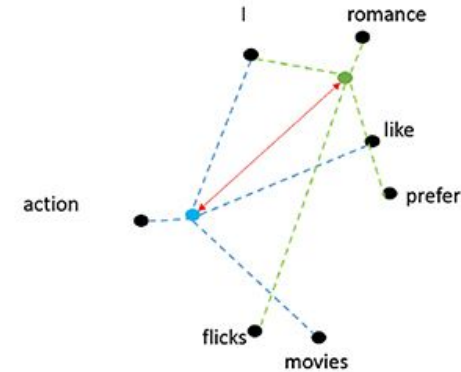
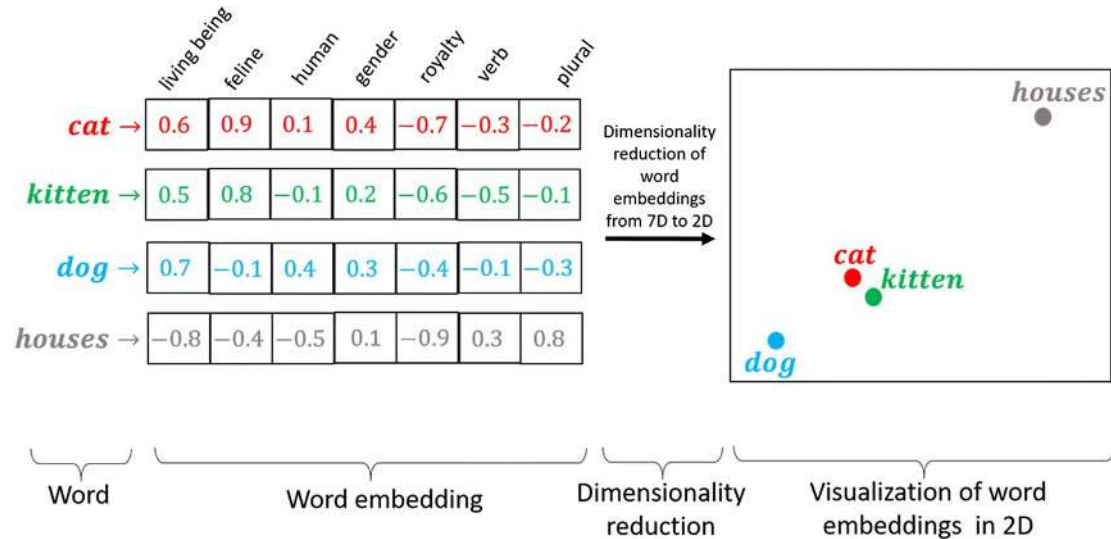


PART II — AI-assisted phylogenomics

Two main 'lines' of development of methods

- Genome/Protein Language Models to recode sequences and 'learn' the *grammar* of genomes

Encoding proteins as numerical vectors ('*embeddings*')



PART II — AI-assisted phylogenomics

Two main ‘lines’ of development of methods

- Genome/Protein Language Models to recode sequences and ‘learn’ the *grammar* of genomes

Protein language models in a nutshell

- Trained on millions of protein sequences
- Learn grammar of evolution implicitly
- No alignments, no trees during training



UniRef50 (ca. 60M non-redundant proteins)

Key insight

- Evolutionary constraints are learnable
- More informative than just the sequence
- Less bias due to indels

Transformer-based models work best (i.e. **ProtT5**, Ankh3)

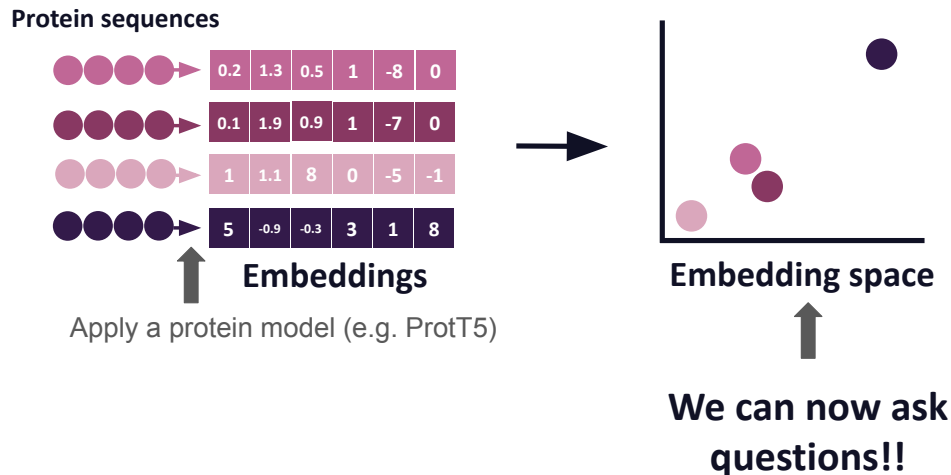
PART II — AI-assisted phylogenomics

Two main 'lines' of development of methods

- Genome/Protein Language Models to recode sequences and 'learn' the *grammar* of genomes

From sequences to embeddings

- Each protein → vector in high-dimensional space
- Similar function/evolution → nearby vectors



PART II — AI-assisted phylogenomics

Two main 'lines' of development of methods

- Genome/Protein Language Models to recode sequences and 'learn' the *grammar* of genomes

Article | [Open access](#) | Published: 14 August 2025

FANTASIA leverages language models to decode the functional dark proteome across the animal tree of life

[Gemma I. Martínez-Redondo](#) ✉, [Francisco M. Perez-Canales](#), [Belén Carbonetto](#), [José M. Fernández](#), [Israel Barrios-Núñez](#), [Marçal Vázquez-Valls](#), [Ildefonso Cases](#), [Ana M. Rojas](#) ✉ & [Rosa Fernández](#) ✉

[Communications Biology](#) 8, Article number: 1227 (2025) | [Cite this article](#)

6933 Accesses | 5 Citations | 97 Altmetric | [Metrics](#)

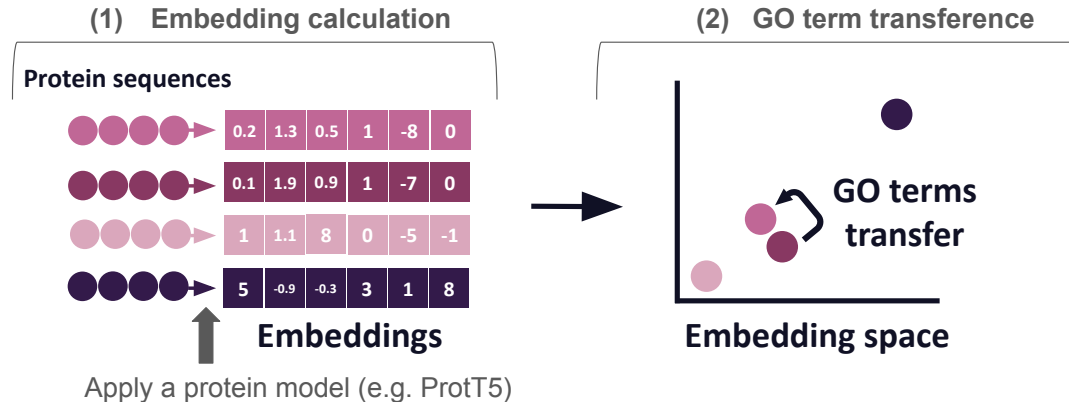
Abstract

Protein functional annotation is crucial in biology, but many protein-coding genes remain uncharacterized, especially in non-model organisms. **FANTASIA (Functional ANnoTation based on embedding space Similarity)** integrates protein language models for large-scale

JOURNAL ARTICLE | EDITOR'S CHOICE

Decoding functional proteome information in model organisms using protein language models

[Israel Barrios-Núñez](#), [Gemma I Martínez-Redondo](#), [Patricia Medina-Burgos](#), [Ildefonso Cases](#) ✉, [Rosa Fernández](#) ✉, [Ana M Rojas](#) ✉ [Author Notes](#)



Different language models (ProtT5, SeqVec, ESM2, Ankh3, etc)

PART II — AI-assisted phylogenomics

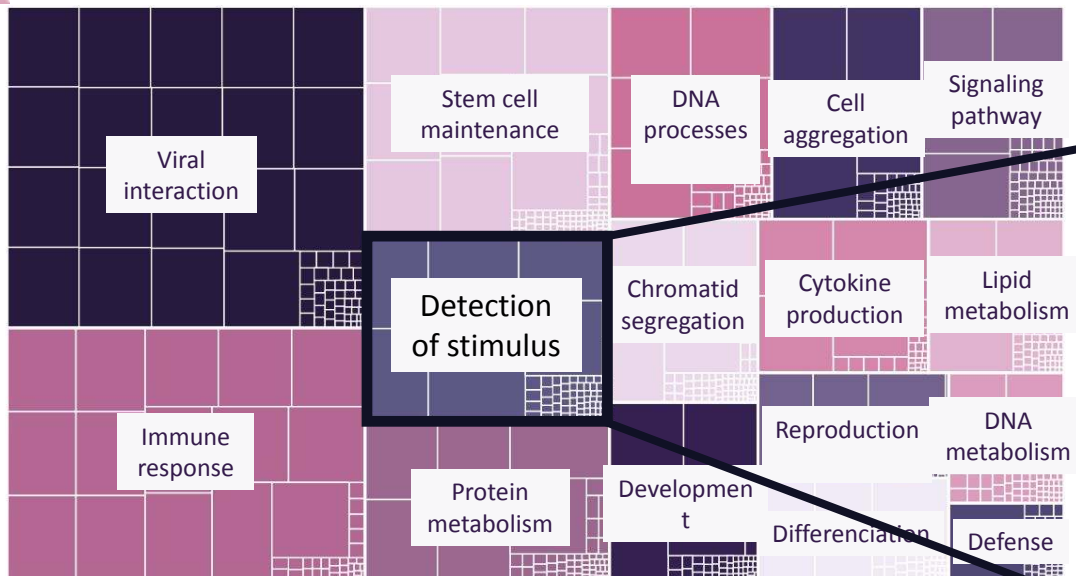
Two main 'lines' of development of methods

- Genome/Protein Language Models to recode sequences and 'learn' the *grammar* of genomes



Investigating the 'dark proteome' of neglected species/lineages

GO enrichment



Ca. 8,000 genes in tardigrades without GO terms based on homology

Cellular response to UV
Response to UV-A
Response to water deprivation
Cellular response to heat
Response to osmotic stress
Cellular response to DNA-damage stimulus
Cellular response to hypoxia

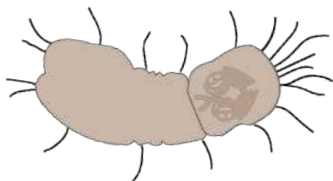
PART II — AI-assisted phylogenomics

Two main 'lines' of development of methods

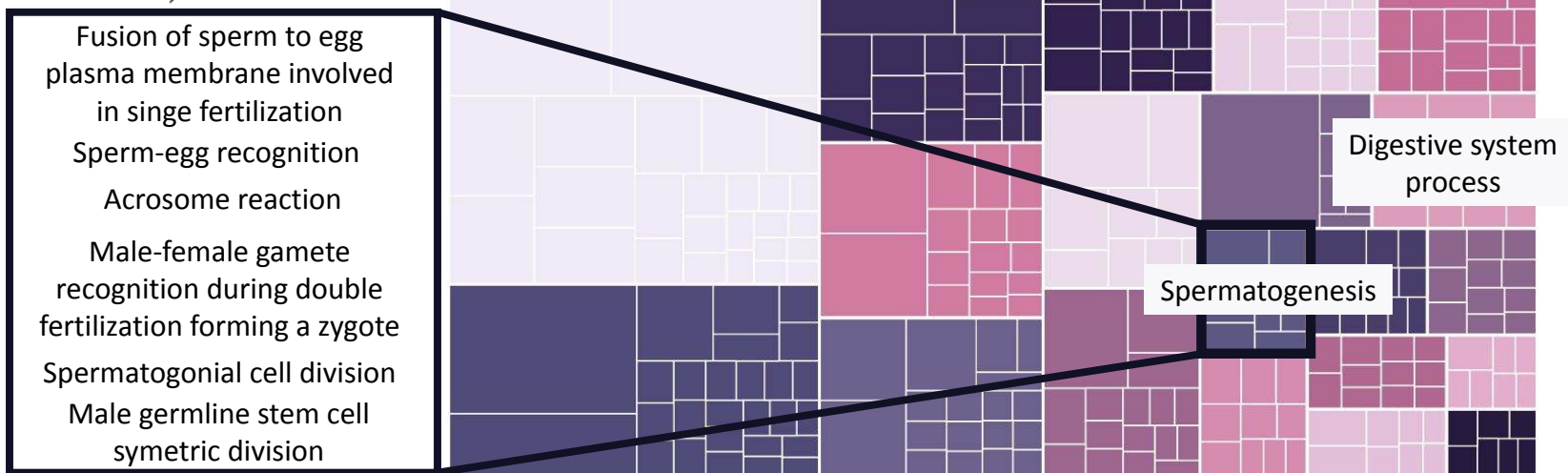
- Genome/Protein Language Models to recode sequences and 'learn' the *grammar* of genomes



Investigating the 'dark proteome' of neglected species/lineages



Micrognathozoa



Fusion of sperm to egg
plasma membrane involved
in single fertilization
Sperm-egg recognition
Acrosome reaction
Male-female gamete
recognition during double
fertilization forming a zygote
Spermatogonial cell division
Male germline stem cell
symetric division

Digestive system
process

Spermatogenesis

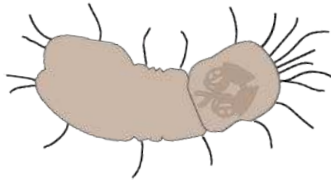
PART II — AI-assisted phylogenomics

Two main 'lines' of development of methods

- Genome/Protein Language Models to recode sequences and 'learn' the *grammar* of genomes



Investigating the 'dark proteome' of neglected species/lineages



Micrognathozoa

Fusion of sperm to egg
plasma membrane involved
in single fertilization
Sperm-egg recognition

Digestive system
process

If all high scores are noise -> No enrichment
Enrichment -> model isn't hallucinating at random

Male germline stem cell
symetric division

PART II — AI-assisted phylogenomics

Two main 'lines' of development of methods

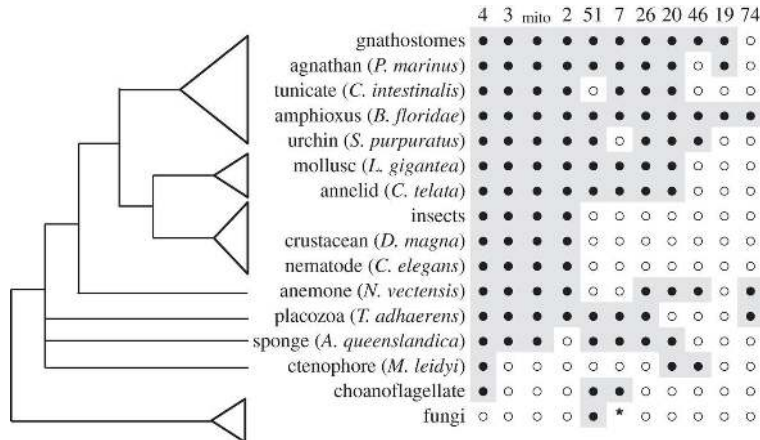
- Genome/Protein Language Models to recode sequences and 'learn' the *grammar* of genomes



Scaling up comparative genomics (exploration of orthogroups/gene families)

1,000 animal genomes from all phyla, 24 million genes, 520K orthogroups ('gene families')

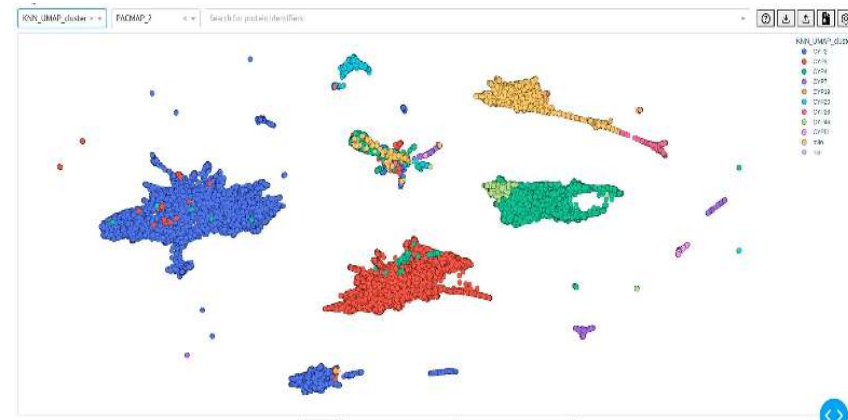
Example 1: Largest orthogroups: **CYTOCHROME P450** (83K genes; **48K** > 300 aa; 11 clans)



Nelson et al. (2013)



→
Embeddings
+
supervised
Machine
Learning



Martínez-Redondo et al. (in prep)

PART II — AI-assisted phylogenomics

Two main 'lines' of development of methods

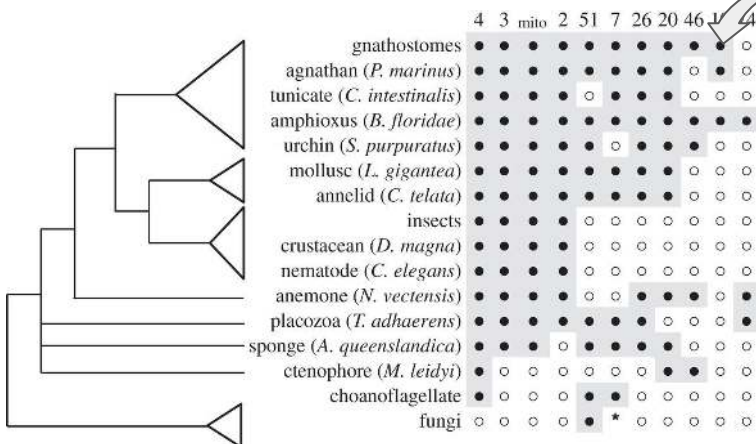
- Genome/Protein Language Models to recode sequences and 'learn' the *grammar* of genomes



Scaling up comparative genomics (exploration of orthogroups/gene families)

1,000 animal genomes from all phyla, 24 million genes, 520K orthogroups ('gene families')

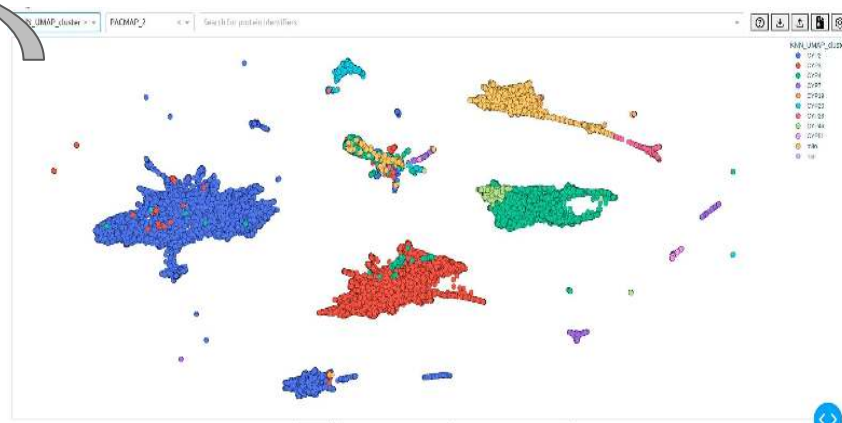
Example 1: Largest orthogroups: **CYTOCHROME P450** (83K genes; **48K** > 300 aa; 11 clans)



Nelson et al. (2013)



Embeddings
+
supervised
Machine
Learning



Martínez-Redondo et al. (in prep)

PART II — AI-assisted phylogenomics

Two main 'lines' of development of methods

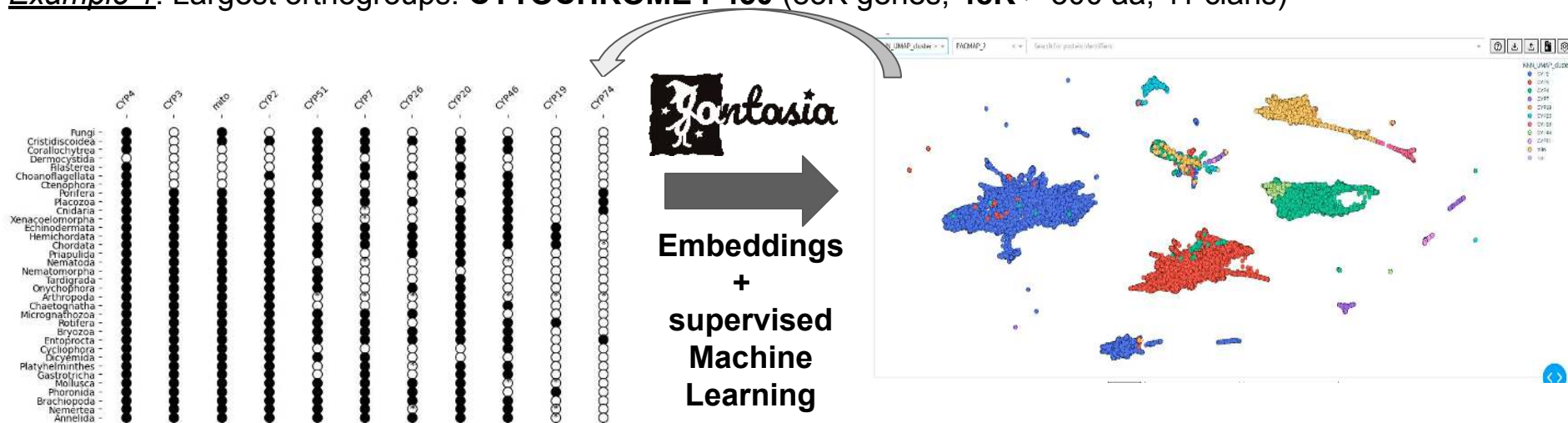
- Genome/Protein Language Models to recode sequences and 'learn' the *grammar* of genomes



Scaling up comparative genomics (exploration of orthogroups/gene families)

1,000 animal genomes from all phyla, 24 million genes, 520K orthogroups ('gene families')

Example 1: Largest orthogroups: **CYTOCHROME P450** (83K genes; **48K** > 300 aa; 11 clans)



PART II — AI-assisted phylogenomics

Two main ‘lines’ of development of methods

- Genome/Protein Language Models to recode sequences and ‘learn’ the *grammar* of genomes

Other potential applications of embeddings

Species trees from embeddings (in progress)

- Aggregate protein embeddings across genomes
- Compute genome–genome distances
- Infer species relationships without MSAs

Caution

- Functional and phylogenetic signals are entangled
- Models also learn dataset biases

PhyloGen: Language Model-Enhanced Phylogenetic Inference via Graph Structure Generation

Chenrui Duan^{1,2*} Zelin Zang^{2*} Siyuan Li^{1,2} Yongjie Xu^{1,2} Stan Z. Li^{2†}
¹Zhejiang University, College of Computer Science and Technology; ²Westlake University
duanchenrui@westlake.edu.cn;
{zangzelin; lisiyuan; xuyongjie; stan.zq.li}@westlake.edu.cn
^{*}Equal contribution [†]Corresponding author

Genome language model (DNABERT2)

JOURNAL ARTICLE

Do protein language models learn phylogeny?

Sanjana Tule, Gabriel Foley, Mikael Bodén ✉

Briefings in Bioinformatics, Volume 26, Issue 1, January 2025, bbaf047,

<https://doi.org/10.1093/bib/bbaf047>

Protein language models
Individual gene trees w/o MSA

PART II — AI-assisted phylogenomics

From gene phylogenies to embedding trees - Conceptual challenges

Embedding-based phylogenomics forces a redefinition of what is being inferred, shifting from explicit models of mutational change to implicit representations of evolutionary constraint. It demands new criteria for interpretation, validation, and trust.

A few (of many) open questions

- **Are embedding distances measures of ancestry, evolutionary constraint, or learned functional similarity? Can these be disentangled?**
- **What replaces explicit models of sequence evolution? What is the implicit evolutionary process acting on embeddings, and can it be formalized and validated?**
- **How should uncertainty and statistical support be defined?**
- **Under which evolutionary regimes (deep time, high divergence, domain reshuffling, convergence) do embeddings provide genuinely new or more reliable signal?**

PART II — AI-assisted phylogenomics

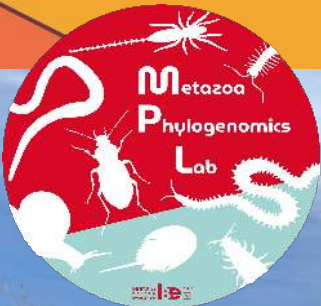
Take-home message

Let's Play!



From: <https://michellekassorla.substack.com/p/an-ai-playground>

Acknowledgements



www.metazomics.com



[@rosafernandez.bsky.social](https://www.bsky.social/@rosafernandez)



rosa.fernandez@ibe.upf-csic.es



European Research Council
Established by the European Commission

SEA2LAND, grant agreement no. 948281



International

Human Frontier
Science Program

Organization

Grant agreement no. RGY0056/2022



EUROPEAN REFERENCE GENOME ATLAS



Agència
de Gestió
d'Ajuts
Universitaris
i de Recerca



GOBIERNO
DE ESPAÑA

MINISTERIO
DE CIENCIA
E INNOVACIÓN

