Lecture 1: Introduction

Jadranka Rota and Niklas Wahlberg

Systematic Biology Group

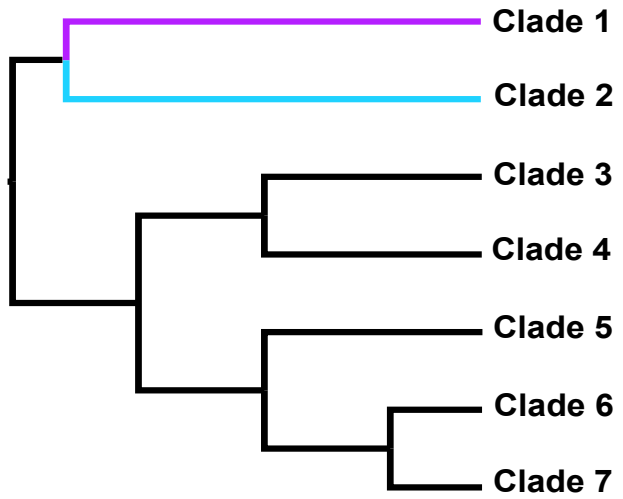
Department of Biology

Lund University

LUND UNIVERSITY

Aims

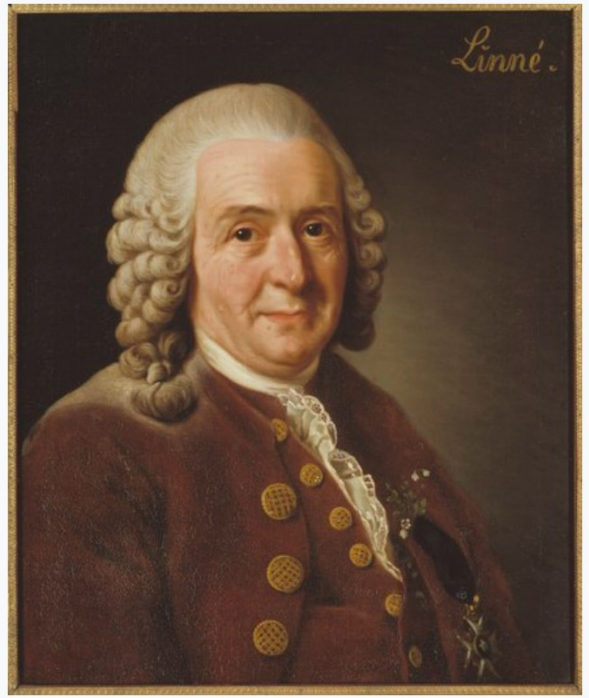
- **Cover key concepts in phylogenetics**
  - **E.g. monophyly, homology, analogy**
- **Explain why evolutionary history is important in biology**
- **Understand the basics of statistical phylogenetic inference**
- **Develop “tree thinking”**



Systematics is ….

- **The scientific study of the kinds and diversity of organisms and of the relationships among them**
- **Traditionally: taxonomy (naming and classifying organisms, Greek "taxis" = arrangement, and "nomos" = law)**
- **Since ca. 1980s: largely based on phylogenetics (first morphological, then molecular)**
- **Most recently: including phylogenomics**
- **Provides essential framework for recognition and study of biodiversity and evolution**

Carl Linnaeus



Carl von Linné by Alexander Roslin, 1775
(oil on canvas, Gripsholm Castle)

Known as the “Father of modern taxonomy”
(Source: Wikipedia)

Human need for taxonomy

- **Naming of organisms around us**
  - **Makes communication a lot easier!**
- **Sorting different groups into higher categories**
  - **Helps us organize the living world around us**
  - **E.g. porcini (Karljohanssvamp) is an edible species of mushrooms, but many other mushrooms are very poisonous, e.g. fly amanita (Röd flugsvamp)**

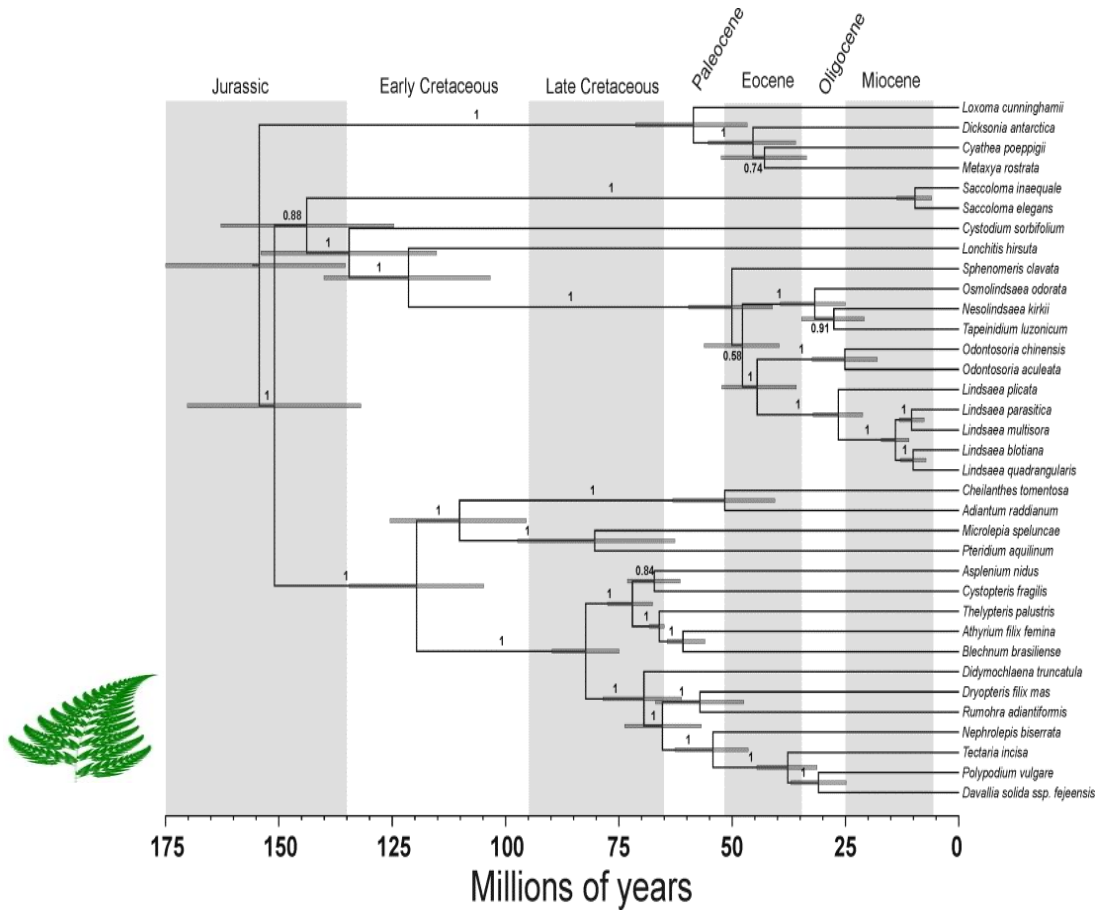
Note: one taxon, many taxa (in some Greek words singular ending is –on, plural is –a, e.g. also phenomenon, phenomena)

Porcini - *Boletus edulis* Bull. (1782)

Fly amanita - Amanita muscaria (L.) Lam. (1783)

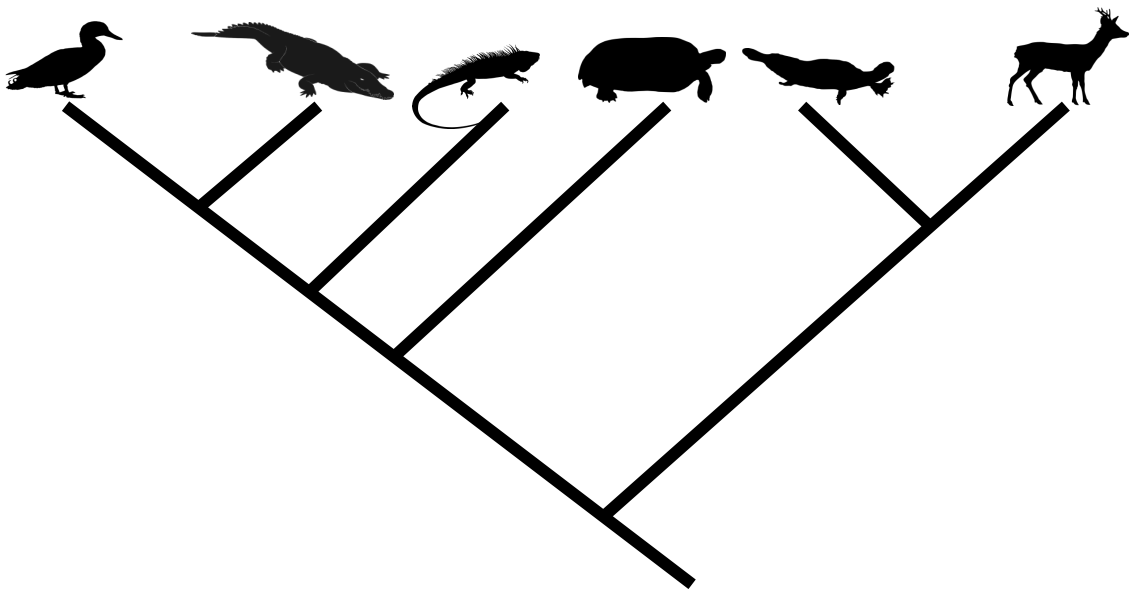
Systematics also includes…

- **The study of character evolution**
- **The study of molecular evolution**
- **The study of speciation/extinction dynamics**
- **The study of historical biogeography**
- **The study of the temporal framework of evolution**



Evolutionary History

- **How do we learn about the evolutionary history of organisms?**
- **Why should we care about it?**

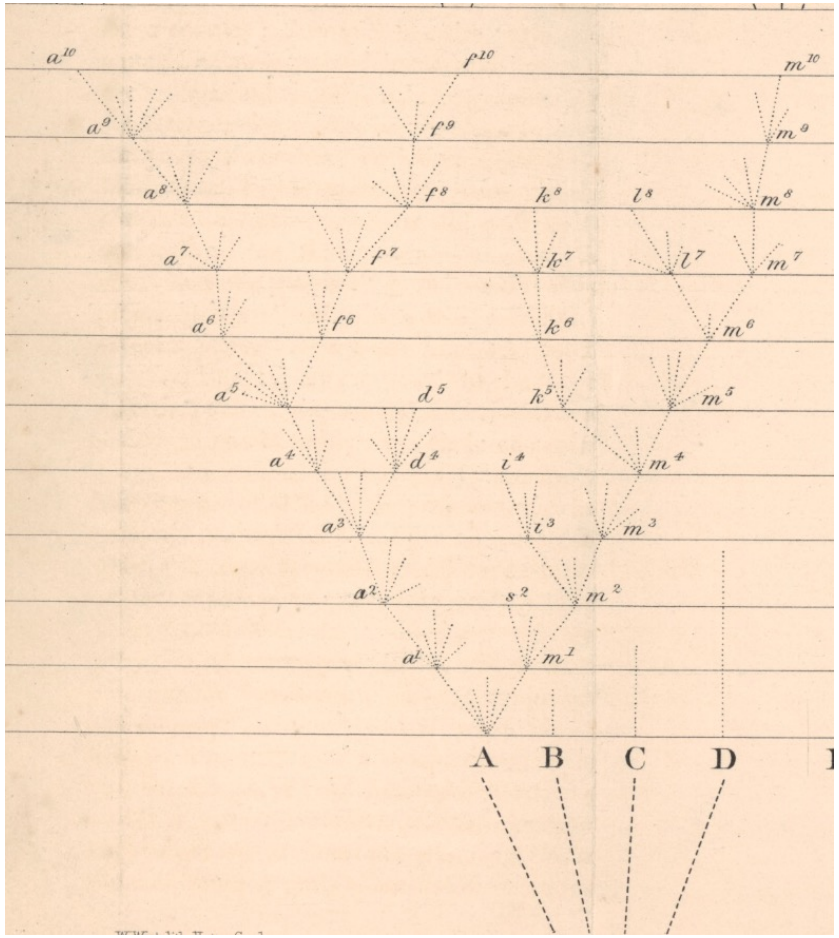


“Nothing in biology makes sense except in the light of evolution”

- Theodosius Dobzhansky, essay written in 1973

“Nothing in evolution makes sense except in the light of phylogeny”

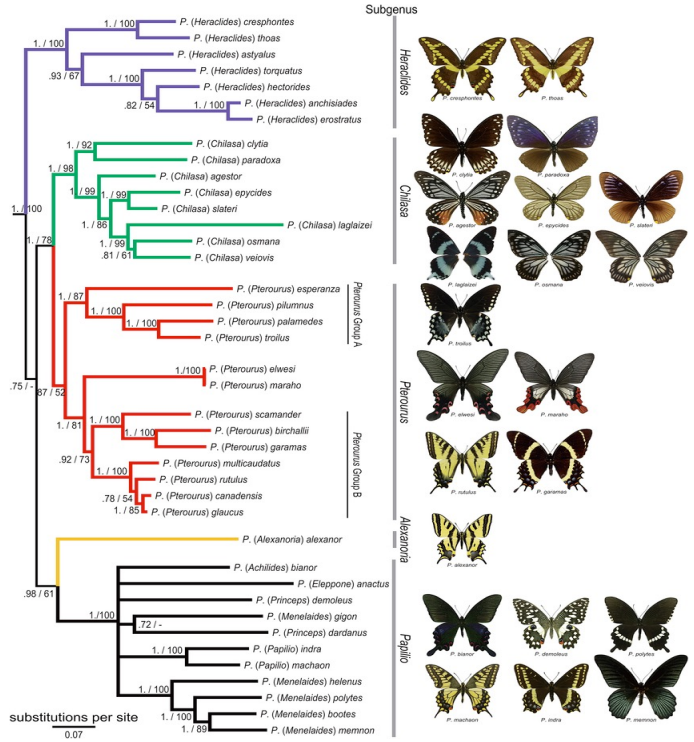
- Quote from Jay Savage, in 1997 Society of Systematic Biology Presidential Address



The only figure in Darwin’s On the Origin of Species

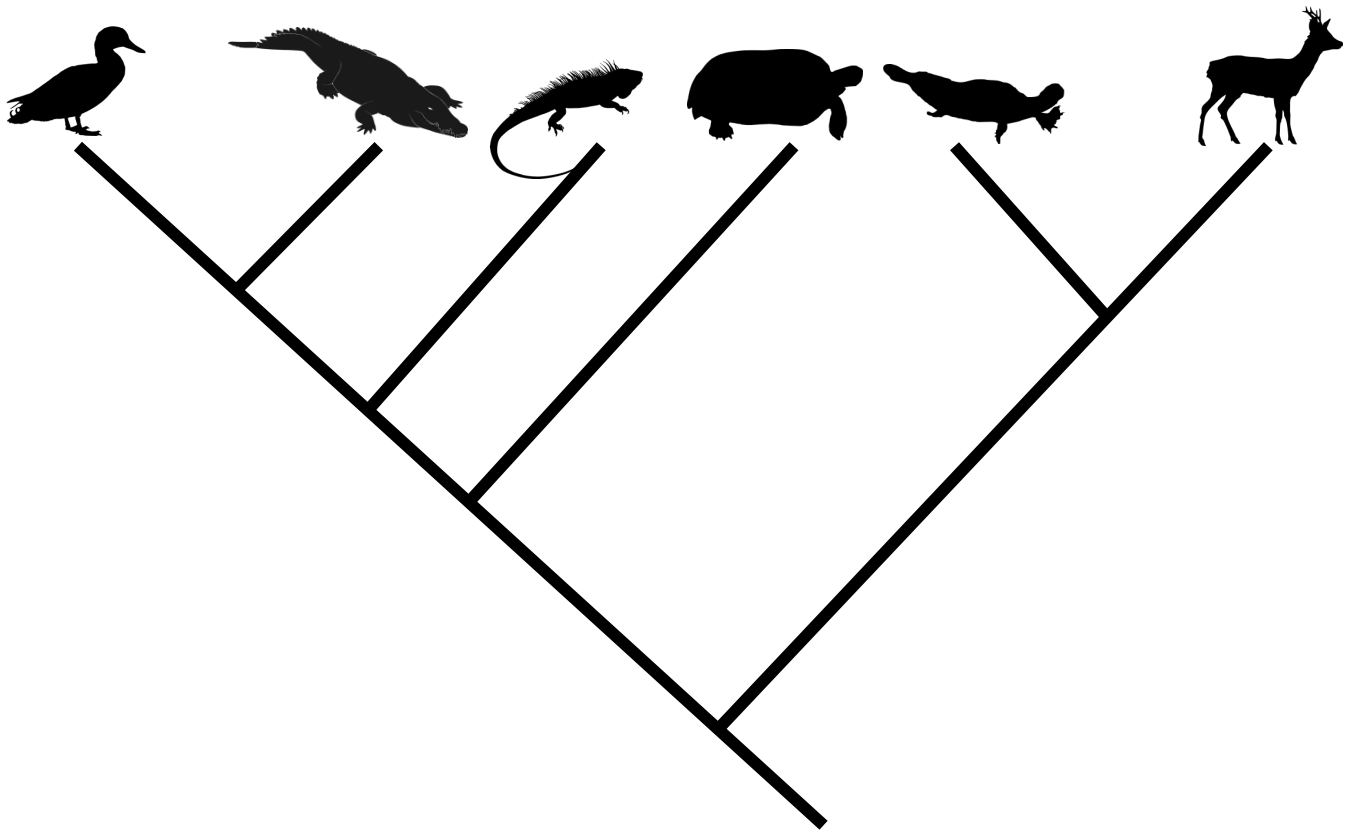
The very basic facts

- **What we see today in nature is the outcome of what has happened in the past**
- **Ecology and evolution are inseparable**
- **“Species” or “genes” are not individual entities without any connections to other species or genes**
  - **phylogeny**



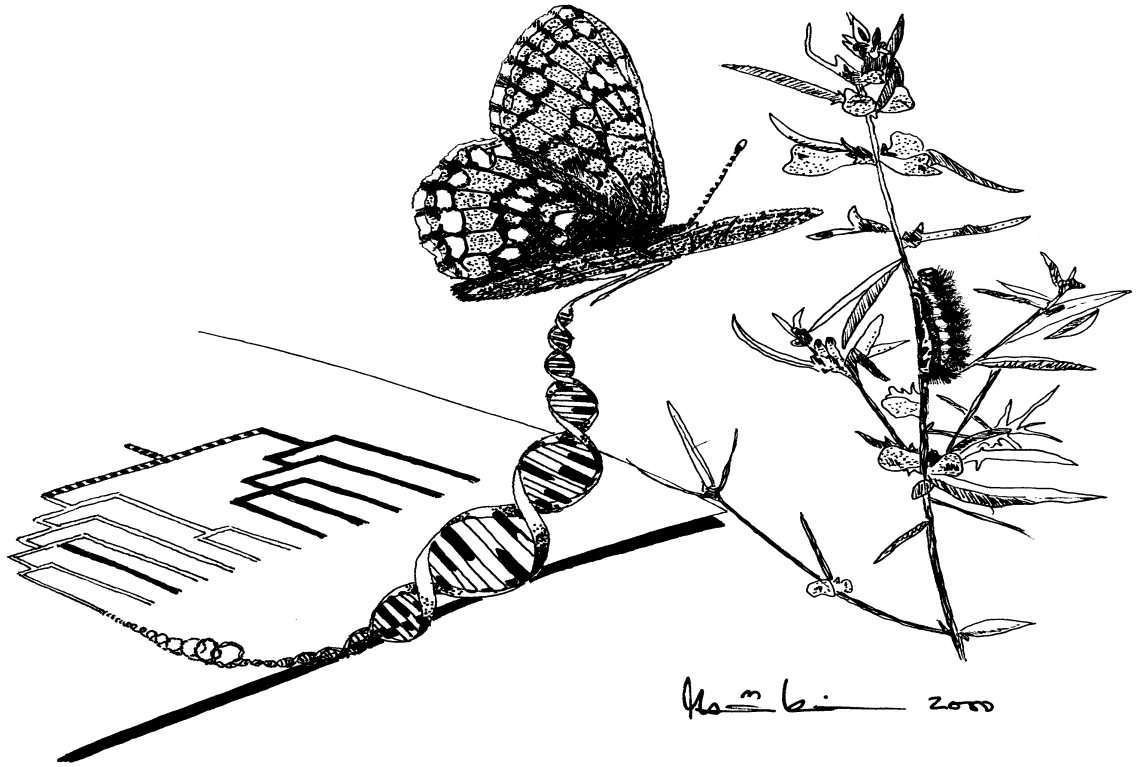
What is a phylogeny?

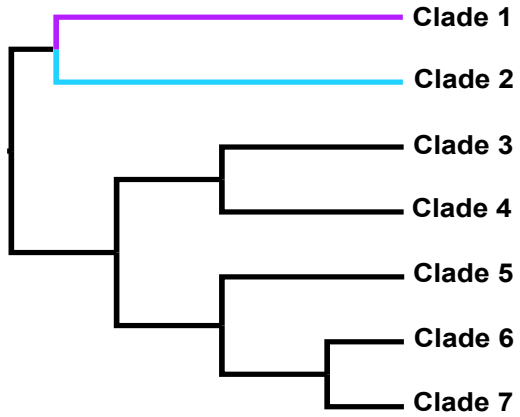
- **A phylogeny is the historical genealogy of a group of species**



A phylogeny is an inference

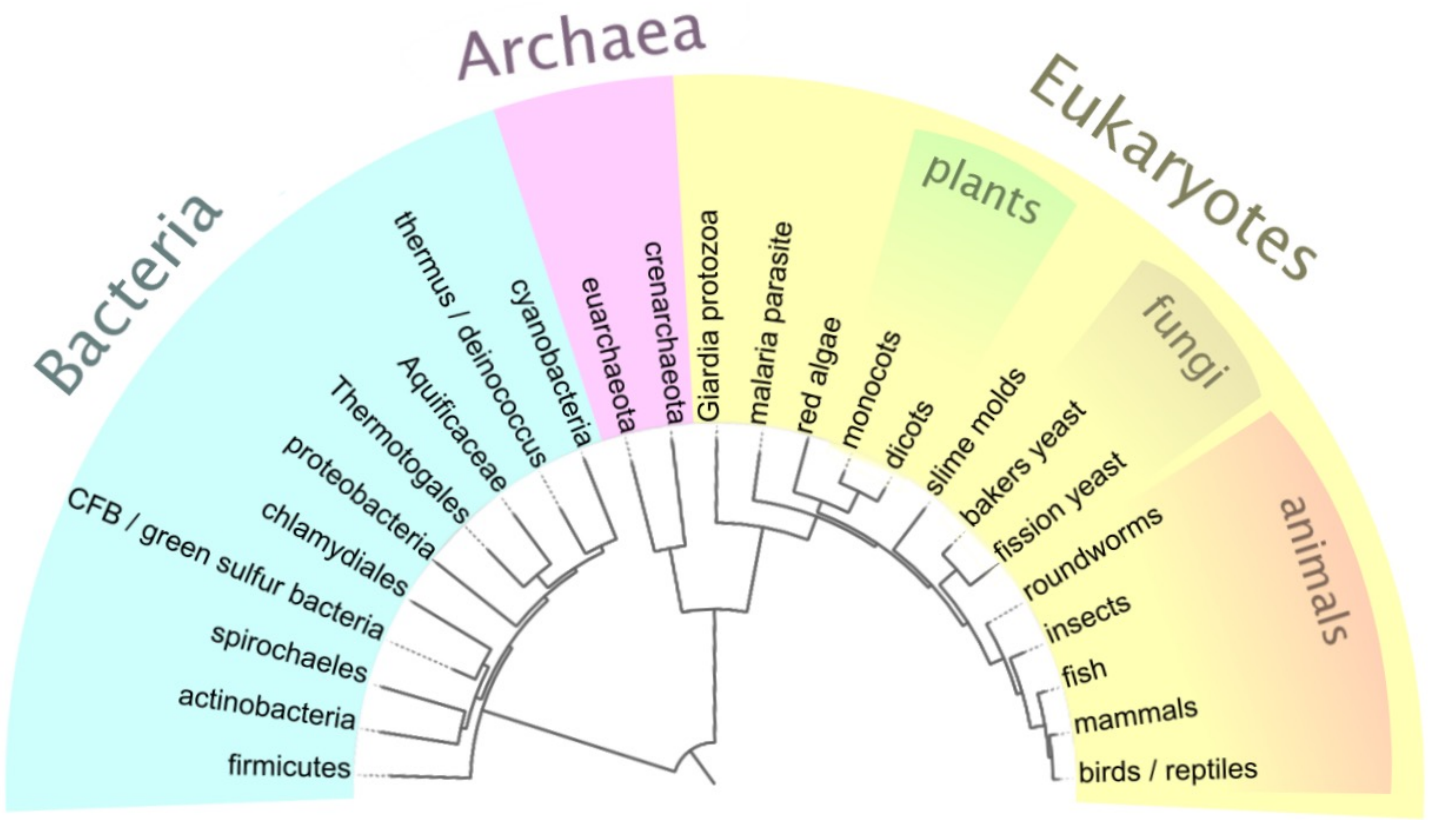
- **Envisioned as a dichotomously branching tree**
- **A phylogeny cannot be observed**
- **A phylogenetic hypothesis can be inferred from observed data**





What we are after

- **Phylogenies – the Tree of Life**
- **With phylogenies we are attempting to get a good working framework for Life**
- **Getting to the root of how evolution has worked**

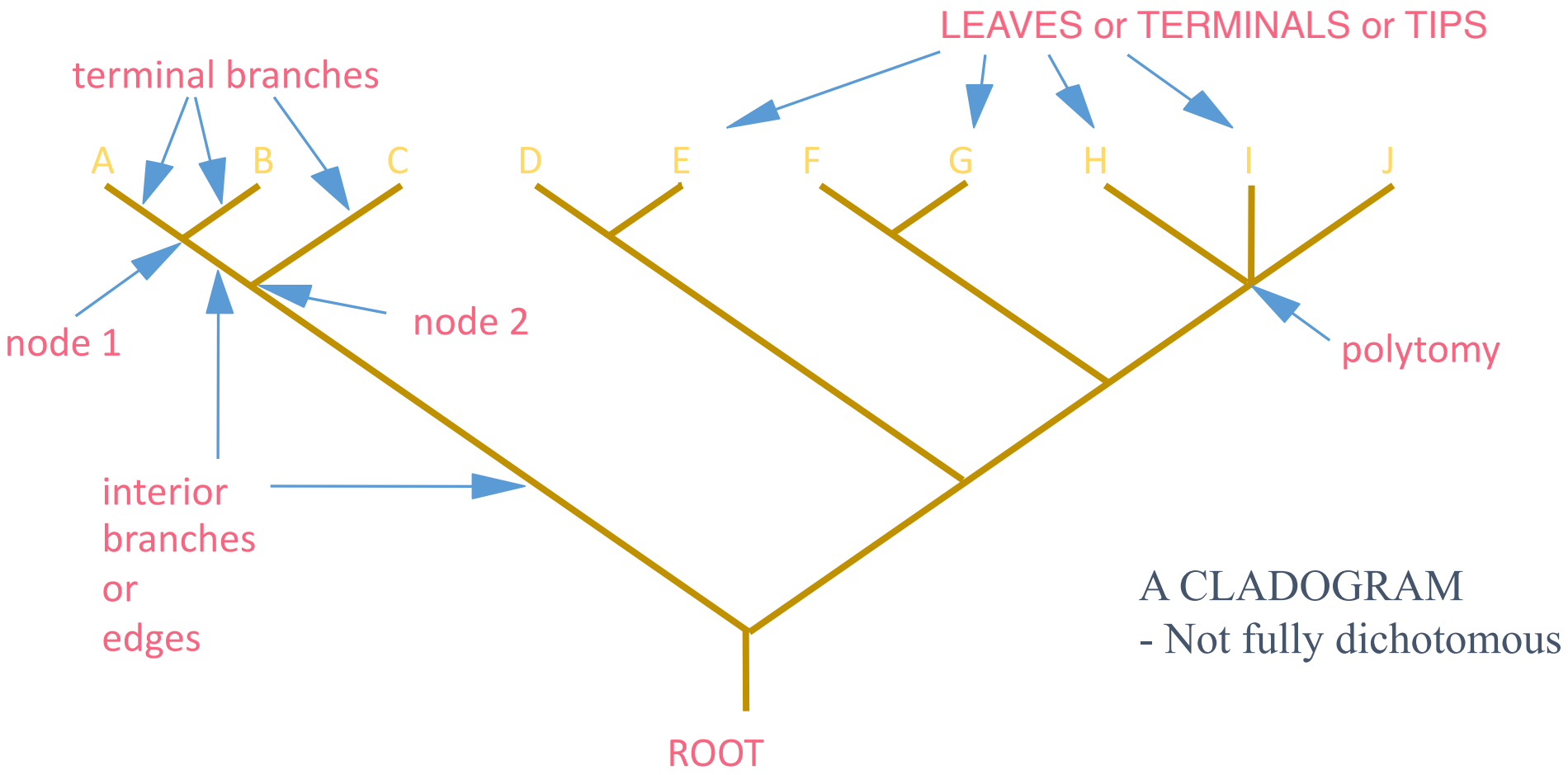


Source: Wikipedia

Some basic concepts

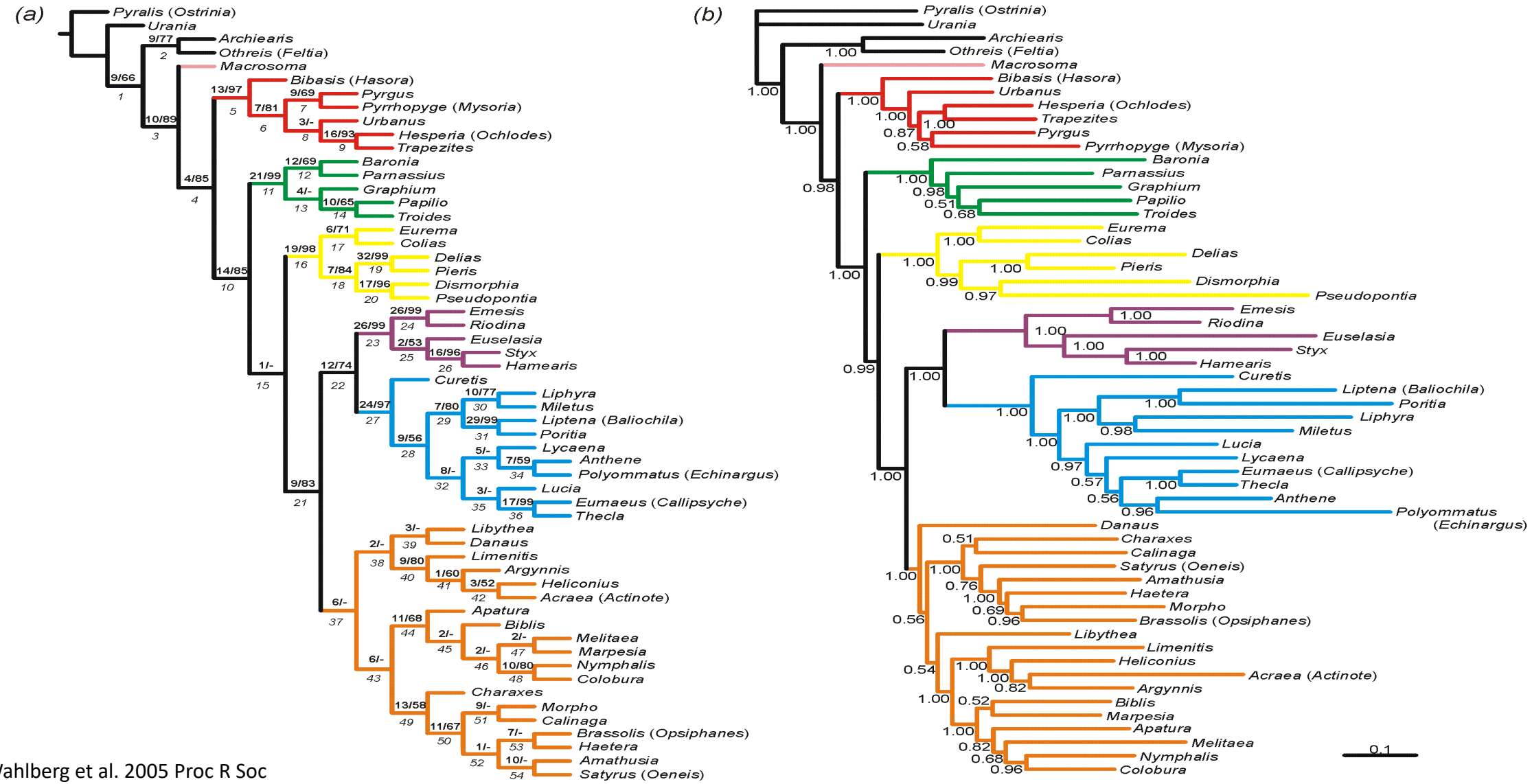
- **Cladogram – a tree diagram which depicts a hypothesised evolutionary history (topology)**
- **Phylogram – a tree which indicates by branch length the degree of change believed to have occurred along each lineage (topology with informative branch lengths)**
- **Chronogram – a tree in which branch lengths are directly in proportion to time (a type of an ultrametric tree – all tips are equidistant from the root)**

Phylogenetic Trees



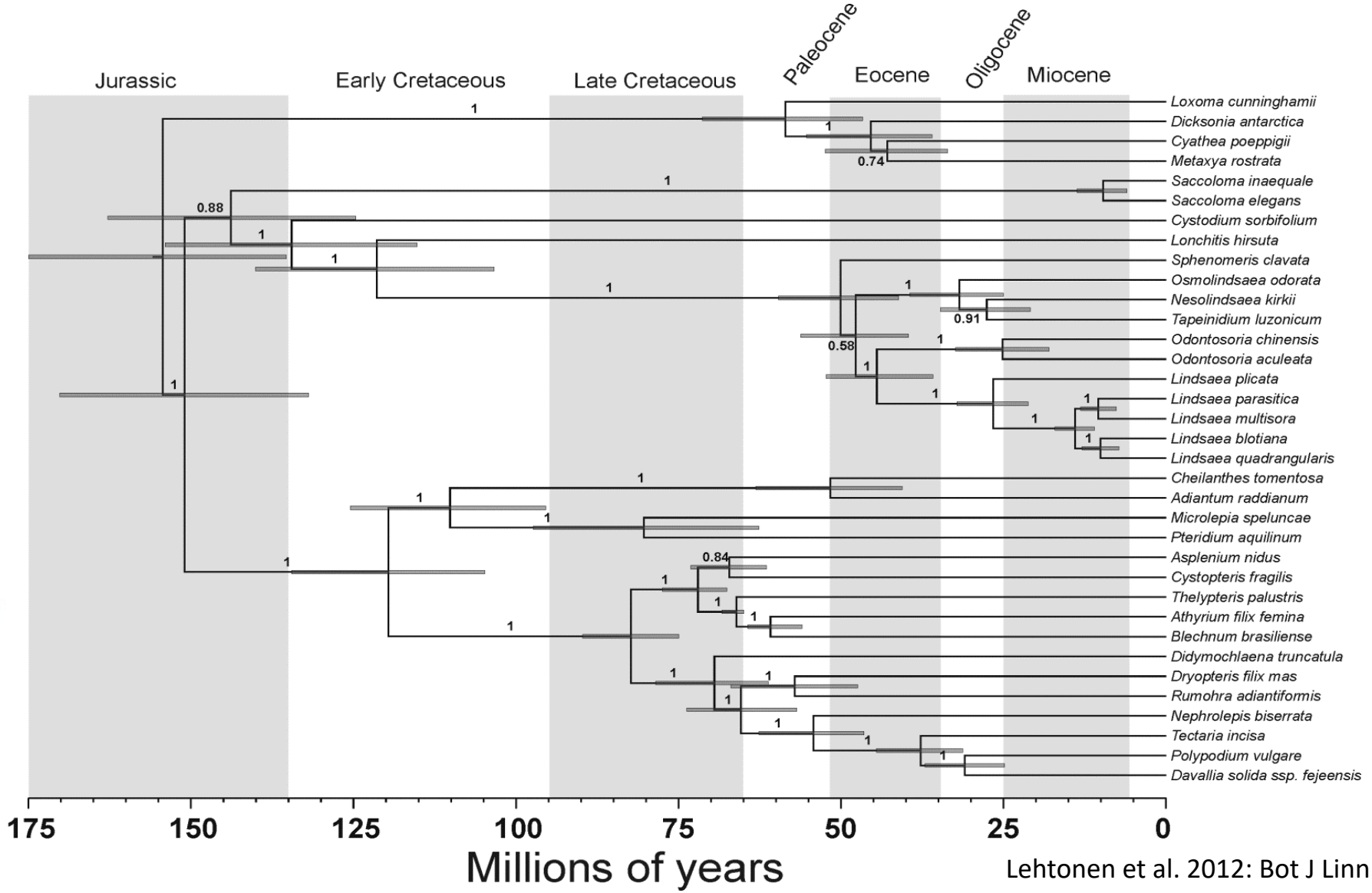
A CLADOGRAM
- Not fully dichotomous

Cladograms and phylograms



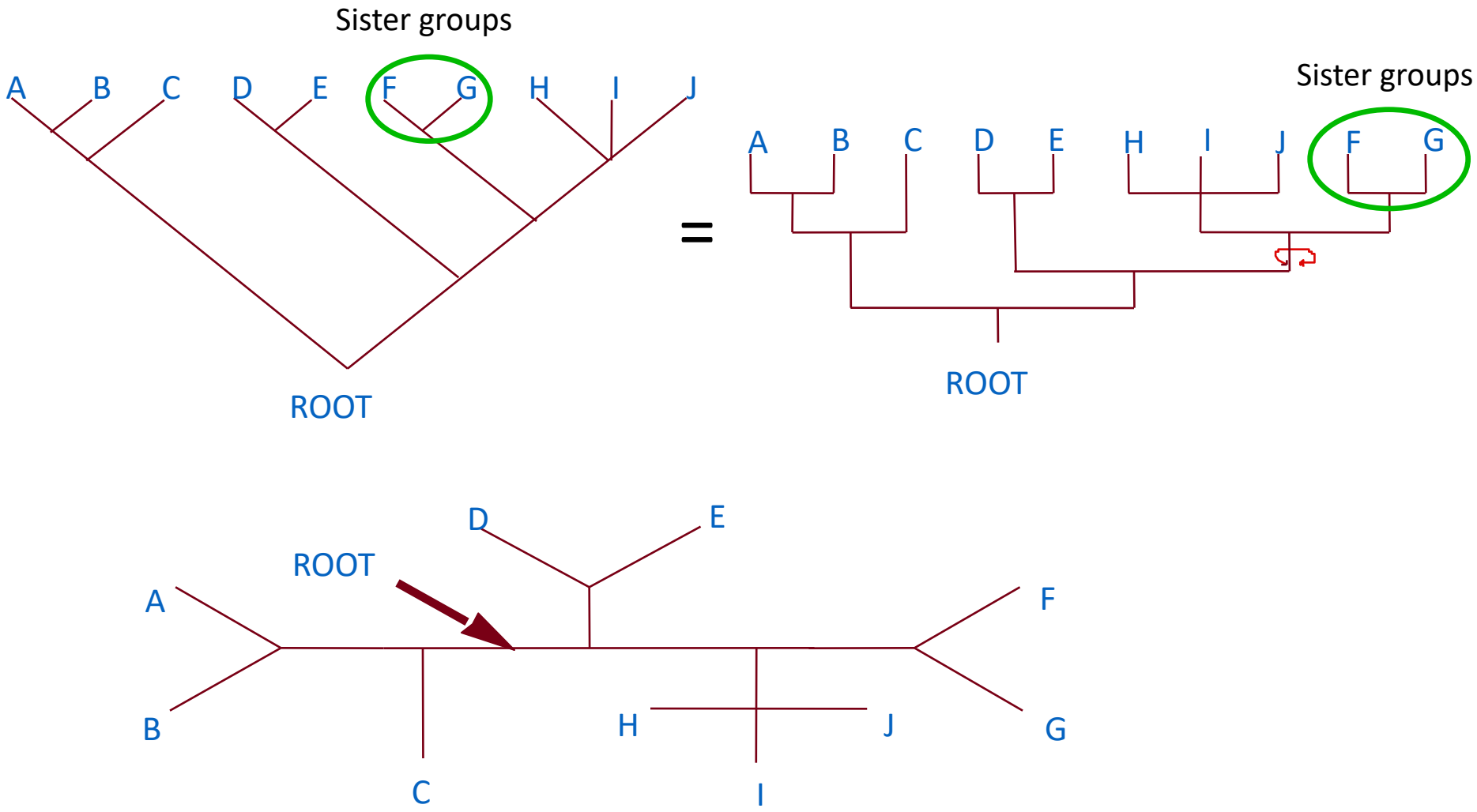
Wahlberg et al. 2005 Proc R Soc

Chronogram
(an ultrametric tree – all tips equidistant from the root)



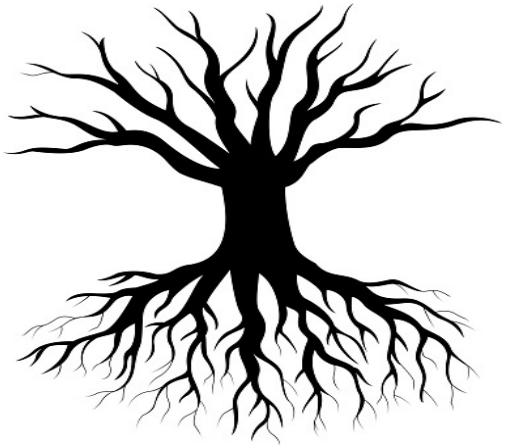
Lehtonen et al. 2012: Bot J Linn Soc 170: 489.

Trees - Rooted and Unrooted



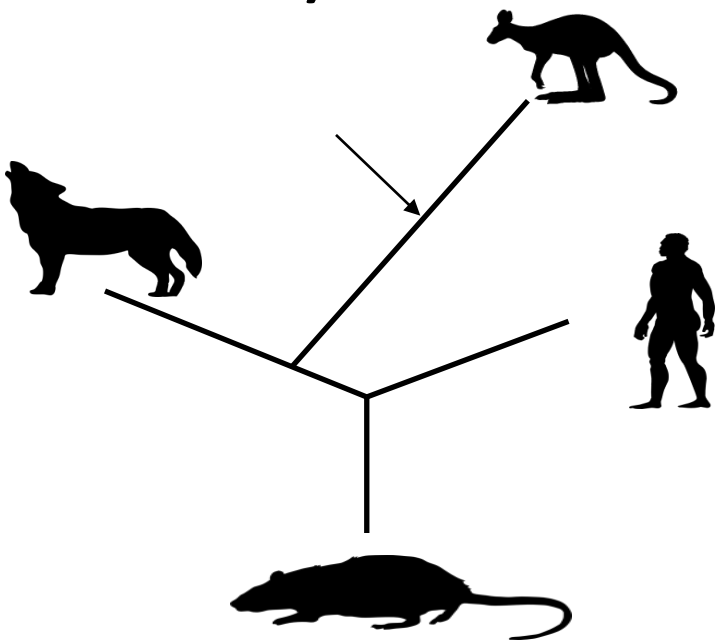
Rooting a tree

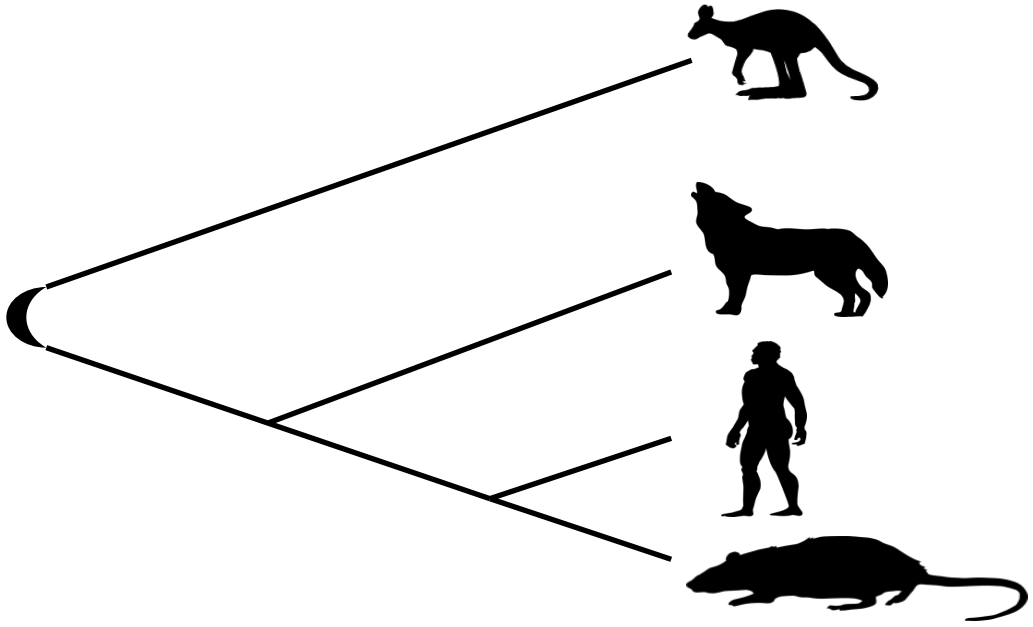
- **Rooting a tree using outgroups**
  - **Commonly we include several outgroups**
  - **Place the root on the branch leading to the outgroup taxon**
- **Other ways of rooting a tree**
  - **Assume a molecular clock**
  - **Midpoint rooting (root on the longest branch)**



Outgroup rooting of unrooted trees

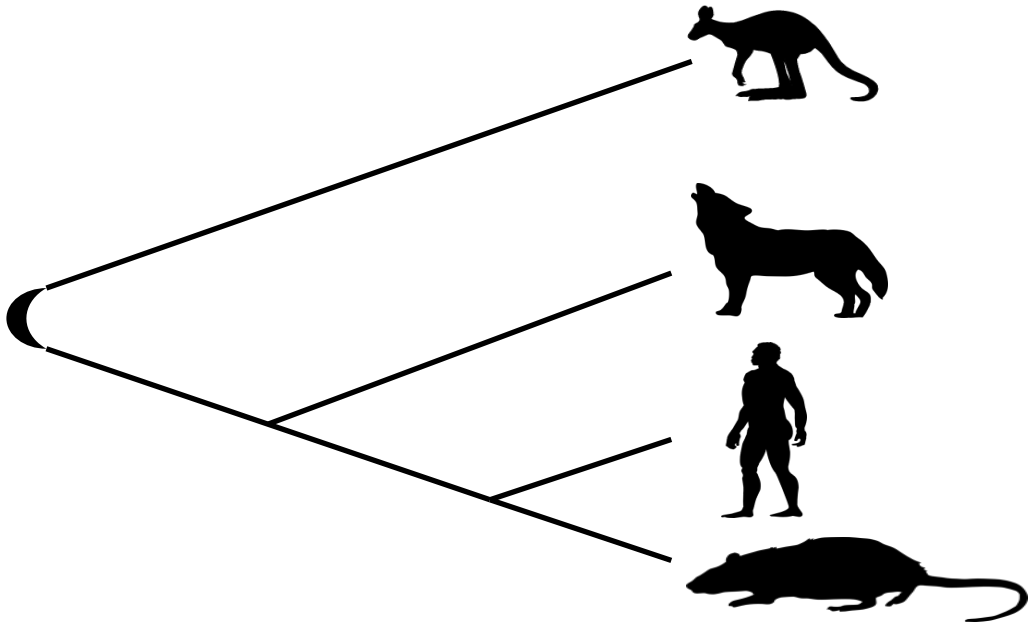
- **Outgroup – related group that definitely diverged earlier (palaeontological evidence)**
- **Not too distantly related (tree method becomes unreliable if it is too distant)**



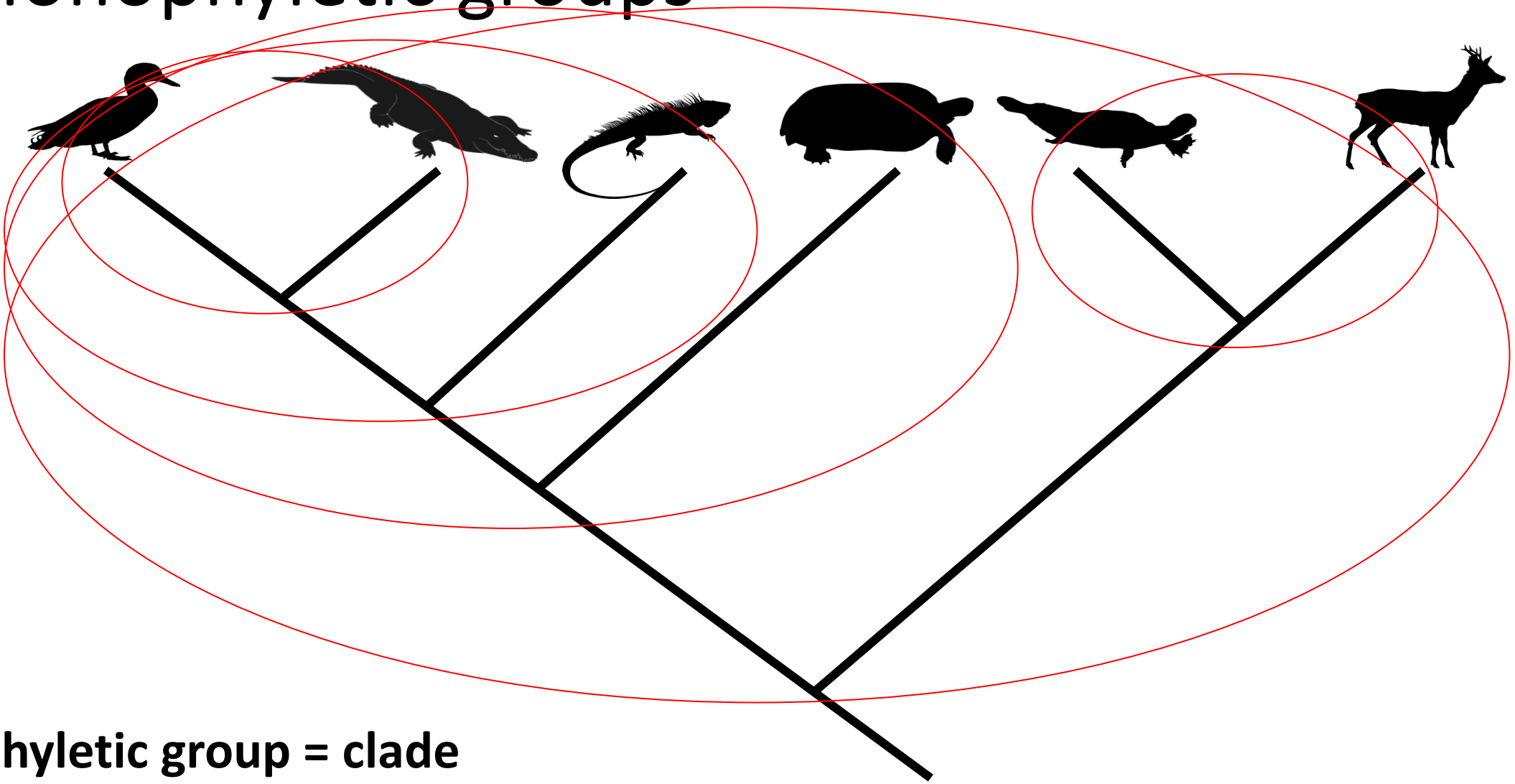


Phylogenetic systematics

- **Uses tree diagrams to portray relationships based upon recency of common ancestry**
- **Monophyletic groups (clades) – contain species which are more closely related to each other than to any outside of the group, including the MRCA (most recent common ancestor)**



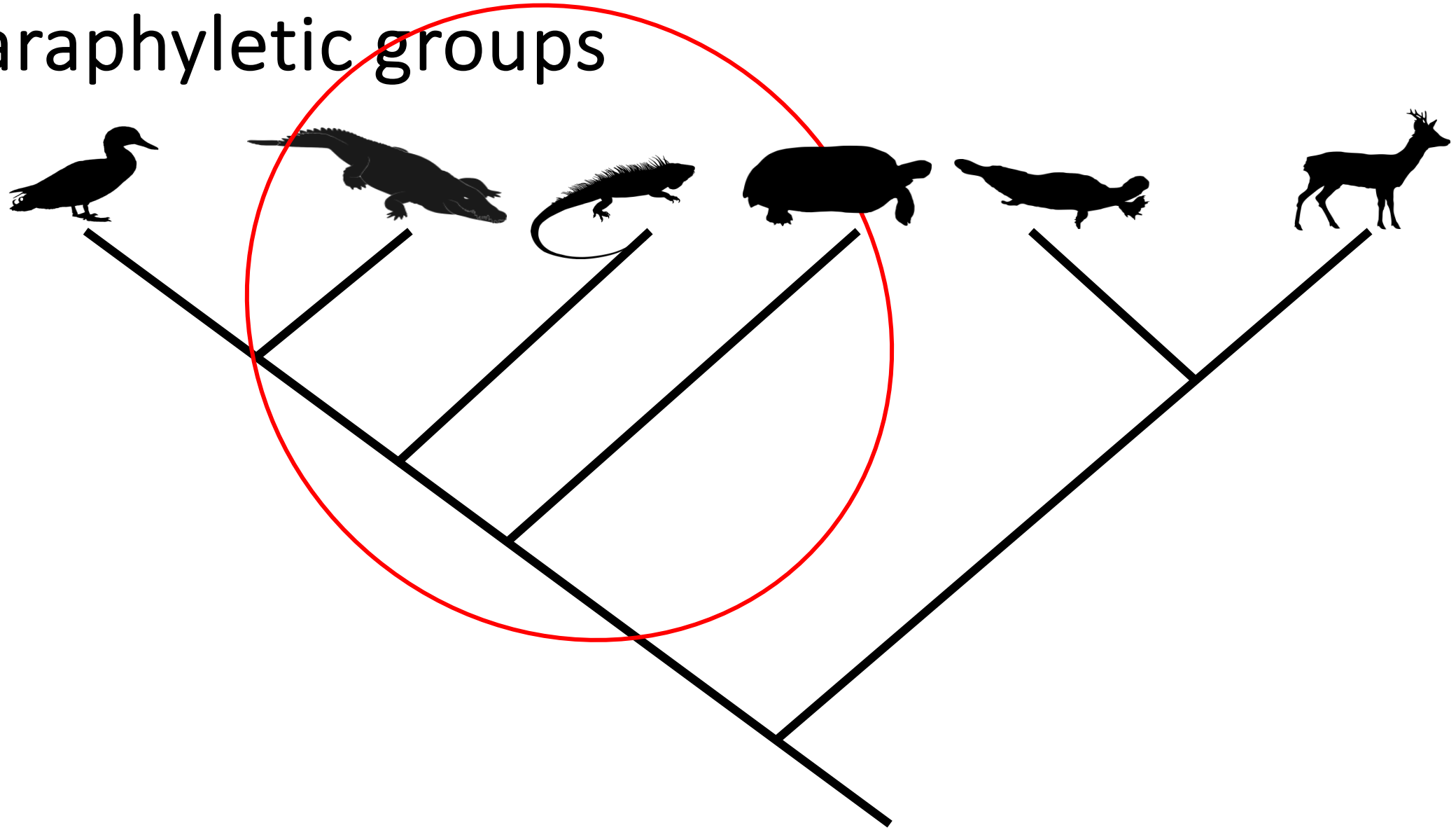
Monophyletic groups



Monophyletic group = clade

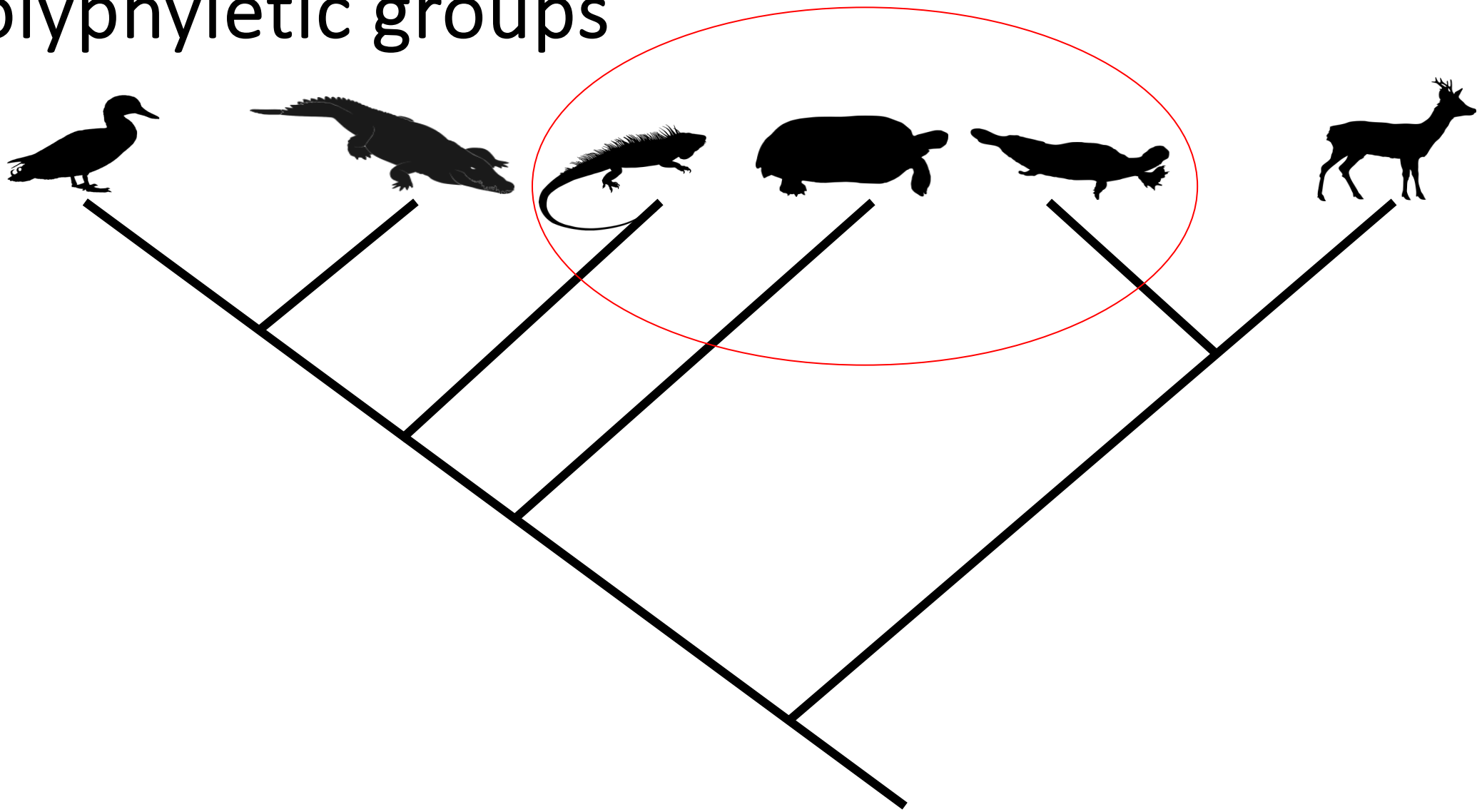
- Include an ancestor and all of its descendants

Paraphyletic groups



- Include ancestor and some but not all of its descendants

Polyphyletic groups



- Include some but not all of the descendants and exclude the ancestor

Sister groups

- **By definition, sister groups are of equal age**
- **Common mistake**
  - **The sister group that has fewer species is referred to as basal**
  - **Possible to have nodes that are more basal than other nodes, but not lineages compared to their sister group**
  - **Rather than saying “this group is basal”, one should say “this group is sister to all other lineages”**

This question is actually not yet considered resolved!

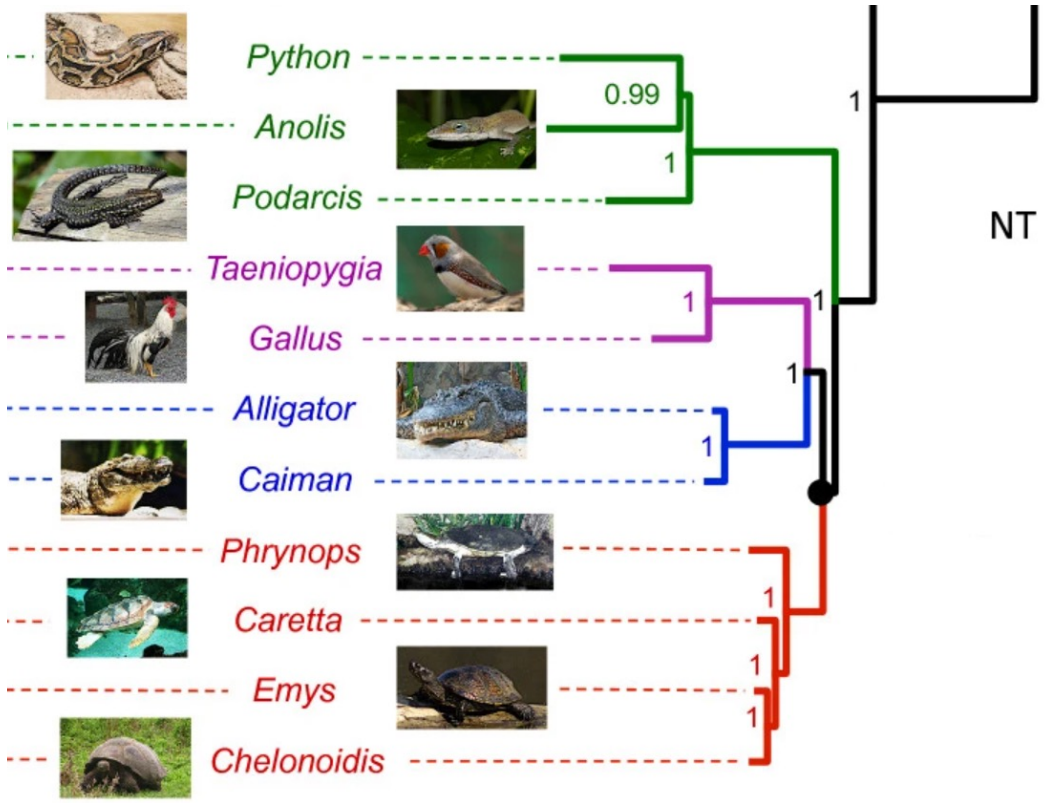
Research article | Open Access | Published: 27 July 2012

Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria)

Ylenia Chiari, Vincent Cahais, Nicolas Galtier & Frédéric Delsuc

BMC Biology **10**, Article number: 65 (2012) | Cite this article

32k Accesses | **241** Citations | **50** Altmetric | Metrics

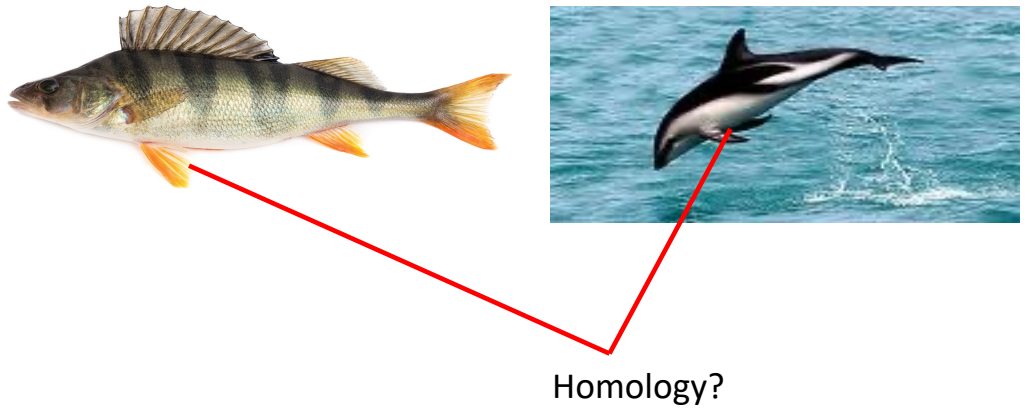


Some premises underlying phylogenetic inferences

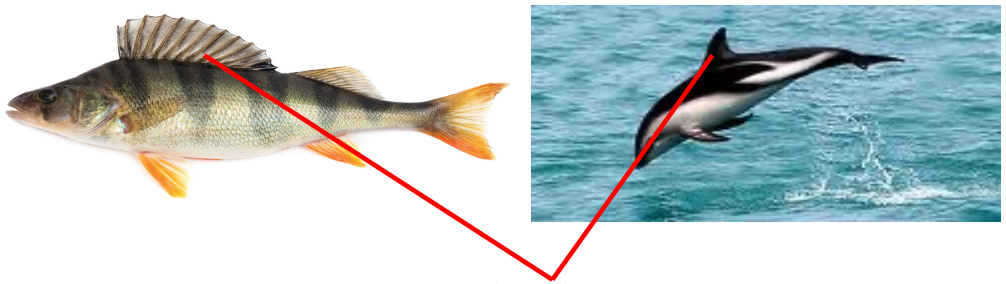
- **Phylogenetic inferences are premised on**
  - **the inheritance of ancestral characters**
  - **the existence of a shared evolutionary history**
- **Homology considered as evidence of common ancestry**
- **A tree-like model of evolution**
  - **There are evolutionary processes that don’t fit this model, e.g. lateral transfer**

Homology

- **The most fundamental concept in inferring phylogeny is homology**
- **We need to be sure the characters we are studying are homologous, i.e. ”the same” character in different organisms**
- **Otherwise our analyses will be misled**



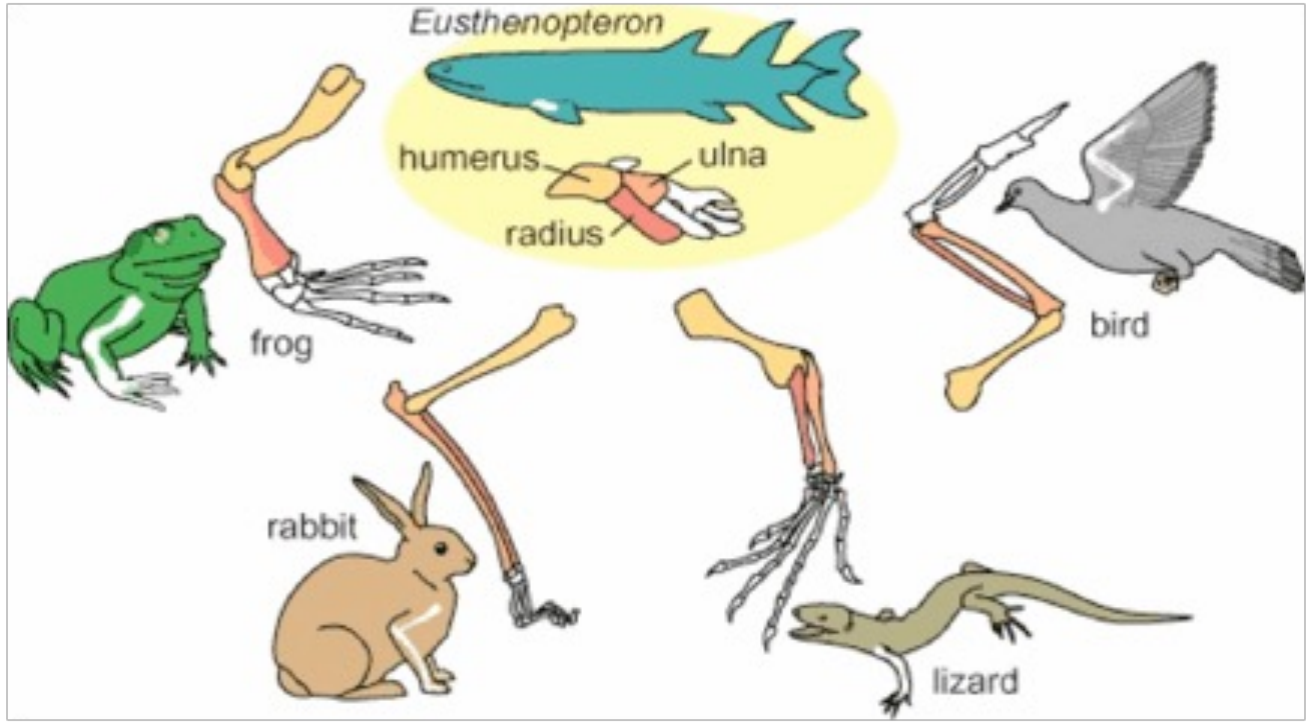
vs.



Owen's definition of homology

Homologue: the same organ under every variety of form and function (true or essential correspondence)

Richard Owen 1843



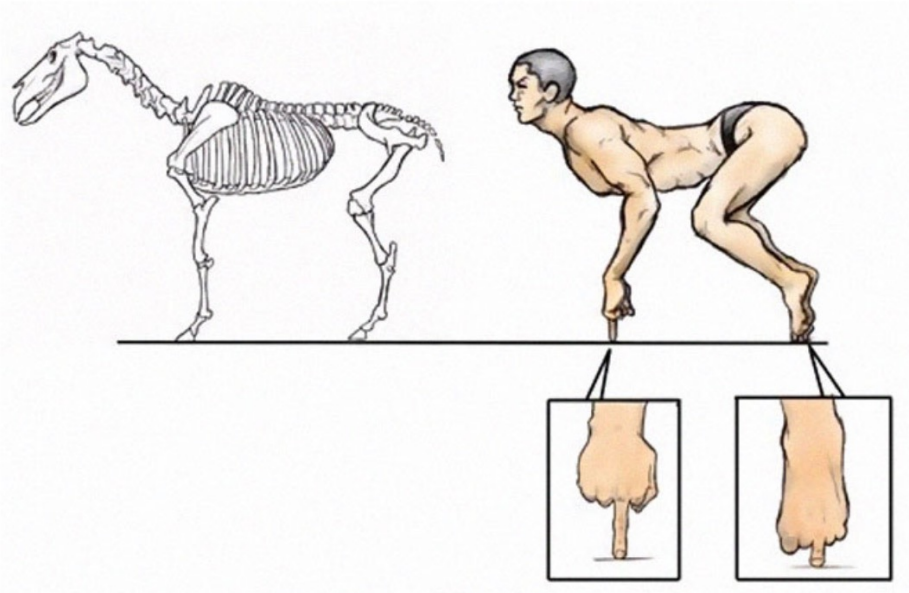


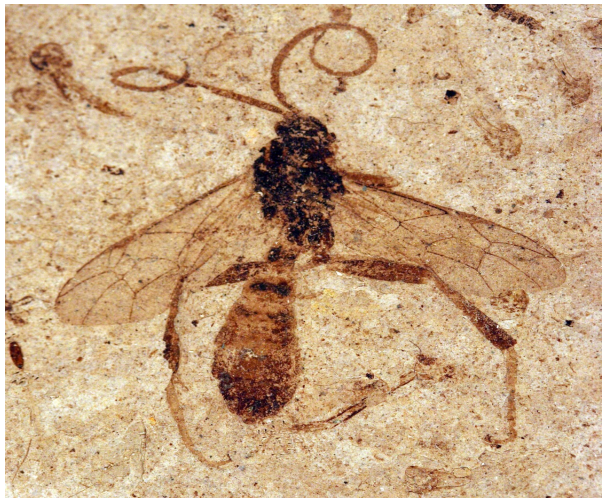
Image source: Satoshi Kawasaki

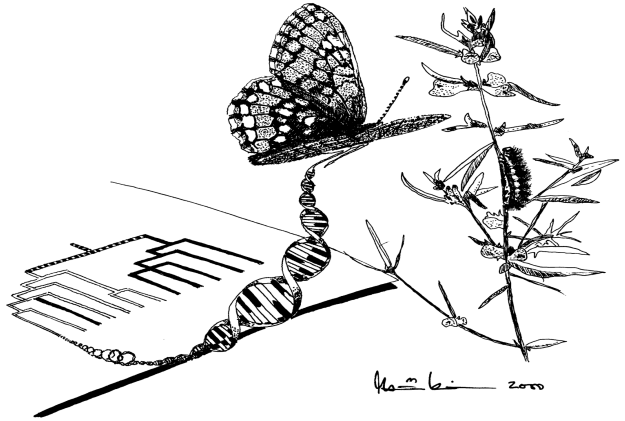
Main kinds of data in phylogenetic inference

- **Morphological**
  - **Traditionally used in phylogenetic inference**
  - **Still necessary for fossils and when molecular data are lacking**
  - **Can also help when molecular data are ambiguous**
- **Molecular**
  - **Most commonly used nowadays**
  - **Ease of sequencing led to a revolution in molecular phylogenetics**

Phylogenetic analysis is an attempt to infer the past

- **Inferring a phylogeny is an attempt to produce a best estimate of an evolutionary history based upon incomplete information**
- **Our direct information about the past is limited**
  - **Fossil record very incomplete**
  - **Access to contemporary species and molecules**

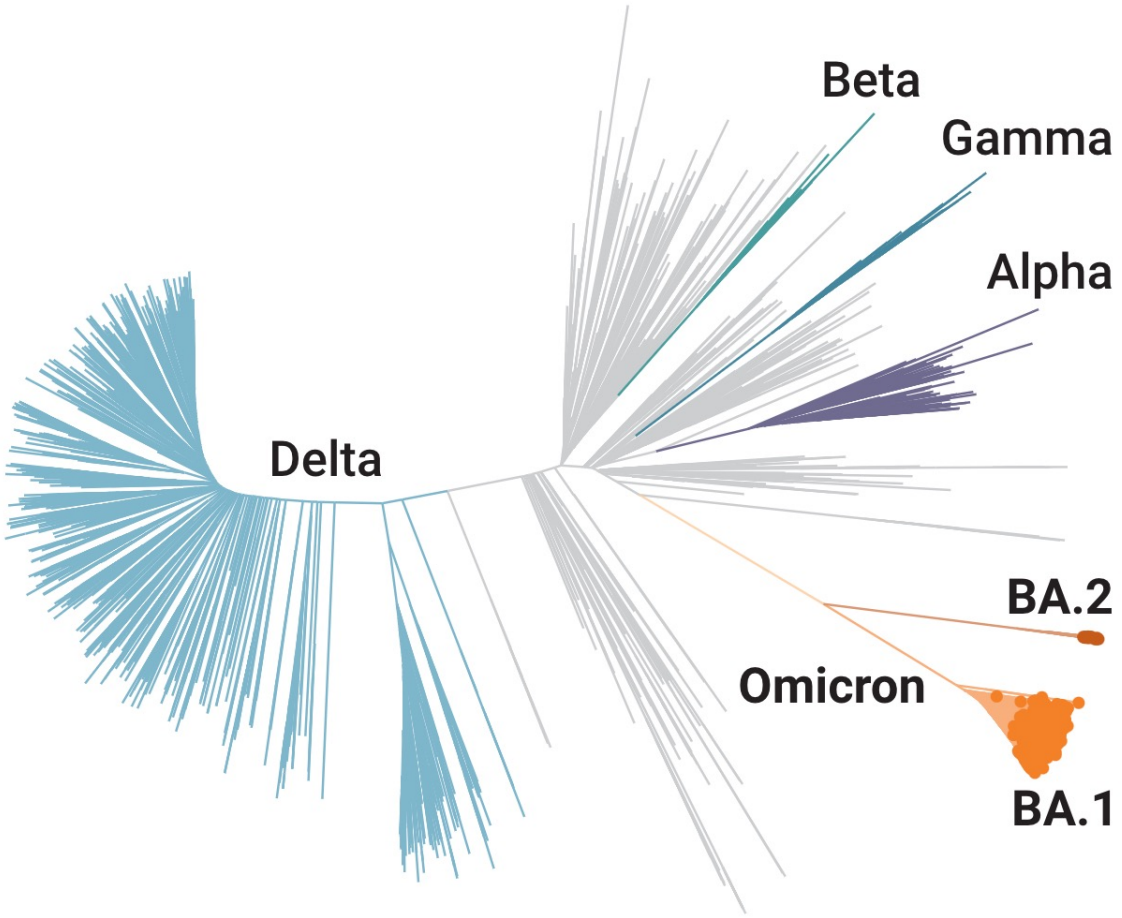




Evolutionary History and Its Importance in Biology

Why do we need phylogenies?

- **What is shown here?**

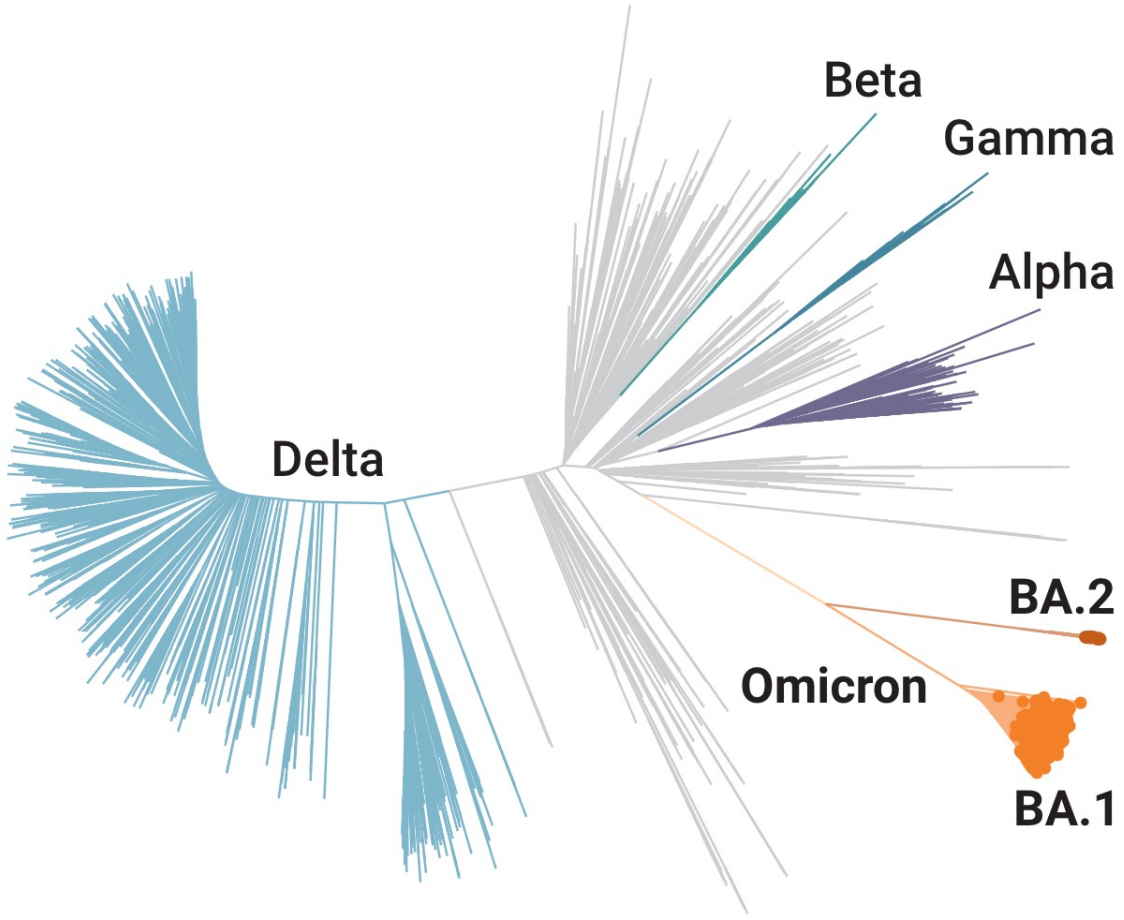


https://www.science.org/content/article/sudden-rise-more-transmissible-form-omicron-catches-scientists-surprise?utm_campaign=SciMag&utm_source=Social&utm_medium=Twitter

Why do we need phylogenies?

- **What is shown here?**

Phylogeny of covid-19 strains



https://www.science.org/content/article/sudden-rise-more-transmissible-form-omicron-catches-scientists-surprise?utm_campaign=SciMag&utm_source=Social&utm_medium=Twitter

Examples relevant to systematics and evolutionary history

- **Snakes and lizards**
- **Dinosaurs and birds**
- **Relationships among the three domains of life**

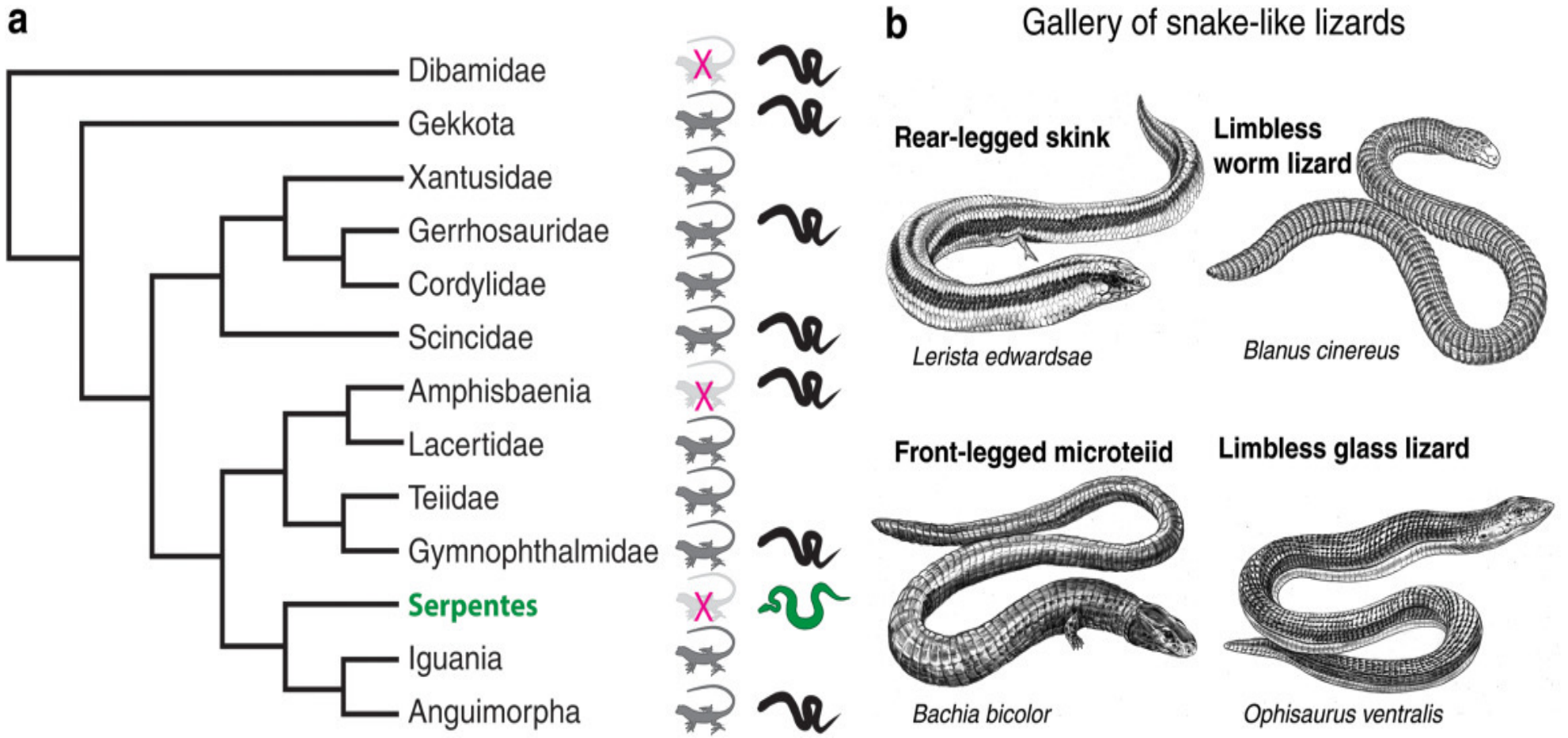
Which statement is true?

a) **Snakes are the sister group of lizards.**

b) **Some lizards are more closely related to the snakes than to other lizards.**

c) **Snakes are reptiles that are not closely related to lizards.**

Snakes are one of several lizard lineages that have lost limbs



Snake-like body plan

Leal & Cohn (2018) Developmental, genetic, and genomic insights into the evolutionary loss of limbs in snakes. Genesis, Volume: 56, Issue: 1, First published: 02 November 2017, DOI: (10.1002/dvg.23077)

Which statement is true?

a) **Snakes are the sister group of lizards.**

b) **Some lizards are more closely related to the snakes than to other lizards.**

c) **Snakes are reptiles that are not closely related to lizards.**

Dinosaurs…

a) … belong to reptiles but are not closely related to birds.

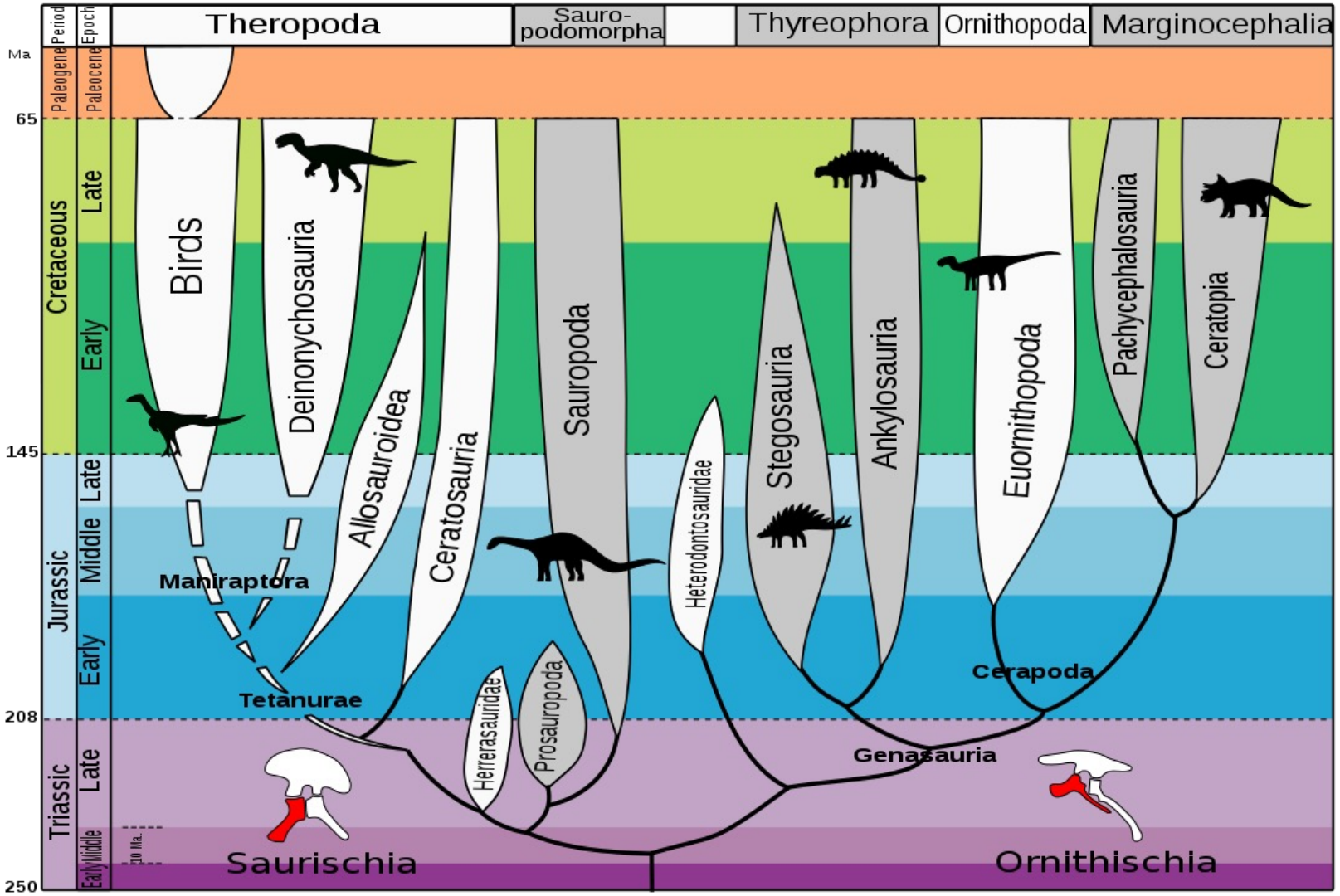
b) … are the sister group of birds.

c) … are paraphyletic with respect to birds.

Image source: Wikipedia

Are dinosaurs extinct?

- **Avian dinosaurs**
- **Non-avian dinosaurs**
- **Dinosaurs are paraphyletic with respect to birds**
- **Pseudoextinction**



Dinosaurs…

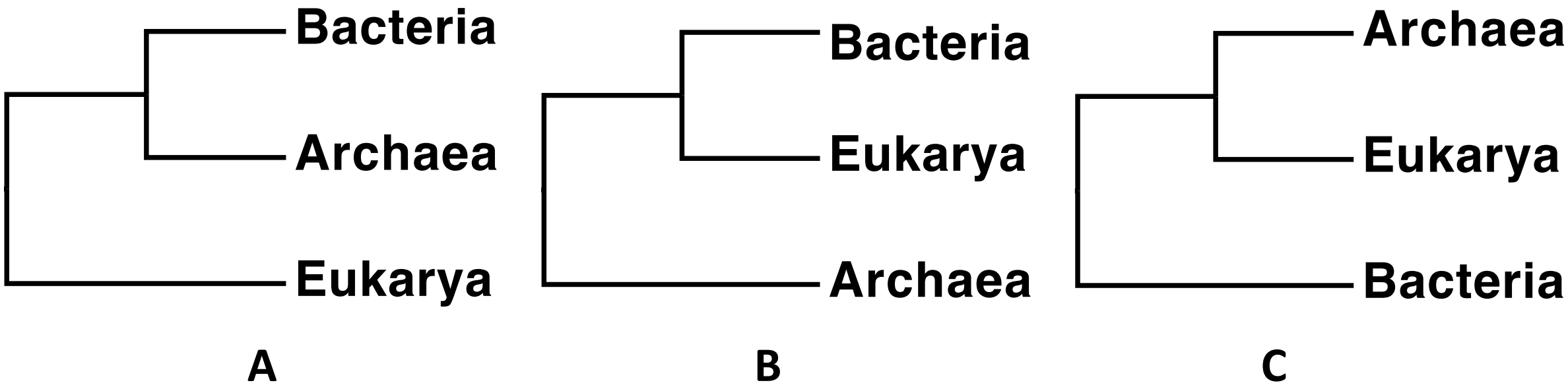
a) … belong to reptiles but are not closely related to birds.

b) … are the sister group of birds.

c) … are paraphyletic with respect to birds.

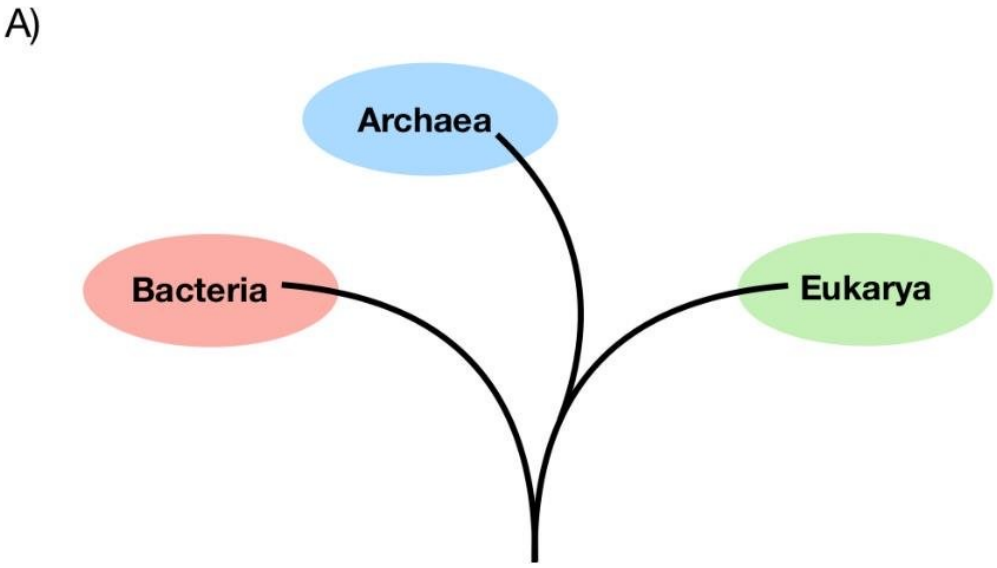
Image source: Wikipedia

Best hypothesis for the relationships among the three domains of life is:

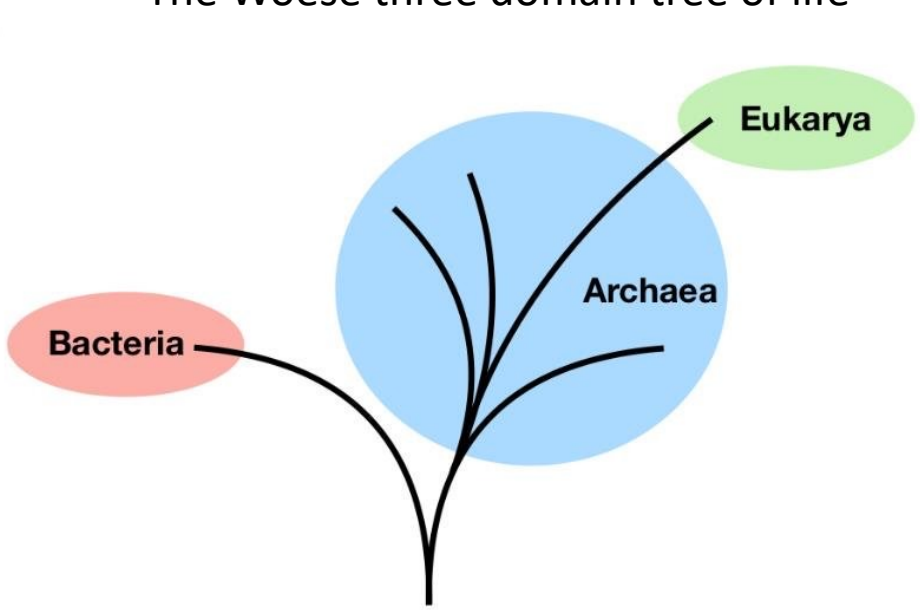


Origin of Eukarya

- **Eukaryotes**
  - **Organisms whose cells have a nucleus enclosed within a nuclear envelope**
  - **Plants, animals, fungi + a lot of single-cell lineages**
- **Which lineage is the sister lineage to Eukarya?**
- **Why is this interesting?**
  - **Where does the complex cell of eukaryotes originate from?**



The Woese three domain tree of life



The Eocyte hypothesis

Fig. 4 from MacLeod et al. 2019 (doi: 10.3934/microbiol.2019.1.48)

Origin of Eukarya

- **Which lineage is the sister lineage to Eukarya?**
- **Discovery of Asgard Archaea is complicating the picture**

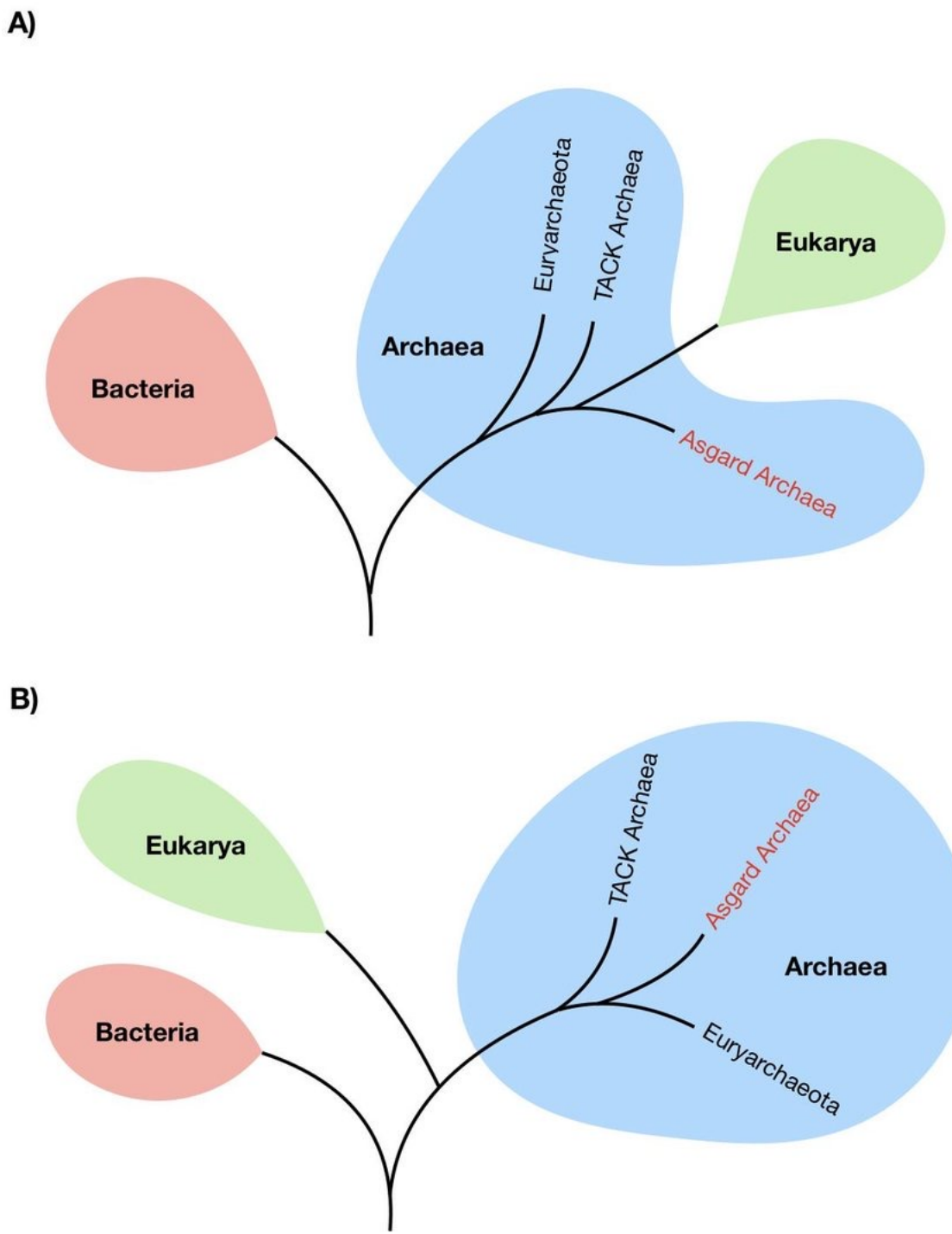
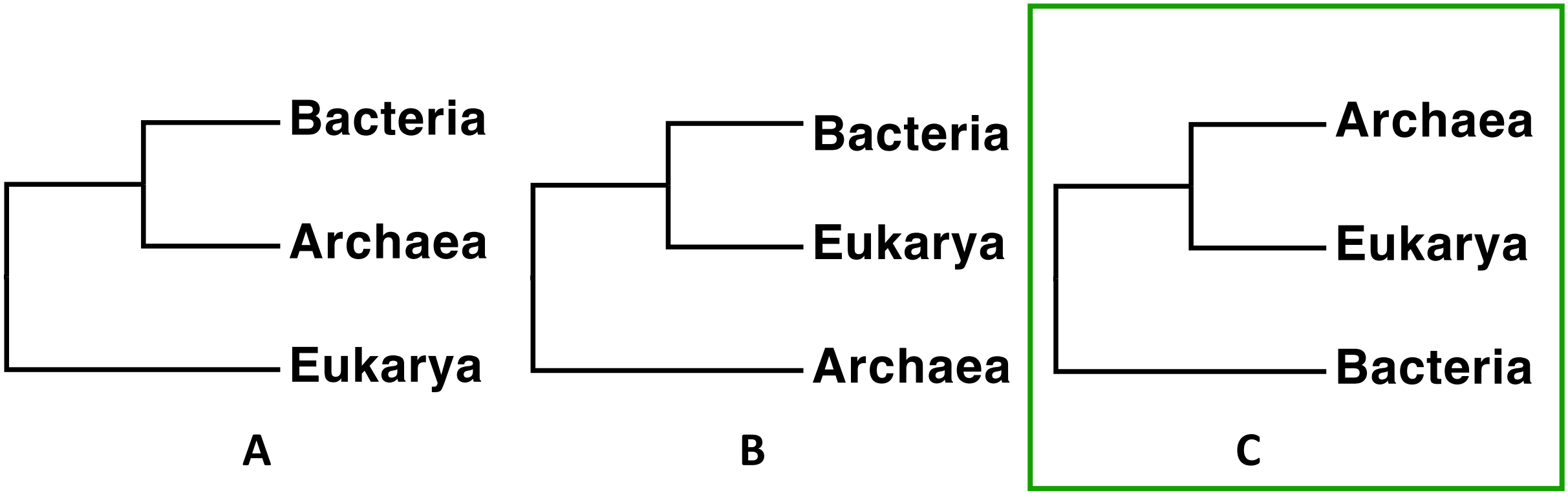


Figure 5. (A) The tree of life resulting from a range of phylogenetic analyses of conserved markers, ribosomal RNA genes and ribosomal proteins, placing Asgard archaea as the closest archaeal relatives of eukaryotes. (B) The tree produced from phylogenetic analyses of Asgard RNA polymerase genes supports the three-domain topology of the tree of life, with Asgard archaea as a sister group to Euryarchaeota.

Fig. 5. MacLeod et al. 2019 (doi: 10.3934/microbiol.2019.1.48)

Best hypothesis for the relationships among the three domains of life is:



Molecular evolution:

arms race between caterpillars and their host plants

- **Plants have secondary metabolites**
- **One of the functions: protection from herbivory**
- **Insects can evolve the ability to metabolize plant secondary metabolites**
- **Then plants can evolve new chemicals, and we have an arms race**
- **Example: Edger et al. 2015 PNAS study**

The butterfly plant arms-race escalated by gene and genome duplications

Patrick P. Edger[a,b,c,1], Hanna M. Heidel-Fischer[d,1], Michaël Bekaert[e], Jadranka Rota[f], Gernot Glöckner[g,h], Adrian E. Platts[i], David G. Heckel[d], Joshua P. Der[j,k], Eric K. Wafula[j], Michelle Tang[a], Johannes A. Hofberger[l], Ann Smithson[m,n], Jocelyn C. Hall[o], Matthieu Blanchette[i], Thomas E. Bureau[p], Stephen I. Wright[q], Claude W. dePamphilis[j], M. Eric Schranz[l], Michael S. Barker[b], Gavin C. Conant[r,s], Niklas Wahlberg[f], Heiko Vogel[d], J. Chris Pires[a,s,2], and Christopher W. Wheat[t,2]

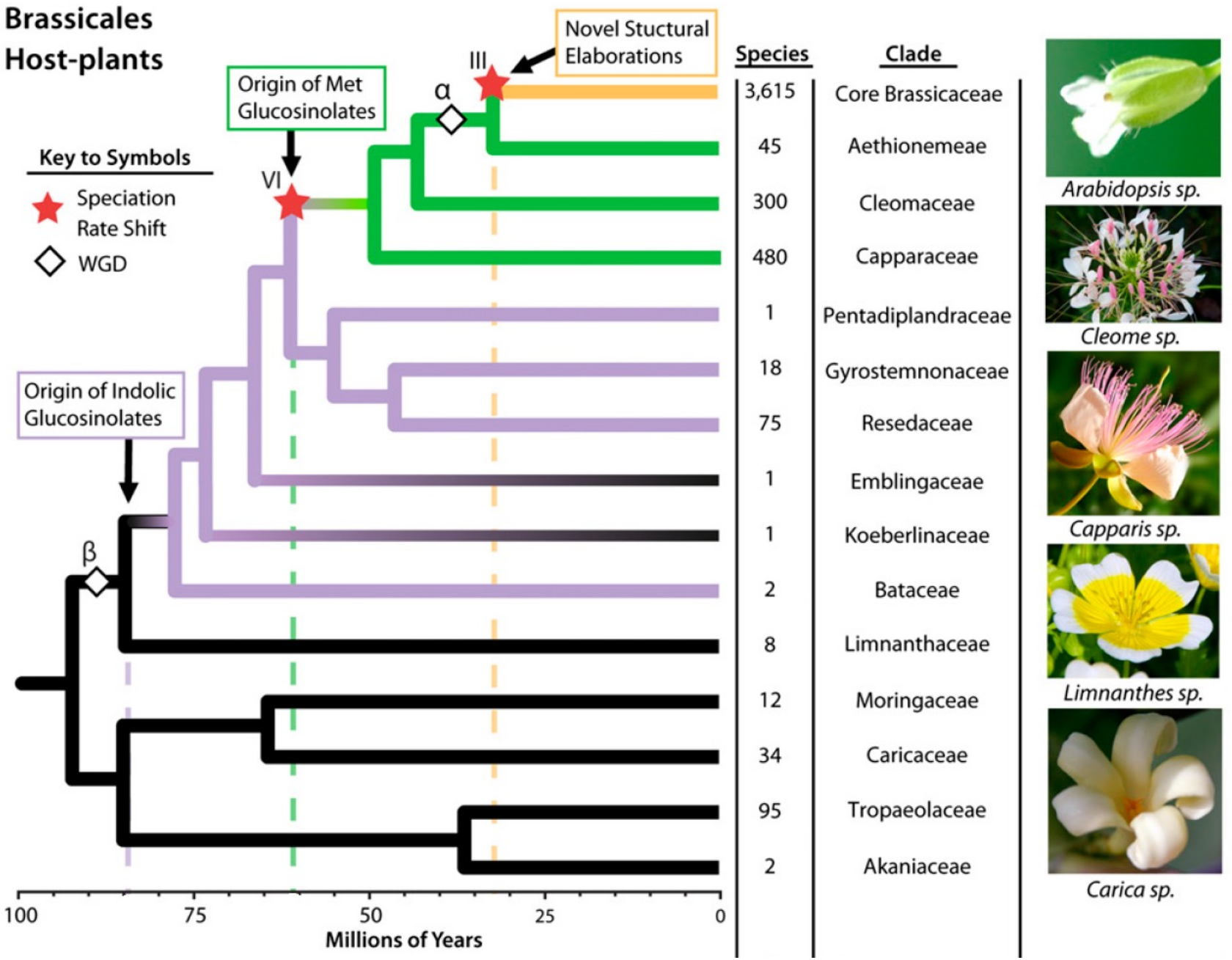
[a]Division of Biological Sciences, University of Missouri, Columbia, MO 65211; [b]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721; [c]Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720; [d]Department of Entomology, Max Planck

Edger et al. 2015. The butterfly plant arms-race escalated by gene and genome duplications

- **Butterflies: family Pieridae**
- **Plants: order Brassicales**
- **A number of gradual changes**
- **BUT also gene and genome duplications**
- **The story: “Nearly 90 million years ago, the ancestors of *Brassica* (mustards, cabbage) and related plants developed a chemical defense called glucosinolates.”**
  - **Glucosinolates – toxic to most insects, for humans source of the sharp taste in wasabi, horseradish, and mustard**

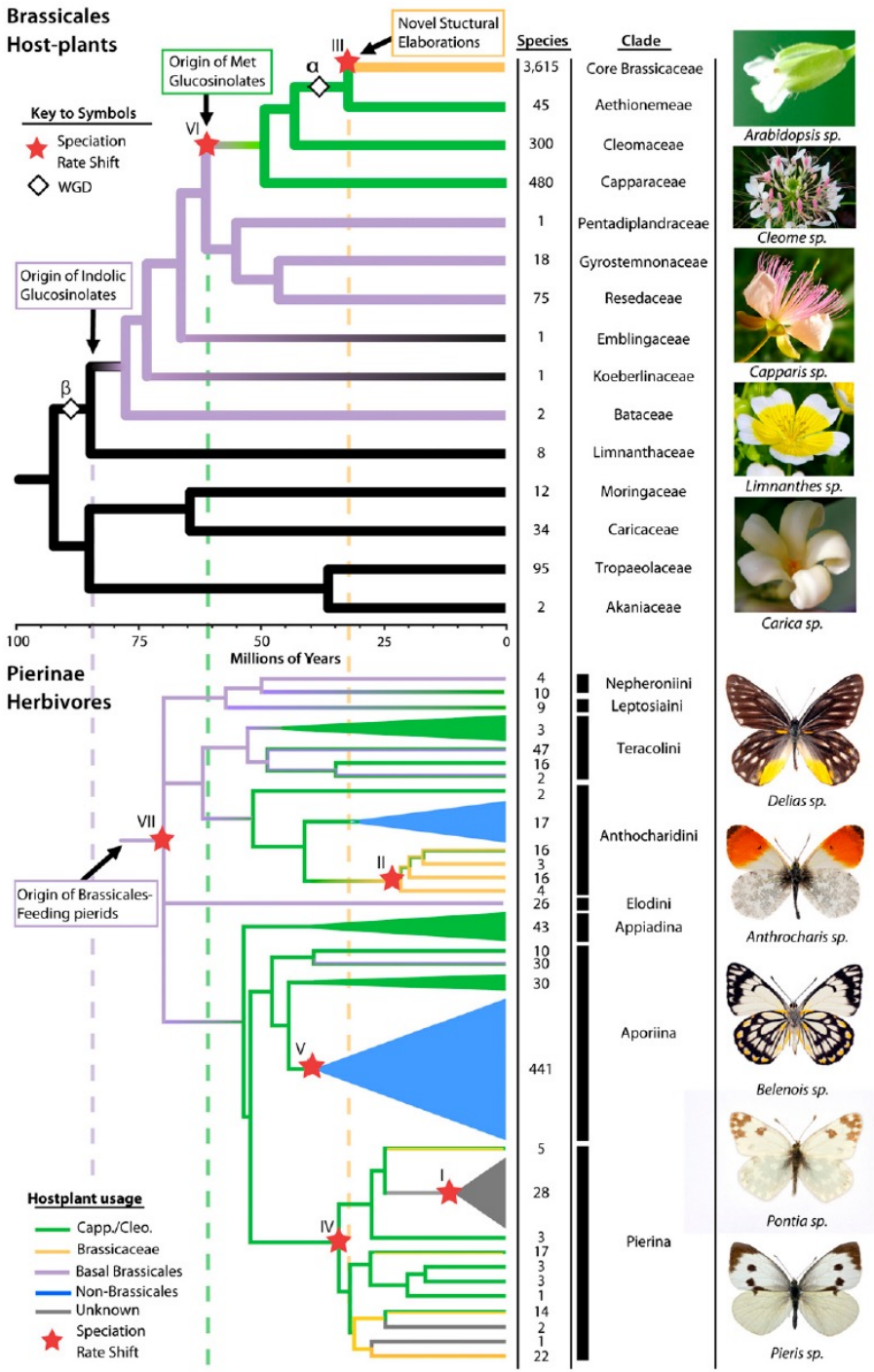
Cabbage Butterfly *Pieris brassicae* (Linnaeus, 1758)

WGD
= whole genome duplications



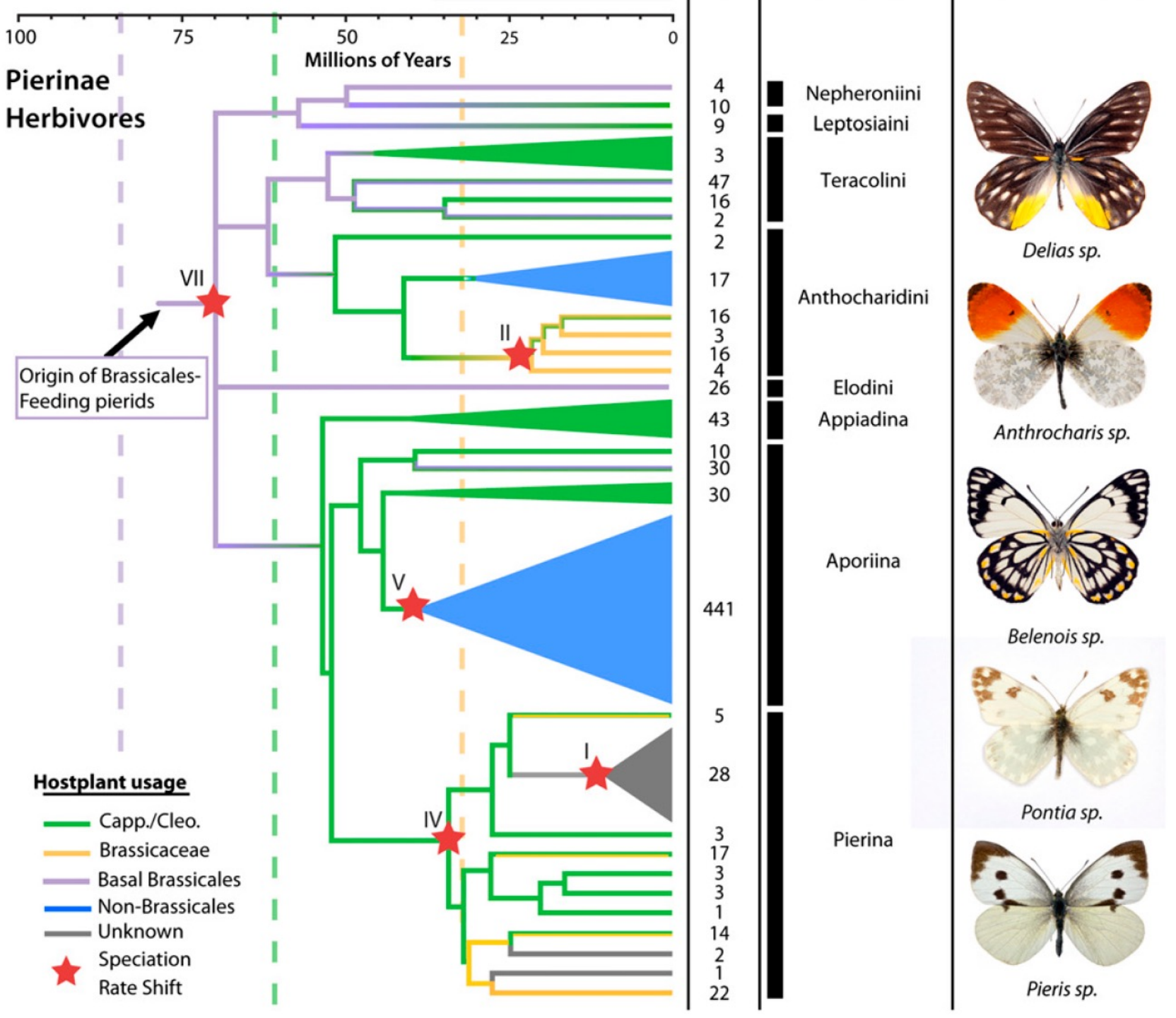
Edger et al. 2015 Fig. 1

- **Timing of divergence for plants and butterflies carried out on independent data**
- **Butterflies colonized the host plants after they had evolved their chemical defences**



Edger et al. 2015 Fig. 1

- **It appears that colonizing chemically protected host plants allowed the butterflies to speciate faster**



Edger et al. 2015 Fig. 1

Figure 2. (B) An illustration of the evolution of core glucosinolate pathways across Brassicales; with substrates tryptophan (Trp), phenylalanine (Phe), and methionine (Met) shown at the top, enzymes depicted as white ovals, and each pathway as black vertical lines.

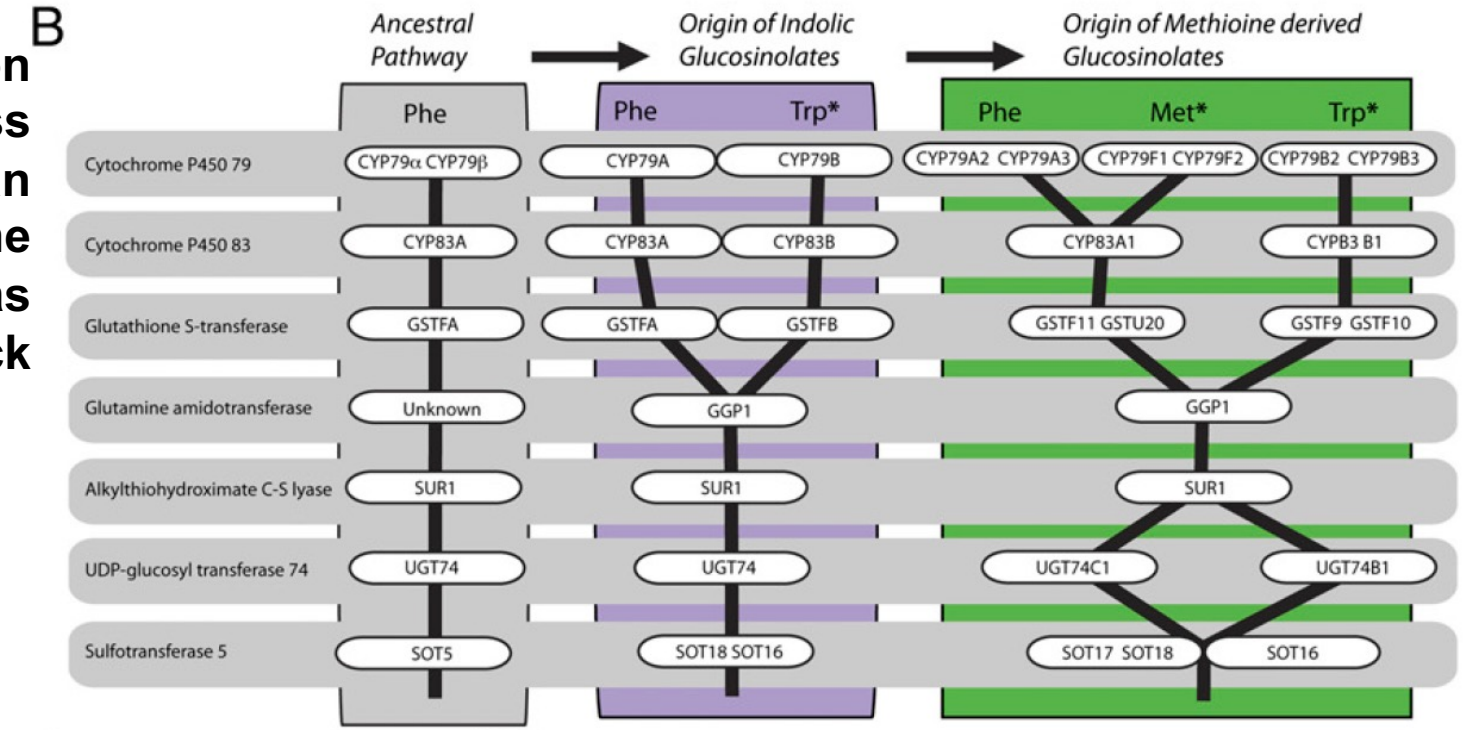
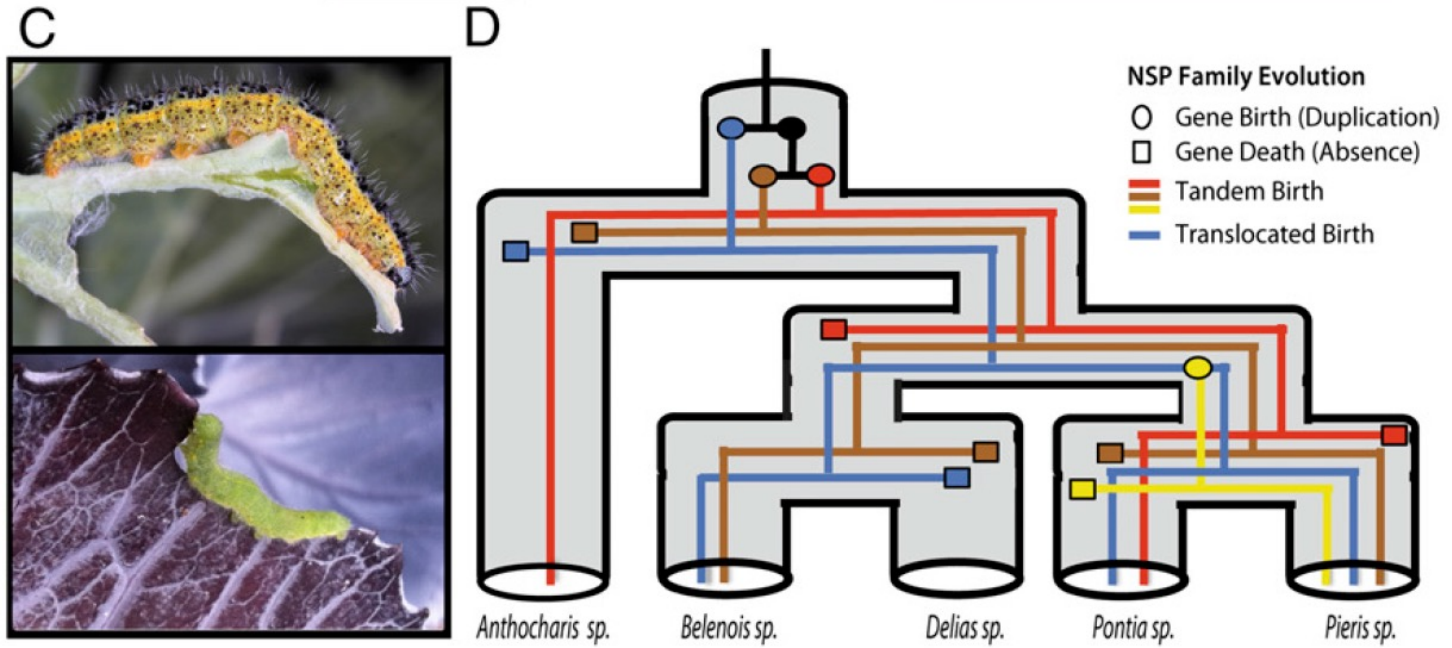


Figure 2. (D) Evolution of the NSP gene family is shown across select Pierinae genera, indicating the birth and death dynamics of four paralogous clades.



***Nitrile Specificer Protein (NSP)**: primary detoxification mechanism used by the butterflies to break down this chemical defense system

Criminal investigations:
evolutionary history of a deliberate HIV infection

- **Source: Herron & Freeman "Evolutionary Analysis" 5th edition**
- **Players in this true story:**
  - **Janet Trahan, nurse – the accuser**
  - **Richard Schmidt, physician – the accused**
  - **David Hillis, evolutionary biologist – providing crucial evidence**
- **Legal case:**
  - **Trahan became sick with HIV and accused Schmidt of deliberately injecting her with blood from a HIV-positive patient as revenge for breaking off their affair**
- **Analysis:**
  - **Evolutionary history of viral samples from various people to establish from where Trahan contracted her virus**

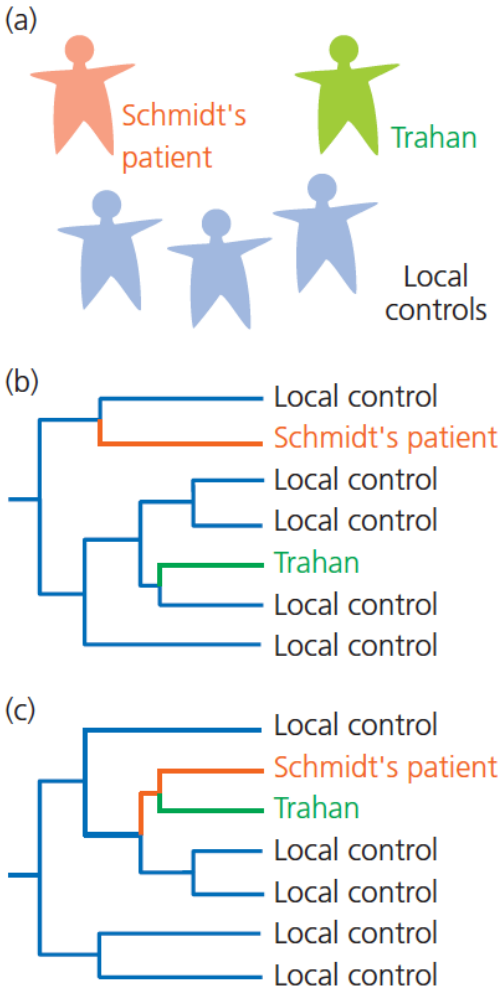


Figure 1.23 Reconstructing evolutionary history tests the prosecution's theory of a crime (a) The individuals from whom HIV samples were collected. (b) The viral evolutionary tree predicted if the suspect is innocent. (c) The evolutionary tree predicted if the suspect is guilty.

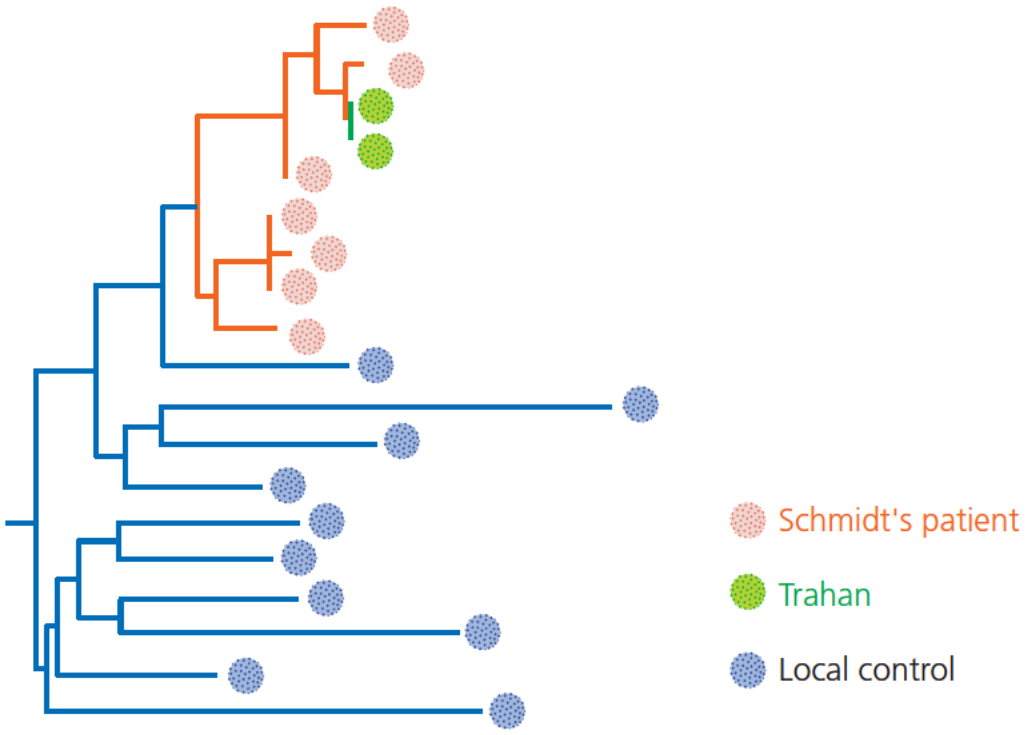


Figure 1.24 Incriminating evidence This reconstructed evolutionary history is consistent with the prosecution's charges. Redrawn from Metzker et al. (2002).

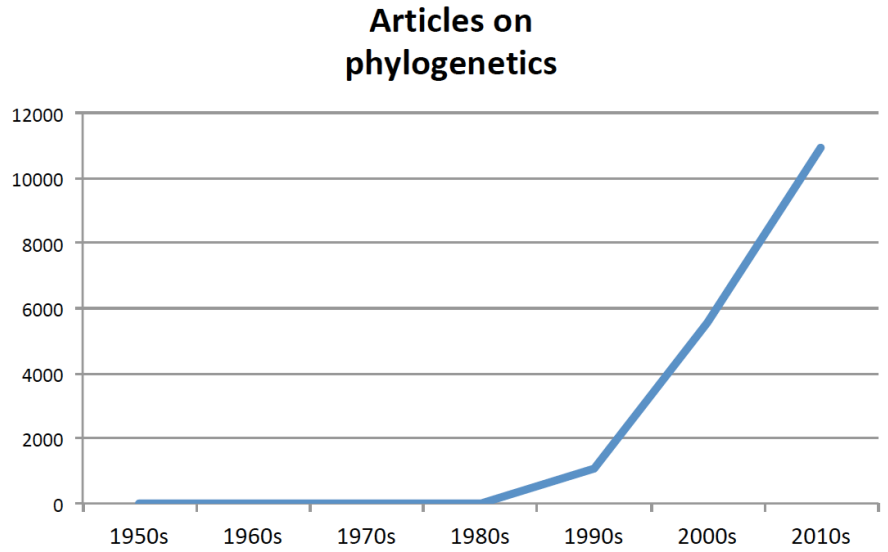
Source: Herron & Freeman "Evolutionary Analysis" 5th edition

Importance of Understanding Evolutionary History:
examples we talked about (there could be many, many more)

- **Viral evolution and its impact on human health**
- **Systematics – understanding the relationships among different groups of organisms and how they acquired their current form and function**
  - **E.g. snakes and lizards, birds and dinosaurs, the main domains of life**
- **Molecular evolution**
  - **Genome and gene duplications as weapons in an arms race between plants and caterpillars**
- **Even criminal investigations!**

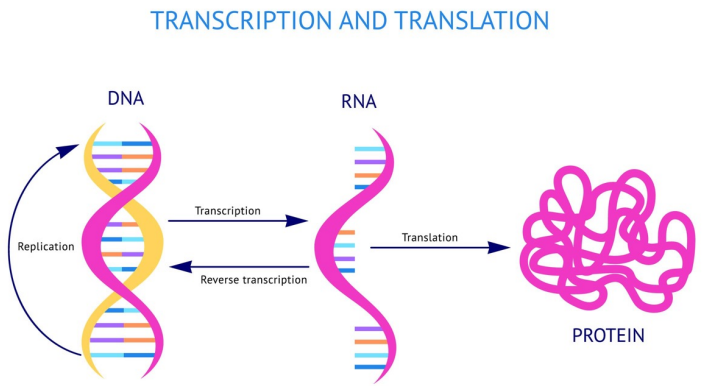
The rise of systematics

- **Within the last 25 years the number of phylogenetic studies has skyrocketed**
- **Largely due to the advent of easy DNA sequencing methods**
- **Is helping us understand biodiversity and evolutionary processes better**



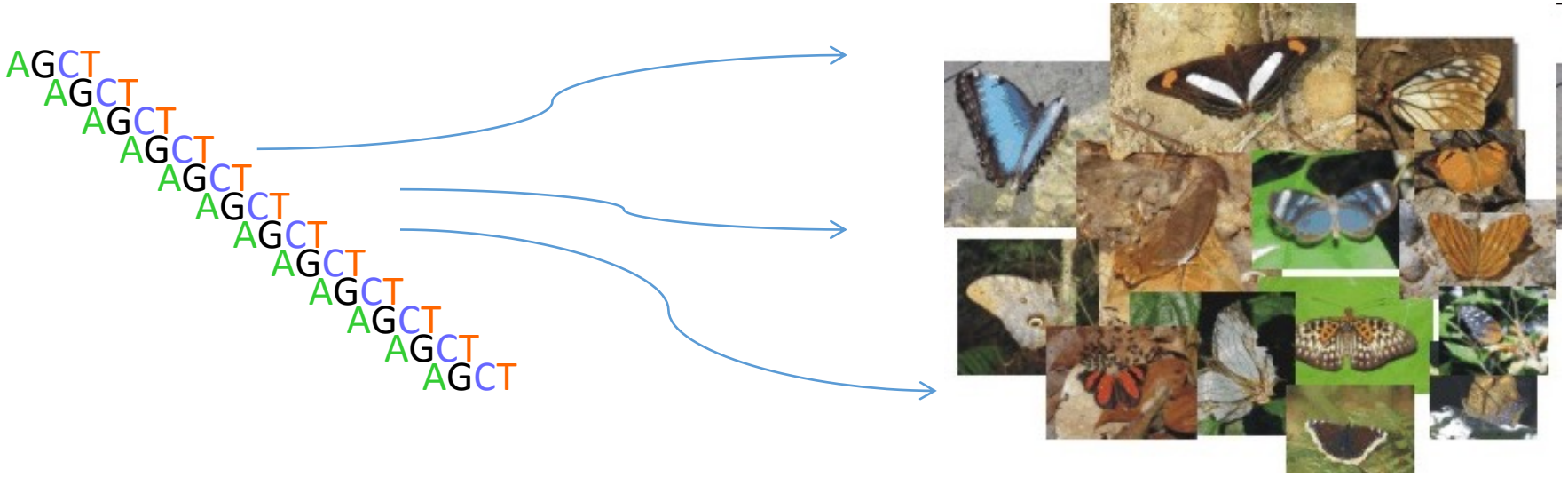
Why DNA is the Ultimate Source of Information

- **Higher levels lack information**
  - **For example, one can infer protein sequence from DNA sequence data, but not complete DNA sequence from protein sequence data**
- **Lower levels provide no additional useful information**
  - **For example, sub-atomic structure does not provide information about historical relationships**



Why *molecular* systematics?

- **Ease of data generation for large numbers of taxa**
- **Ease of generating a large number of independent data sets for given taxa**
- **Molecular characters behind the morphological characters we see**



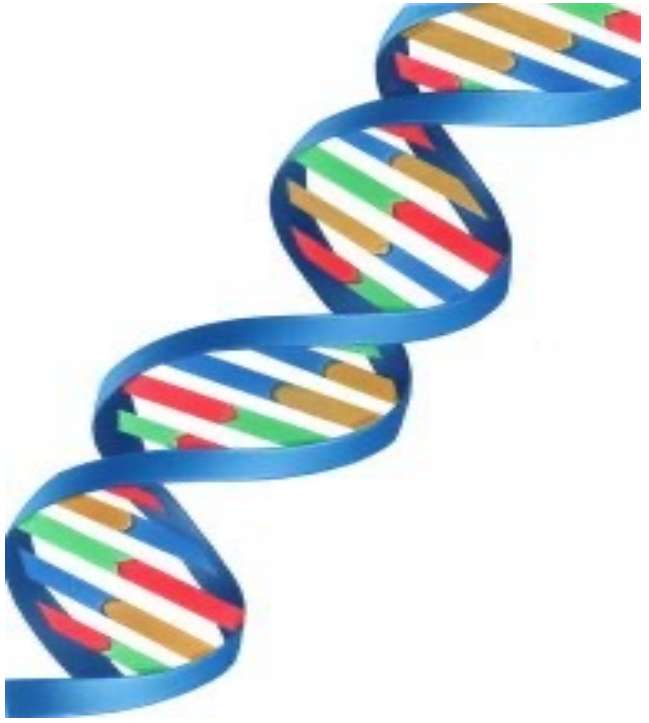
Molecular systematics as a part of understanding evolution

- **Biochemistry — basic low-level processes (e.g., nucleotide substitution, amino acid interactions)**
- **Molecular genetics — fundamental genetic processes (e.g., DNA replication, recombination)**
- **Population genetics — micro-evolutionary processes**
- **Systematics — macro-evolutionary processes**

What is “Molecular” in Molecular Systematics?

- Carbohydrates — No
- Lipids — No
- Secondary metabolites — No
- **Proteins (amino acids) — Yes**
- **Nucleic Acids (DNA and RNA) — Yes**

Proteins and DNA are sometimes described as “informational macromolecules”



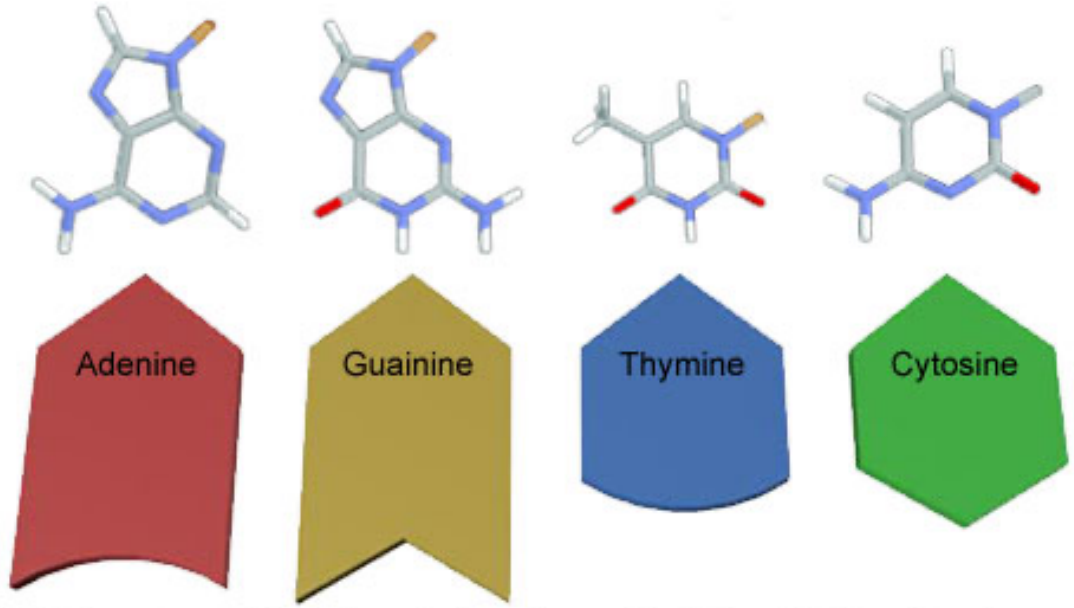
DNA as a source of information

- **DNA has four characters**

Purines

Pyrimidines

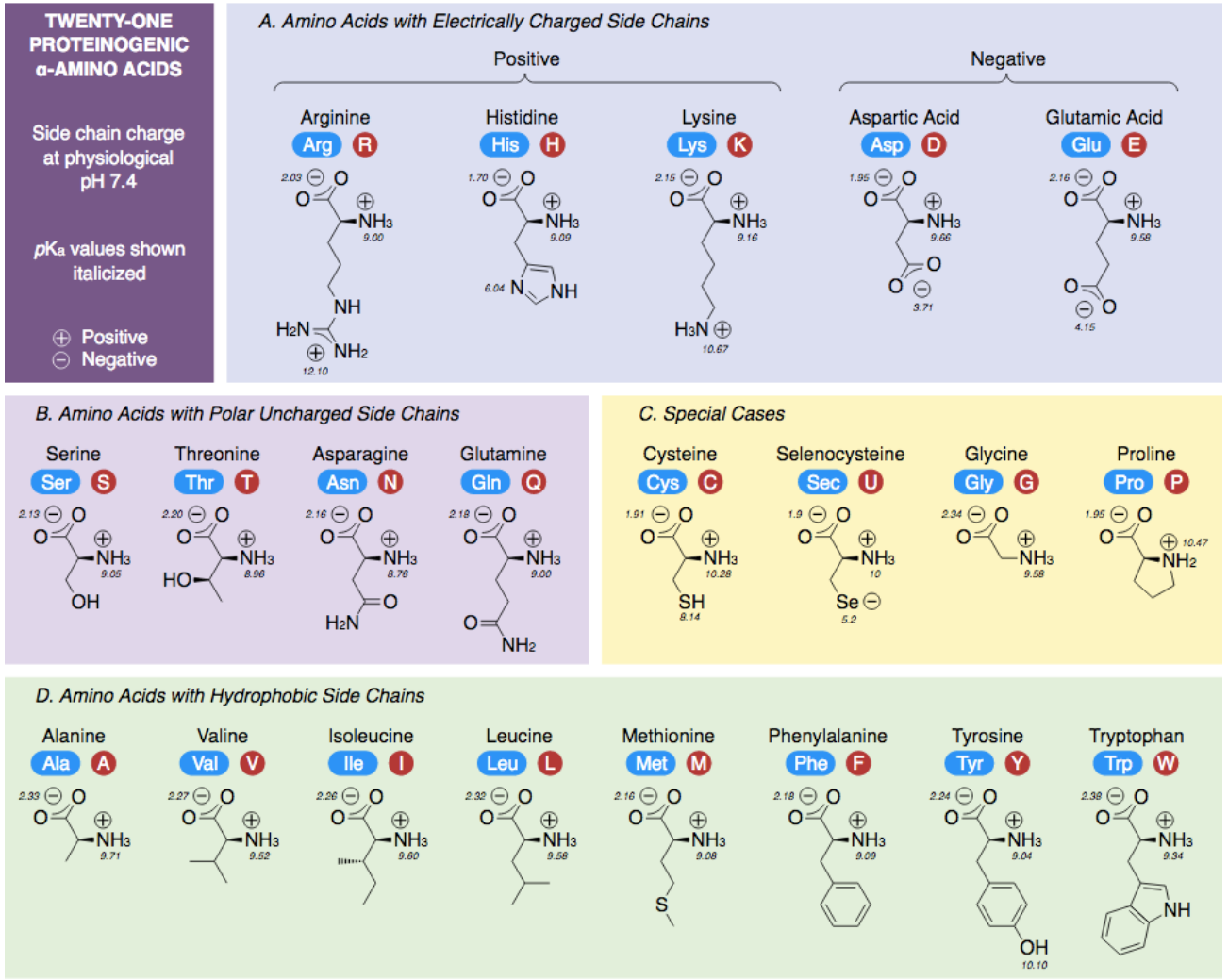
Figure B-3: The Four Nitrogenous Bases



Each base has a distinct shape that can be used to distinguish it form the others. 3D representations of the four bases are shown, with the corresponding chemical structures drawn above.

Proteins as a source of information

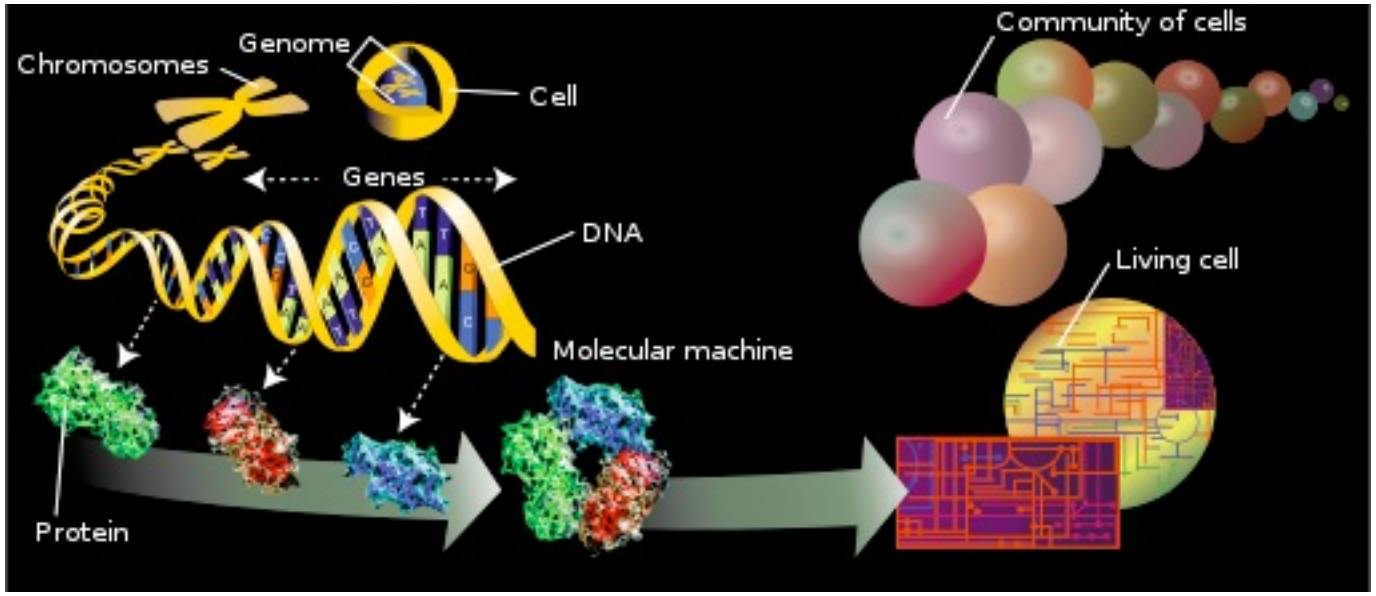
Protein sequences have 21 characters



Source: Wikipedia

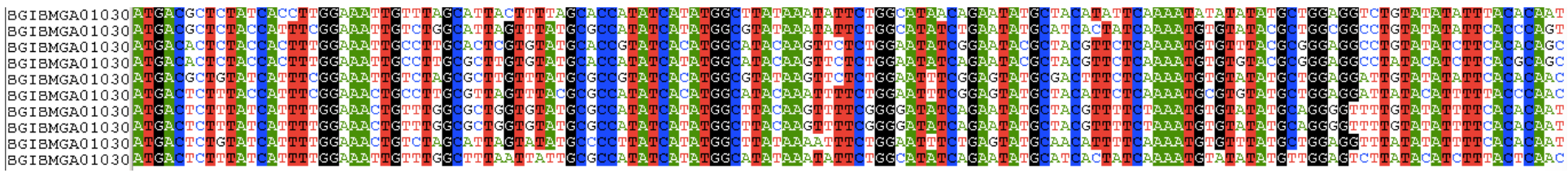
DNA found as various entities in the genome

- **Protein-coding genes**
  - **introns and exons**
- **Ribosomal DNA**
- **Repetitive elements**
- **Regulatory regions**
- **Junk DNA**
- **etc**



Homology in DNA sequences!

- **Two steps:**
- **Are we looking at the same region of the genome in our species of interest?**
  - **Orthology vs paralogy**
- **Are we looking at the same site within our chosen marker in our species of interest?**
  - **Alignment**

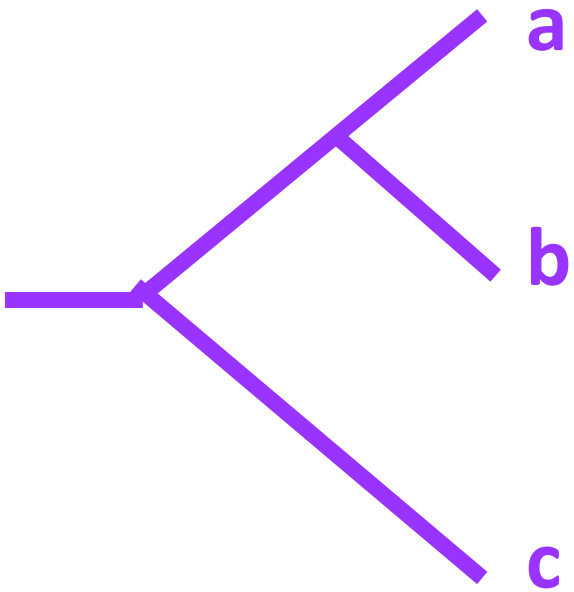


Orthology or paralogy?

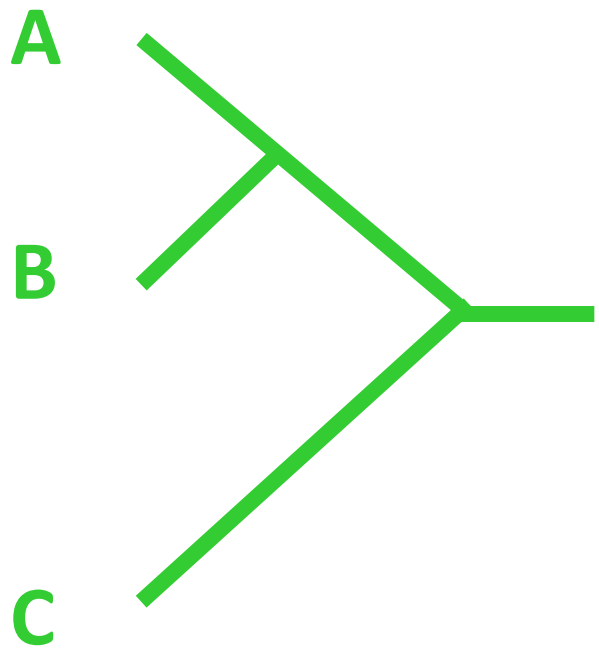
- **Are the genome regions sequenced from different species the same (homologous)?**
- **Gene duplication**
  - **1) duplicate gene degenerates - pseudogene**
  - **2) duplicate gene aquires new function**
- **A problem particularly accute currently as we analyze phylogenomic data**

Orthology:
gene trees and species trees

Gene phylogeny



Organism phylogeny

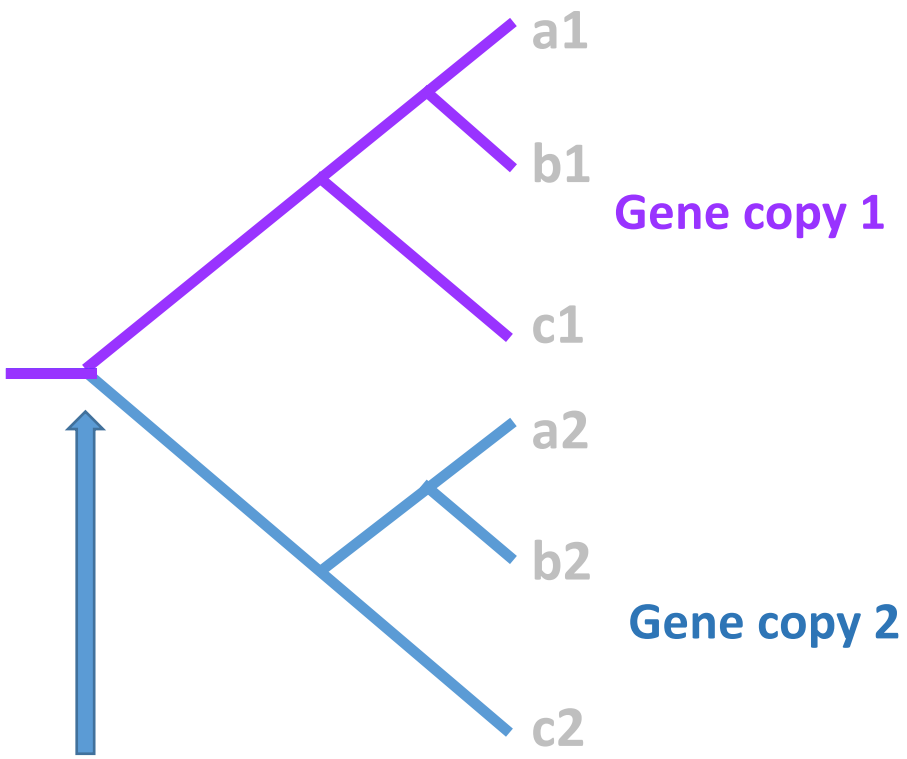


ORTHOLOGY

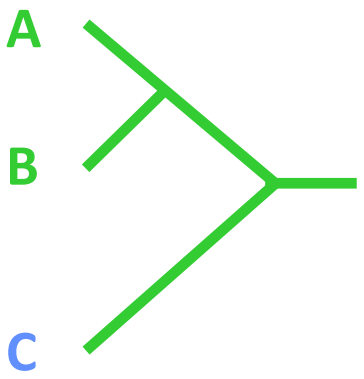
Orthologs: genes that arose due to speciation

Paralogy: can produce misleading trees

Gene phylogenies



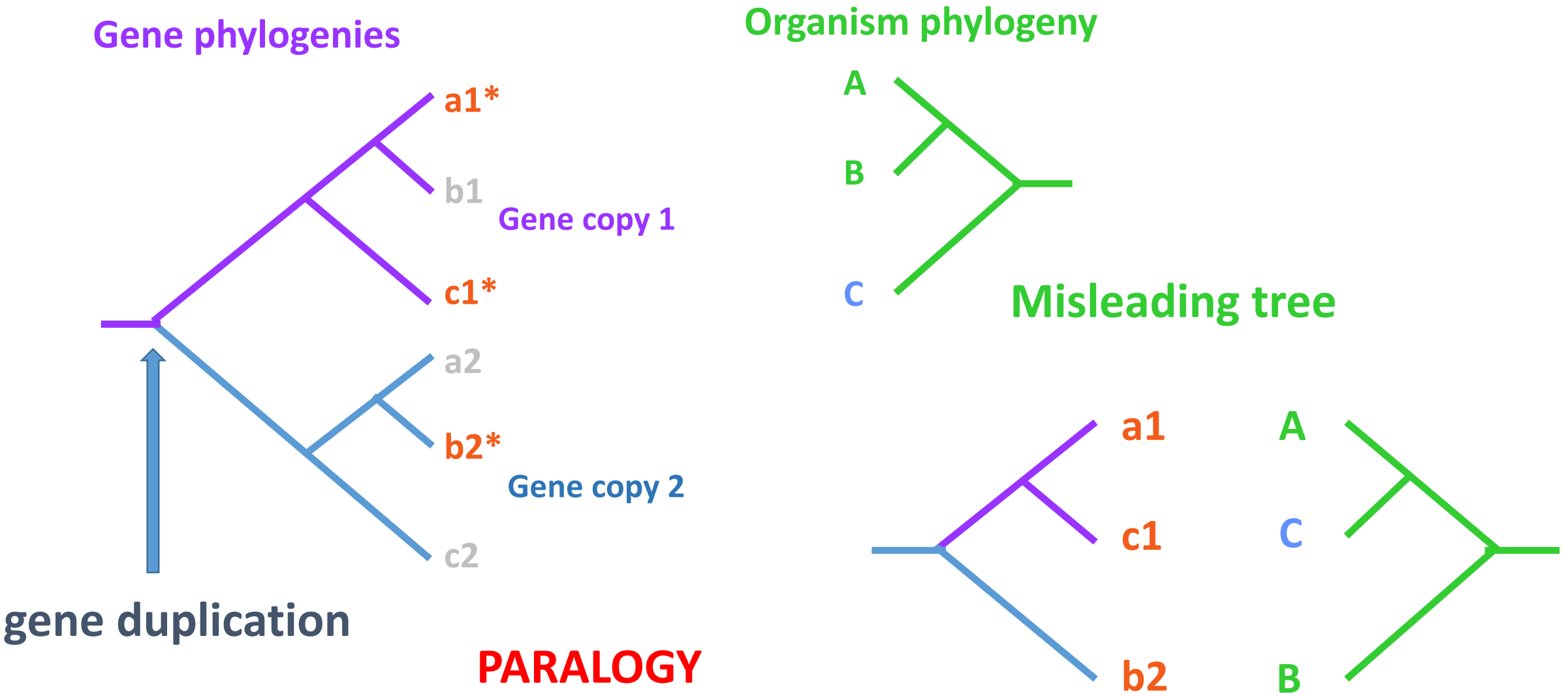
Organism phylogeny



PARALOGY

Paralogs: genes that arose due to duplication events

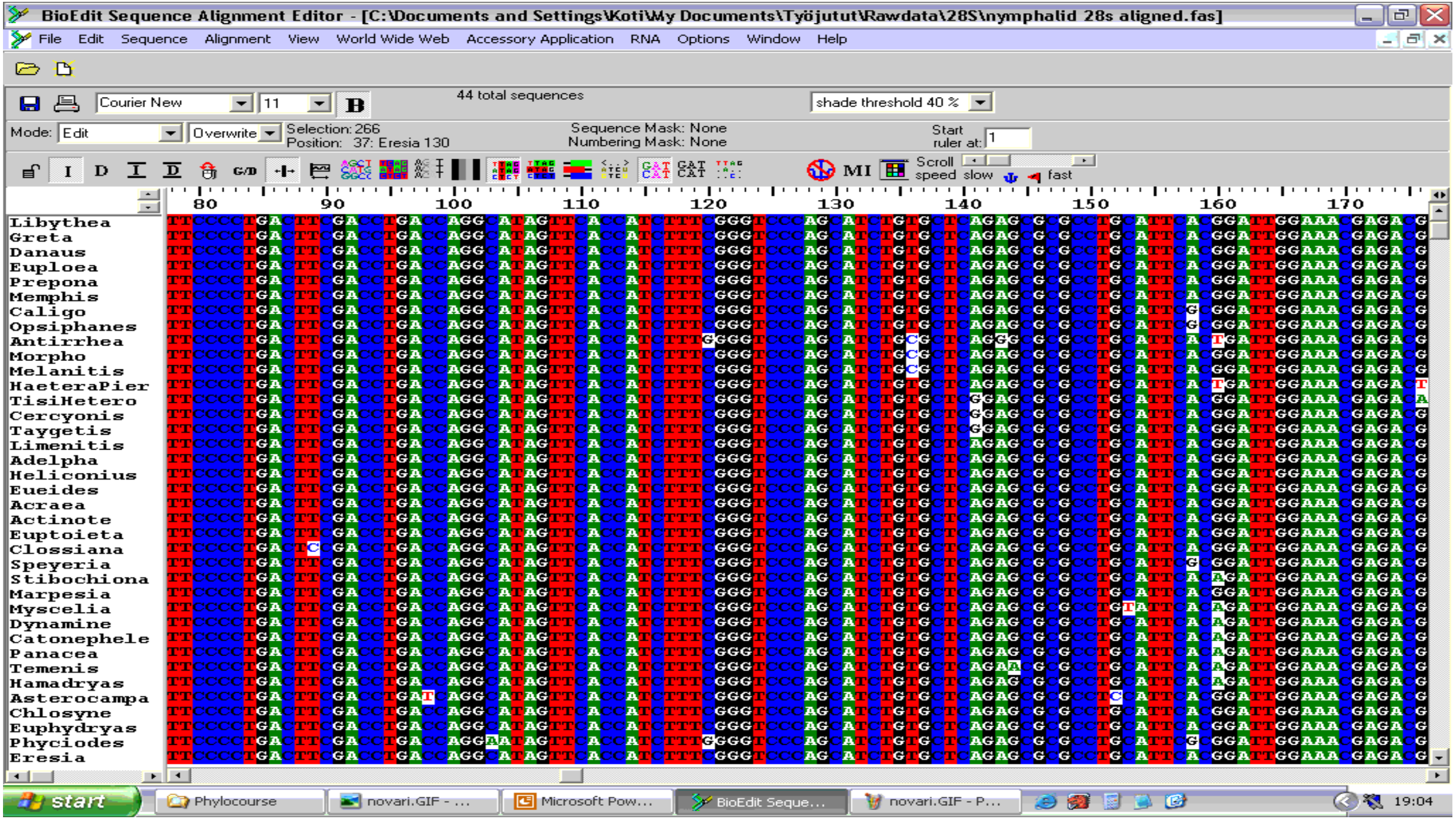
Paralogy: can produce misleading trees

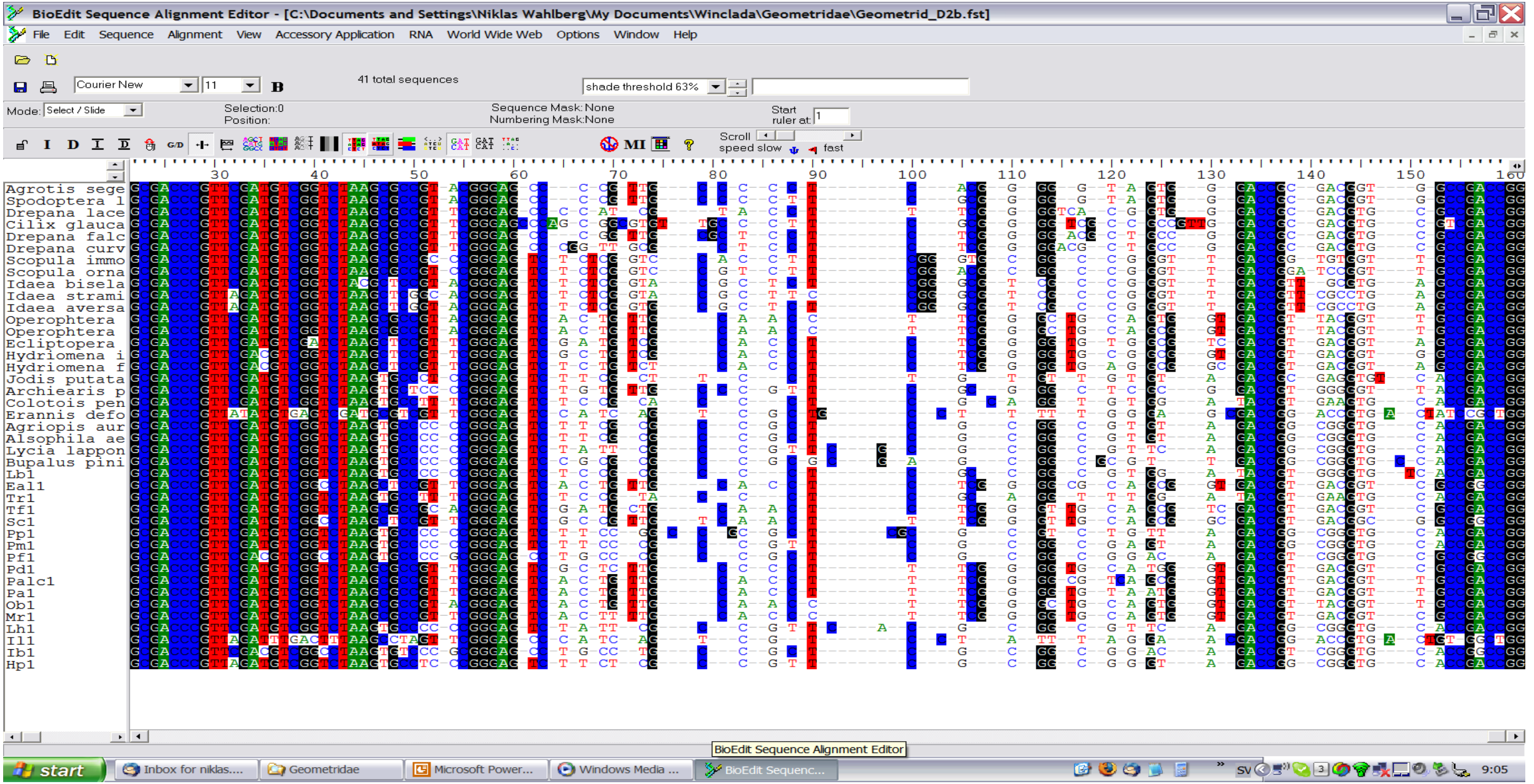


Paralogs: genes that arose due to duplication events

What sequences should we use?

- **Choice of sequence - appropriate for question (fast or slow evolving - close or distant relationships).**
- **Many sequences are a mosaic of different rates**
  - **rRNA different structural regions evolve at different rates**
  - **Proteins - synonymous (silent) rate (codon position 3) is often faster than nonsynonymous (positions 1 & 2 - changes aa) rate of change**
  - **Transitions occur more readily than transversions**


BioEdit Sequence Alignment Editor - [C:\Documents and Settings\Koti\My Documents\Työjutut\Rawdata\28S\nymphalid 28s aligned.fas]
File Edit Sequence Alignment View World Wide Web Accessory Application RNA Options Window Help
Courier New
11
44 total sequences
shade threshold 40 %
Mode: Edit
Overwrite
Selection: 266
Position: 37: Eresia 130
Sequence Mask: None
Numbering Mask: None
Start ruler at: 1
Scroll speed slow fast
80 90 100 110 120 130 140 150 160 170
Libythea
Greta
Danaus
Euploea
Prepona
Memphis
Caligo
Opsiphanes
Antirrhea
Morpho
Melanitis
HaeteraPier
TisiHetero
Cercyonis
Taygetis
Limenitis
Adelpha
Heliconius
Eueides
Acraea
Actinote
Euptoieta
Clossiana
Speyeria
Stibochiona
Marpesia
Myscelia
Dynamine
Catonephele
Panacea
Temenis
Hamadryas
Asterocampa
Chlosyne
Euphydryas
Phyciodes
Eresia
start
Phylocourse
novari.GIF - ...
Microsoft Pow...
BioEdit Seque...
novari.GIF - P...
19:04


BioEdit Sequence Alignment Editor - [C:\Documents and Settings\Niklas Wahlberg\My Documents\Winclada\Geometridae\Geometrid_D2b.fst]
File Edit Sequence Alignment View Accessory Application RNA World Wide Web Options Window Help
Courier New
11
41 total sequences
shade threshold 63%
Mode: Select / Slide
Selection:0
Position:
Sequence Mask: None
Numbering Mask:None
Start ruler at: 1
Scroll speed slow fast
Agrotis sege
Spodoptera l
Drepana lace
Cilix glauca
Drepana falc
Drepana curv
Scopula immo
Scopula orna
Idaea bisela
Idaea strami
Idaea aversa
Operophtera
Operophtera
Ecliptopera
Hydriomena i
Hydriomena f
Jodis putata
Archiearis p
Colotois pen
Erannis defo
Agriopis aur
Alsophila ae
Lycia lappon
Bupalus pini
Lb1
Eal1
Tr1
Tf1
Sc1
Pp1
Pm1
Pf1
Pd1
Palc1
Pa1
Ob1
Mr1
Lh1
Il1
Ib1
Hp1
BioEdit Sequence Alignment Editor
start
Inbox for niklas....
Geometridae
Microsoft Power...
Windows Media ...
BioEdit Sequenc...
SV
9:05


BioEdit Sequence Alignment Editor - [C:\Documents and Settings\Koti\My Documents\Työjutut\Rawdata\Unchecked\NymphalidaeCOI.fst]
File Edit Sequence Alignment View World Wide Web Accessory Application RNA Options Window Help
Courier New
11
55 total sequences
shade threshold 40 %
Mode: Edit
Overwrite
Selection: 0
Position: 341
Sequence Mask: None
Numbering Mask: None
Start ruler at: 1
MI
Scroll speed slow fast
310 320 330 340 350 360 370 380 390 400
Libythea71 1
Actinote90 1
Adelpha107 1
Aglais63 3
Agraulis 24
Amnosia101 1
Anartia36 3
Antanartia65
Anthanassa12
Antirrhea109
Araschnia39
Archaeoprepo
Asteroca82 1
Caligo70 10
Calinaga64 3
Castilia76 2
Catacropte88
Catonephele6
Cercyonis8 1
Chersonesia1
Chlosyne62 1
Clossiana76
Colobura68 1
Cyr thyodama
Danaus108 21
Dichorragia1
Doleschallia
Doxocopa lau
Dynamine115
Eresia92 5
Eueides proc
Euphydryas13
Euploea70 8
Euptoieta94
Gnathotric89
Greta70 9
Hamadryas62
start
Poy
Molecular methods to...
BioEdit Sequence Alig...
20:32

Data

- **For long, the field of systematics was restricted by the amount of data**
- **10 years ago, datasets comprising 3-5 genes were the norm**
- **5 years ago the genomic revolution swung into full effect**
- **We are now faced with an abundance of potential data, but what can we do with it?**

The Era of Phylogenomics

- **Genomes can now be sequenced relatively easily**
- **Whole genomes contain a lot of information that is irrelevant for systematics, especially at deep levels**
- **The field of systematics is still trying to figure out how best to utilize genomic level data**
  - **What parts of the genome should be used?**
  - **How can we get at those parts in the most efficient way?**
  - **Where can we access specimens for our chosen methods?**

Phylogenomic data

- **RADseq**
  - **Restriction-site Associated DNA sequencing**
- **Transcriptomes**
  - **All genes that are being expressed in a certain tissue at a certain time**
- **Ultra-Conserved Elements**
  - **Probes to pull out UCEs and flanking regions**
- **Anchored Hybrid Enrichment**
  - **Probes for e.g. exons of protein coding sequences**
- **Whole Genome Sequencing**

The Data

File formats

Computer programs

- **Multitude of programs available for free!**
- **Most have their own input format**
- **Many are ”black box” programs**
- **Input files are always simple text files!!!**

No good online resource available

http://evolution.gs.washington.edu/phylip/software.html

was an attempt but not updated for a long time

Computer programs - ML

- **IQ-TREE (recommended)**
- **RAxML (recommended)**
- **PHYML**
- **GARLI**

Computer programs- Bayesian inference

- **MrBayes (recommended)**
- **BEAST (recommended)**
- **BAMBE**
- **BayesPhylogenies**

Viewing trees

- **FigTree (recommended)**
- **TreeView**
- **Winclada**
- **Dendroscope (for large trees >200 taxa)**
- **ITOL (Interactive Tree of Life) - https://itol.embl.de/ - an online resource that has a lot of different options**

Three most common data formats

- **FASTA**
- **Phylip**
- **Nexus**

Input format - FASTA

```
>Papilio_glaucus_69_3
GAGaTGGAaGACAaGGTTTCGTCGACCCTGTCCGGCCTCGAGGGCGAACT
>Hamearis84_13
GGaATGGAaGAGAAaGTCTCCACAACCCTCTCCGGACTCGAAGGTGAGCT
>Danaus_plexippus108_21
GAGAtGGAGGAGAaGGTCTCCTCCACCCTCTCAGGTCTCGAAGGTGAACT
>Greta_oto70_9
GGAATGGAaGAGAaGGTCTCCTCGACCCTCTCAGGCCTTGAAGGTGAACT
>Amathusia_phidippus114_17
GGaATGGAaGACAAaGTCTCCTCAaCCCTCTCCGGTCTTGAGGGTGAACT
>Morpho_peleides66_5
GGaATGGAGAGAAAaGTCTCTACTACCCTGTCTGGCCTCGAAGGCGAACT
>BrintesiaB01
GGAATGGAaGACAAaGTCTCGTCCACCCTCTCCGGGCTGGAAGGCGAGCT
>Elymnias_casiphone121_20
GAGAwGGaAGAcaAAGTATCCTCCACCCTCTCTGGTCTTGAAGCTGAACT
>Erebia_oemeEW24_7
gGaATGGAaGACAAaGTCTCCTCGACTCTCTCTGGCCTCGAAGGCGAGCT
```

Input format – PHYLIP

```
9 50
Papilio_gl GAGaTGGAaGACAaGGTTTCGTCGACCCTGTCCGGCCTCGAGGGCGAACT
Hamearis84 GGaATGGAaGAGAAaGTCTCCACAACCCTCTCCGGACTCGAAGGTGAGCT
Danaus_ple GAGAtGGAGGAGAaGGTCTCCTCCACCCTCTCAGGTCTCGAAGGTGAACT
Greta_oto7 GGAATGGAaGAGAaGGTCTCCTCGACCCTCTCAGGCCTTGAAGGTGAACT
Amathusia_ GGaATGGAaGACAAaGTCTCCTCAaCCCTCTCCGGTCTTGAGGGTGAACT
Morpho_pel GGaATGGAGAGAAAaGTCTCTACTACCCTGTCTGGCCTCGAAGGCGAACT
BrintesiaB GGAATGGAaGACAAaGTCTCGTCCACCCTCTCCGGGCTGGAAGGCGAGCT
Elymnias_c GAGAwGGaAGAcaAAGTATCCTCCACCCTCTCTGGTCTTGAAGCTGAACT
Erebia_oem gGaATGGAaGACAAaGTCTCCTCGACTCTCTCTGGCCTCGAAGGCGAGCT
```

Input format - NEXUS

```
#NEXUS
BEGIN DATA;
   DIMENSIONS  NTAX=9 NCHAR=50;
   FORMAT DATATYPE=DNA MISSING=? GAP=- INTERLEAVE=No;
   Matrix

[ArgKin 596]
Papilio_glaucus_69_3       GAGaTGGAaGACAaGGTTTCGTCGACCCTGTCCGGCCTCGAGGGCGAACT
Hamearis84_13              GGaATGGAaGAGAAaGTCTCCACAACCCTCTCCGGACTCGAAGGTGAGCT
Danaus_plexippus108_21     GAGAtGGAGGAGAaGGTCTCCTCCACCCTCTCAGGTCTCGAAGGTGAACT
Greta_oto70_9              GGAATGGAaGAGAaGGTCTCCTCGACCCTCTCAGGCCTTGAAGGTGAACT
Amathusia_phidippus114_17 GGaATGGAaGACAAaGTCTCCTCAaCCCTCTCCGGTCTTGAGGGTGAACT
Morpho_peleides66_5        GGaATGGAGAGAAAaGTCTCTACTACCCTGTCTGGCCTCGAAGGCGAACT
BrintesiaB01               GGAATGGAaGACAAaGTCTCGTCCACCCTCTCCGGGCTGGAAGGCGAGCT
Elymnias_casiphone121_20   GAGAwGGaAGAcaAAGTATCCTCCACCCTCTCTGGTCTTGAAGCTGAACT
Erebia_oemeEW24_7          gGaATGGAaGACAAaGTCTCCTCGACTCTCTCTGGCCTCGAAGGCGAGCT
;
end;
```

Input format – NEXUS interleaved

```
#NEXUS
BEGIN DATA;
   DIMENSIONS  NTAX=9 NCHAR=121;
   FORMAT DATATYPE=DNA MISSING=? GAP=- INTERLEAVE=Yes;
   Matrix

[ArgKin 50 bp]
Papilio_glaucus_69_3       GAGaTGGAaGACAaGGTTTCGTCGACCCTGTCCGGCCTCGAGGGCGAACT
Hamearis84_13              GGaATGGAaGAGAAaGTCTCCACAACCCTCTCCGGACTCGAAGGTGAGCT
Danaus_plexippus108_21     GAGAtGGAGGAGAaGGTCTCCTCCACCCTCTCAGGTCTCGAAGGTGAACT
Greta_oto70_9              GGAATGGAaGAGAaGGTCTCCTCGACCCTCTCAGGCCTTGAAGGTGAACT
Amathusia_phidippus114_17 GGaATGGAaGACAAaGTCTCCTCAaCCCTCTCCGGTCTTGAGGGTGAACT
Morpho_peleides66_5        GGaATGGAGAGAAAaGTCTCTACTACCCTGTCTGGCCTCGAAGGCGAACT
BrintesiaB01               GGAATGGAaGACAAaGTCTCGTCCACCCTCTCCGGGCTGGAAGGCGAGCT
Elymnias_casiphone121_20   GAGAwGGaAGAcaAAGTATCCTCCACCCTCTCTGGTCTTGAAGCTGAACT
Erebia_oemeEW24_7          gGaATGGAaGACAAaGTCTCCTCGACTCTCTCTGGCCTCGAAGGCGAGCT

[COI 71 bp]
Papilio_glaucus_69_3       taAagAtaTTgGaACATTATACTTTATTTTTGGAATTTGAGCAAGAATATTAGGAACTTCTTTAAGTTTAT
Hamearis84_13              ???????????????????????????????????TGAGCAGGAATAGTAGGAACATCATTAAGATTAC
Libythea_celtis71_1        ???????????????????????????????????TGAGCAGGAATAGTAGGAACTTCATTAAGTCTAT
Danaus_plexippus108_21     ???????????????????????????????????TGAGCAGGAATAGTTGGGACATCTTTAAGTCTTT
Greta_oto70_9              ???????????????????????????????????TGAGCAGGAATAGTAGGAACATCTTTAAGTTTAT
Amathusia_phidippus114_17  ???????????????????????????????????TGATCTGGAATAGTAGGAACATCCCTCAGTCTTA
Morpho_peleides66_5        ???????????????????????????????????TGAGCCGGTATAATTGGTACATCCCTAAGTCTTA
BrintesiaB01               ???????????????????????????????????TGAGCAGGTATAGTAGGAACATCTCTTAGTTTAA
Elymnias_casiphone121_20   ???????????????????????????????????TGATCAGGAATAGTAGGAACTTCCCTCAGTCTTA
Erebia_oemeEW24_7          ???????????????????????????????????TGAGCAGGTATAGTAGGTACTTCCCTTAGTCTTA
;
end;
```