

Molecular Phylogenetics, Monash University Malaysia, 2024-11-12

An introduction to molecular phylogenetics

Niklas Wahlberg

Systematic Biology Group

Department of Biology

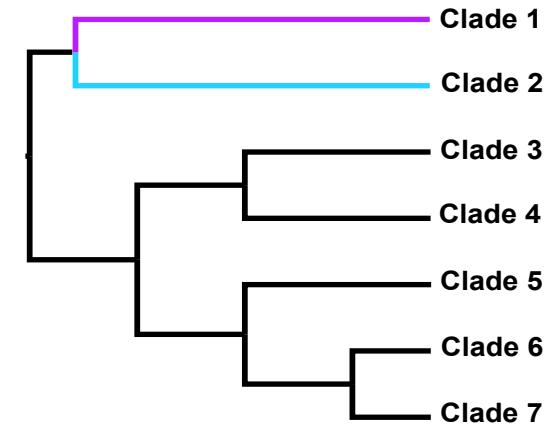
Lund University



LUND
UNIVERSITY

Aims

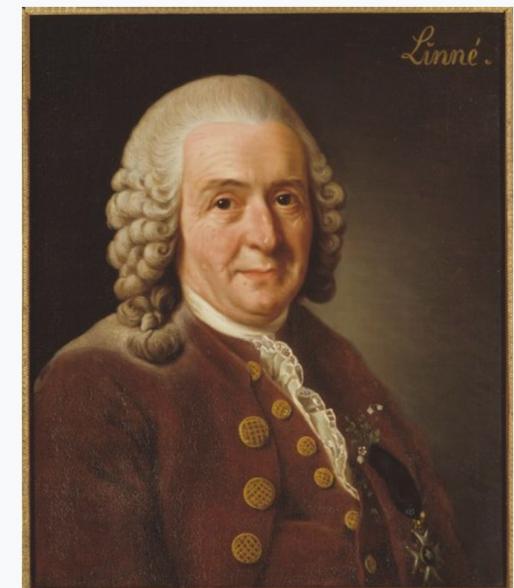
- Cover key concepts in phylogenetics
 - E.g. monophyly, homology, analogy
- Explain why evolutionary history is important in biology
- Understand the basics of statistical phylogenetic inference
- Develop “tree thinking”



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 is an edible species of mushrooms, but many other mushrooms are very poisonous, e.g. fly amanita

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



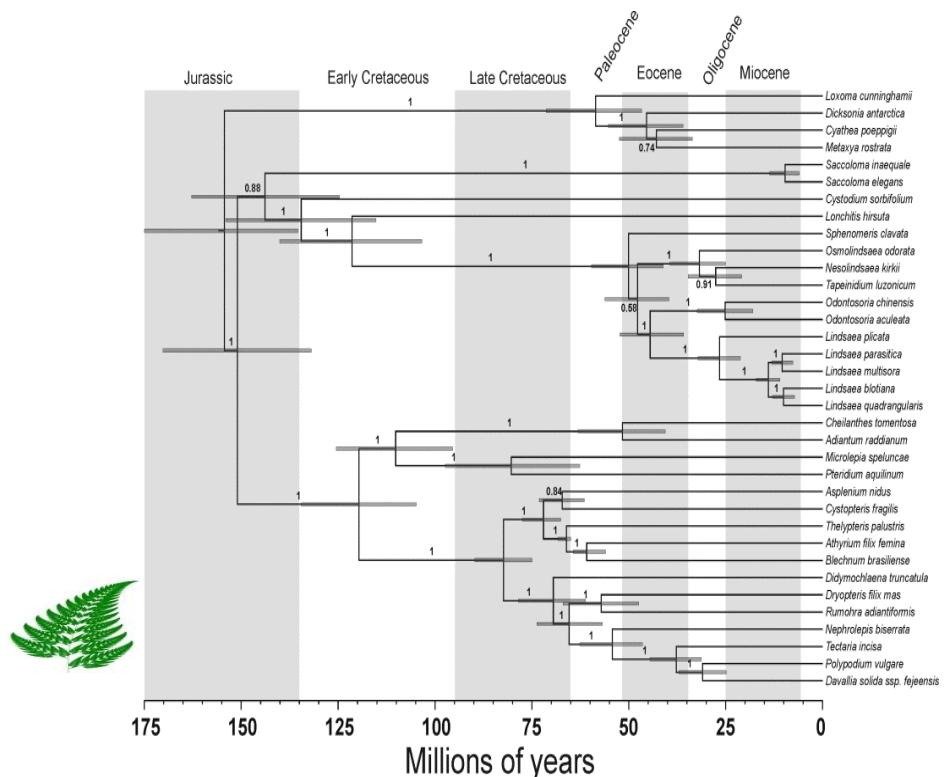
Porcini - *Boletus edulis* Bull. (1782)



Fly agaric - *Amanita muscaria* (L.) Lam. (1783)

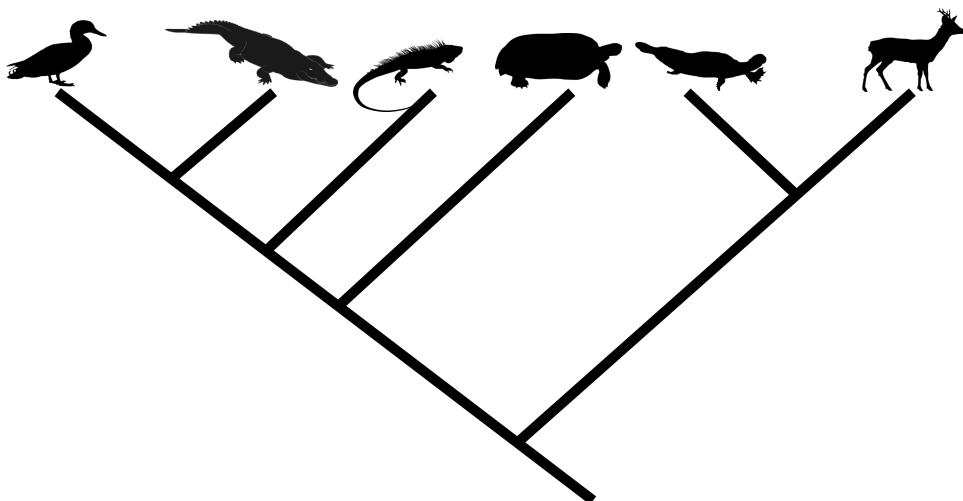
Systematics also includes...

- 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?

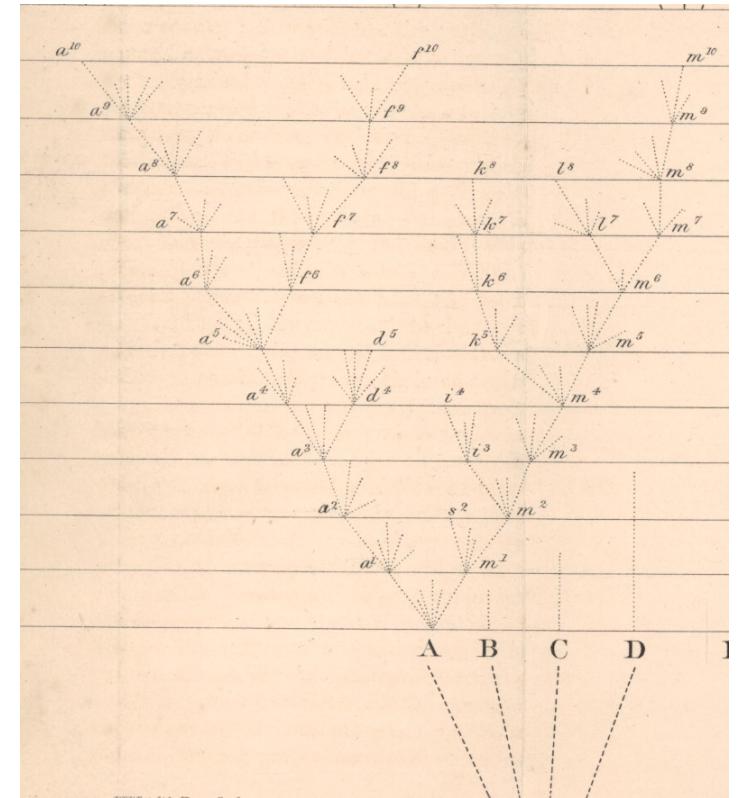


“Nothing in biology makes sense except in the light of evolution”

- Theodosius Dobzhansky, essay written in 1973

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

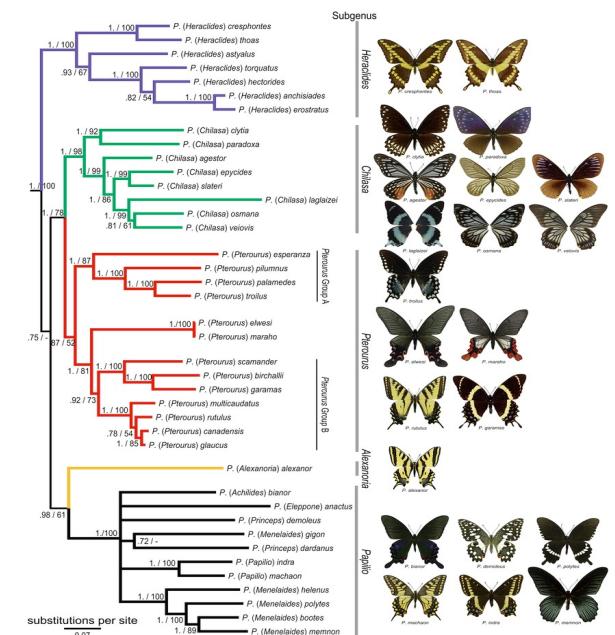
- Quote from Jay Savage, in 1997 Society of Systematic Biology Presidential Address



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

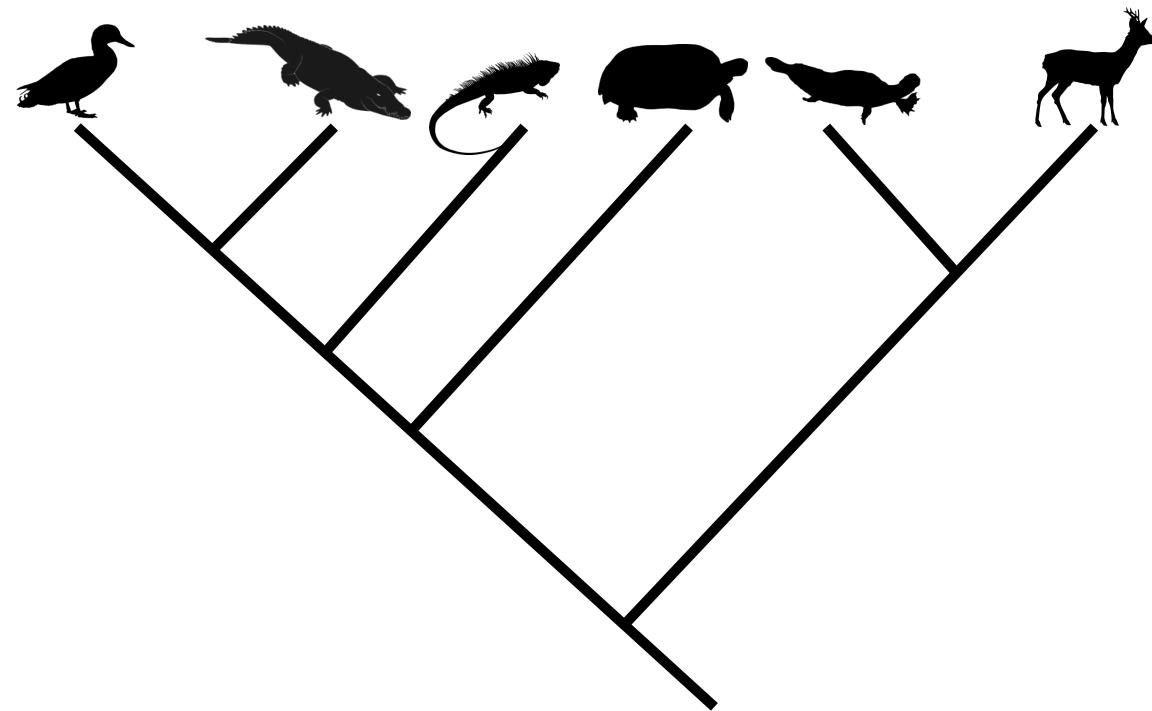
The very basic facts

- 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



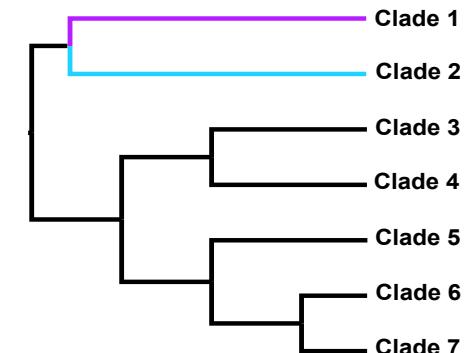
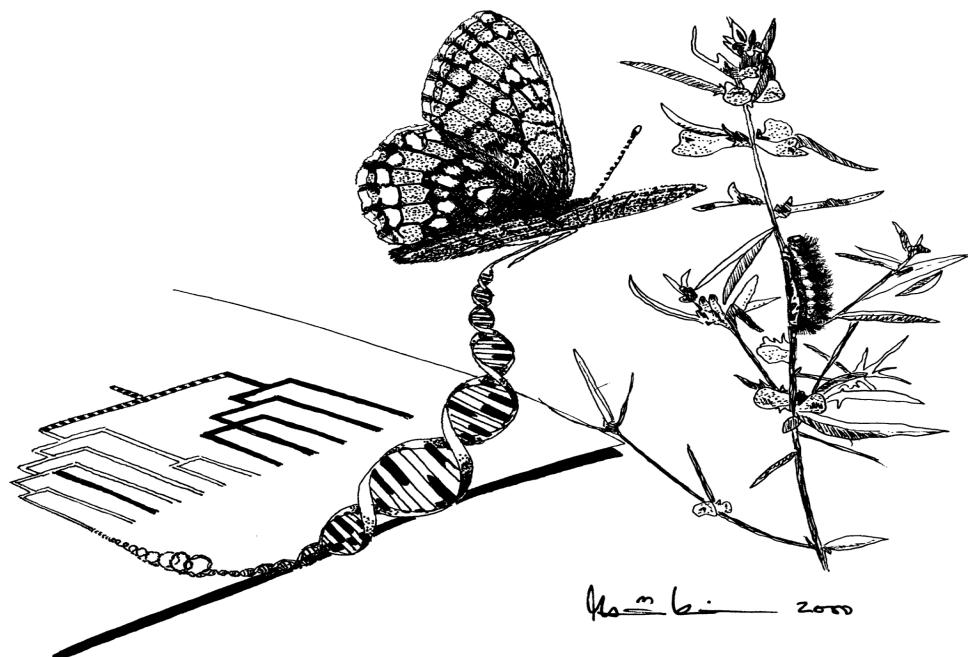
What is a phylogeny?

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



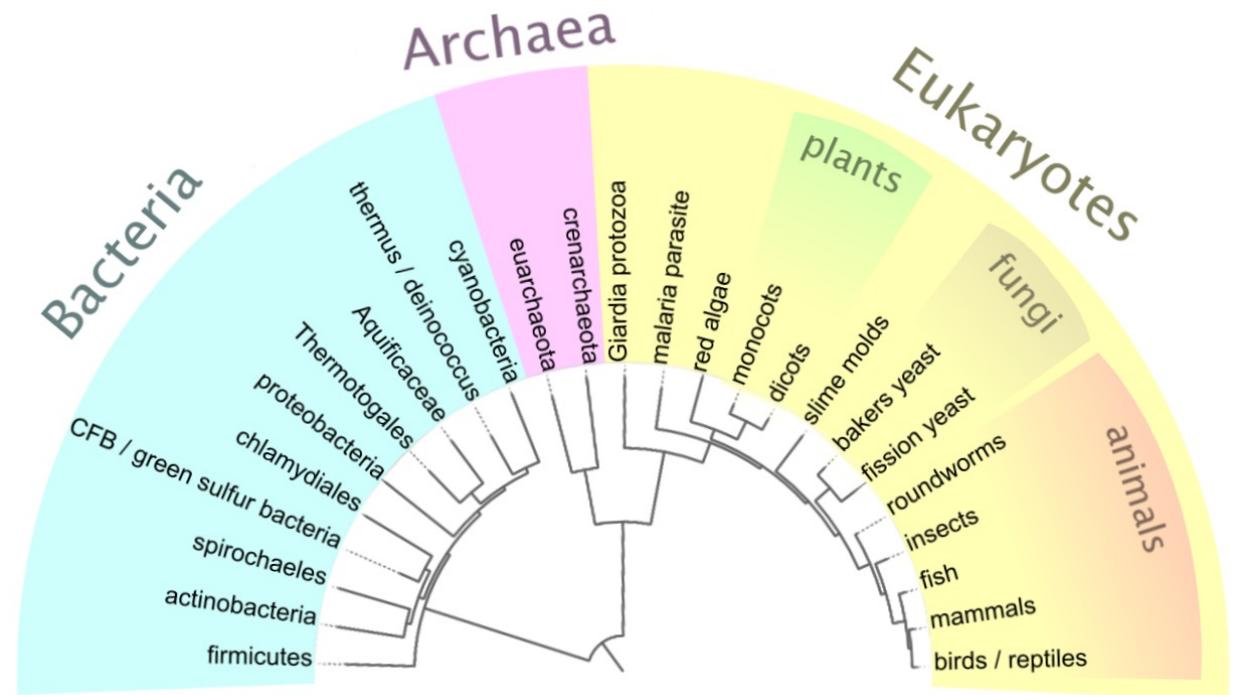
A phylogeny is an inference

- 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

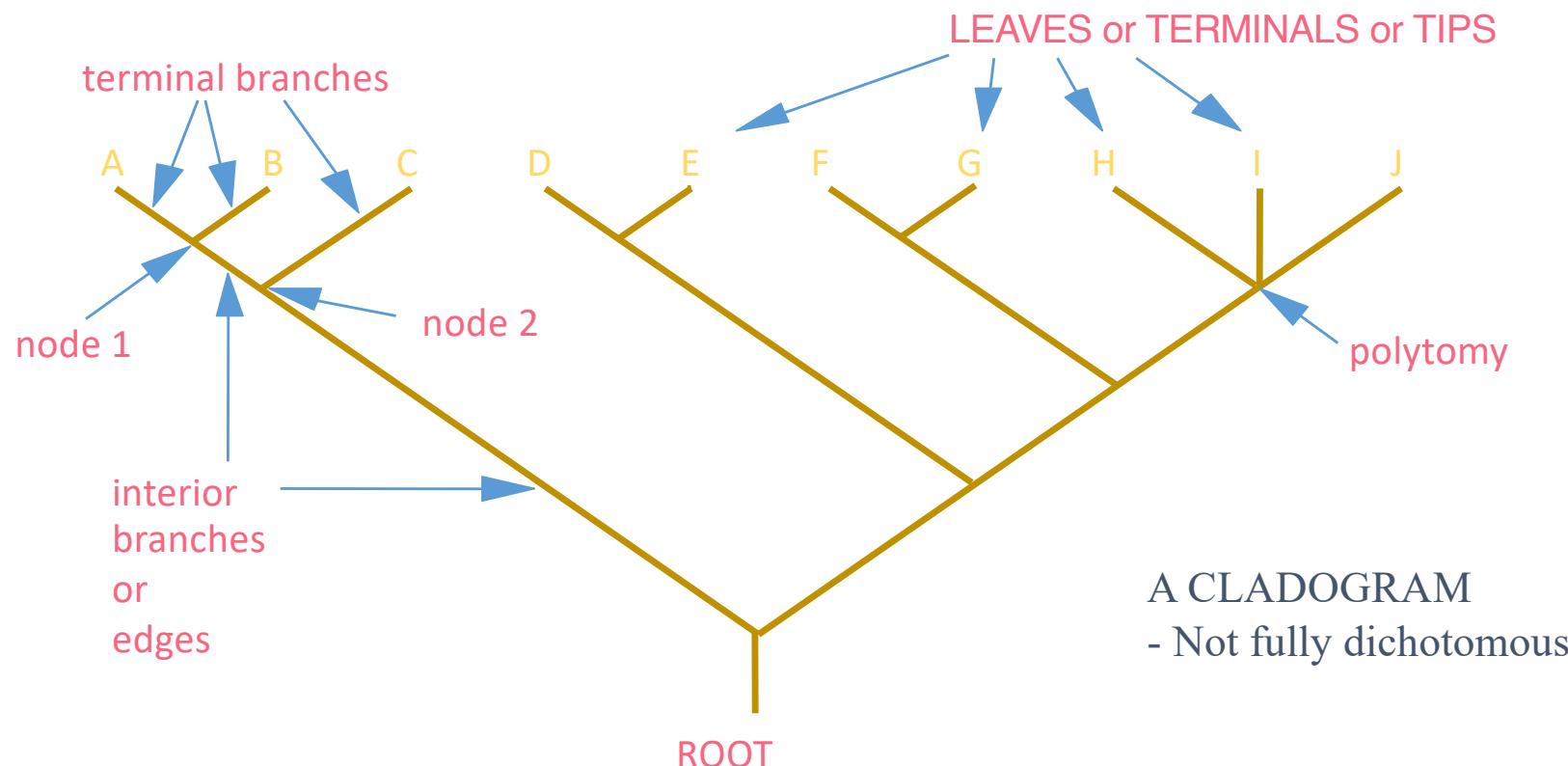


Source: Wikipedia

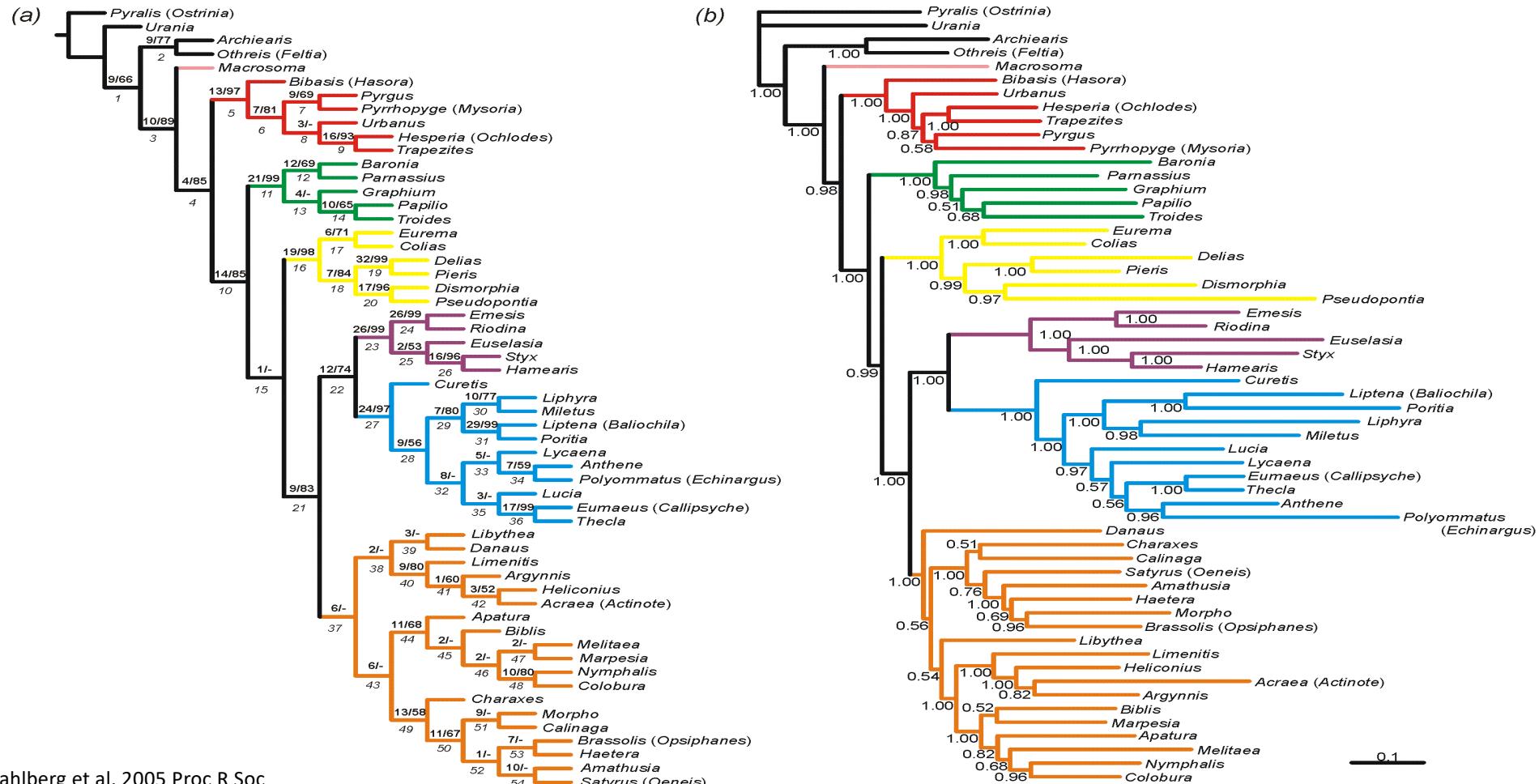
Some basic concepts

- **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

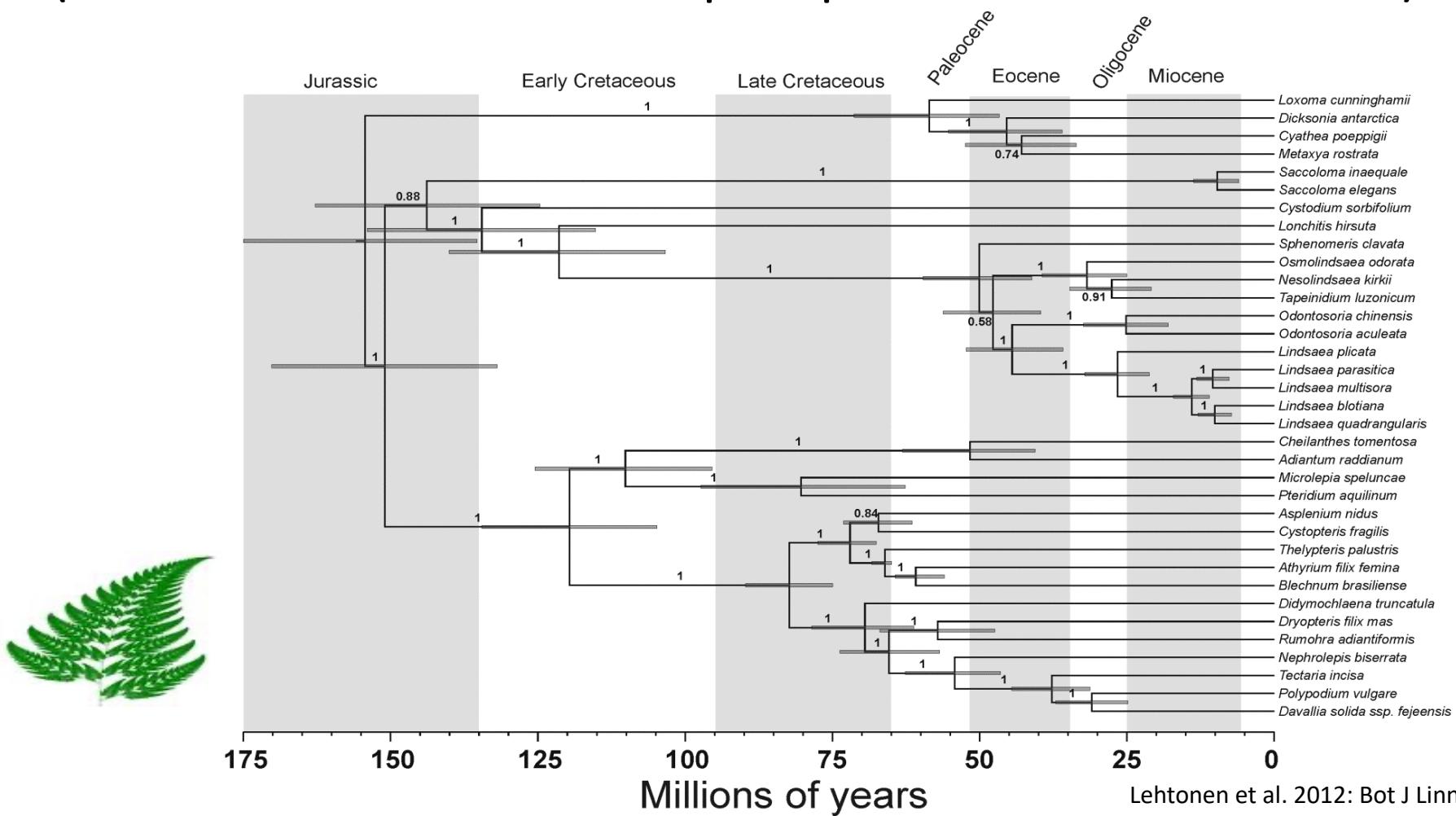


Cladograms and phygrams

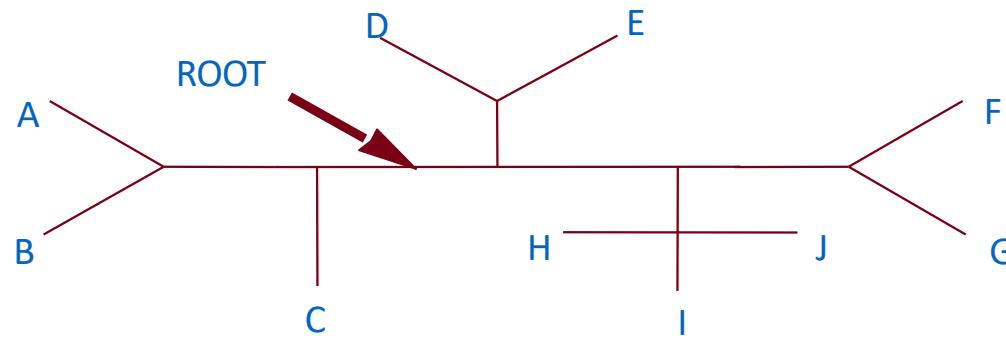
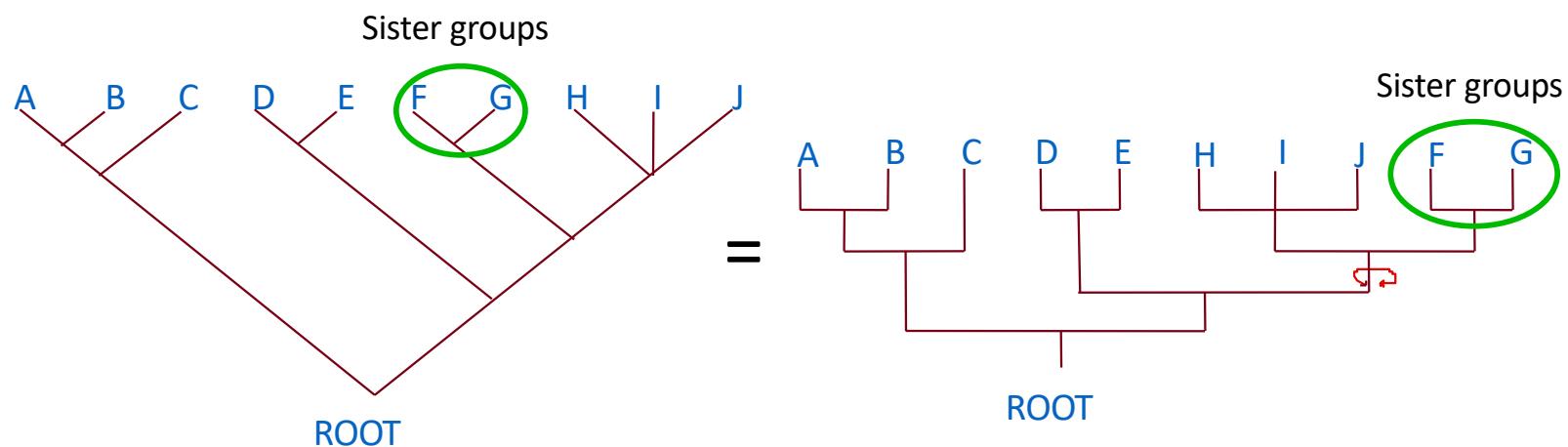


Chronogram

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

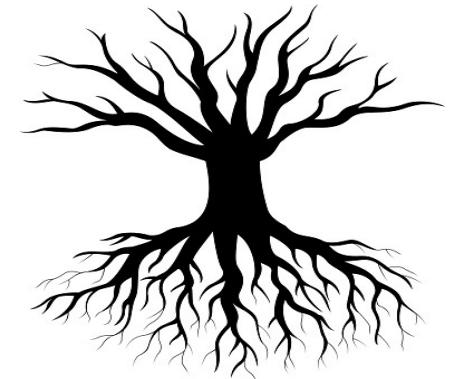


Trees - Rooted and Unrooted



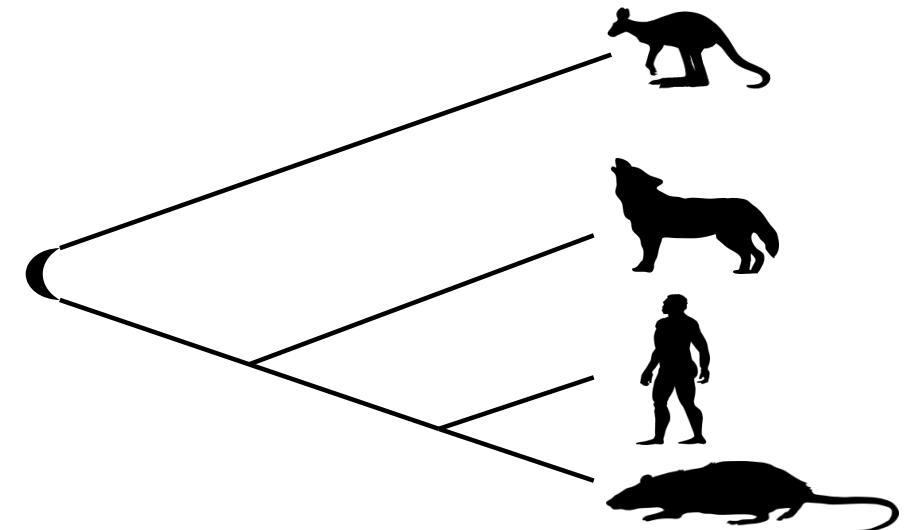
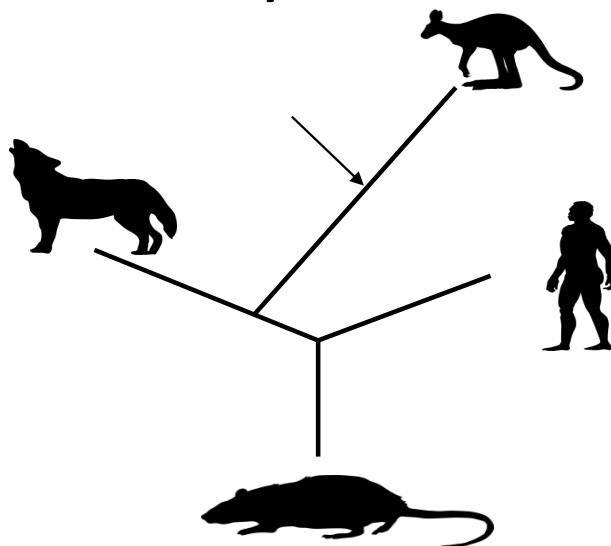
Rooting a tree

- **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)



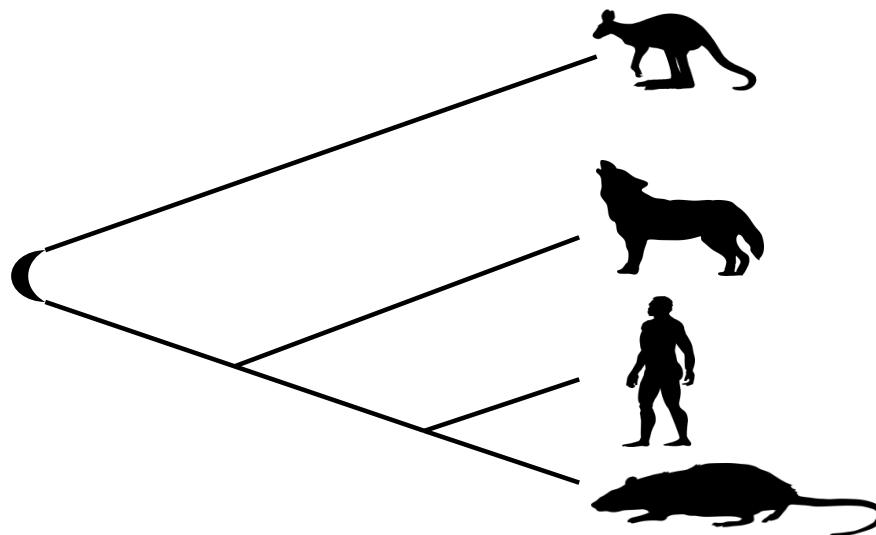
Outgroup rooting of unrooted trees

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

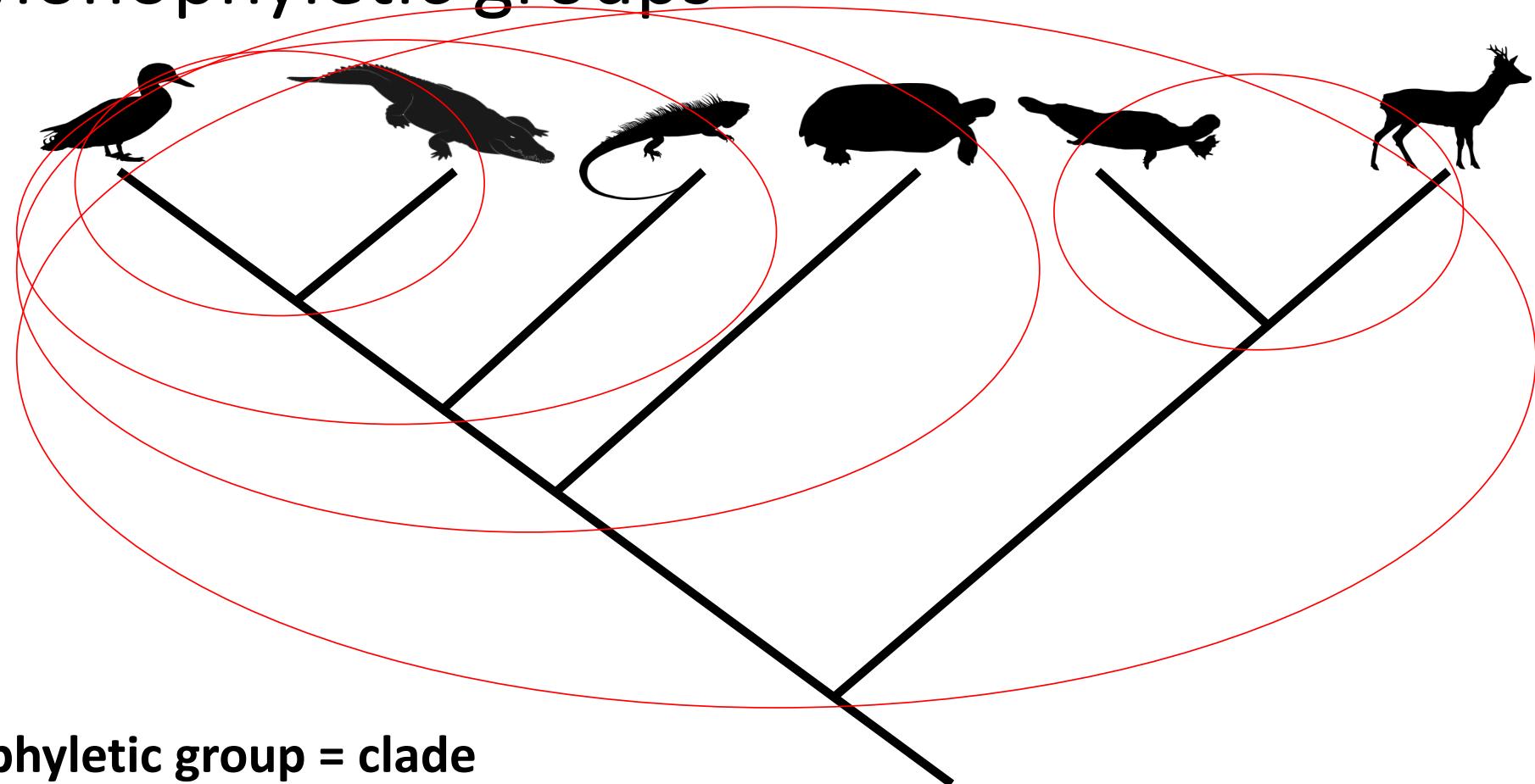


Phylogenetic systematics

- 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)



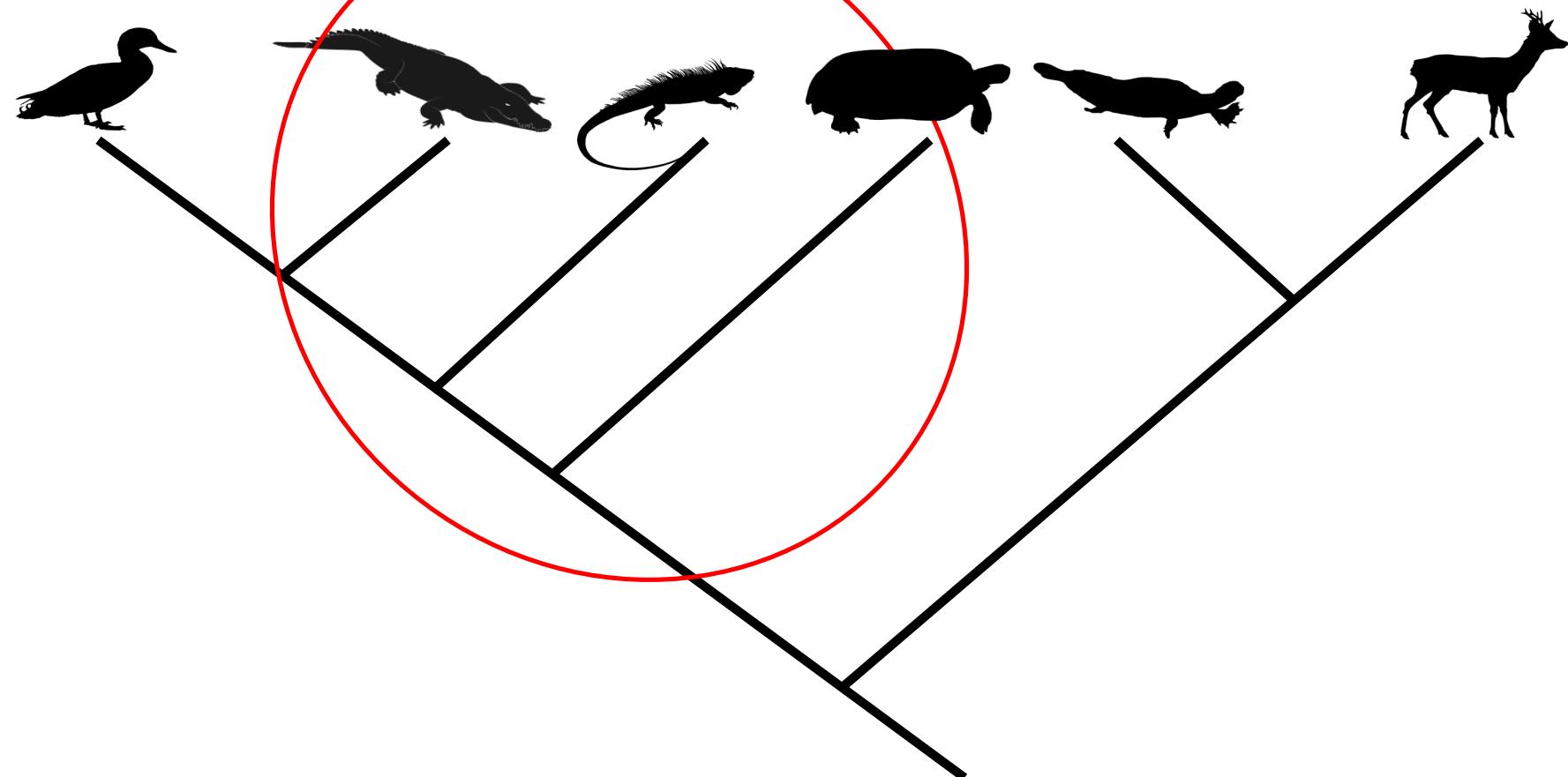
Monophyletic groups



Monophyletic group = clade

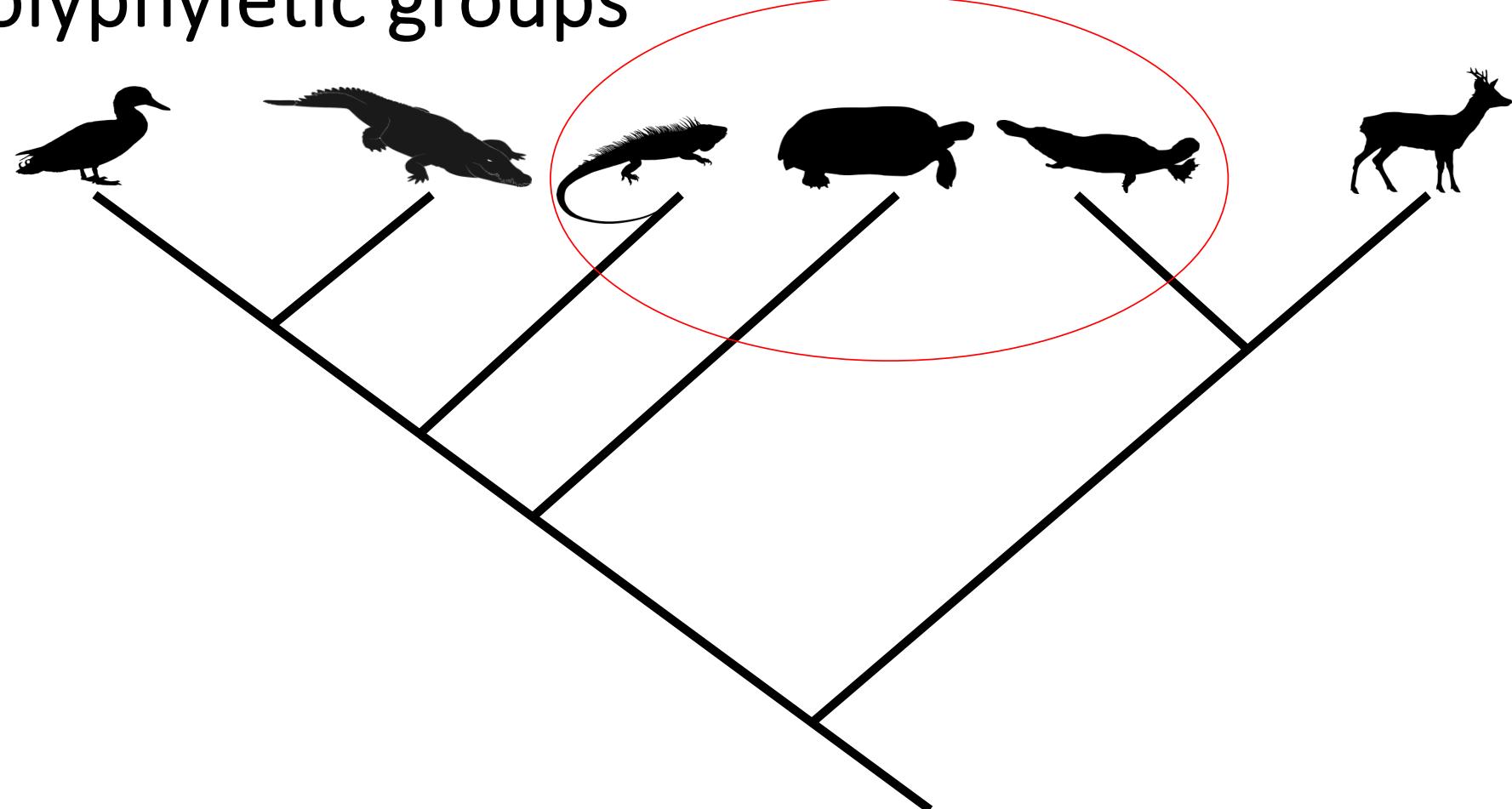
- Include an ancestor and all of its descendants

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!

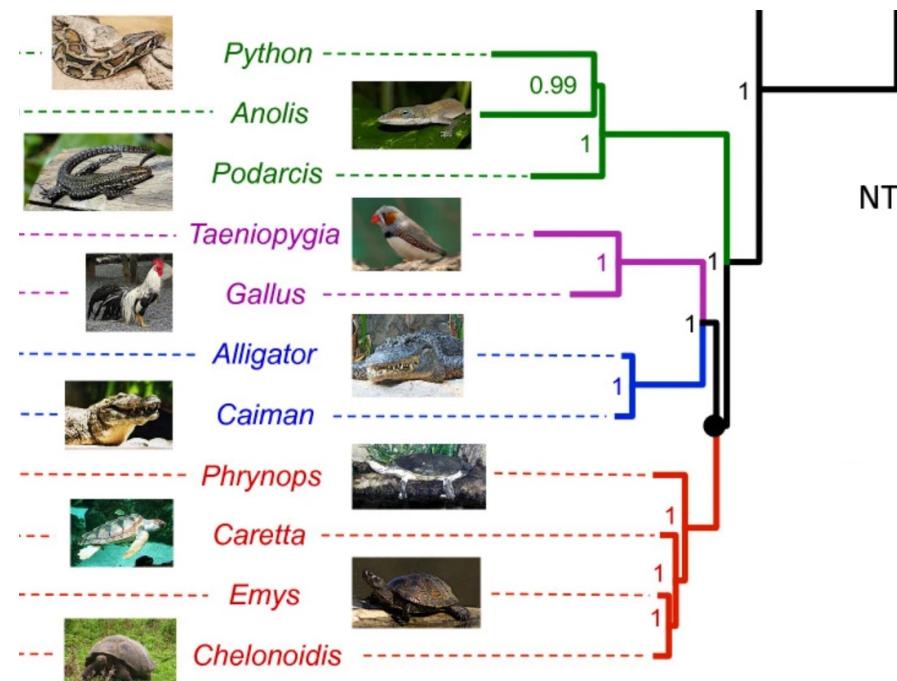
Research article | Open Access | Published: 27 July 2012

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](#)

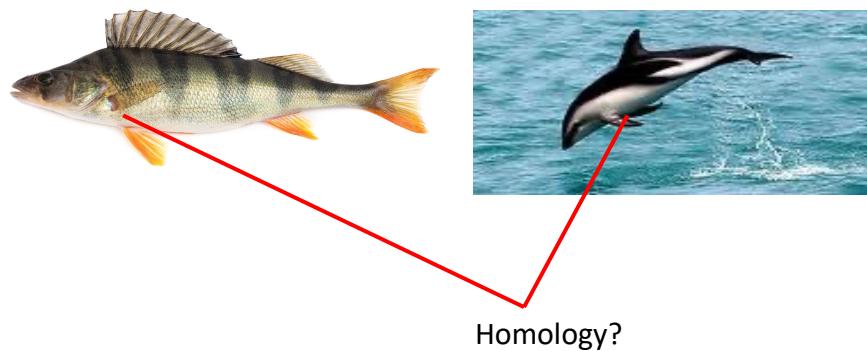


Some premises underlying phylogenetic inferences

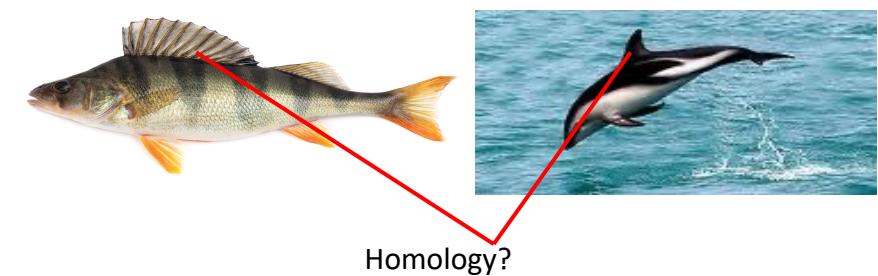
- Phylogenetic inferences are premised on
 - the inheritance of ancestral characters
 - the existence of a shared evolutionary history
- Homology considered as evidence of common ancestry
- A tree-like model of evolution
 - There are evolutionary processes that don't fit this model, e.g. lateral transfer

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.

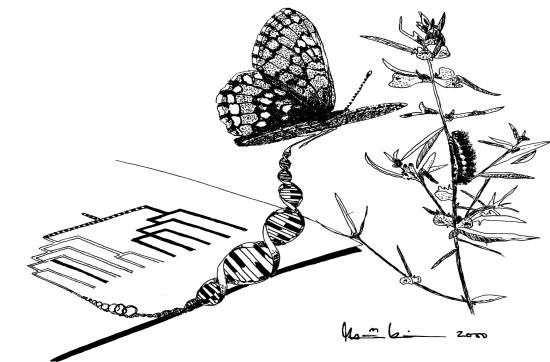
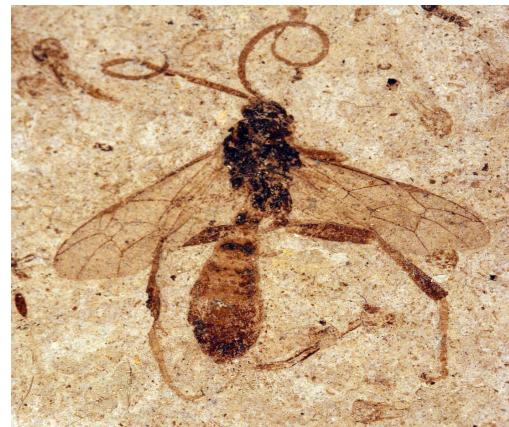


Main kinds of data in phylogenetic inference

- **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

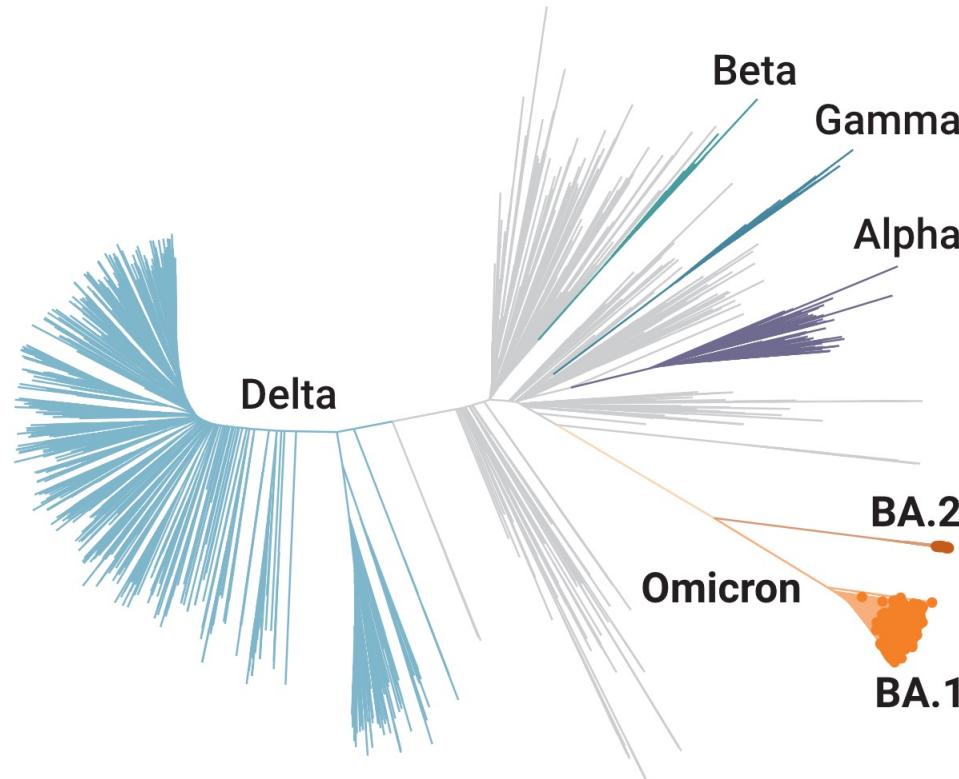
Phylogenetic analysis is an attempt to infer the past

- 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



Why do we need phylogenies?

- What is shown here?

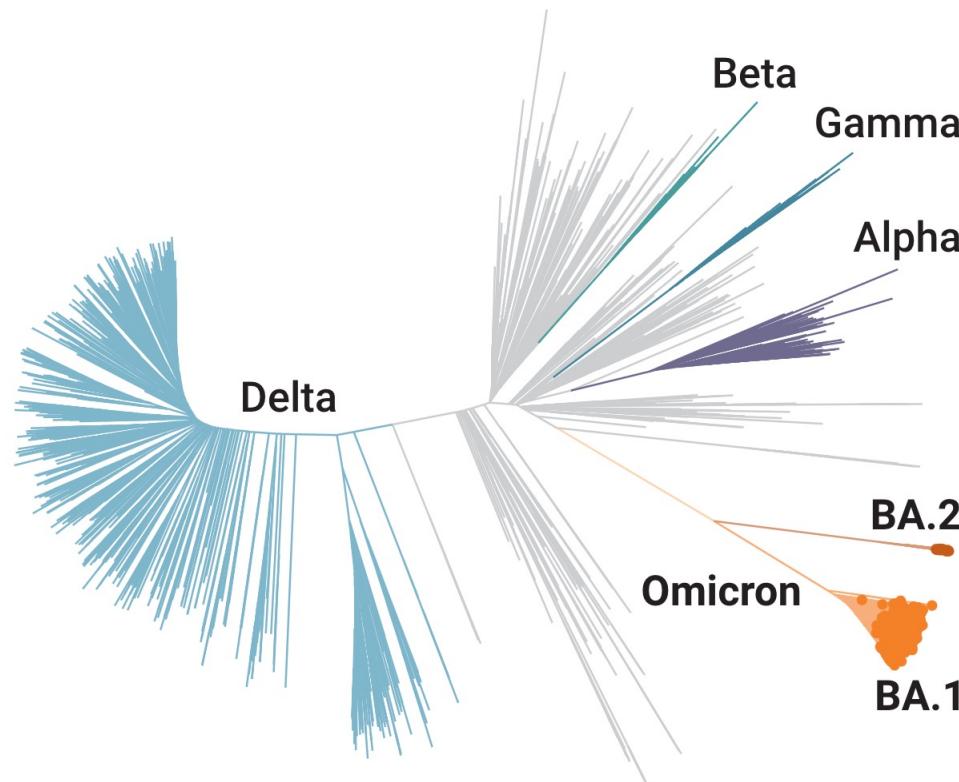


https://www.science.org/content/article/sudden-rise-more-transmissible-form-omicron-caught-scientists-surprise?utm_campaign=SciMag&utm_source=Social&utm_medium=Twitter

Why do we need phylogenies?

- What is shown here?

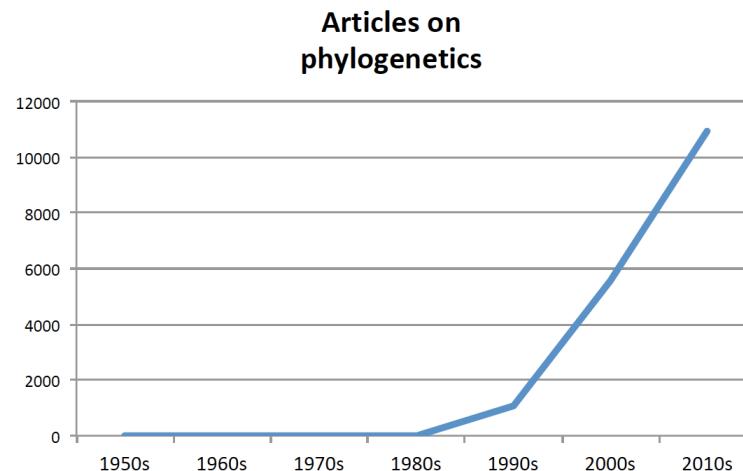
Phylogeny of covid-19 strains



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

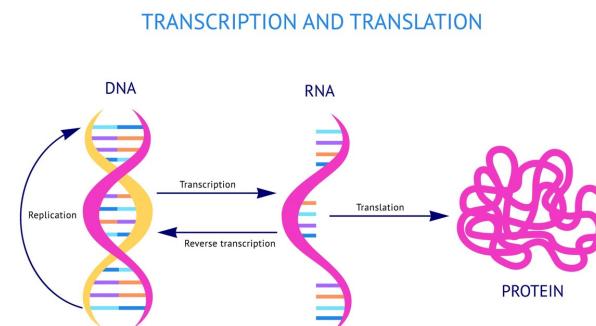
The rise of systematics

- 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



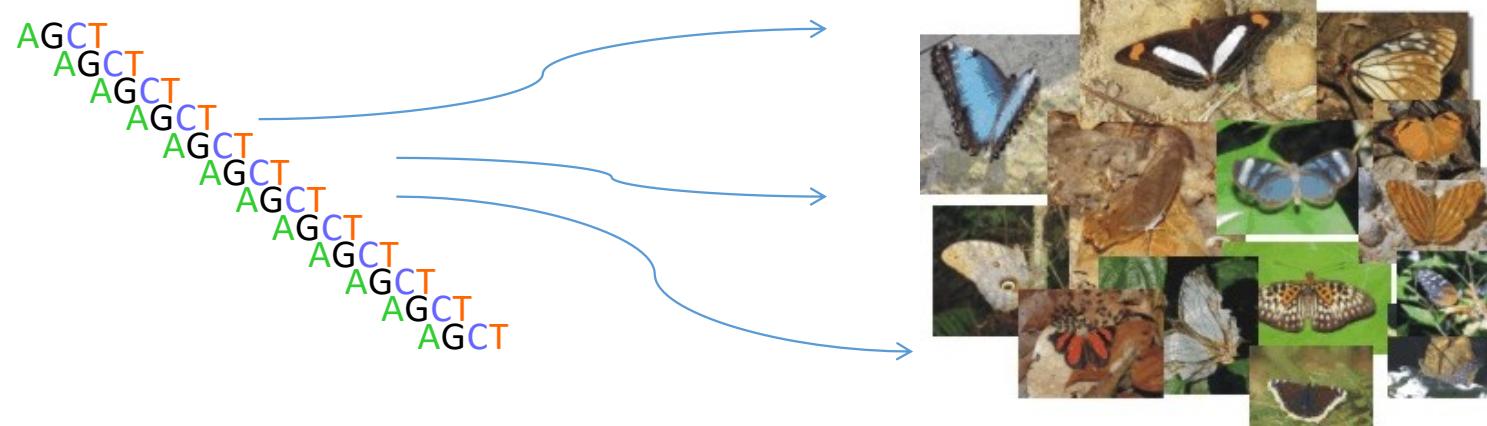
Why DNA is the Ultimate Source of Information

- **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

DNA as a source of information

- ▶ DNA has four characters

Purines

Pyrimidines

Figure B-3: The Four Nitrogenous Bases



Adenine

Guanine

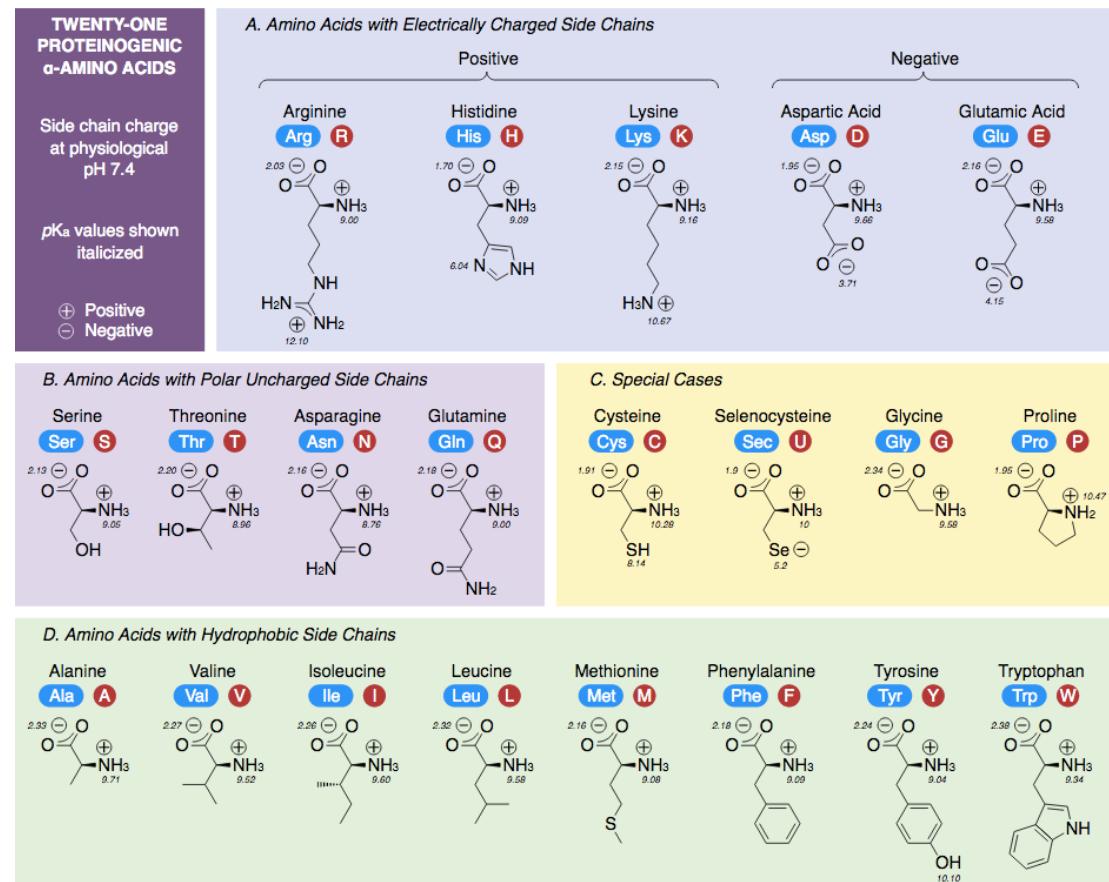
Thymine

Cytosine

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

Proteins as a source of information

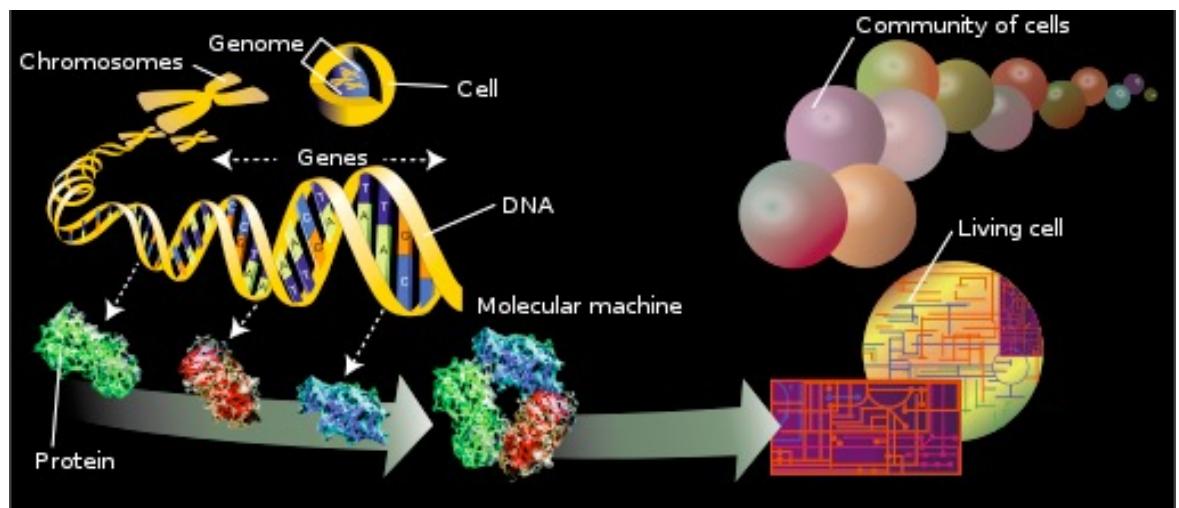
Protein sequences
have 21 characters



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!

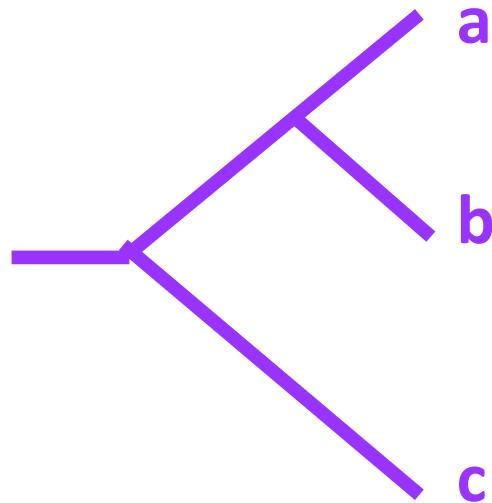
- Two steps:
 - Are we looking at the same region of the genome in our species of interest?
 - Orthology vs paralogy
 - Are we looking at the same site within our chosen marker in our species of interest?
 - Alignment

Orthology or paralogy?

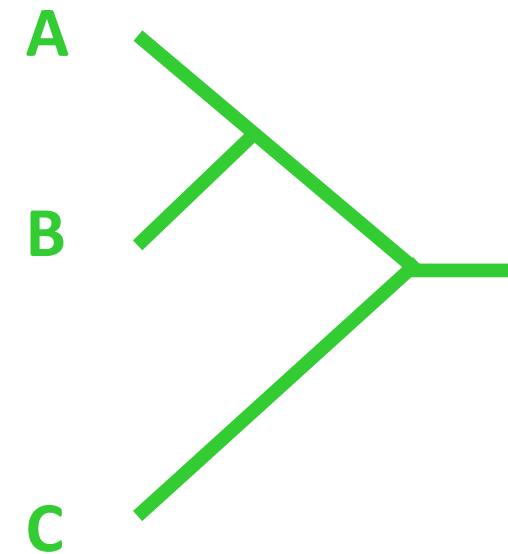
- Are the genome regions sequenced from different species the same (homologous)?
- Gene duplication
 - 1) duplicate gene degenerates - pseudogene
 - 2) duplicate gene acquires new function
- A problem particularly acute currently as we analyze phylogenomic data

Orthology: gene trees and species trees

Gene phylogeny



Organism phylogeny

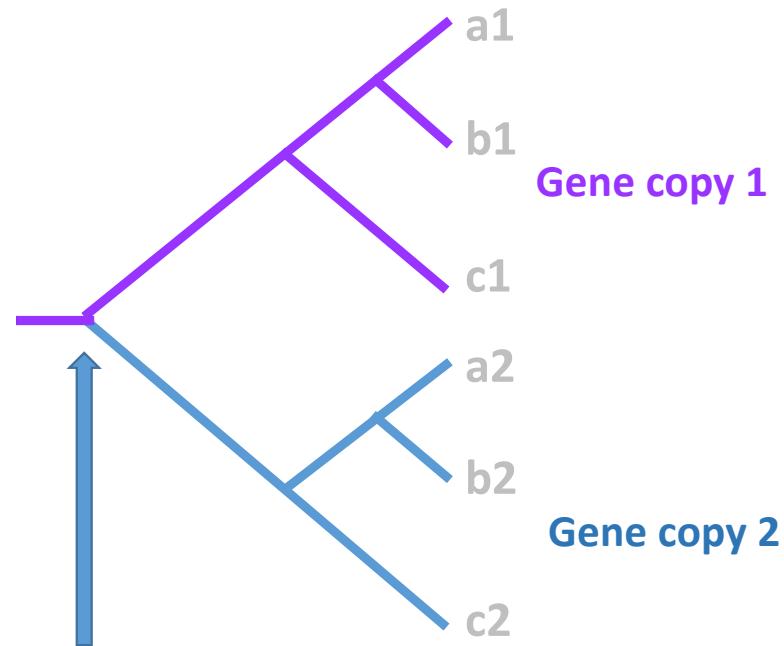


ORTHOLOGY

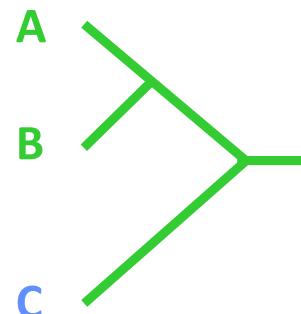
Orthologs: genes that arose due to speciation

Paralogy: can produce misleading trees

Gene phylogenies



Organism phylogeny



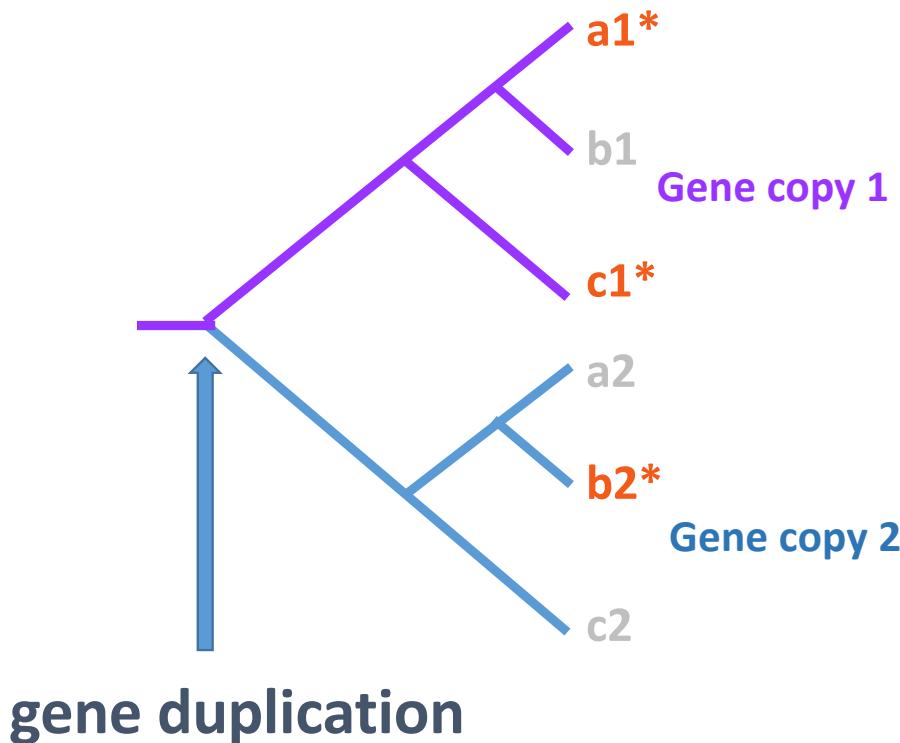
gene duplication

PARALOGY

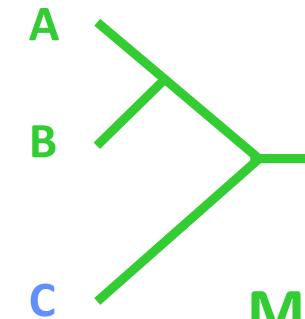
Paralogs: genes that arose due to duplication events

Paralogy: can produce misleading trees

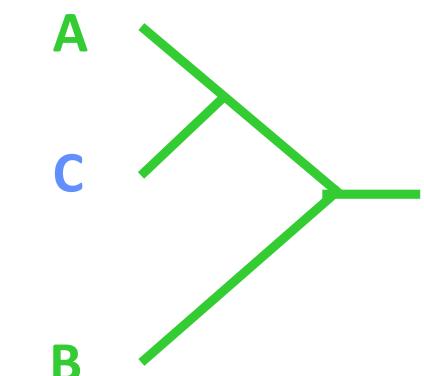
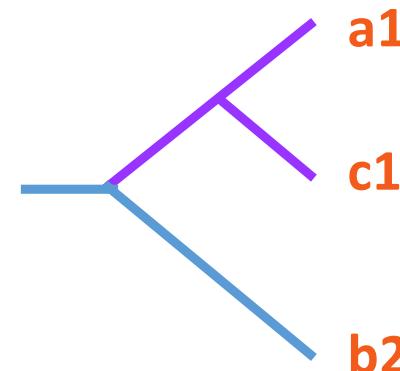
Gene phylogenies



Organism phylogeny



Misleading tree



Paralogs: genes that arose due to duplication events

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 process of estimating a phylogenetic tree

- Acquire the data → lab work or public databases
- Assemble the relevant data into dataset(s) → determine orthology!
- Align the sequences → determine homologous positions!
- Analyse the data in an appropriate program → models of evolution!
- Interpret the results

What do we model in DNA sequence evolution?

- Nucleotide substitutions
 - The rate at which each nucleotide is replaced by each alternative nucleotide

What is the challenge?

- DNA has only four characters

Figure B-3: The Four Nitrogenous Bases

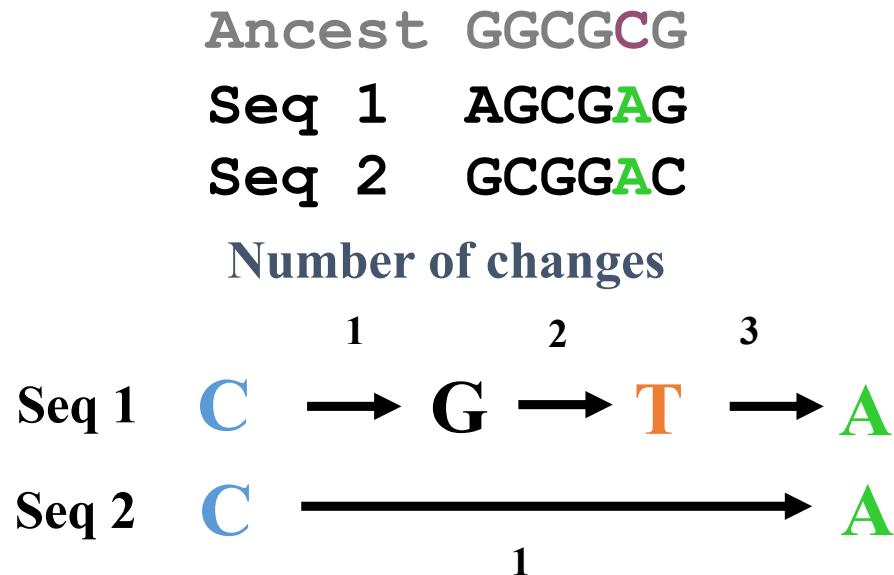


Each base has a distinct shape that can be used to distinguish it from the others. 3D representations of the four bases are shown, with the corresponding chemical structures drawn above.

Multiple changes at a single site
- hidden changes

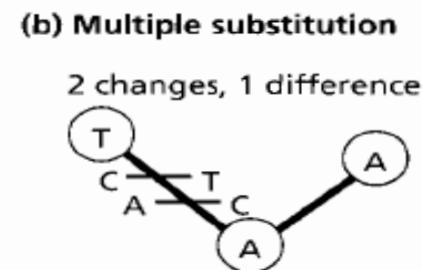
Seq 1 AGCGAG
Seq 2 GC~~G~~AC

Multiple changes at a single site - hidden changes



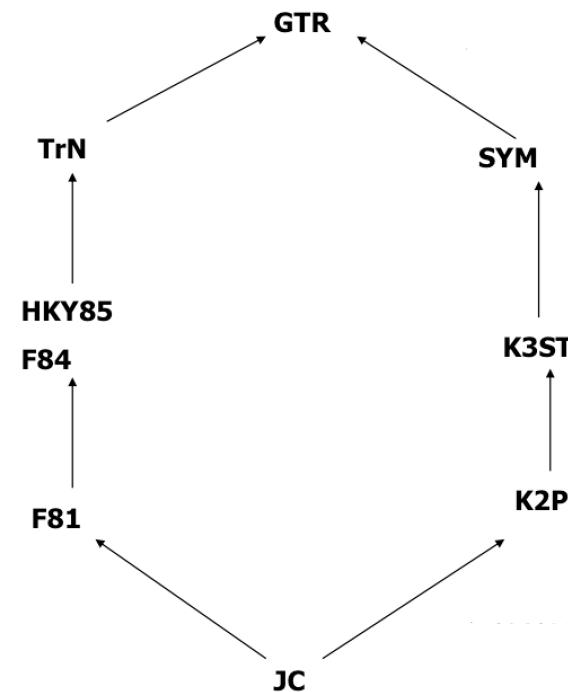
Saturation in sequence data:

- Saturation is due to **multiple substitutions at the same site** subsequent to lineage splitting
- Models of evolution attempt to infer the missing information through correcting for “**multiple hits**”
- Most data will contain some fast evolving sites which are potentially saturated
 - e.g. in protein-coding genes codon position 3



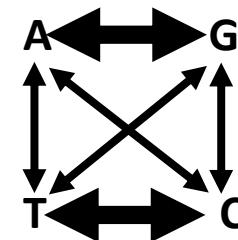
Saturation in sequence data (cont.)

- In severe cases the data become essentially random and all information about relationships can be lost
- Probabilistic models of sequence evolution are used to calculate expected distances



Substitution types

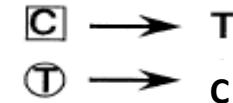
- Purines: A, G
- Pyrimidines: C, T
- Transversions
 - Pu --> Pyr
 - Pyr --> Pu
- Transitions – more common
 - Pu --> Pu
 - Pyr --> Pyr



Pur - Pyr mispairs lead to transitions



In next round of replication



Modelling nucleotide substitutions

- These dynamics can be modelled over a tree and they are incorporated into distance methods, maximum likelihood, and Bayesian inference
- Models incorporate information about the **rates at which each nucleotide is replaced** by each alternative nucleotide
 - For DNA this can be expressed as a 4×4 rate matrix (known as the Q matrix)
- Other model parameters may include:
 - Site by site rate variation (aka among-site rate variation – ASRV)

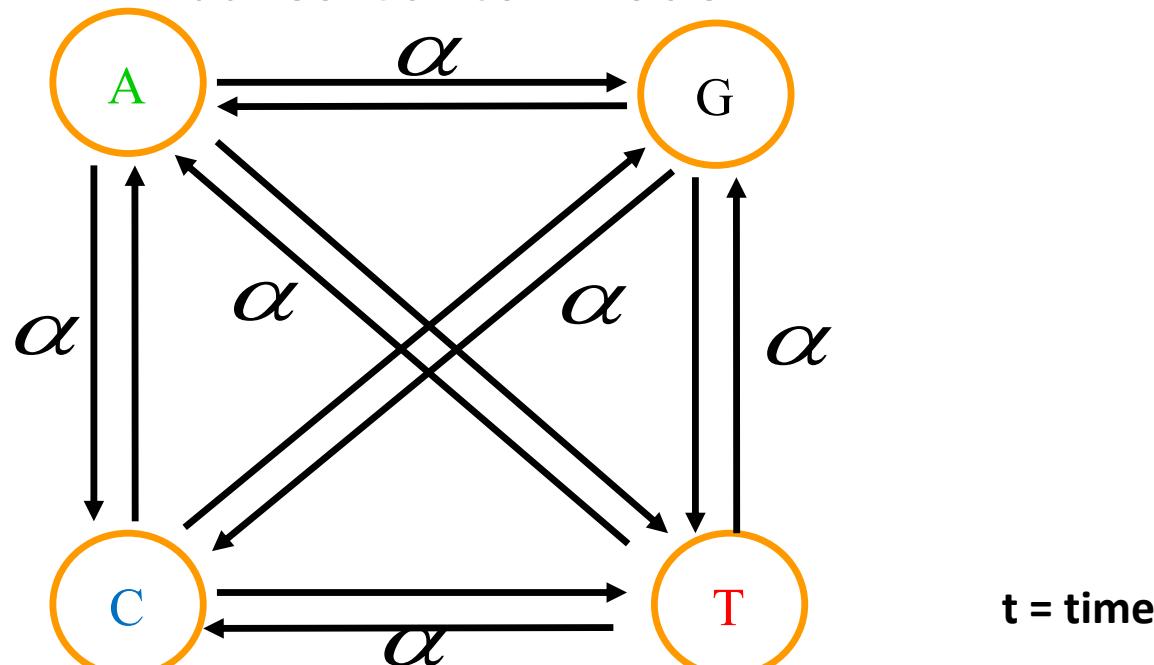
Corrections for multiple substitutions: First DNA substitution model

Jukes & Cantor (1969) assumptions:

1. **A = T = G = C No nucleotide bias**
2. **Every base changes to every other base with equal probability (no TS/TV bias)**
3. **All sites change with the same probability (no ASRV - among-site rate variation)**

Also: probability of substitution & base composition remains constant over time/across lineages

Jukes-Cantor model



$t = \text{time}$

- $\alpha = \text{the rate of substitution}$ (α changes from A to G every t)
- The rate of substitution for each nucleotide is 3α
- In t steps there will be $3\alpha t$ changes

The Q matrix

		To			
		A	C	G	T
From	A	-3 α	α	α	α
	C	α	-3 α	α	α
	G	α	α	-3 α	α
	T	α	α	α	-3 α

The Jukes-Cantor model: the simplest model

	A	C	G	T	
A	-3 α	α	α	α	JC model: one parameter model
C	α	-3 α	α	α	1) It assumes that all bases are equally frequent ($p=0.25$)
G	α	α	-3 α	α	2) It assumes that all sites can change and they do so at the same rate of α
T	α	α	α	-3 α	

The Jukes-Cantor model: the simplest model

	A	C	G	T	
A	—	α	α	α	JC model: one parameter model
C	α	—	α	α	1) It assumes that all bases are equally frequent ($p=0.25$)
G	α	α	—	α	2) It assumes that all sites can change and they do so at the same rate of α
T	α	α	α	—	

Improvements on Jukes-Cantor

- Allow **base frequencies** to be unequal to accommodate e.g. sequences such as these

AAACCTGGATTACCGAGATTAAAGCGATATATTGCAATGC

34% A

17% C

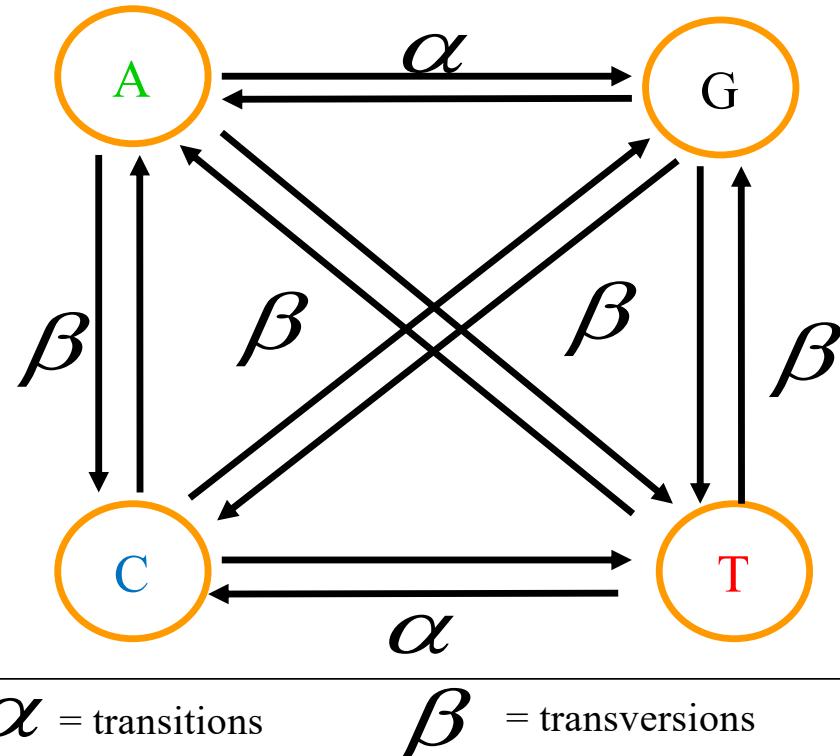
29% T

20% G

- Allow **transitions** to be more common than **transversions**, in fact, allow separate estimates of the probability of change of **all six possible nucleotide substitutions**
- Allow the **probability of substitution to change along the molecule - ASRV**

		2nd position				3rd position
1st position	U	C	A	G		
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr stop stop	Cys Cys stop Trp	U C A G	
	Leu Leu Leu Leu	Pro Pro Pro Pro	His Gln Gln Gln	Arg Arg Arg Arg	U C A G	
	Ile Ile Met	Thr Thr Thr	Asn Lys Lys	Ser Ser Arg	U C A G	
	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G	
Amino Acids						
Ala: Alanine Arg: Arginine Asp: Aspartagine Asn: Asparagine Asu: Asparagine cod Asx: Aspartic acid Cys: Cysteine		Gln: Glutamine Gly: Glycine Dcy: Dihydroxy Dye: Dihydroxy Ile: Isoleucine Leu: Leucine Lys: Lysine Met: Methionine Phe: Phenylalanine Pro: Proline		Ser: Serine Thr: Threonine Trp: Tryptophane Val: Valine		

Kimura (1980) model: K2P



The Kimura model has 2 parameters

	A	C	G	T
A	-	β	α	β
C	β	-	β	α
G	α	β	-	β
T	β	α	β	-

K2P model is more realistic, but
still

- 1) It assumes that all bases are equally frequent ($p=0.25$)
- 2) There are two substitution types (transitions – α and transversions - β)

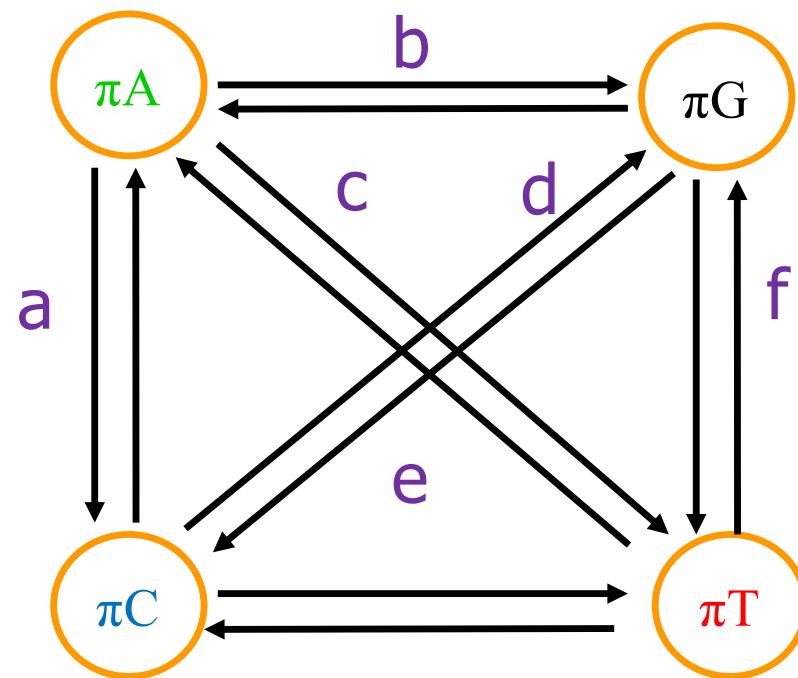
The Hasegawa-Kishino-Yano model

	A	C	G	T
A	—	$\pi_C\beta$	$\pi_G\alpha$	$\pi_T\beta$
C	$\pi_A\beta$	—	$\pi_G\beta$	$\pi_T\alpha$
G	$\pi_A\alpha$	$\pi_C\beta$	—	$\pi_T\beta$
T	$\pi_A\beta$	$\pi_C\alpha$	$\pi_G\beta$	—

HKY model:

- 1) Base frequencies are allowed to vary: π_A , π_C , π_G , π_T
- 2) There are two substitution types (transitions – α and transversions – β)

The General Time-Reversible model



The General Time-Reversible model (GTR)

	A	C	G	T
A	—	$\pi_C a$	$\pi_G b$	$\pi_T c$
C	$\pi_A a$	—	$\pi_G d$	$\pi_T e$
G	$\pi_A b$	$\pi_C d$	—	$\pi_T f$
T	$\pi_A c$	$\pi_C e$	$\pi_G f$	—

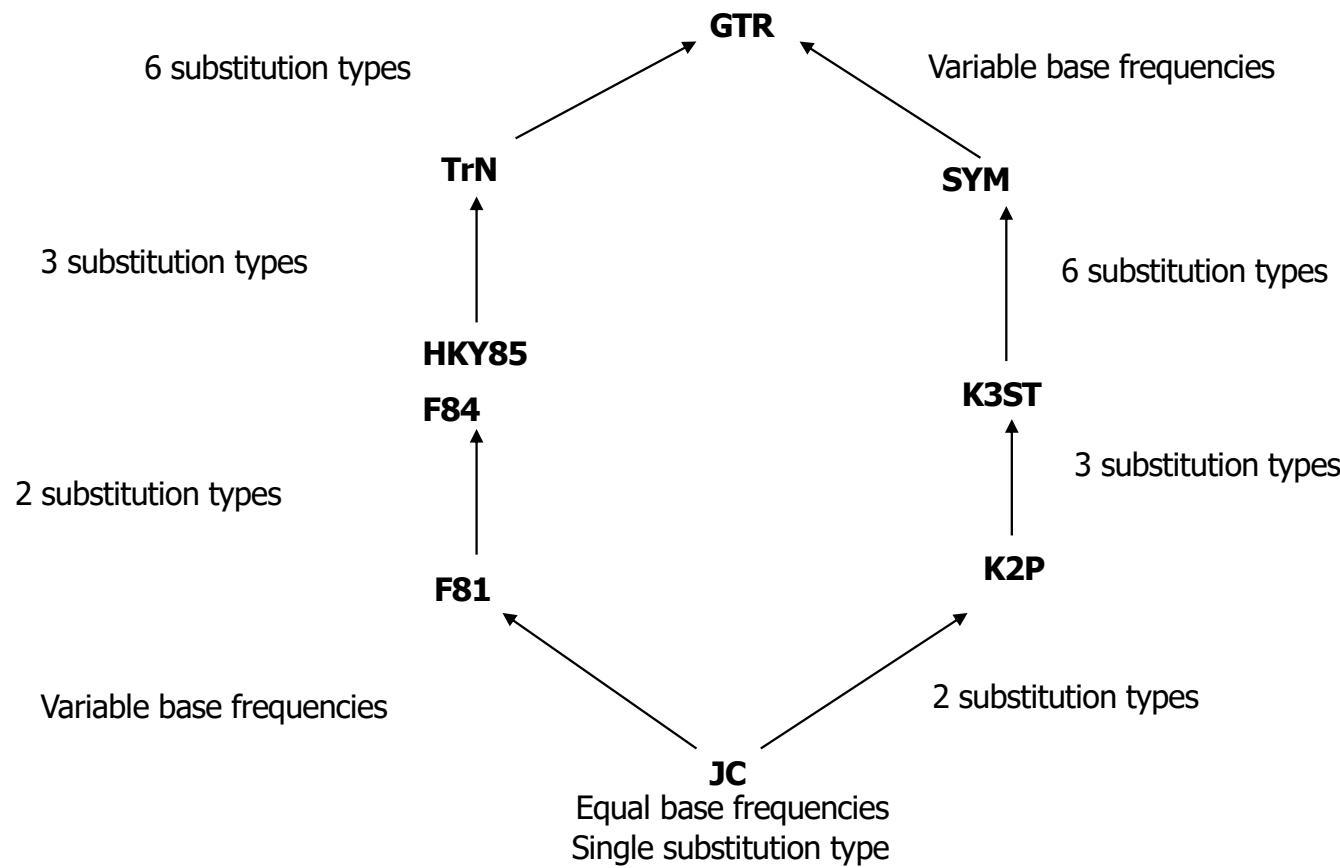
GTR model:

- 1) Base frequencies are allowed to vary: π_A , π_C , π_G , π_T
- 2) There are six substitution types: a, b, c, d, e, f

The most commonly used models

- Almost all models used are special cases of one model:
 - The general time reversible model - GTR

ACA_GGTGAGGCTCA_GCCAAT_TTGAGC_TTTGTCGATAGGT



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
GAGATGGAAgACAAaGGTTCGTCGACCCCTGTCCGGCCTCGAGGGCGAACT
>Hamearis84_13
GGATGGAAgAGAAaGTCTCCACAACCCTCTCCGGACTCGAAGGTGAGCT
>Danaus_plexippus108_21
GAGATGGAGAGAAgGGTCTCCTCACCCCTCTCAGGTCTCGAAGGTGAACT
>Greta_oto70_9
GGAATGGAAgAGAAaGGTCTCCTCGACCCCTCTCAGGCCTTGAGGGTGAAC
>Amathusia_phidippus114_17
GGATGGAAgACAAaGTCTCCTCAaCCCTCTCCGGTCTTGAGGGTGAAC
>Morpho_peleides66_5
GGATGGAGAGAAaGTCTCTACTACCCTGTCTGGCCTCGAAGGCGAACT
>BrintesiaB01
GGAATGGAAgACAAaGTCTCGTCCACCCTCTCCGGGCTGGAAGGCGAGCT
>Elymnias_casiphone121_20
GAGAwGGAAgACAAaAAGTATCCTCACCCCTCTGGTCTTGAGCTGAAC
>Erebia_oemeEW24_7
gGaATGGAAgACAAaGTCTCCTCGACTCTCTGGCCTCGAAGGCGAGCT
```

Input format – PHYLIP

```
9 50
Papilio_g1 GAGaTGGAAgACAAggTTCGTCGACCCTGTCCGGCCTCGAGGGCGAACT
Hamearis84 GGaATGGAAgAGAAaGTCTCCACAACCCCTCTCGAAGGTGAGCT
Danaus_ple GAGAtGGAGGAGAaGGTCTCCTCCACCCCTCTCAGGTCTCGAAGGTGAACt
Greta_oto7 GGAATGGAAgAGAAaGGTCTCCTCGACCCTCTCAGGCCTTGAAAGGTGAACt
Amathusia_ GGaATGGAAgACAAaGTCTCCTCAaCCCTCTCCGGTCTTGAGGGTGAACt
Morpho_pel GGaATGGAGAGAAAaGTCTCTACTACCCTGTCTGGCCTCGAAGGCGAACT
BrintesiaB GGAATGGAAgACAAaGTCTCGTCCACCCCTCTCGGGCTTGAAGGGCAGCT
Elymnias_c GAGAwGGaAGAcAaAAGTATCCTCCACCCCTCTGGTCTTGAAAGCTGAACt
Erebia_oem gGaATGGAAgACAAaGTCTCCTCGACTCTCTGGCCTCGAAGGCGAGCT
```

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      GAGaTGGAAgACAAggTTCTGACCCCTGTCCGGCCTCGAGGGCGAACT
  Hamearis84_13              GGaATGGAAgAGAAAGTCTCCACAACCCCTCTCCGGACTCGAAGGTGAGCT
  Danaus_plexippus108_21    GAGAtGGAGGAGAAggTCTCCTCCACCCCTCTCAGGTCTCGAAGGTGAACT
  Greta_oto70_9               GGAATGGAAgAGAAggTCTCCTCGACCCCTCTCAGGCCTTGAAGGTGAACT
  Amathusia_phidippus114_17  GGaATGGAAgACAAAGTCTCCTCAACCCCTCTCCGGTCTTGAGGGTGAACT
  Morpho_peleides66_5        GGaATGGAGAGAAAAGTCTCTACTACCCCTGTCTGGCCTCGAAGGCGAACT
  BrintesiaB01               GGAATGGAAgACAAAGTCTCGTCCACCCCTCTCCGGGCTTGAAGGCAGCT
  Elymnias_casiphone121_20  GAGAwGGAAgACAAAGTATCCTCCACCCCTCTGGTCTTGAAGCTGAACT
  Erebia_oemeEW24_7          gGaATGGAAgACAAAGTCTCCTCGACTCTCTGGCCTCGAAGGCGAGCT
;
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      GAGaTGGAAgACAAggTTCGTCAccCTGTCGGCCTCGAGGGCGAAct
Hamearis84_13              GGaATGGAAgAGAAaGTCTCCACAACCCCTCTCGGACTCGAAGGTGAGCT
Danaus_plexippus108_21    GAGAtGGAGGAGAAggTCTCCTCACCCCTCTCAGGTCTCGAAGGTGAACT
Greta_oto70_9               GGAATGGAAgAGAAgGTCTCCTCGACCCTCTCAGGCCTTGAAGGTGAACT
Amathusia_phidippus114_17  GGAATGGAAgACAAaGTCTCCTCAaCCCTCTCGGTCTTGAAGGTGAACT
Morpho_peleides66_5        GGAATGGAGAGAAAaGTCTCTACTACCCCTGTGGCCTCGAAGGCAGAAct
BrintesiaB01                GGAATGGAAgACAAaGTCTCGTCCACCCCTCTCAGGGCTGGAAGGGCAGCT
Elymnias_casiphone121_20   GAGAwGGAAgACAAaAGTATCCTCACCCCTCTGGTCTTGAAGCTGAACT
Erebia_oemeEW24_7           gGaATGGAAgACAAaGTCTCCTCGACTCTCTGGCCTCGAAGGCAGCT

  [COI 71 bp]
Papilio_glaucus_69_3      taAagAtaTTgGaACATTATACTTTATTTGGAAATTGAGCAAGAACATTAGGAACCTTTAAAGTTAT
Hamearis84_13              ?????????????????????????????????????????TGAGCAGGAATAGTAGGAACATCATTAAAGATTAC
Libythea_celtis71_1         ?????????????????????????????????????????TGAGCAGGAATAGTAGGAACCTTCATTAAAGTCTAT
Danaus_plexippus108_21    ?????????????????????????????????????TGAGCAGGAATAGTAGGAACATCTTAAGTCTT
Greta_oto70_9               ?????????????????????????????????TGAGCAGGAATAGTAGGAACATCTTAAGTTAT
Amathusia_phidippus114_17  ?????????????????????????????????TGATCTGGAAATAGTAGGAACATCCCTCAGTCTTA
Morpho_peleides66_5        ?????????????????????????????????TGAGCCGGTATAATTGGTACATCCCTAAGTCTTA
BrintesiaB01                ?????????????????????????????????TGAGCAGGTATAGTAGGAACATCTCTAGTTAA
Elymnias_casiphone121_20   ?????????????????????????????TGATCAGGAATAGTAGGAACCTCCCTCAGTCTTA
Erebia_oemeEW24_7           ?????????????????????????TGAGCAGGTATAGTAGGTACTCCCTTAGTCTTA
;
end;
```