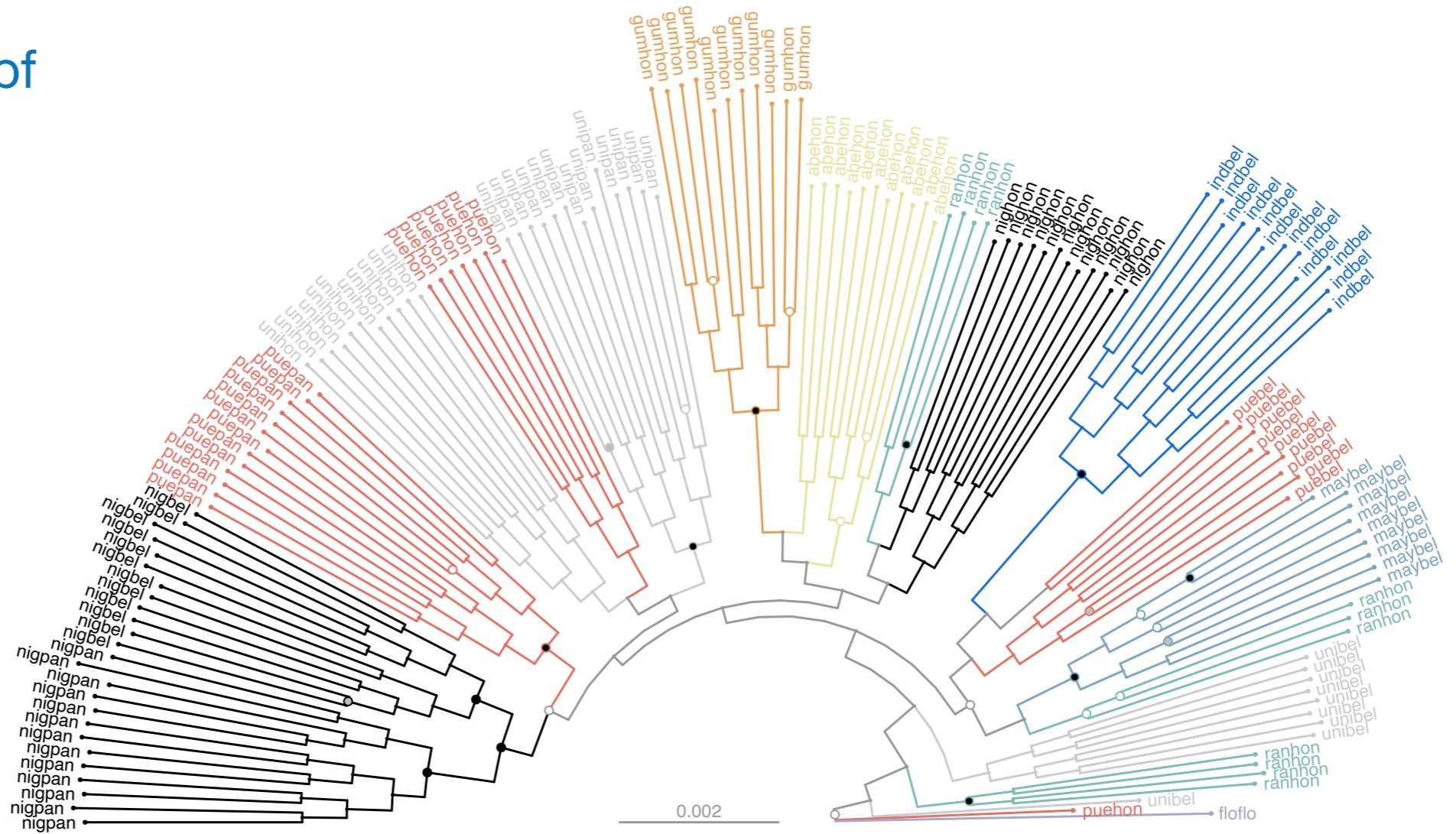


# Introduction to phylogenetics

Marine ecological genetics | July 11, 2025

Martin Helmkampf



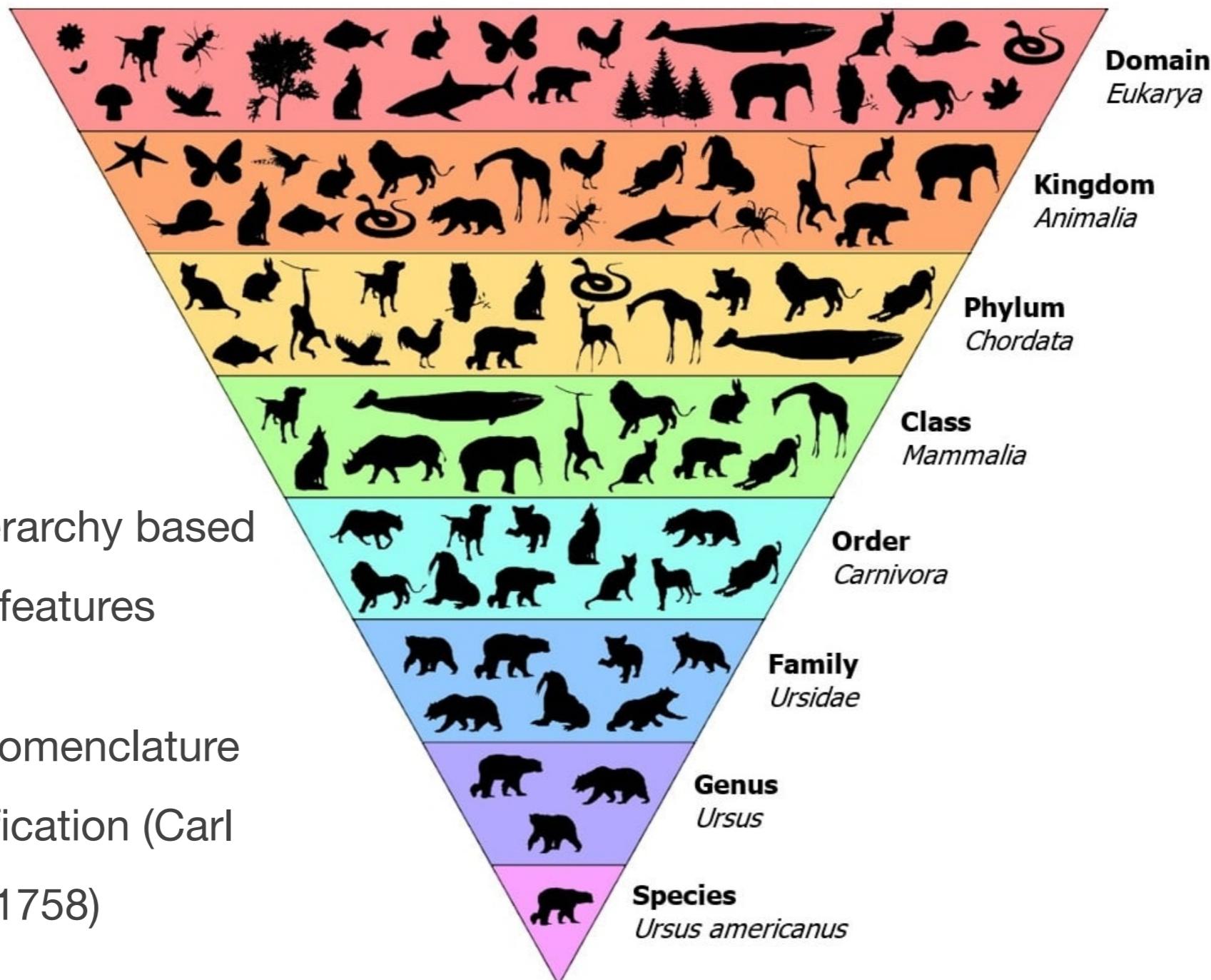
# Basic concepts and applications

# How can we describe the immense diversity of life?



[resonanceglobal.com](http://resonanceglobal.com)

# Organisms can be grouped by similarity



# Descent with modification ...

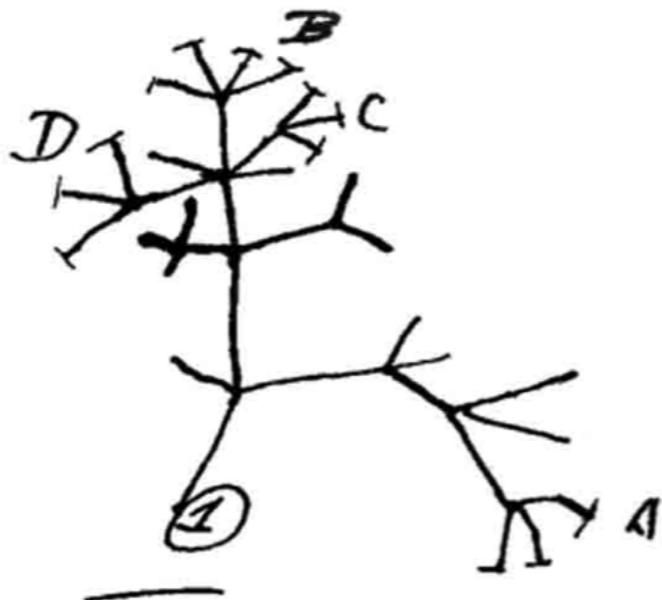
- All organisms are related through the evolutionary process of descent with modification (**splitting of lineages + divergence**)
- This **branching pattern** generates nested hierarchies of similarity over evolutionary time

# ... is best represented by a tree

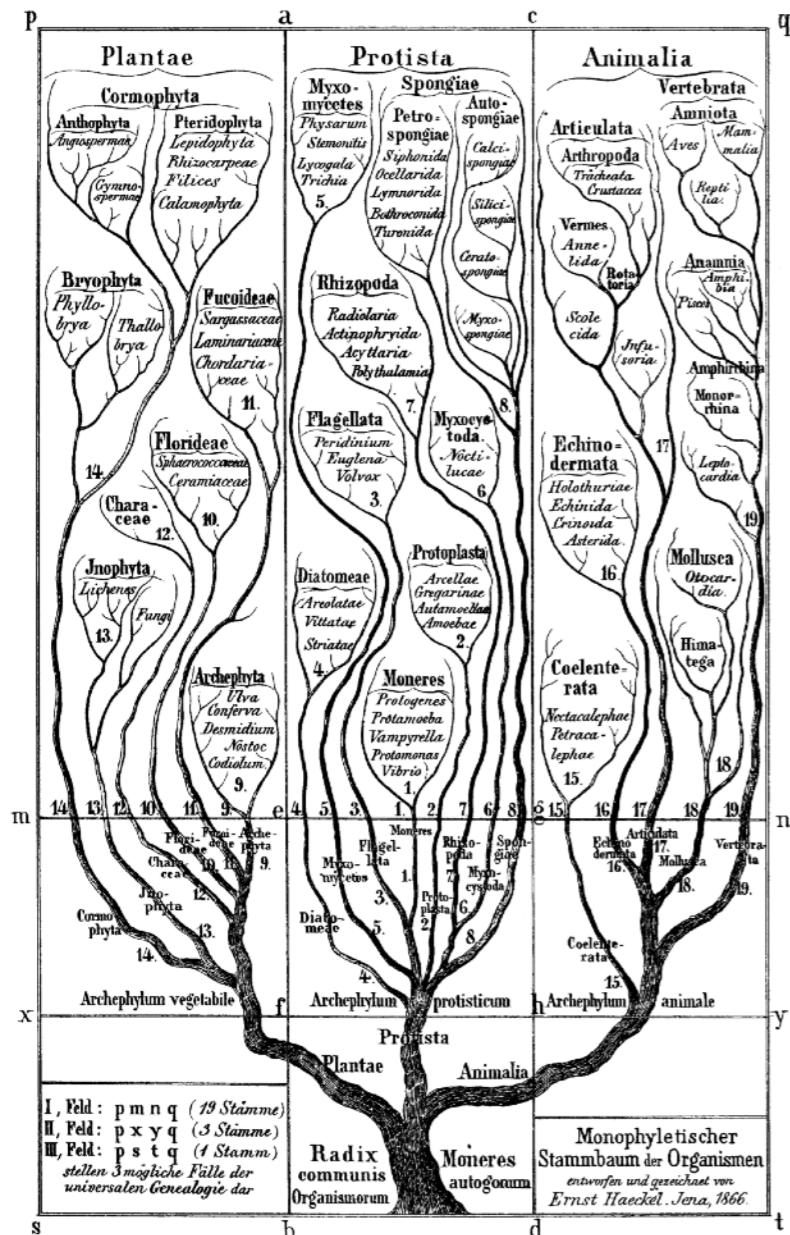
“The affinities of all the beings of the same class have sometimes been represented by a great tree ... The green and budding twigs may represent existing species; and those produced during each former year may represent the long succession of extinct species”

*I think*

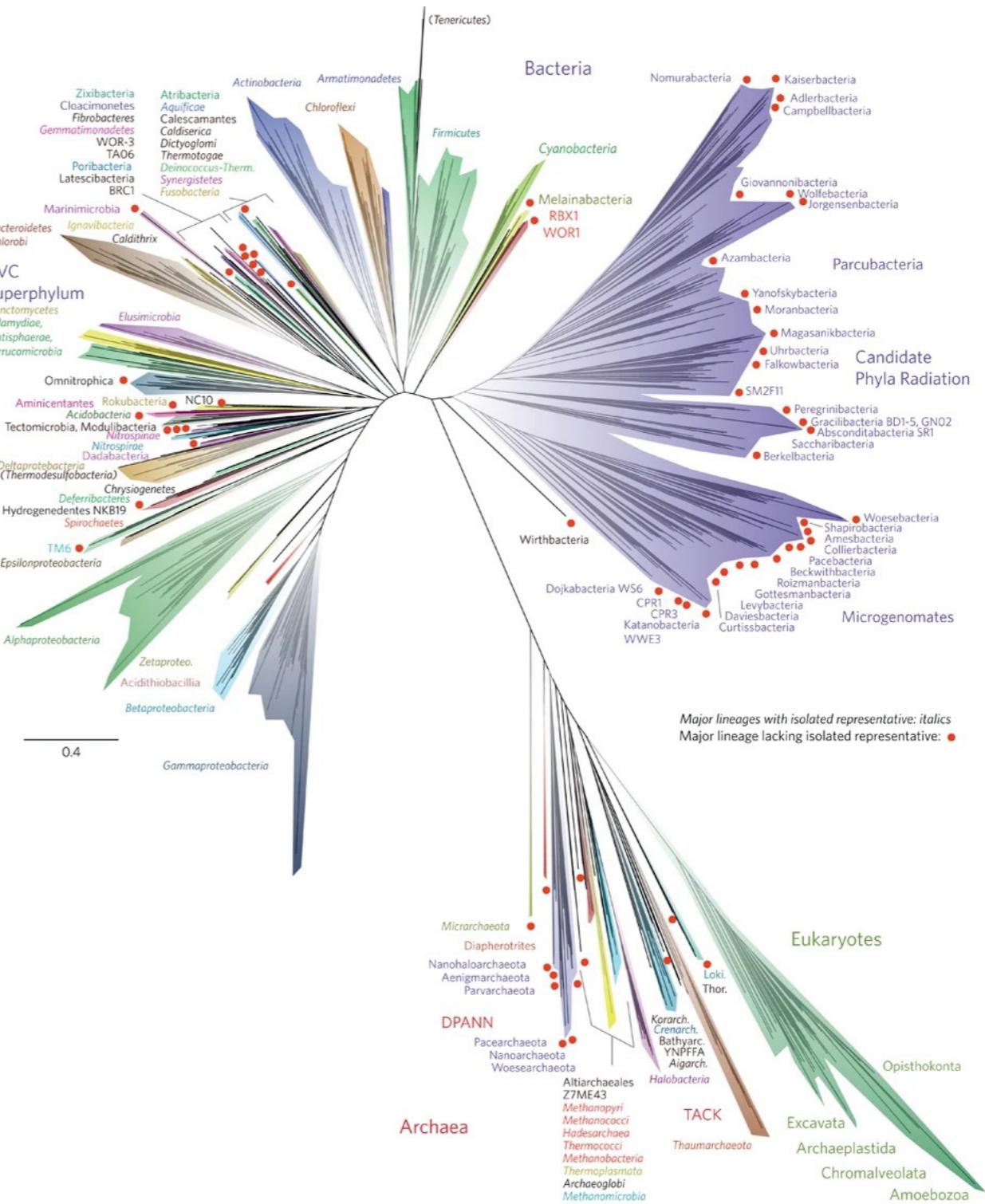
— C. Darwin, 1859



# The tree of life



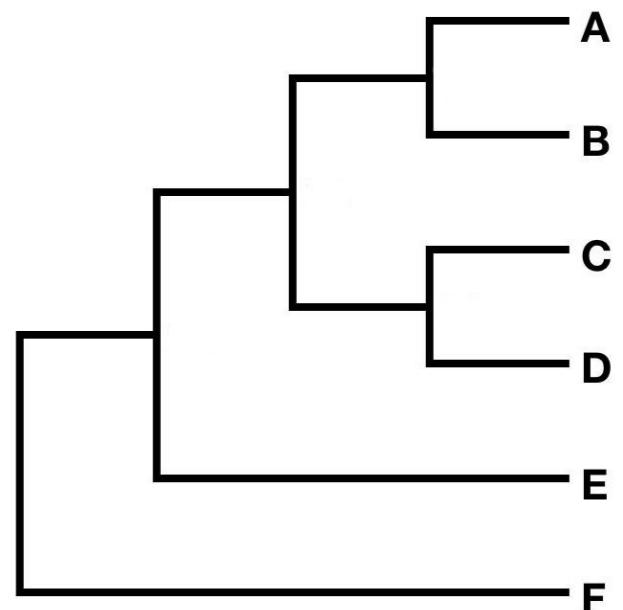
E. Haeckel, 1866



Hug et al. 2016, *Nature Microbiology*

# What is a phylogenetic tree / phylogeny?

- Graphical representation of the **evolutionary relationships** of a group of organisms or genes
- Shows the **history of descent** from common ancestors, in particular the order and timing of branching events
- Typically inferred by comparing **DNA sequences**
- Almost always a **hypothesis** subject to change



# Why study phylogenies

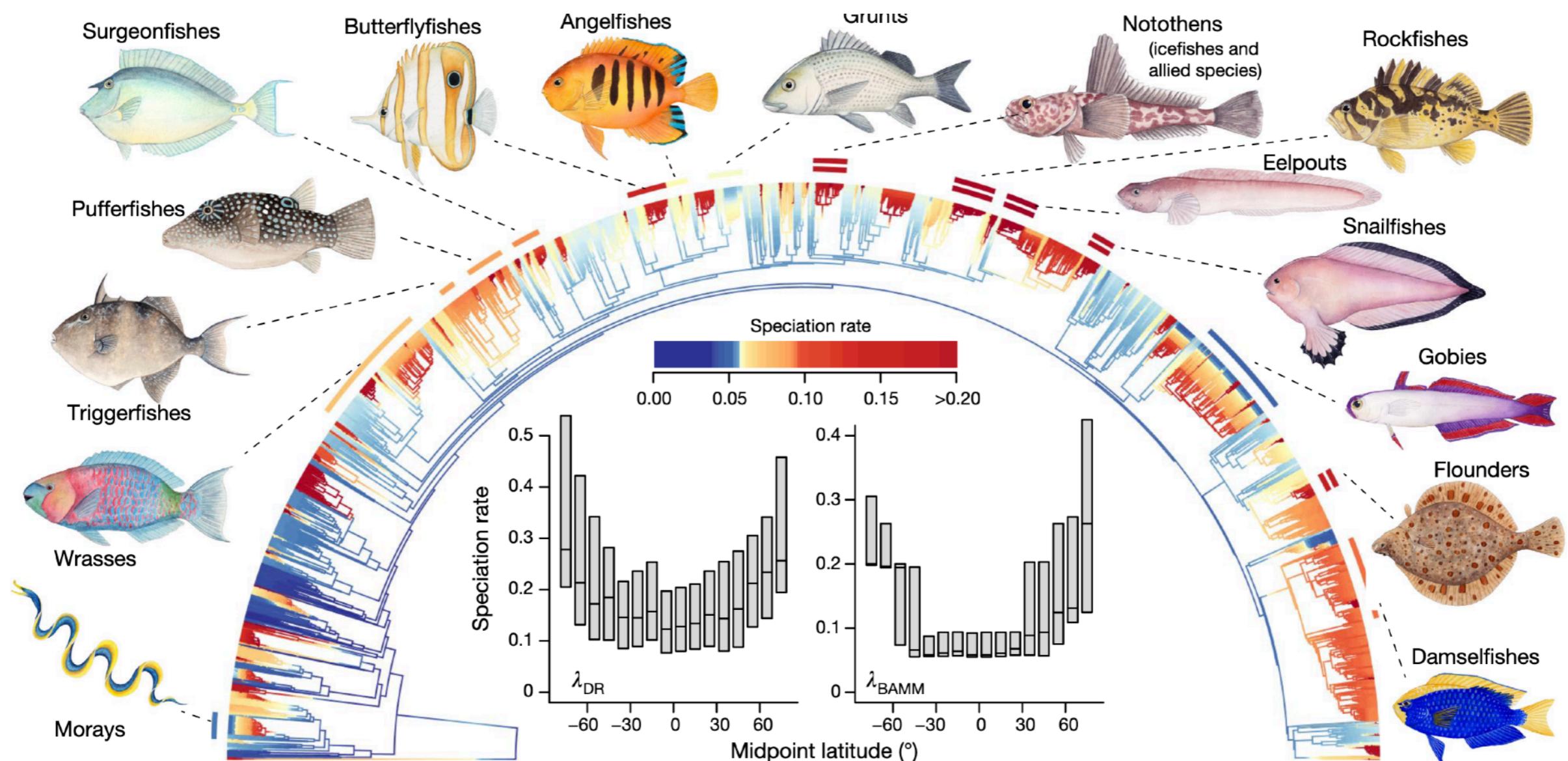
- Reconstruct past evolutionary events and the **history of life**
- Study how **traits, genes, and genomes** evolve
- Provide evolutionary framework to **organize biodiversity** (systematics)
- **Identify species** and clarify species boundaries (taxonomy)
- Highlight lineages with unique evolutionary history (conservation)

# Why study phylogenies

- Reconstruct past evolutionary events and the **history of life**
- Study how **traits, genes, and genomes** evolve
- Provide evolutionary framework to **organize biodiversity** (systematics)
- **Identify species** and clarify species boundaries (taxonomy)
- Highlight lineages with unique evolutionary history (conservation)

Phylogenies are central to understanding the patterns (relationships) and processes (mechanisms) of evolution, revealing how organisms came to be the way they are today

# Application examples



5000+ marine fishes | 27 genes | maximum likelihood analysis

Rabosky et al. 2018, *Nature*

# Application examples

## Nuclear markers confirm taxonomic status and relationships among highly endangered and closely related right whale species

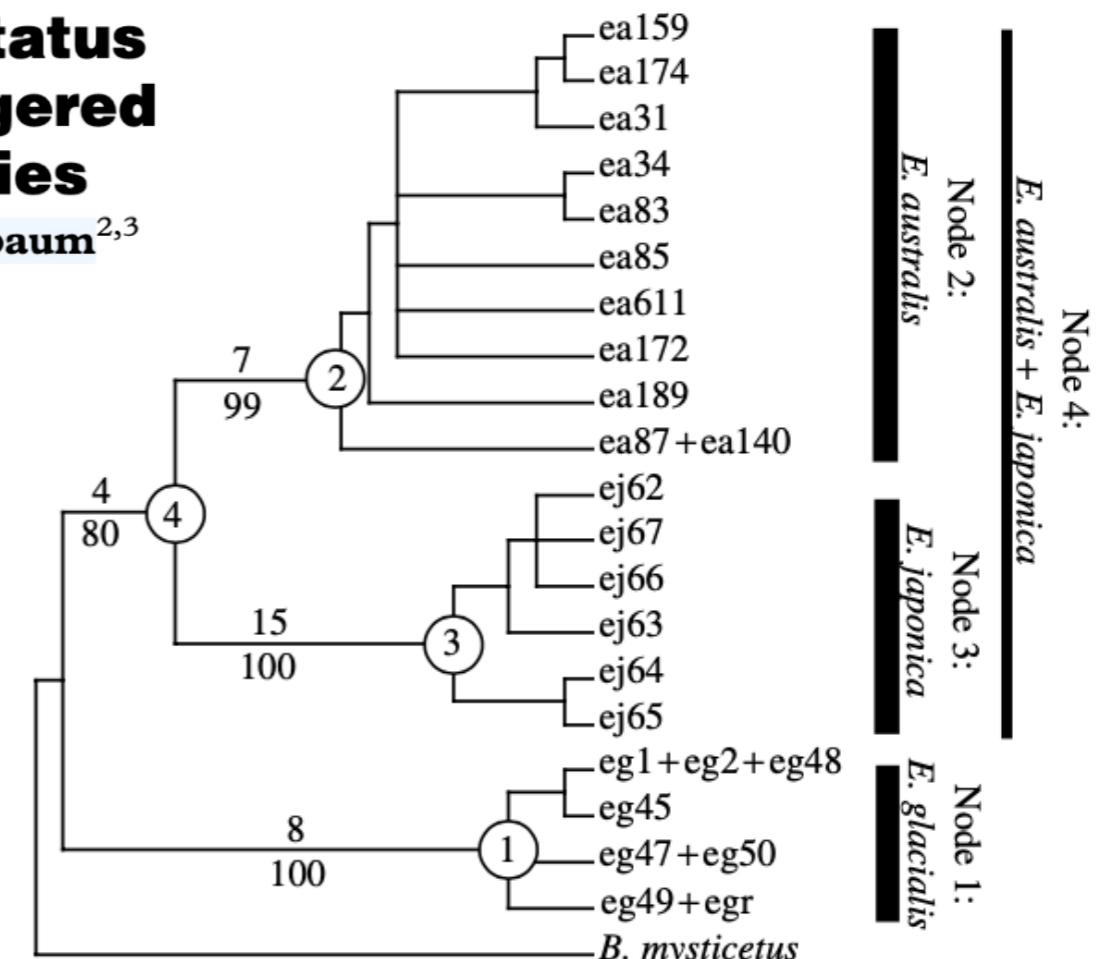
C.A. Gaines<sup>1,2,3,\*</sup>, M.P. Hare<sup>4</sup>, S.E. Beck<sup>4</sup> and H.C. Rosenbaum<sup>2,3</sup>

25 right whale specimens

14 genes



maximum parsimony analysis



Gaines et al. 2005, Proc. R. Soc. B, image by Frédérique Lucas

# Application examples

The NEW ENGLAND JOURNAL of MEDICINE

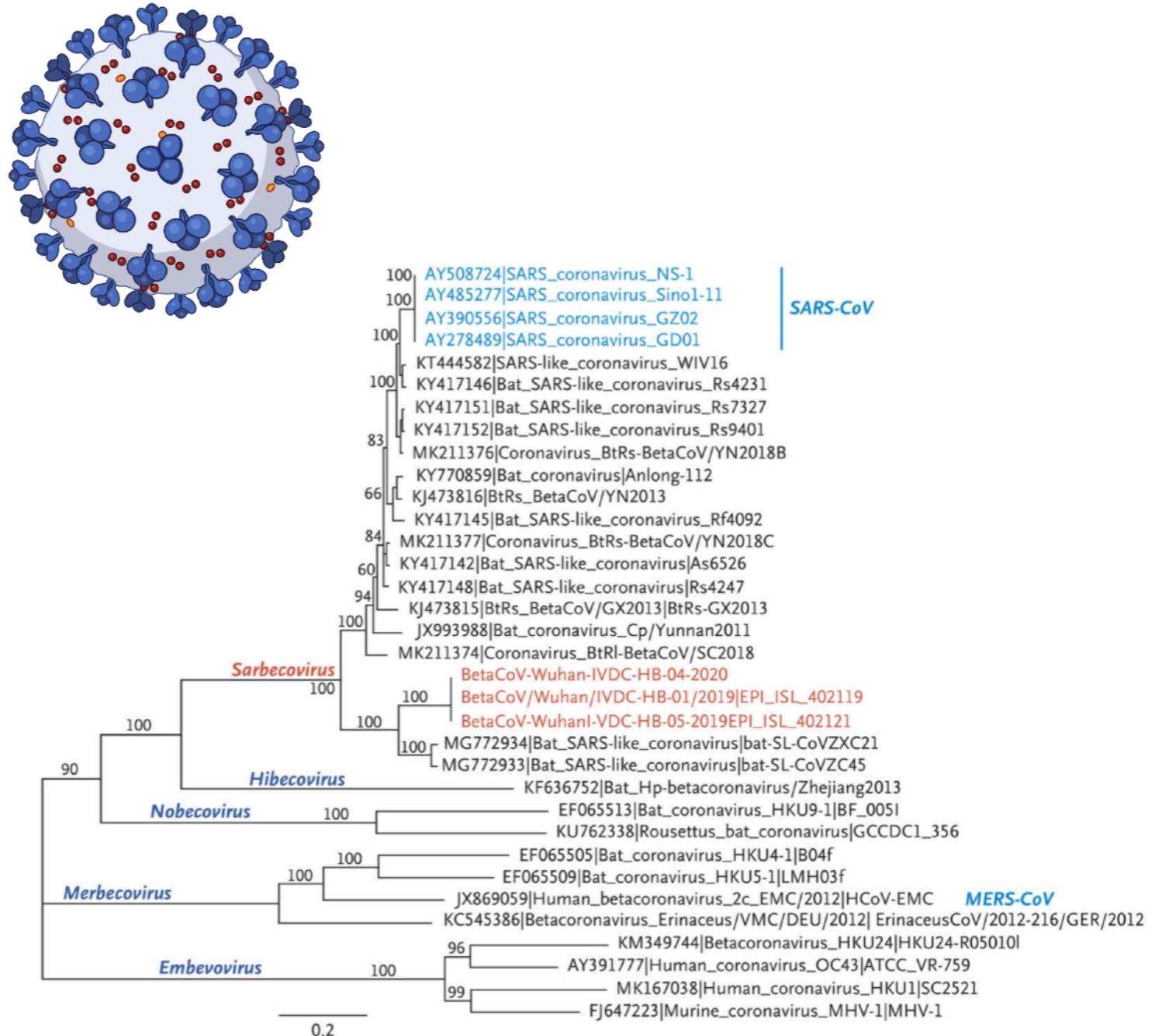
## BRIEF REPORT

### A Novel Coronavirus from Patients with Pneumonia in China, 2019

Na Zhu, Ph.D., Dingyu Zhang, M.D., Wenling Wang, Ph.D., Xingwang Li, M.D., Bo Yang, M.S., Jingdong Song, Ph.D., Xiang Zhao, Ph.D., Baoying Huang, Ph.D., Weifeng Shi, Ph.D., Roujian Lu, M.D., Peihua Niu, Ph.D., Faxian Zhan, Ph.D., Xuejun Ma, Ph.D., Dayan Wang, Ph.D., Wenbo Xu, M.D., Guizhen Wu, M.D., George F. Gao, D.Phil., and Wenjie Tan, M.D., Ph.D., for the China Novel Coronavirus Investigating and Research Team

#### SUMMARY

In December 2019, a cluster of patients with pneumonia of unknown cause was linked to a seafood wholesale market in Wuhan, China. A previously unknown betacoronavirus was discovered through the use of unbiased sequencing in samples from patients with pneumonia. Human airway epithelial cells were used to isolate a novel coronavirus, named 2019-nCoV, which formed a clade within the subgenus sarbecovirus, Orthocoronavirinae subfamily. Different from both MERS-CoV and SARS-CoV, 2019-nCoV is the seventh member of the family of coronaviruses that infect humans. Enhanced surveillance and further investigation are ongoing. (Funded by the National Key Research and Development Program of China and the National Major Project for Control and Prevention of Infectious Disease in China.)

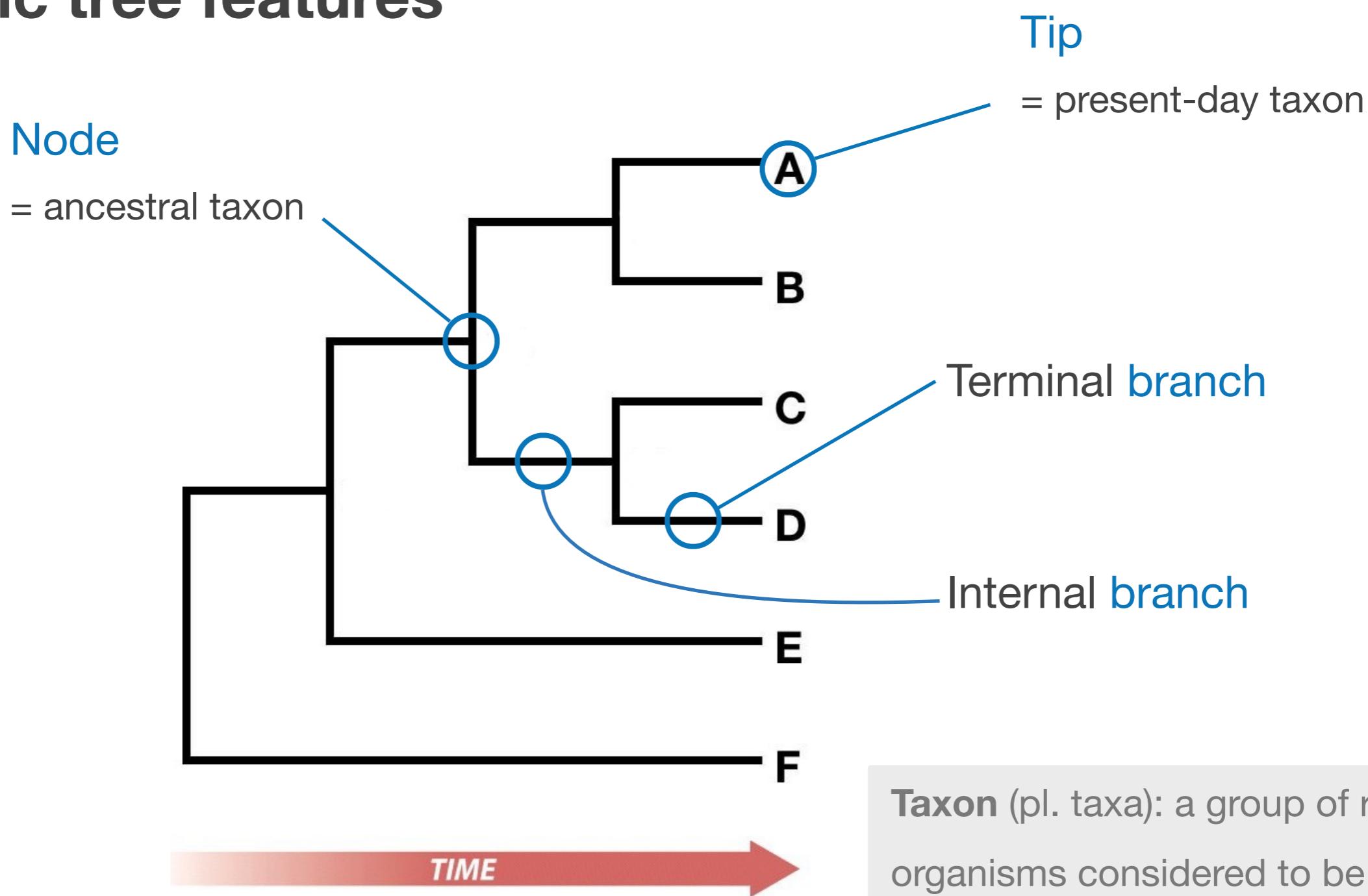


### Whole genomes of coronaviruses | maximum likelihood analysis

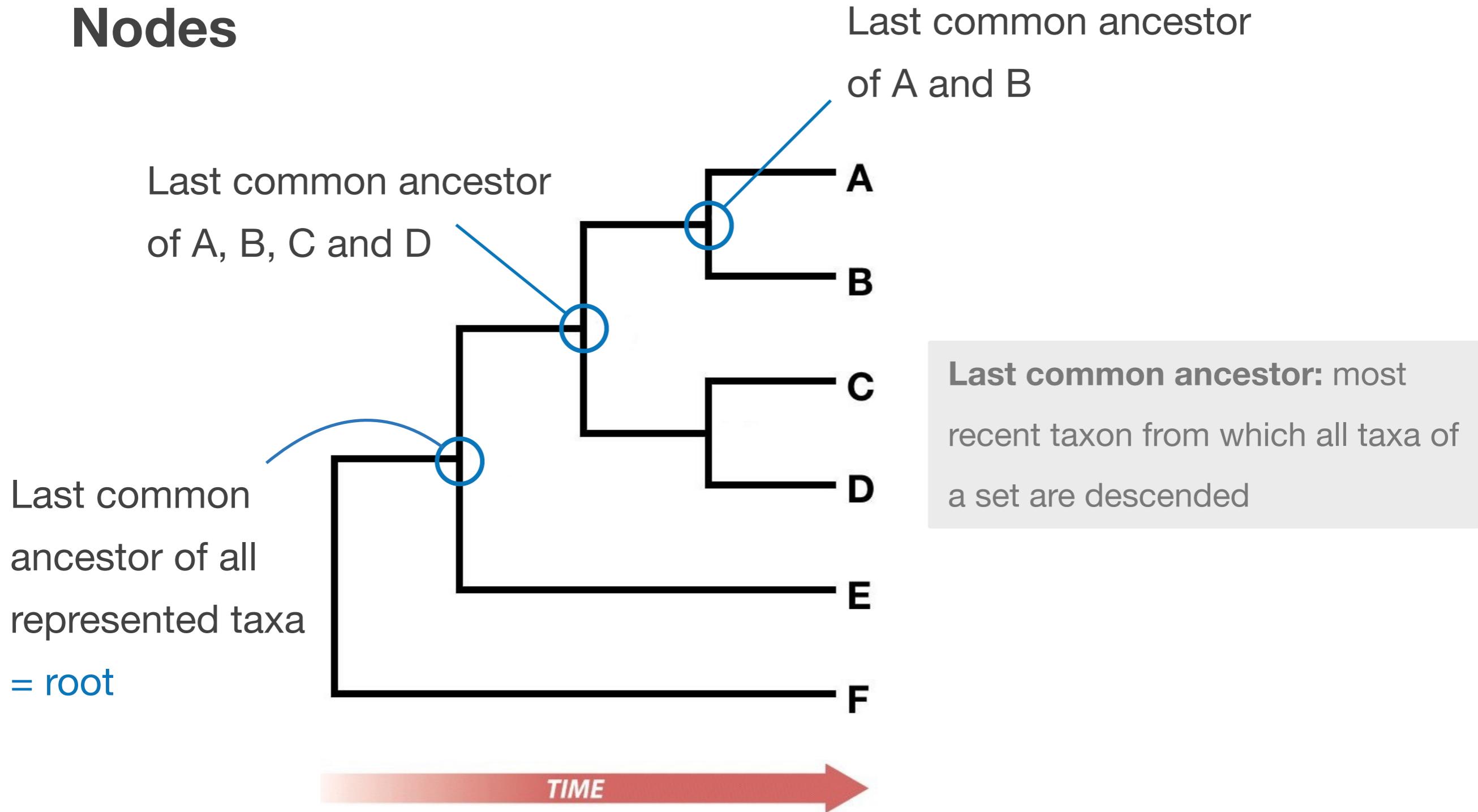
Zhu et al. 2020, *N. Engl. J. Med.*, image by Jerry Gu / BioRender

# Reading phylogenetic trees

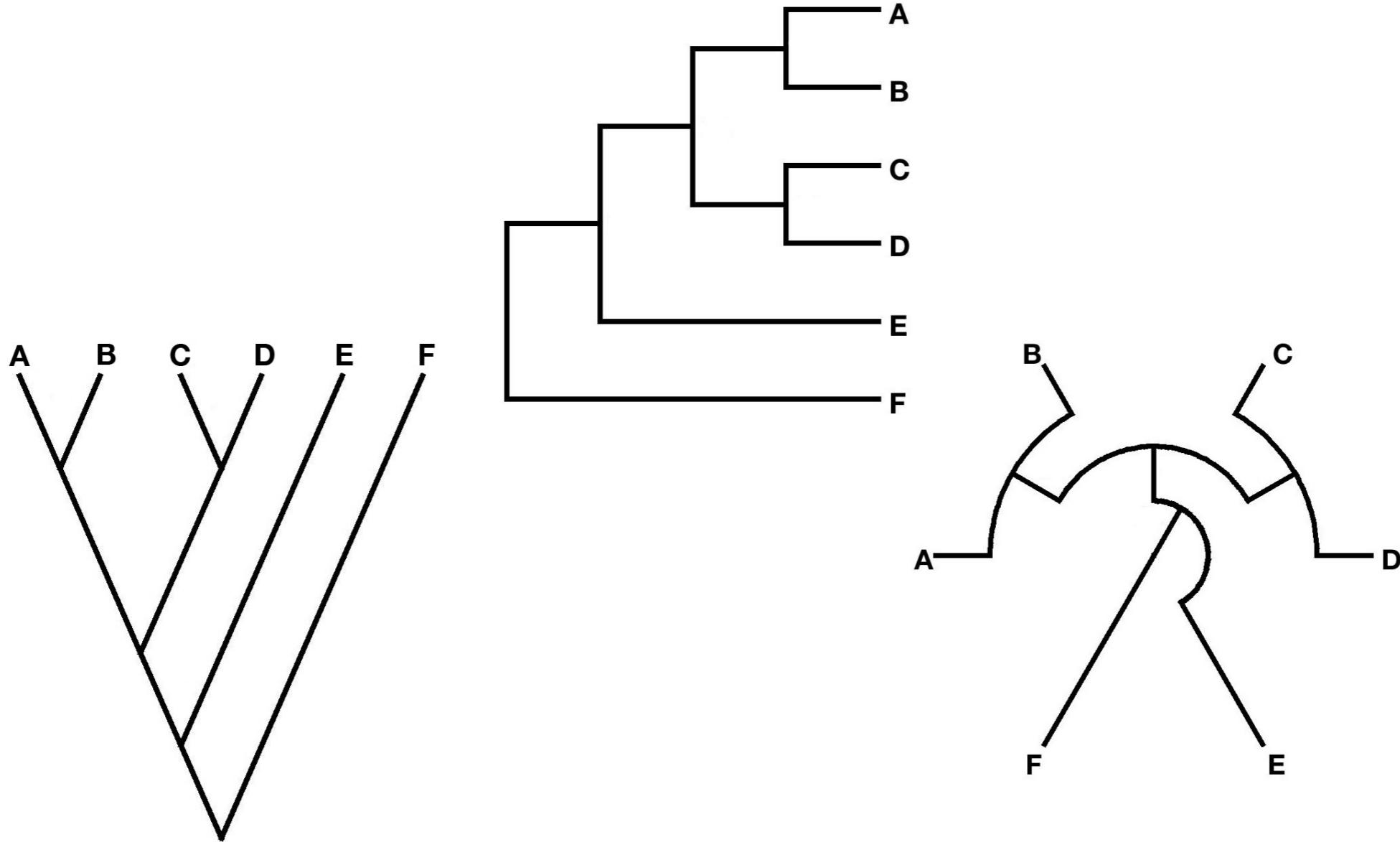
# Basic tree features



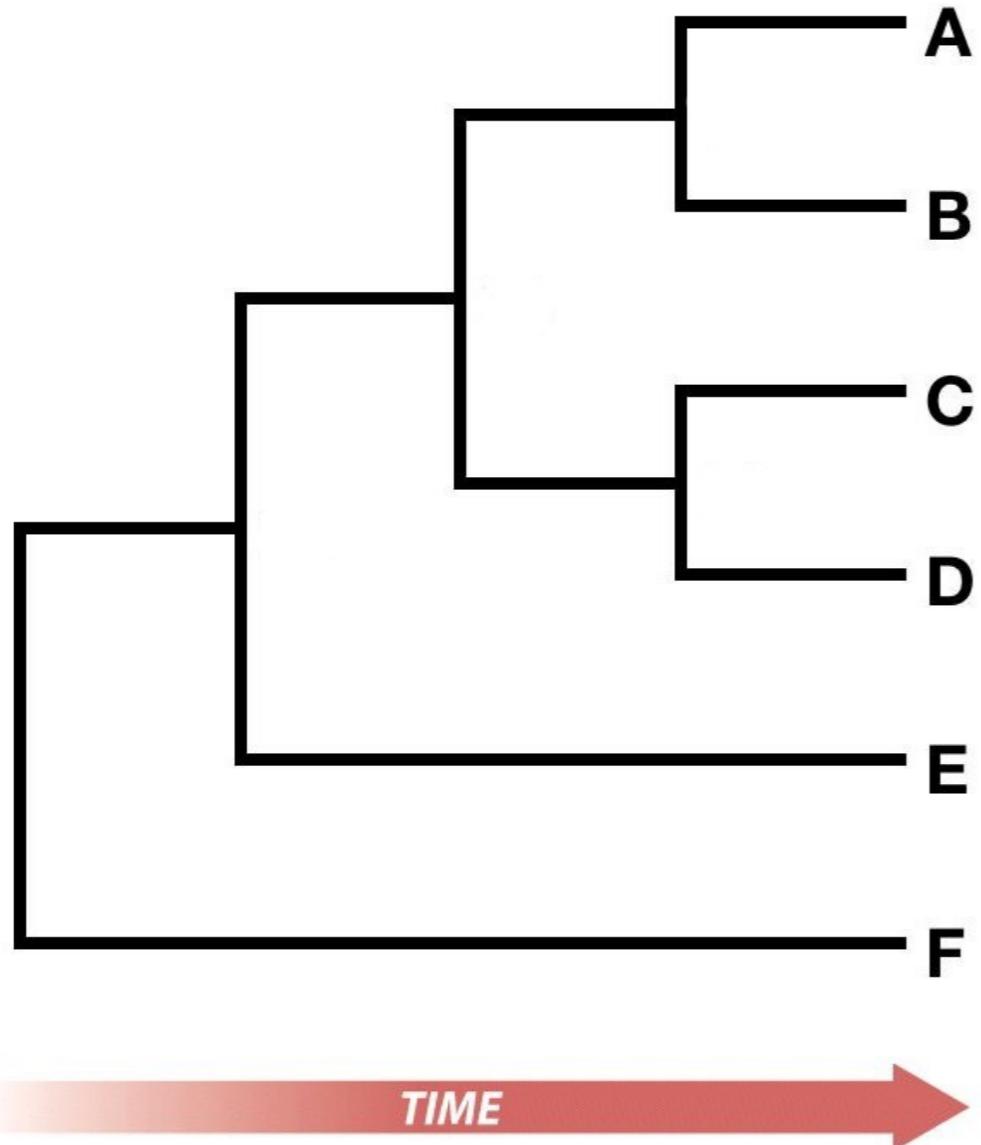
# Nodes



# Trees come in many shapes

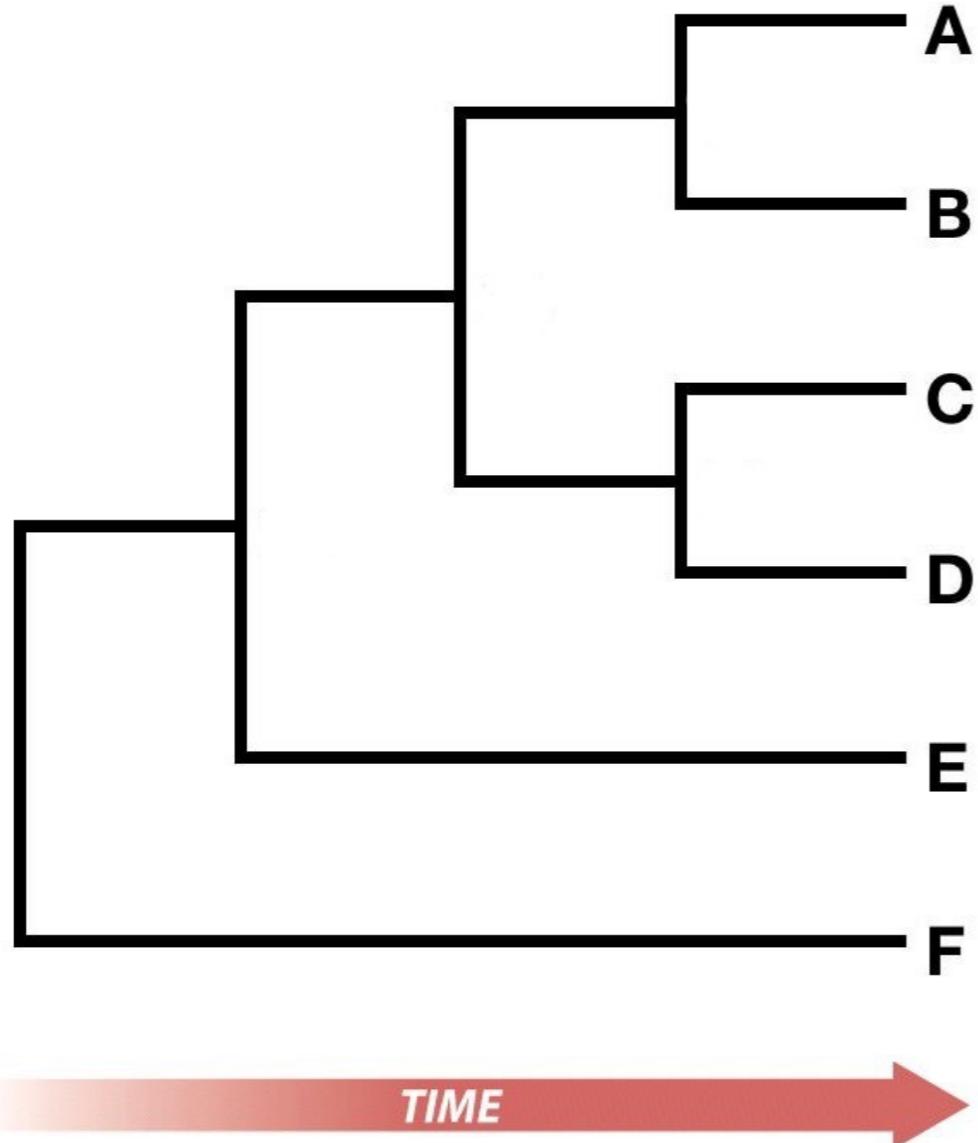


# ... what matters is the branching pattern



The **topology** or branching order represents degrees of relatedness

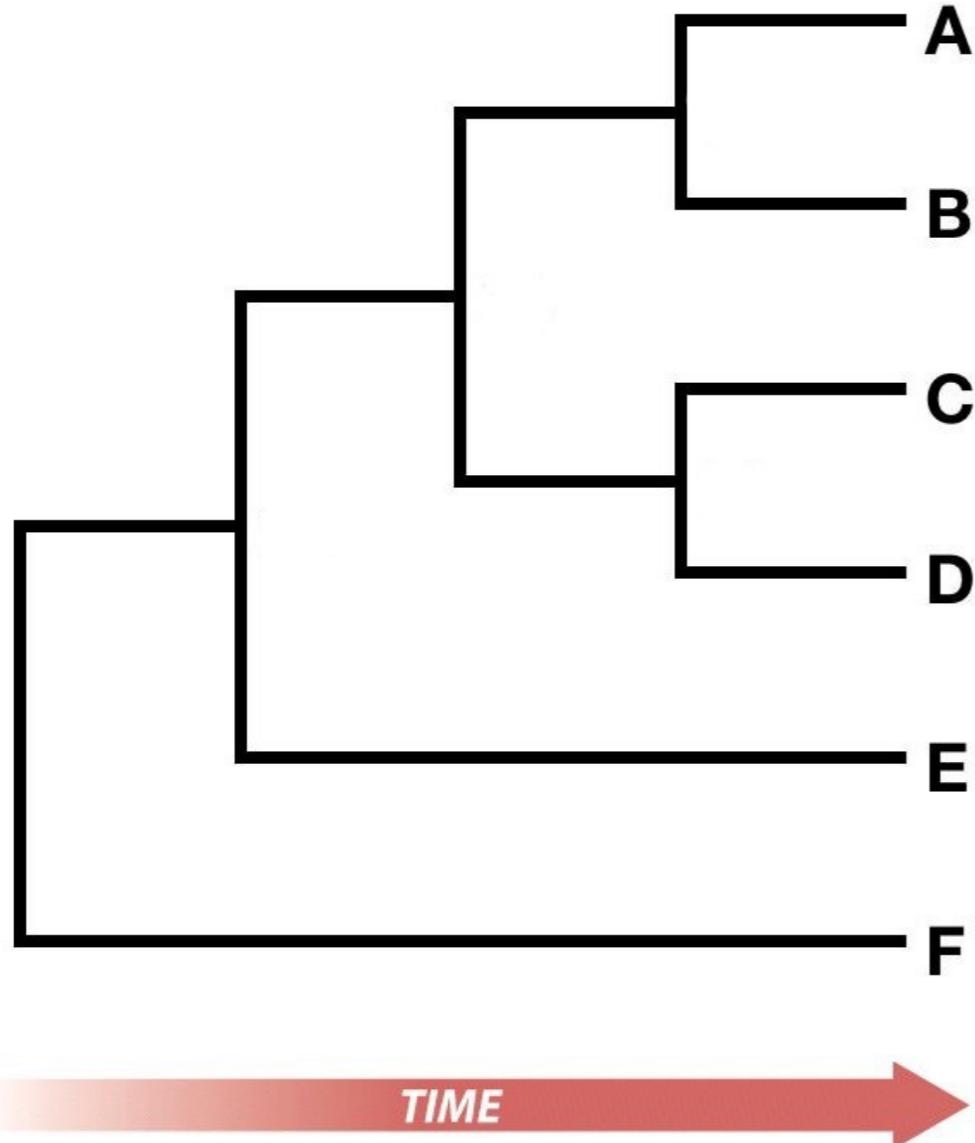
# ... what matters is the branching pattern



The **topology** or branching order represents degrees of relatedness

Is D more closely related to A or to E?

# ... what matters is the branching pattern

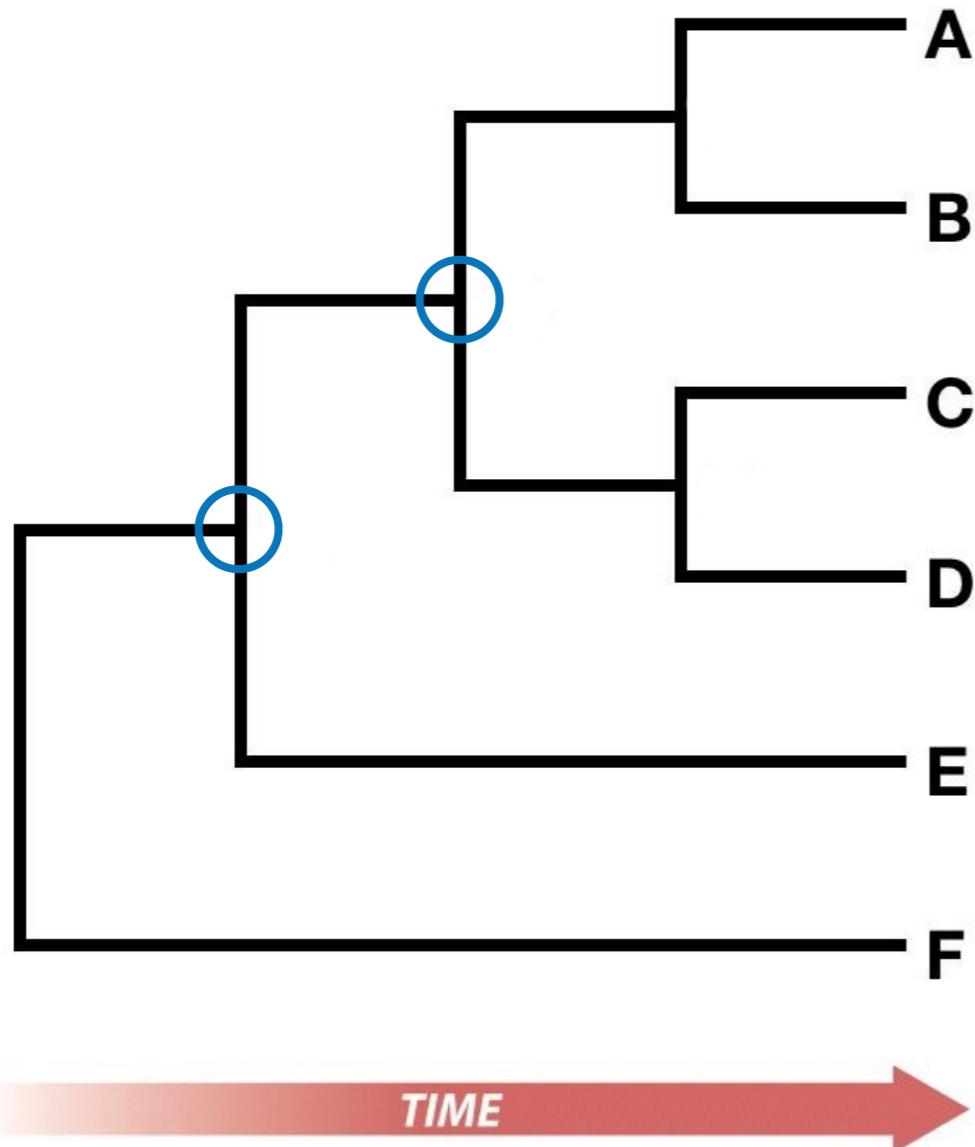


The **topology** or branching order represents degrees of relatedness

Is D more closely related to A or to E?

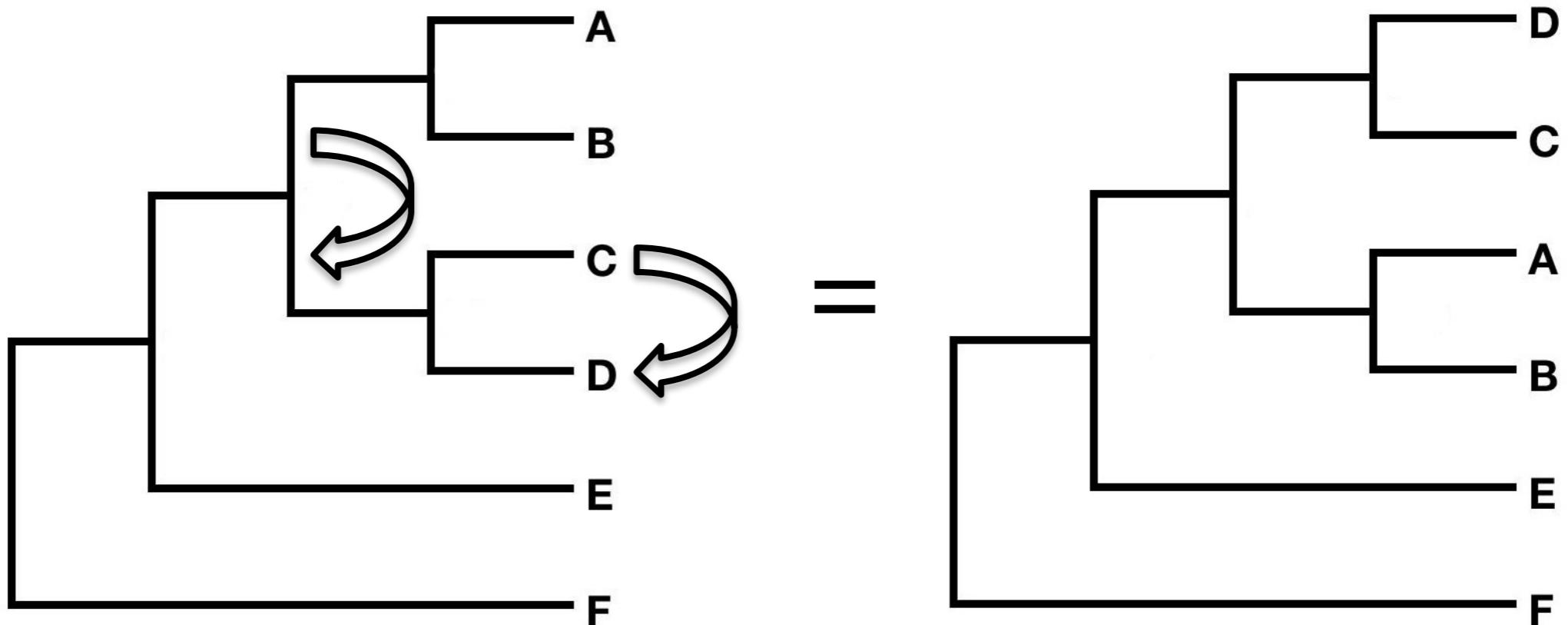
Which common ancestor is more recent?

# ... what matters is the branching pattern

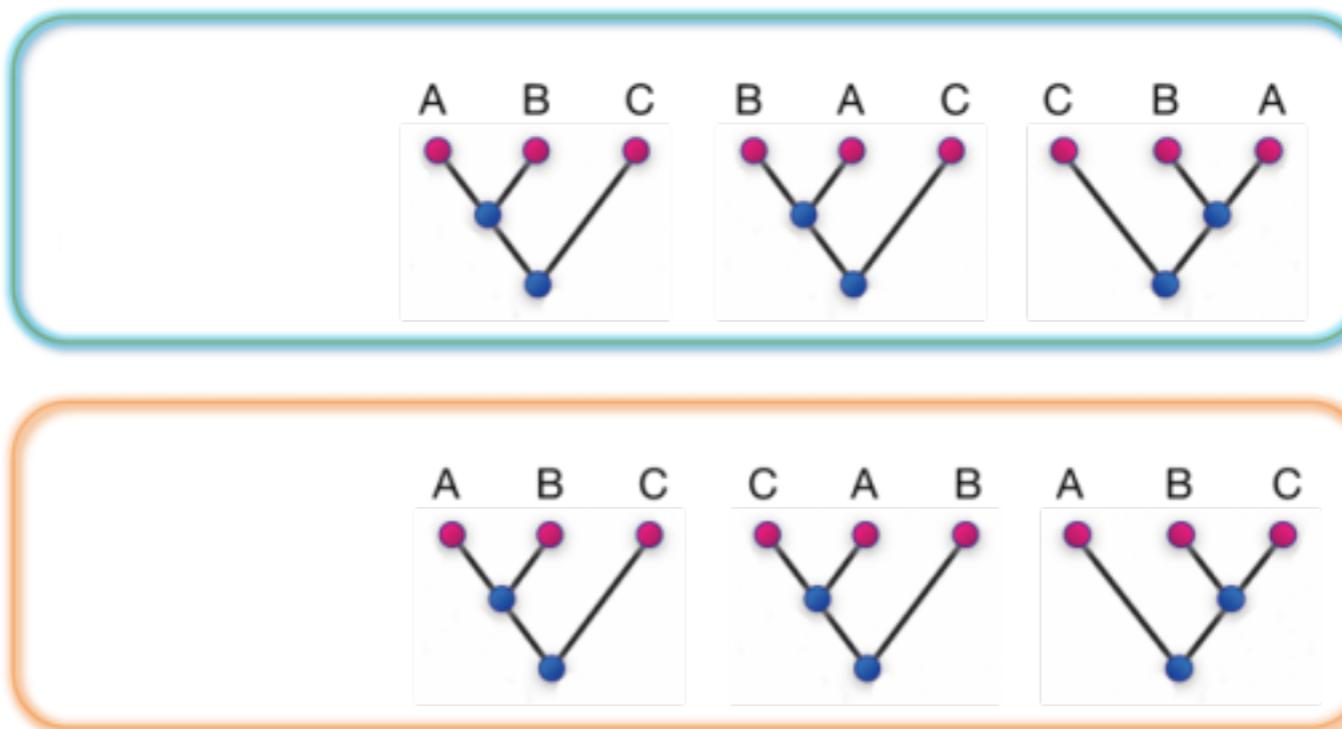


D is more closely related to A  
because A and D share a more  
recent common ancestor than  
either does with E

# Node rotation is arbitrary



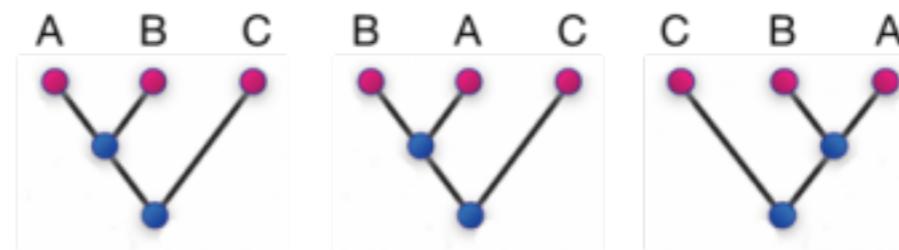
# Same or different topology?



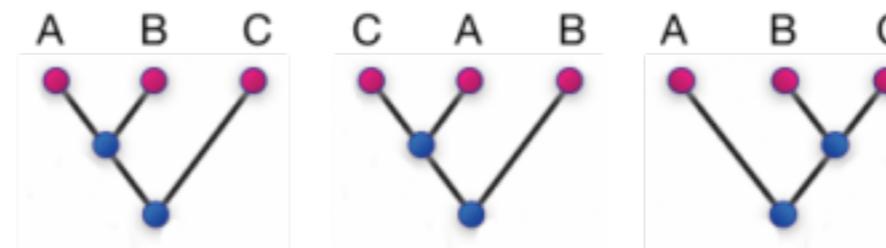
Adapted from EMBL-EBI, Creative Commons

# Same or different topology?

Same

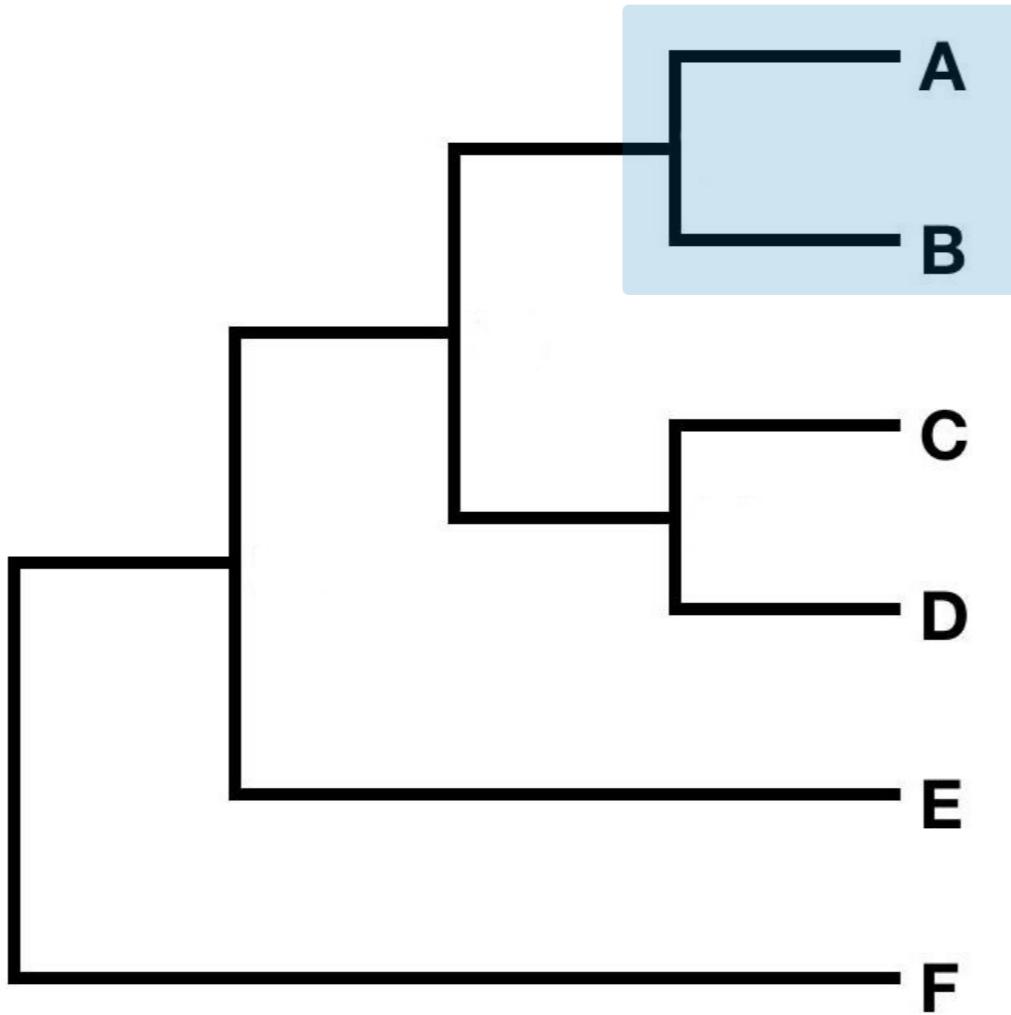


Different



Adapted from EMBL-EBI, Creative Commons

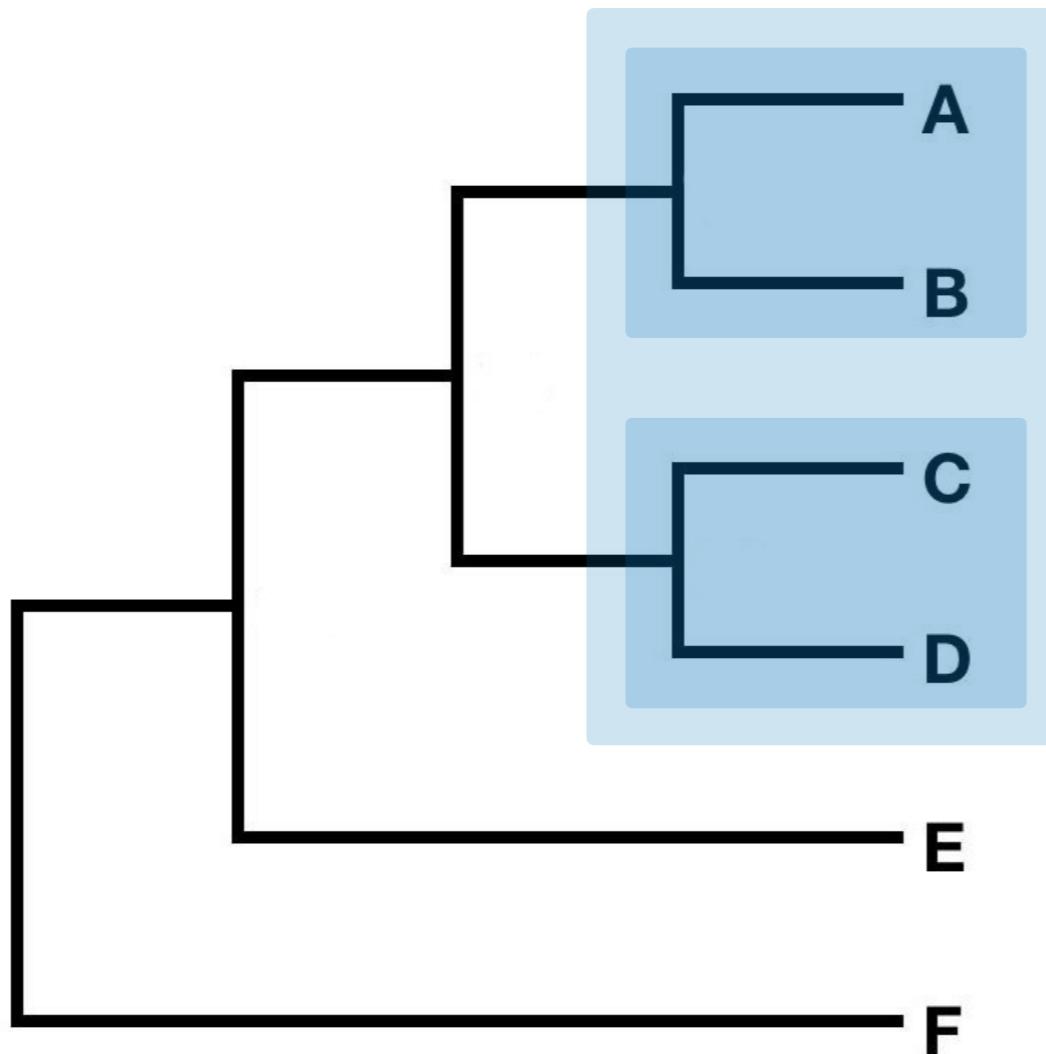
# Sister taxa



A and B are sister taxa

**Sister taxon:** a taxon's most closely related taxon, split from the same ancestral node

# Sister taxa

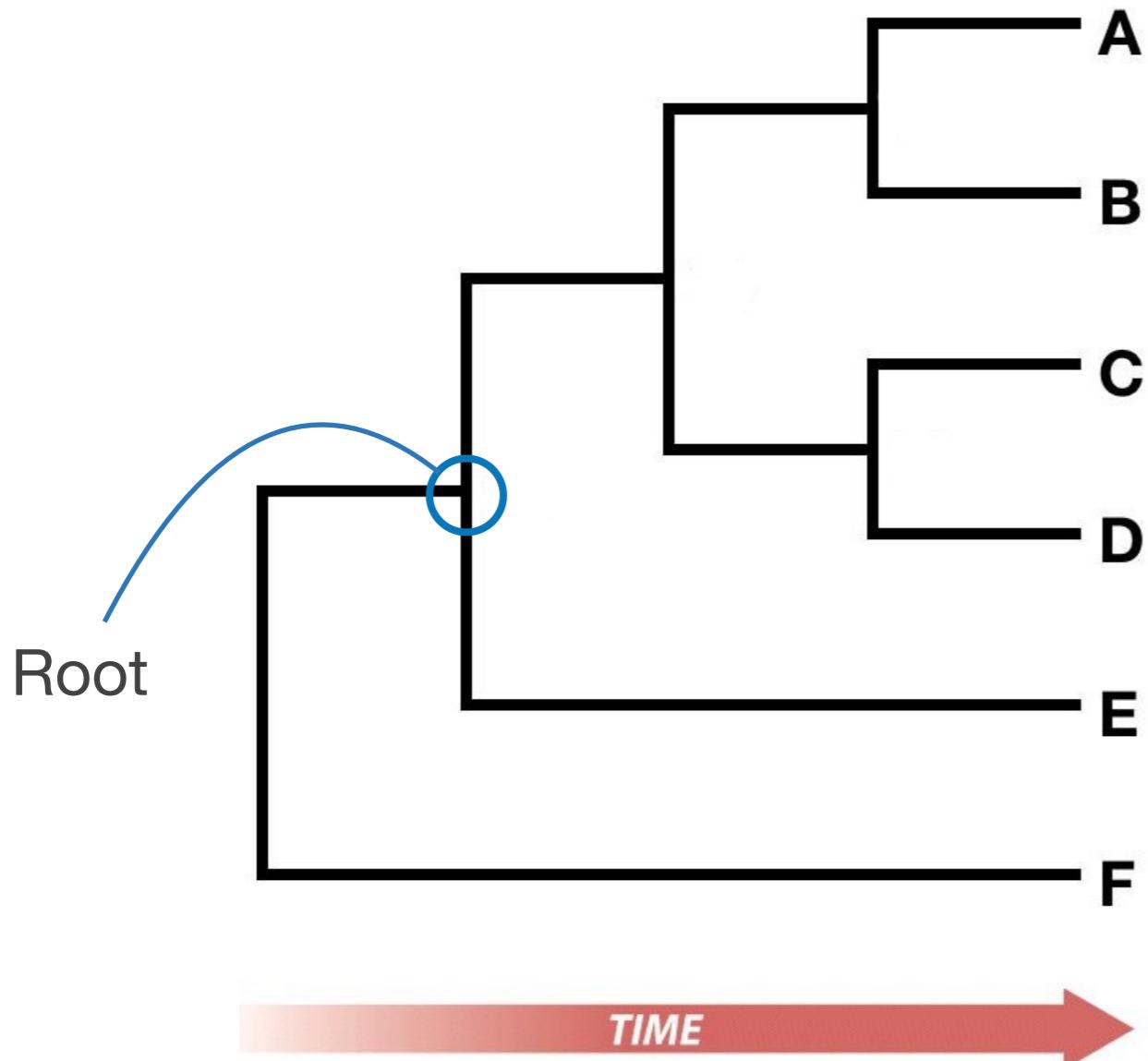


A and B are sister taxa

A+B's sister taxon is C+D

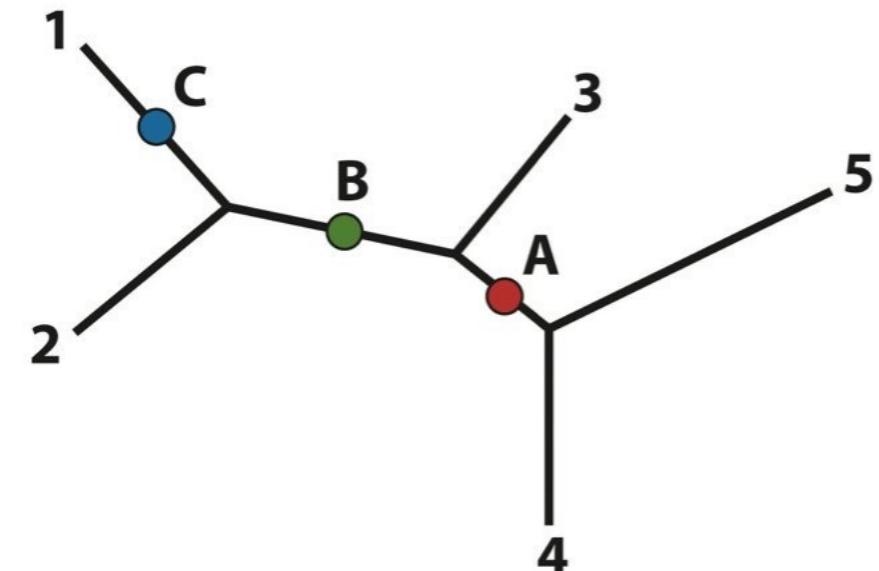
**Sister taxon:** a taxon's most closely related taxon, split from the same ancestral node

# Root and outgroup



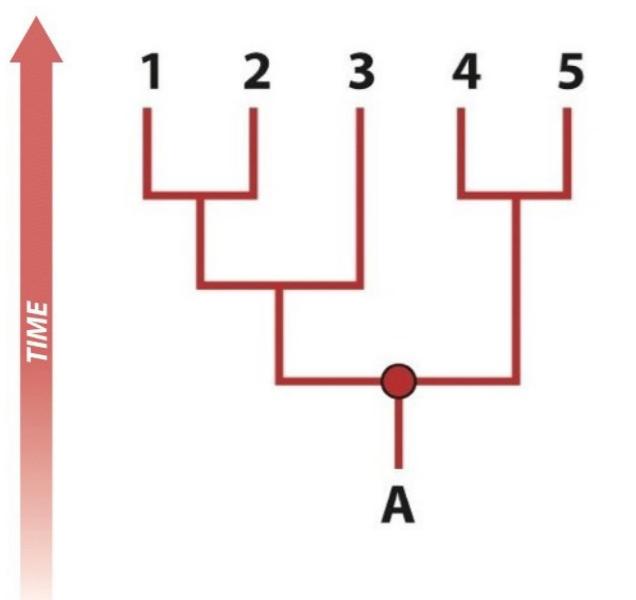
**Root:** node at base of the tree from which all other taxa are derived, indicating direction of time

# Root and outgroup



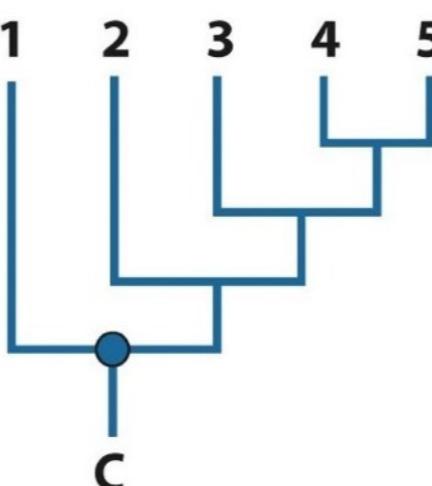
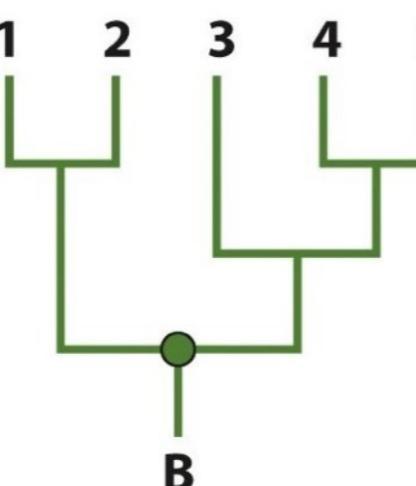
Unrooted tree

Ancestor, direction of evolution unknown



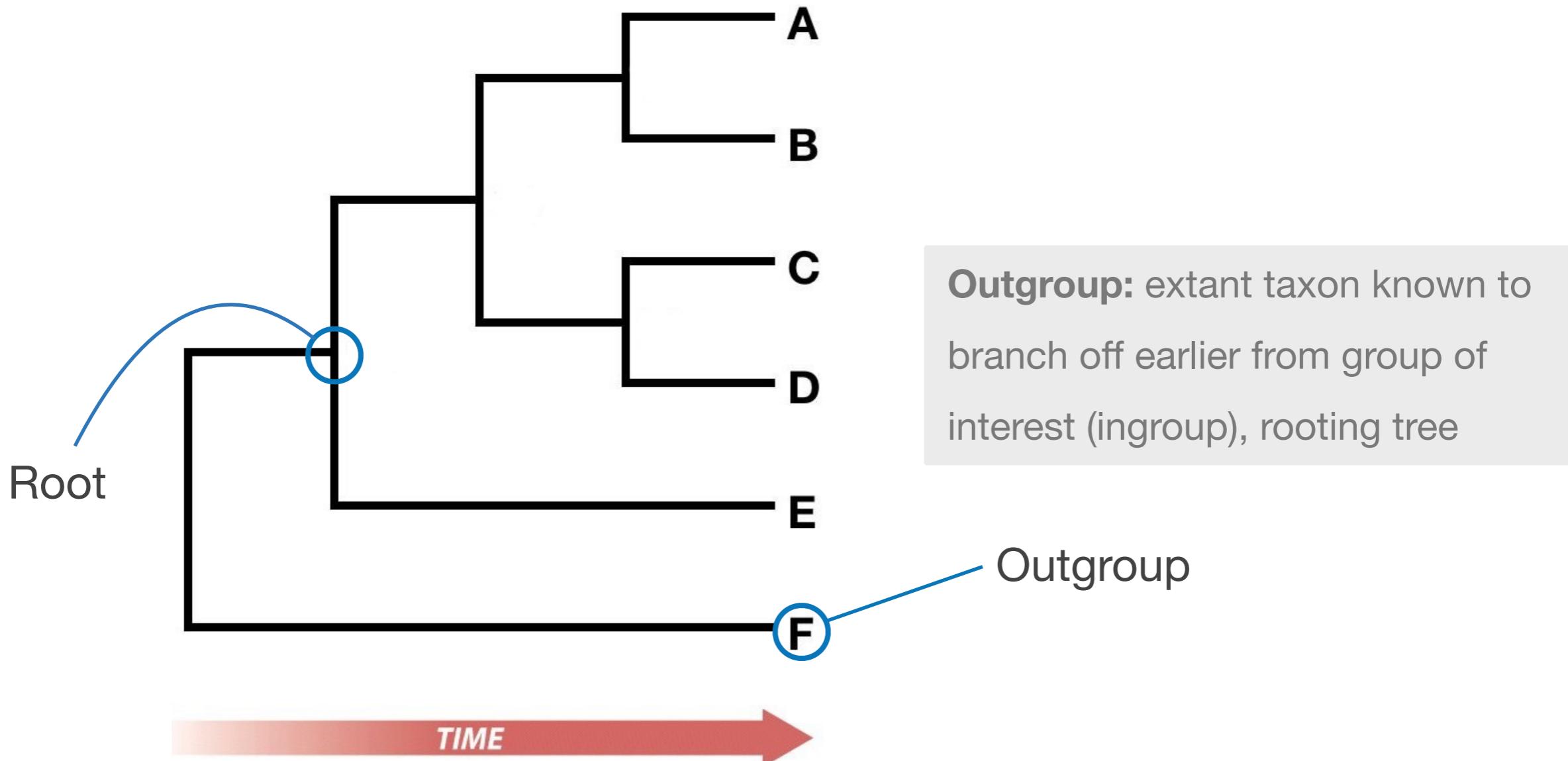
Rooted trees

Root indicates  
ancestral taxon,  
direction of evolution



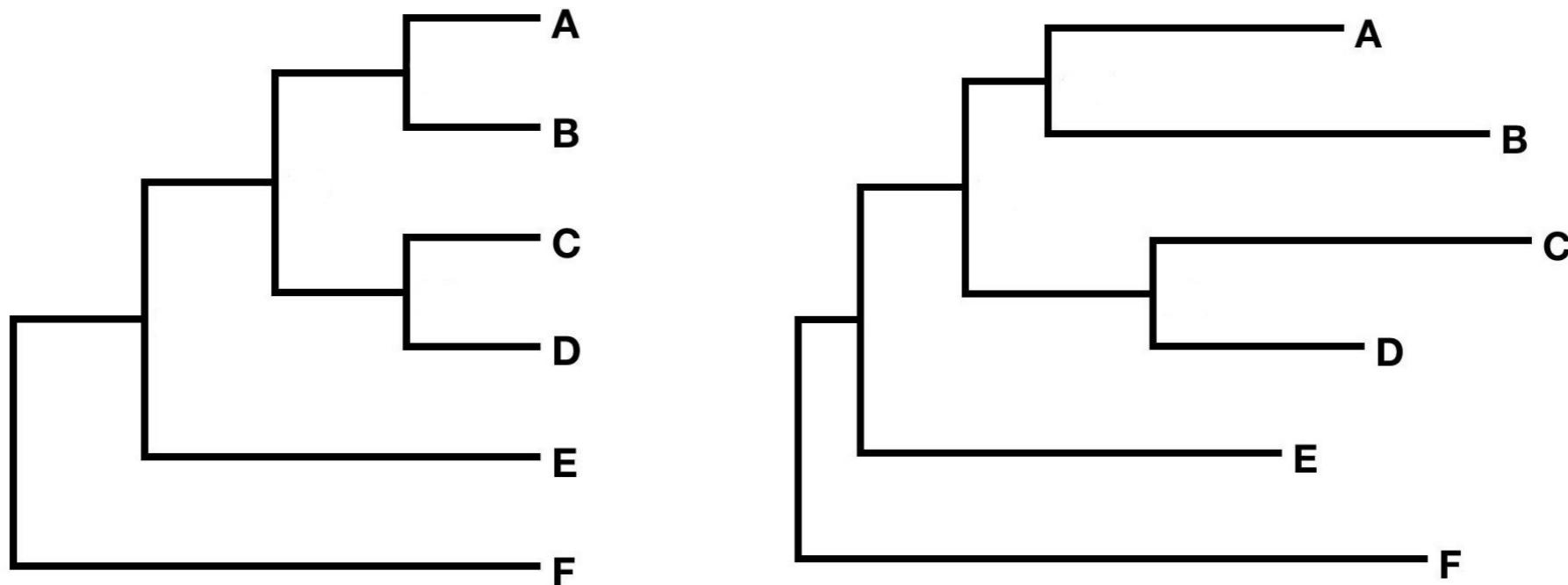
Bergstrom and Dugatkin, Evolution (Norton & Co., 2012)

# Root and outgroup



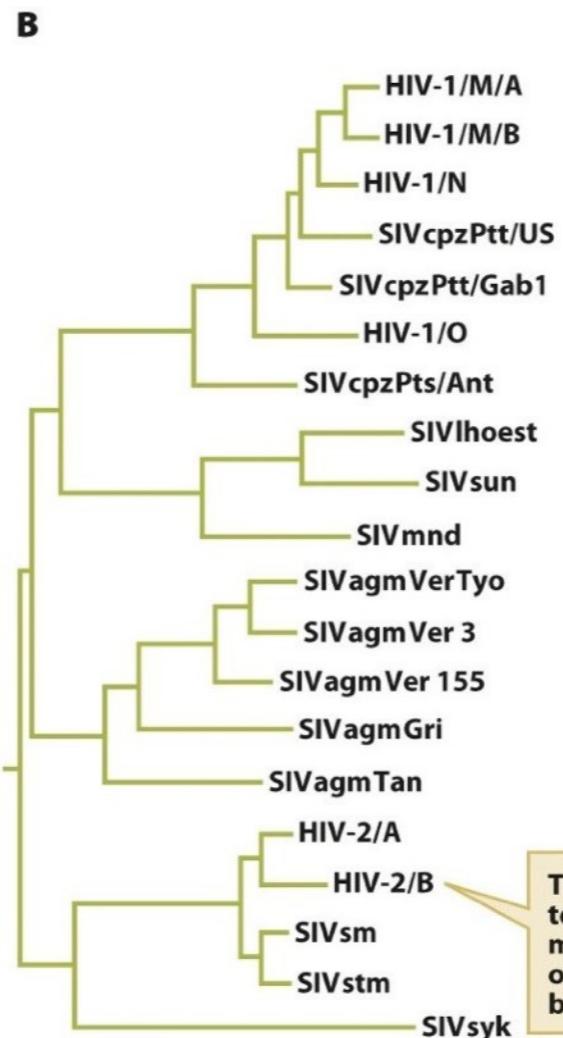
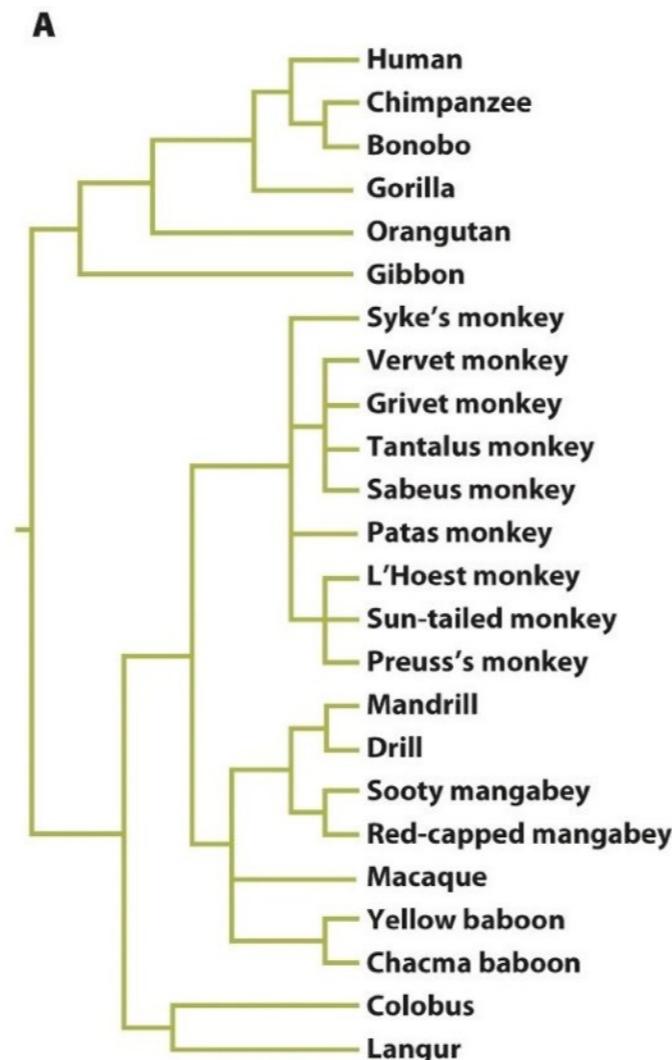
# Branch length

- Can be meaningless (arbitrary to space taxa evenly)
- ... or represent **evolutionary change** or time



# Branch length

No meaning:  
cladogram



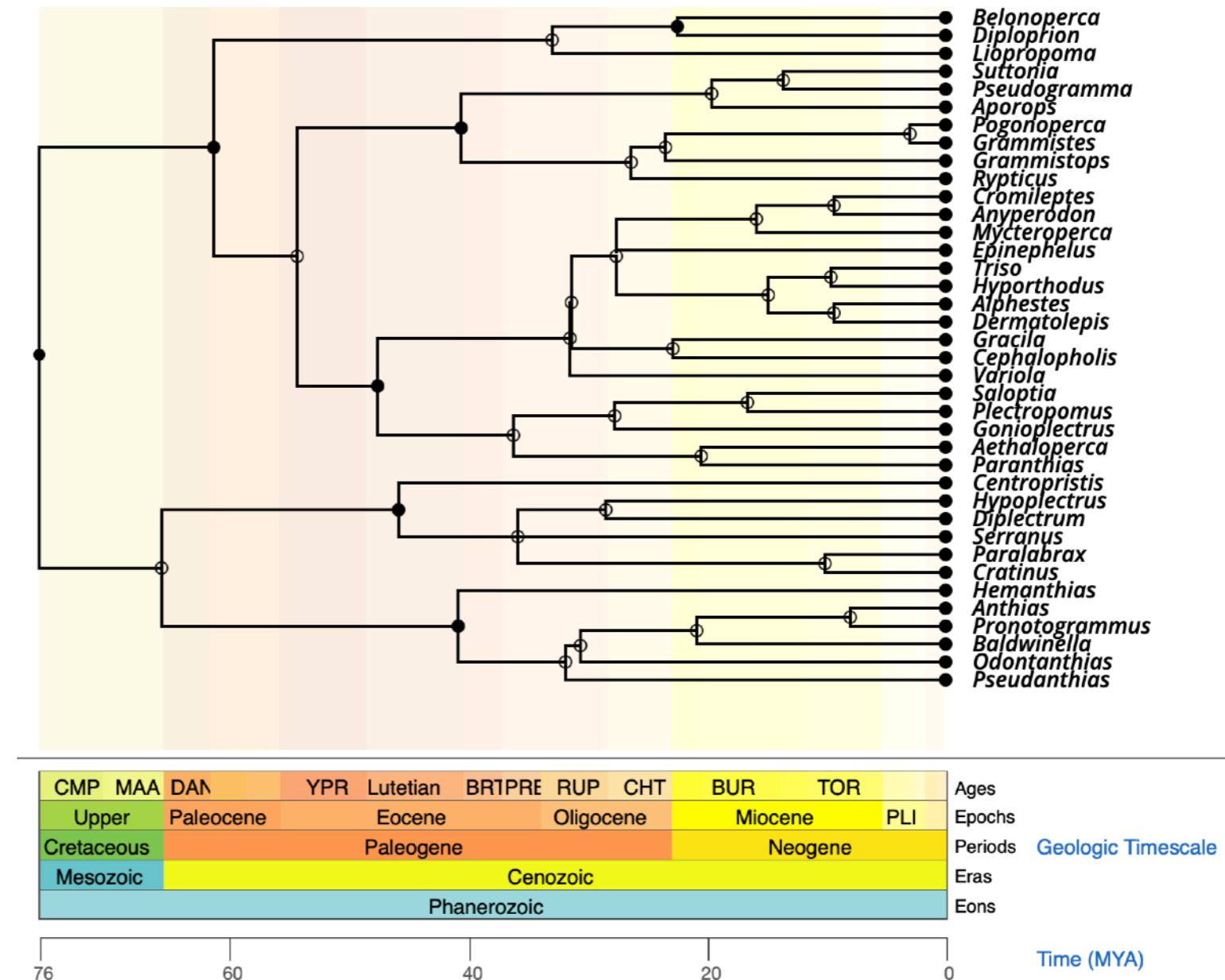
Evolutionary  
change:  
phylogram

The longer branch leading to HIV-2/B indicates that more change has occurred on that branch than on the branch leading to HIV-2/A

Bergstrom and Dugatkin, Evolution (Norton & Co., 2012)

# Branch length

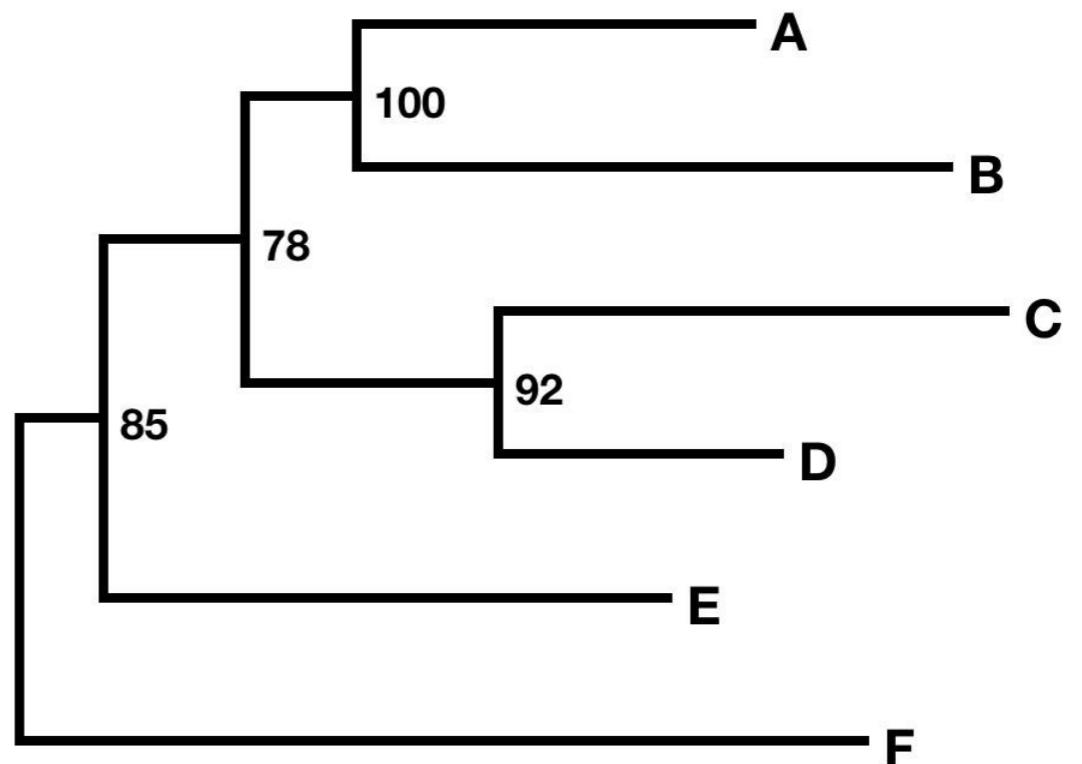
Time: chronogram



Serranidae, [timetree.org](http://timetree.org)

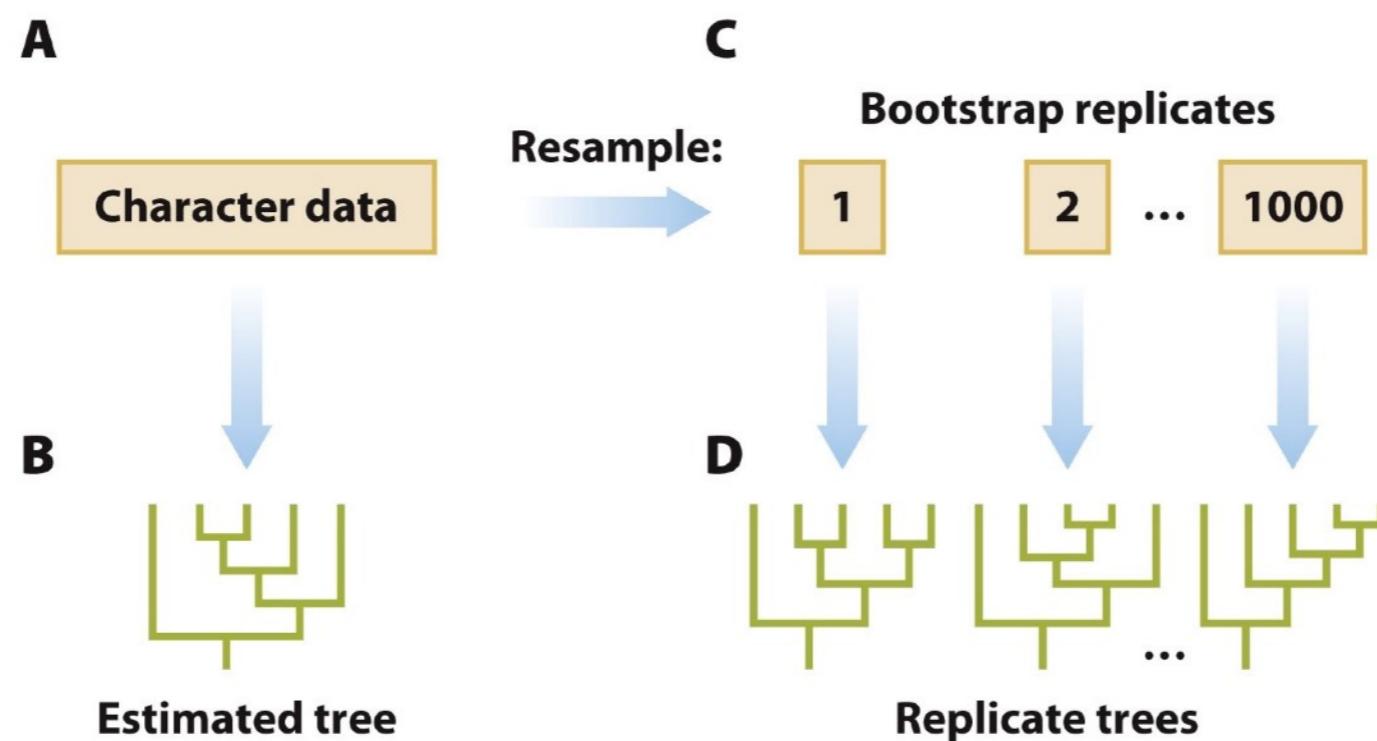
# Confidence values

- Measure statistical confidence (**support**) in the branching pattern
- **Bootstrap** support value: percentage from 0 (no support/confidence) to 100 (high support/confidence)



# Bootstrapping

- Calculated as percentage of replicate trees that contain the same node (i.e. the same taxa diverge from this node, though not necessarily in the same internal branching order)

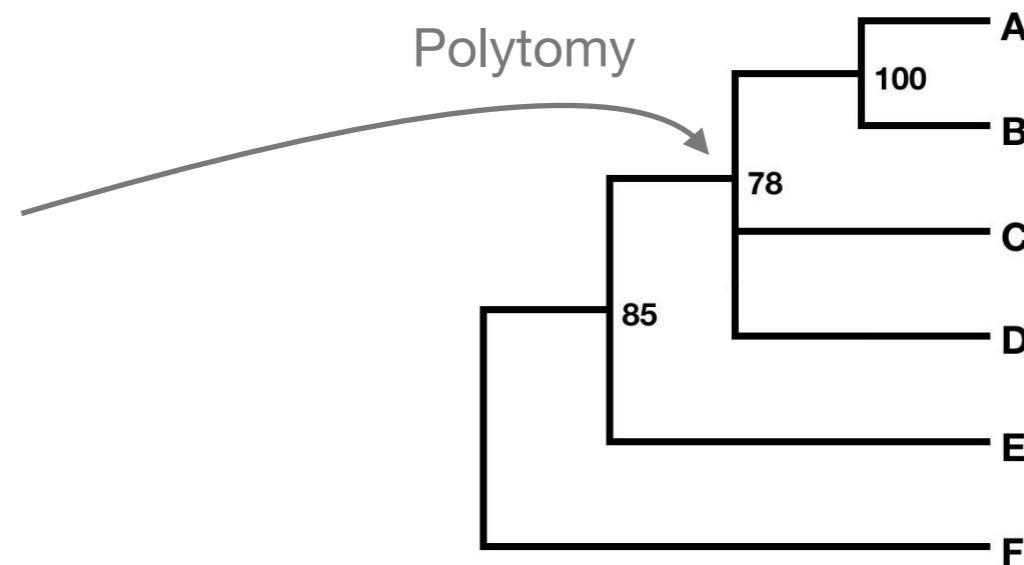


Bergstrom and Dugatkin, Evolution (Norton & Co., 2012)

# Bootstrapping

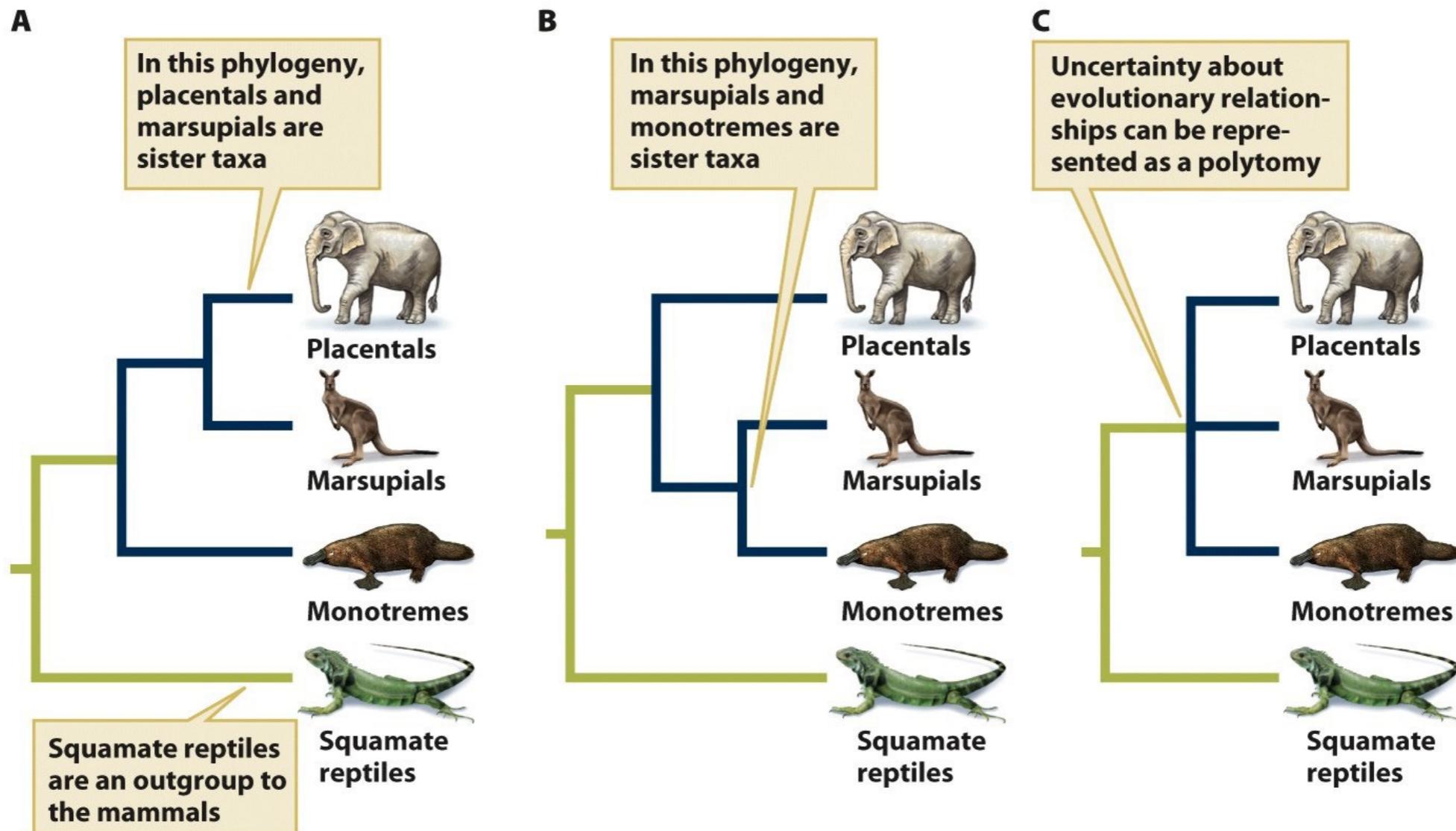
- Rule of thumb to interpret bootstrap values:

Value	Support
< 50	none
50–70	weak
70–90	moderate
> 90	high



- Important: measures how **consistently** the data set supports the tree, not whether the tree is true (e.g. the data set might be biased)

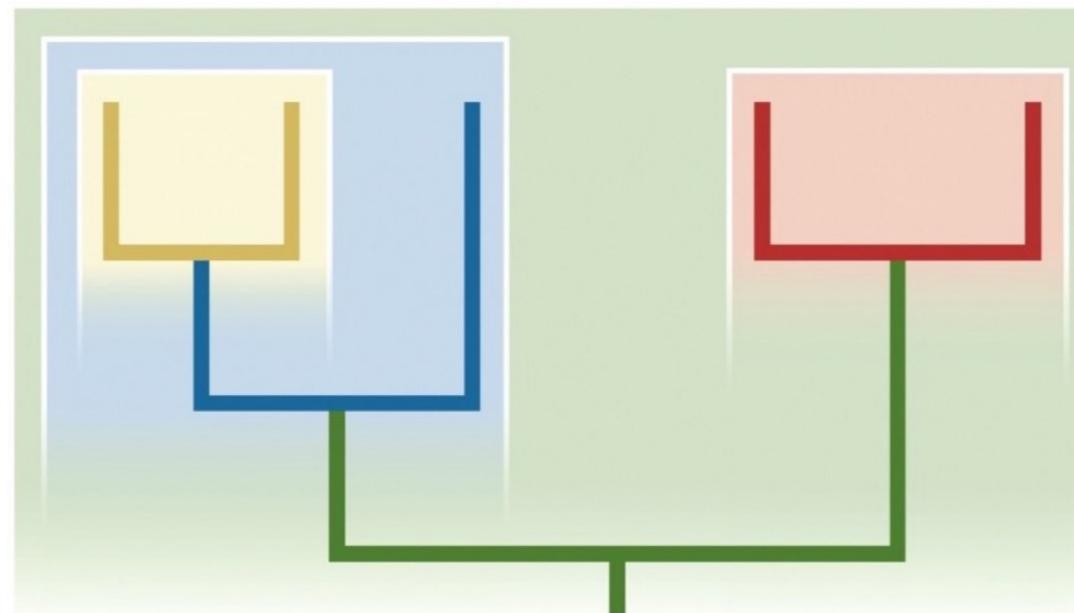
# Uncertainty reminder: trees are hypotheses



Bergstrom and Dugatkin, Evolution (Norton & Co., 2012)

# Monophyly

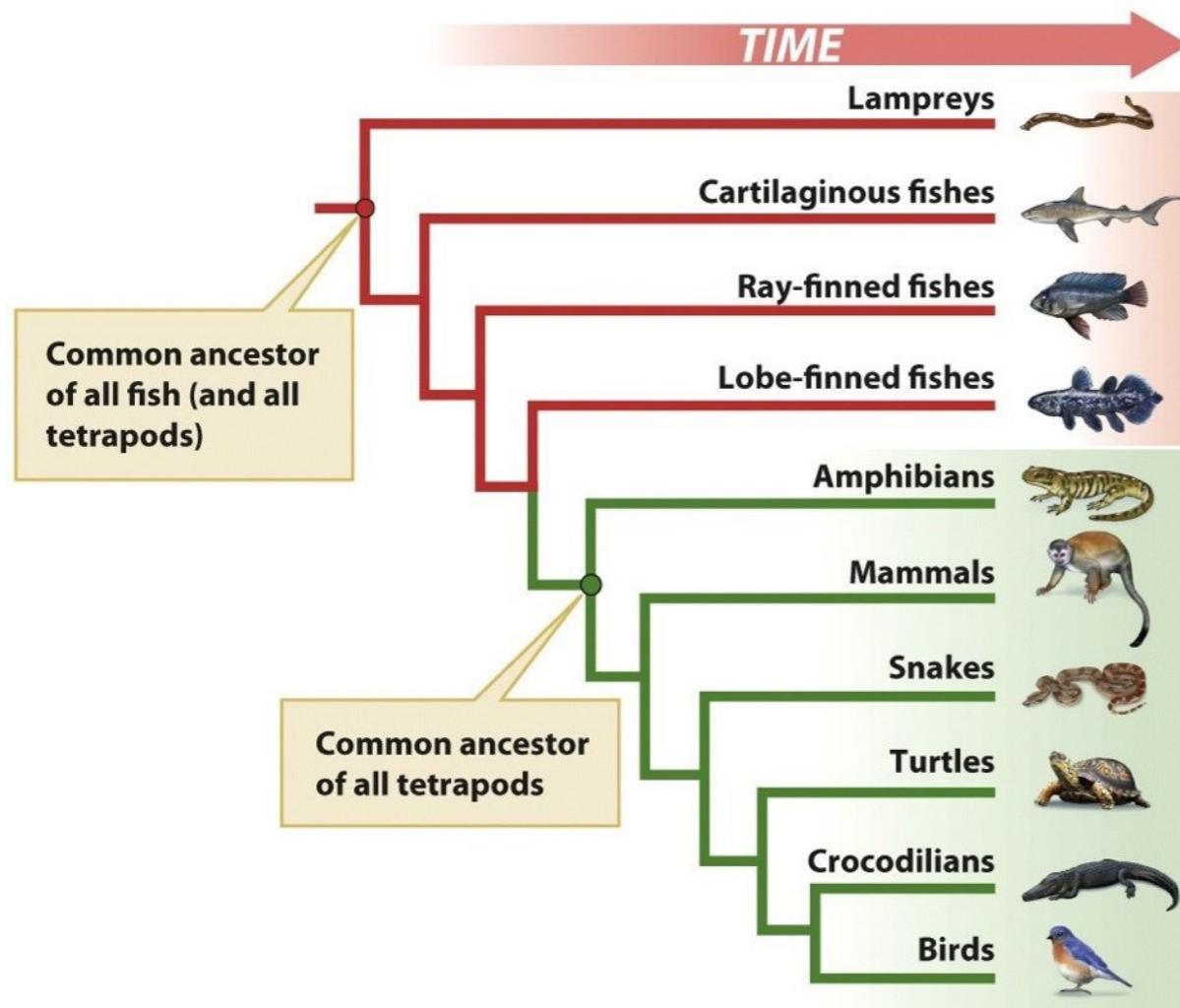
- Monophyletic groups or **clades** consist of a common ancestor, all of its descendants, and no other member
- Basis of modern systematics



Bergstrom and Dugatkin, Evolution (Norton & Co., 2012)

# Monophyly

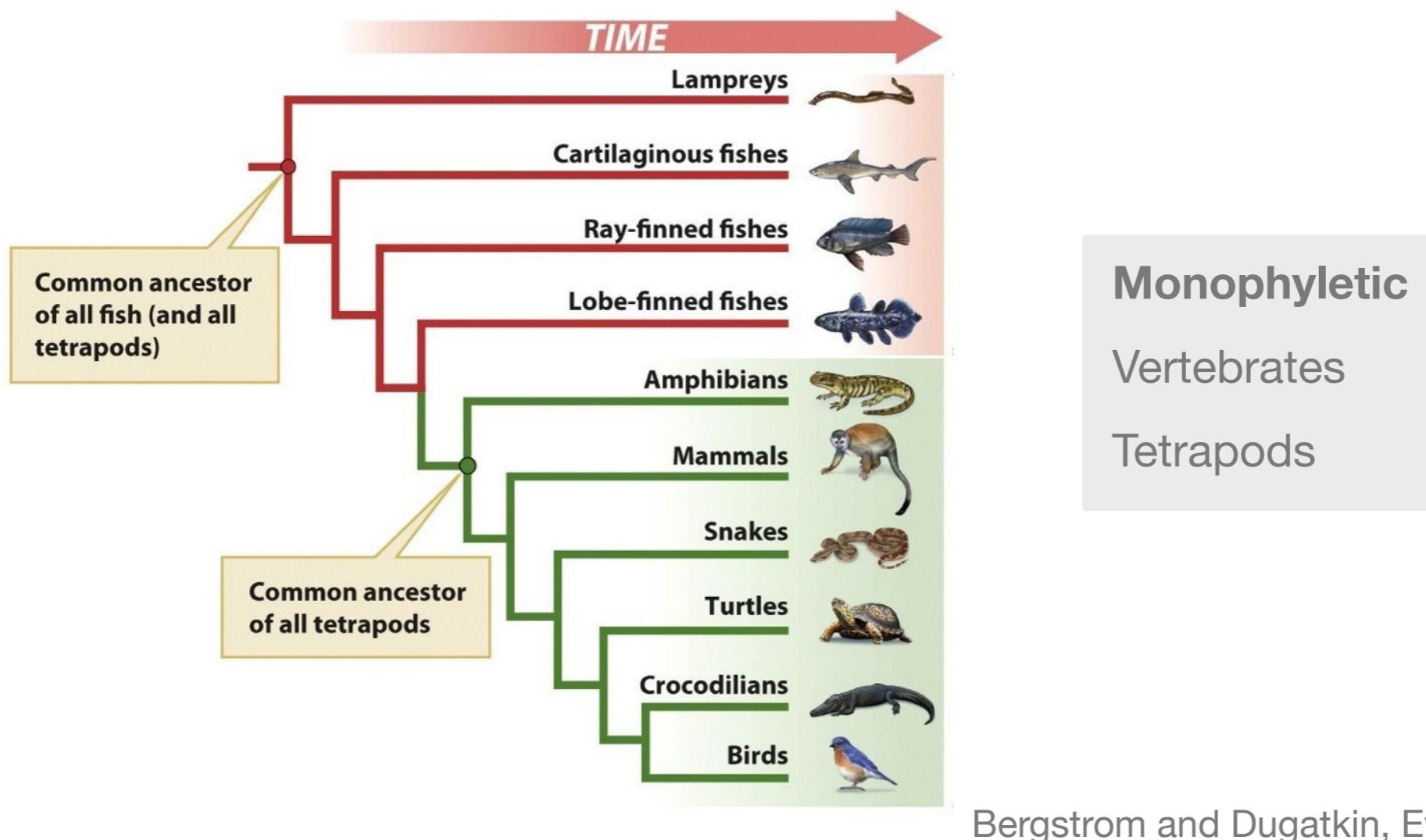
- Monophyletic groups or **clades** consist of a common ancestor, all of its descendants, and no other member



Bergstrom and Dugatkin, Evolution (Norton & Co., 2012)

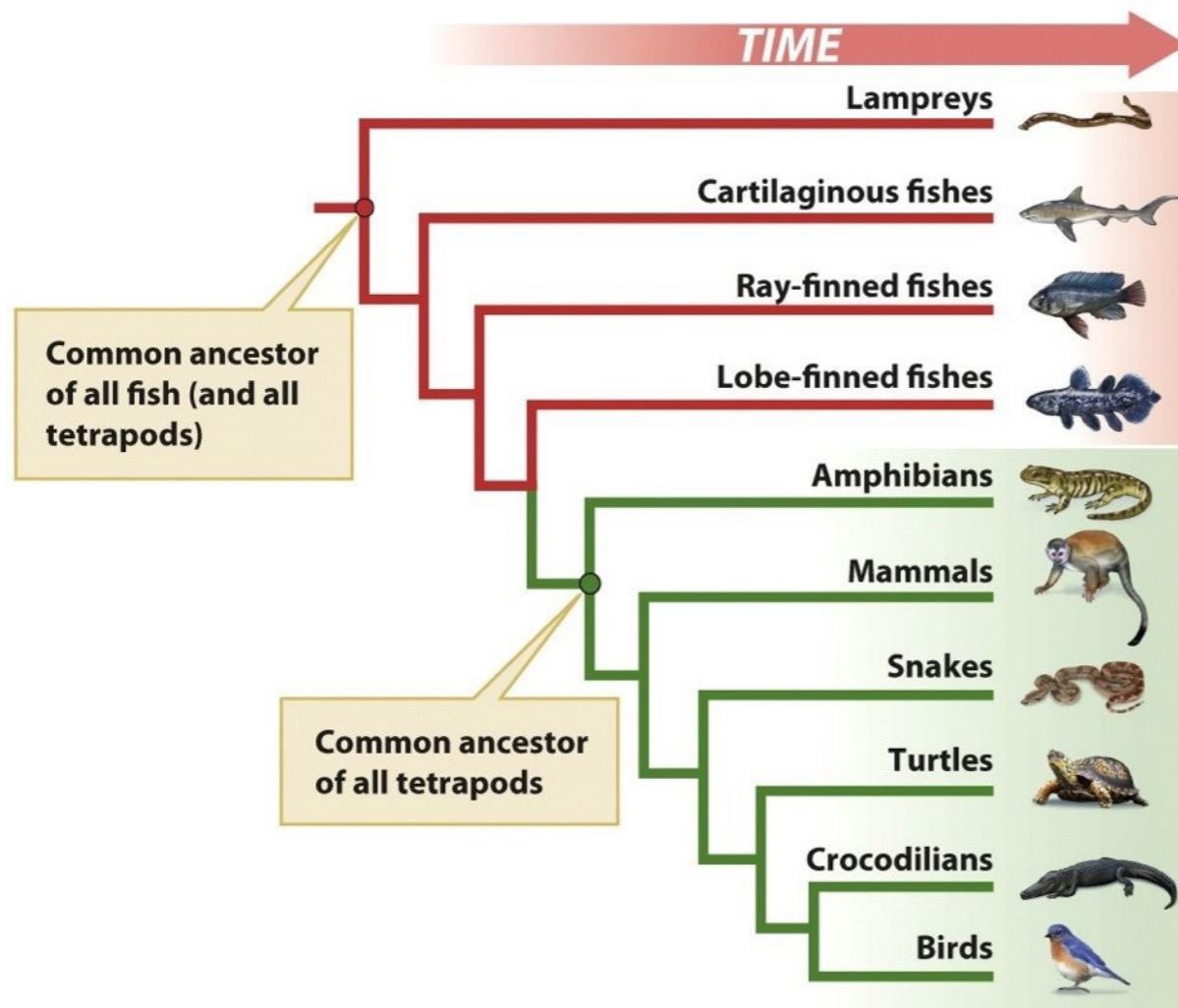
# Monophyly

- Monophyletic groups or **clades** consist of a common ancestor, all of its descendants, and no other member



# Paraphyly

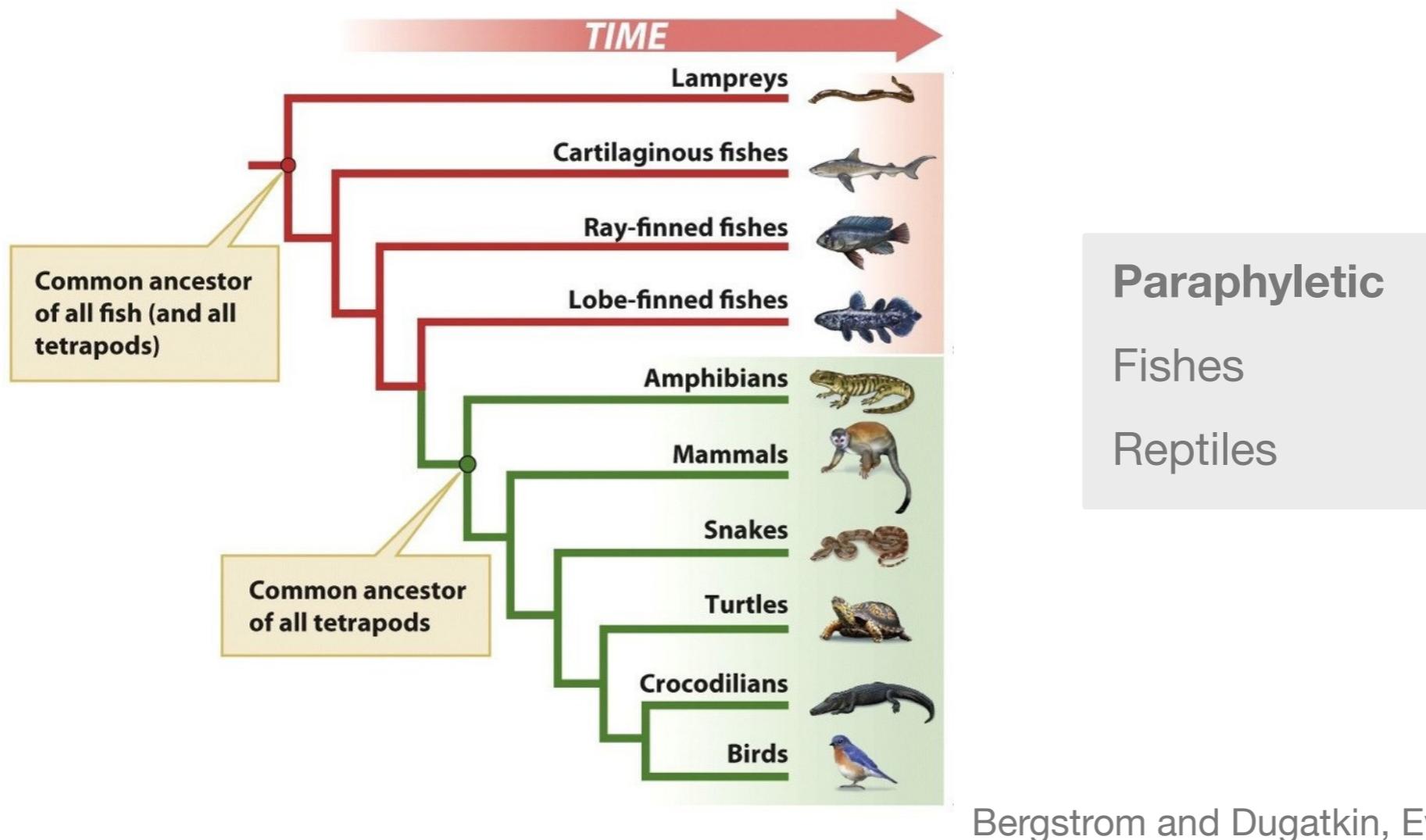
- Paraphyletic groups contain a common ancestor, and some but not all of its descendants



Bergstrom and Dugatkin, Evolution (Norton & Co., 2012)

# Paraphyly

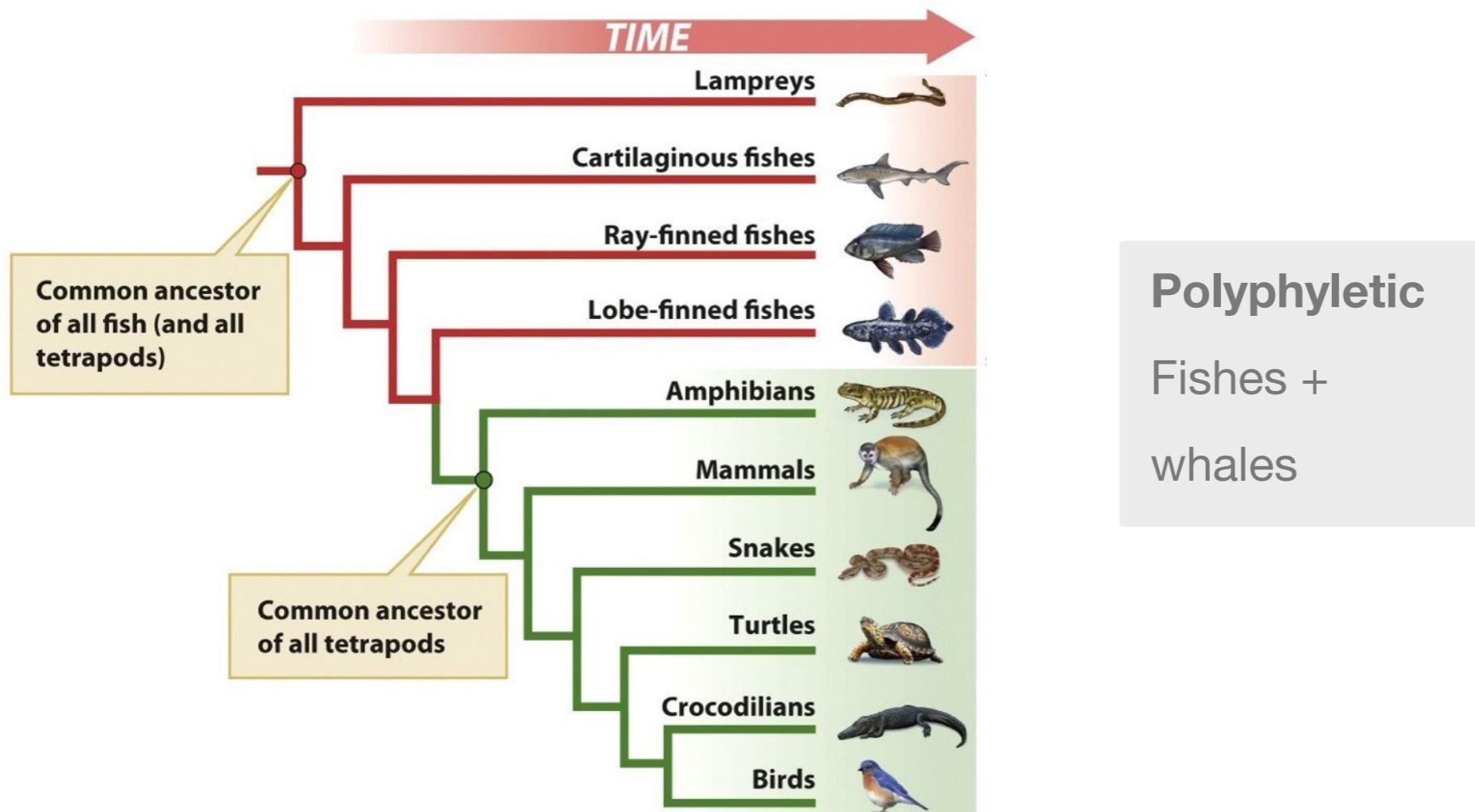
- Paraphyletic groups contain a common ancestor, and some but not all of its descendants



Bergstrom and Dugatkin, Evolution (Norton & Co., 2012)

# Polyphyly

- Polyphyletic groups contain neither the common ancestor of their members, nor all of its descendants (not accepted in systematics)



Bergstrom and Dugatkin, Evolution (Norton & Co., 2012)

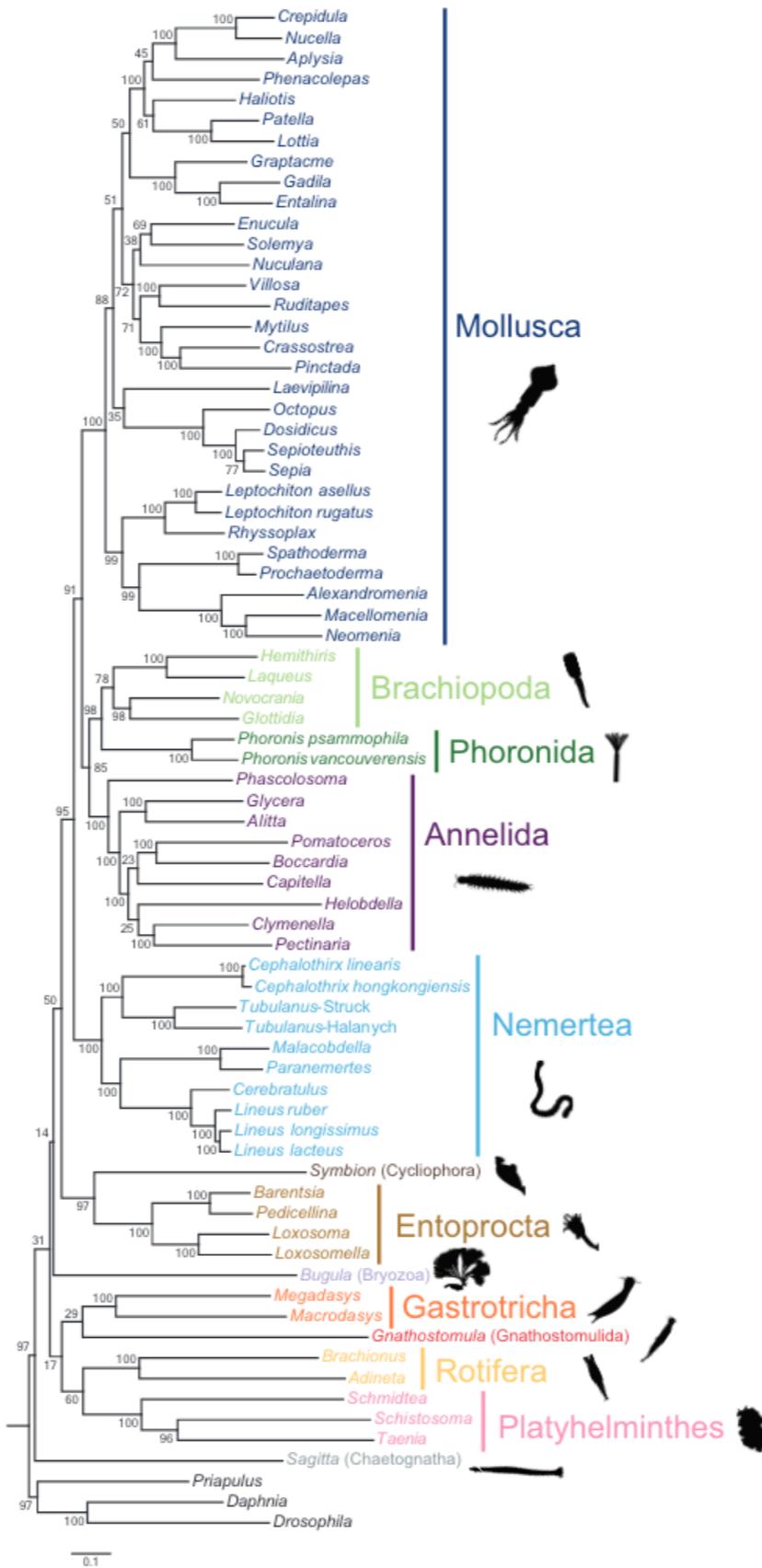
# Interpreting trees

Exercise 1

Animal phyla (Lophotrochozoa)

100+ protein-coding genes

Maximum likelihood analysis



Kocot et al. 2016, *Sys. Biol.*

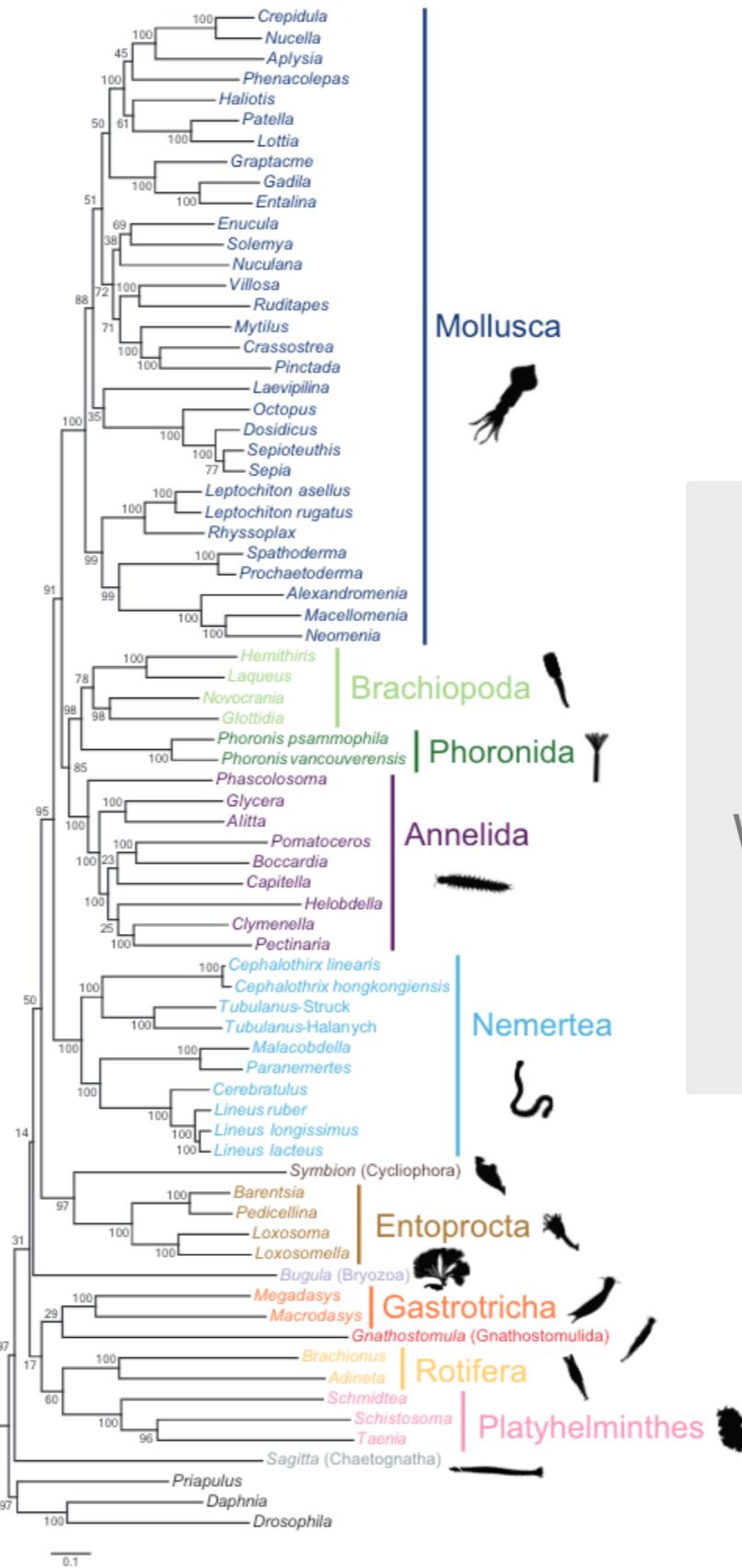
# Interpreting trees

Exercise 1

Animal phyla (Lophotrochozoa)

100+ protein-coding genes

Maximum likelihood analysis



- Identify:  
Branch lengths  
Well-supported clade  
Weakly supported group  
Sister-group  
Outgroup

Kocot et al. 2016, *Sys. Biol.*

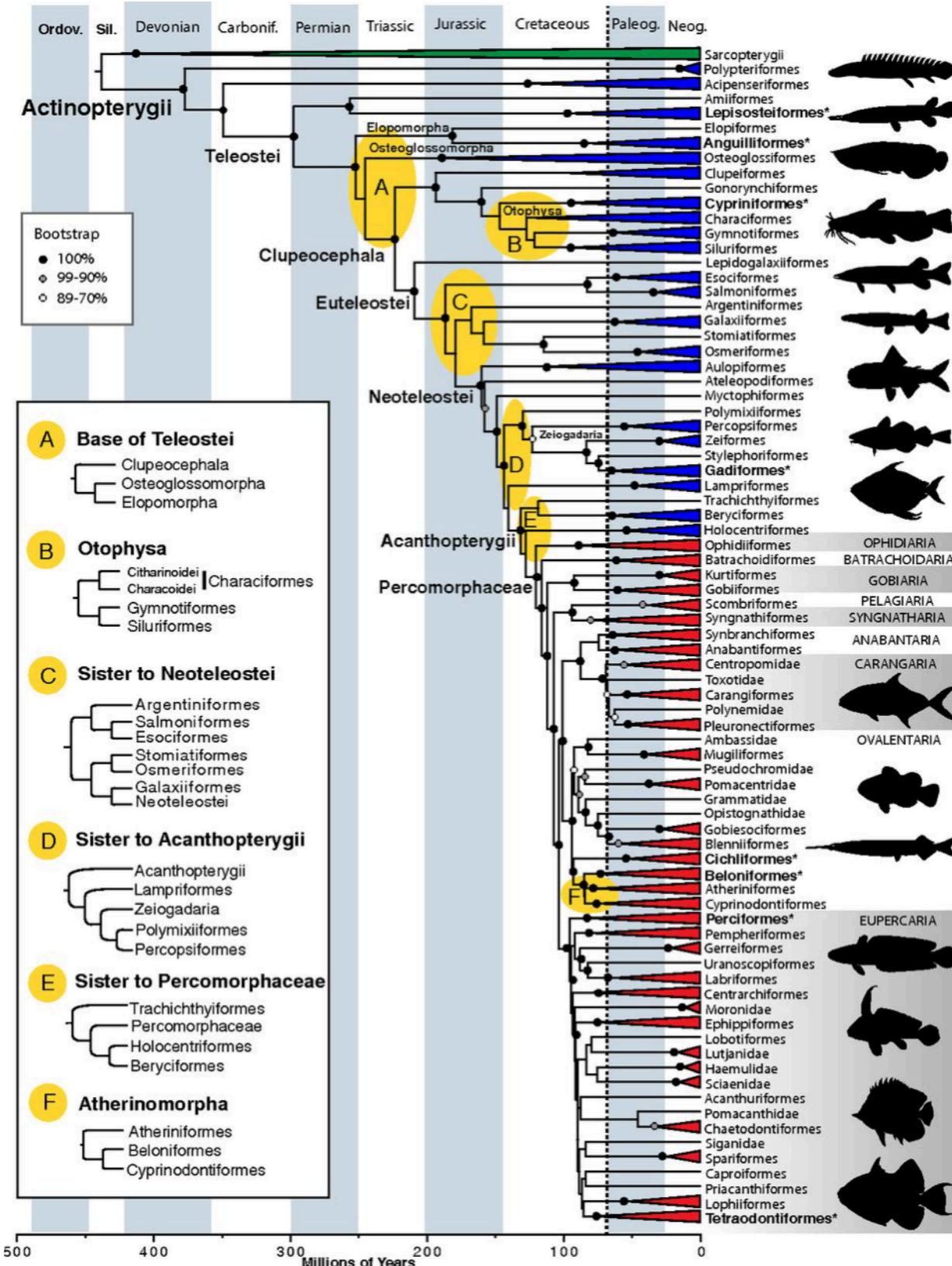
# Interpreting trees

Ray-finned fishes

1100+ exons

Maximum likelihood

Exercise 1

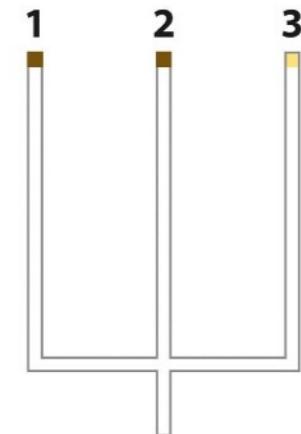


Hughes et al. 2018, PNAS

# Inferring phylogenetic trees

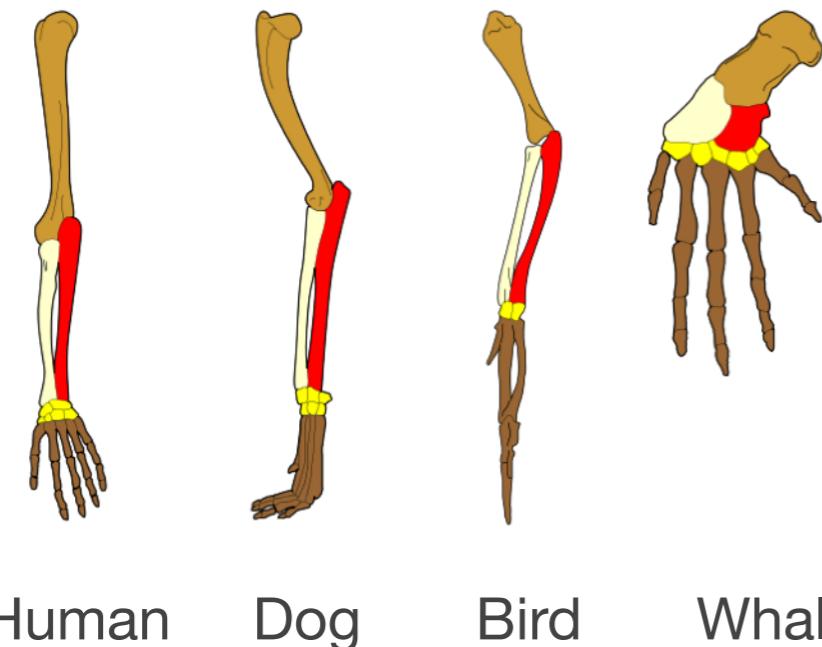
# The principle of phylogenetic inference

- Evolutionary relationships can be inferred by **observing traits** (characters) and **measuring their degree of similarity**
- Rationale: Taxa with many traits in common are more likely to be closely related
- Key assumption: traits are **homologous**



# Homology

- Homologous traits are similar between species because they are **inherited from a common ancestor**
- Therefore, homologous traits can be used to reconstruct evolutionary relationships

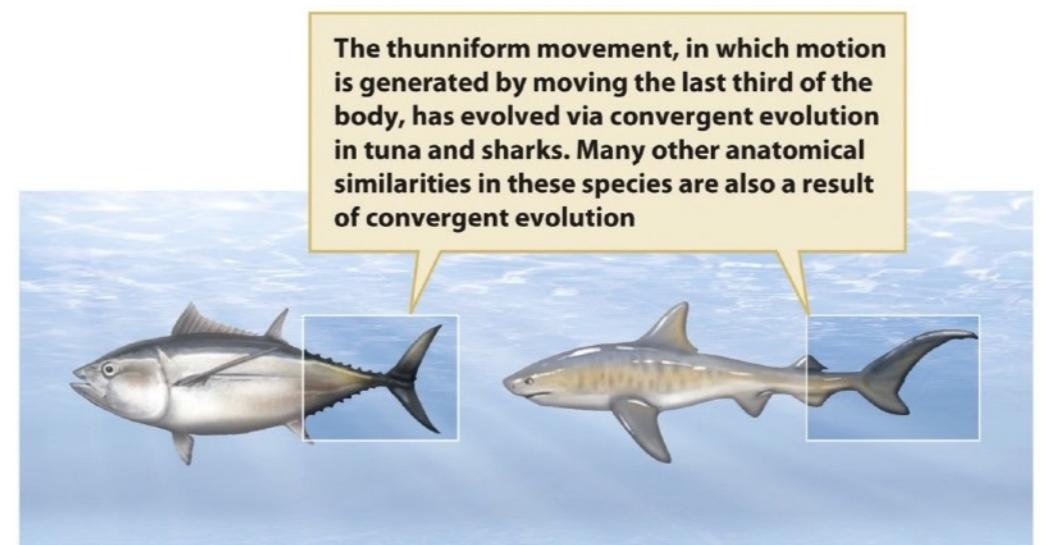
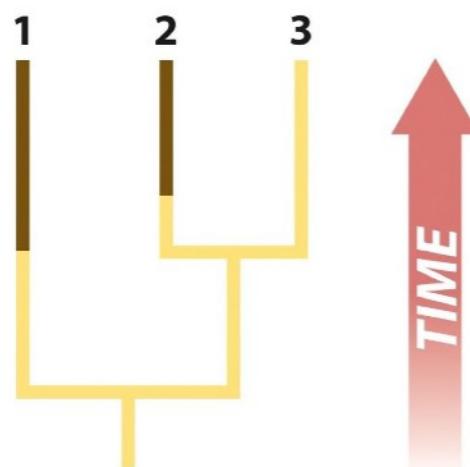


CGTTAGGATGCT  
CGTCA-GATGCT  
CGCTAGGG-GC-

Sequence homology

# Homoplasy

- Analogous trait, e.g. resulting from convergent evolution, not common ancestry
- Misinterpreting homoplasies as homologies can lead to wrong conclusions about evolutionary relationships



Bergstrom and Dugatkin, Evolution (Norton & Co., 2012)

# Homologous or analogous?

Bat and moth wings?



Bat and bird wings?



# Homologous or analogous?

Bat and moth wings?



Did the last common ancestor have this trait?

Bat and bird wings?



# Homologous or analogous?

Bat and moth wings?      no → analogous



Did the last common ancestor have this trait?

Bat and bird wings?



# Homologous or analogous?

Bat and moth wings?      no → analogous



Did the last common ancestor have this trait?

Bat and bird wings?

as forelimbs, yes → homologous

as wings, no → analogous



# Phylogenetic markers

Can be morphological or molecular features, but are typically DNA sequences:

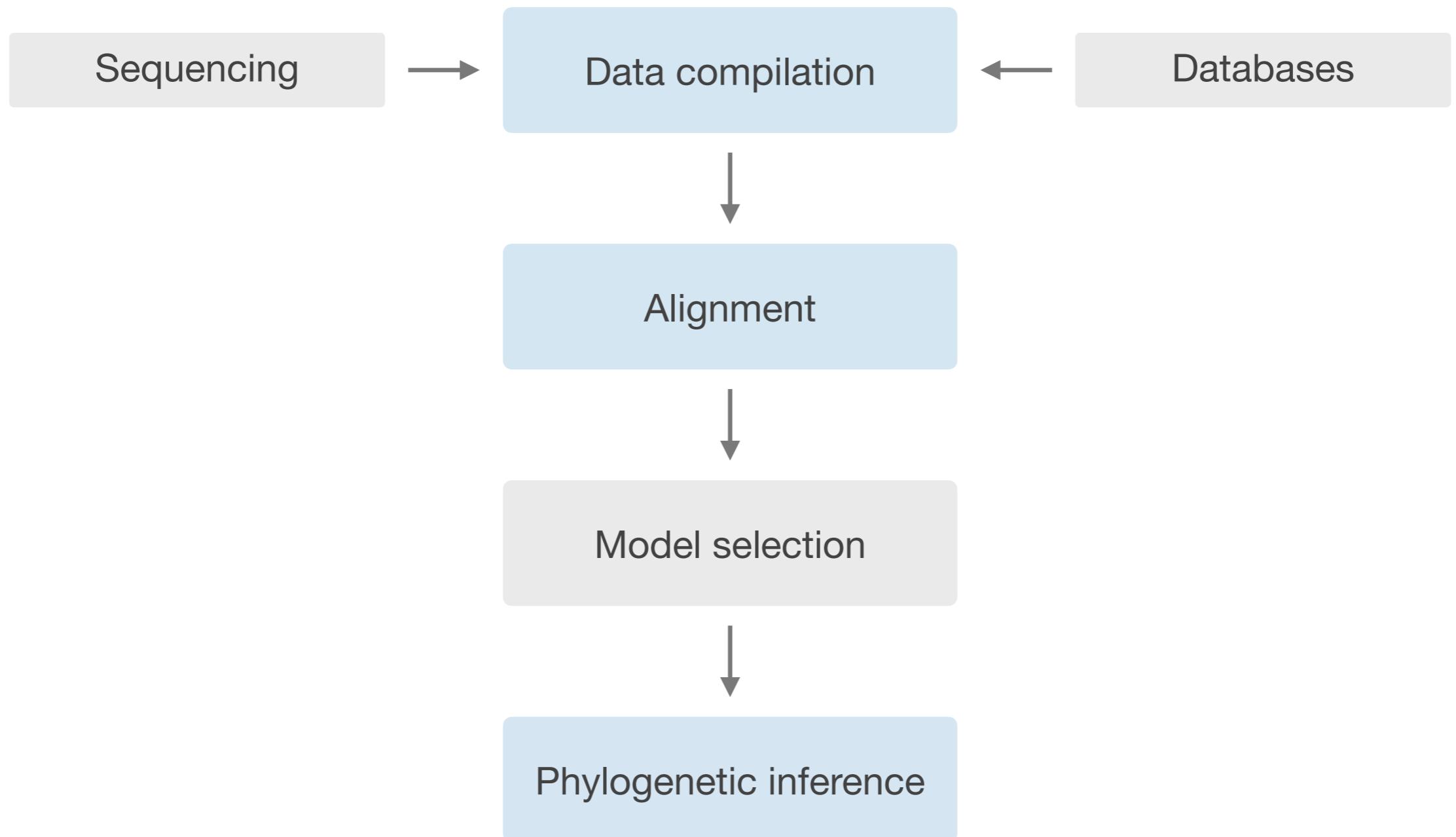
- Ribosomal RNA genes (rRNA), e.g. 16S, 18S
- Mitochondrial protein-encoding genes, e.g. COI
- Nuclear protein-encoding genes
- Introns, microsatellites and SNPs
- Whole genomes (max resolution, all levels)



Conserved /  
phylum level

Evolving fast /  
population level

# Molecular phylogeny workflow



# Alignment

```
G A T G T T C G A A  
G A T C - - - G A A  
G A C C - T C G - T
```

Arranges nucleotide or amino acid sequences  
so that the number of mismatches are minimized

- Accomplished by introducing gaps (-), which represent insertions or deletions (**indels**) / account for sequence length differences due to mutations over time
- Represents our best guess at which parts of sequences are evolutionarily equivalent, i.e. homologous
- Precisely: a **hypothesis of homology**, where each column indicates sites that are inferred to have descended from the same position in a common ancestor

# Alignment

Example: Insect desaturase genes, amino acid sequence

Hs(CG9747a)	---PNWDSNFEIHDEEDLKH-LRLTSATIDKRKCEQYEHKTRKKV.KG--FE.K.LNI.M...LGTL.SLF.INMIGNIT.FF.TVFLGAVQG....V.....LP..Y..L
Hs(CG9747b)	LGTQRCITTEAYIVANDDSKTI-TFDSSNVSAATKEVYRMKRSGEF.TK---I.SMA.L...TVGF.GIL.FDYLQNLK.TS.IFCMHVLTV.....A.....LP..I..-
Lh(CG9747a)	TDIFSIEPKICTANNFEKS-LTIAKMTEDENGIQNENKVHTKV.KNFKTE.K.MNI.H...GGCL.SLF.FSWSENWK.TV.IFIVGTLQA....I.....WQ..V..L
Lh(CG9747i)	SSSVTEDN-EDQKLLQLEHENKNSILDQPP.EHIVQP.I.RNI.G...FLAP.FFI.RFFQIKFW.WI.GFTYGFPTT....A.....LP..I..A
Lh(CG9747k)	PVKIVDTD-NDATDKKLEKIKLADF.TN---I.YIV.Q...TIGI.ALL.FNYMTNPM.TL.IIGMYIISN....A.....LP..I..F
Cf(CG9747a)	VEENFGKE-EFSTITTDDENDIQTRDETRKKF.EKFKTE.K.LNV.S...MIML.ACF.VEWYKDLR.NI.MFIVSVIGA....A.....WQ..I..L
Cf(CG9747c)	NSTITEGN-KSRKLTLR-TENGKNSILDQPP.ENIVQP.I.RNI.A...IITI.LFV.RYNEAKFW.WI.TISYEILGG....A.....LP..I..T
Pb(CG9747a)	PNISTTKTKILTEEDFNES-LTTKIINDENDTQYKREVVYVKT.K--TMK.K.KNI.F...ISFV.VCF.FKFFENLK.LA.IFIMYLIGG....A.....WP..I..V
Pb(CG9747b)	SSTITEDN-KNRKSVQPGTKNGRNSILDQPP.EHIVQP.I.RNV.G...FVAI.LFV.RYHKATLW.WI.STVYGYTAG....L.....LP..I..A
Pb(CG9747c)	CDDVKCKEEAMTEKKM.N---D.FAI.F...VIGL.GLV.FNFFQNWK.TL.CLILYELGC....A.....LP..L..V
Ac(CG9747a)	SSMITEGN-KNQMLVQPGTENGRDSNLNQPP.-HMEQP.I.RNI.V...IVAI.LFA.RYREAKFW.WM.SILHLLSAG....L.....LP..I..V
Ac(CG9747c)	PNILSIDSKILTAKKFEESIIEKTTDENDENDVQNRSESYIKS.G--KTE.K.KNI.T...LCFV.VFF.FKFFENLK.TA.IYGMILIGG....A.....WQ..I..V
Ac(CG9747d)	RDDIKKSKMIKLNKF.TD---I.FIA.Q...AIGI.GLL.FNYWQNLI.TI.IIVMYIISN....A.....LP..I..L
Prim.cons.	LGTPNI522ES2I2TEDNF22SS25K4TKL22DET2NQYESILDQK2EE2I2Q2LIWRNII3ILHIIGIY22FTFNFF3NLKTTIWIFIM22IGGIGVTAGAHLWTHRKA KLPLRILL2
	130 140 150 160 170 180 190 200 210 220 230 240
Hs(CG9747a)	IFYTIAGQNKI...D..V...T..DA...SN....S.....I.K..E.I.K.T..S..LE.VI.GD..FMV.K.I.C.A..IM.P..G...TWIR.ISIQ.FFRYVFG..CT..L
Hs(CG9747b)	NSV...N..L...C..DA...TN....S.....M.N..K.L.K.L..S..LS.VV.SE..FLP.Q.L.S.V..TI.P..F...TWGR.IISQ.FIRYMIT..SI..I
Lh(CG9747a)	IFYCIAGQNNP...D..V...T..DA...SN....S.....M.K..E.I.K.I..T..MA.FV.GR..FLI.K.V.A.L..IS.P..A...IWSR.VMSQ.FR-YVLG..FT..A
Lh(CG9747i)	IFYCMAGQTHL...V..I...T..SA...SN....S.....M.H..A.K.Y.V..S..AA.IR.FD..YPP.M.S.C.V..TL.Q..L..TWDI.ITST.VR-YVWS..AT..F
Lh(CG9747k)	SMLQRMCTDTI...N..M...C..NA...IK....S.....M.E..E.T.K.L..S..MS.VT.GE..FLP.Q.F.G.I..TA.P..L..TWTR.IMSQ.FIRYIIT..SV..I
Cf(CG9747a)	IFYATAGQNNL...D..V...T..NA...SN....S.....L.K..D.I.K.I..R..LA.AM.GY..FSI.K.I.A.L..IM.P..G...TWSR.CKSQ.IR-YVLS..GT..V
Cf(CG9747c)	CFYCMVGQTHL...V..T...T..SA...AS...S.....V.C..A.K.G.V..S..AA.IR.FD..YLP.M.S.C.V..TL.P..I..TWIV.IGAT.SR-YLVT..AA..F
Pb(CG9747a)	IMYSSSGLNNL...D..V...T..DA...SN....S.....M.K..E.I.R.T..S..LT.AI.CC..YSI.R.V.A.L..VM.P..G...TLYR.LVSQ.IIRYTLL..II..A
Pb(CG9747b)	YLYCMAGMTHP...V..T...T..SA...SN....S.....M.H..A.I.Y.V..S..AA.IR.VD..YSP.M.T.T.V..AI.P..L..TWEI.IGAV.IR-YVLL..ST..F
Pb(CG9747c)	LCFSTTAEISI...T..V...V..HL...SR...C.....R.H..E.V.K.T..R..LT.LA.QK..PGL.N.I.S..I..TL.P..F...TWSR.FISQ..MRVLFV..AS..F
Ac(CG9747a)	LLYCLAGQTHP...V..T...T..CA...AN...S.....M.H..A.K.Y.V..S..AA.IR.VD..YAL.M.P.C.V..AI.P..V..TWDI.ISSV.IR-YVWA..ST..F
Ac(CG9747c)	VLYASTGMNDI...D..V...T..DA...SN....S.....M.K..D.I.K.V..S..LA-AA.SV..FLI.K.I.T.L..VM.P..G...TWYR.FVSQ.ILRYVLL..FI..V
Ac(CG9747d)	ICFSASAQNTI...N..M...C..NA...SQ...G.....M.E..E.I.K.L..S..LS.VV.GE..FLL.Q.L.G.I..TT.P..L..TWNR.IVSQ.FIRYMIS..AV..I
Prim.cons.	IFYCMAGQNHIVVRDHRVHHKYTDADPHNSNRGFFFSHVGWLMMKKHPEVIEK2VDMSDLAD2RGDKYFLPL2LIFCFVLPTMVPVY2WNETWSRAIVSQI2RRYVL3LN2TWNSF

# Phylogenetic inference methods

- Maximum parsimony
- Distance methods
- Maximum likelihood
- Bayesian inference



implicit assumptions about evolutionary change

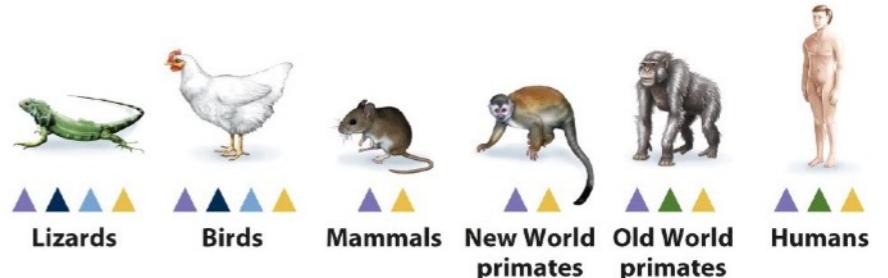


explicit statistical models of evolutionary change

# Maximum parsimony

Principle: the best tree is the one explaining the observed characters while requiring the **fewest evolutionary changes**

- ⊕ Can be used for morphological data



- ⊖ Simplistic assumption of the evolutionary process (no explicit modeling)

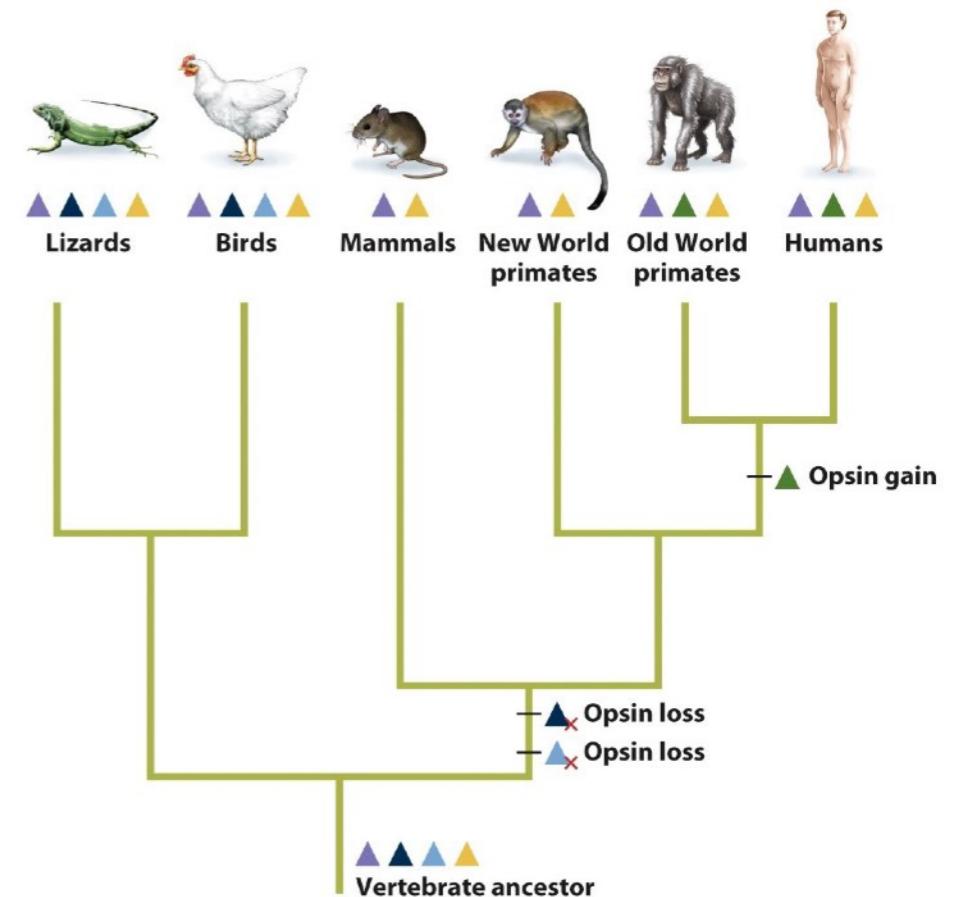
Adapted from Frentiu and Briscoe 2008

# Maximum parsimony

Principle: the best tree is the one explaining the observed characters while requiring the **fewest evolutionary changes**

⊕ Can be used for morphological data

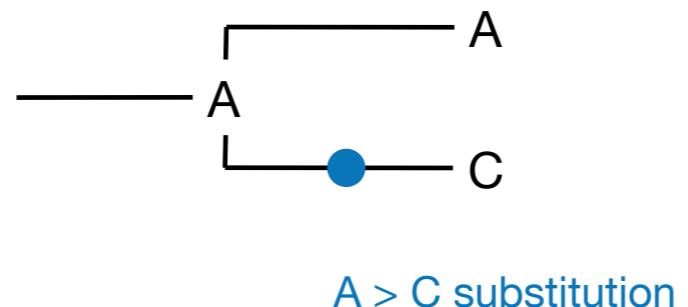
⊖ Simplistic assumption of the evolutionary process (no explicit modeling)



Adapted from Frentiu and Briscoe 2008

# Distance methods

- Construct phylogenetic trees by measuring the **evolutionary distance** between each pair of taxa and minimizing total distance between taxa
- Evolutionary distance = number of differences (**substitutions**) between two homologous sequences



# Distance methods

- Common measure: **p-distance** (substitutions relative to sequence length)

Human	VLSPADKTNV	KAAWGKVGAH	AGEYGAEALE	RMFLSFPTTK	TYFPHFDLSH	GSAQVKGHGK	KVADALTNAV
Cow	....A..G..	.....G.	A.....	.....	.....A	...A..K..	
Kangaroo	...A...KH..	..I.....K..	.....A..G..	..T.H.....	IQA....	I.....GQ..	

# Distance methods

- Common measure: **p-distance** (substitutions relative to sequence length)

Human	VLSPADKTNV	KAAWGKVGAH	AGEYGAEALE	RMFLSFPTTK	TYFPHFDLSH	GSAQVKGHGK	KVADALTNAV
Cow	....A..G..	.....G.	A.....	.....	.....A	...A..K..	
Kangaroo	...A...KH..	..I.....K..	.....A..G..	..T.H.....	.....IQA..	...I.....GQ..	

e.g.

Human–Cow

7/70 = **0.10**

# Distance methods

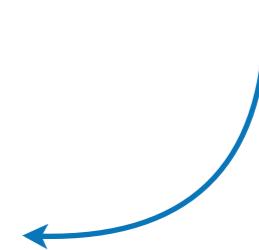
- Common measure: **p-distance** (substitutions relative to sequence length)

Human	VLSPADKTNV	KAAWGKVGAH	AGEYGAEALE	RMFLSFPTTK	TYFPHFDLSH	GSAQVKGHGK	KVADALTNAV
Cow	....A...G..	.....G.	A.....	.....	.....	A .. A .. K ..	
Kangaroo	...A...KH..	I.....K..	.....A..G..	T.H.....	.....	IQA... .I...GQ..	

e.g.  
Human–Cow  
 $7/70 = 0.10$

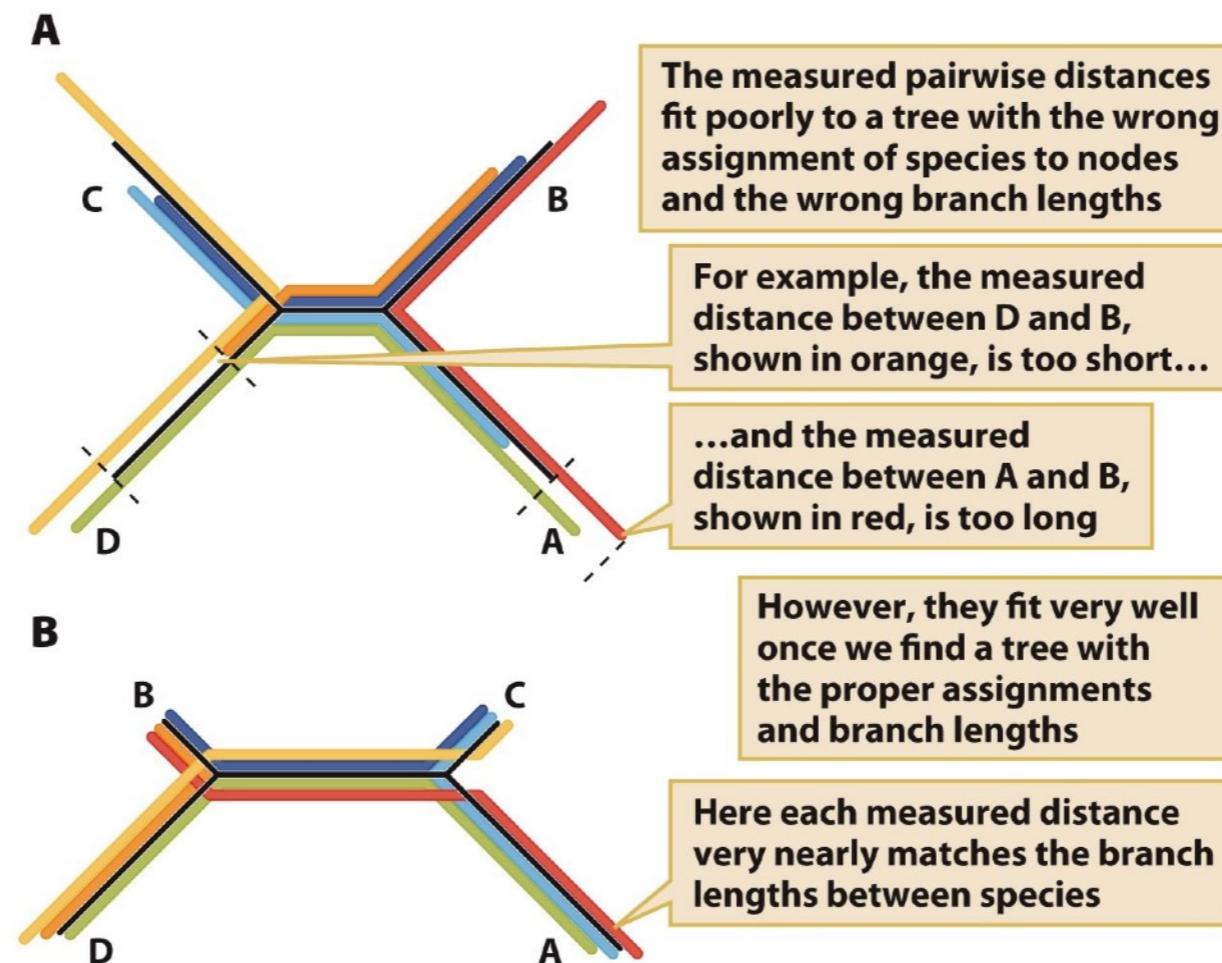
- Compile distance matrix

	Human	Cow	Kangaroo
Human	0	-	-
Cow	0.10	0	-
Kangaroo	0.21	0.23	0



# Distance methods

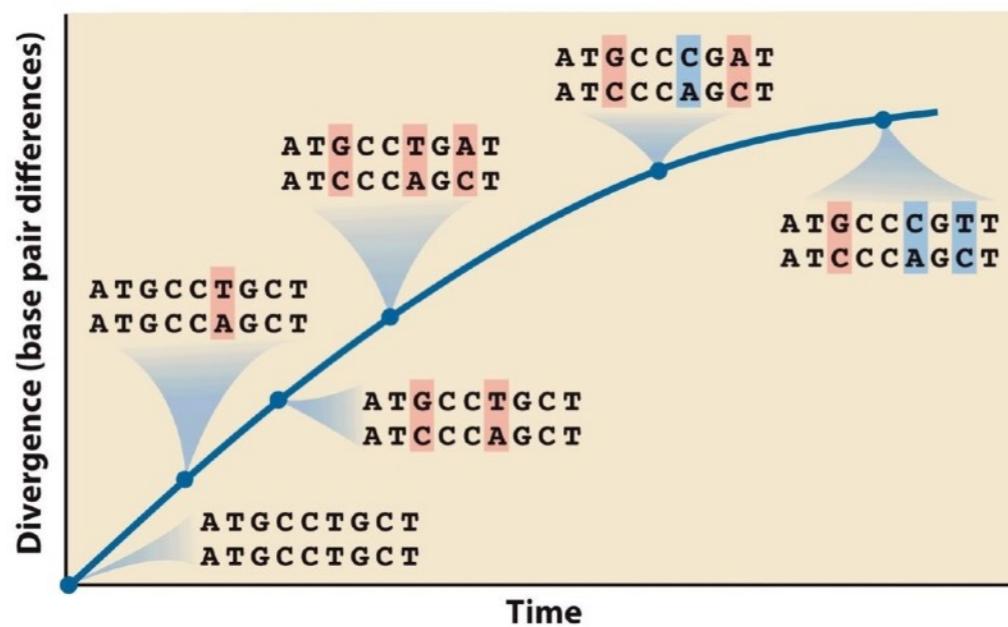
- Reconstruct the tree by minimizing total distance



Bergstrom and Dugatkin, Evolution (Norton & Co., 2012)

# Distance methods

- ⊕ Extremely fast, easy to run (Neighbor Joining, UPGMA)
- ⊖ Loss of information
- ⊖ The more substitutions, the less accurate ([undetected multiple substitutions](#))



Bergstrom and Dugatkin, Evolution (Norton & Co., 2012)

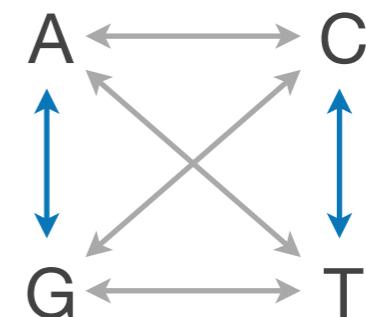
# Substitution models

- Probabilistically describe molecular evolution over time to account for homoplasies
- Parameters (for DNA):

Nucleotide frequencies  $\pi_A, \pi_C, \pi_G, \pi_T$

Exchange rates:  $\rho_{AC}, \rho_{AG}, \rho_{AT}, \rho_{CA}, \dots, \rho_{TG}$

→ Rate matrix Q



- Most general: GTR, General Time Reversible model

# Probabilistic methods

- Evaluate phylogenetic hypotheses in a statistical framework
- Find parameter values (rate matrix, tree topology, branch lengths) that maximize probability function

**Maximum likelihood:**

Probability of observing the data given the model

Bayesian inference:

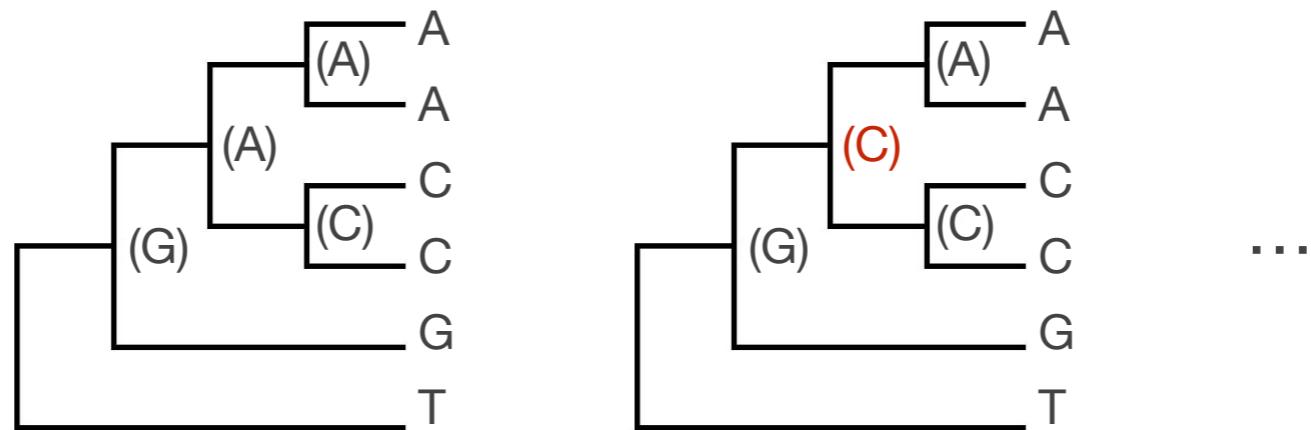
Probability of observing the model given the data

# Maximum likelihood algorithm

- Choose parameter combination (model, including tree topology)
- Calculate likelihood to observe data at each alignment position under this model



For this, sum over the probabilities of every possible ancestral character state (scenarios how data could have evolved)



# Maximum likelihood algorithm

- Total likelihood is product of individual likelihoods at each position
- Tweak model parameters, repeat calculation\*
- Best model (tree) is the one with the highest total likelihood

# Maximum likelihood algorithm

- Total likelihood is product of individual likelihoods at each position
- Tweak model parameters, repeat calculation\*
- Best model (tree) is the one with the highest total likelihood

\* For all possible tree topologies?

# How many trees are there?

$$N = \frac{(2t - 5)!}{2^{t-3}(t - 3)!}$$

Stars in the milky way ←  
Prokaryotes on Earth ←  
Atoms in the universe ←

$t$ (taxa)	$N$ (trees)
4	3
5	15
8	10,395
10	2,027,025
20	$2.2 \times 10^{20}$
30	$8.7 \times 10^{36}$
40	$1.3 \times 10^{55}$
50	$2.8 \times 10^{74}$
60	$5.0 \times 10^{94}$

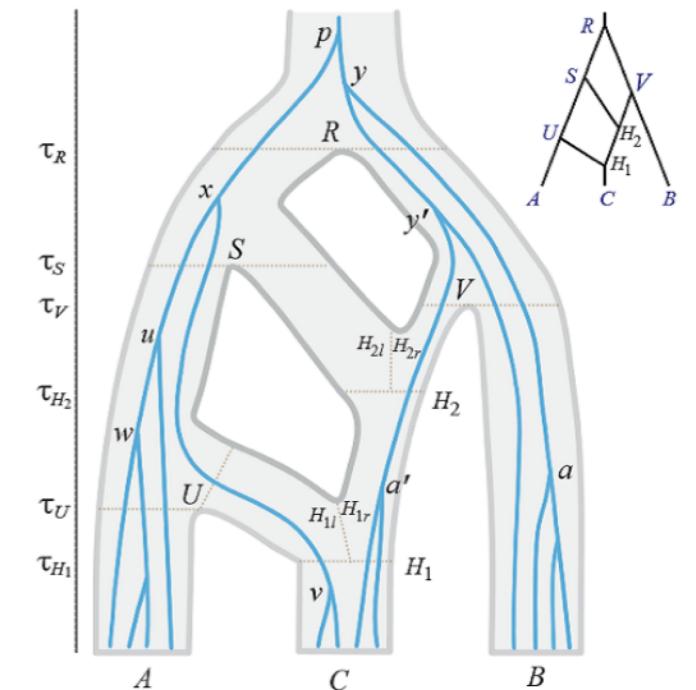
- Only **heuristic** tree search algorithms make ML approach feasible

# Maximum likelihood

- ⊕ Powerful, versatile (models are continually improved)
- ⊕ Alternative hypotheses can be tested
- ⊕ Many implementations (e.g. RAxML, IQTREE, PhyML)
- ⊖ Complex, computationally intense

# Why are so many phylogenies still unresolved?

- Lack or loss of phylogenetic information
- Evolution is **heterogeneous**: taxa evolve at different speeds, are biased in terms of amino acid composition etc.
- **Non-tree like evolutionary processes** are common (hybridization, horizontal gene transfer)



Ziheng Yang, University College London

# Summary

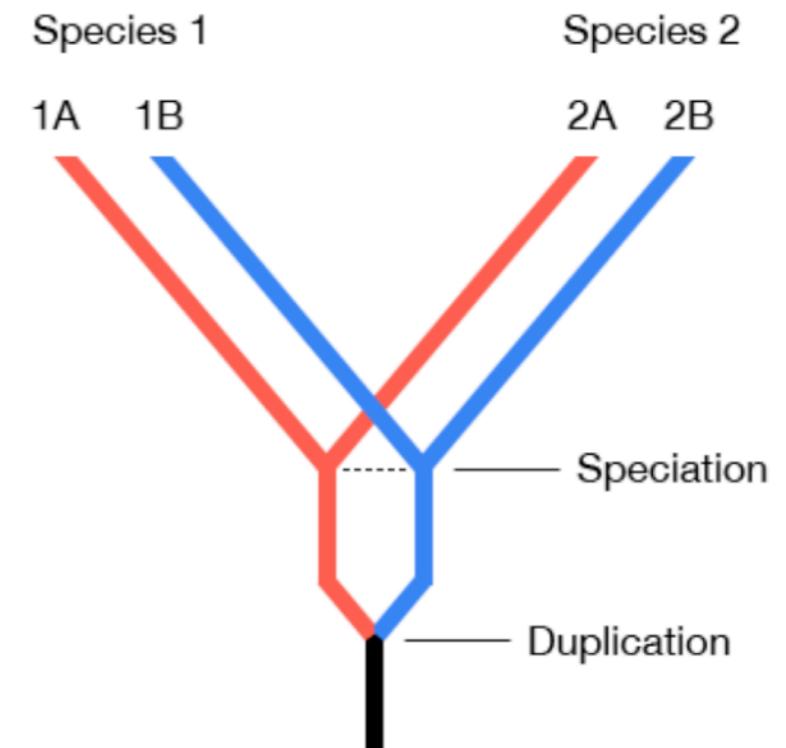
- Phylogenies are central to understanding the patterns and processes of evolution
- Phylogenetic trees are **hypotheses** about the evolutionary relationships between taxa or genes
- They can be inferred from **homologous characters** (molecular or morphological) and measuring their degree of similarity
- There are different phylogenetic approaches to infer trees, including probabilistic methods using **models of sequence evolution**

# Bonus material

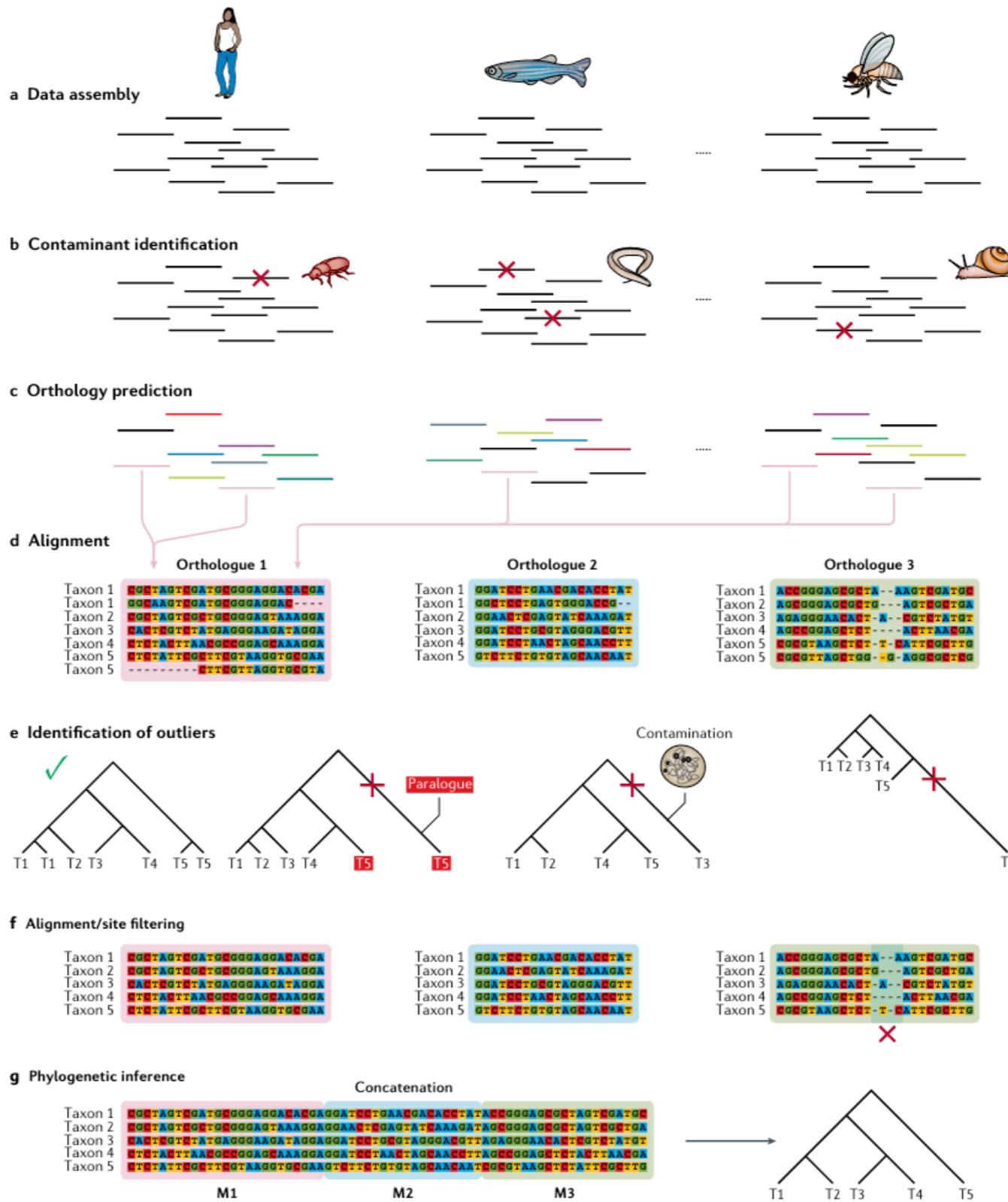
# Orthology and paralogy

Genes / molecular sequence may be homologous (share ancestry) for one of several reasons

- Speciation > orthologous (e.g. 1A–2A)
- Duplication > paralogous (e.g. 1A–1B or 2B)
- Horizontal gene transfer



Only orthologs faithfully reflect species events



Kapli et al. 2020 (Nature Reviews)