

---

## Lecture 1.1

---

# **Introduction to Molecular Phylogenetics**

---

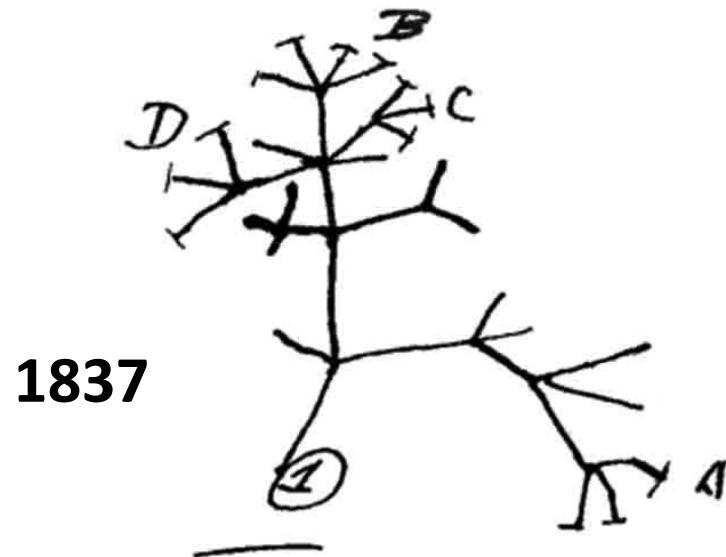
# Phylogenetic Trees

# What is a phylogenetic tree?

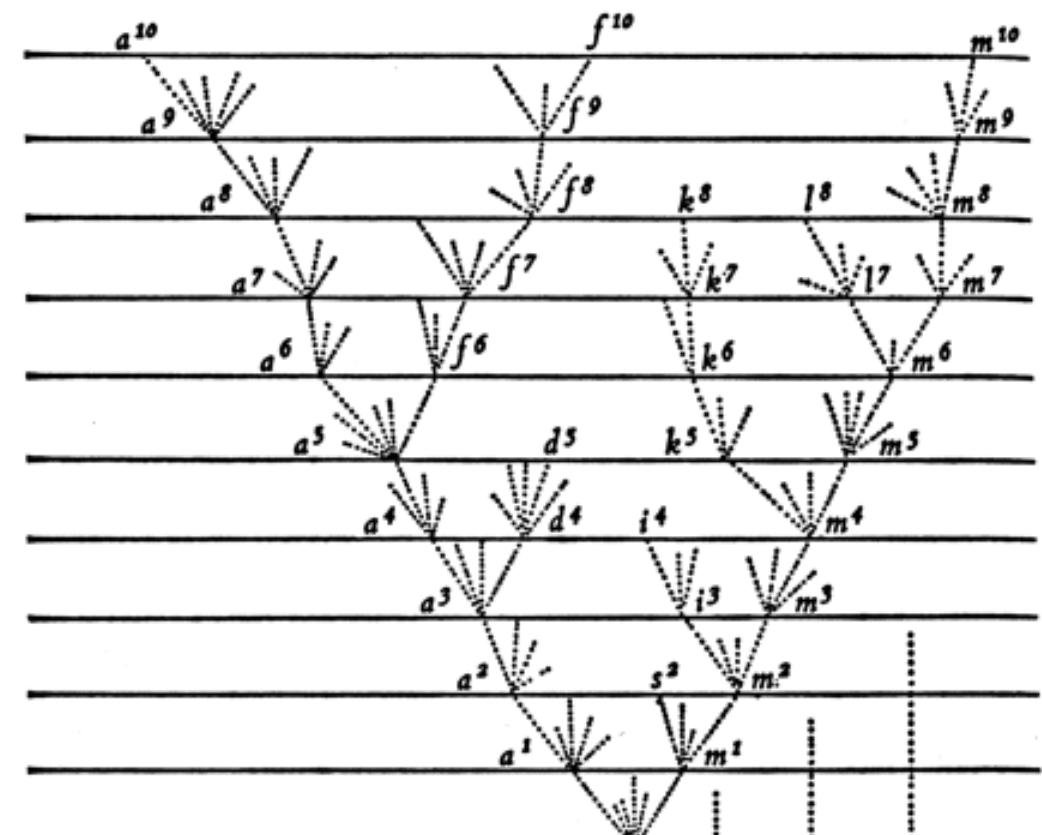
## Phylogeny

evolutionary relationships  
among a set of organisms

I think



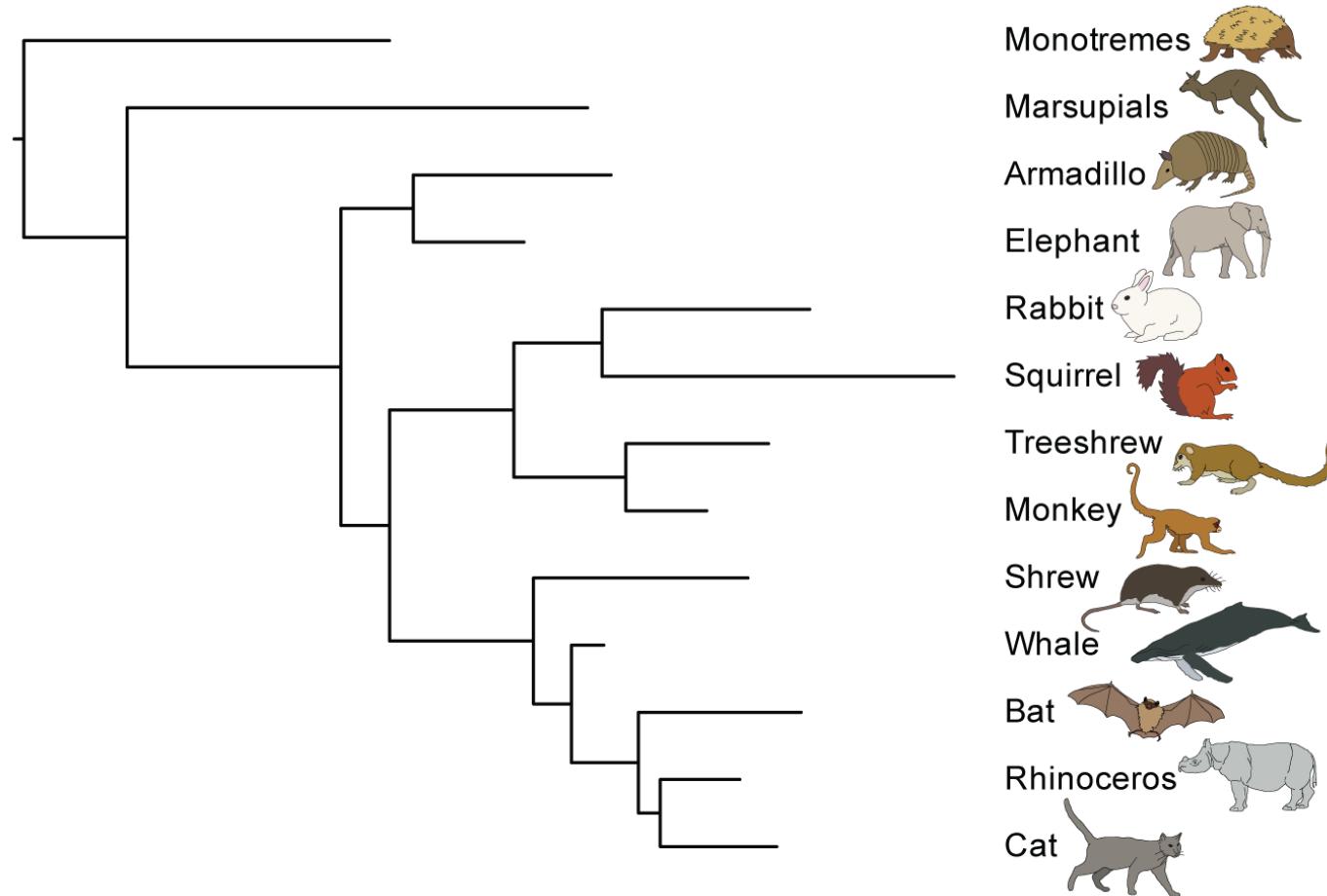
1837

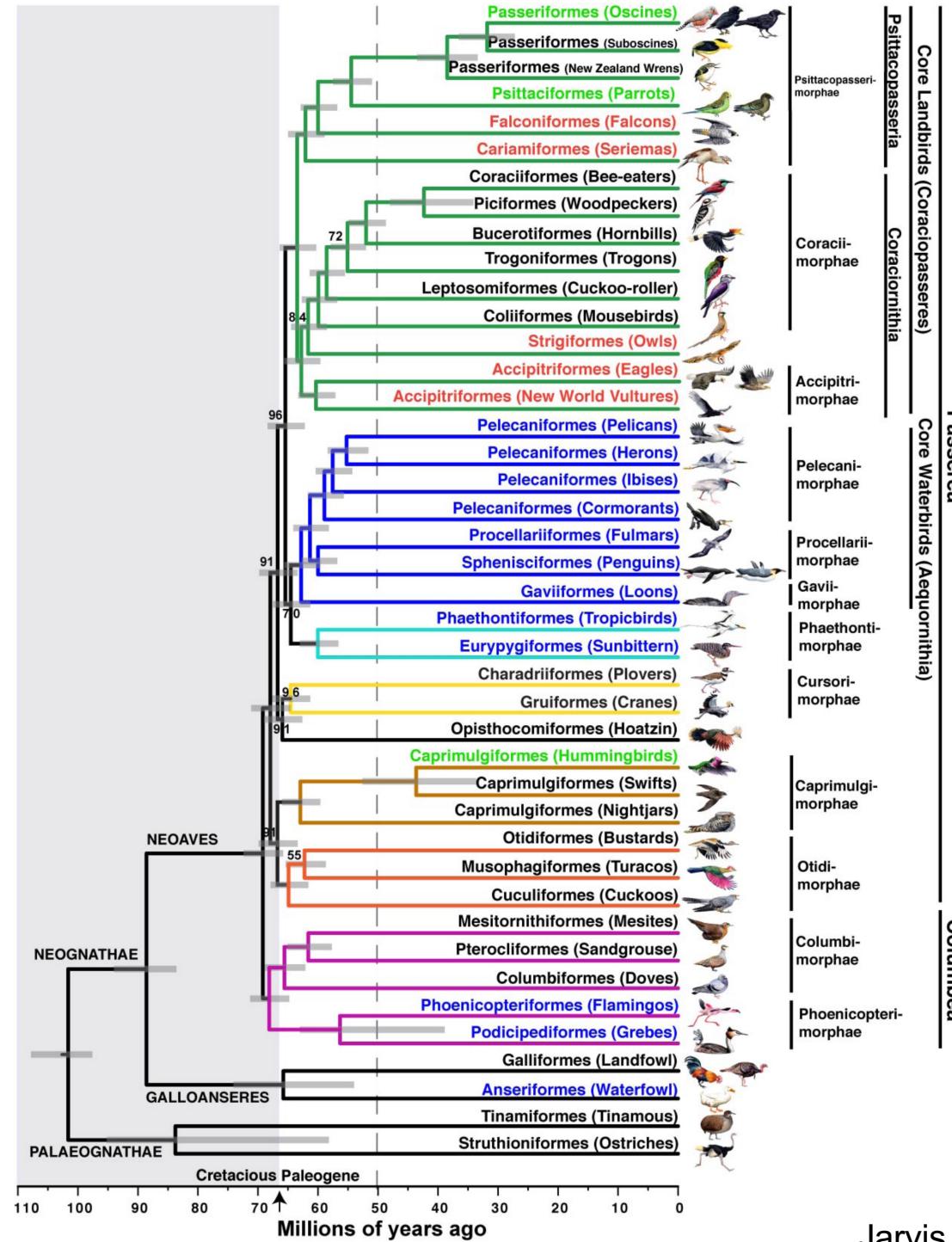


1859

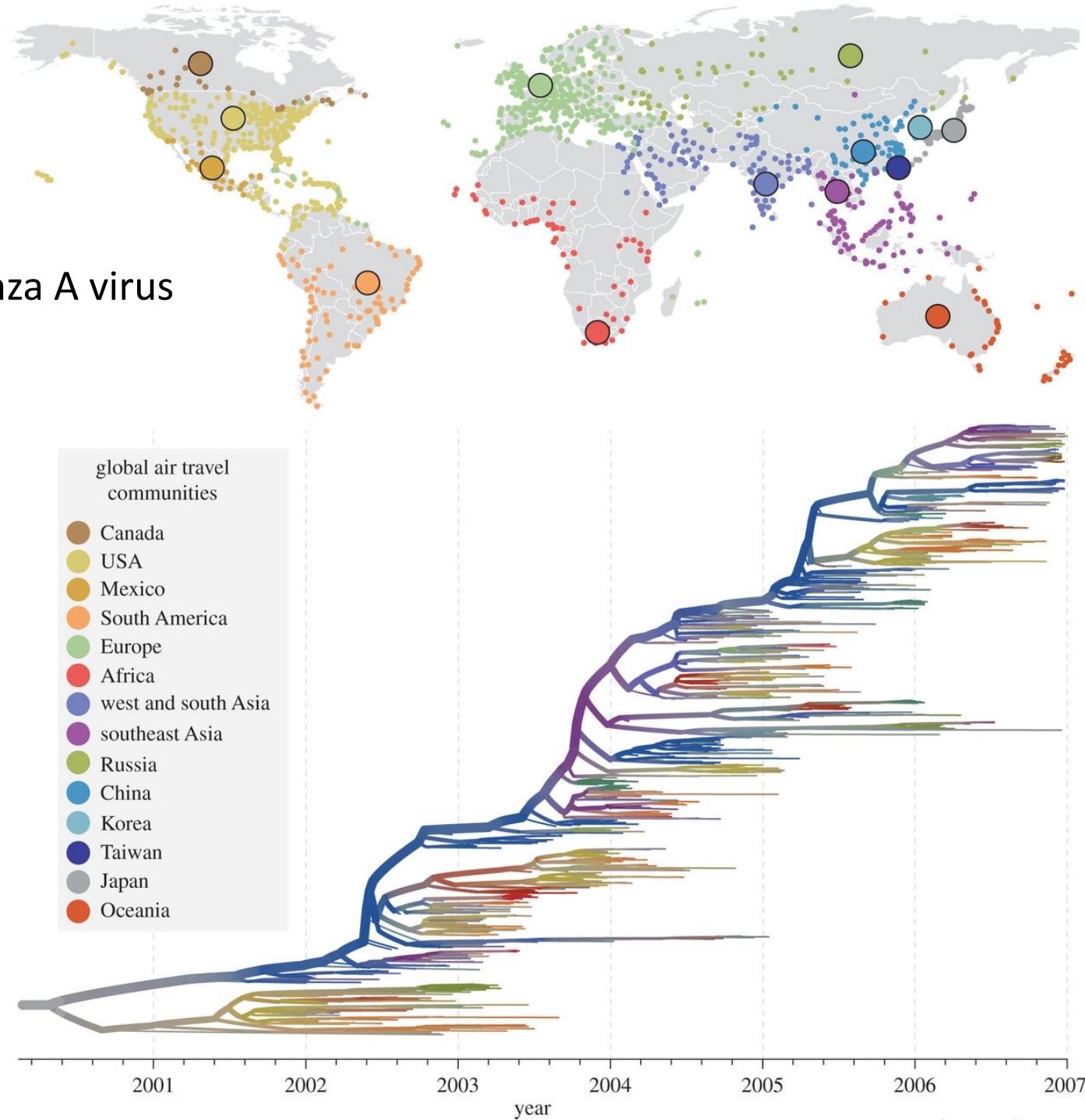
# Phylogenetic trees

- Topology (relationships)
- Branch lengths (amount of evolutionary change or time)





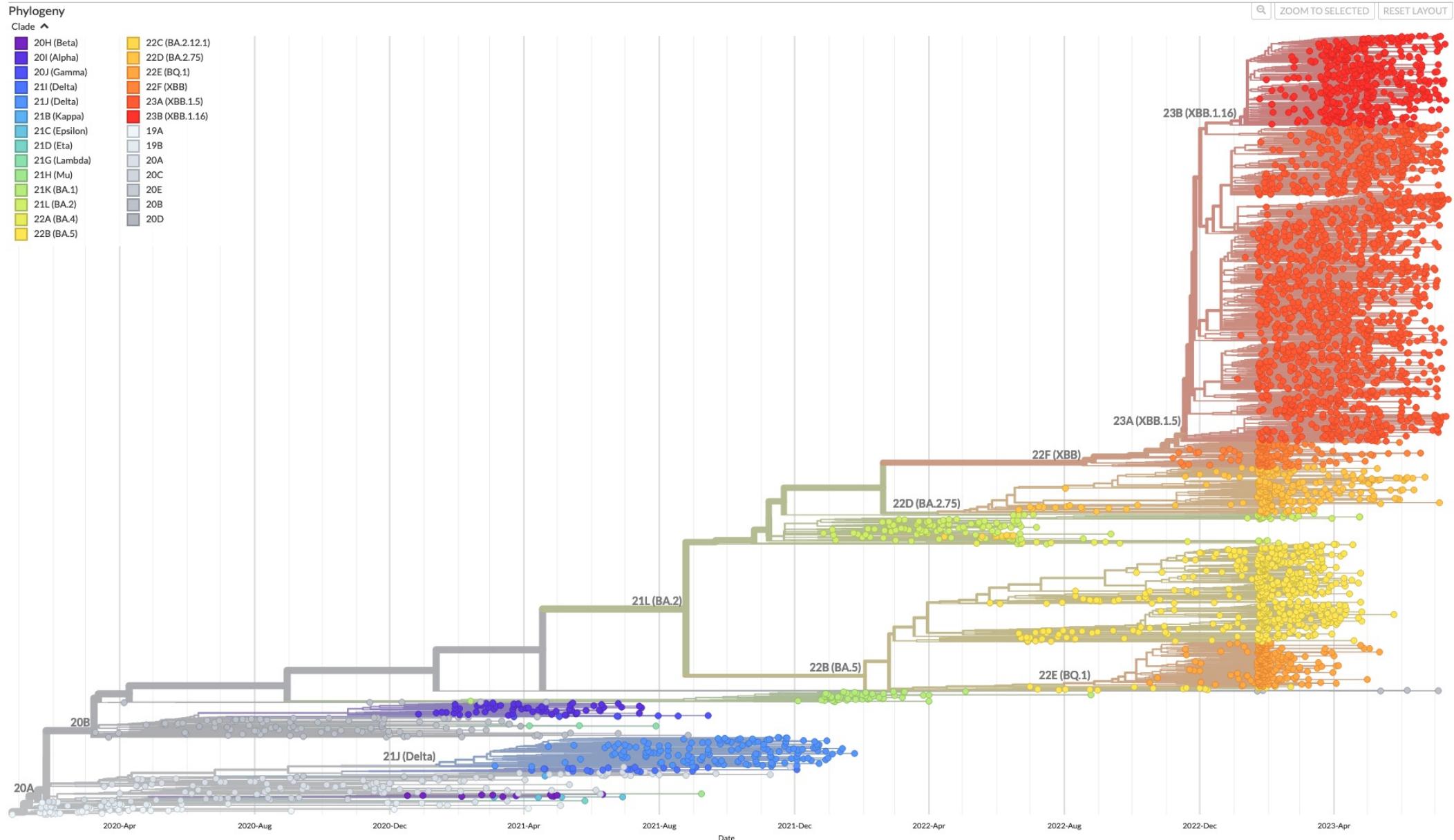
## Human influenza A virus Subtype H3N2



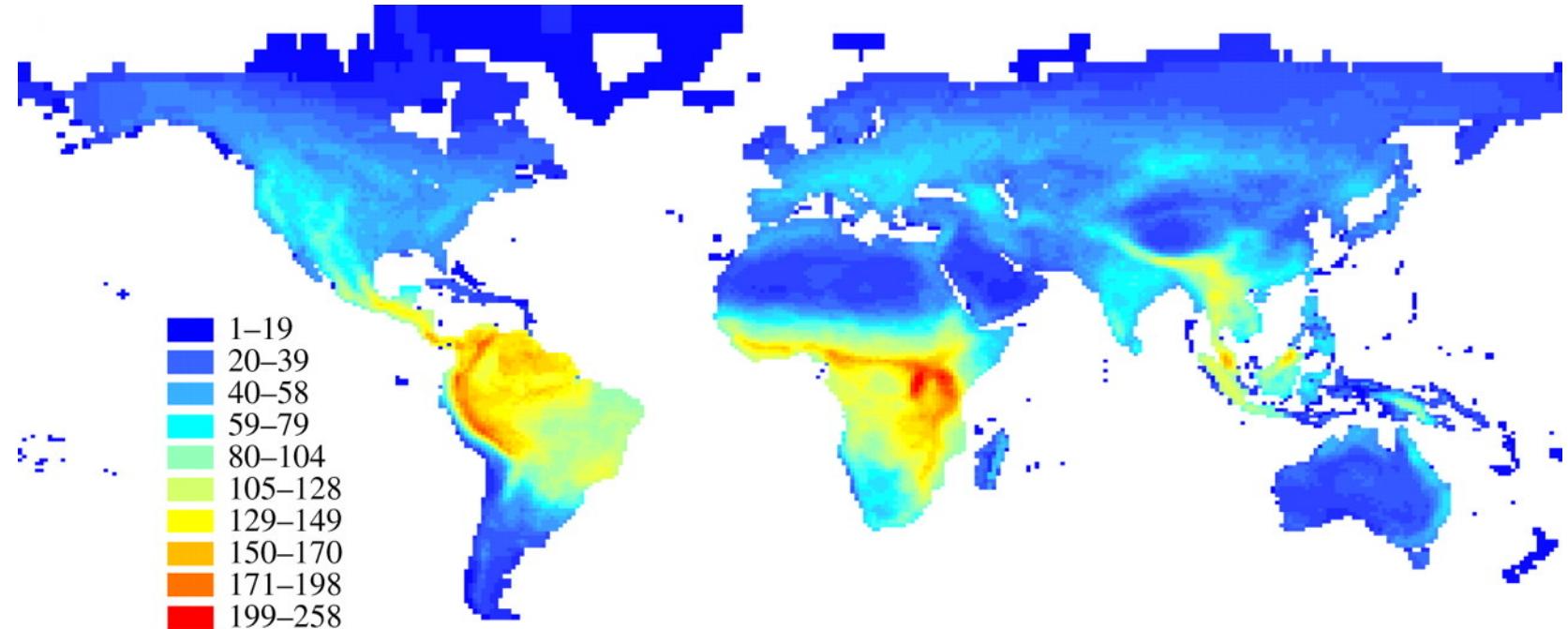
## Genomic epidemiology of SARS-CoV-2 with subsampling focused globally over the past 6 months

Built with nextstrain/ncov. Maintained by the Nextstrain team. Enabled by data from [GISAYD](#).

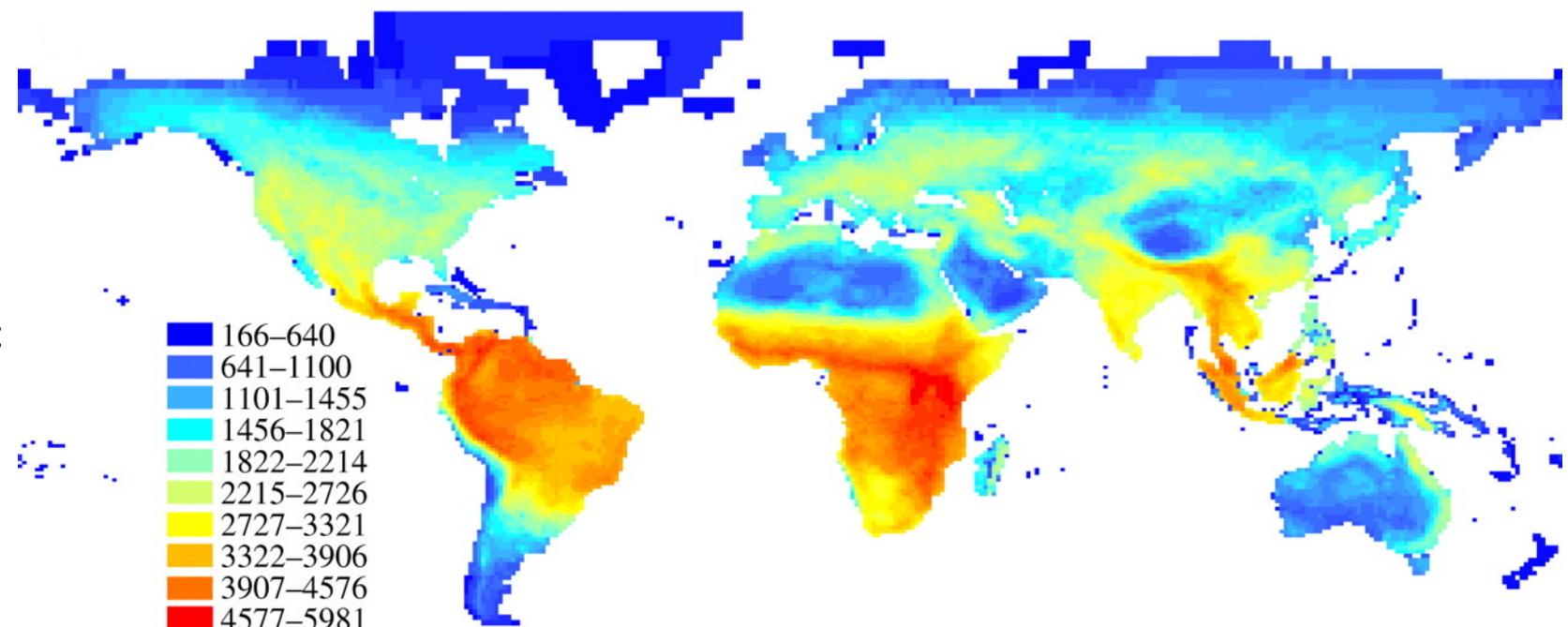
Showing 3801 of 3801 genomes sampled between Dec 2019 and Jul 2023.



Mammal  
species  
richness

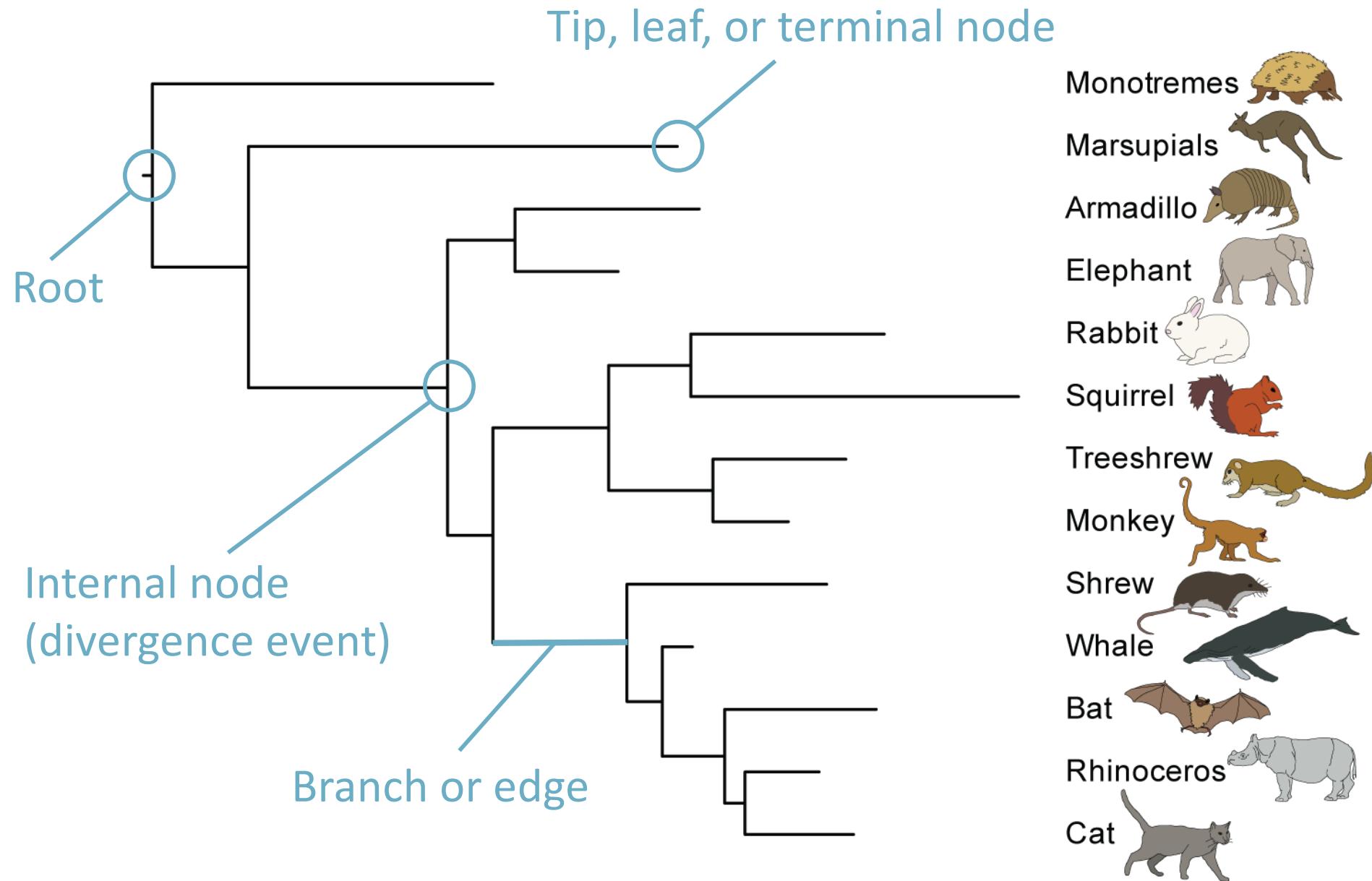


Mammal  
phylogenetic  
diversity

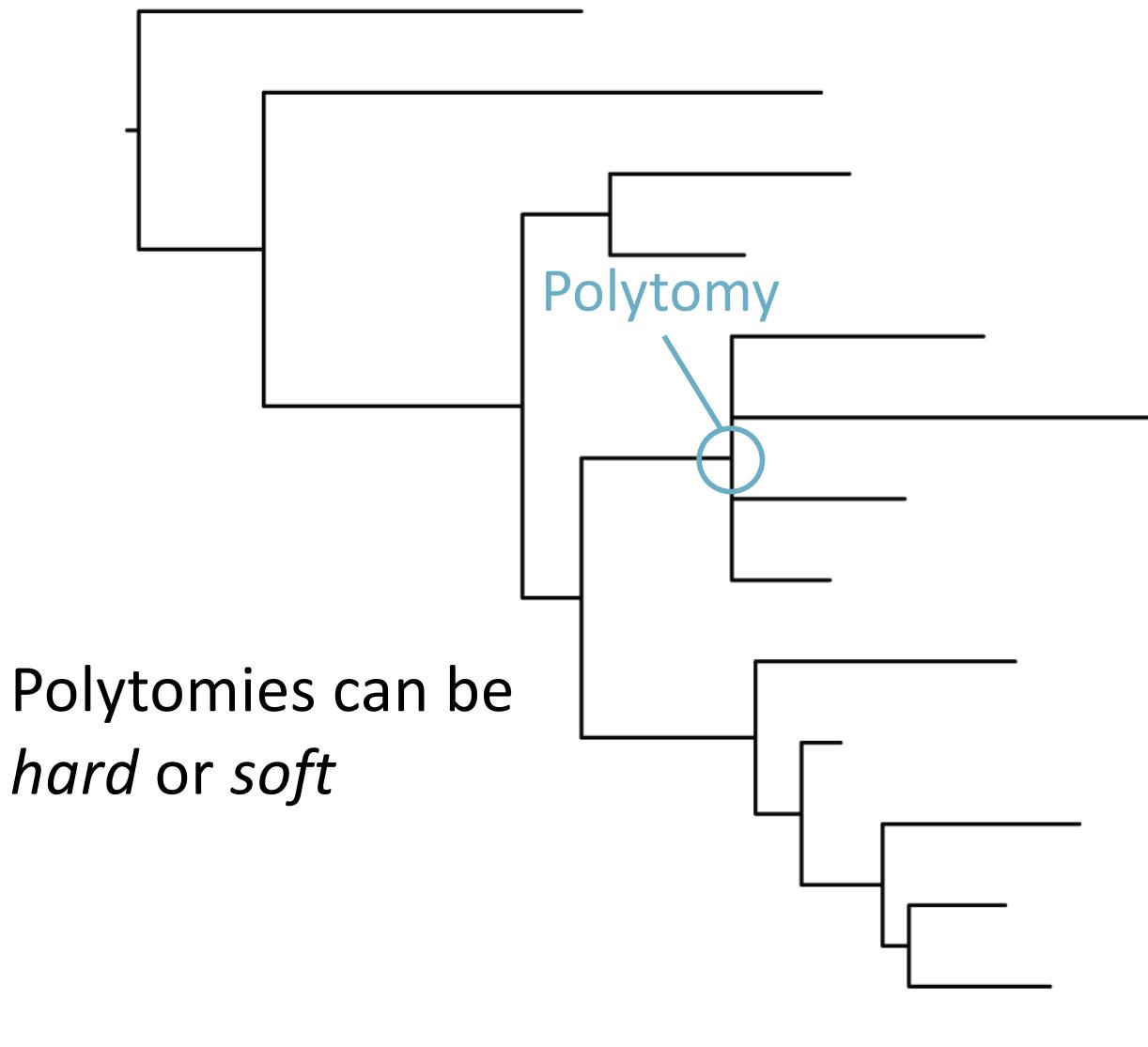


# Tree Thinking

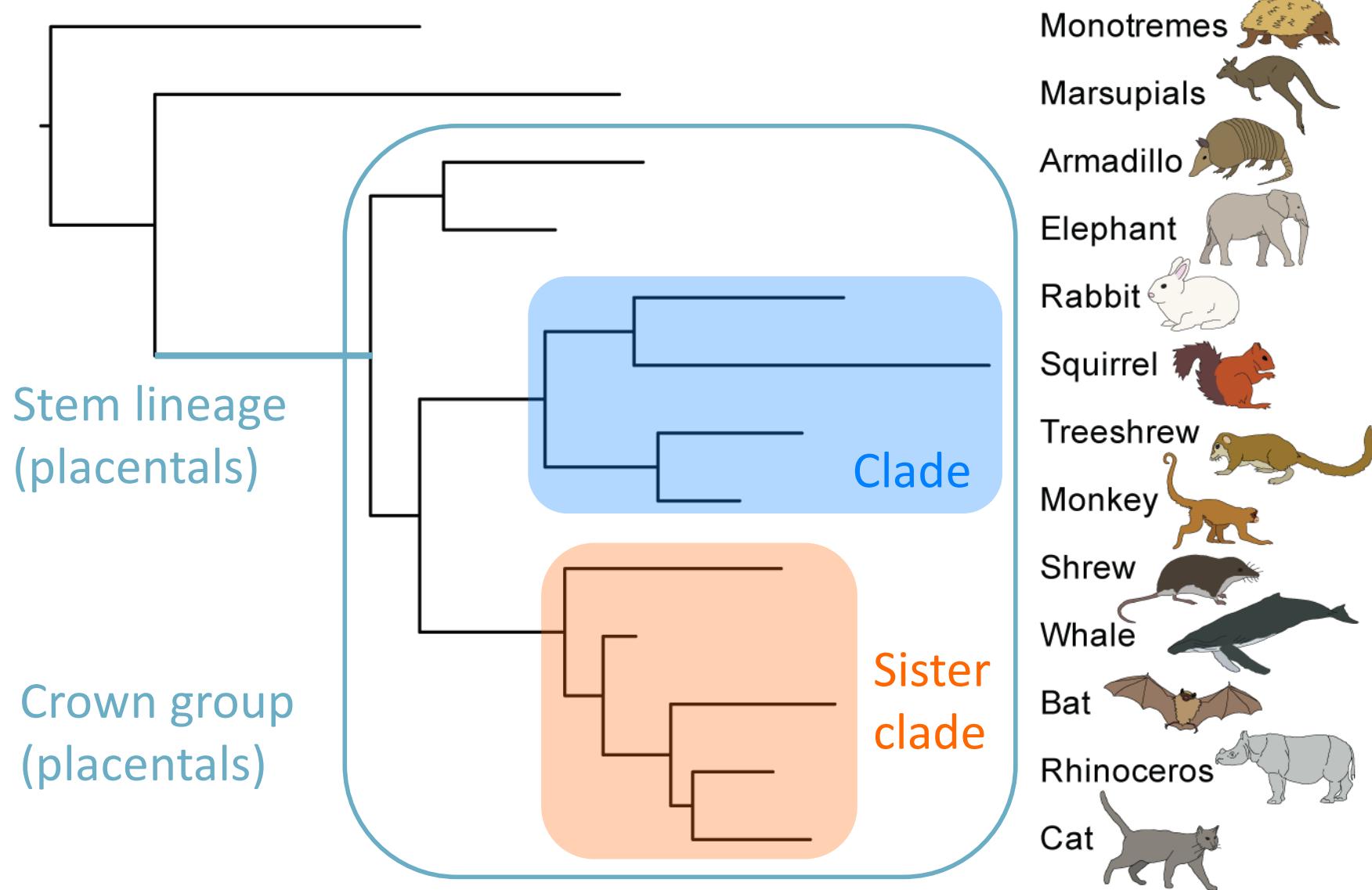
# Phylogenetic trees



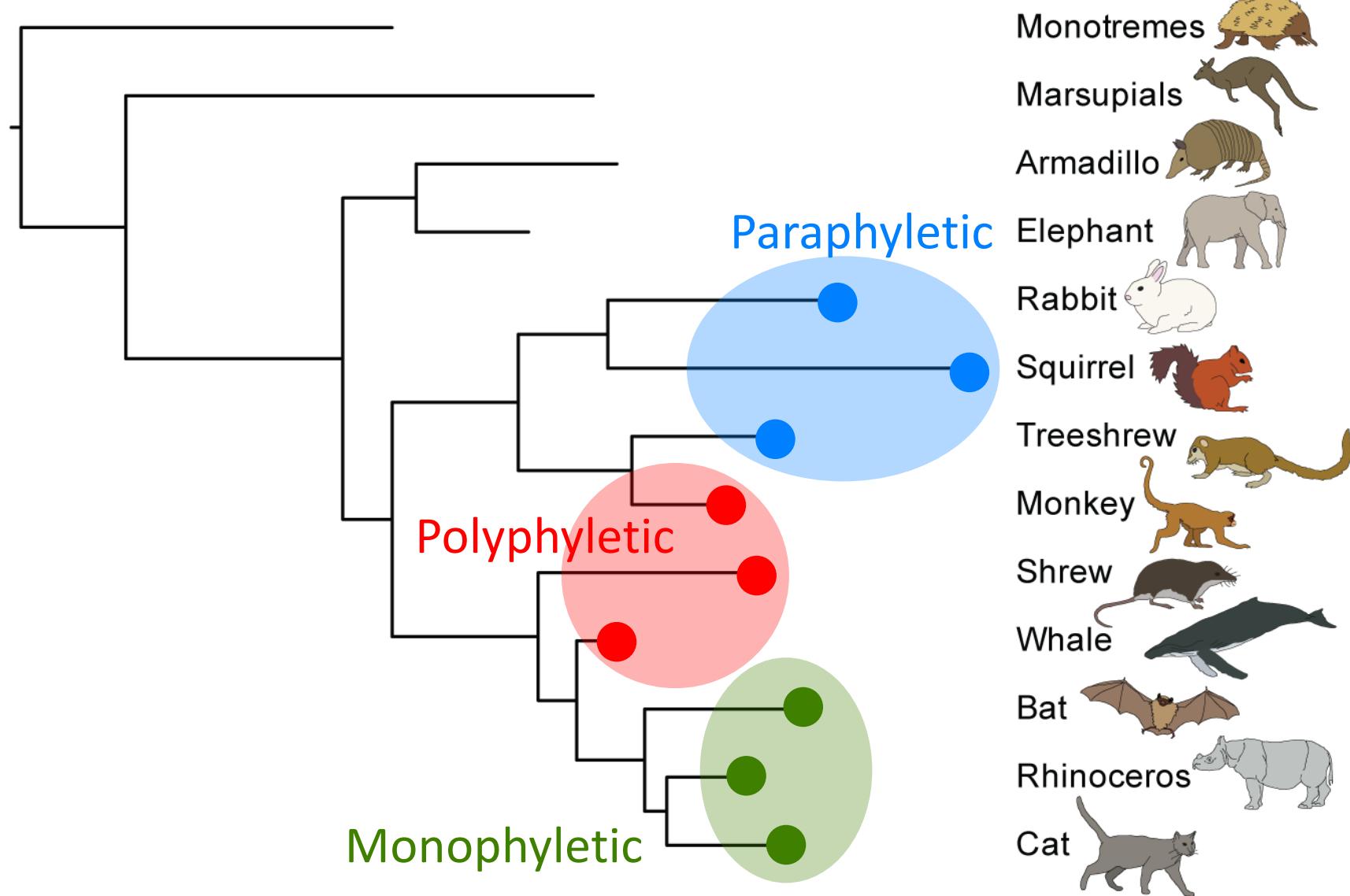
# Phylogenetic trees



# Phylogenetic trees

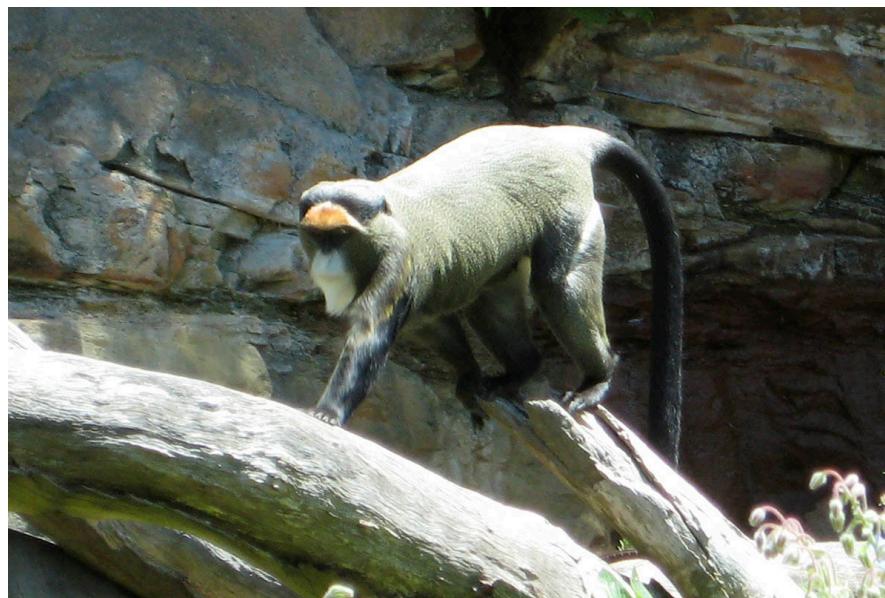


# Cladistic terms

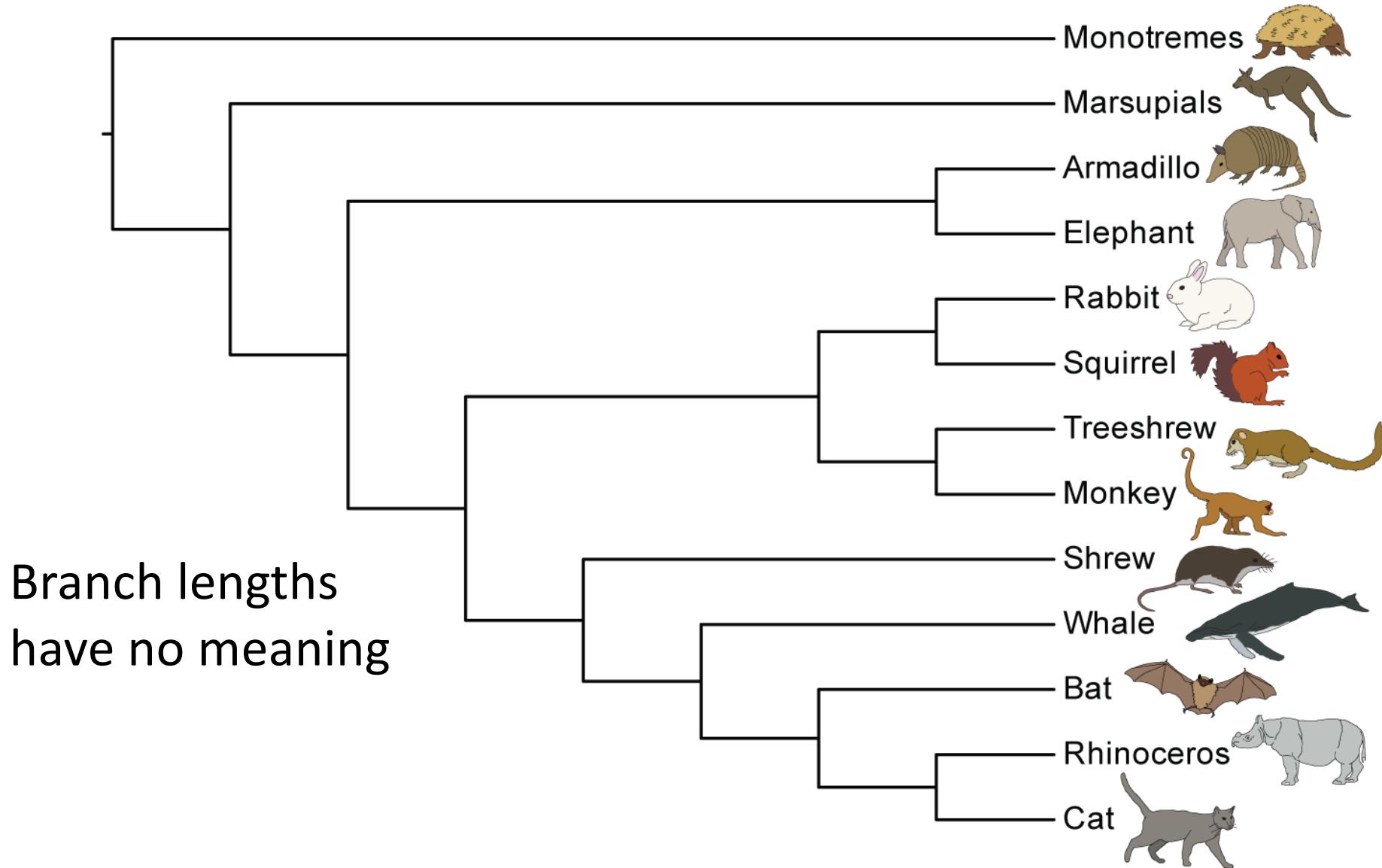


# Paraphyletic groups

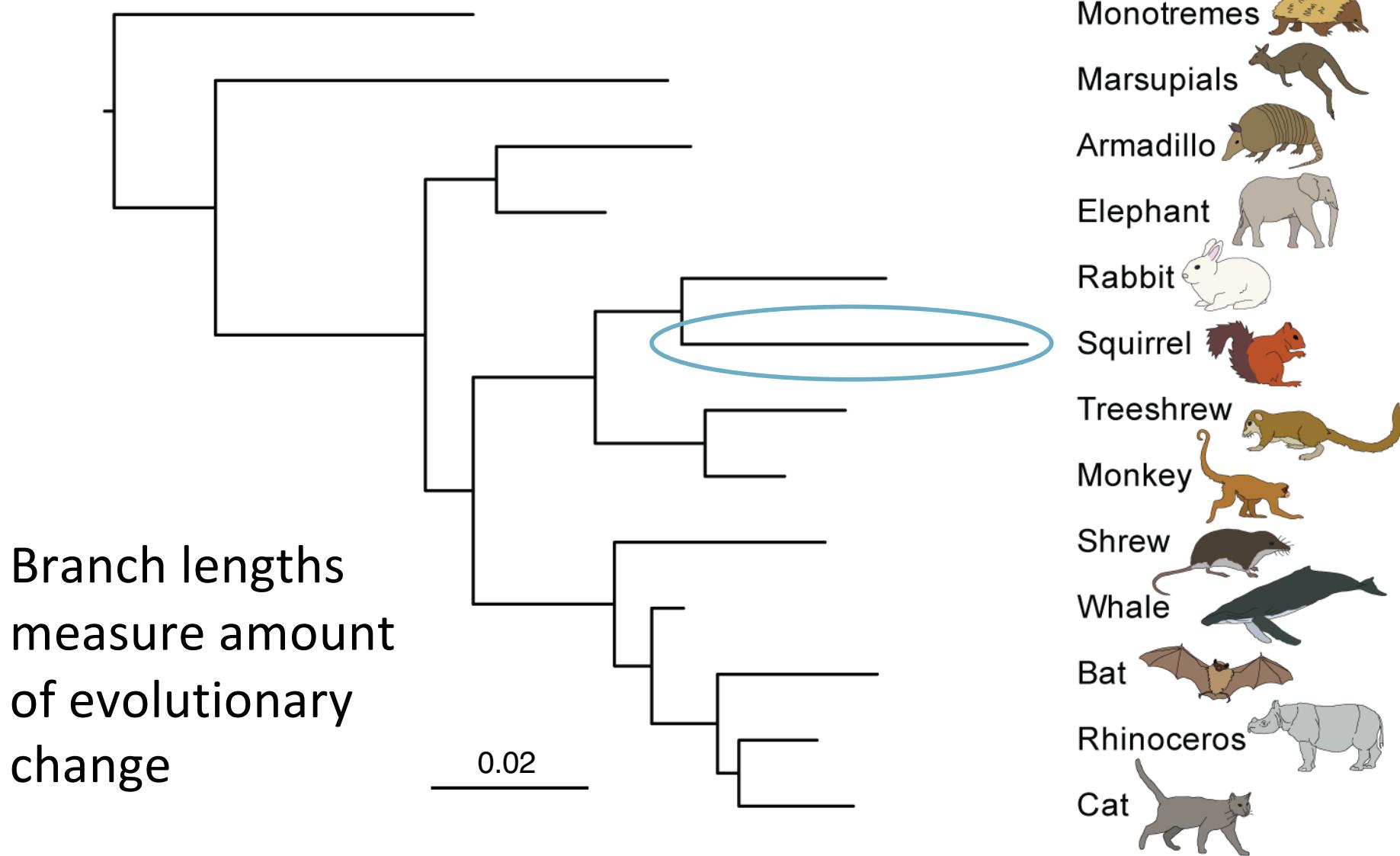
---



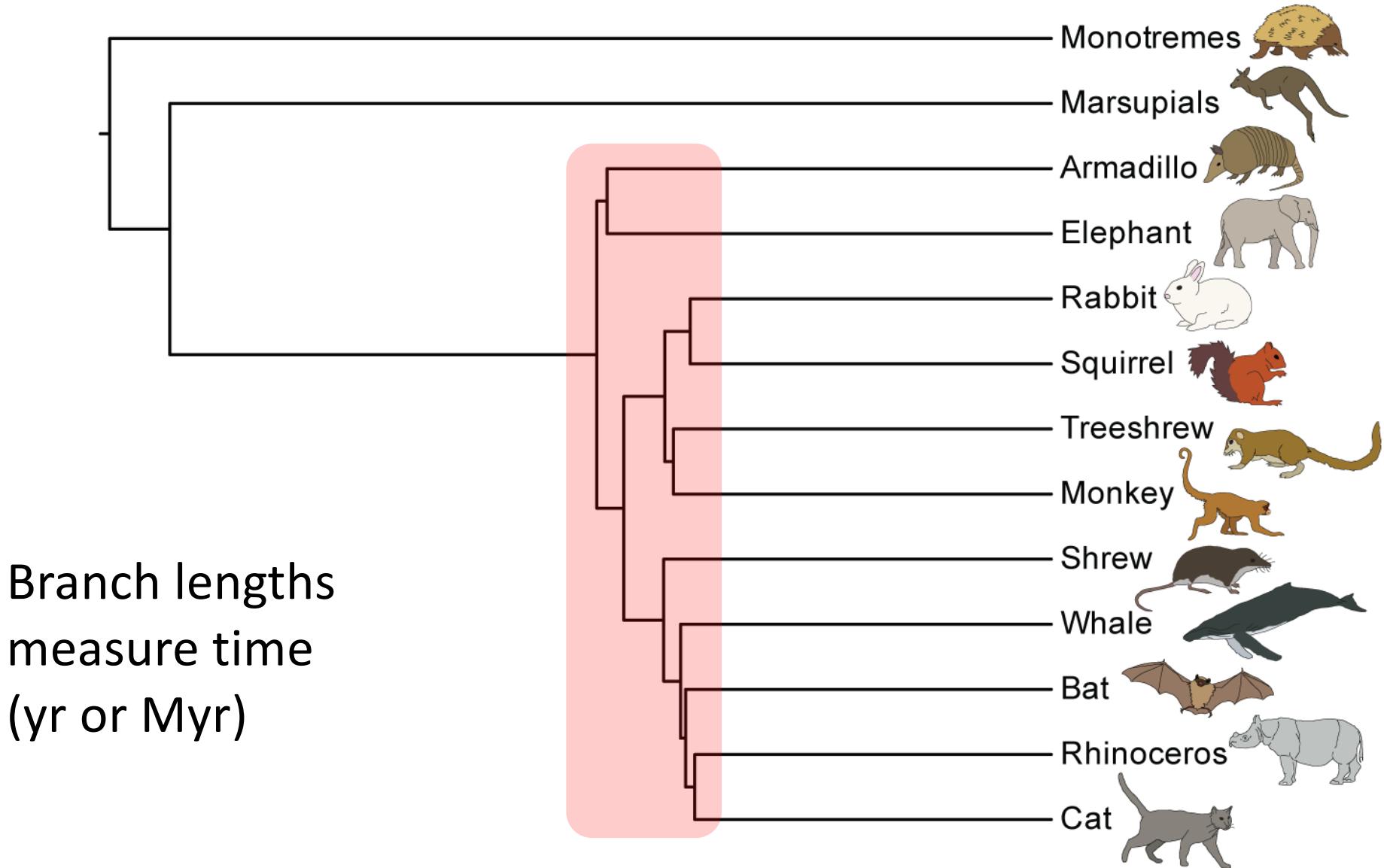
# Trees: Cladogram



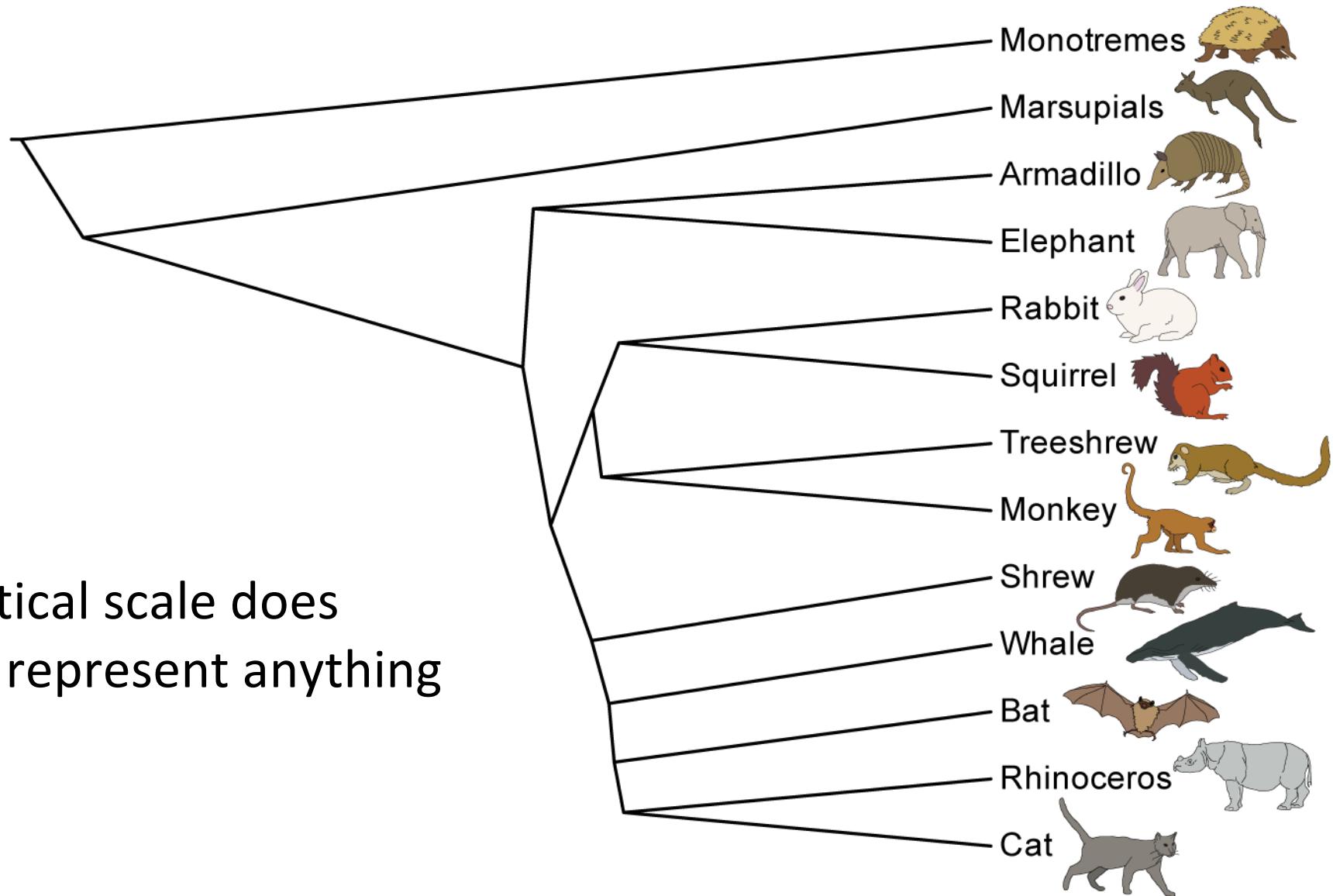
# Trees: Phylogram



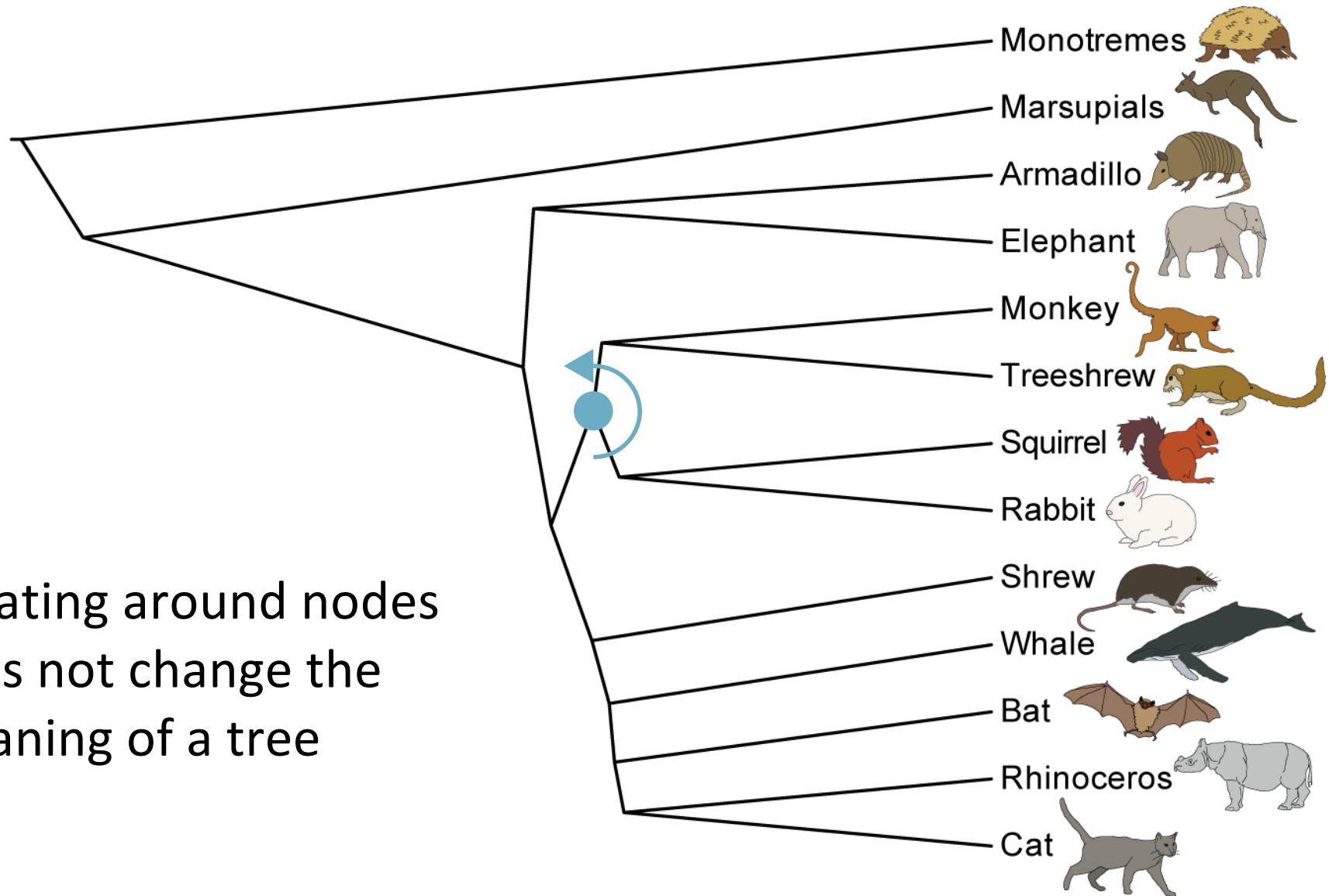
# Trees: Chronogram or time-tree



# Phylogenetic trees

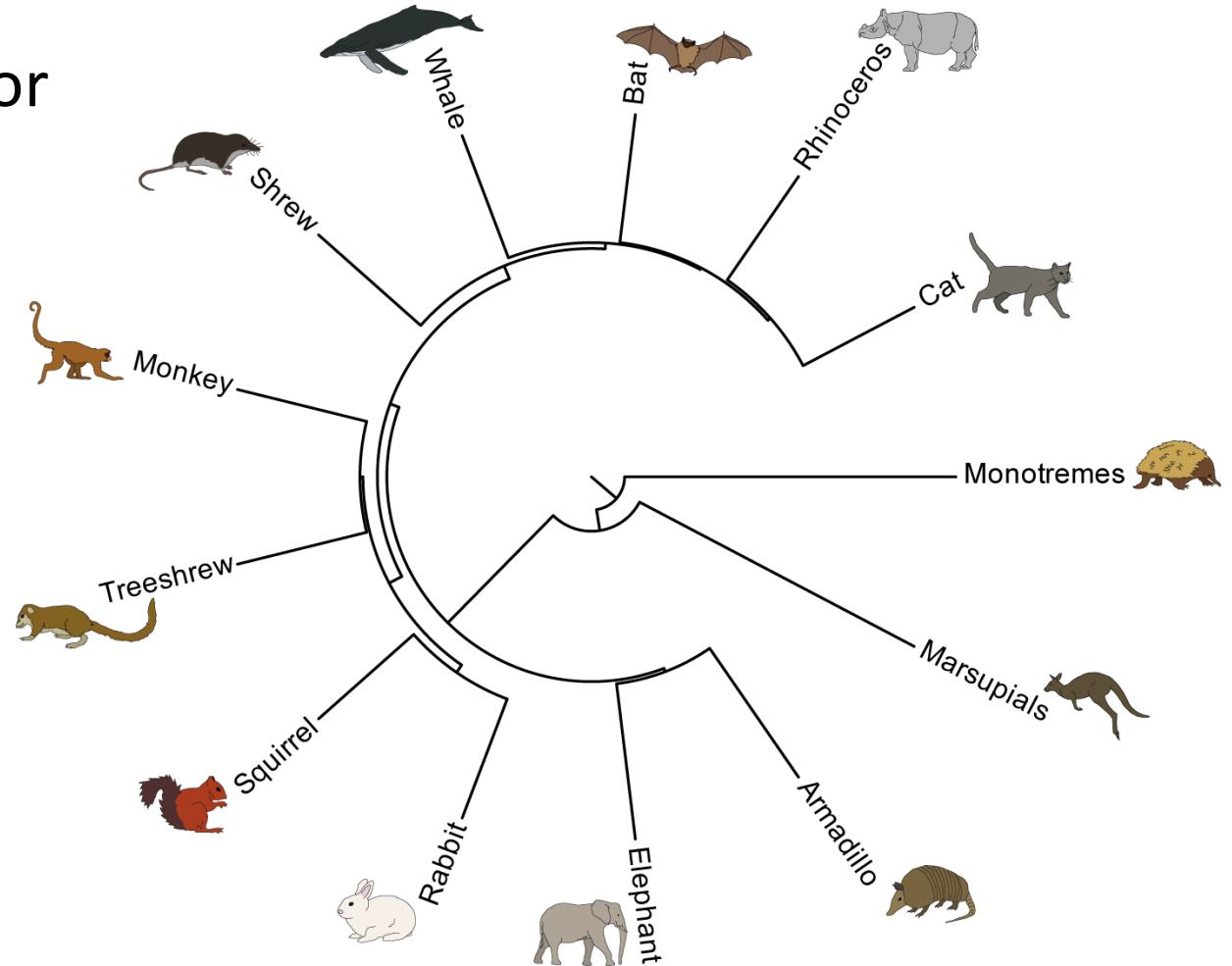


# Phylogenetic trees

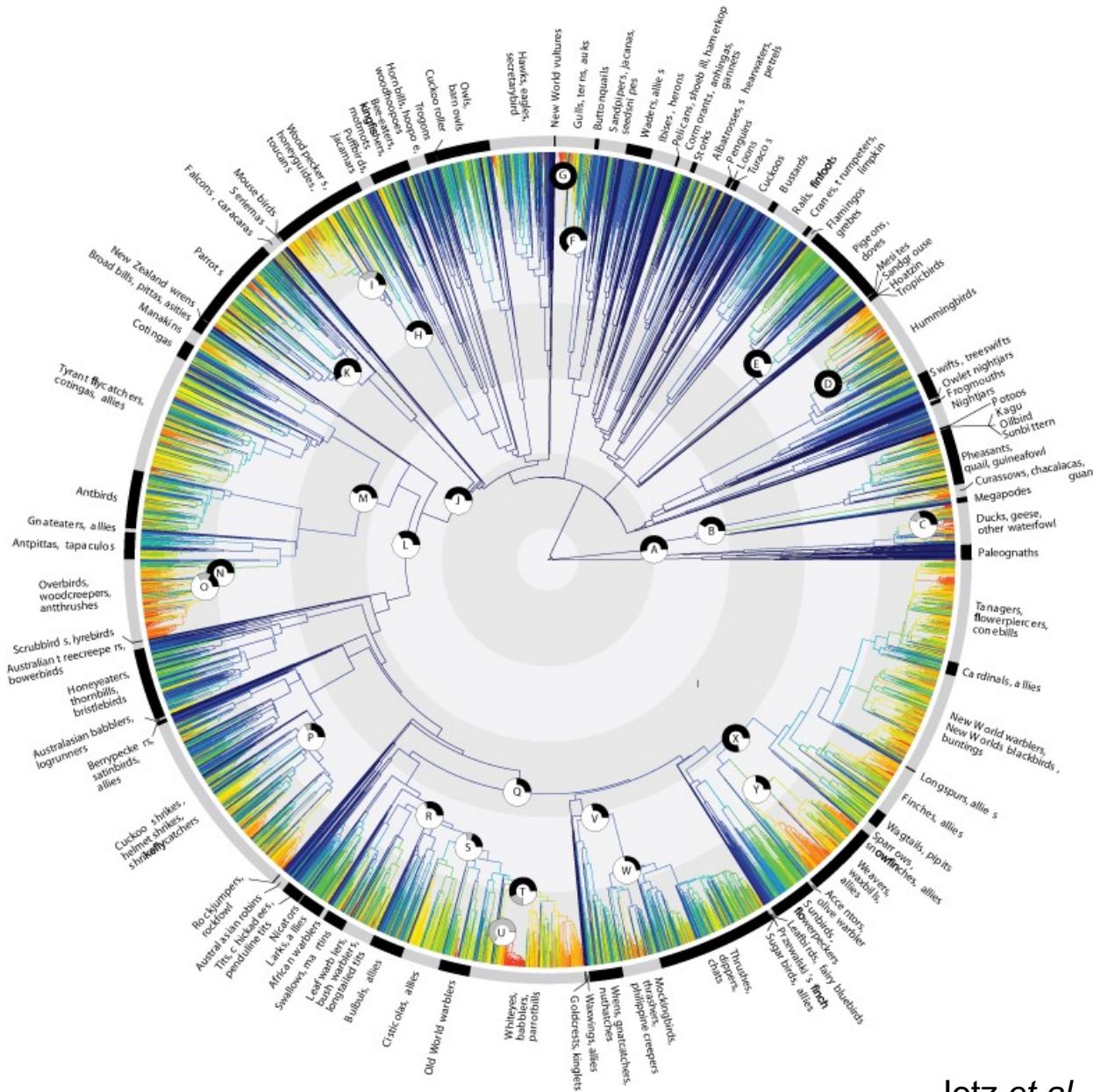


# Phylogenetic trees: Circular

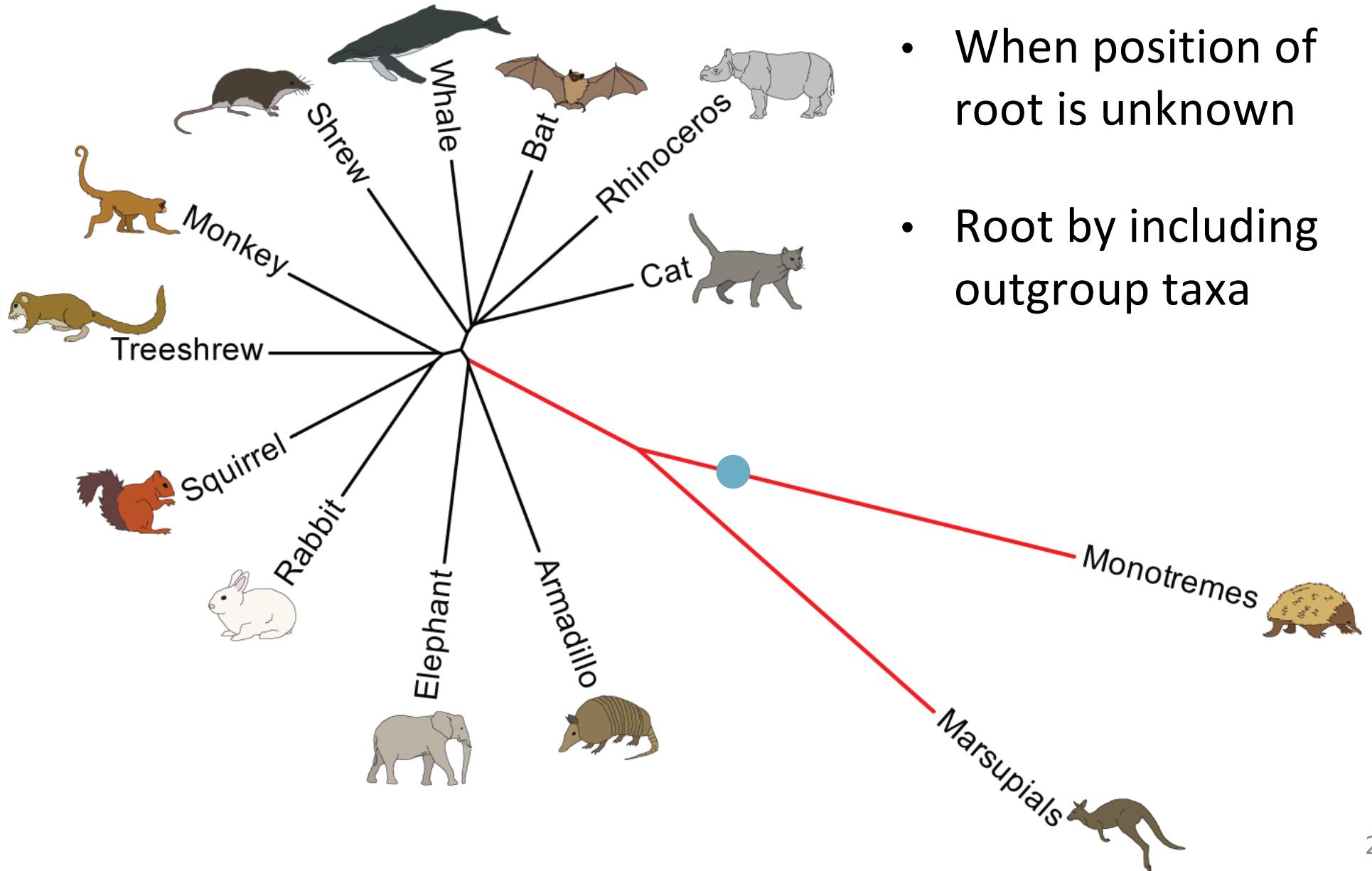
- Root is placed in centre
- Cladogram, phylogram, or chronogram
- Often used to display large trees
- Difficult to interpret



# Phylogenetic trees: Circular



# Phylogenetic trees: Unrooted



# Rooting

---

- **Include outgroup taxa**
  - Taxon closely related to ingroup
  - Taxon is not part of ingroup
- **Root at midpoint**
  - Highly unreliable if internal branches are short
- **Use a molecular clock**
  - Phylogenetic analysis infers a rooted tree

# Phylogenetic trees: Newick format

---

- Without branch lengths (cladogram):
  - (Monotremes,(Marsupials,((Elephant,Armadillo),(((Squirrel,Rabbit),(Monkey,Treeshrew)),(Shrew,(Whale,(Bat,(Cat,Rhinoceros)))))));
- With branch lengths (phylogram/chronogram):
  - (Monotremes:12.0,(Marsupials:11.0,((Elephant:1.0,Armadillo:1.0):9.0,(((Squirrel:1.0,Rabbit:1.0):2.0,(Monkey:1.0,Treeshrew:1.0):2.0):5.0,(Shrew:4.0,(Whale:3.0,(Bat:2.0,(Cat:1.0,Rhinoceros:1.0):1.0):1.0):1.0):4.0):2.0):1.0);

# Molecular Phylogenetics

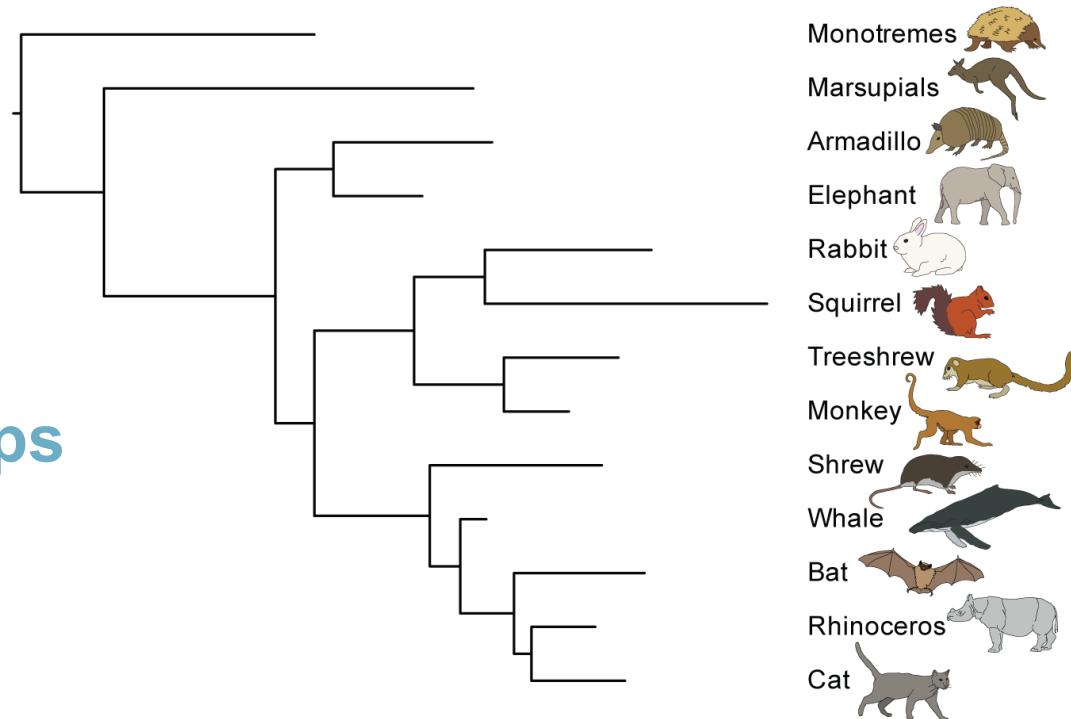
# Phylogenetic analysis

---

- We rarely know the actual evolutionary history of a set of individuals or species
  - Viral transmission histories
  - Pedigrees (humans, domesticated animals, lab organisms, *etc.*)
- Usually we must analyse data from present-day organisms to infer phylogenies

# Fundamental assumptions

- Phylogenetic methods make several fundamental assumptions:
  - Relationships among taxa can be represented by a tree
  - Homologous characters are being compared
  - Characters are mutually independent

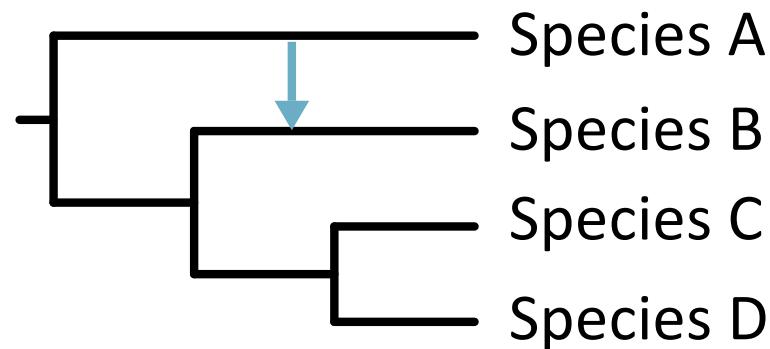


When might relationships  
not be treelike?

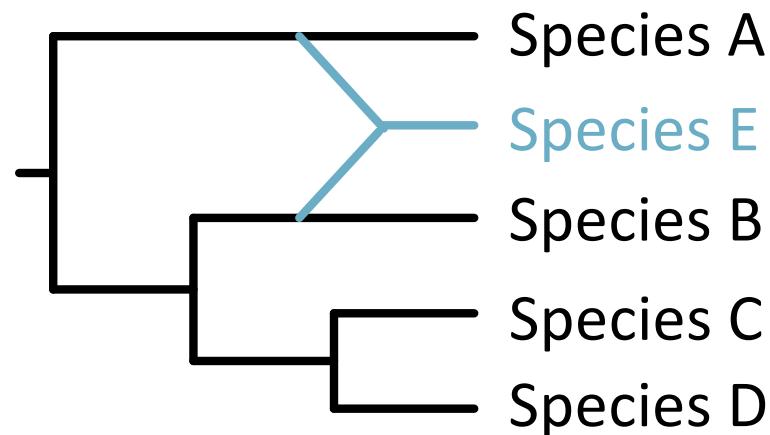
# Non-treelike evolution

---

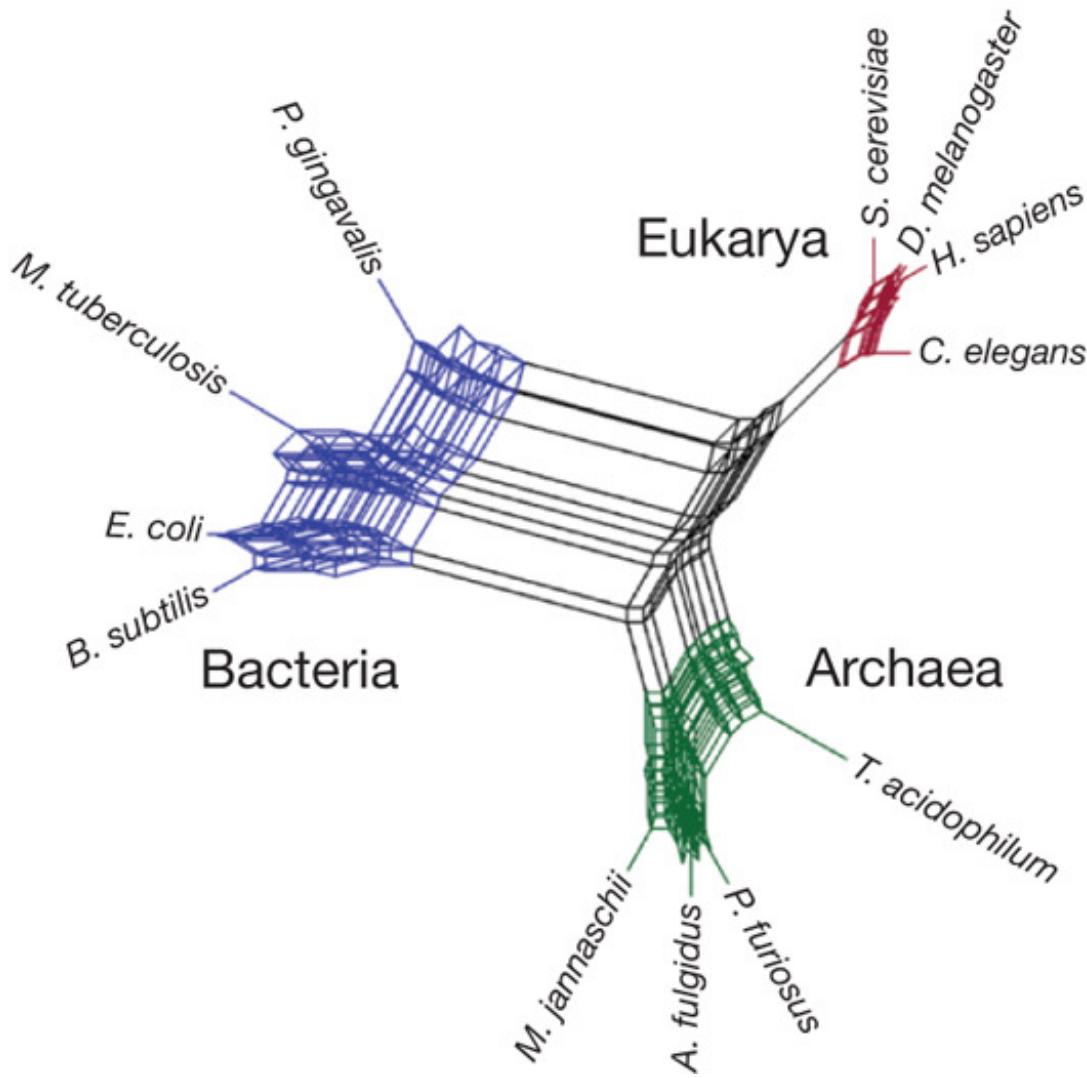
Horizontal gene transfer



Hybrid speciation

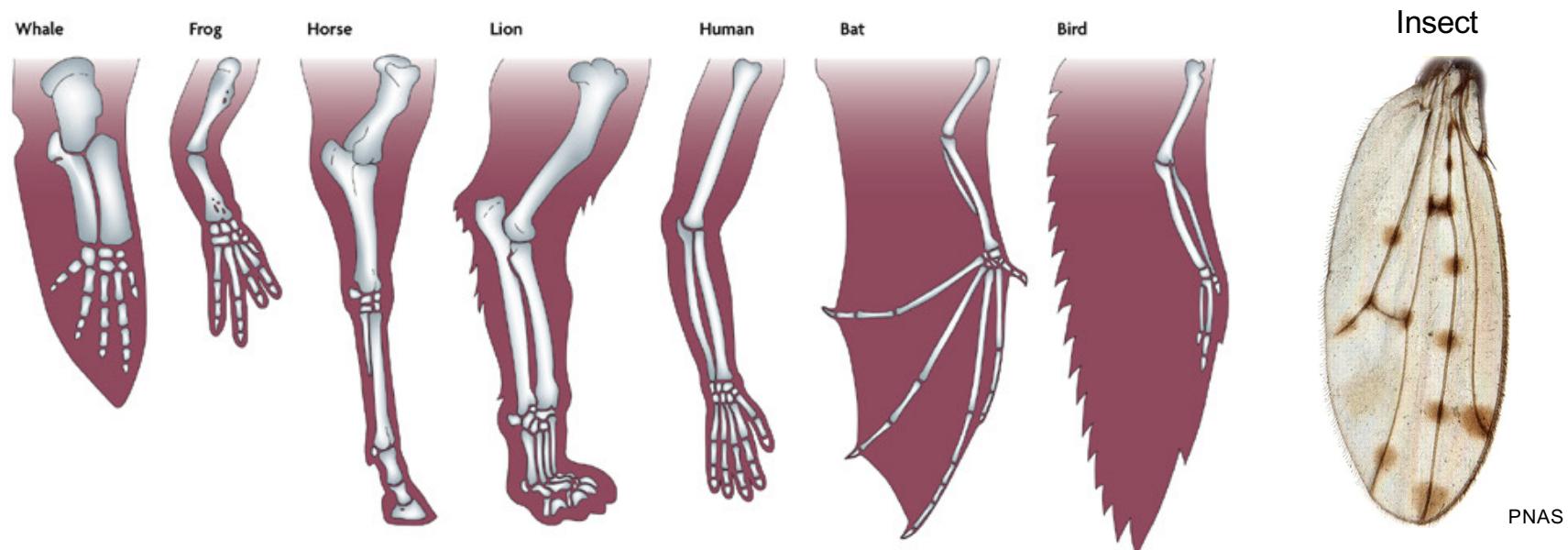


# Phylogenetic networks



# Fundamental assumptions

- Phylogenetic methods make several fundamental assumptions:
  - Relationships among taxa can be represented by a tree
  - Homologous characters are being compared
  - Characters are mutually independent



# Character homology

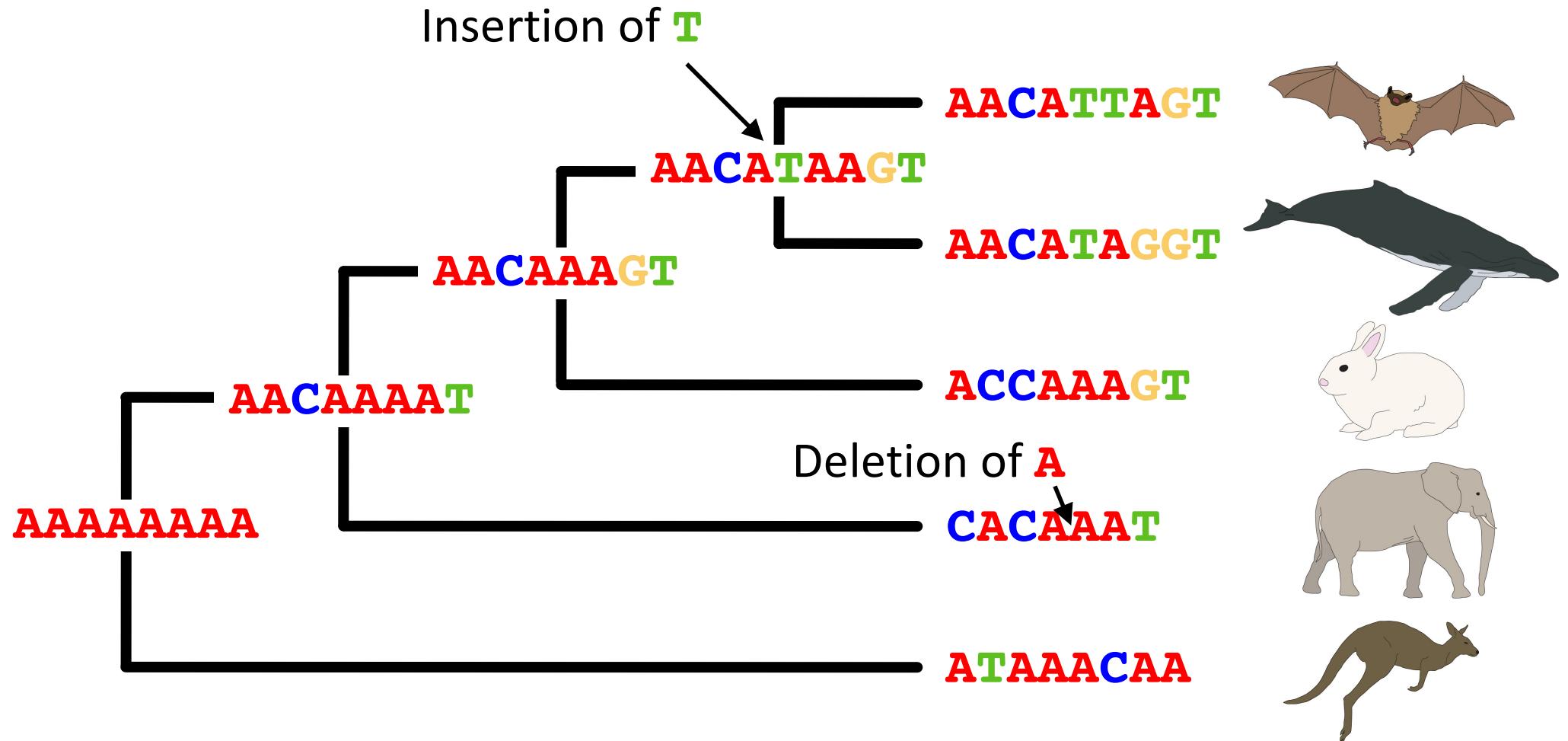
- Comparing strings of nucleotides (or amino acids)
- Each nucleotide site is a character
- But DNA sequences can vary in length

bat  
whale  
rabbit  
elephant  
kangaroo

**CGTTAGTACACT**  
**CGATAGTTCACT**  
**CGTTAGTTTACC**  
**CATTGGATTACT**  
**CATTGGTTTACT**



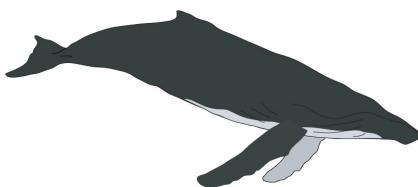
# DNA sequence evolution



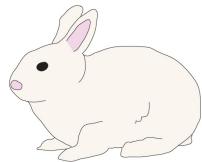
# DNA sequence alignment



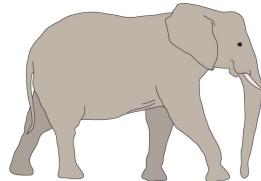
**AACAT<sub>T</sub>TAGT**



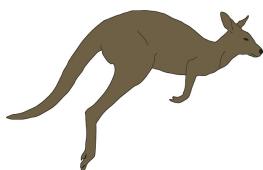
**AACAT<sub>A</sub>TAGGT**



**ACCAAAAGT**



**CACAAAT**



**ATAAAACAA**



**AACATTAGT**

**AACATAGGT**

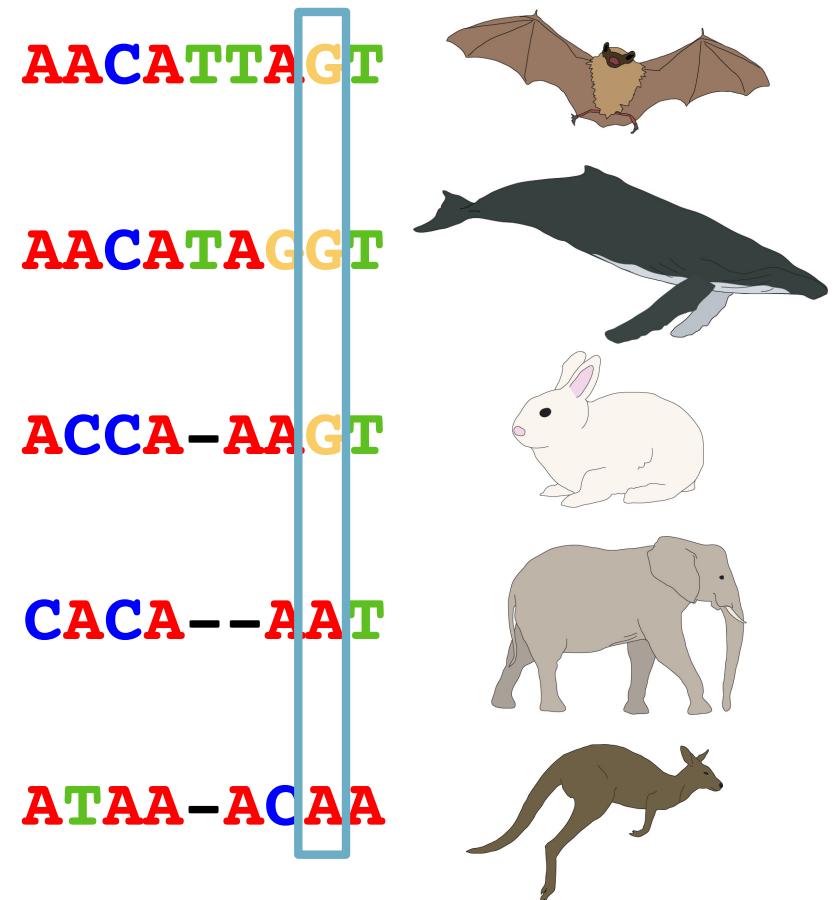
**ACCA-AAGT**

**CACA--AAT**

**ATAA-ACAA**

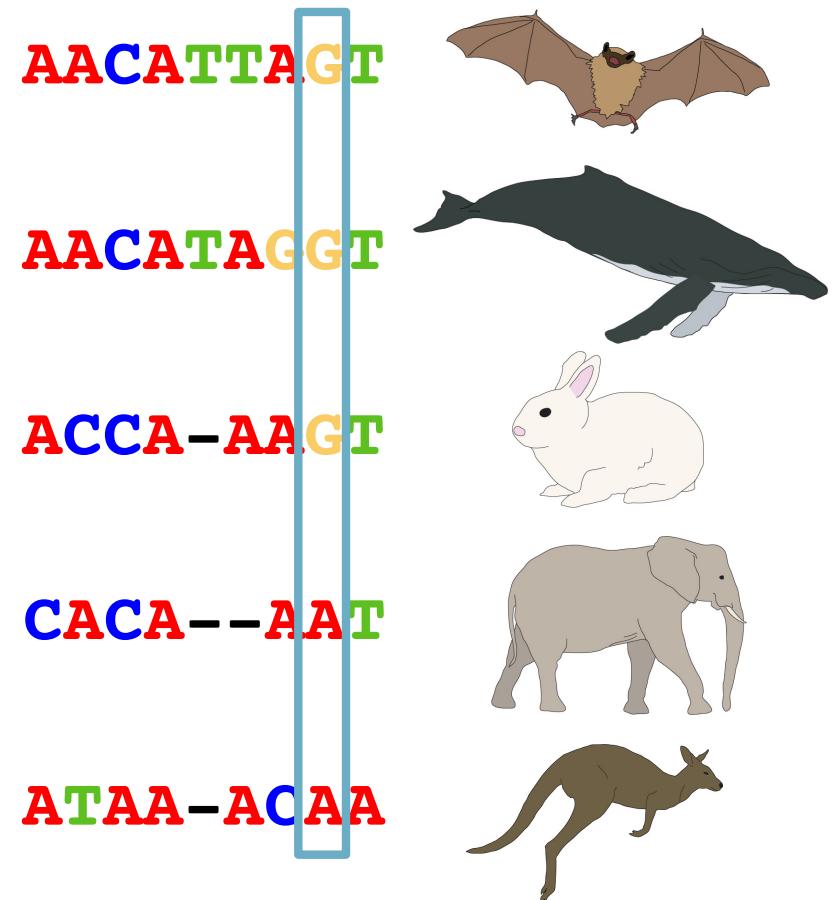
# DNA sequence alignment

- Homologous site
- Inherited from the common ancestor of all sequences in the alignment
- The aim of sequence alignment is to maximise the number of sites for which you can infer homology



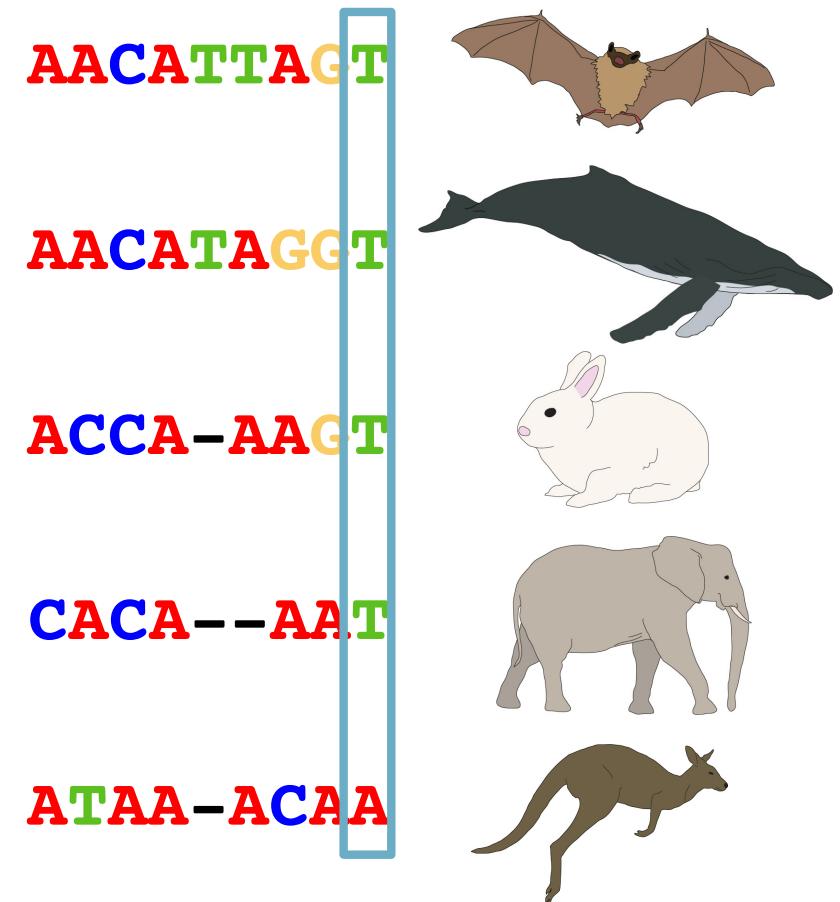
# DNA sequence alignment

- Groups together the first 3 sequences
- Groups together the last 2 sequences
- Informative for all phylogenetic methods



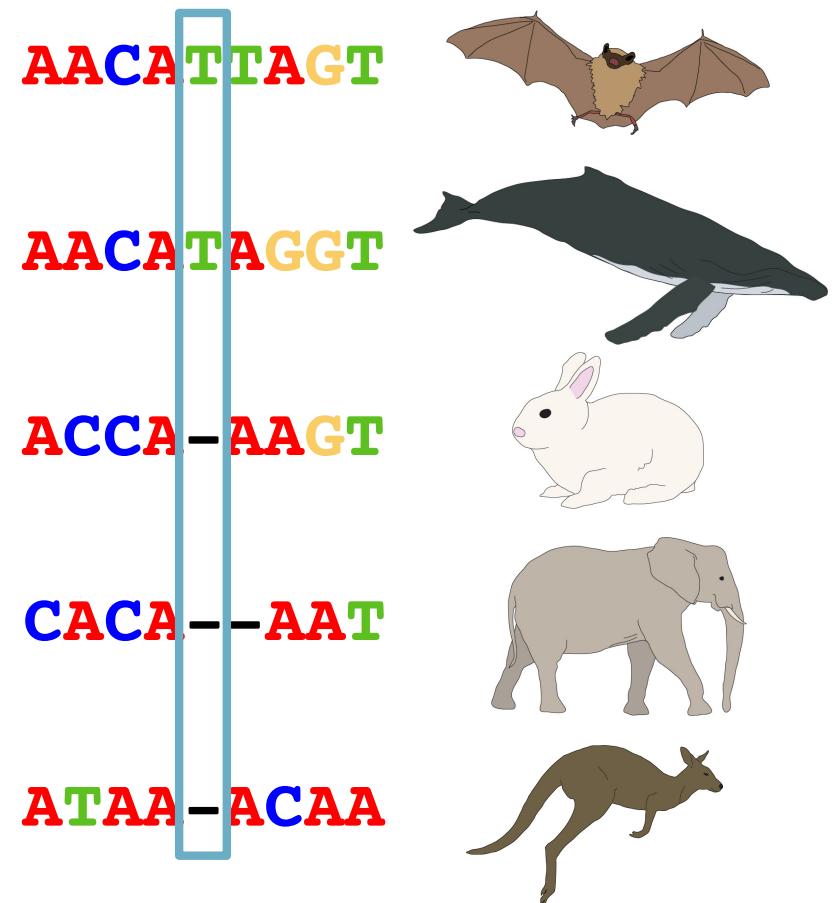
# DNA sequence alignment

- Does not group any sequences
  - Not useful for maximum parsimony
- But informative for estimating amount of evolutionary change
  - Useful for other methods



# DNA sequence alignment

- Indel – insertion or deletion
- Potentially informative
- Most phylogenetic methods do not use indel data



# A practical approach

---

Align sequences using automated methods

---

**CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**

---

Julie D.Thompson, Desmond G.Higgins<sup>+</sup> and Toby J.Gibson\*

Software

**MUSCLE: a multiple sequence alignment method with reduced time and space complexity**  
Robert C Edgar\*

**MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform**

Kazutaka Katoh, Kazuharu Misawa<sup>1</sup>, Kei-ichi Kuma and Takashi Miyata\*

# A practical approach

Align sequences using automated methods



Adjust alignments by eye

CTATGTGGCACCCAGCCCCATGCA--AGC

ATATGTGGCA-----CCCAGGCA--AG-

ATATGTGGCACCCAGCCCCATGCATTT--

# A practical approach

Align sequences using automated methods



Adjust alignments by eye

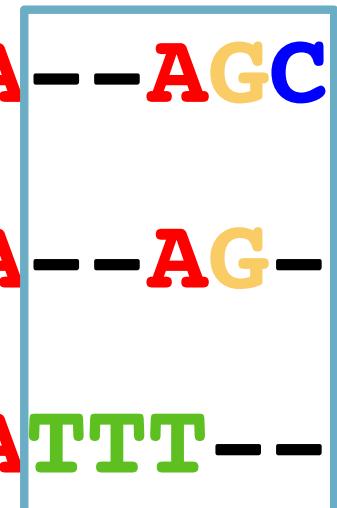


Delete sites with uncertain homology

CTATGTGGCACCCAGGCCATGCA -- AGC

ATATGTGGCA ----- CCCAGGCAG -- AG -

ATATGTGGCACCCAGGCCATGCA TTT --



# Useful references

---

