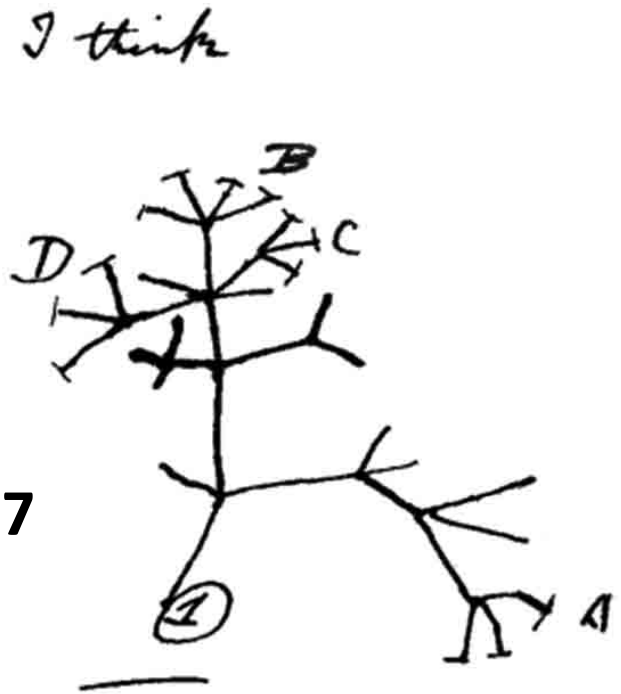
Lecture 1.1

Introduction to Molecular Phylogenetics

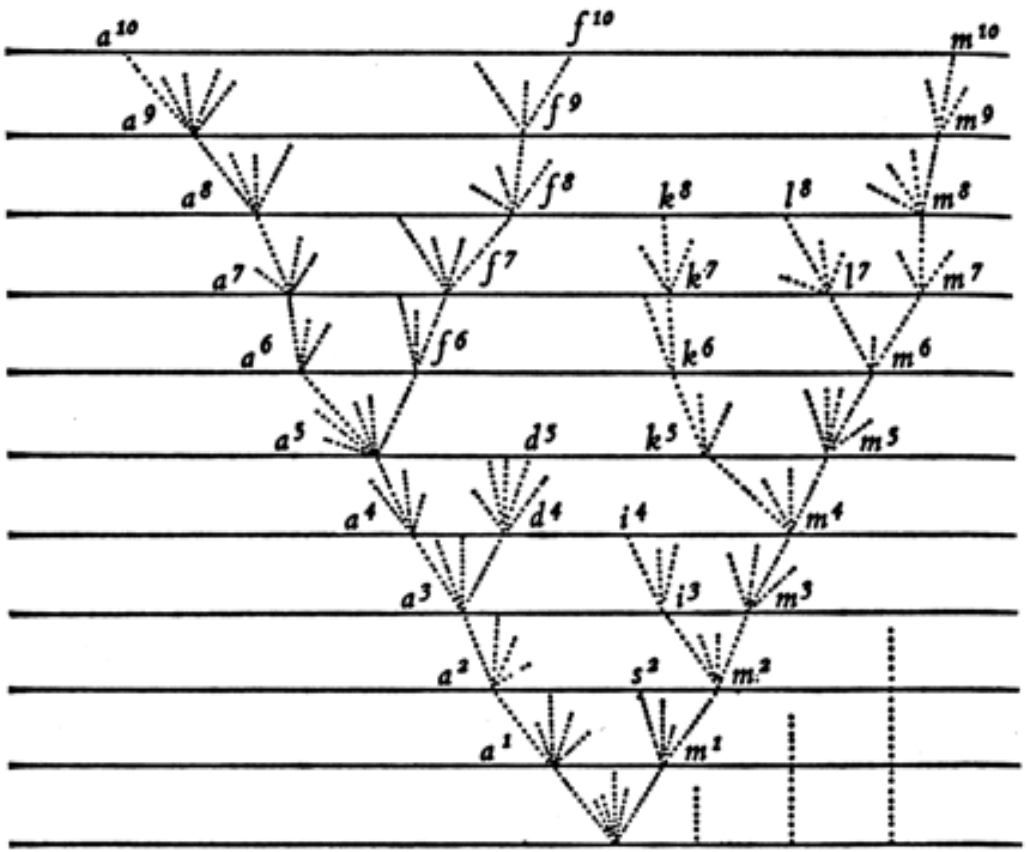
Phylogenetic Trees

What is a phylogenetic tree?

Phylogeny
evolutionary relationships among a set of organisms



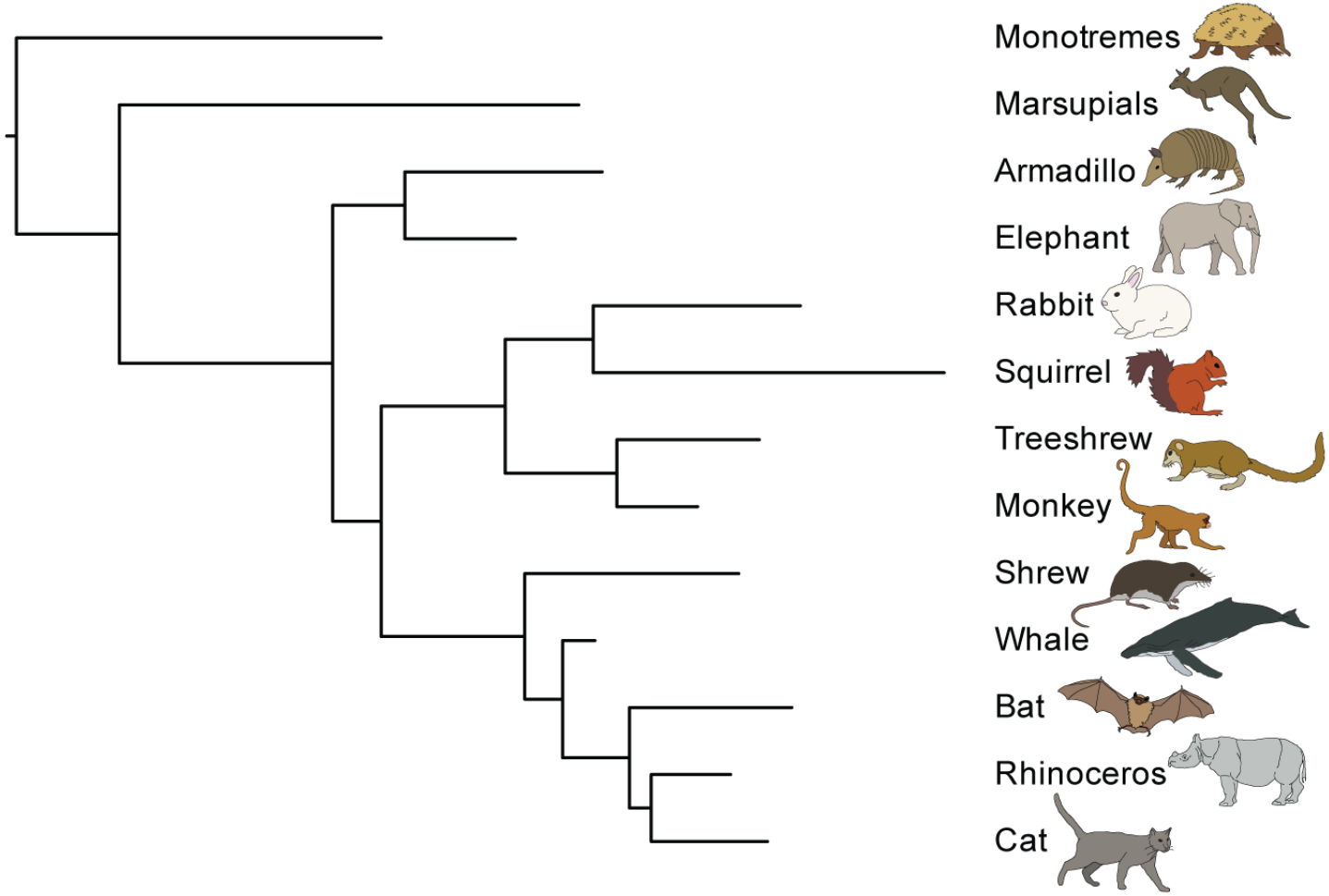
1837

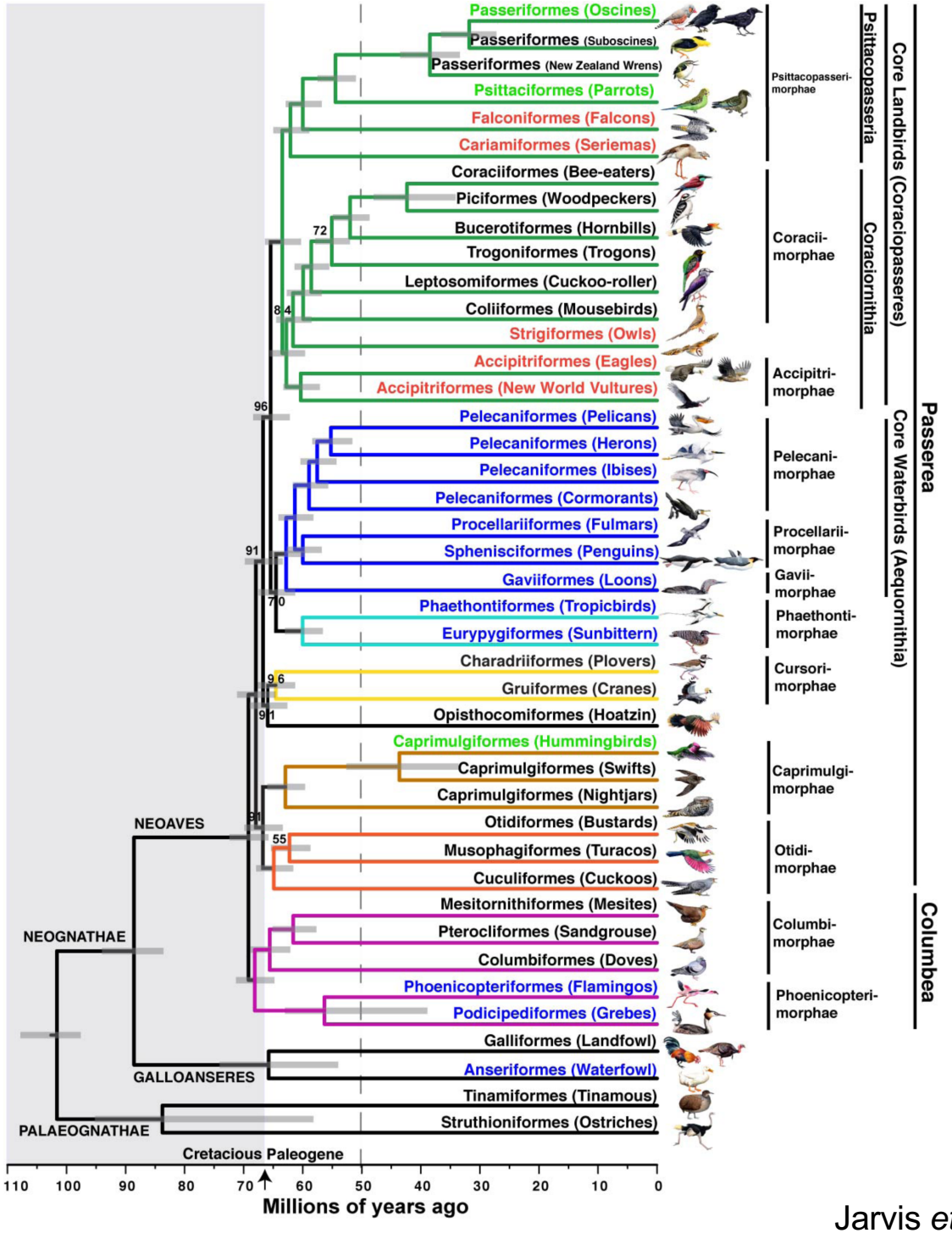


1859

Phylogenetic trees

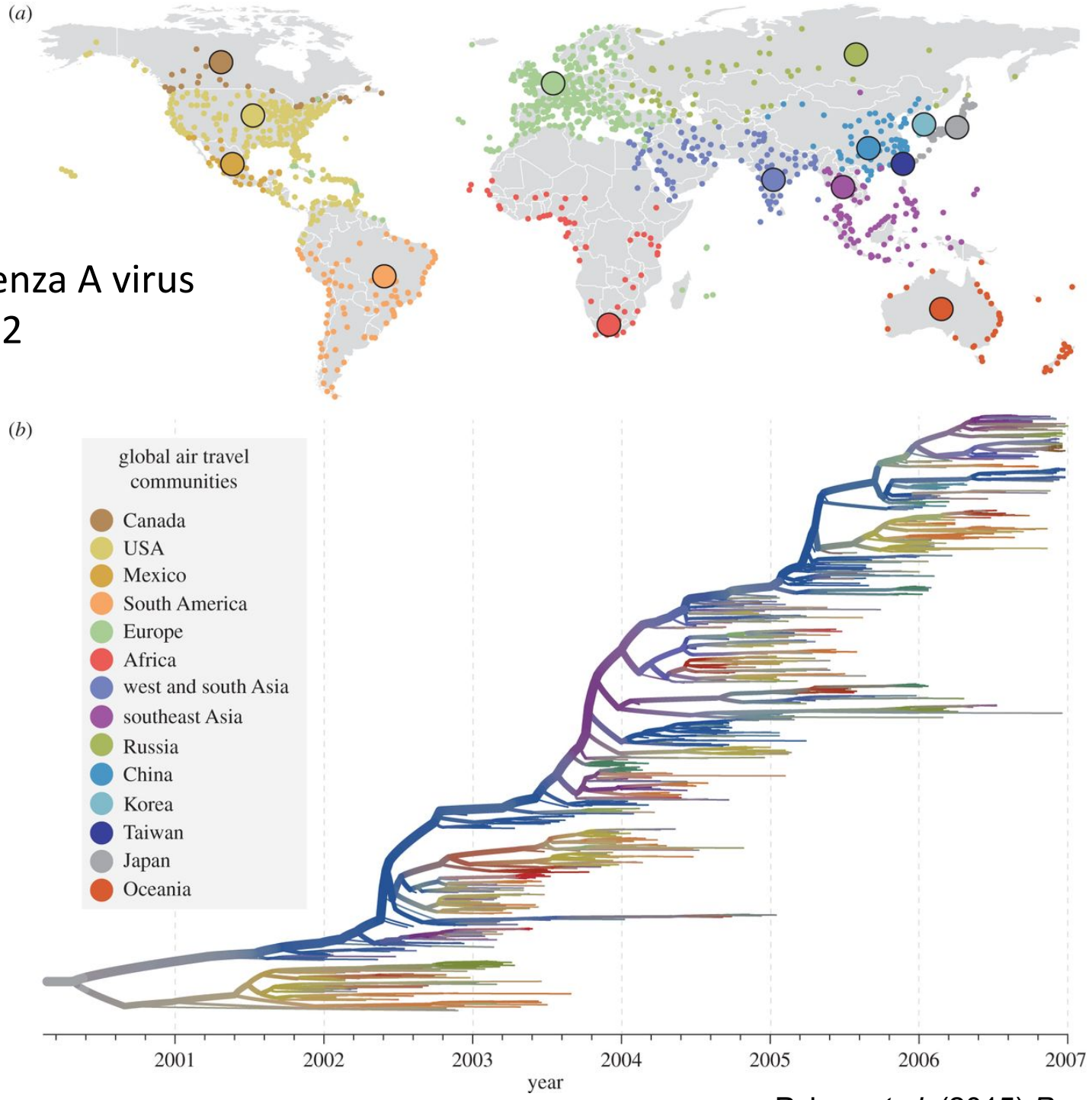
- Topology (relationships)
- Branch lengths (amount of evolutionary change or time)



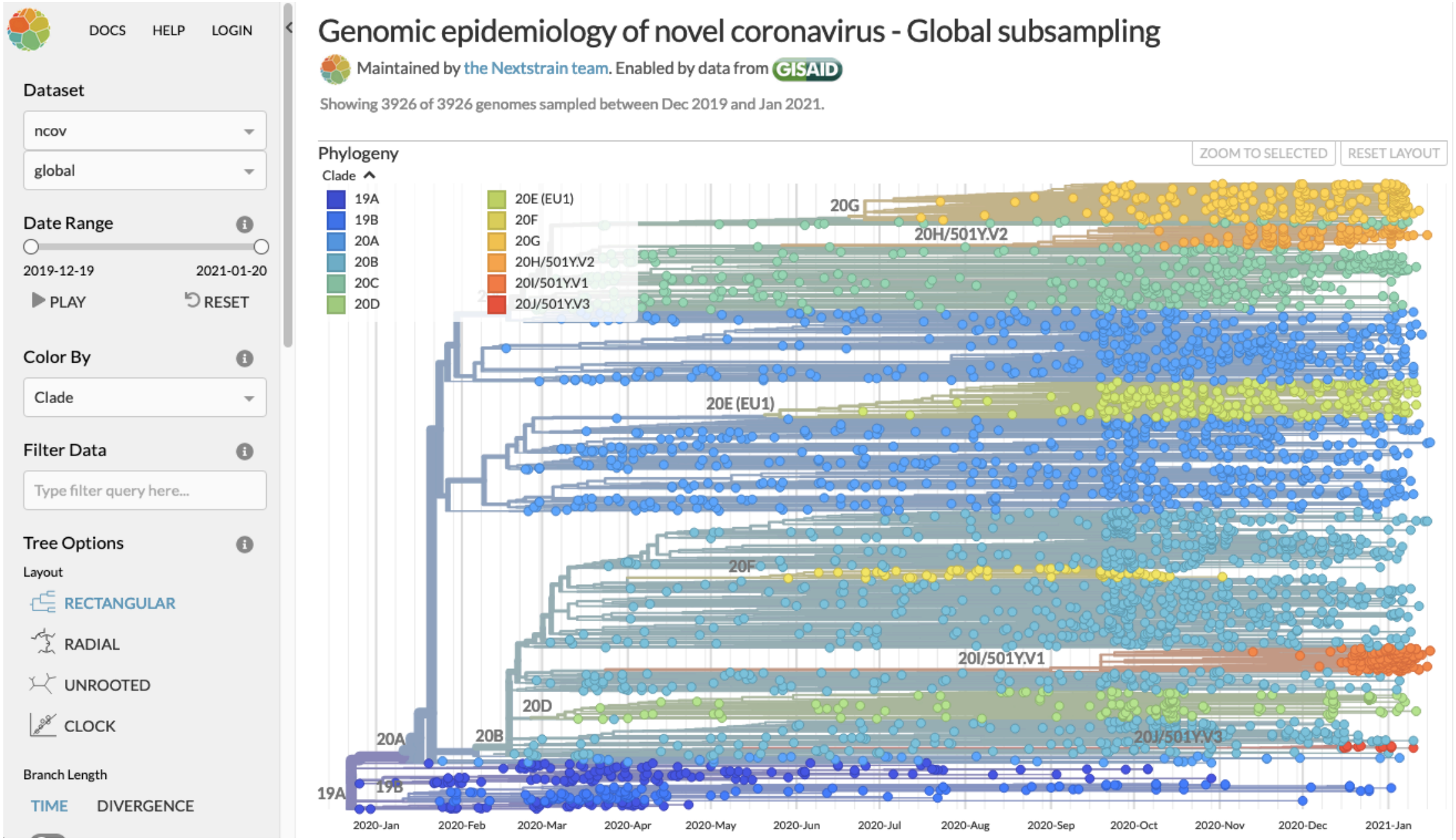


Jarvis *et al*. (2014) *Science*

Human influenza A virus
Subtype H3N2

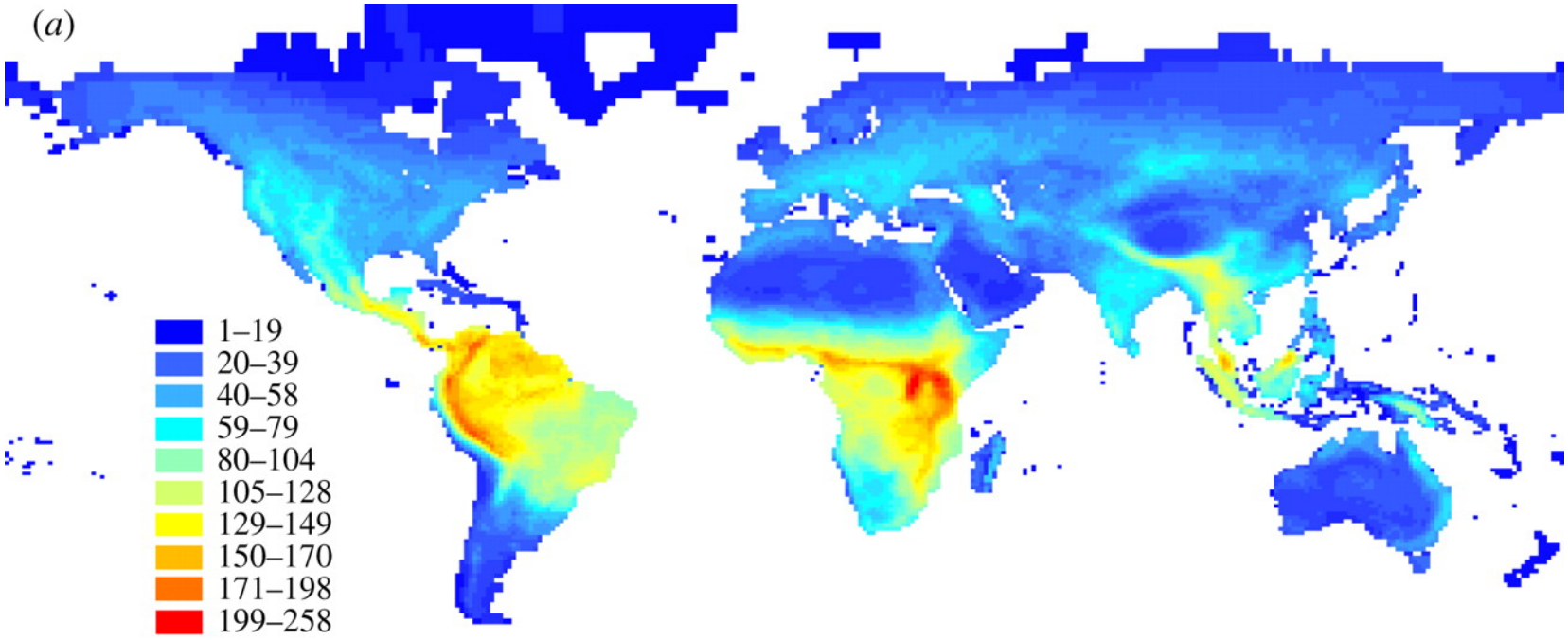


Pybus *et al*. (2015) *Proc Roy Soc B*

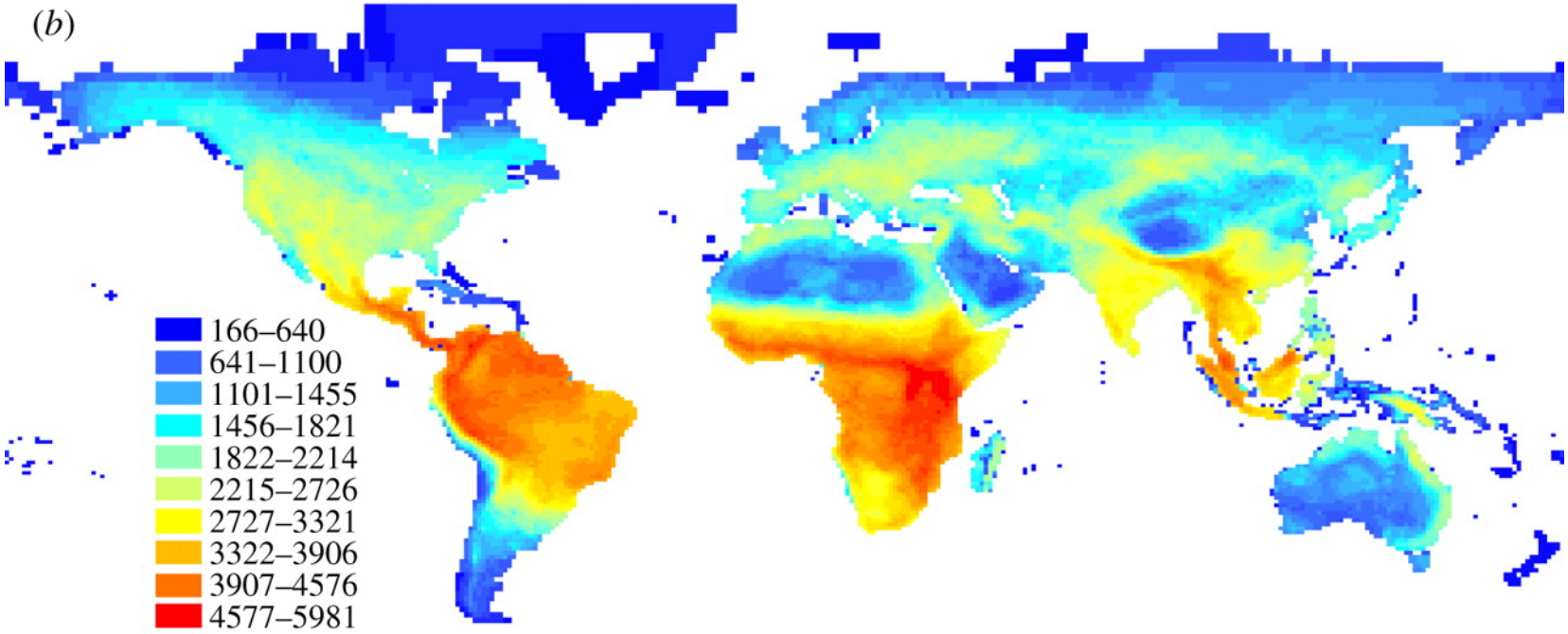


nextstrain.org, January 2021

Mammal species richness



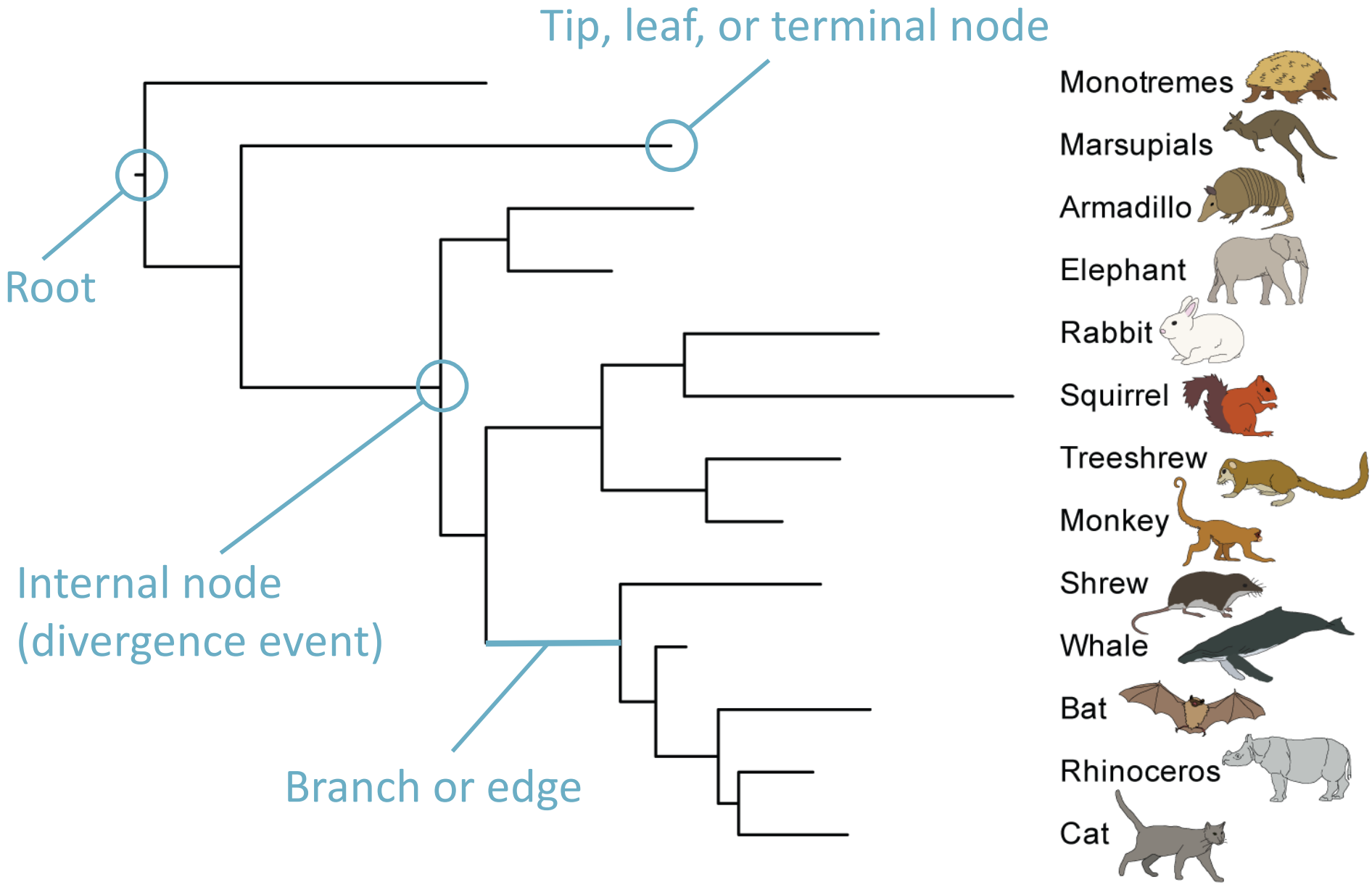
Mammal phylogenetic diversity



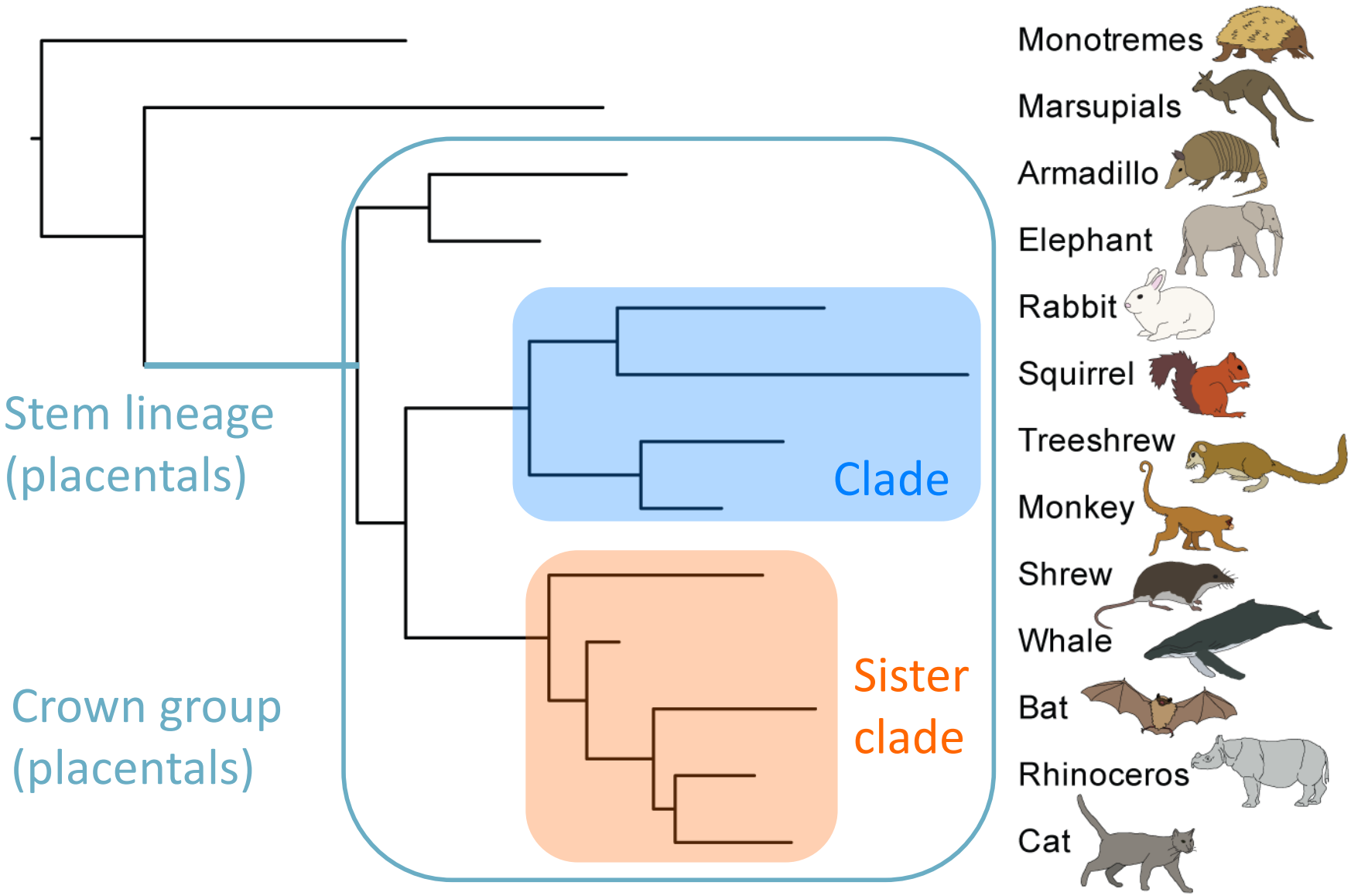
Davies & Buckley (2011) *Phil Trans Roy Soc B*

Tree Thinking

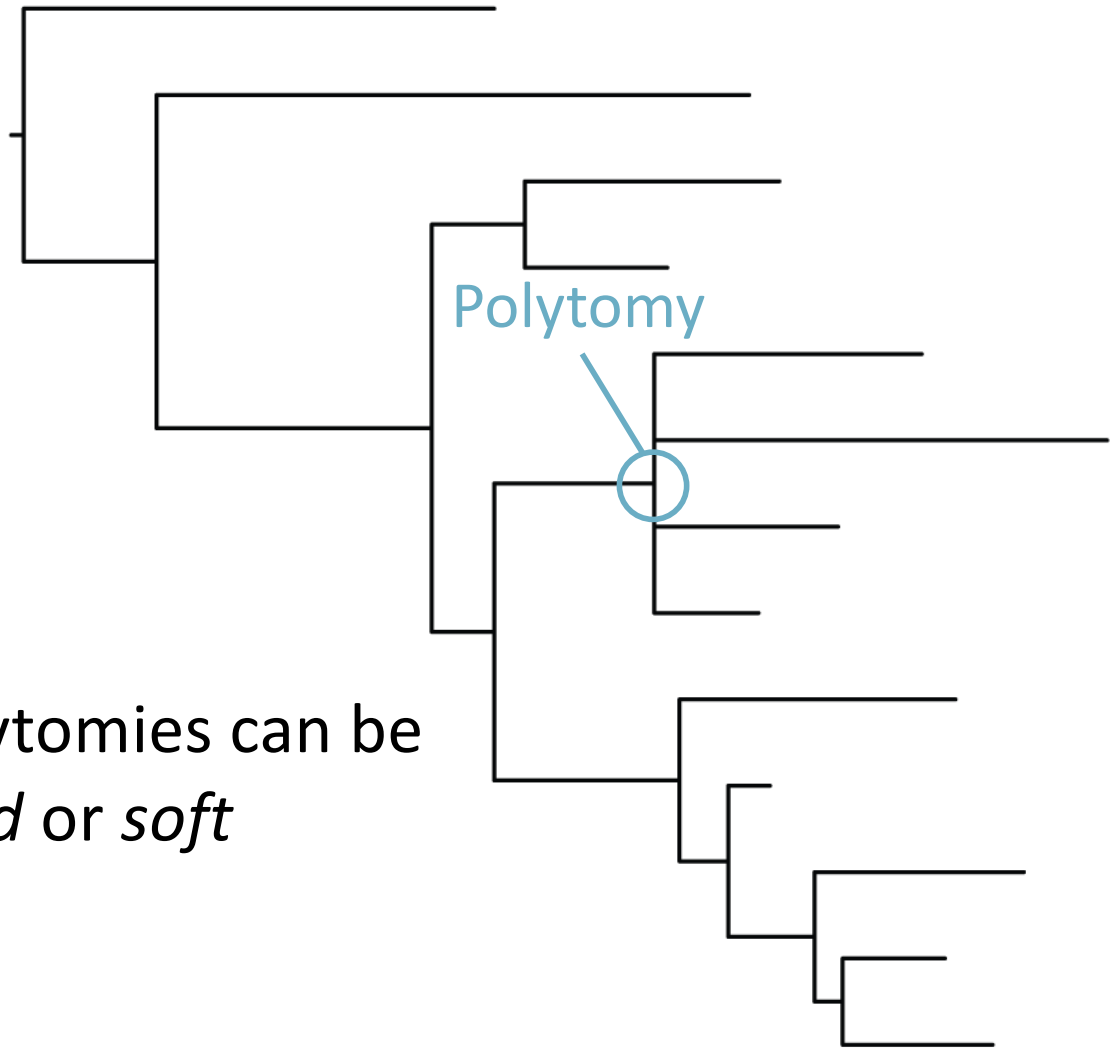
Phylogenetic trees



Phylogenetic trees

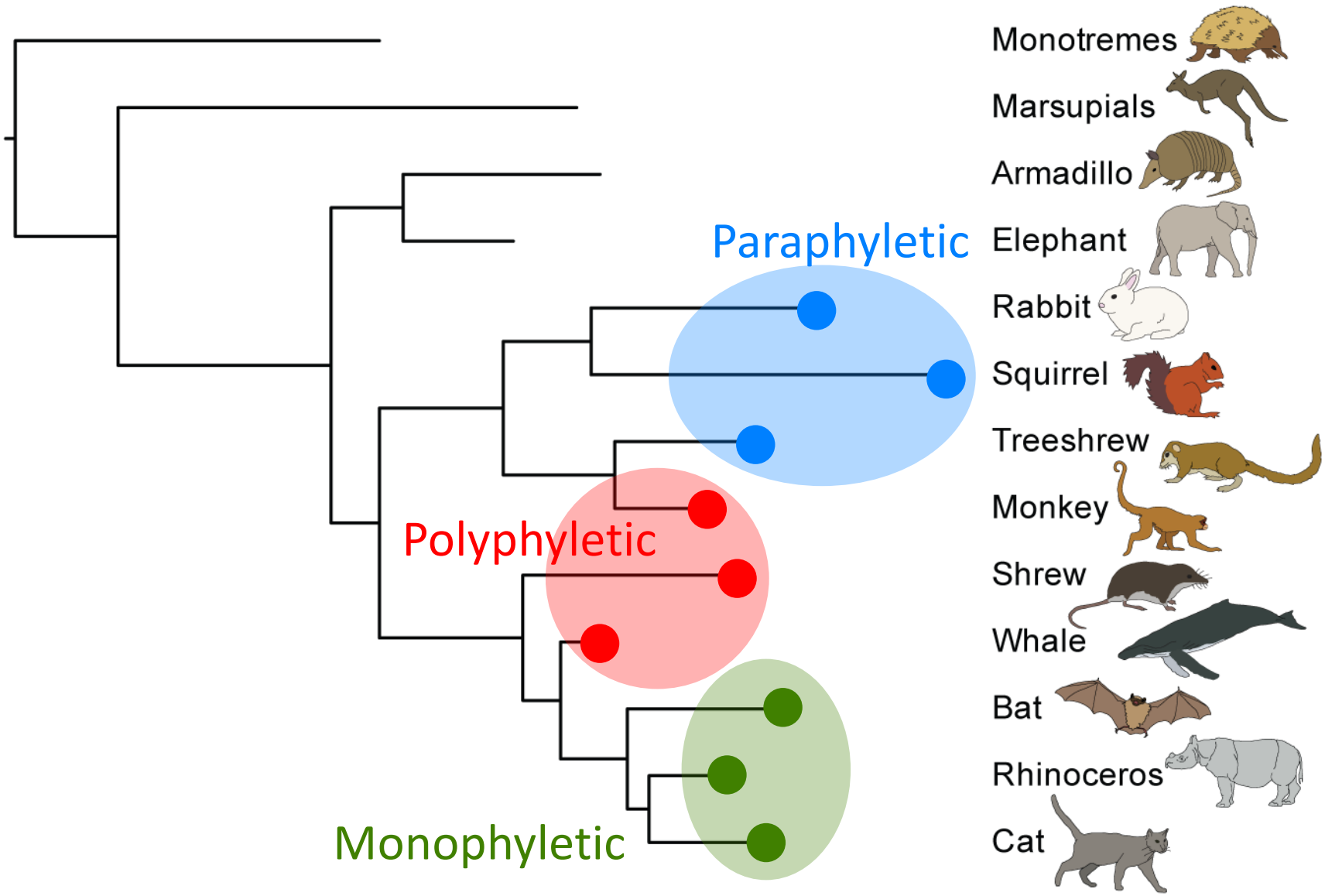


Phylogenetic trees

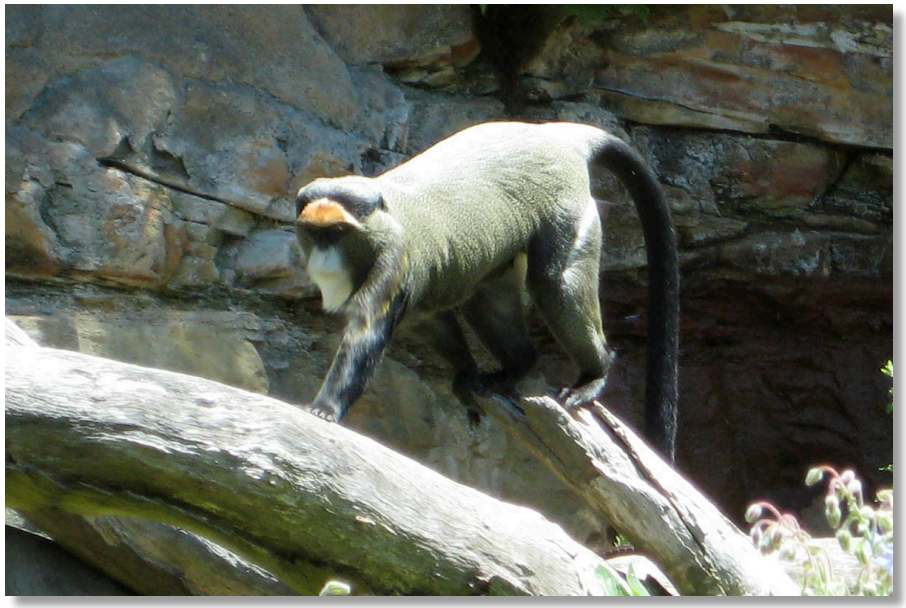


Polytomies can be
hard or *soft*

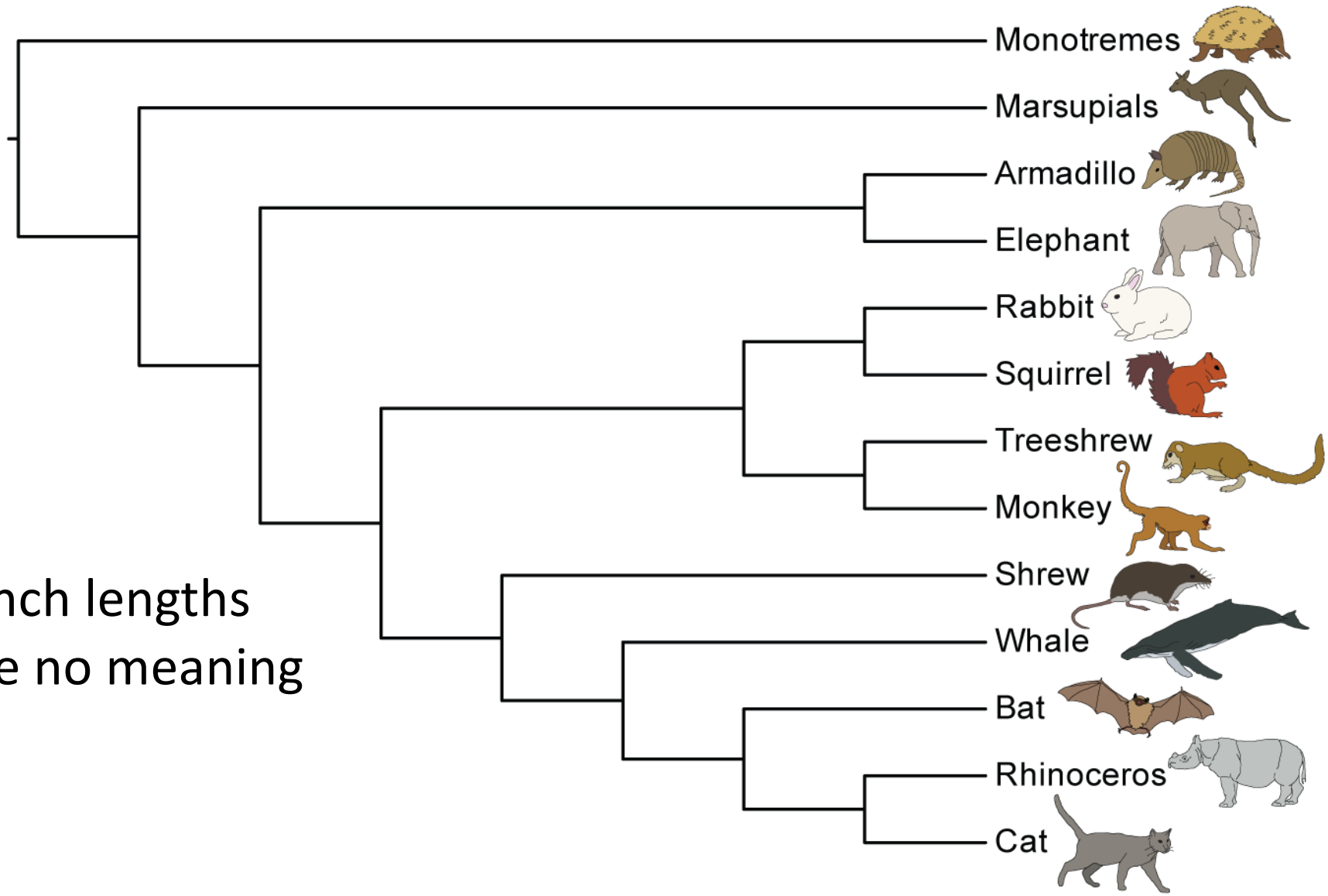
Cladistic terms



Paraphyletic groups

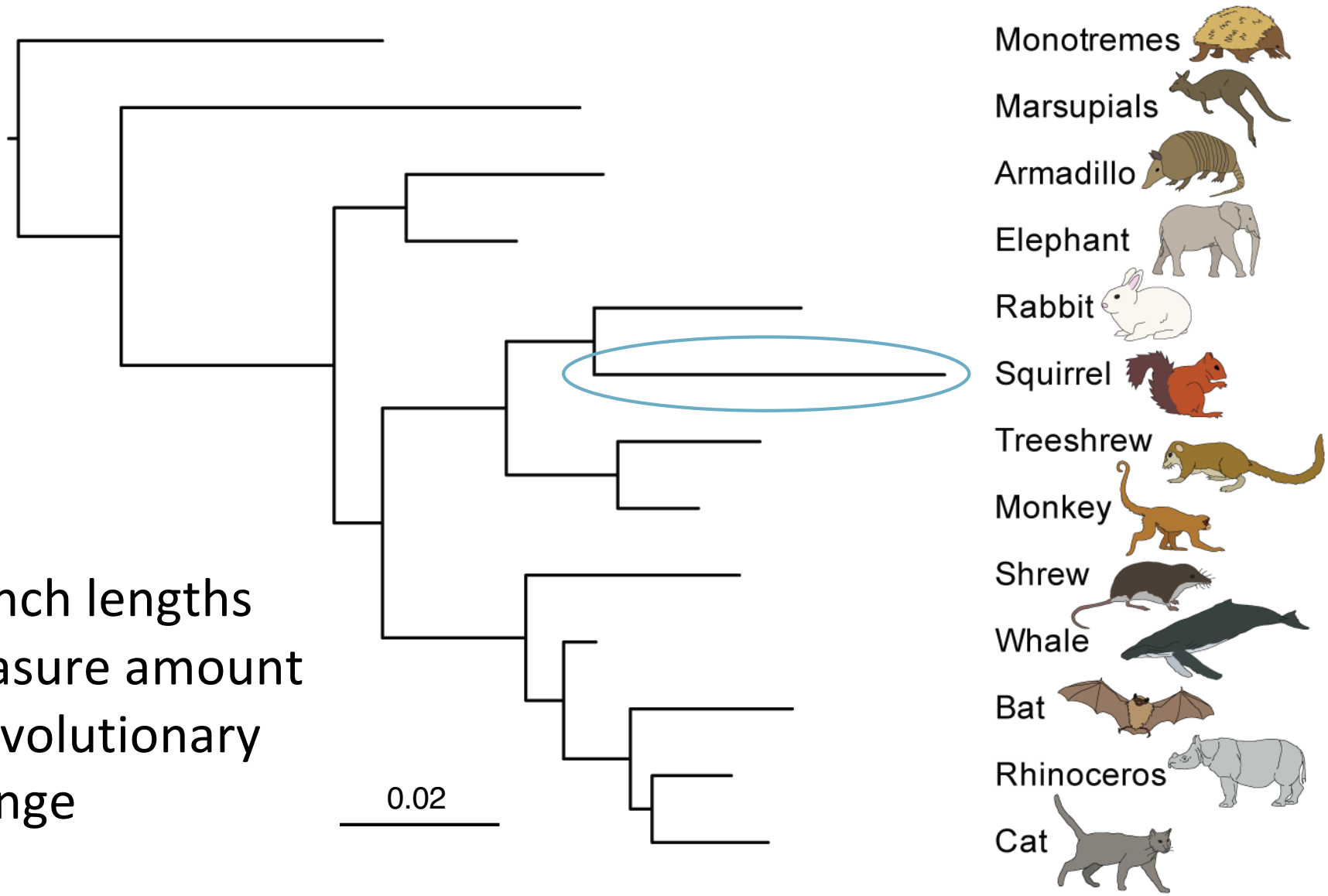


Trees: Cladogram



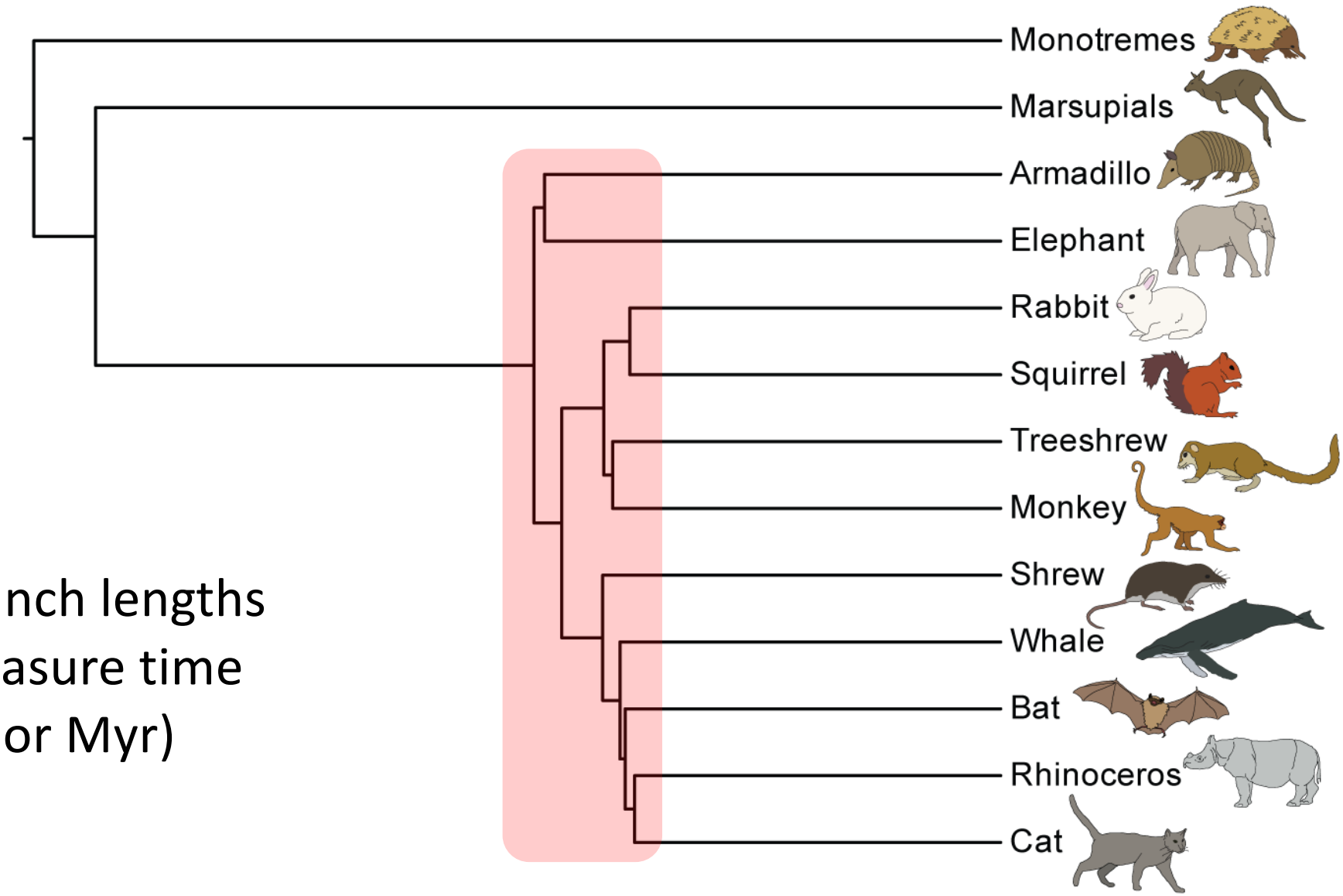
Branch lengths have no meaning

Trees: Phylogram



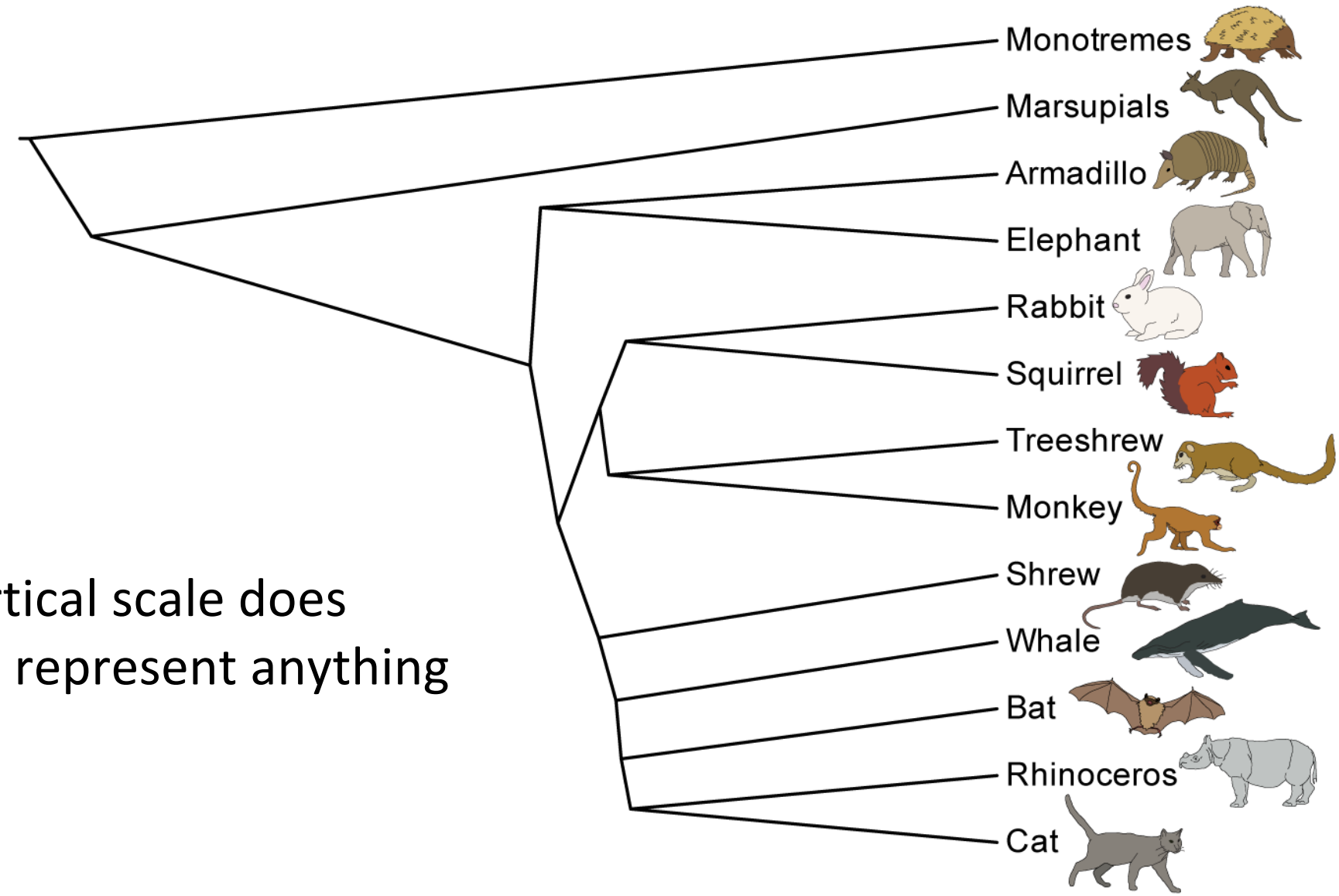
Branch lengths measure amount of evolutionary change

Trees: Chronogram or time-tree



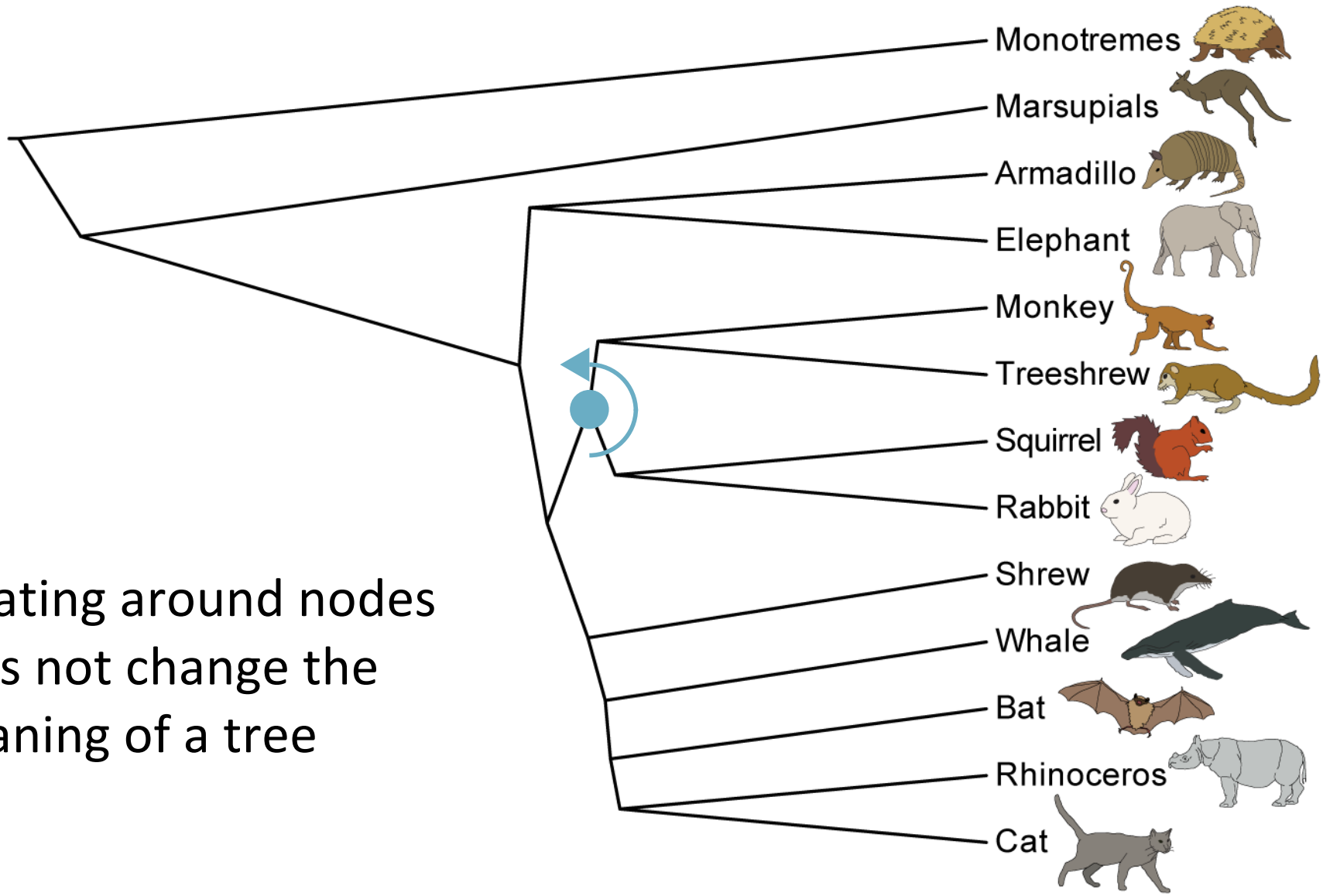
Branch lengths measure time (yr or Myr)

Phylogenetic trees



Vertical scale does not represent anything

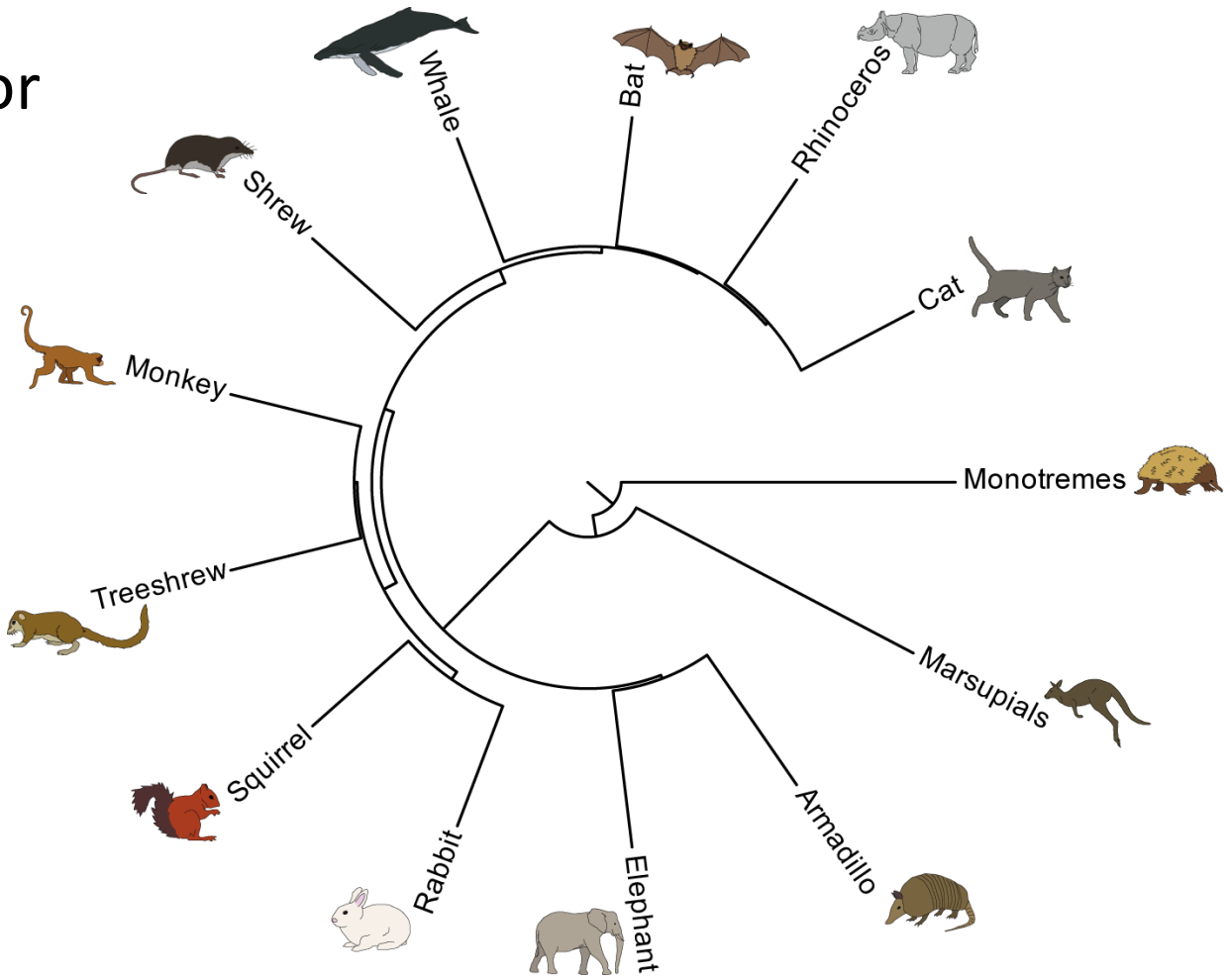
Phylogenetic trees



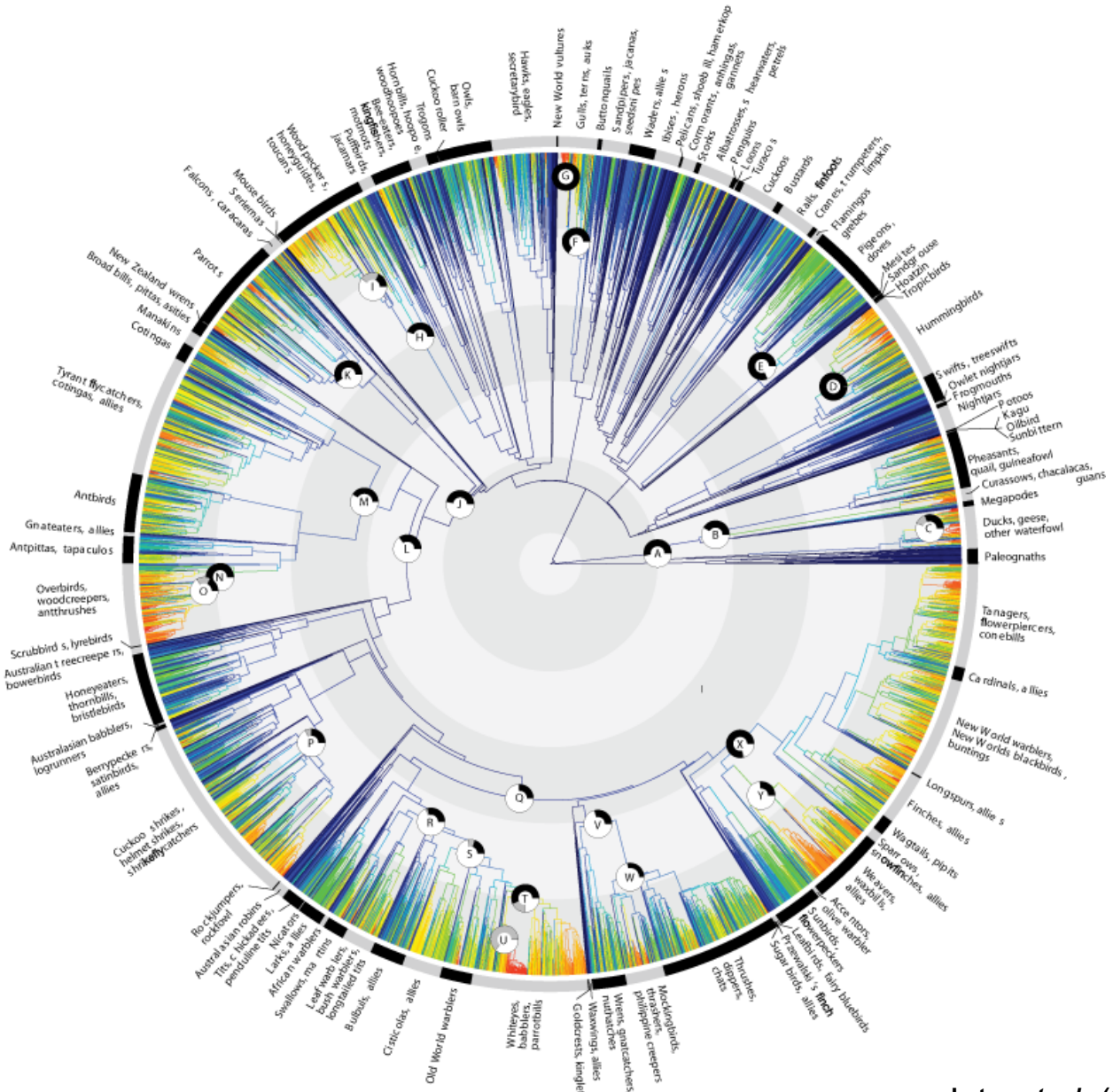
Rotating around nodes does not change the meaning of a tree

Phylogenetic trees: Circular

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

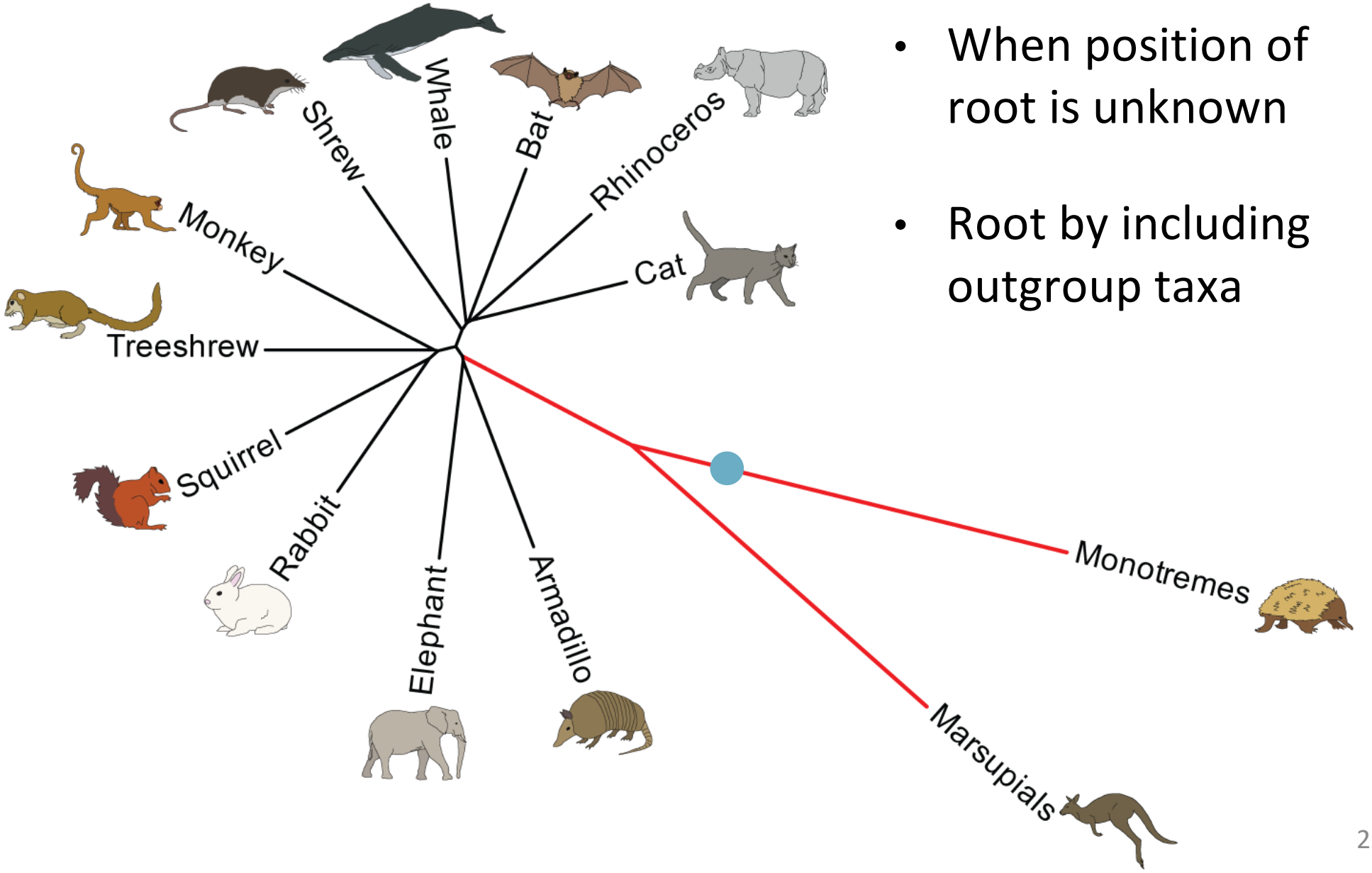


Phylogenetic trees: Circular



Jetz *et al*. (2012) *Nature*

Phylogenetic trees: Unrooted



- When position of root is unknown
- Root by including outgroup taxa

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**
  - Automatically estimates position of root

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):1.0);

Molecular Phylogenetics

Phylogenetic analysis

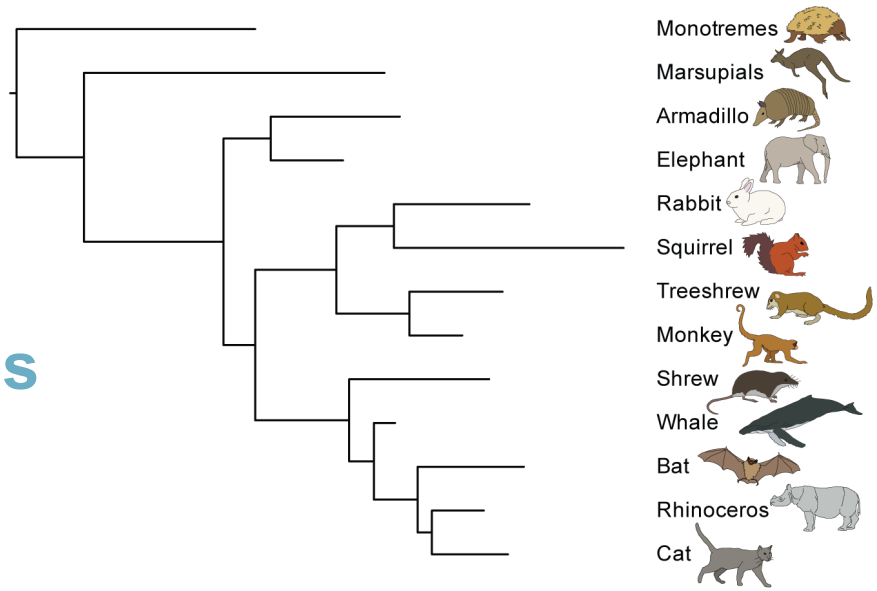
- Sometimes we know the phylogeny
  - Viral transmission histories
  - Pedigrees (humans, domesticated animals, lab organisms, *etc*.)

- Usually we do not know the phylogeny but we can estimate it
  - Morphological data
  - Molecular data

Fundamental assumptions

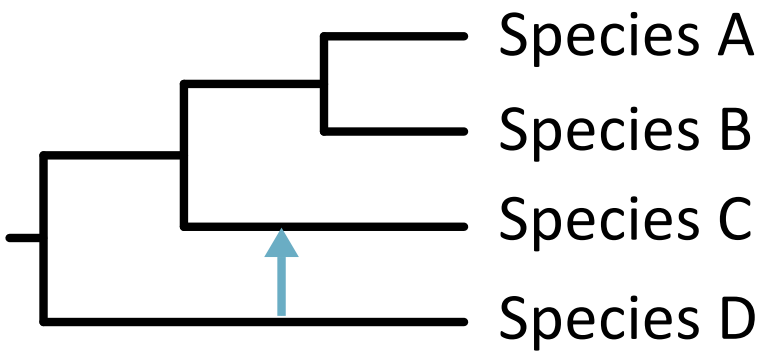
- Phylogenetic methods make several fundamental assumptions:
  - Relationships among taxa can be represented by a tree
  - Homologous characters are being compared
  - Characters are mutually independent

When might relationships not be treelike?

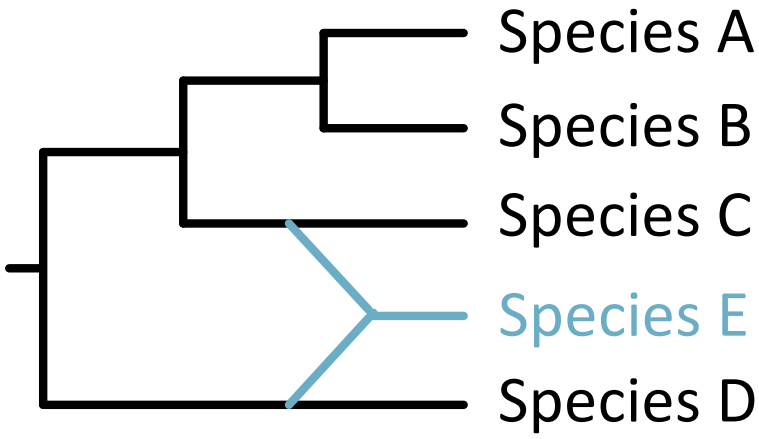


Non-treelike evolution

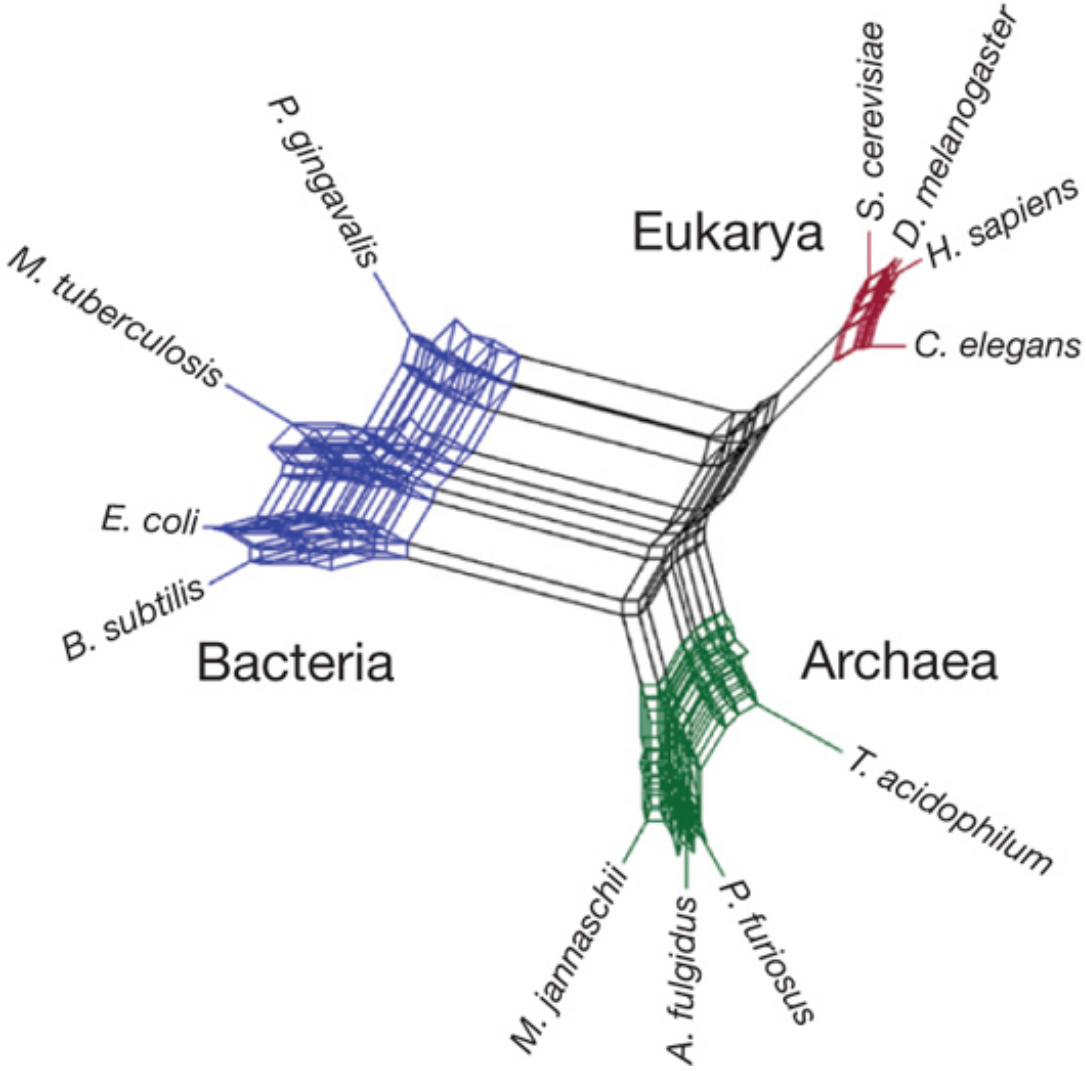
Horizontal gene transfer

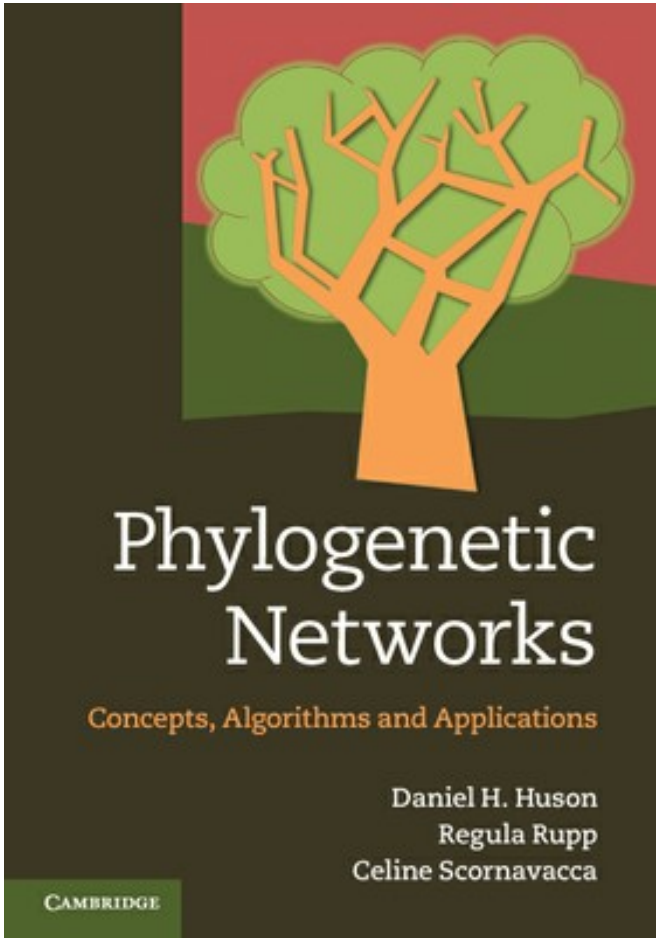


Hybrid speciation



Phylogenetic networks

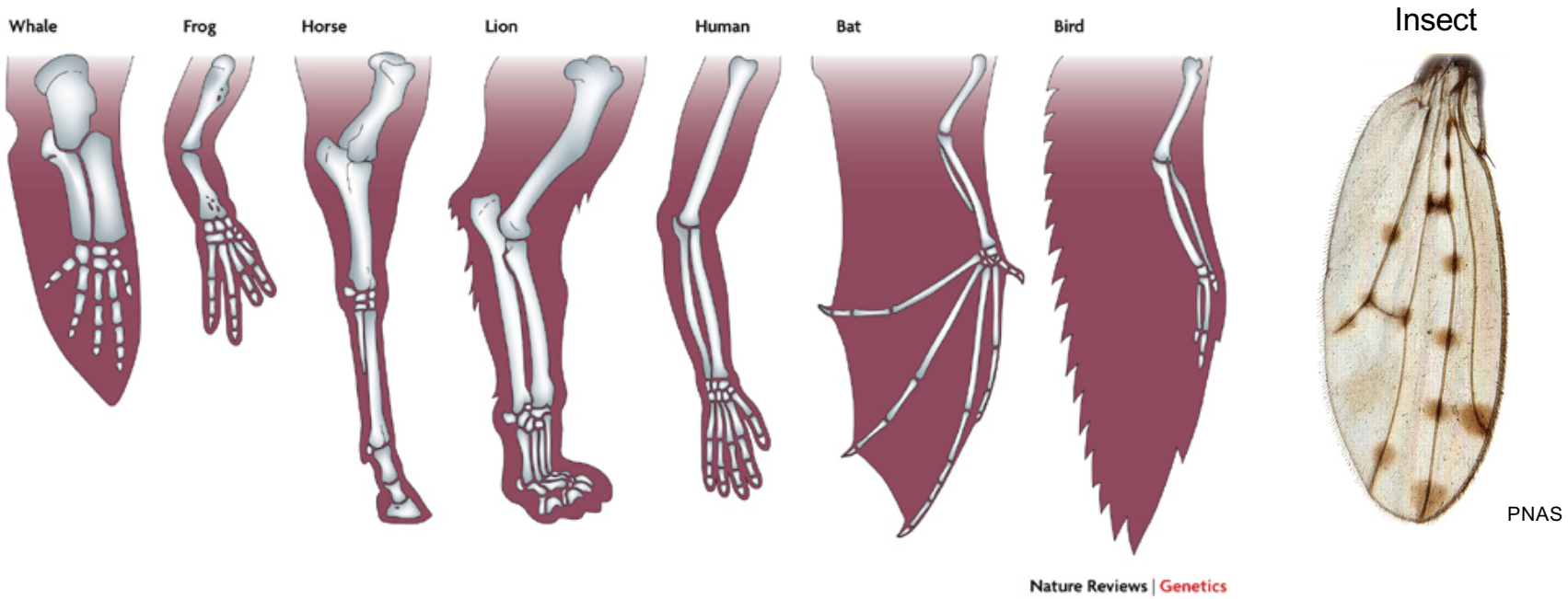




Theobald (2010) *Nature*

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

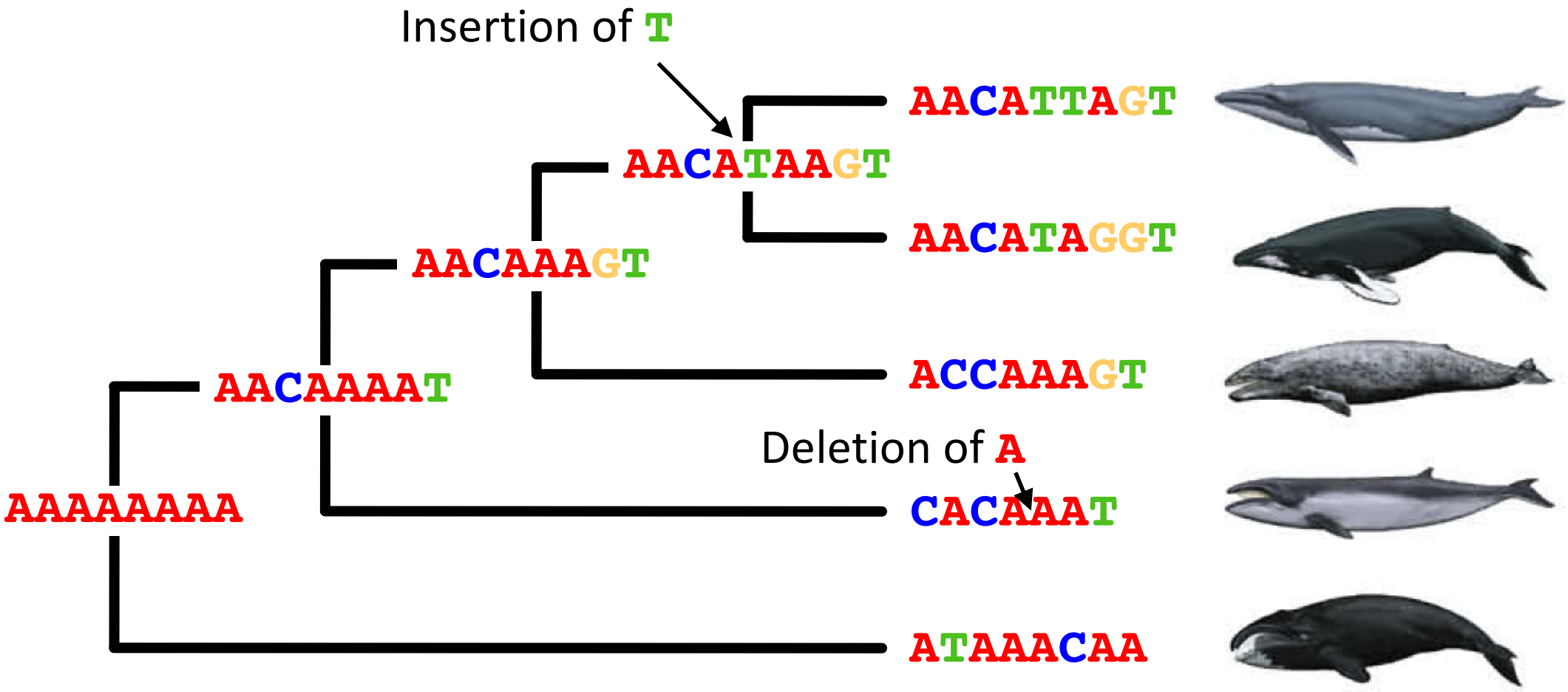


Character homology

- Comparing strings of nucleotides
- Each nucleotide site is a character
- But DNA sequences can vary in length

blue whale	CGTTAGTACACT
humpback whale	CGATAGTTCACT
gray whale	CGTTAGTTTACC
fin whale	CATTGGATTACT
right whale	CATTGGTTTACT

Example: Whales



DNA sequence alignment

AACATTAGT

AACATAGGT

ACCAAAGT

CACAAAT

ATAAACAA

AACATTAGT

AACATAGGT

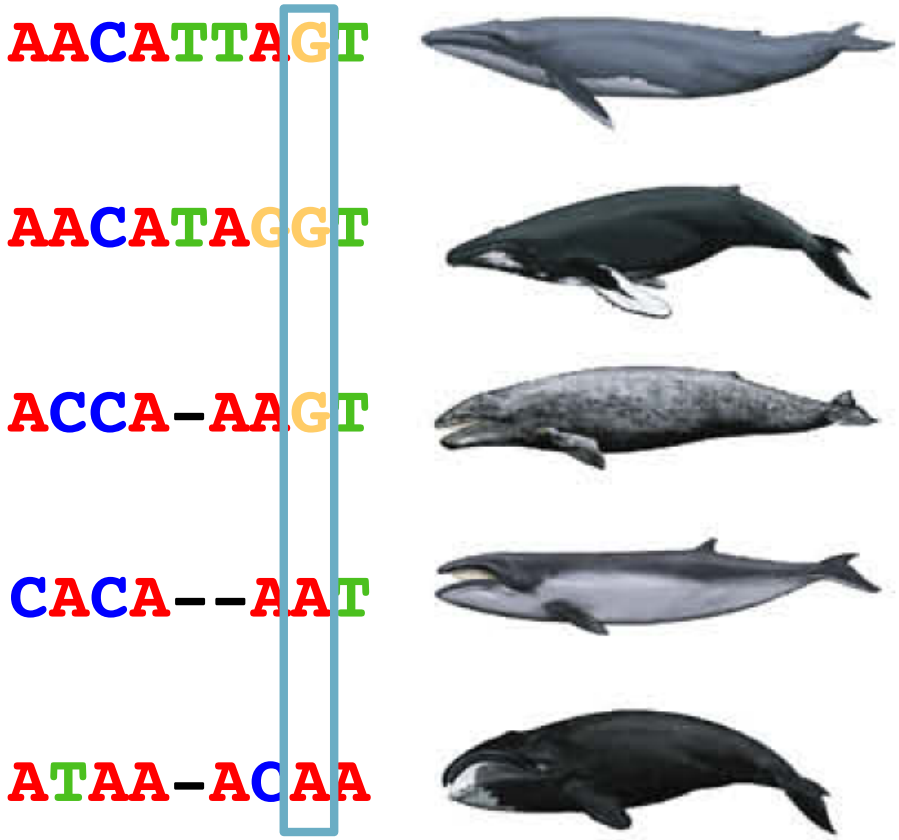
ACCA-AAGT

CACA--AAT

ATAA-ACAA

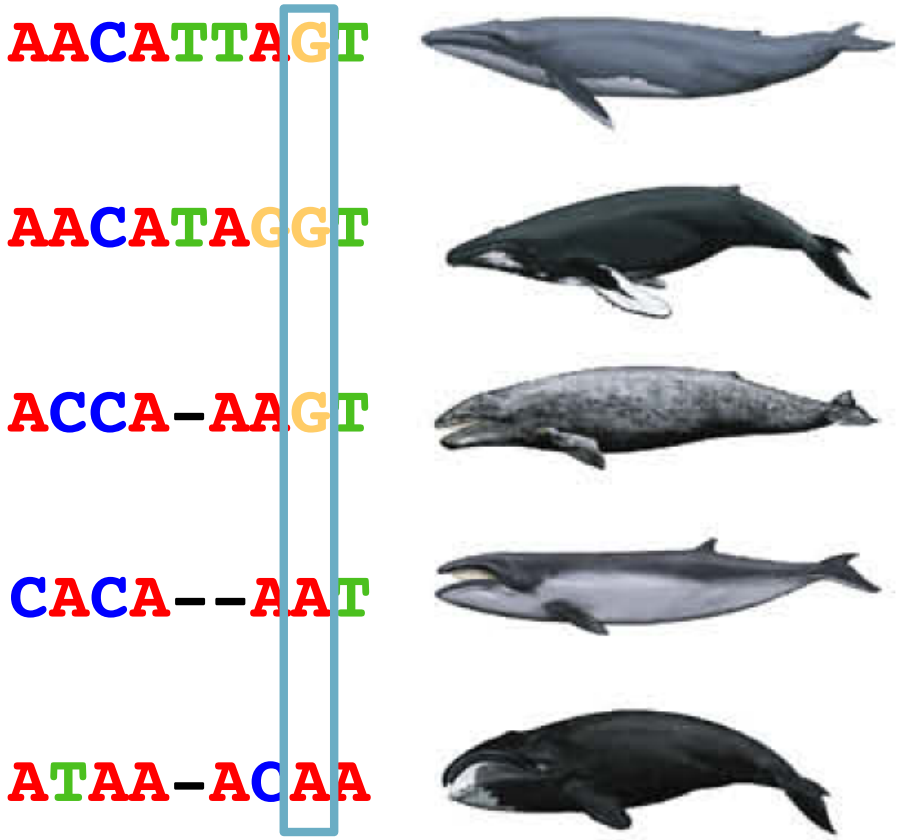
DNA sequence alignment

- 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



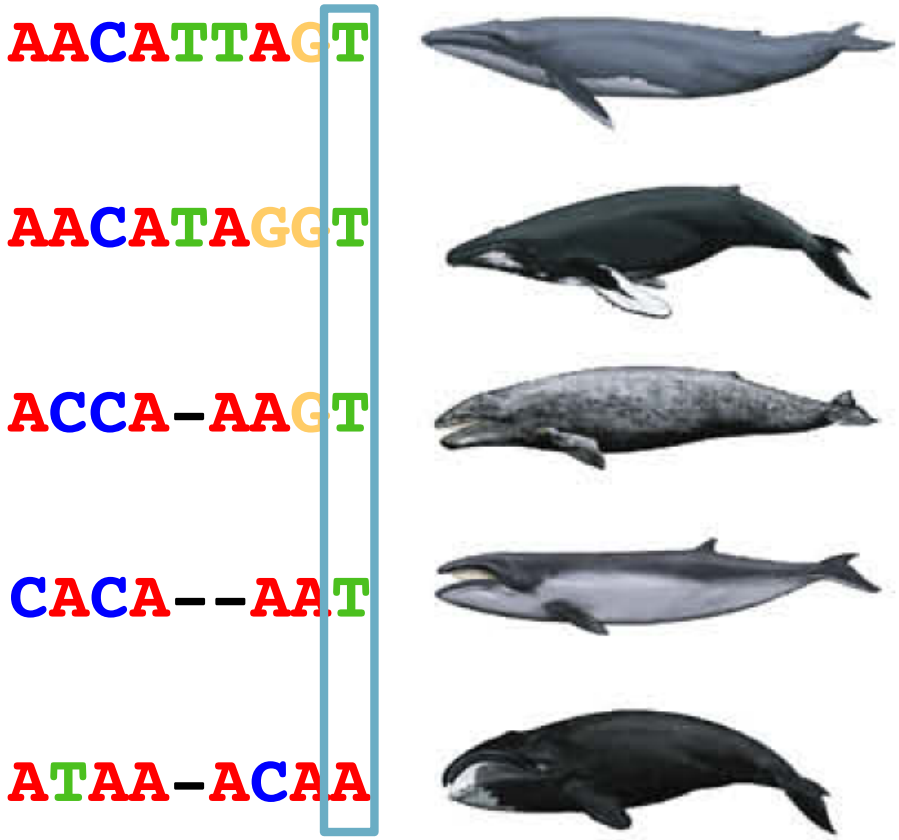
DNA sequence alignment

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



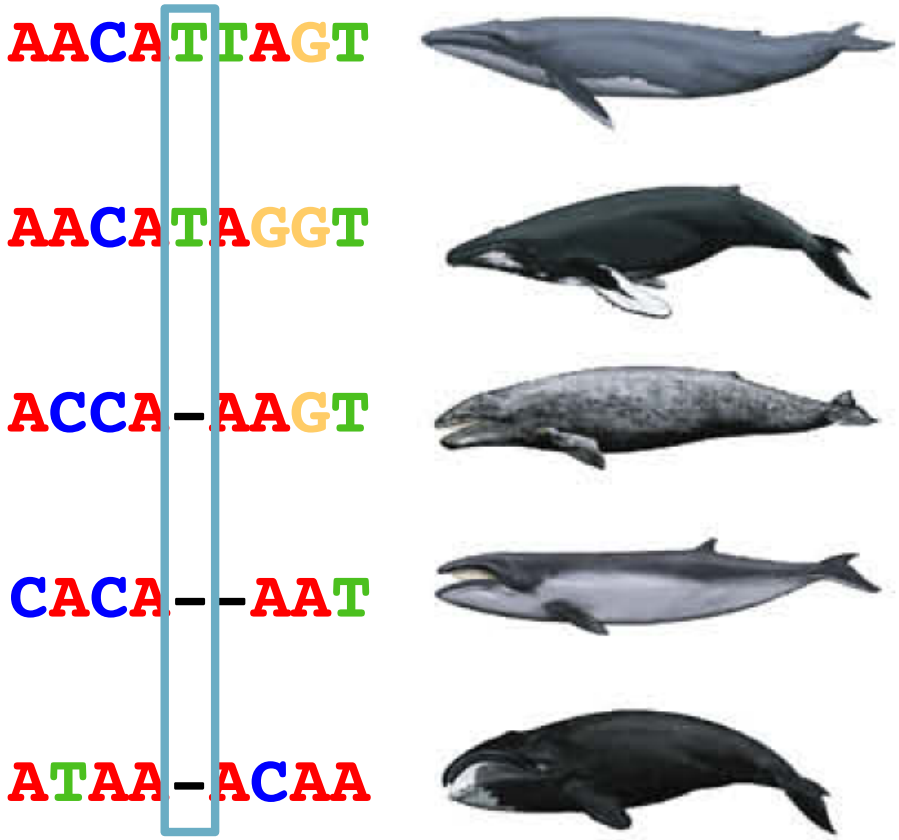
DNA sequence alignment

- 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 really 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+ and Toby J.Gibson*

Software

Open Access

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[1], Kei-ichi Kuma and Takashi Miyata*

A practical approach

Align sequences using automated methods

↓

Adjust alignments by eye

```
CTATGTGGCACCCAGCCCATGCA--AGC
ATATGTGGCA-----CCCAGGCA--AG-
ATATGTGGCACCCAGCCCATGCATTT--
```

A practical approach

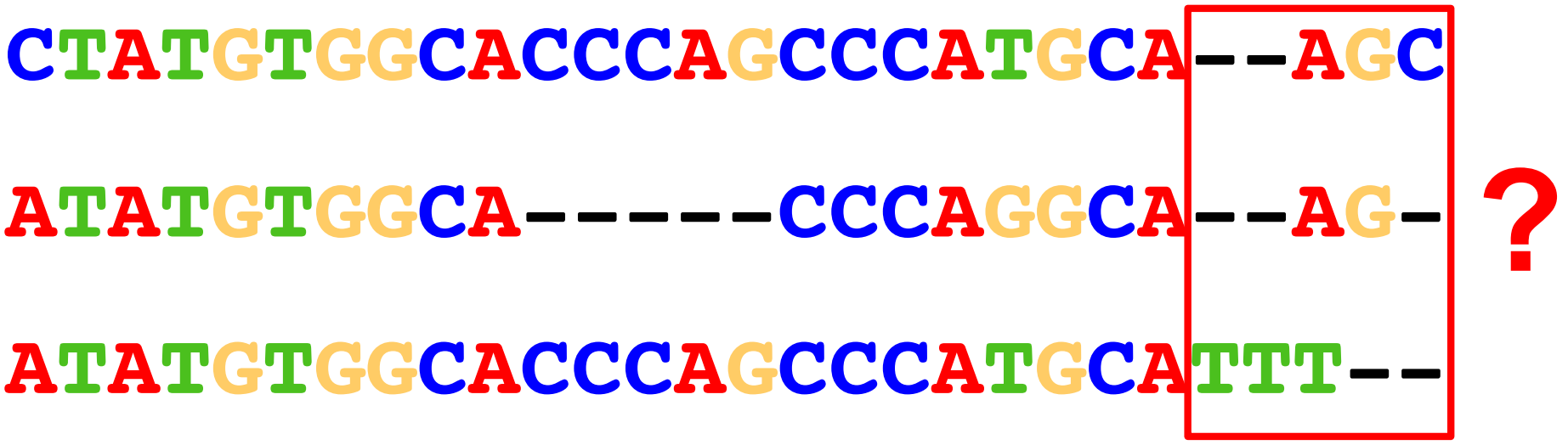
Align sequences using automated methods

↓

Adjust alignments by eye

↓

Delete sites with uncertain homology



Useful references