THE UNIVERSITY OF MELBOURNE

COMPUTATIONAL BIOLOGY RESEARCH

A Hallmark Research Initiative

Introduction to Pathogen Phylogenetic analyses

November 20 - 22
University of Melbourne

Contributors

21 workshops, 5 locations

Several contributors over the years

University of Sydney
Simon Ho
Nathan Lo
Luana Lins
Jun Tong
Charles Foster
David Duchene
Matt Brandley
Martyna Molak

Australian National University
Robert Lanfear

Queensland University of Technology
Matt Phillips

Sebastian Duchene
McKenzie Postdoctoral Fellow
Dept of Biochemistry and Molecular Biology

Jane Hawkey
Postdoctoral Fellow
Dept of Biochemistry and Molecular Biology

Andrew Siebel
Academic Convenor, CBRI
Melbourne Integrative Genomics

Remco Bouckaert
University of Auckland and
Max Planck Institute for the Science of Human History

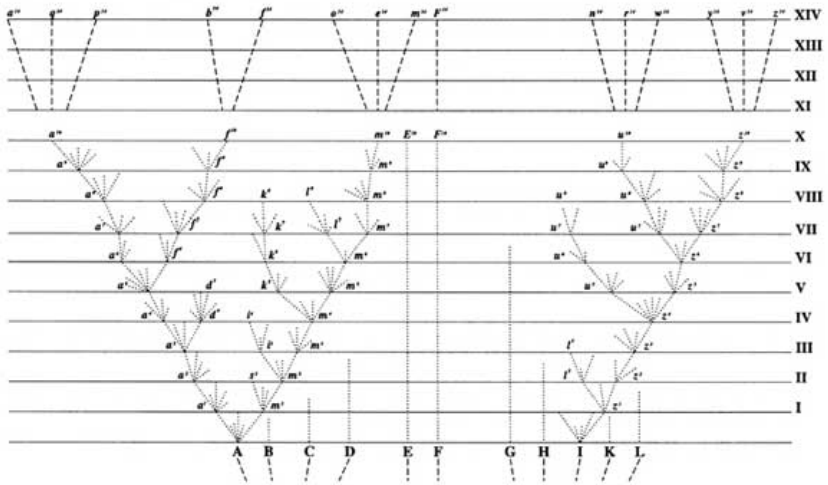
Bayesian phylogenetic methods, Language evolution, BEAST2

John-Sebastian Eden
University of Sydney

Infectious diseases, virus evolution

Workshop overview

- Introduction to phylogenetic analysis
  - Interpreting phylogenetic trees
  - Phylogenetic methods
  - Estimating evolutionary rates and timescales
  - Bayesian phylogenetic analysis
  - Phylodynamics



MEGA

- *Molecular Evolutionary Genetics Analysis*
- Koichiro Tamura and Sudhir Kumar
- Population genetics and phylogenetics
  - Maximum parsimony
  - Distance-based methods
  - Maximum likelihood

PhyML

- Maximum likelihood phylogenetic tree inference
- Guindon S, Dufayard, JF, Lefort V, Anisimova M, Hordijk W, Gascuel, O.
- Command line interface
- Rapid branch support

TempEst

- Visualisation tool to assessing temporal structure
- Rambaut A, Lam T, de Carvalho L, Pybus O.
- Graphical interface

BEAST

- *Bayesian Evolutionary Analysis by Sampling Trees*
- Beast 1:Alexei Drummond & Andrew Rambaut
- **Beast 2:** Bouckaert et al.
- Bayesian phylogenetic analysis
  - Implements many evolutionary models – flexible but complex

BEAST 2 (and BEAST 1.6x)

BEAST 1

R

- Programming language, designed for statistics and graphics
- Ross Ihaka and Robert Gentleman
- We will only use it for a few demonstrations and programming knowledge is not necessary.

Workshop programme – Day 1

09:15 – 09:30	**Arrival**
09:45 – 10:30	**Lecture 1.1**: Introduction to molecular phylogenetics
10:30 – 11:00	**Practical 1a**: Sequence alignment and using *MEGA*
	– Coffee break –
11:30 – 12:30	**Lecture 1.2**: Substitution models
	– Lunch –
13:30 – 14:15	**Lecture 1.3**: Phylogenetic methods
14:15 – 14:45	**Practical 1b**: Model selection in *MEGA*
14:45 – 15:15	**Practical 1c**: Maximum likelihood analysis in *PhyML*
15:20 – 16:30	Public Seminar by JS Eden (Agar theatre)

Workshop programme – Day 2

09:50 – 9:30	Arrival
09:30 – 10:15	Lecture 2.1: The molecular clock
10:15 – 11:00	Practical 2a: Assessing temporal structure in *TempEst*
	– Coffee break –
11:30 – 12:30	Lecture 2.2: Bayesian phylogenetics
	– Lunch break –
13:30 – 14:30	Lecture 2.3: Priors in Bayesian phylogenetics
14:30 – 15:00	Practical 2b: Markov Chain Monte Carlo
15:00 – 15:30	Lecture 2.4: Demographic priors and model selection
15:30 – 16:30	Practical 3a: Molecular dating using *BEAST2*

Workshop programme – Day 3

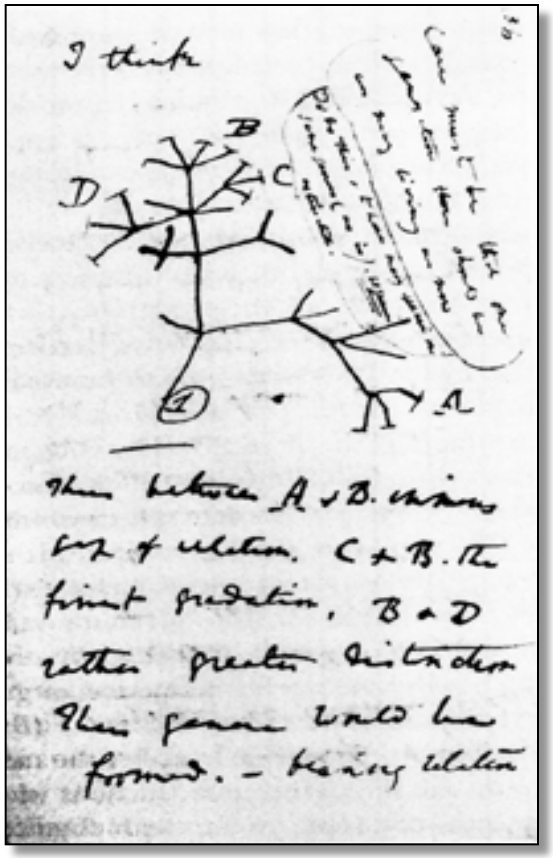
09:15 – 9:30	Arrival
09:30 – 10:00	A tour of *BEAST2* (Remco Bouckaert)
10:00 – 11:00	Lecture 3.1: Infectious disease phylodynamics
	– Coffee break –
11:30 – 12:30	Practical 3a: Phylodynamics in BEAST2 Coalescent and Birth-Death Models
	– Lunch break –
13:30 – 14:30	Practical 3b: Phylodynamics in BEAST Birth-Death SIR
14:30 – 15:30	BEAST2 Clinic (Remco Bouckaert)
15:45 – 16:45	Public seminar by Remco Bouckaert (Agar theatre)
17:30 – 19:00	Networking at Naughtons

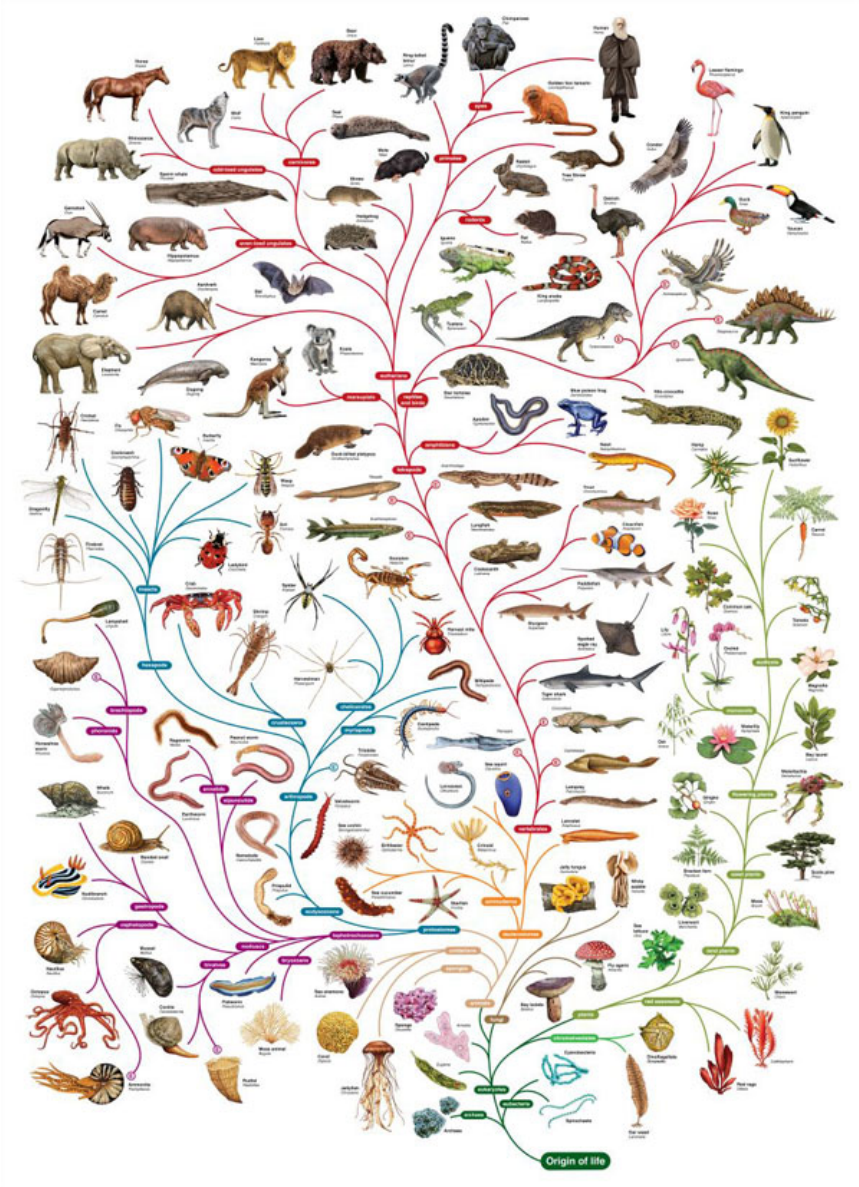
Lecture 1.1:
Introduction to Molecular Phylogenetics

Understanding Phylogenetic Trees

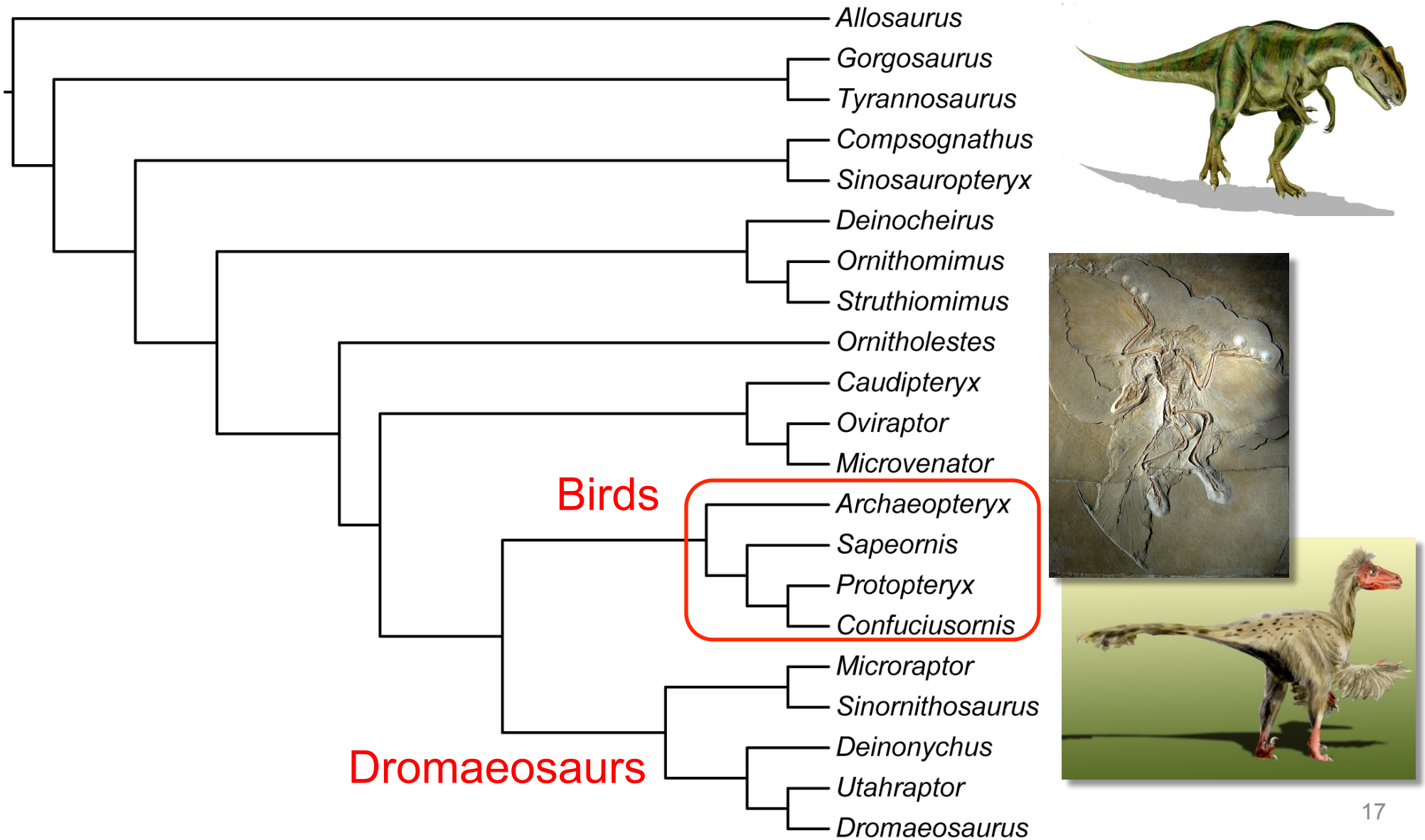
What is a phylogenetic tree?

- The phylogeny refers to the true evolutionary relationships among a set of organisms

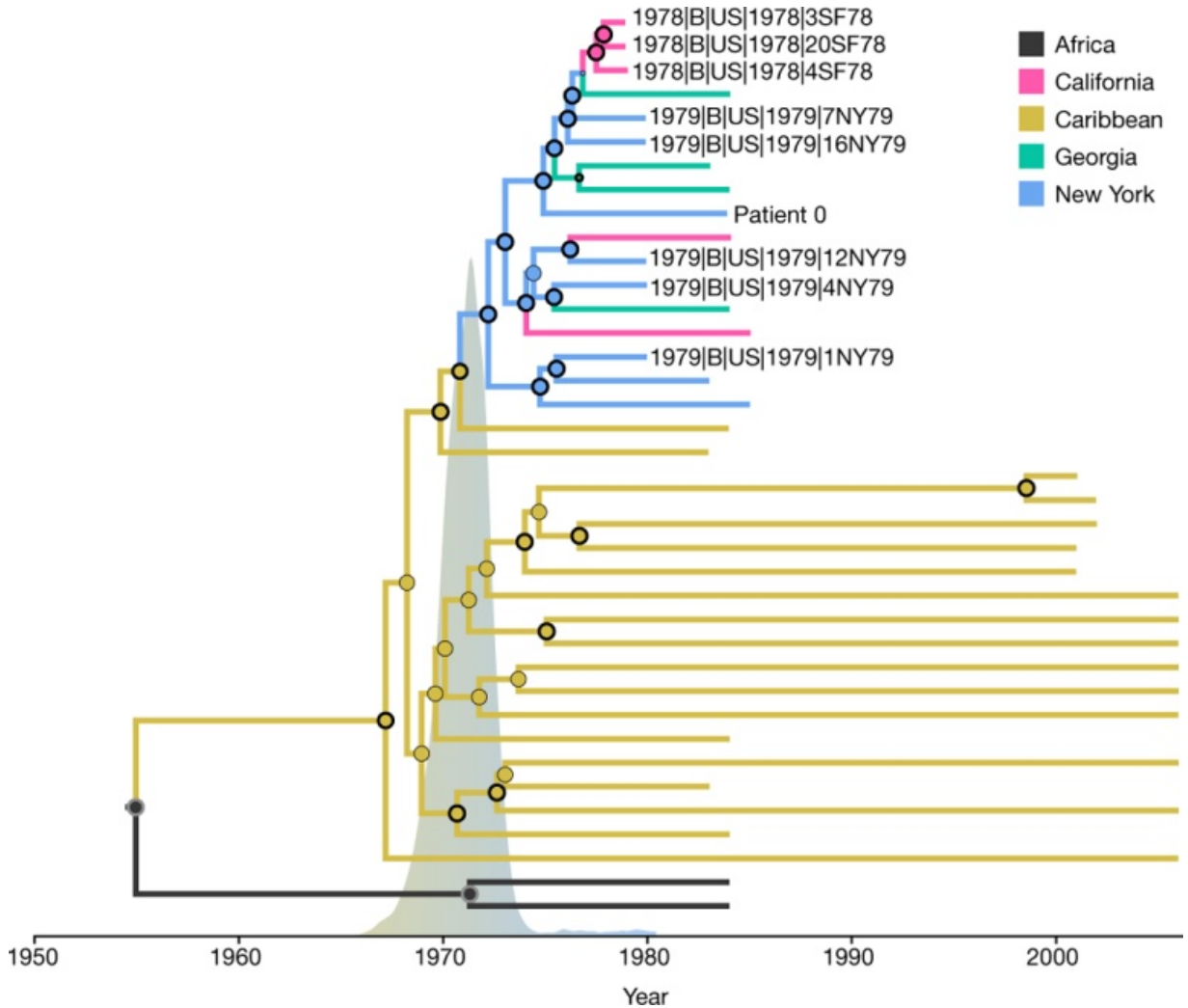




Evolutionary relationships



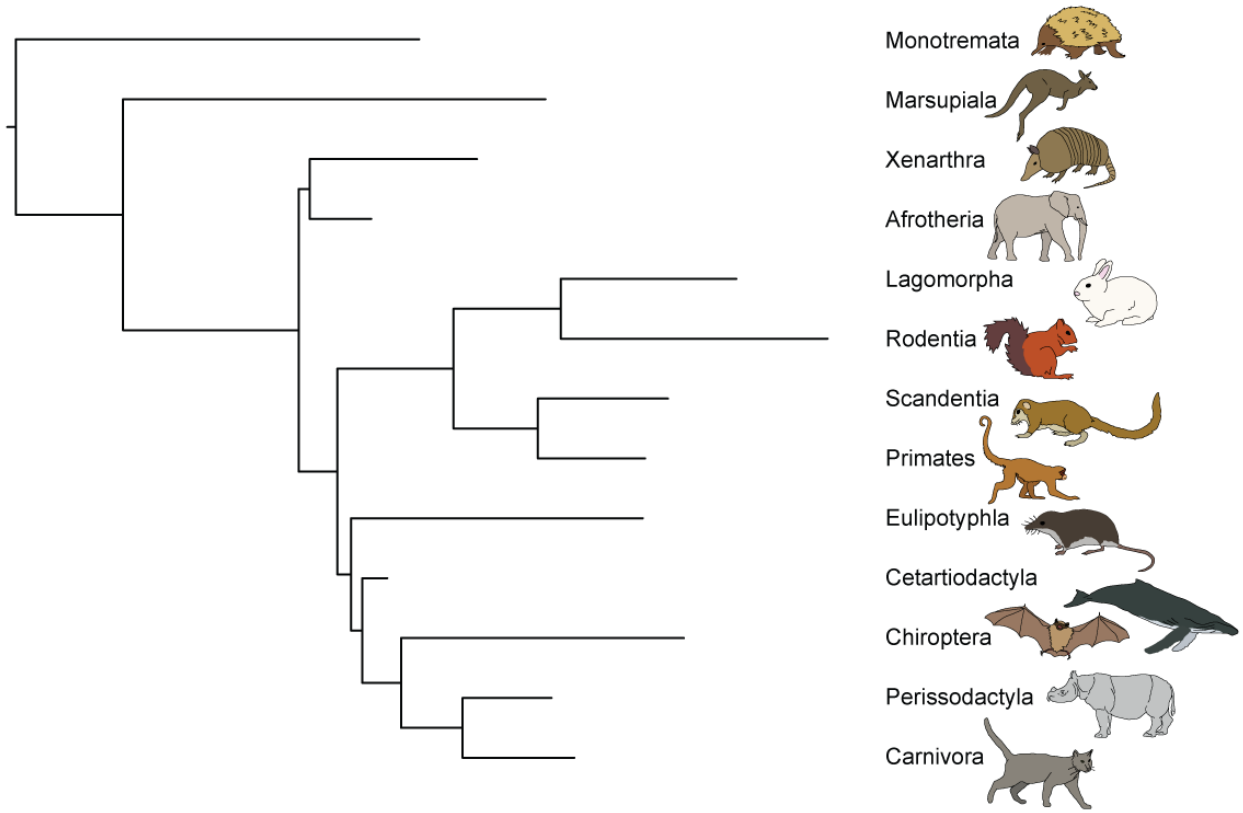
Evolutionary radiations



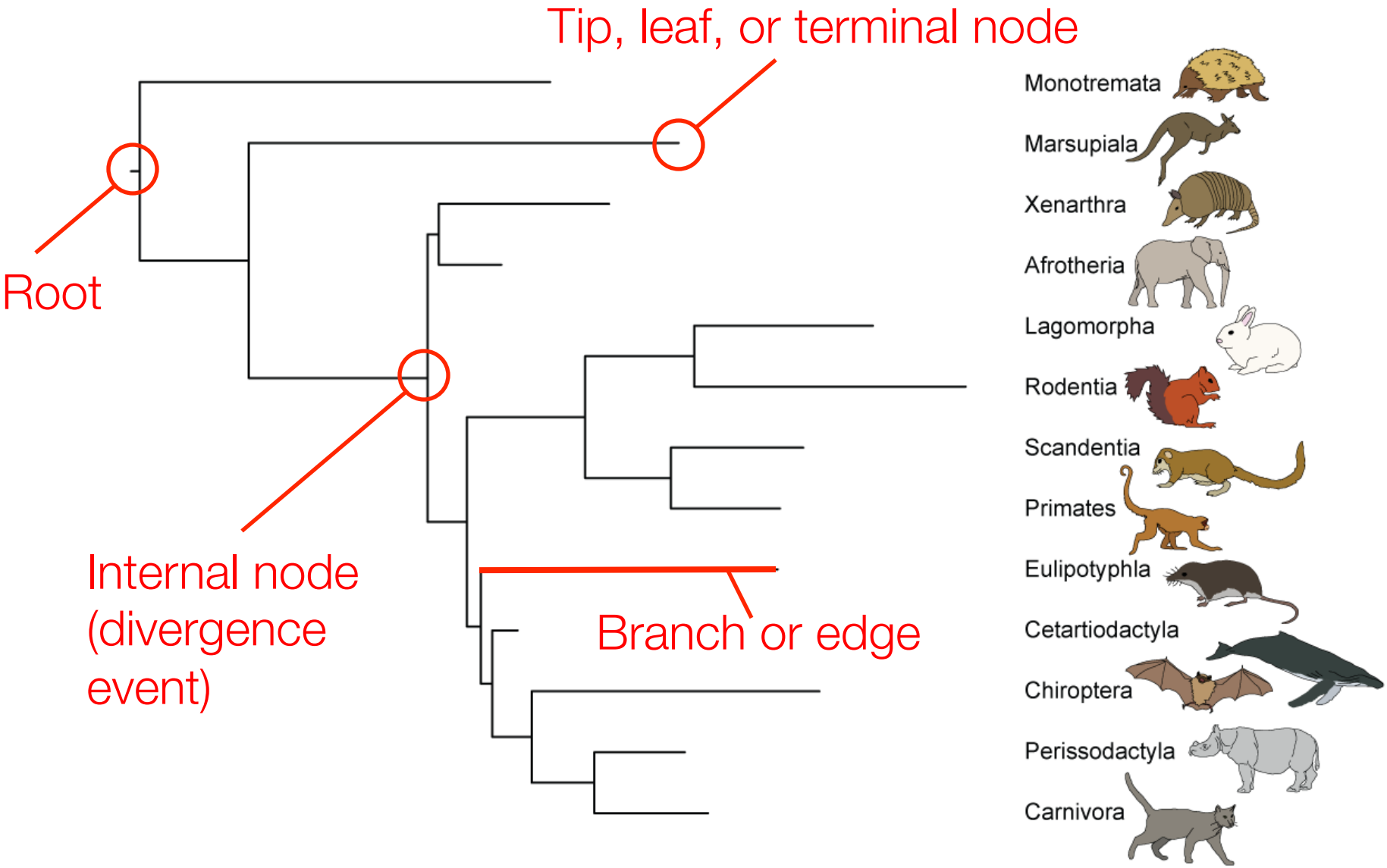
From Worobey et al. 2016 *Nature*

What is a phylogenetic tree?

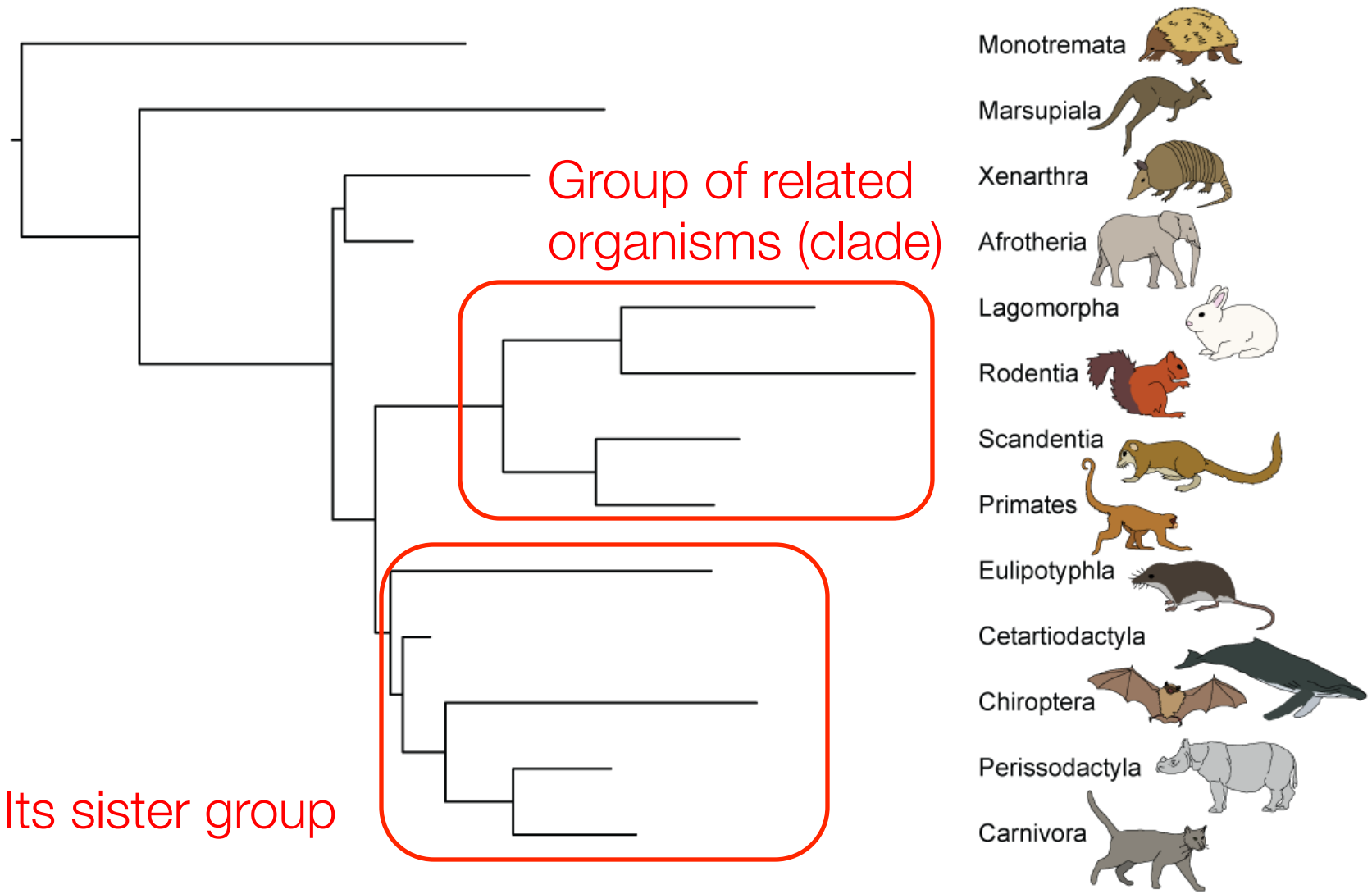
- A phylogenetic tree has two major components
  - Topology (relationships)
  - Branch lengths (amount of evolutionary change or time)



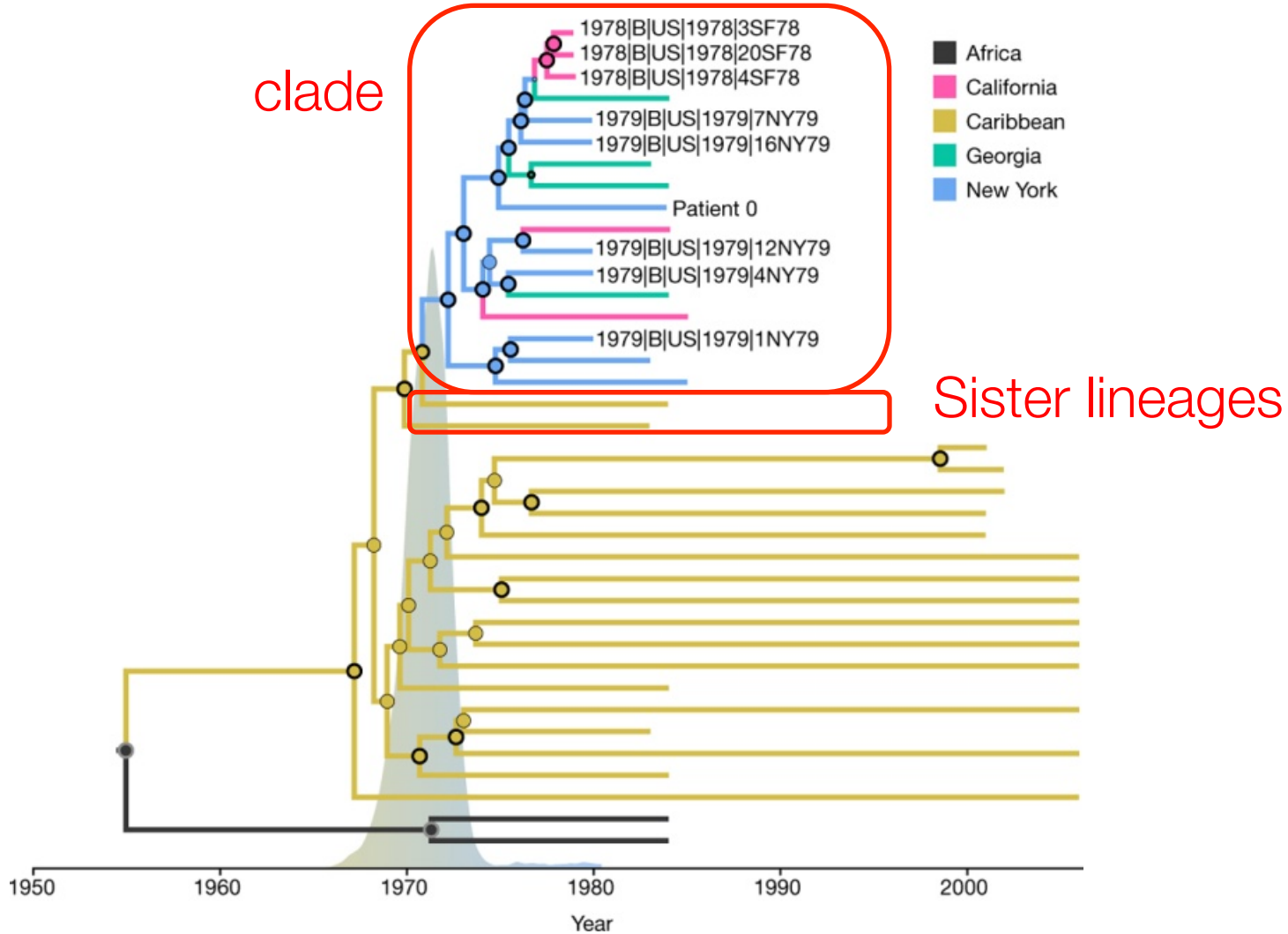
Phylogenetic trees



Phylogenetic trees

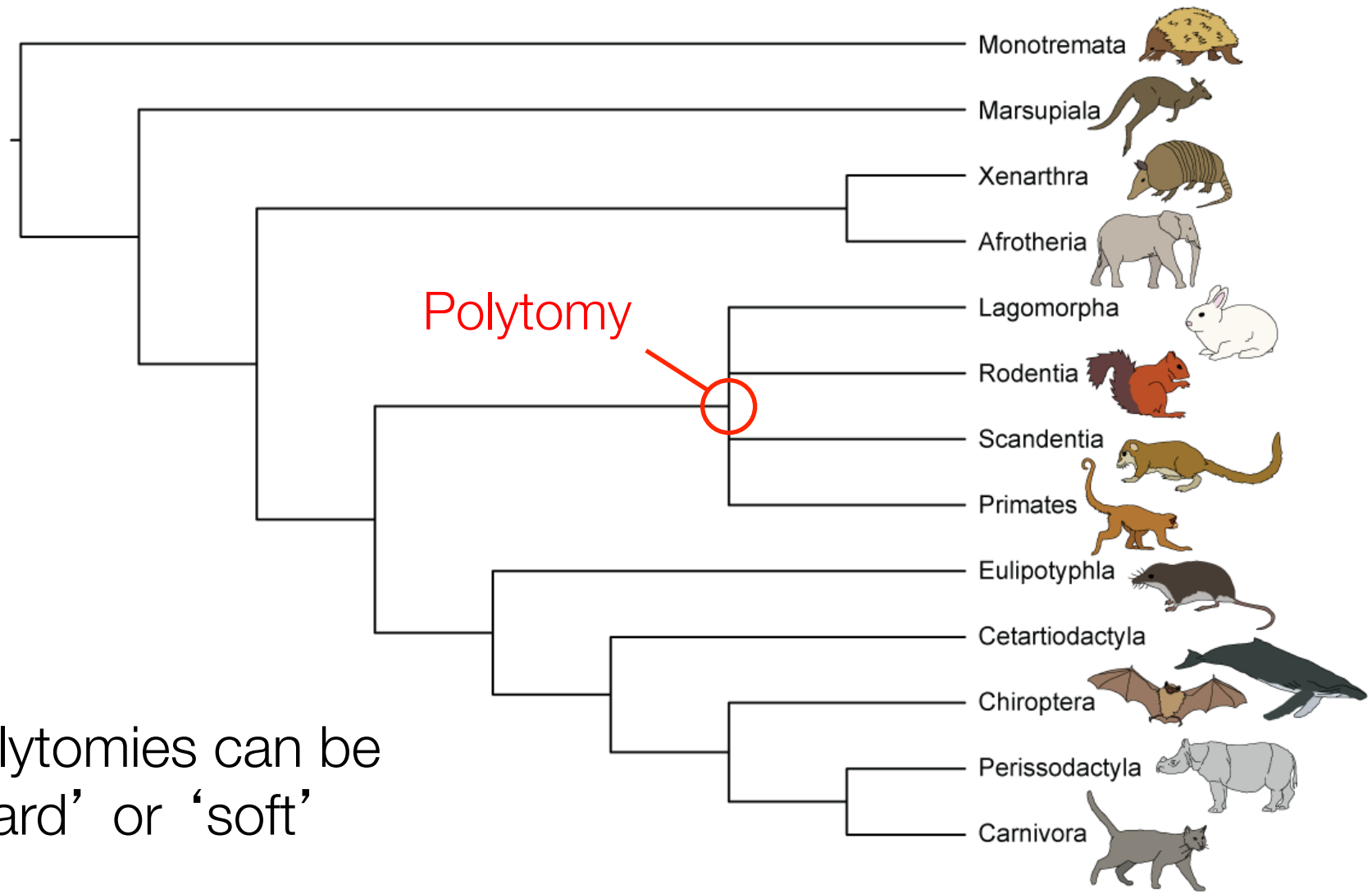


Phylogenetic trees



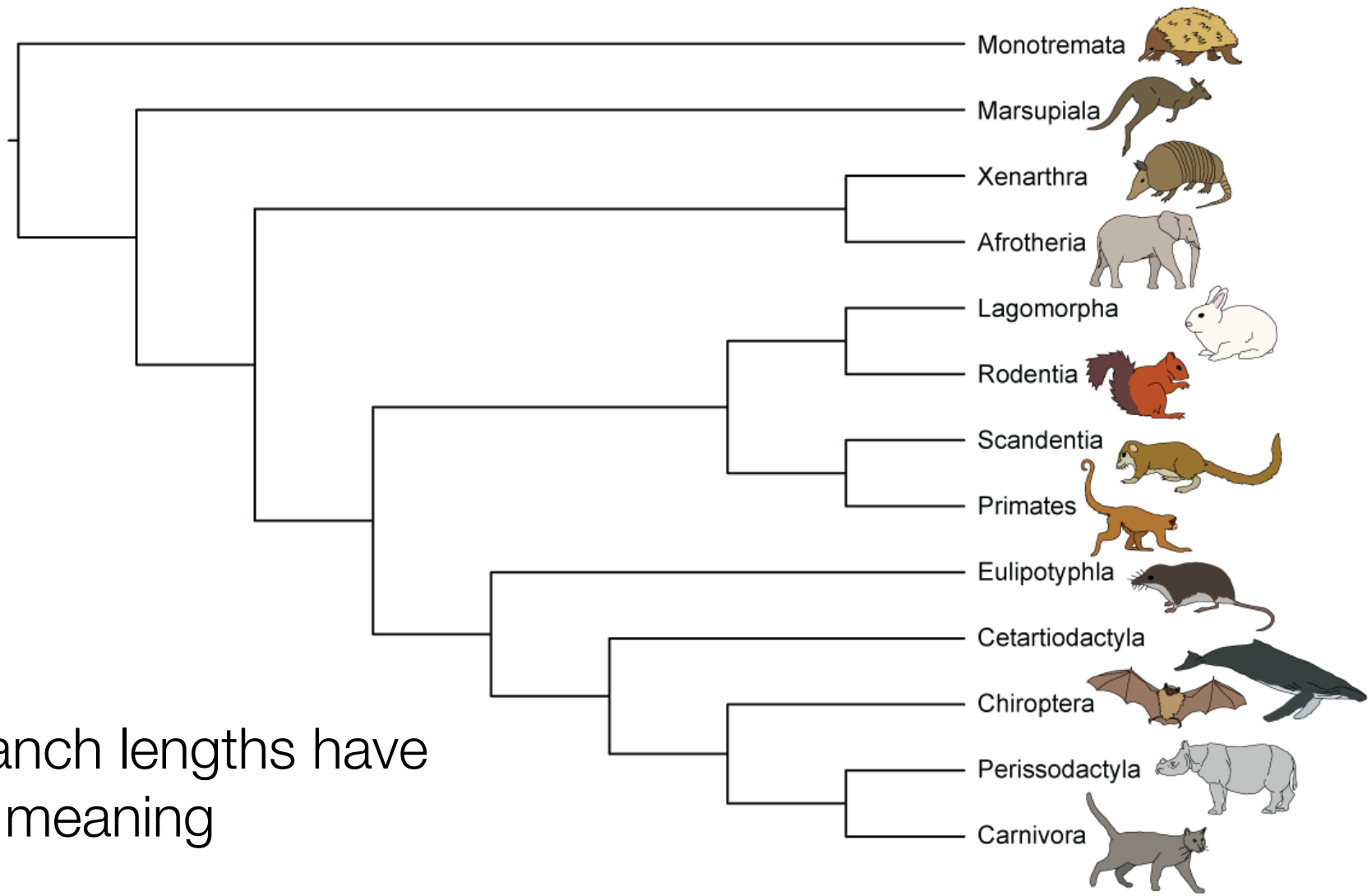
From Worobey et al. 2016 *Nature*

Phylogenetic trees



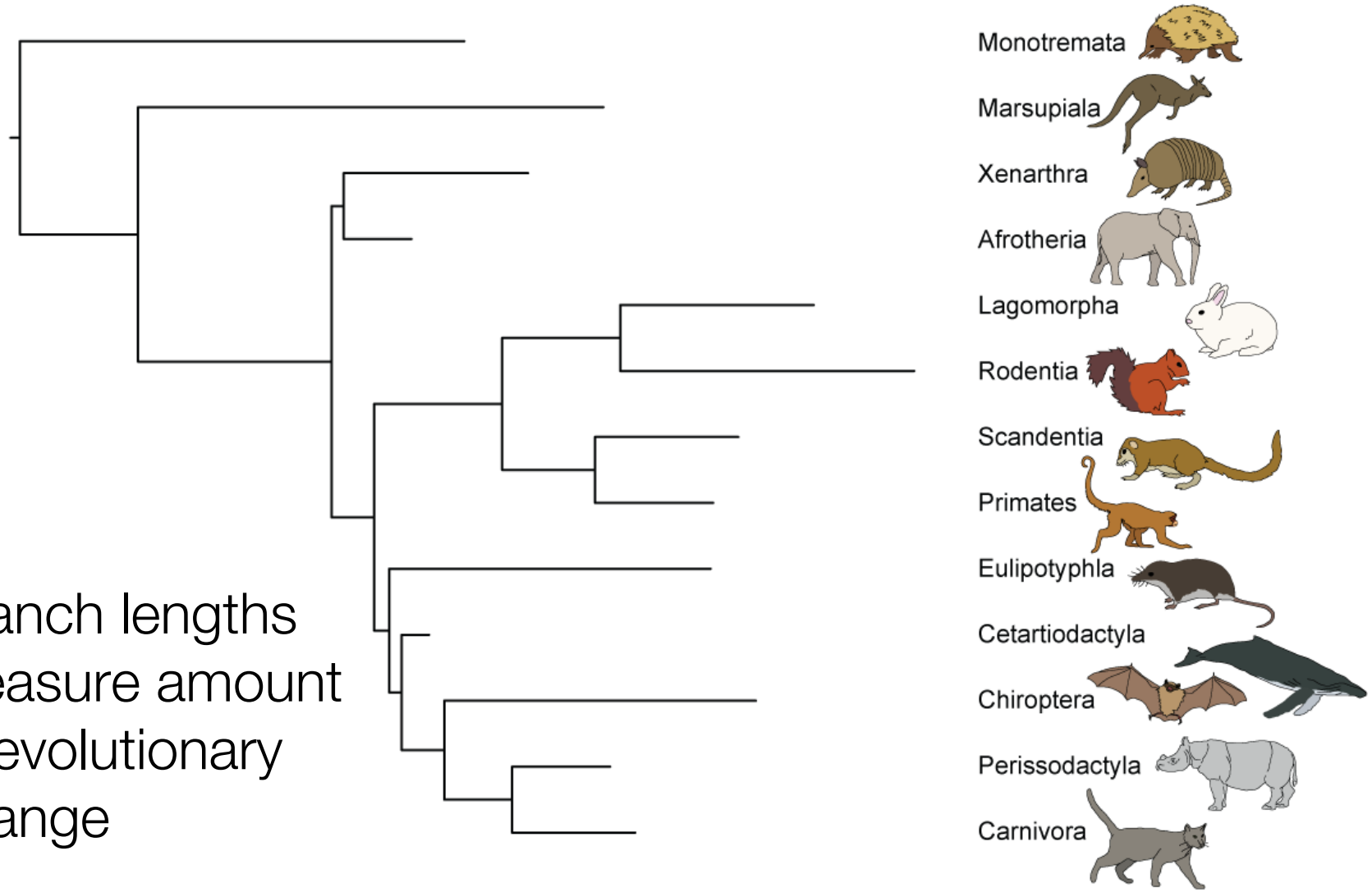
Polytomies can be ‘hard’ or ‘soft’

Phylogenetic trees: Cladogram



Branch lengths have no meaning

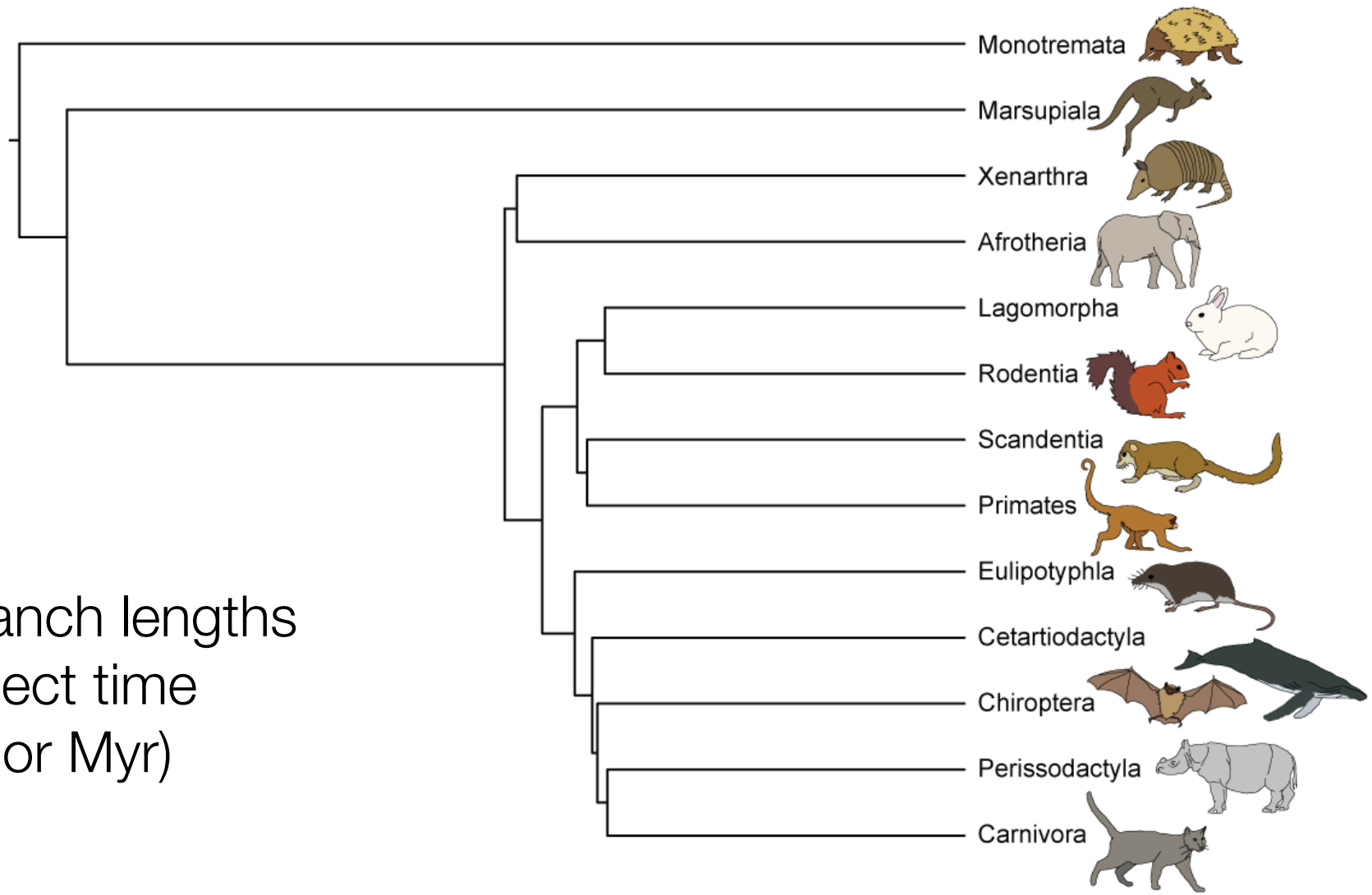
Phylogenetic trees: Phylogram



Branch lengths measure amount of evolutionary change

Phylogenetic trees: Chronogram

Ultrametric

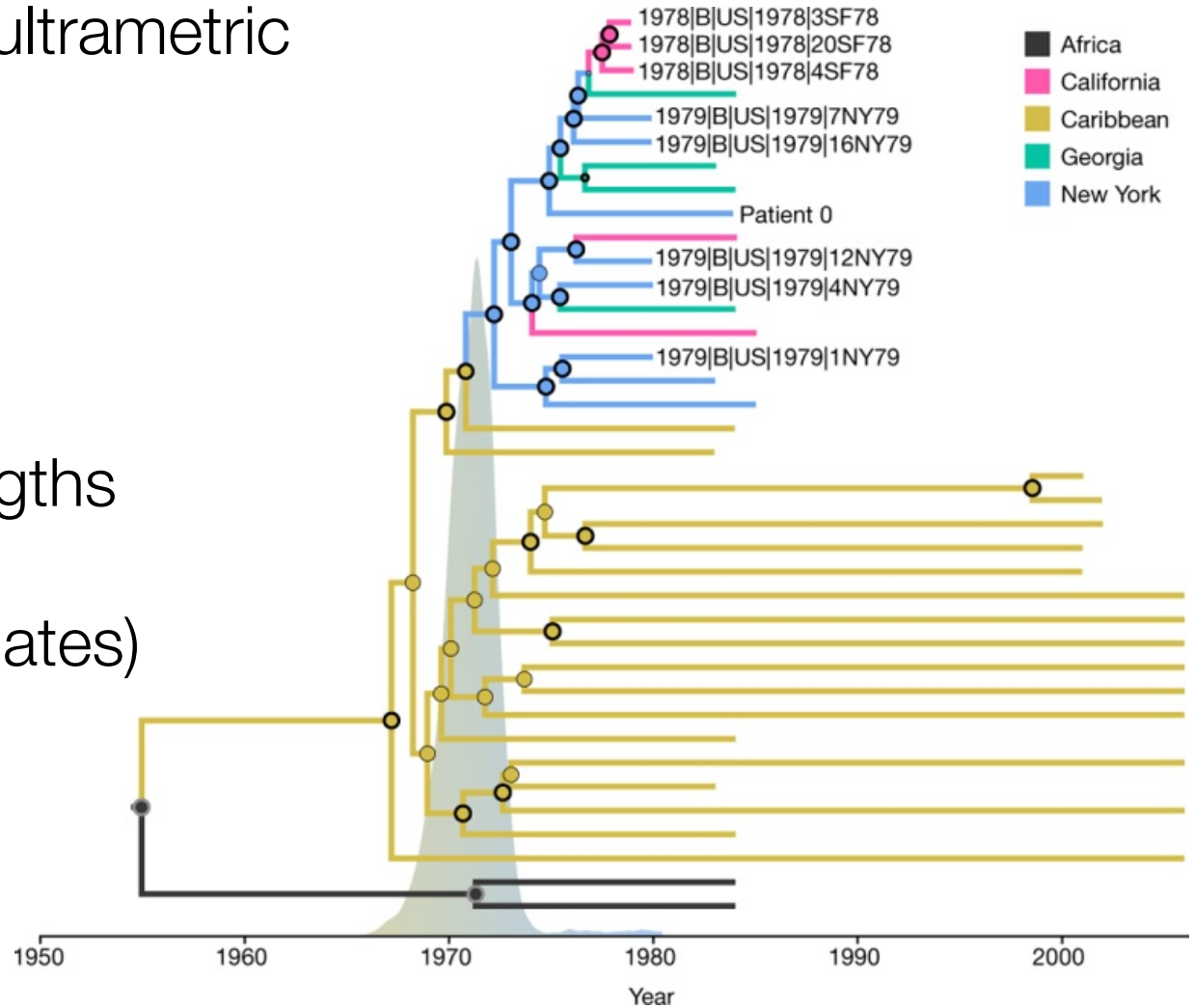


Branch lengths reflect time (yr or Myr)

Phylogenetic trees: Chronograms

Non-ultrametric

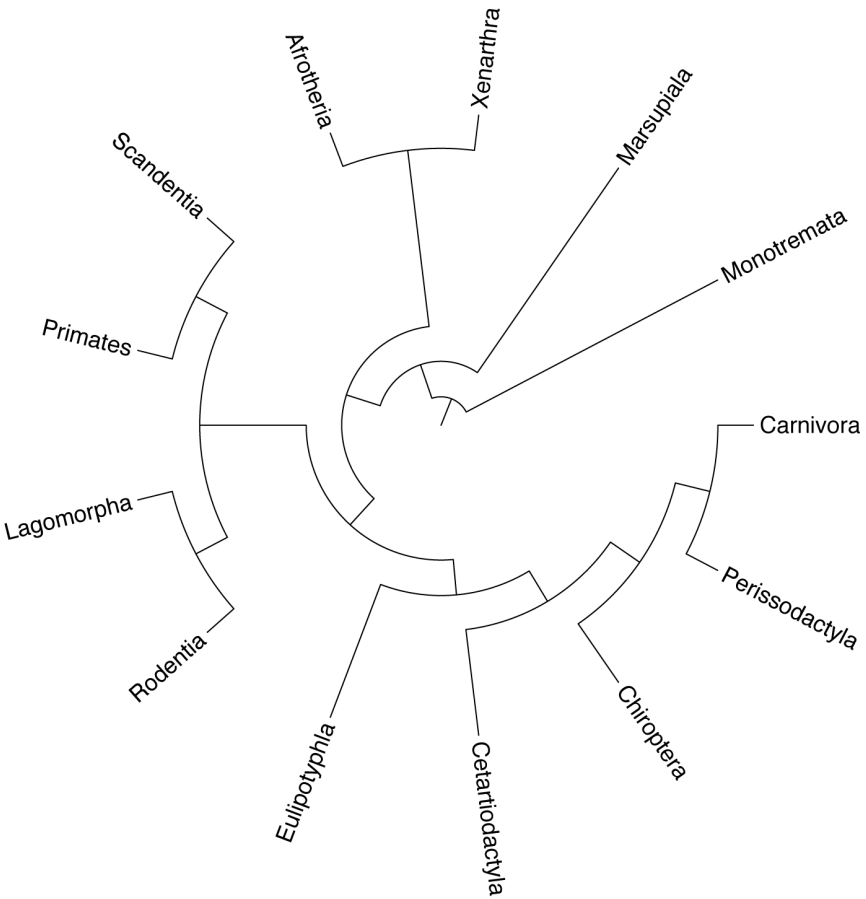
Branch lengths reflect time (calendar dates)



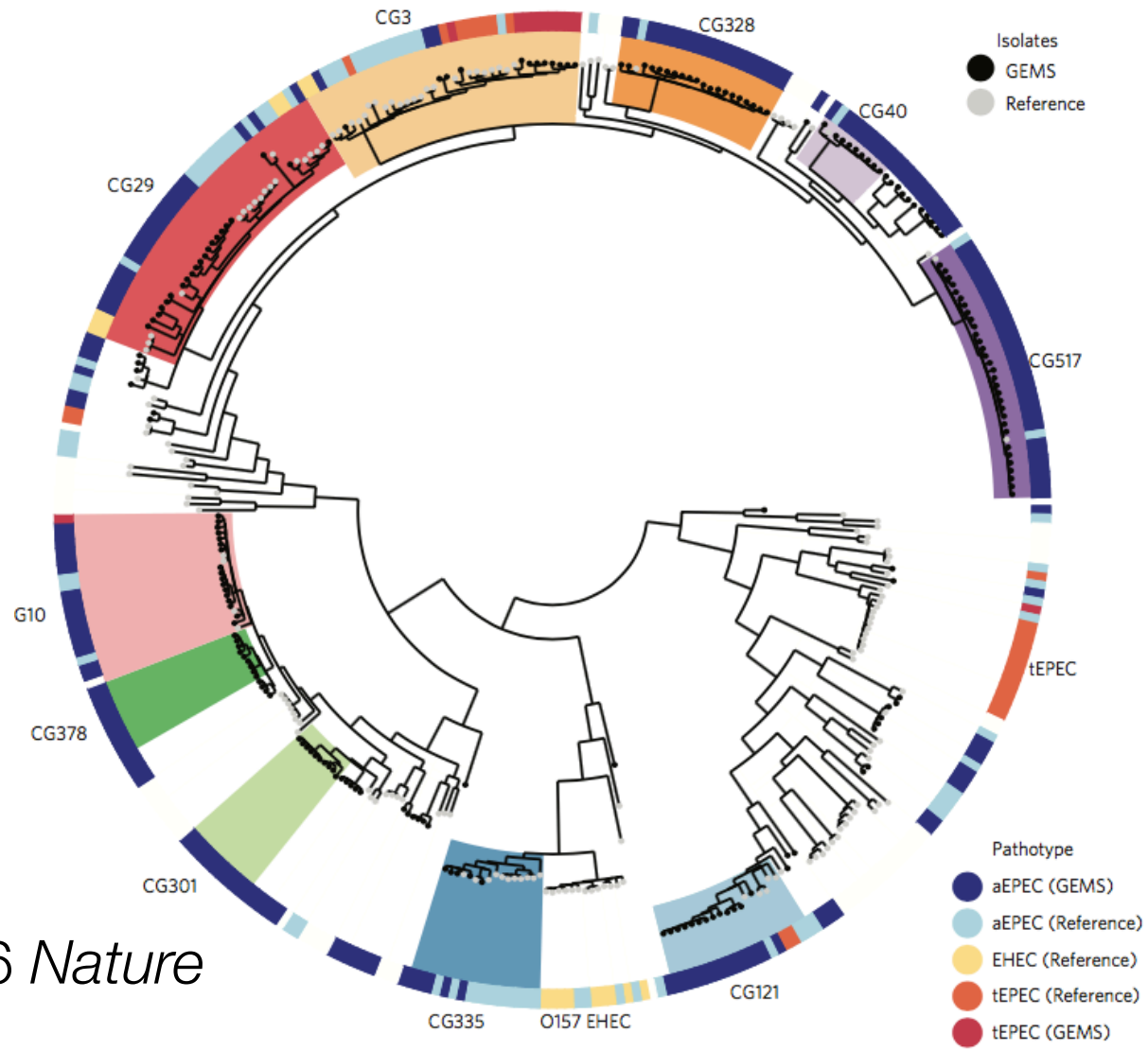
From Worobey et al. 2016 *Nature*

Phylogenetic trees: Circular

- Root is placed in centre
- Cladogram, phylogram, or chronogram
- Often used to visualise large trees
- Can be difficult to interpret
- Visually appealing



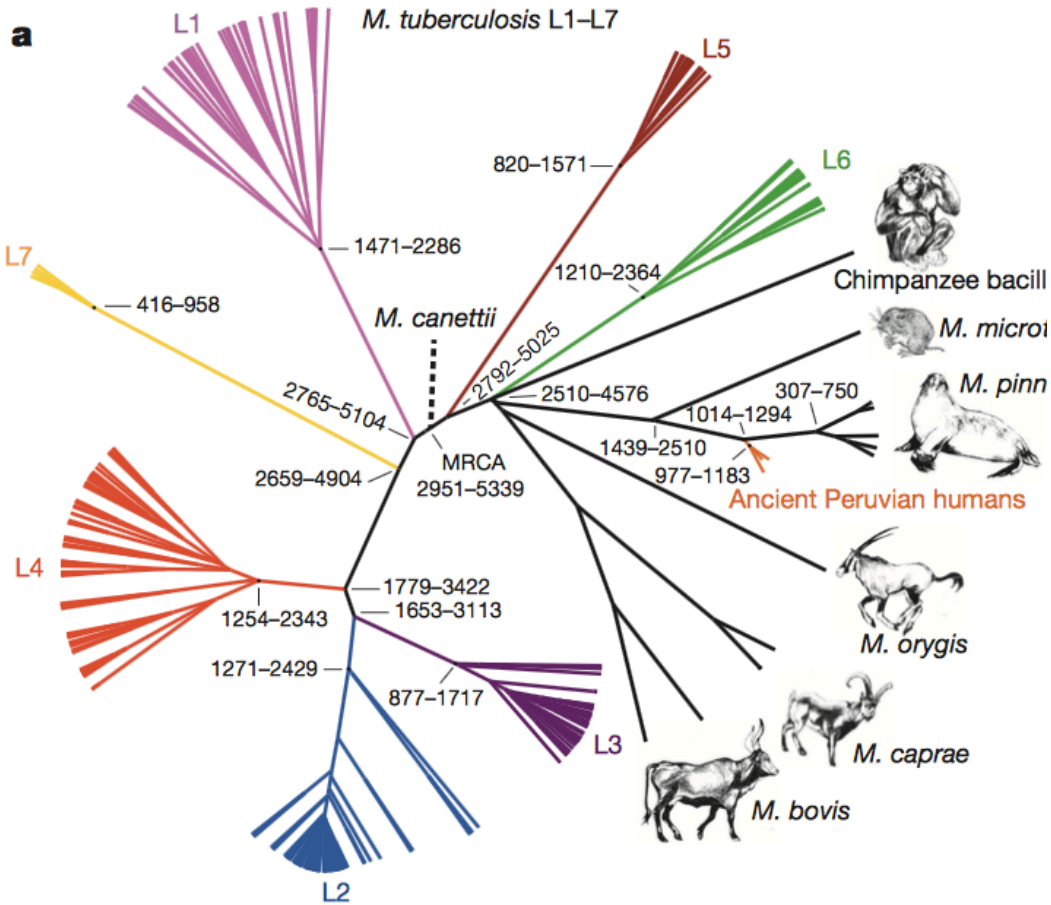
Phylogenetic trees: Circular



From Ingle et al. 2016 *Nature Microbiology*

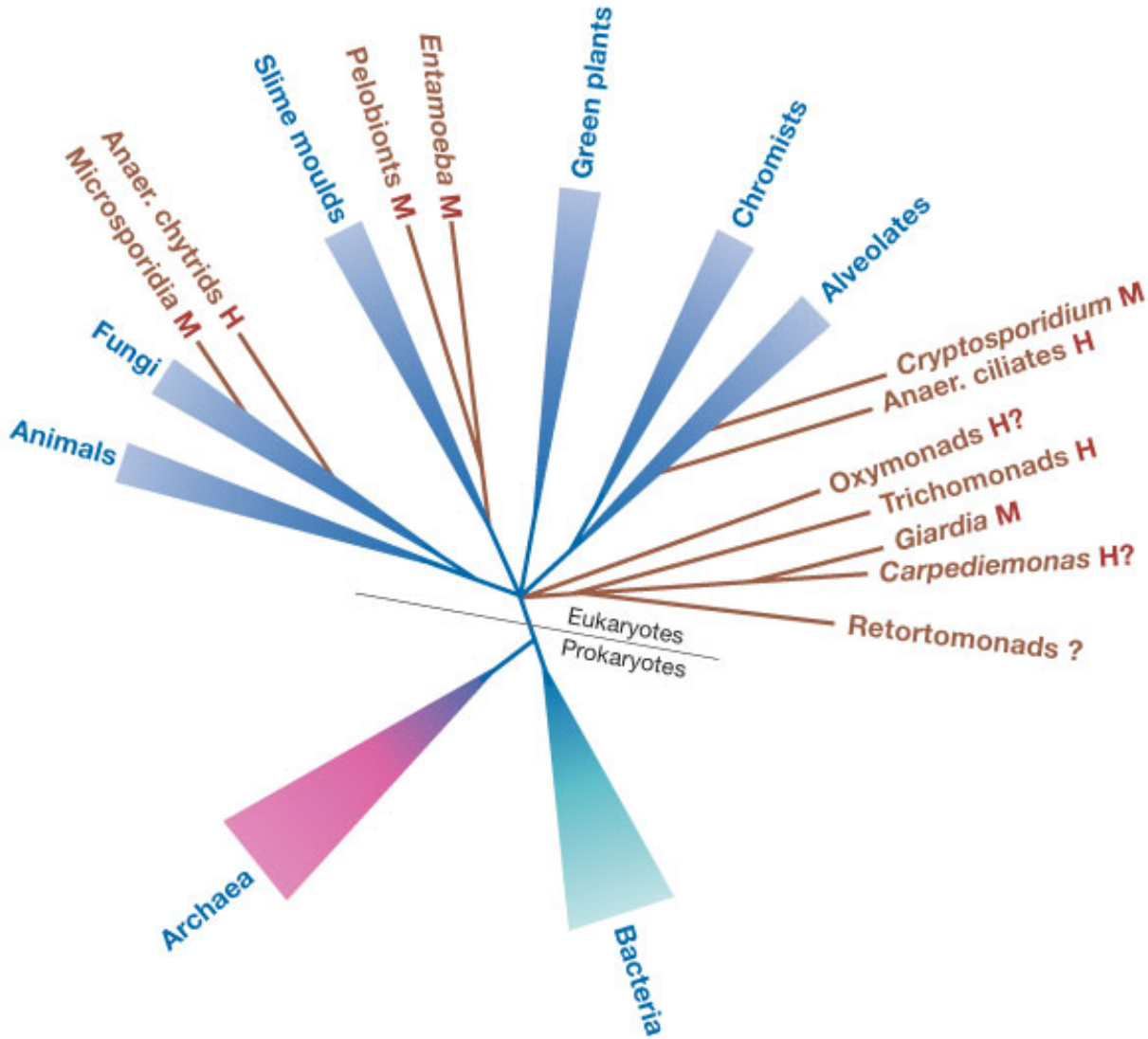
Phylogenetic trees: Unrooted

- Position of root is unknown
- Branch lengths usually represent amount of genetic change (substitutions/site)



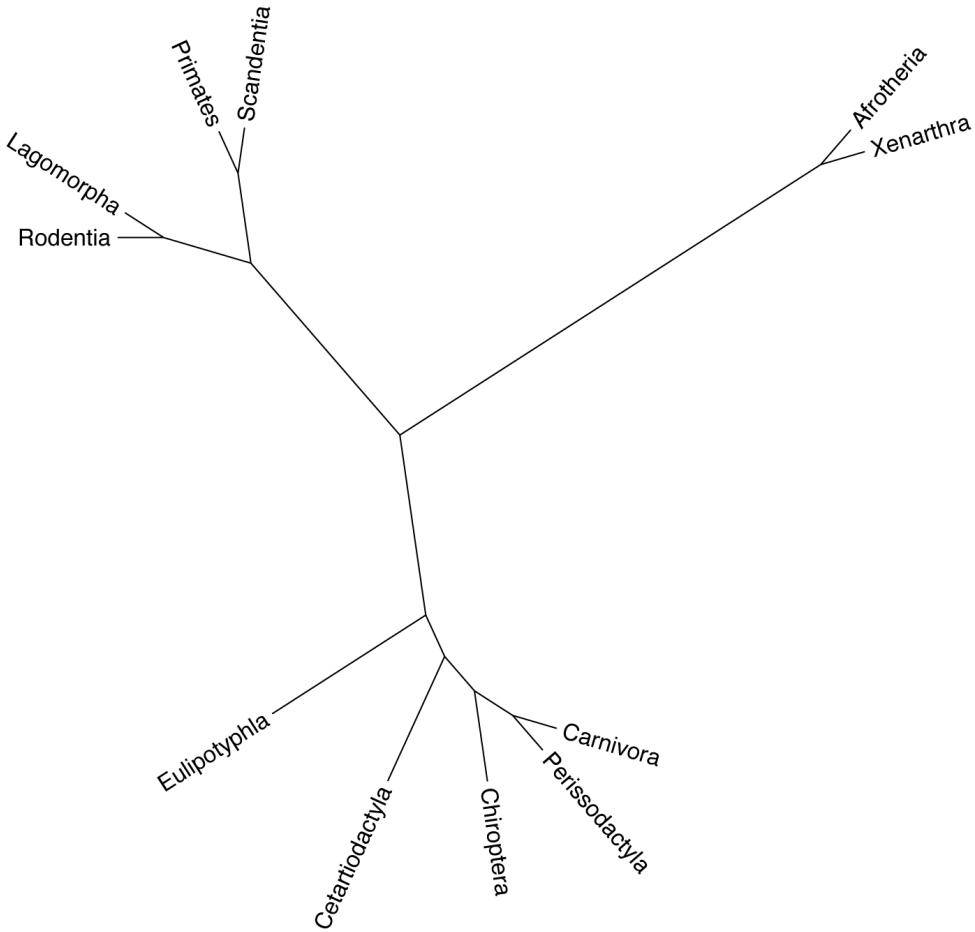
From Bos et al. 2014 *Nature*

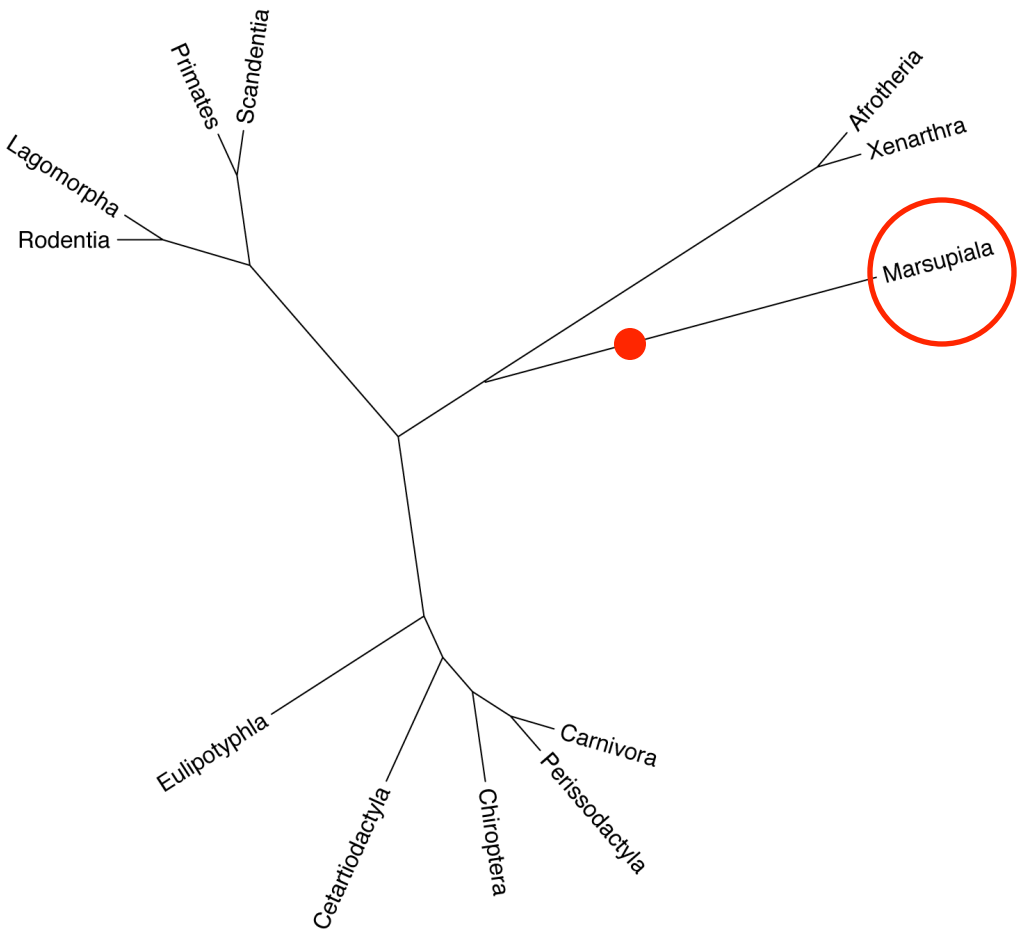
Phylogenetic trees: Unrooted



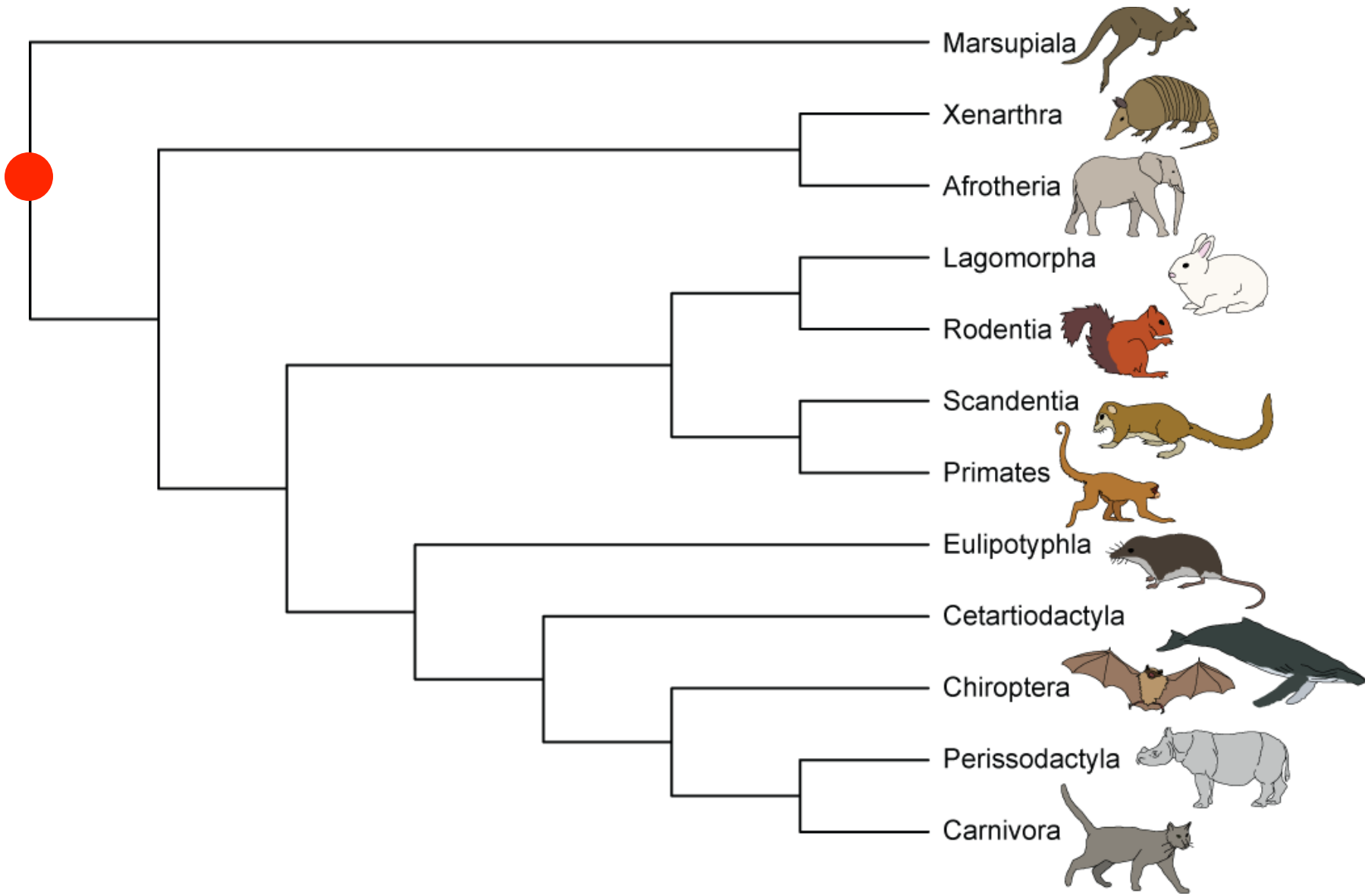
Rooting

- Can root a tree by including an outgroup taxon





Rooting


Marsupiala
Xenarthra
Afrotheria
Lagomorpha
Rodentia
Scandentia
Primates
Eulipotyphla
Cetartiodactyla
Chiroptera
Perissodactyla
Carnivora

Rooting

- Three methods for estimating the root of the tree
  1. Include an outgroup sequence
  2. Root at the midpoint of the tree
  3. Use a molecular clock

Phylogenetic trees: Newick format

- Without branch lengths (cladogram):
  - (Monotremata,(Marsupiala,((Afrotheria,Xenarthra),(((Rodentia,Lagomorpha),(Primates,Scandentia)),(Eulipotyphla,(Cetartiodactyla,(Chiroptera,(Carnivora,Perissodactyla))))))));

- With branch lengths (phylogram/chronogram):
  - (Monotremata:12.0,(Marsupiala:11.0,((Afrotheria:1.0,Xenarthra:1.0):9.0,(((Rodentia:1.0,Lagomorpha:1.0):2.0,(Primates:1.0,Scandentia:1.0):2.0):5.0,(Eulipotyphla:4.0,(Cetartiodactyla:3.0,(Chiroptera:2.0,(Carnivora:1.0,Perissodactyla:1.0):1.0):1.0):1.0):4.0):2.0):1.0):1.0);

Phylogenetic Analysis and Sequence Alignment

Phylogenetic analysis

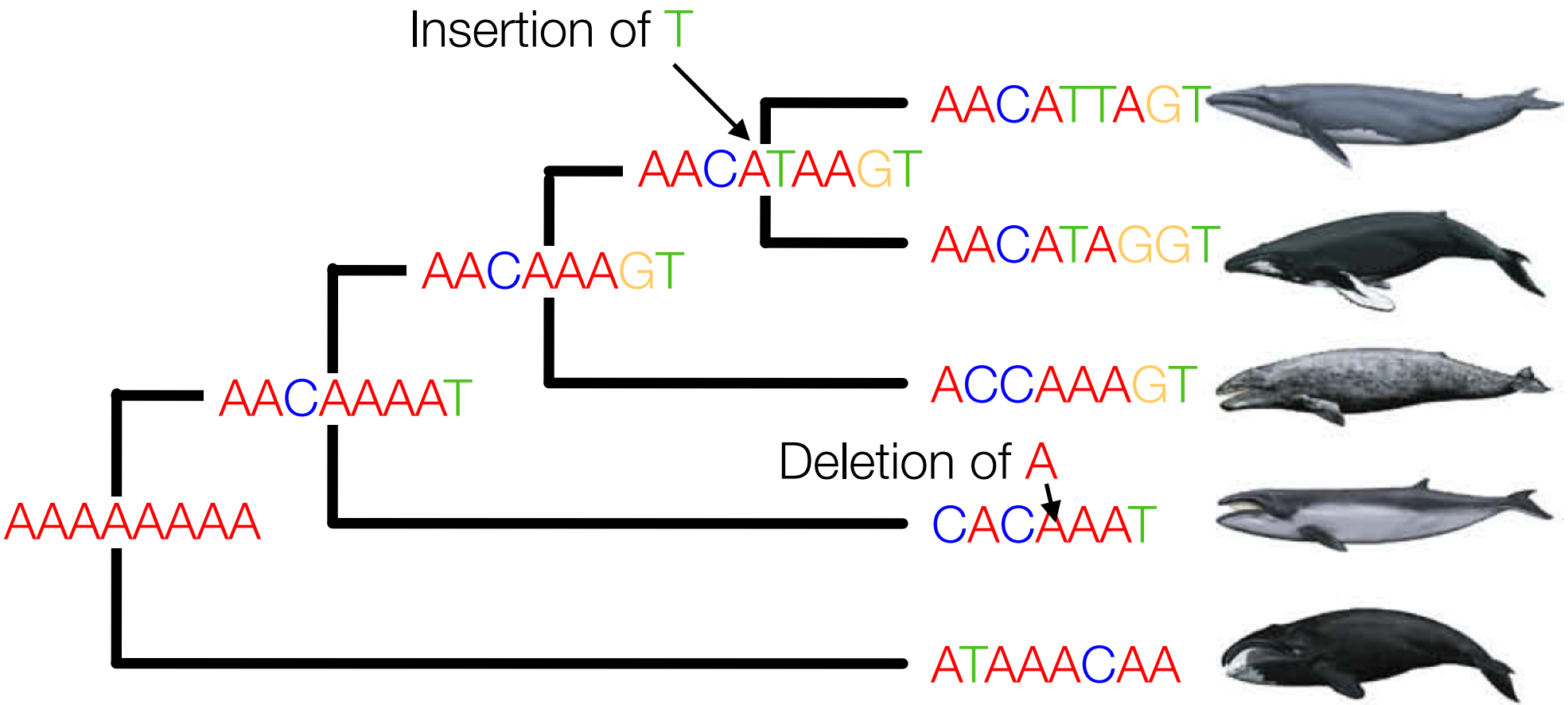
- Sometimes we know the phylogeny
  - Viral transmission histories
  - Documented pedigrees (family histories – humans, domesticated animals, lab organisms, etc.)
- Usually we do not know the phylogeny but we can estimate it
  - Morphological data
  - Molecular data
- Two fundamental results:
  - Estimate of evolutionary relationships
  - Estimate of evolutionary rates and time-scales

Fundamental assumptions

- Phylogenetic methods make several fundamental assumptions:
  - Each aligned site represents a set of orthologous characters
  - Sites in an alignment evolve independently
  - Lineages evolve independently
  - The relationships among the sequences can be represented by a bifurcating (binary) tree
  - Our models of evolution are accurate

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

Example: Whales



DNA sequence alignment

- Homologous sites need to be aligned
- Inferring insertions and deletions "indels"

AACATTAGT

AACATAGGT

ACCA-AAGT

CACA--AAT

ATAA-ACAA

Nucleic Acids Research, 1994, Vol. 22, No. 22 4673–4680

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*
European Molecular Biology Laboratory, Postfach 102209, Meyerhofstrasse 1, D-69012 Heidelberg, Germany

BMC Bioinformatics

BioMed Central

Software

Open Access

MUSCLE: a multiple sequence alignment method with reduced time and space complexity

Robert C Edgar*

prank-msa

Phylogeny-aware progressive sequence alignment method

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

```
AACATTAGT
AACATAGGT
ACCA-AAGT
CACA--AAT
ATAA-ACAA
```

DNA sequence alignment

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

```
AACATTAGT
AACATAGGT
ACCA-AAGT
CACA--AAT
ATAA-ACAA
```

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

```
AACATTAGT
AACATAGGT
ACCA-AAGT
CACA--AAT
ATAA-ACAA
```

DNA sequence alignment

- Indel – insertion or deletion
- Potentially informative
- Most phylogenetic methods do not really use indel data
- Maximum-likelihood and Bayesian methods typically treat them in the same way as missing data

```
AACATTAGT
AACATAGGT
ACCA-AAGT
CACA--AAT
ATAA-ACAA
```

MEGA

- *Molecular Evolutionary Genetics Analysis*
- Koichiro Tamura and Sudhir Kumar
- Population genetics
- Phylogenetics
  - Sequence alignment
  - Model selection
  - Maximum parsimony
  - Distance-based methods
  - Maximum likelihood

Go to Practical 1: Sequence alignment in *MEGA*

MoPad, Data and prac in github

https://public.etherpad-mozilla.org/p/phyloworkshop_melbourne

https://github.com/sebastianduchene/Phyloworkshop_Melbourne_2017