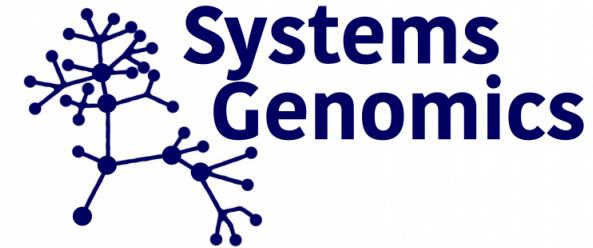




PRISM²



Introduction to Pathogen Phylogenetic analyses

Contributors

19 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 and Macquarie University

Robert Lanfear



Queensland University of Technology

Matt Phillips



Presenters



Sebastian Duchene

McKenzie Postdoctoral Fellow
Centre for Systems Genomics



Jane Hawkey

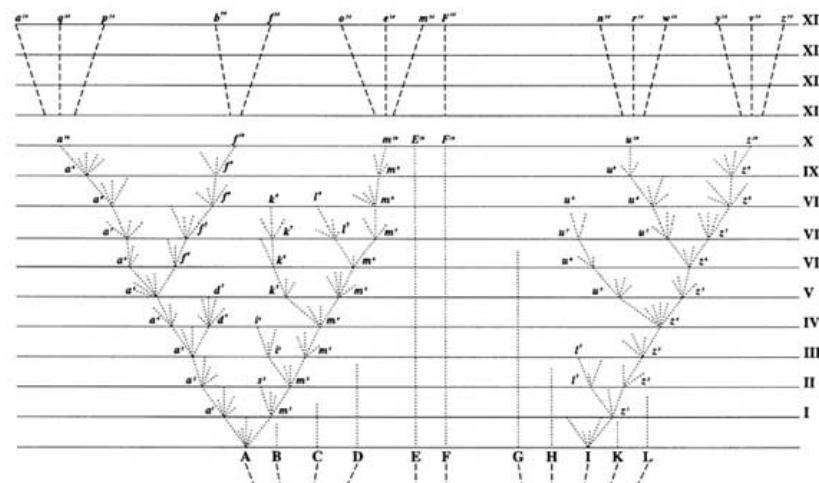
PhD candidate
Centre for Systems Genomics



Alex Strich, James McCaw, and Jodie McVernon

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 – 14:30 Lecture 1.3: Phylogenetic methods

14:30 – 15:00 Practical 1b: Model selection in *MEGA*

15:00 – 15:15 Practical 1c: Maximum likelihood analysis in *PhyML*

Workshop programme – Day 1

15:15 – 15:30 Questions and discussion

– *Coffee break* –

15:45 – 16:00 **Lecture 1.4:** The molecular clock

16:00 – 16:30 **Practical 1d:** Assessing temporal structure in *TempEst*

16:30 – 17:00 Questions and discussion

Workshop programme – Day 2

09:15 – 9:30 Arrival

09:30 – 10:30 **Lecture 2.1:** Bayesian phylogenetic analysis

10:30 – 11:00 **Lecture 2.2:** Models and priors

– *Coffee break* –

11:30 – 12:00 **Practical 2:** Markov Chain Monte Carlo

12:00 – 12:30 **Lecture 2.3:** Estimating evolutionary rates and timescales

– *Lunch break* –

13:30 – 14:30 **Practical 3:** Molecular dating using *BEAST*

14:30 – 14:45 Questions and discussion

Workshop programme – Day 2

14:45 – 15:15 **Lecture 2.4:** Infectious disease phylodynamics

– *Coffee break* –

15:30 – 16:15 **Practical 4:** Molecular epidemiology in *BEAST*

16:00 – 17:00 Phylogenetics clinic and open questions

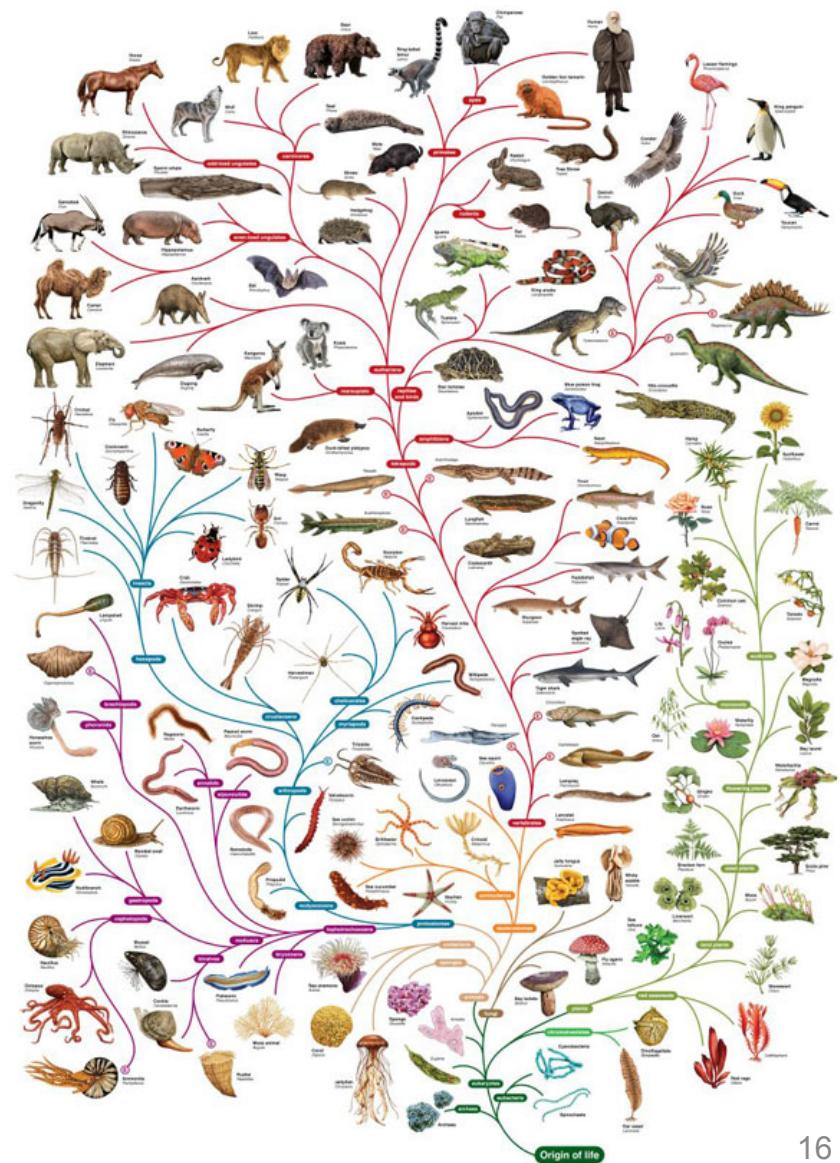
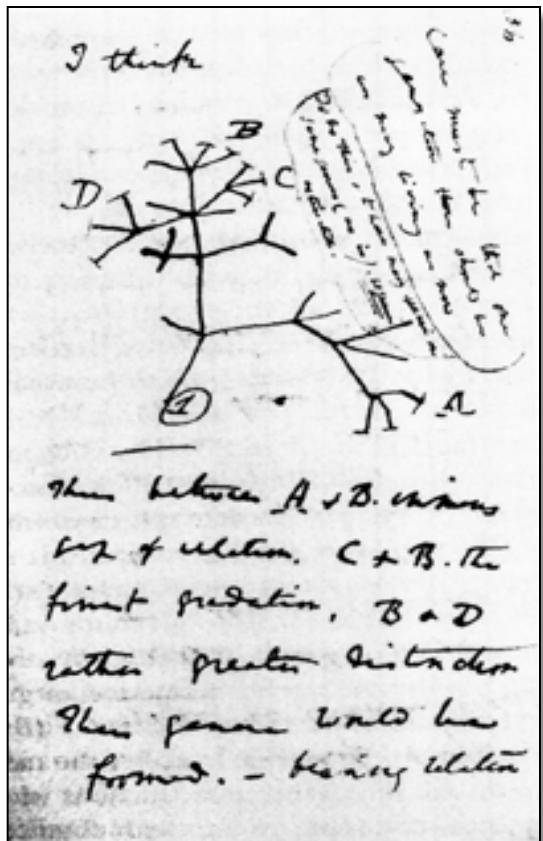
Lecture 1.1:

Introduction to Molecular Phylogenetics

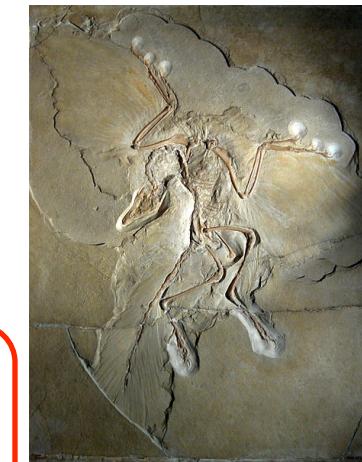
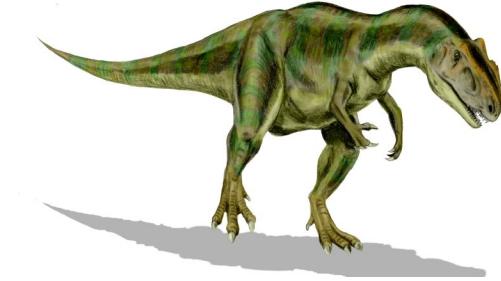
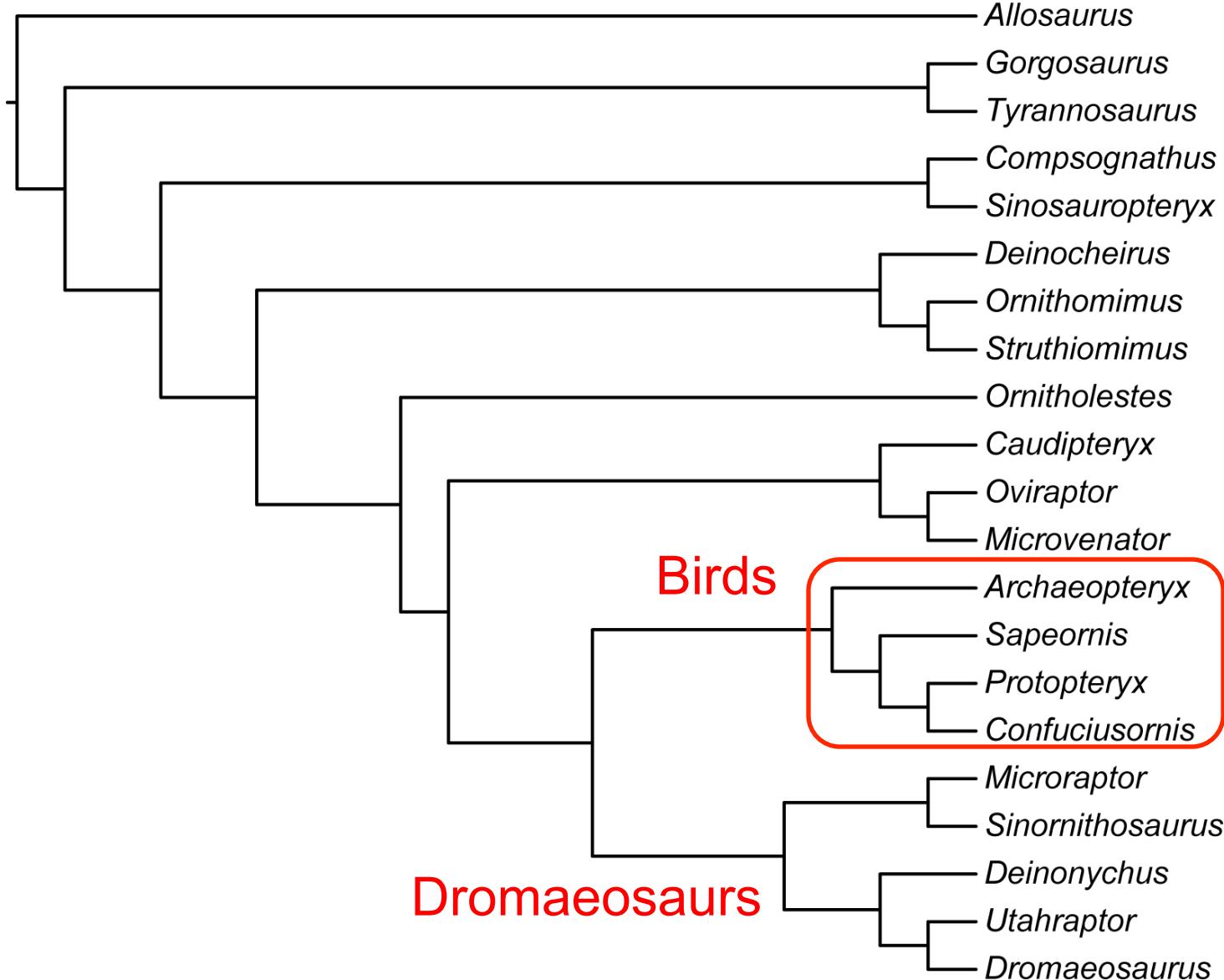
Understanding Phylogenetic Trees

What is a phylogenetic tree?

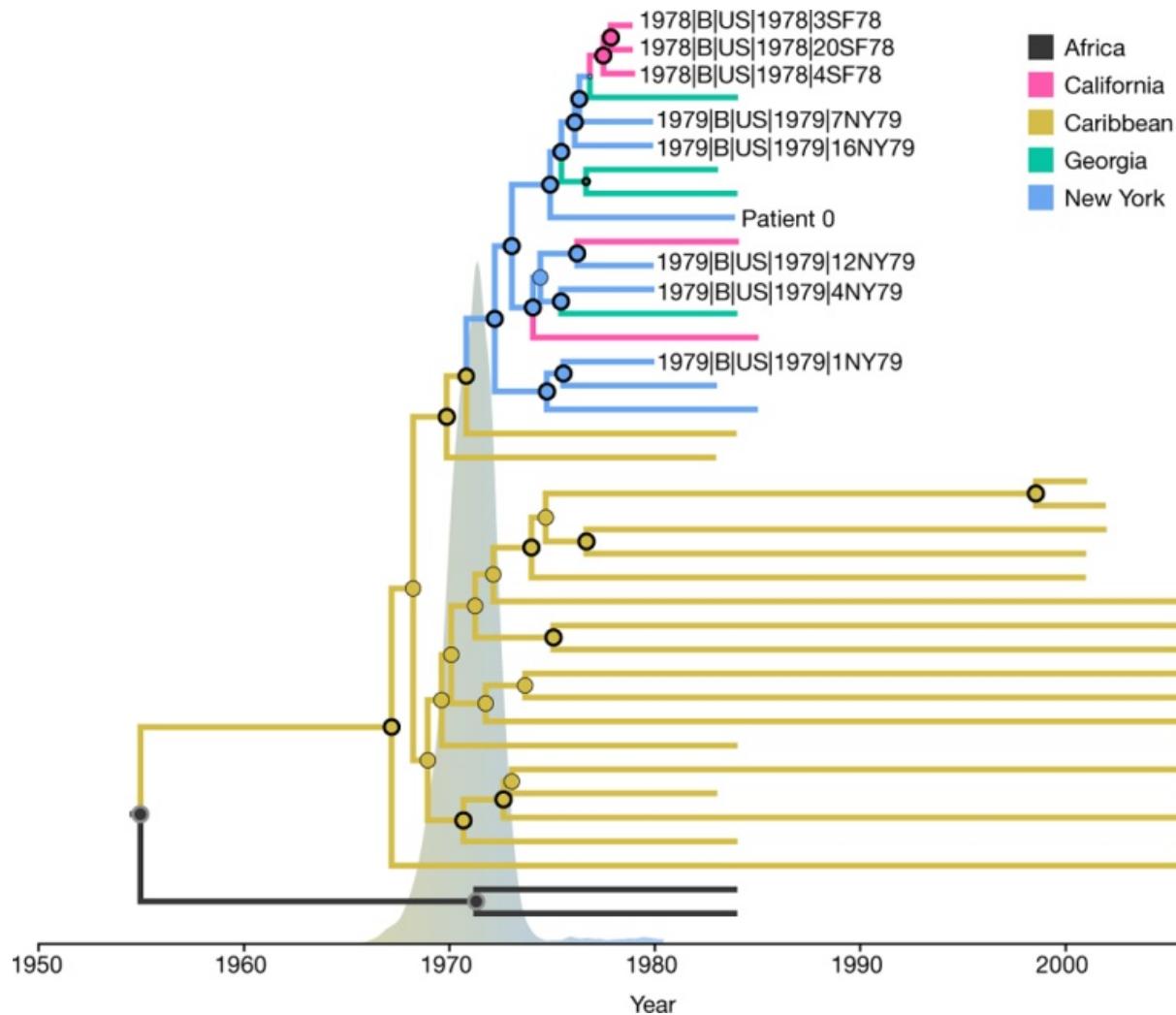
- The phylogeny refers to the true evolutionary relationships among a set of organisms



Evolutionary relationships



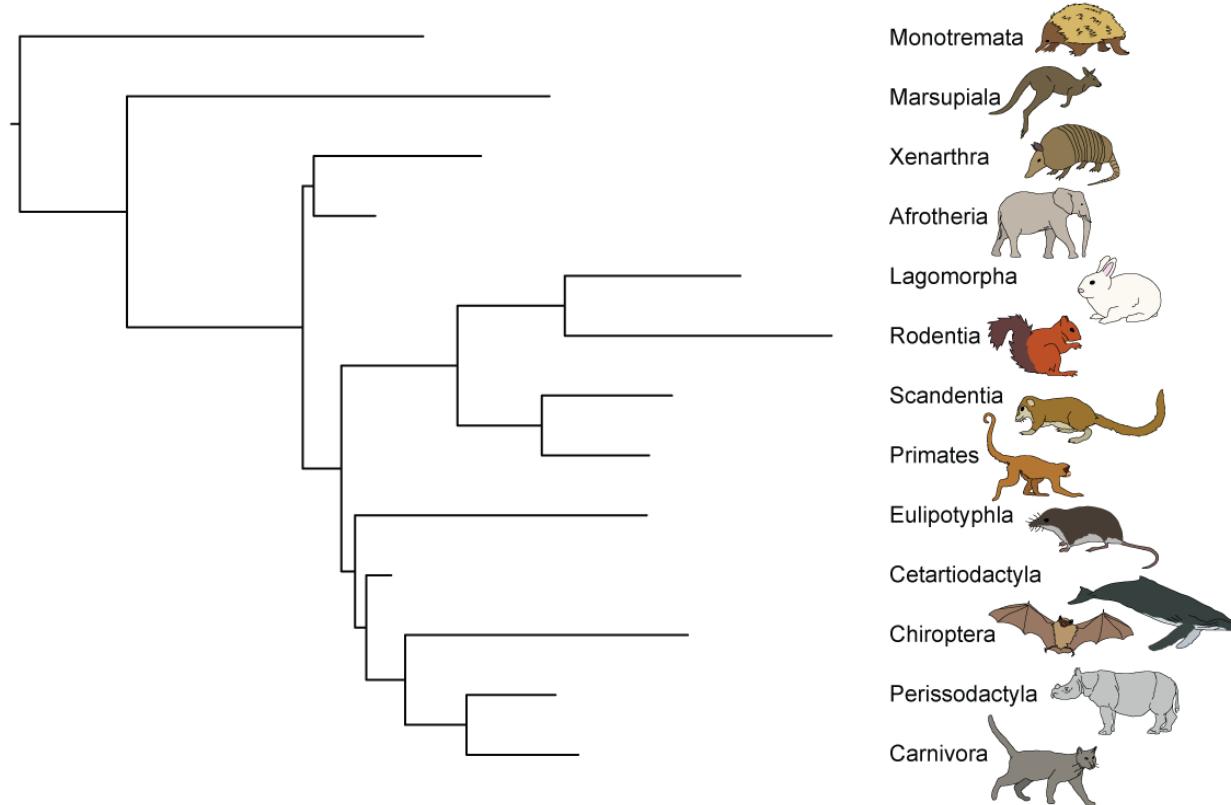
Evolutionary radiations



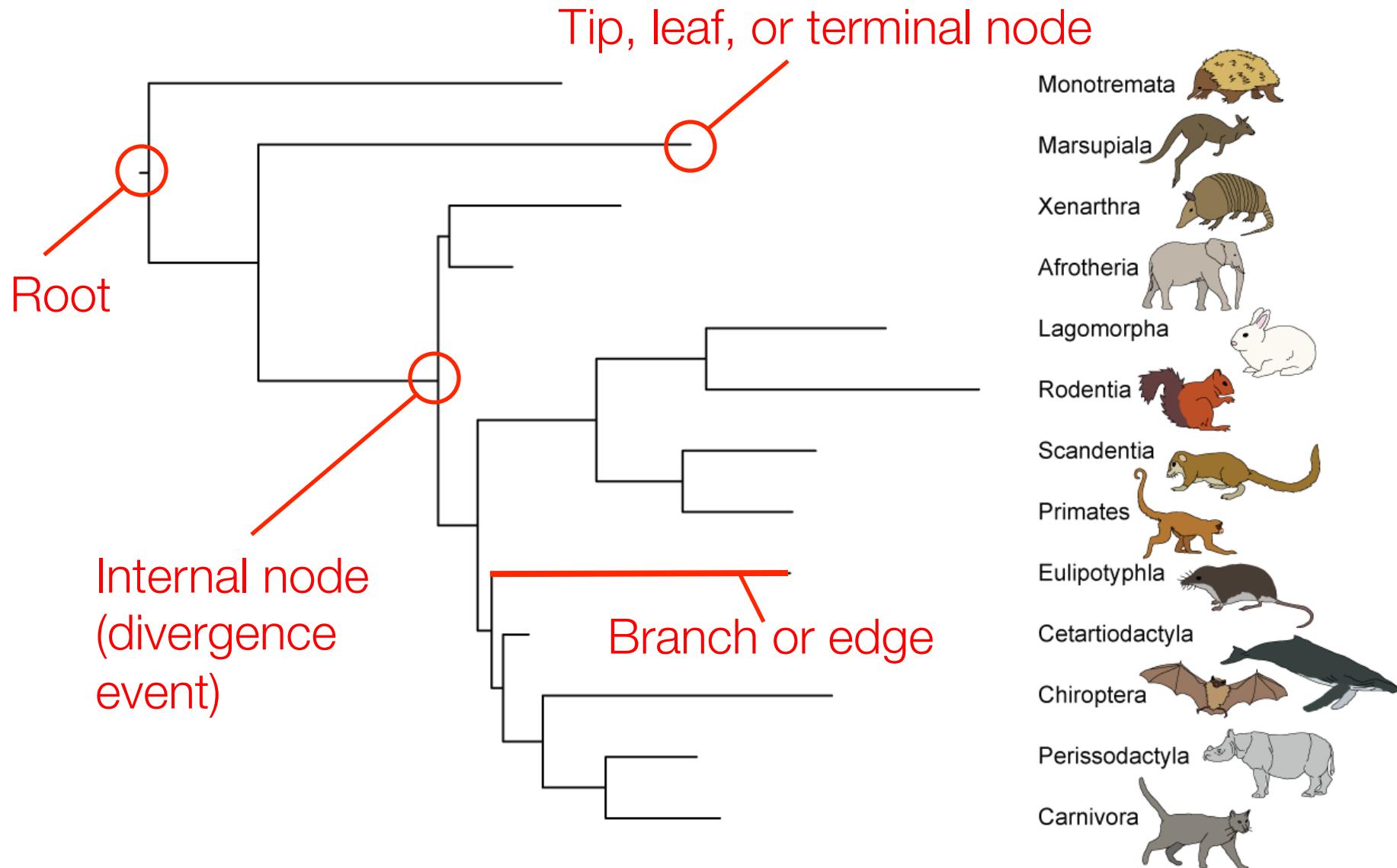
From Worobey et al. 2016 *Nature*

What is a phylogenetic tree?

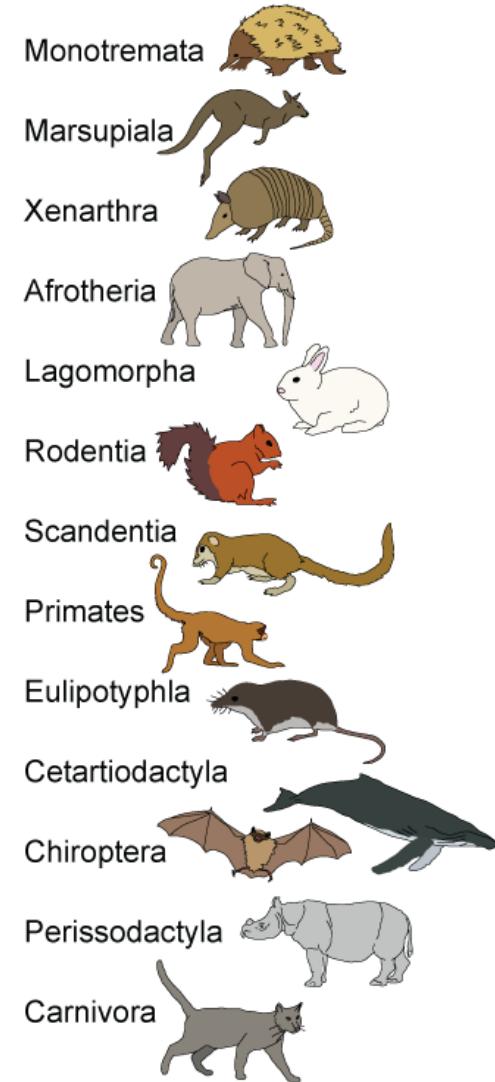
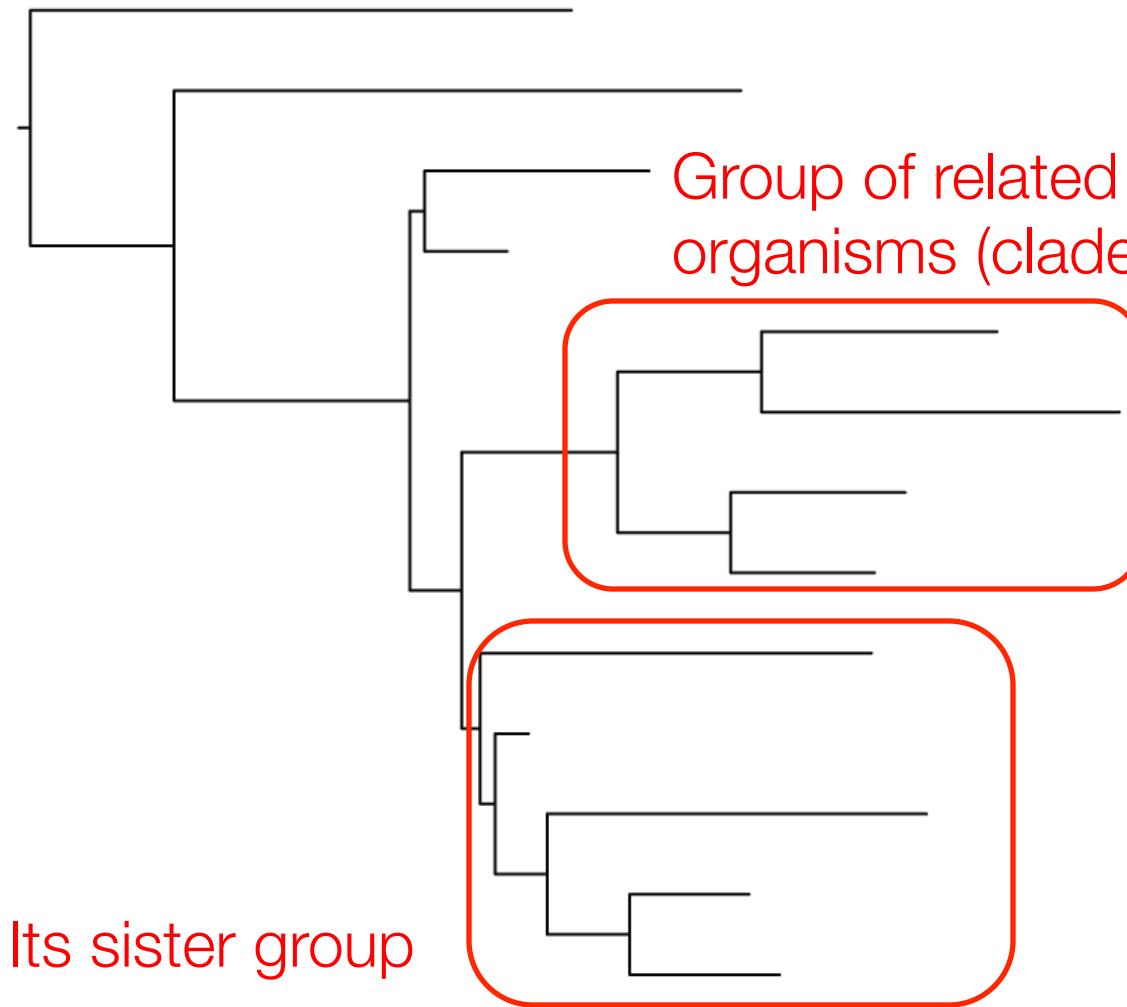
- A phylogenetic tree has two major components
 - Topology (relationships)
 - Branch lengths (amount of evolutionary change or time)



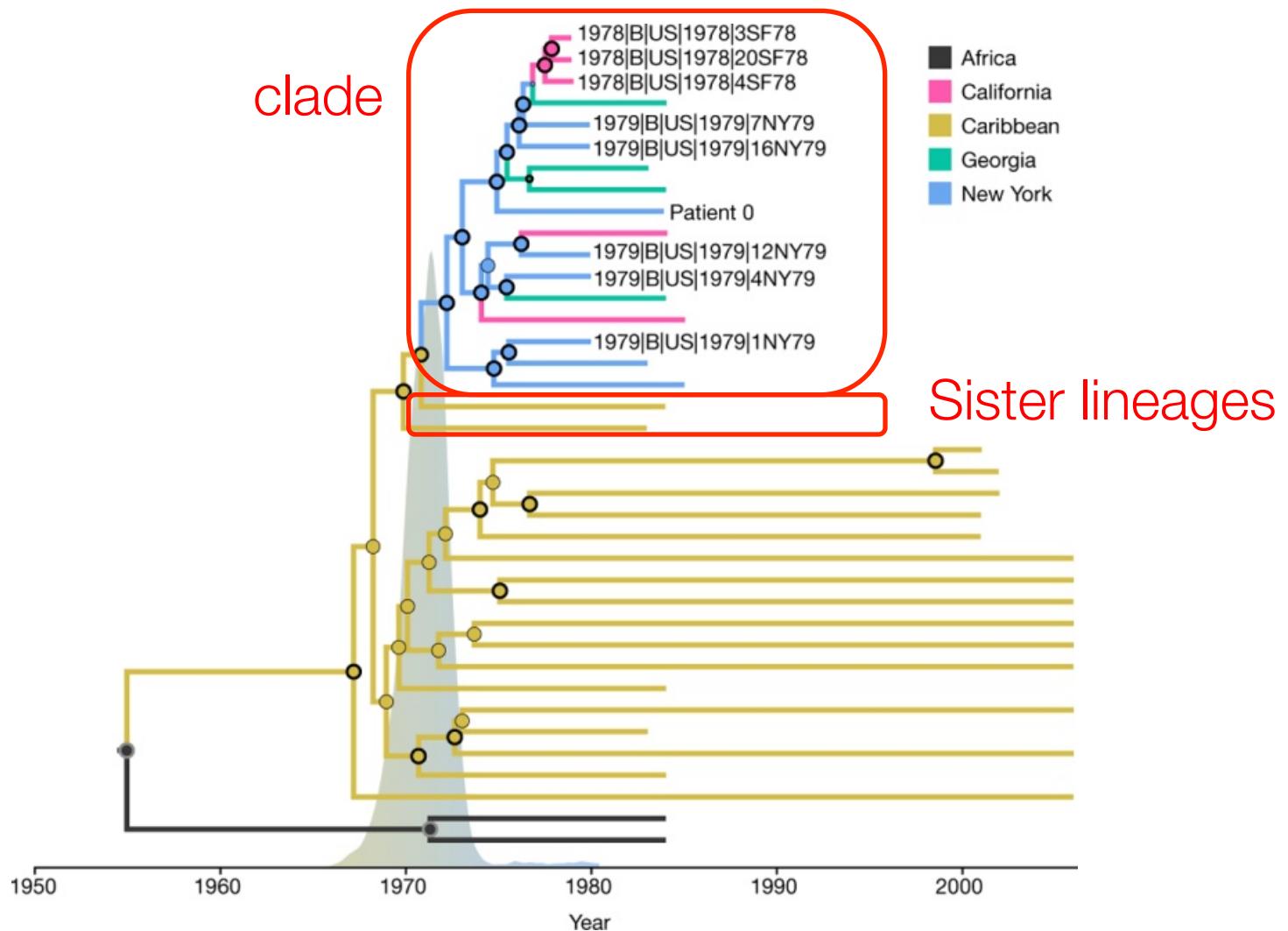
Phylogenetic trees



Phylogenetic trees

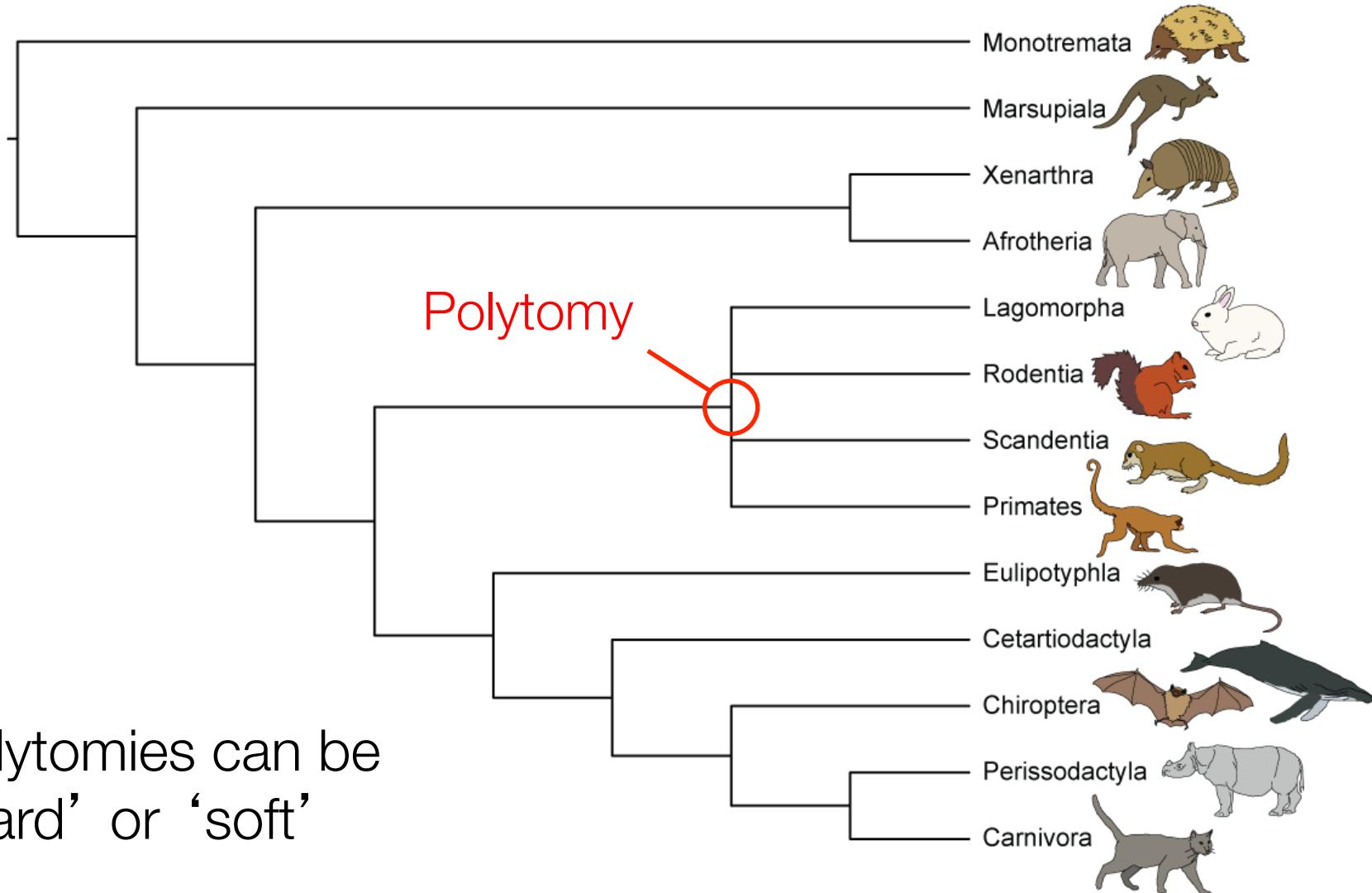


Phylogenetic trees

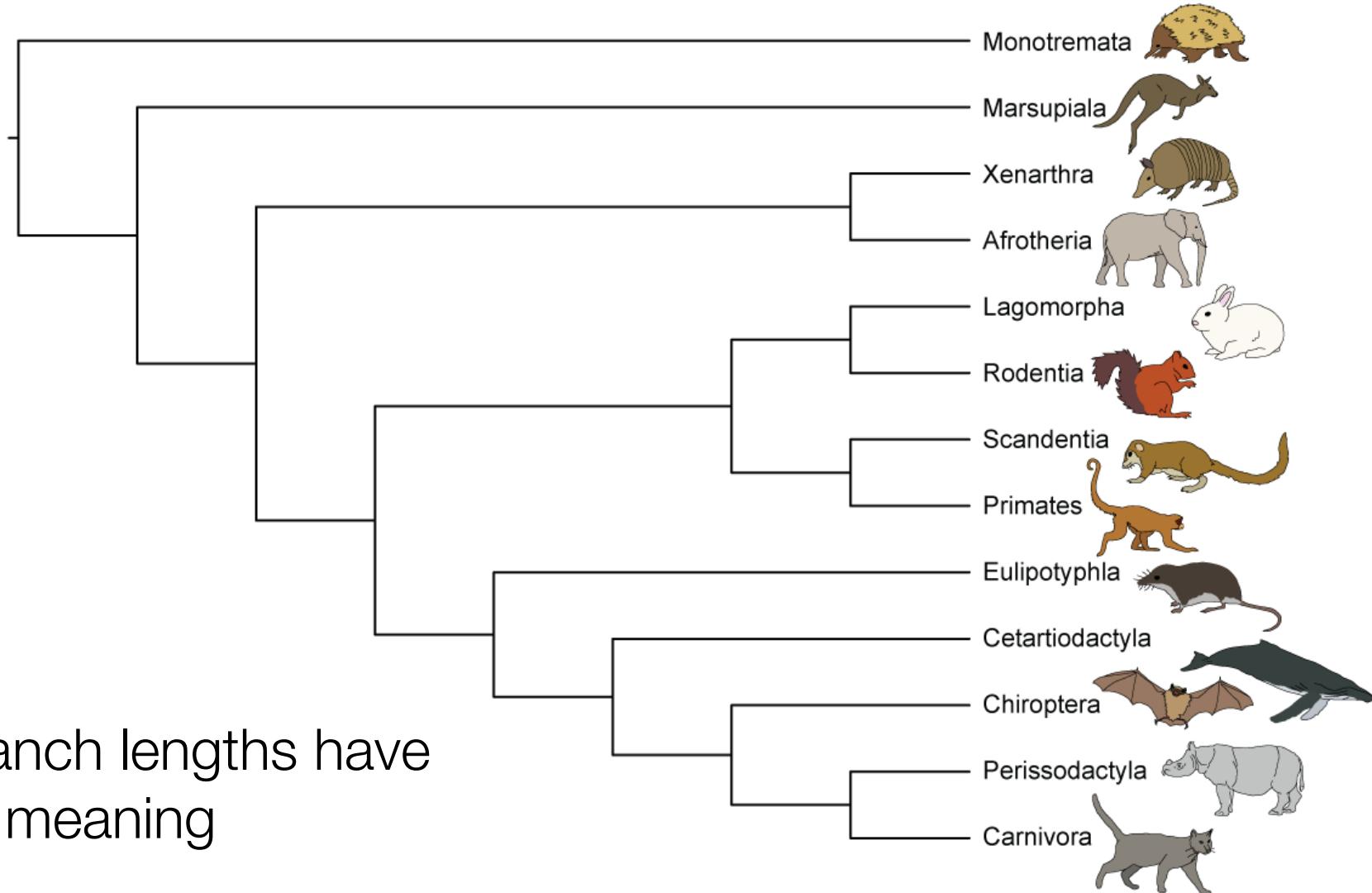


From Worobey et al. 2016 *Nature*

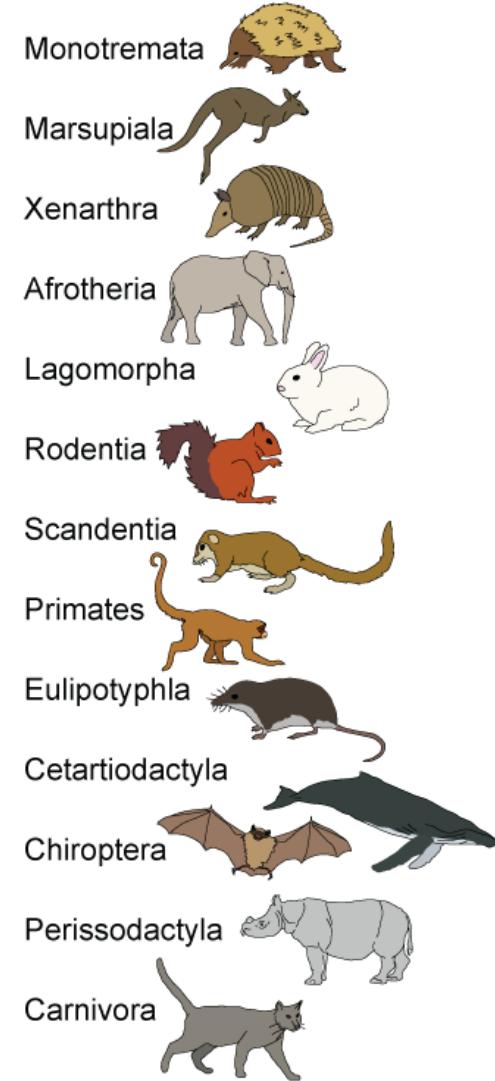
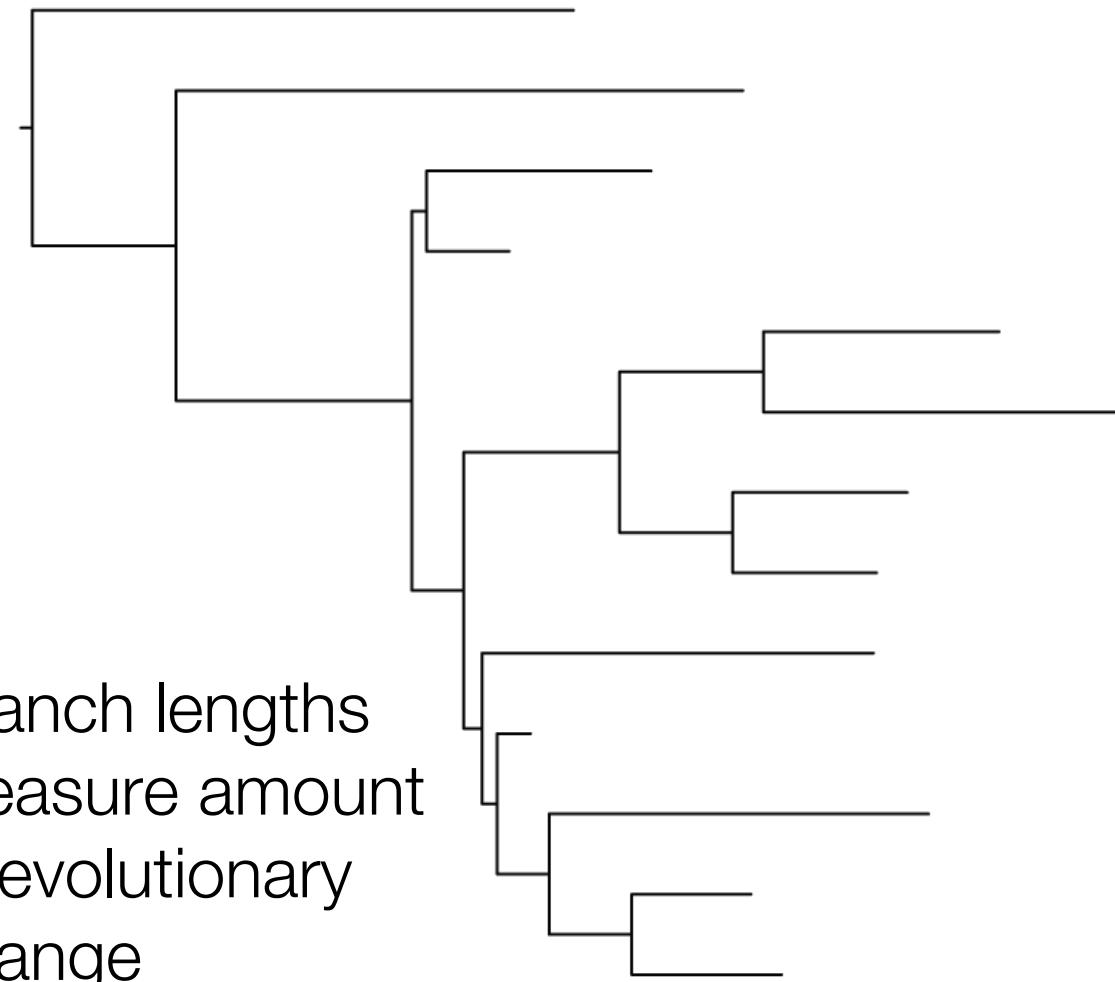
Phylogenetic trees



Phylogenetic trees: Cladogram

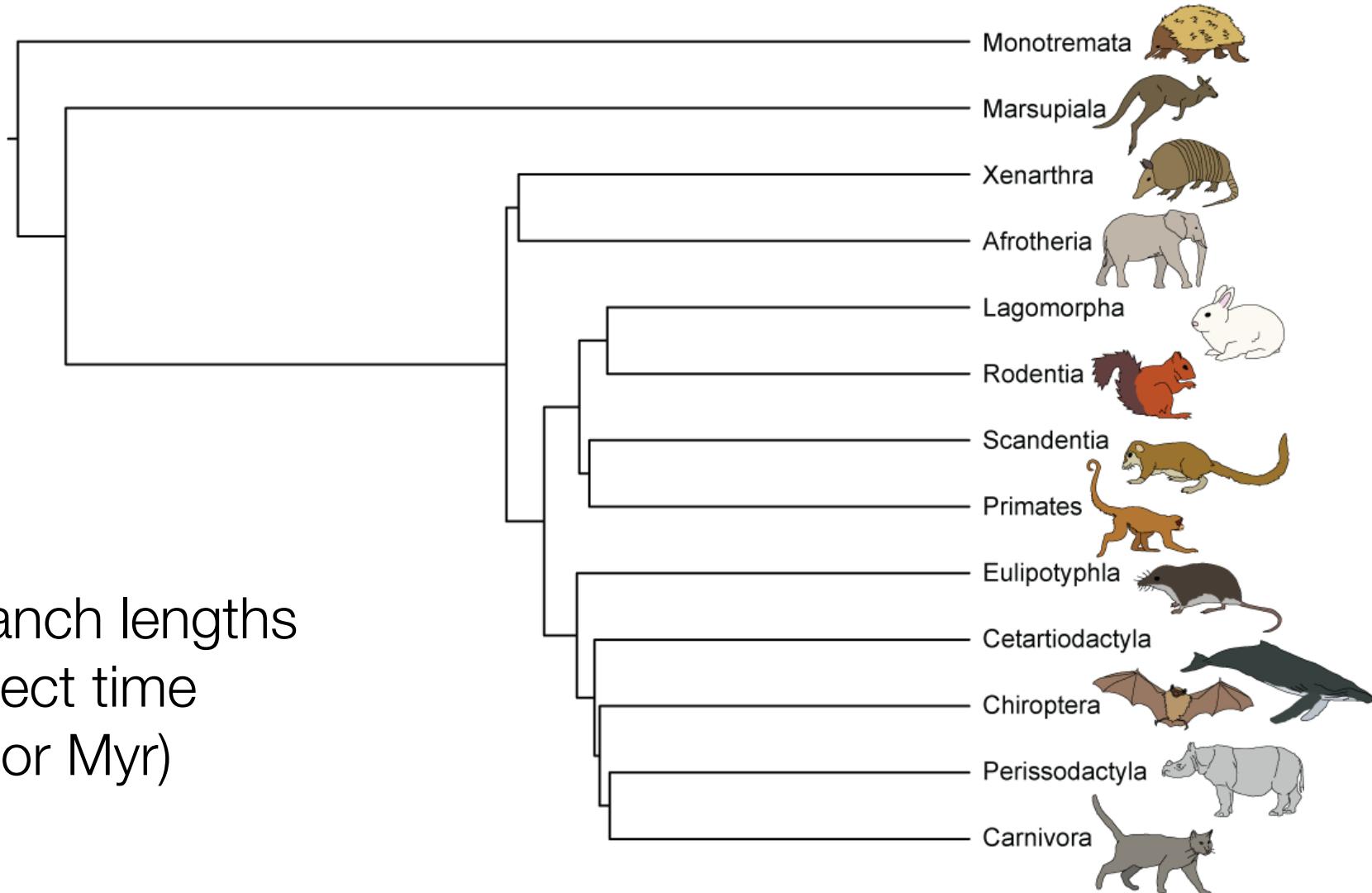


Phylogenetic trees: Phylogram



Phylogenetic trees: Chronogram

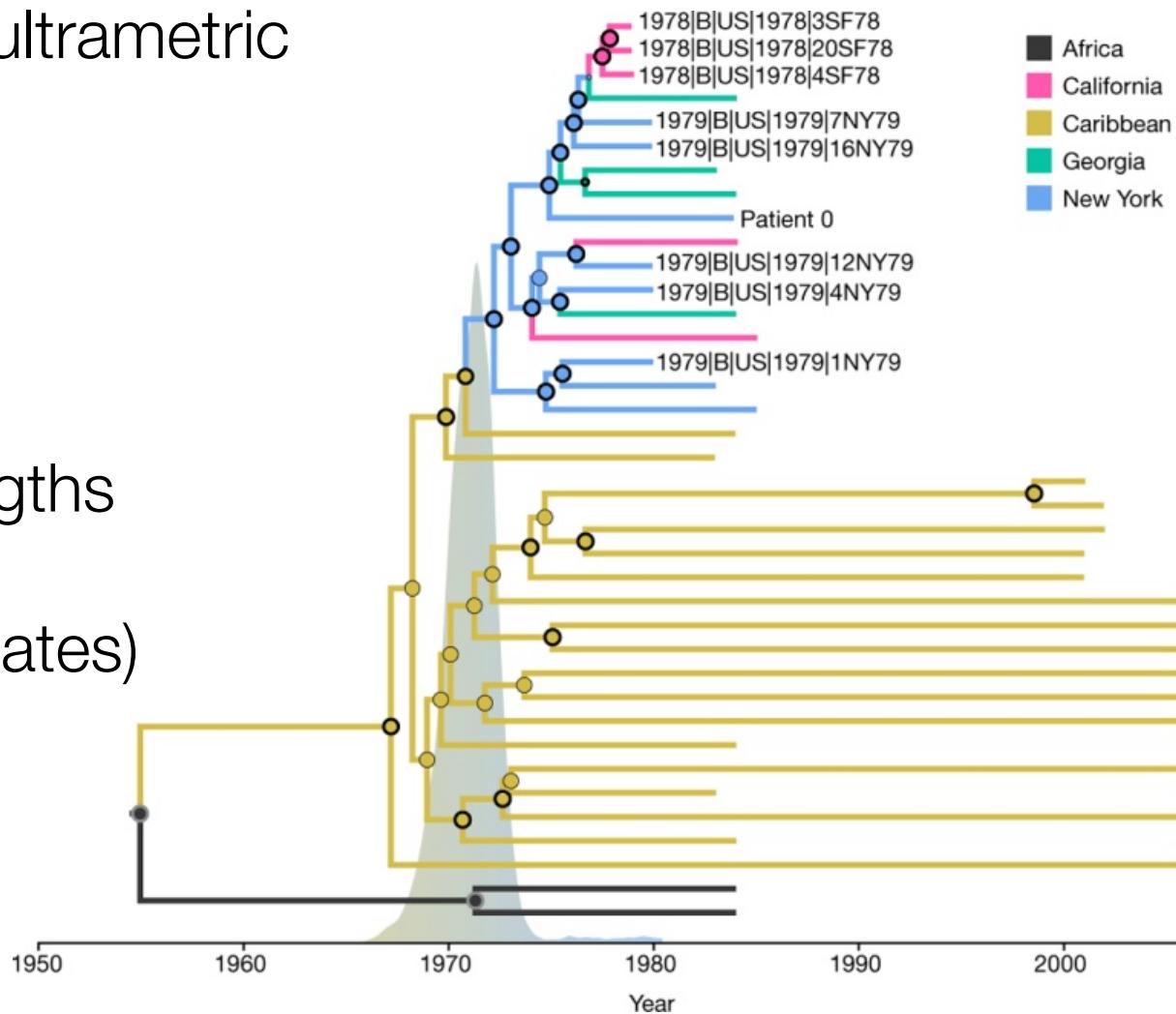
Ultrametric



Phylogenetic trees: Chronograms

Non-ultrametric

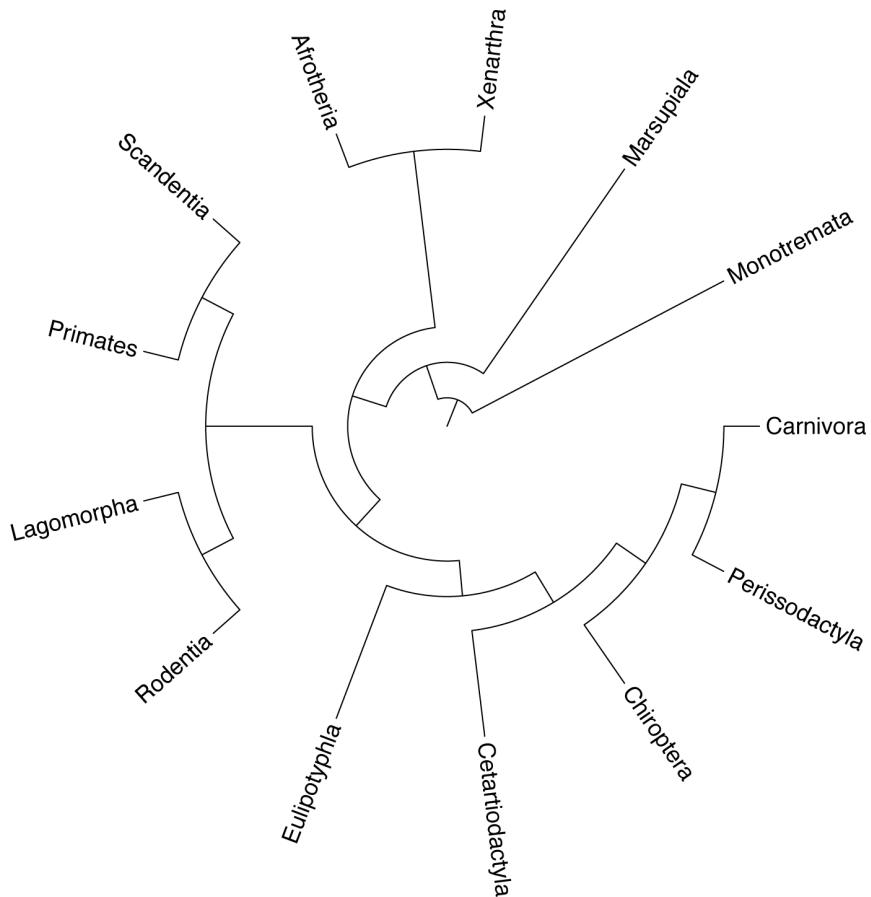
Branch lengths
reflect time
(calendar dates)



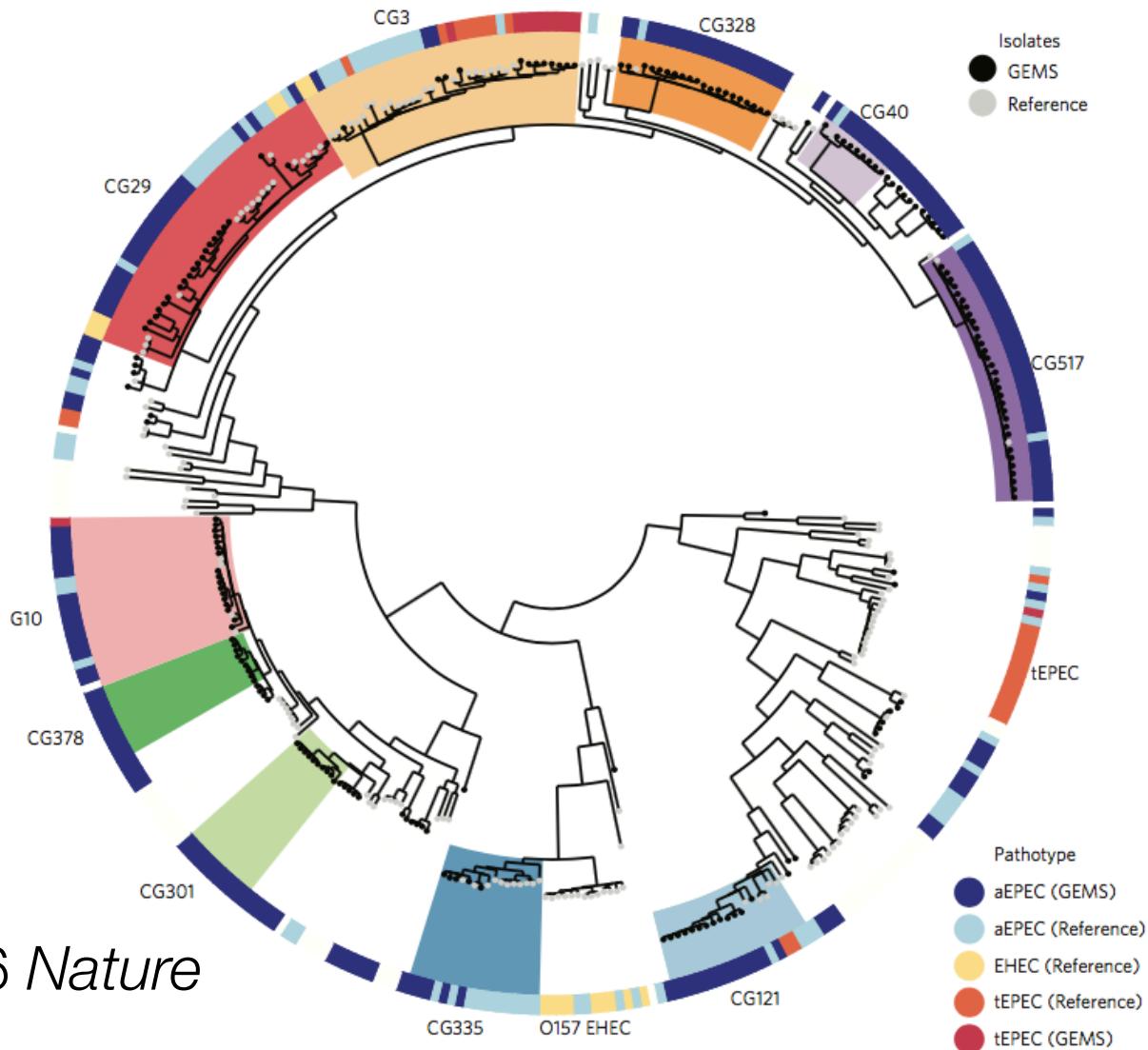
From Worobey et al. 2016 *Nature*

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



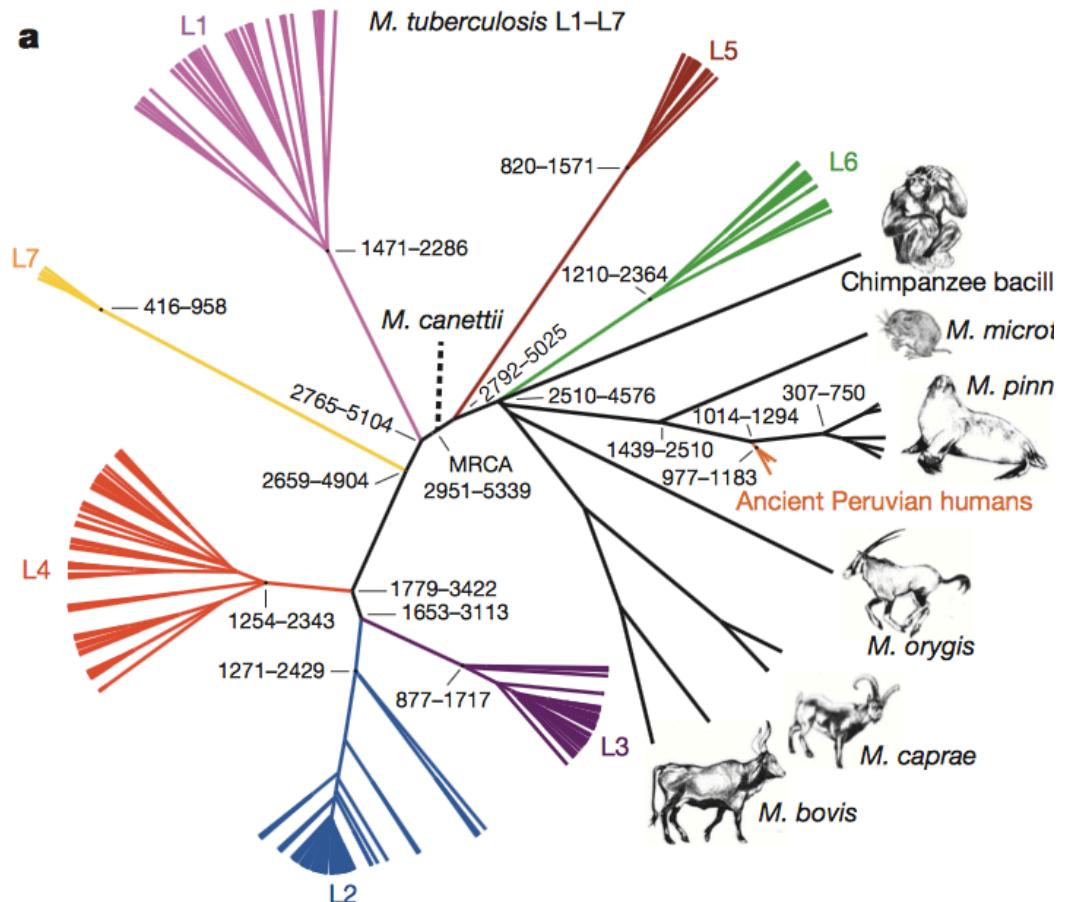
Phylogenetic trees: Circular



From Ingle et al. 2016 *Nature Microbiology*

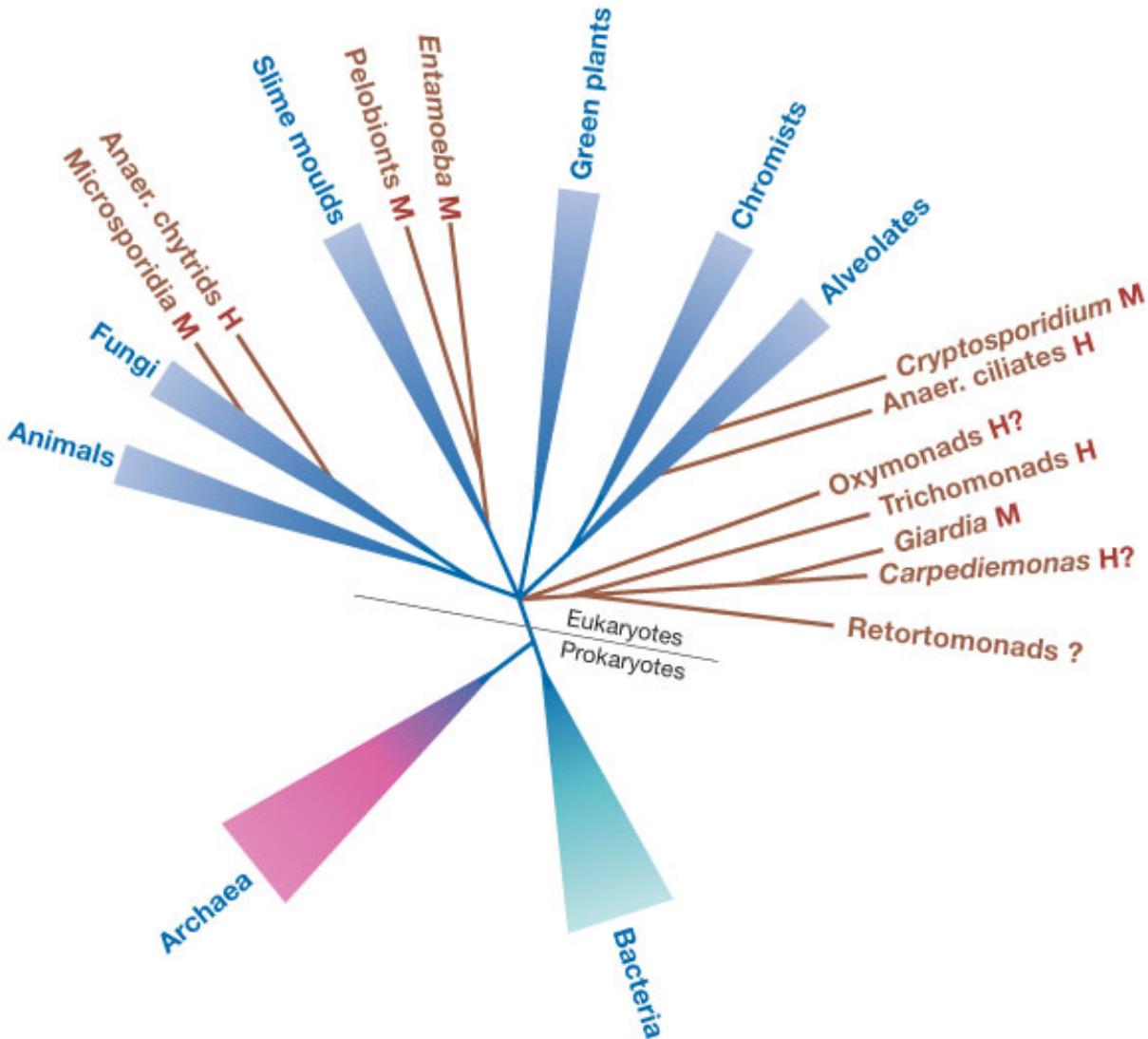
Phylogenetic trees: Unrooted

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



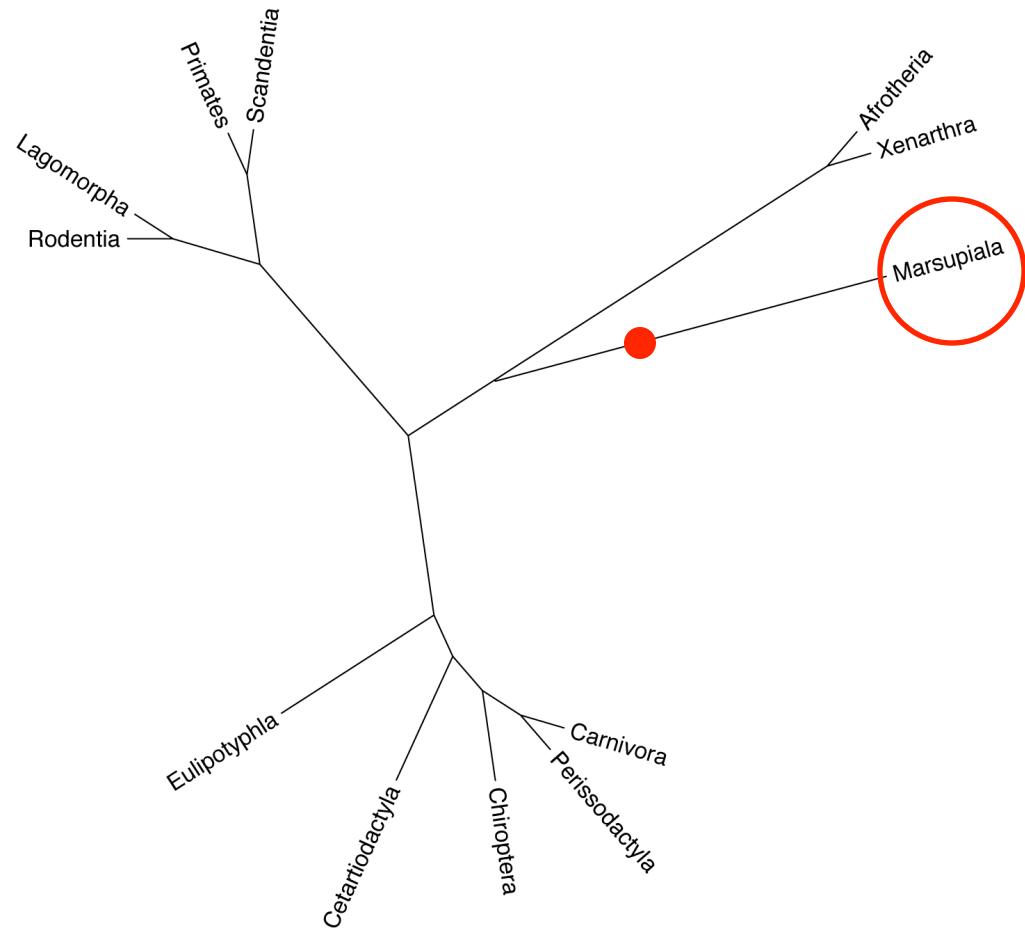
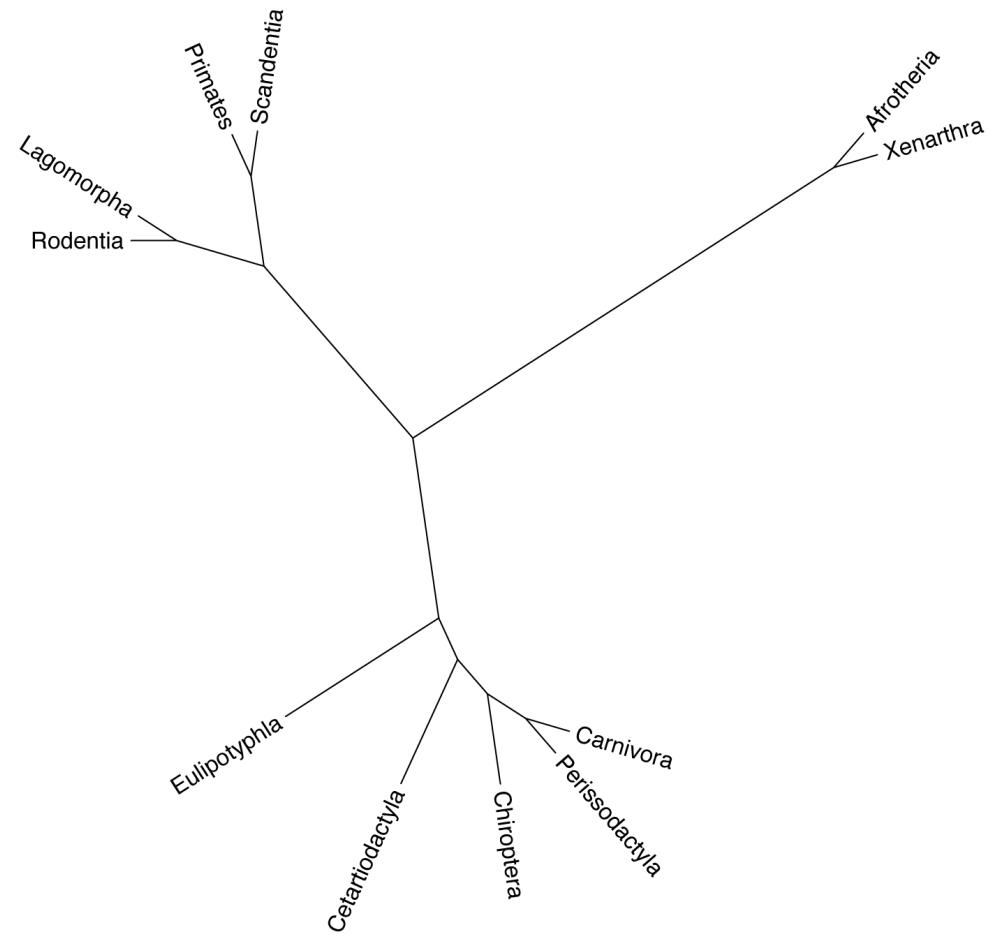
From Bos et al. 2014 *Nature*

Phylogenetic trees: Unrooted

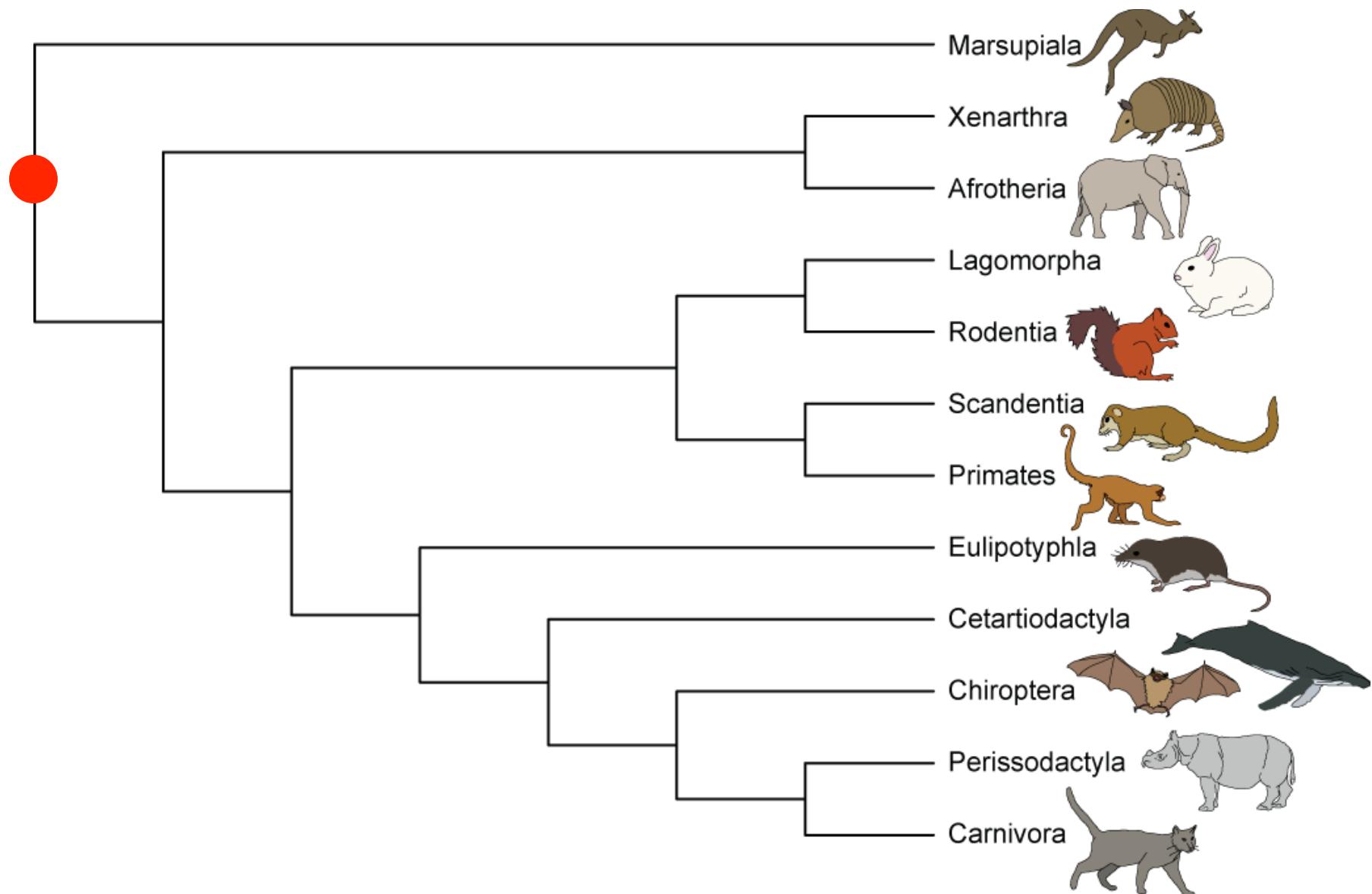


Rooting

- Can root a tree by including an outgroup taxon



Rooting



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

CGATAAGTTCACT

gray whale

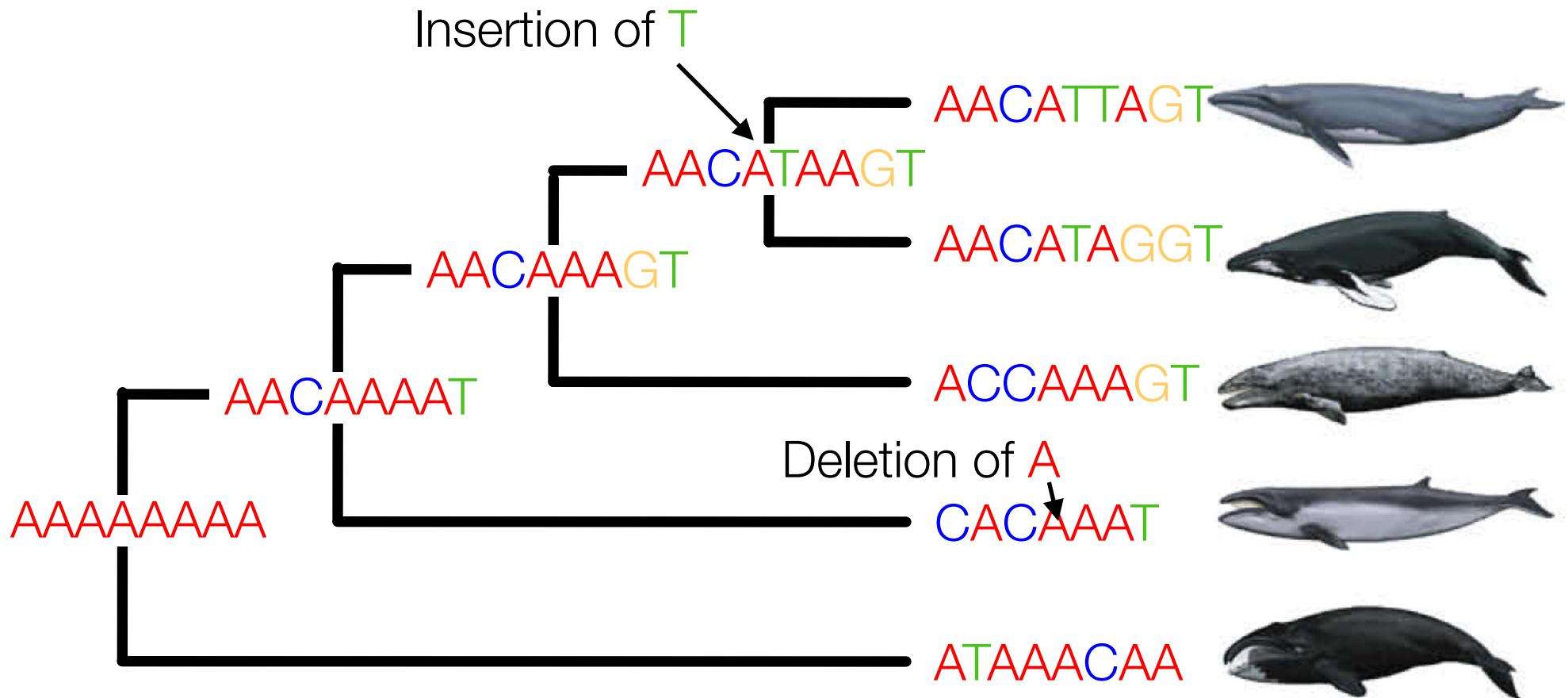
CGTTAGTTTACC

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

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



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

The diagram shows five DNA sequences aligned vertically. The sequences are color-coded: red for A, green for C, blue for G, and yellow for T. A vertical black line is drawn through the sequences, highlighting a specific position where all four bases (A, C, G, T) appear in all five sequences. This indicates a highly conserved homologous site.

AACATTA^GT

AACATA^GGT

ACCA-~~AAG~~T

CACA--~~A~~A^T

ATAA-~~A~~CAA

DNA sequence alignment

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

AACATTAGT
AACATAAGT
ACCA-AAGT
CACAA--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
AACATAAGGT
ACCA-AAGT
CACAA--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

A vertical black line is positioned between the first and second sequences. The sequences are color-coded: red for A, green for C, blue for G, and yellow for T.

AACAT	TTAGT
AACAT	AGGT
ACCA	AAAGT
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