Lecture: Evolutionary biology (reconstructing evolution using phylogenetics)

Sebastian Duchene
Australian Research Council Fellow
Peter Doherty Institute for Infection and Immunity
University of Melbourne

THE UNIVERSITY OF SYDNEY

2015 - PhD Computational Biology
2016 - Postdoc (virus phylogenetics)

THE UNIVERSITY OF MELBOURNE

2016 - Bio21 Institute McKenzie Fellow
(bacterial genomics)

2019 - Doherty Institute
Australian Research Council
DECRA Fellow

Some contributors:

@unimelb

Ashleigh Porter
Wytamma Wirth
Leo Featherstone

Previous workshops:

Simon Ho (Sydney Uni)
Rob Lanfear (ANU)
Matt Phillips (QUT)

For more in-depth phylogenetics:

Taming the beast
(taming-the-beast.org/)

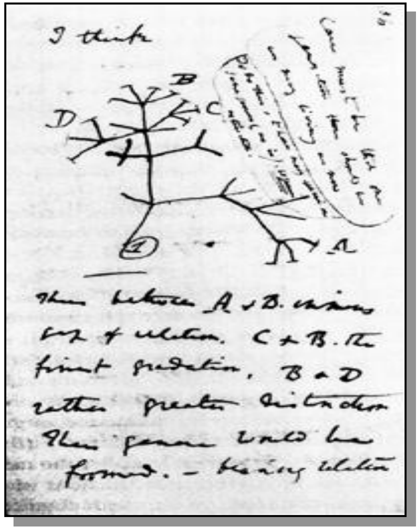
Sydney Phylogenetics workshop
(meep.sydney.edu.au/people/)

Melbourne pathogen
phylodynamics workshop
(TBA. email: sduchene@unimelb.edu.au)

Interpreting phylogenetic trees

What is a phylogenetic tree?

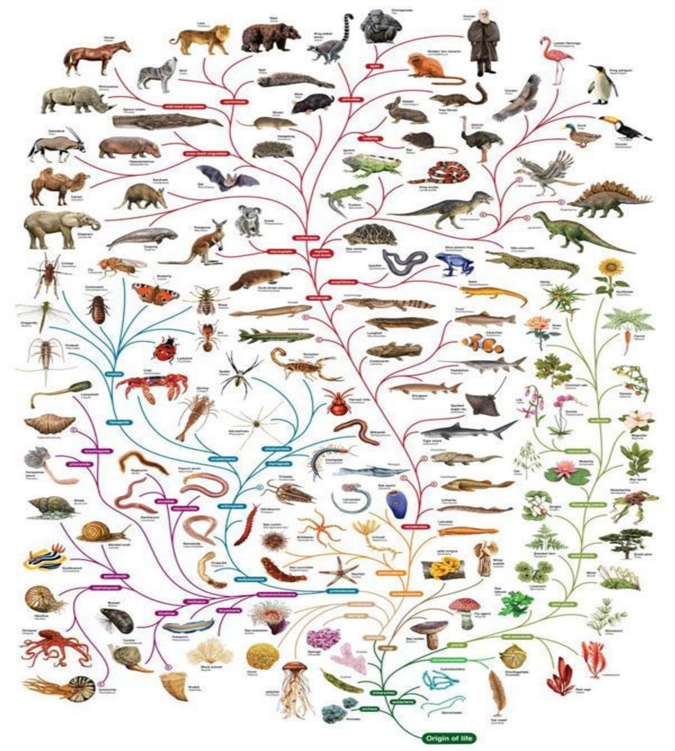
The phylogeny refers to the **true evolutionary relationships** among a set of organisms



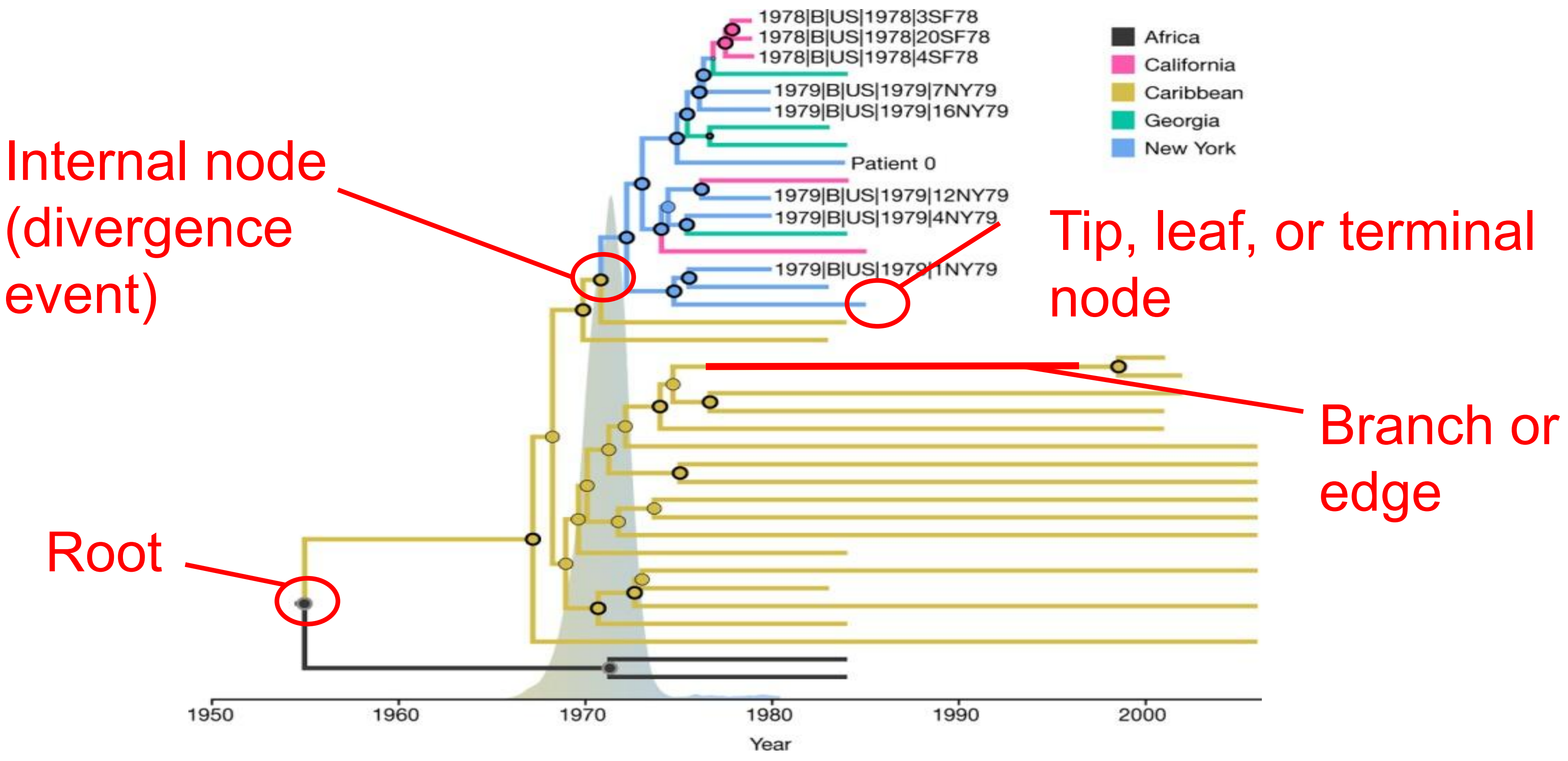
I think

Case must be that one generation [illegible] as many living as now [illegible]

Thus between A & B immense gap of relation. C & B the finest gradation, B & D rather greater distinction. Thus genera would be formed. — bearing relation

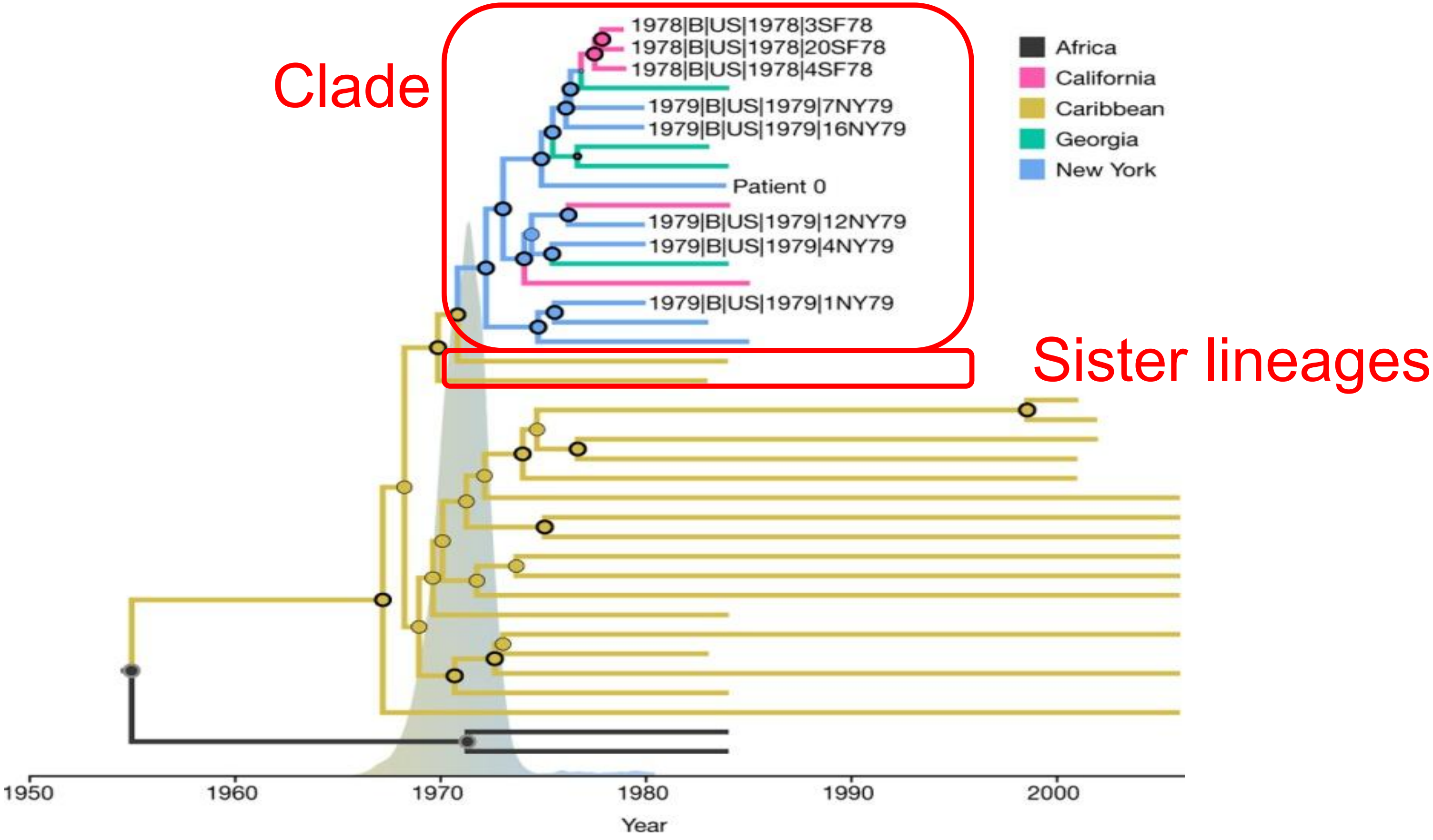


Phylogenetic trees



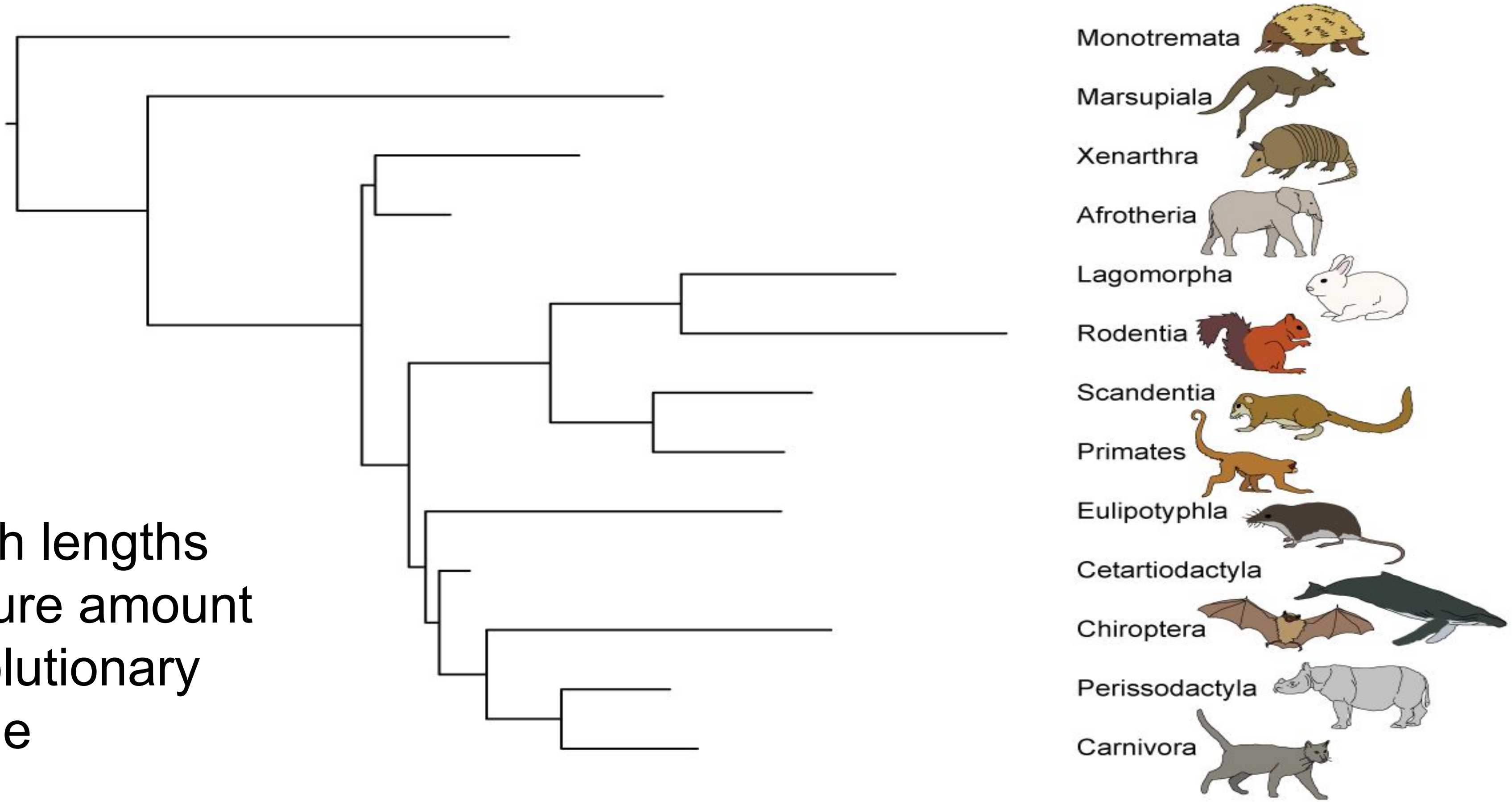
From Worobey et al. 2016 *Nature*

Phylogenetic trees



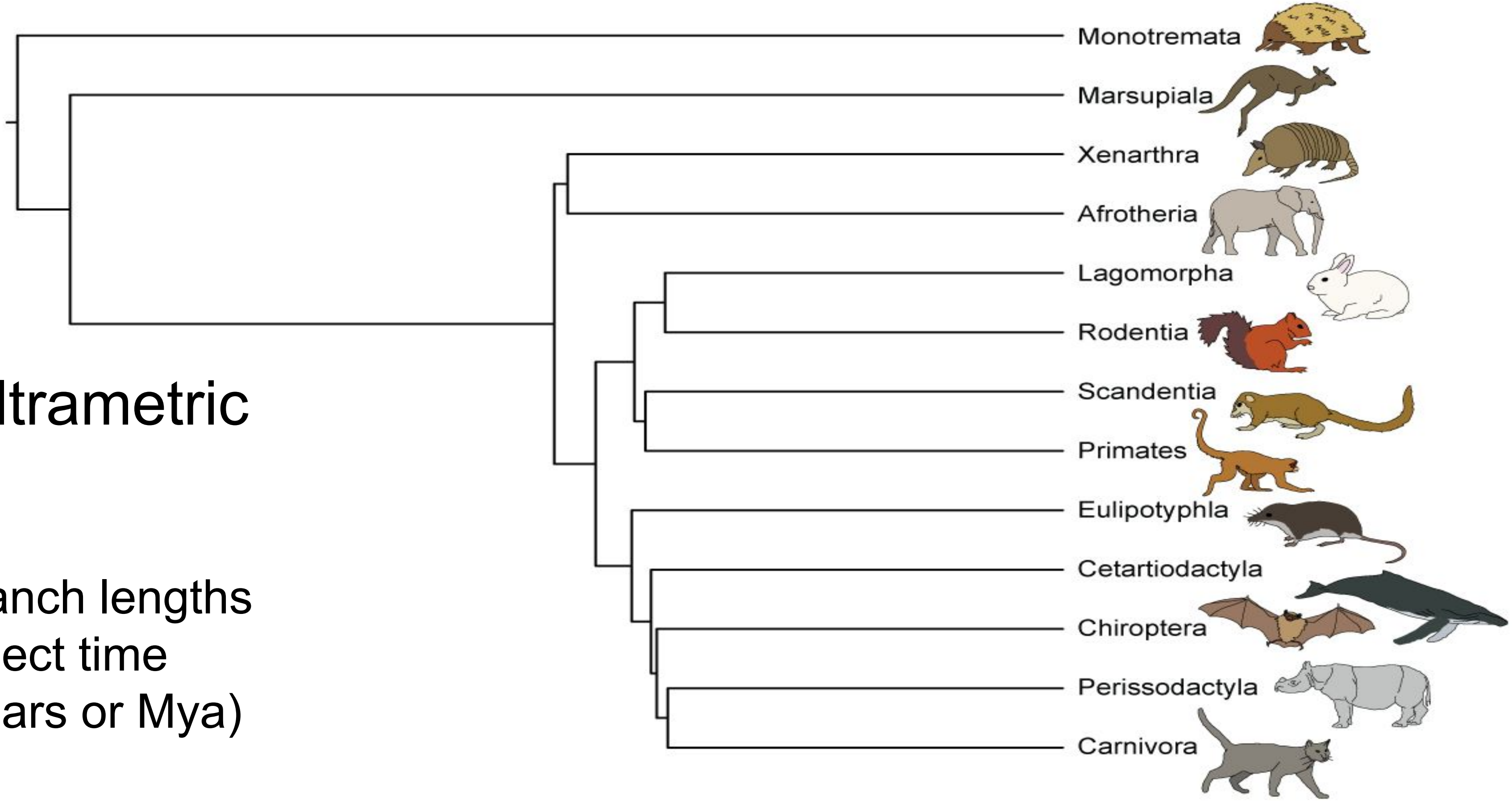
From Worobey et al. 2016 *Nature*

Phylogenetic trees: Phylogram



Branch lengths measure amount of evolutionary change

Phylogenetic trees: Chronograms



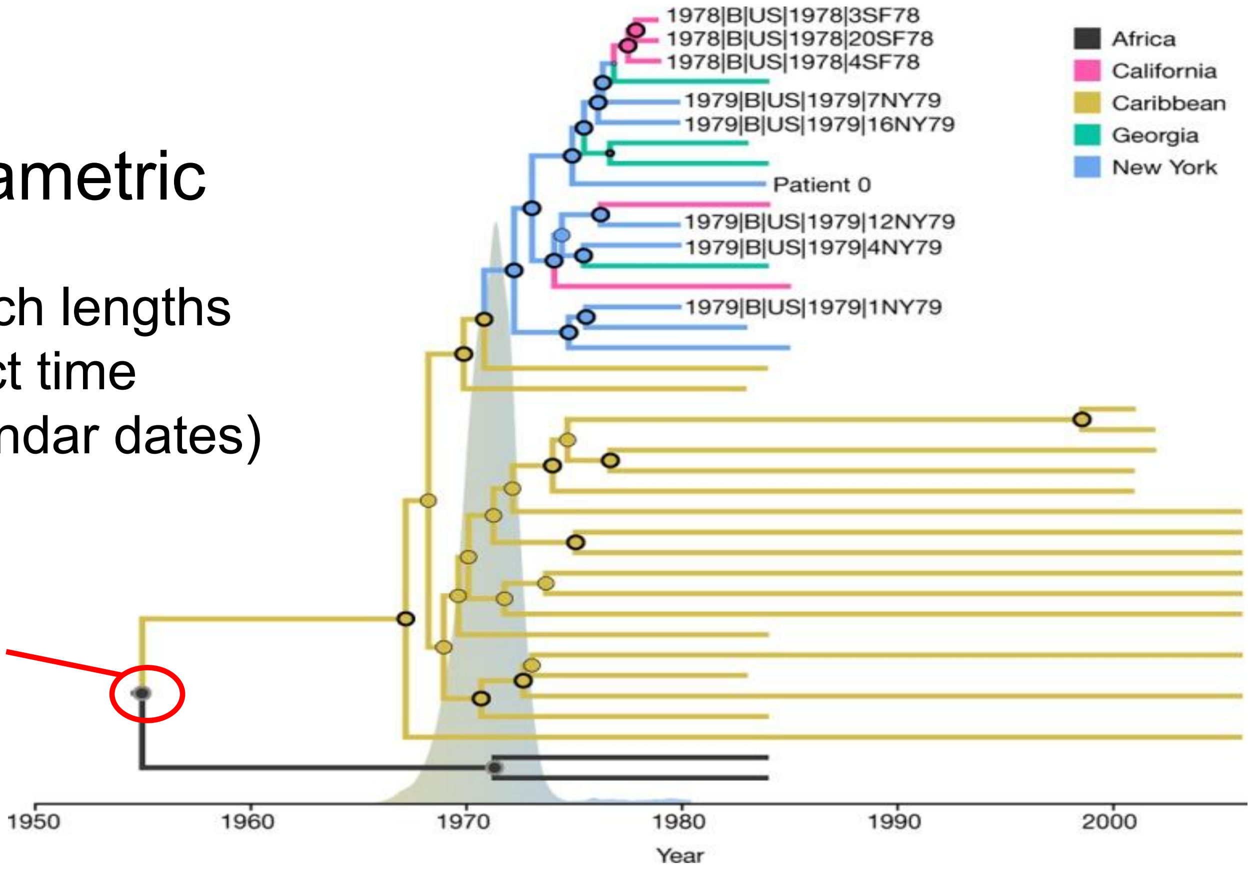
Ultrametric

Branch lengths
reflect time
(years or Mya)

Phylogenetic trees: Chronograms

Non-ultrametric

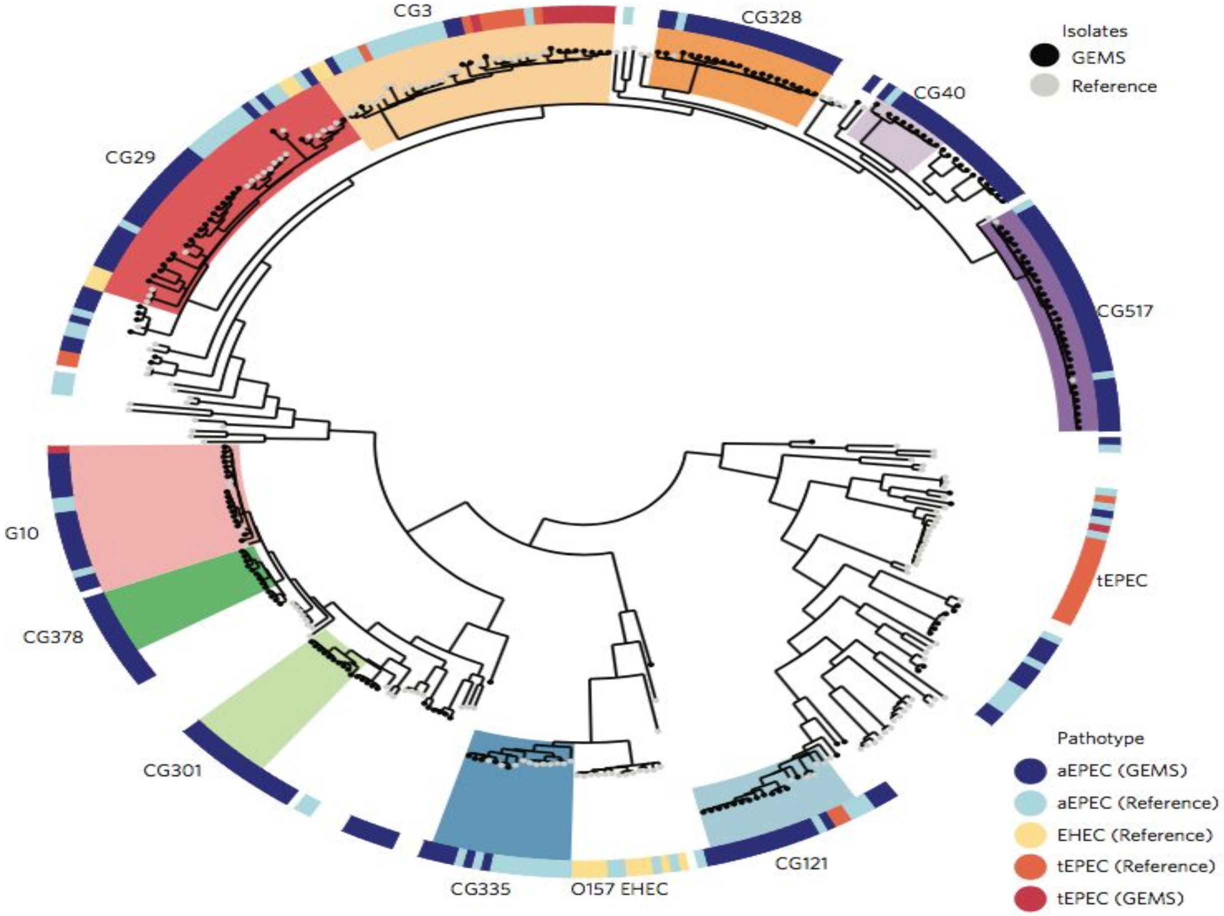
Branch lengths reflect time (calendar dates)



Time to the most recent common ancestor

Phylogenetic trees: Circular

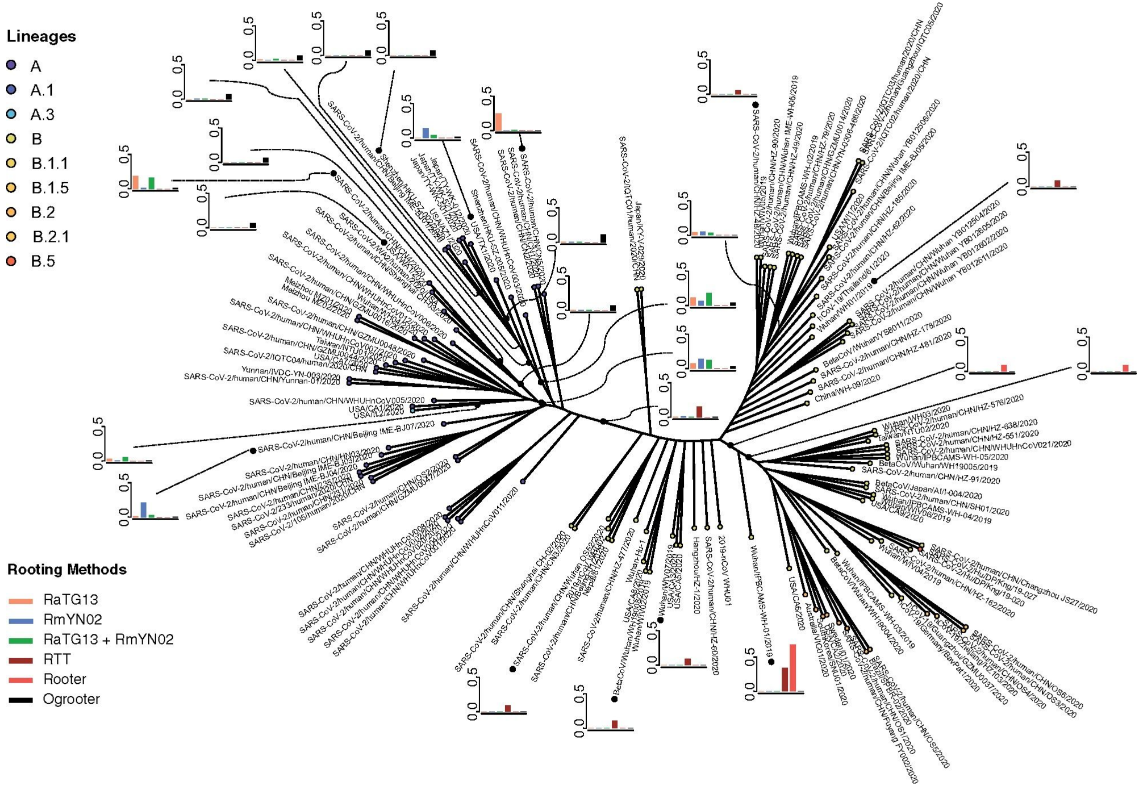
- Good for very large trees
- Branches can be genetic distance or time



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 Pipes et al. 2020 Mol Bio Evol

Concept review

Parts of trees:

- root node
- internal nodes (divergence events)
- tips
- branches or ‘edges’

Types of trees:

- phylograms (branches usually in subs/site)
- chronograms (branches in units of time)
- *cladograms (branches have no meaning)

Inferring phylogenetic trees

Inferring phylogenetic trees

1. Maximum parsimony
2. Distance-based methods
3. Maximum likelihood
4. Bayesian inference

Beast2
Bayesian evolutionary analysis by sampling trees

BEAST
Bayesian Evolutionary Analysis Sampling Trees

IQ-TREE

RAxML

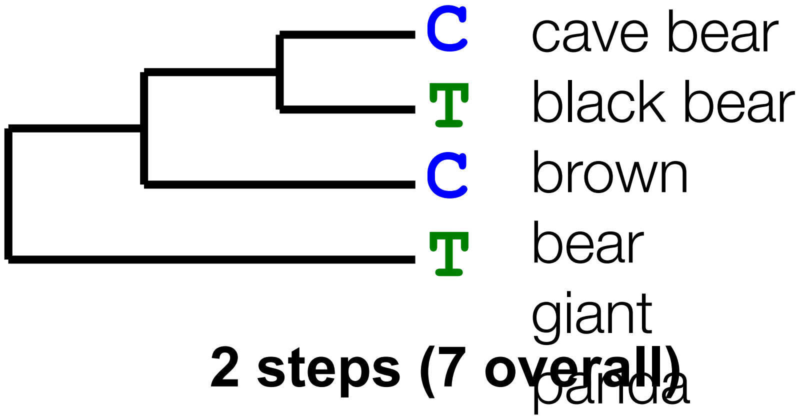
MEGA Molecular Evolutionary Genetics Analysis

Maximum parsimony

brown bear	CGTTAGTACACT
cave bear	CGATAGTTCACT
black bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

Maximum parsimony

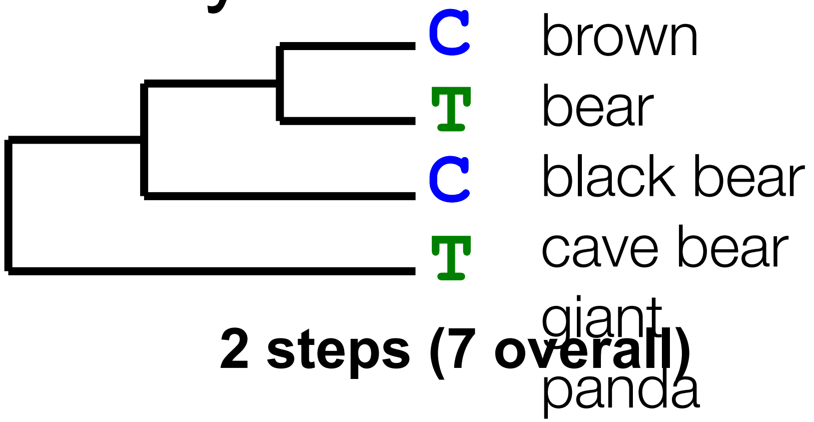
brown bear CGTTAGTACACT
cave bearCGATAGTTCACT
black bear CGTTAGTTTACC
giant panda CATTGGTTTACT



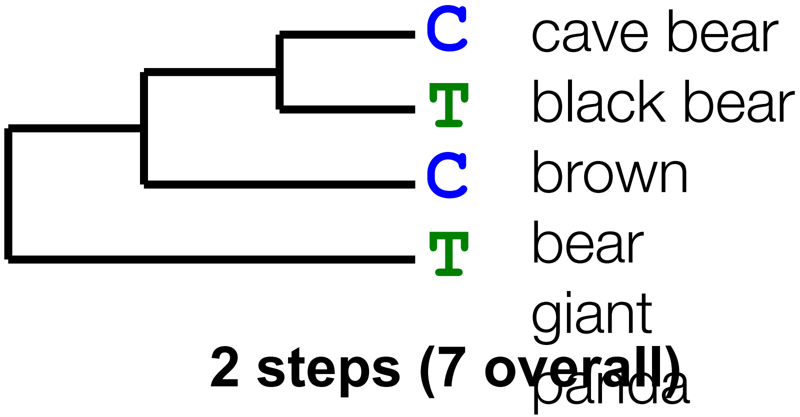
2 steps (7 overall)

Maximum parsimony

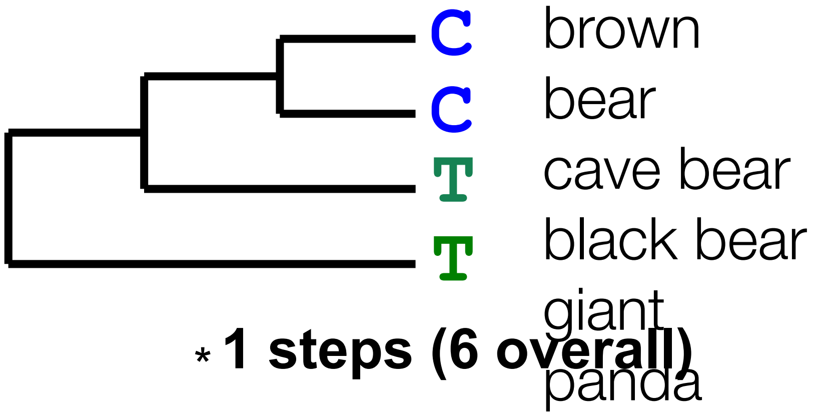
brown bear CGTTAGTACACT
cave bear CGATAGTTCACT
black bear CGTTAGTTTACC
giant panda CATTGGTTTACT



2 steps (7 overall)



2 steps (7 overall)



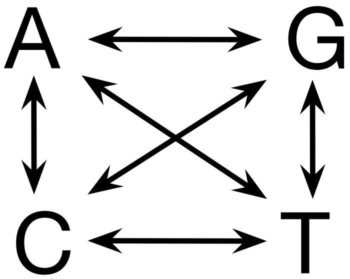
* 1 steps (6 overall)

Maximum parsimony

- Identifies the tree topology that can explain the sequence data, using the smallest number of inferred substitution events
- Commonly used for morphological data
- Now rarely used for analysing genetic data
  - Cannot estimate evolutionary rates or timescales
  - Effects of multiple substitutions

Evolutionary models

Rate Matrix



Base Frequencies

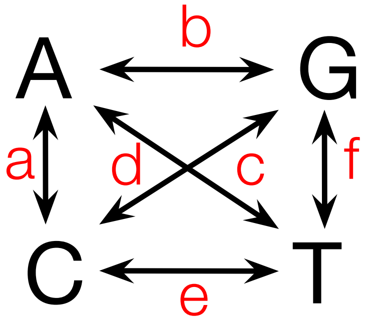
, $\pi_A + \pi_C + \pi_G + \pi_T =$
1,

Site Rates

+ I +
G

Evolutionary models

Rate Matrix



Base Frequencies

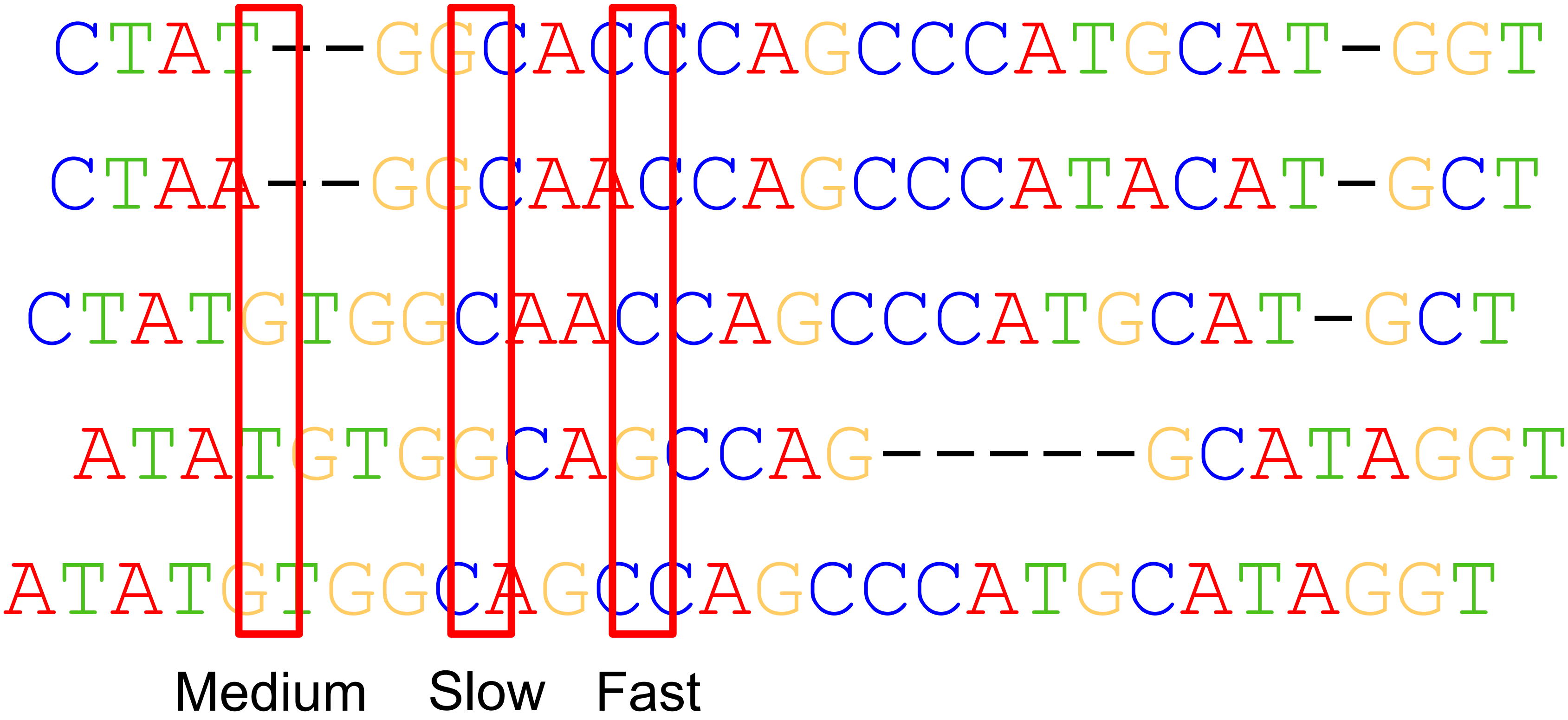
, $\pi_A + \pi_C + \pi_G + \pi_T =$ 1,

Site Rates

+ I + G

JC	**HKY**	**GTR**	**GTR+I+G**
a=b=c=d=e=f	a=c=d=f, b=e	a, b, c, d, e, f	a, b, c, d, e, f
$\pi_A=\pi_C=\pi_G=\pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$
No I or G	No I or G	No I or G	I, G
0 free parameters	4 free parameters	8 free parameters	10 free parameters

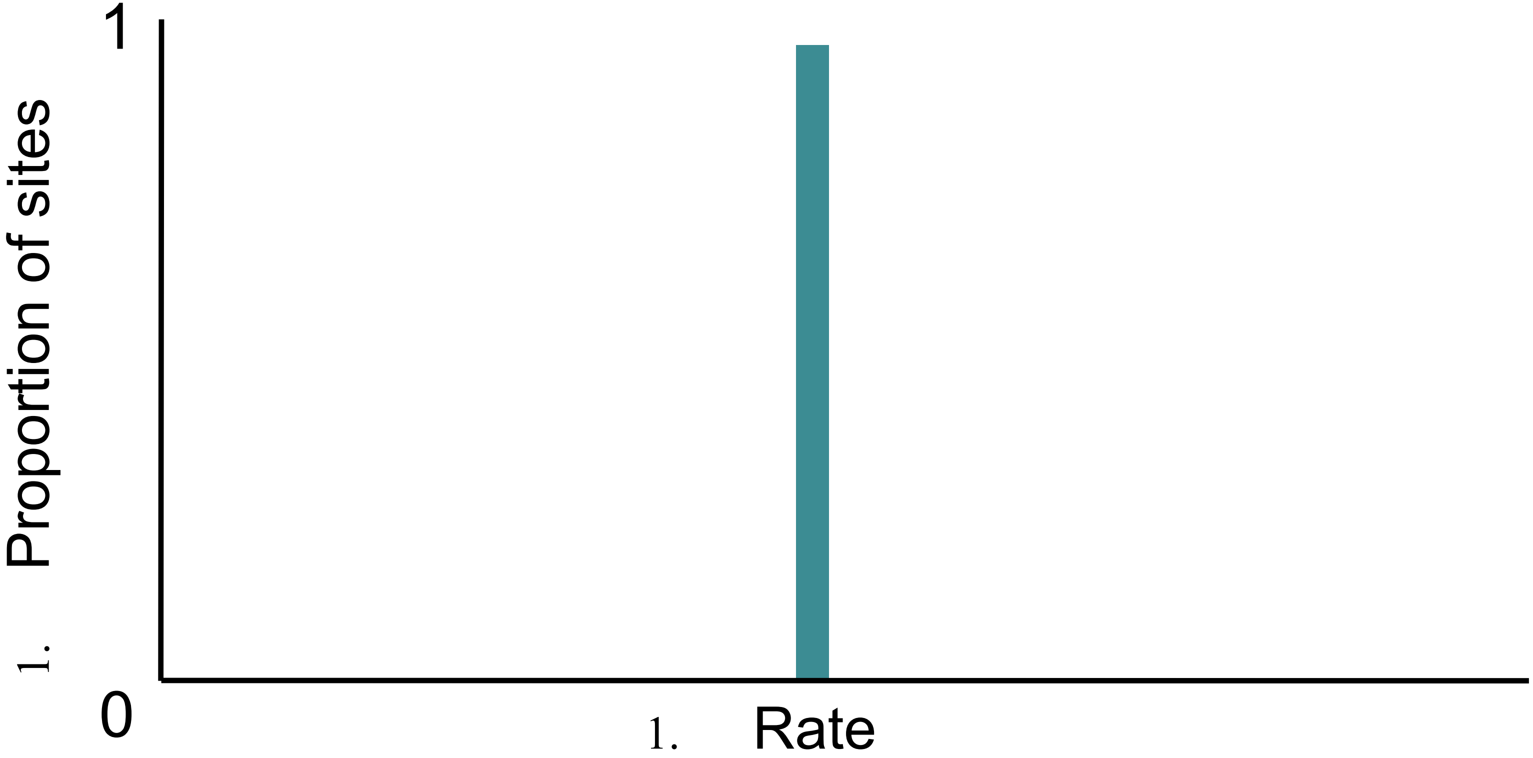
Rate variation among sites



```
CTAT--GGCACCCAGCCCATGCAT-GGT
CTAA--GGCAACCAGCCCATACAT-GCT
CTATGTGGCAACCAGCCCATGCAT-GCT
ATATGTGGCAGCCAG-----GCATAGGT
ATATGTGGCAGCCAGCCCATGCATAGGT
```

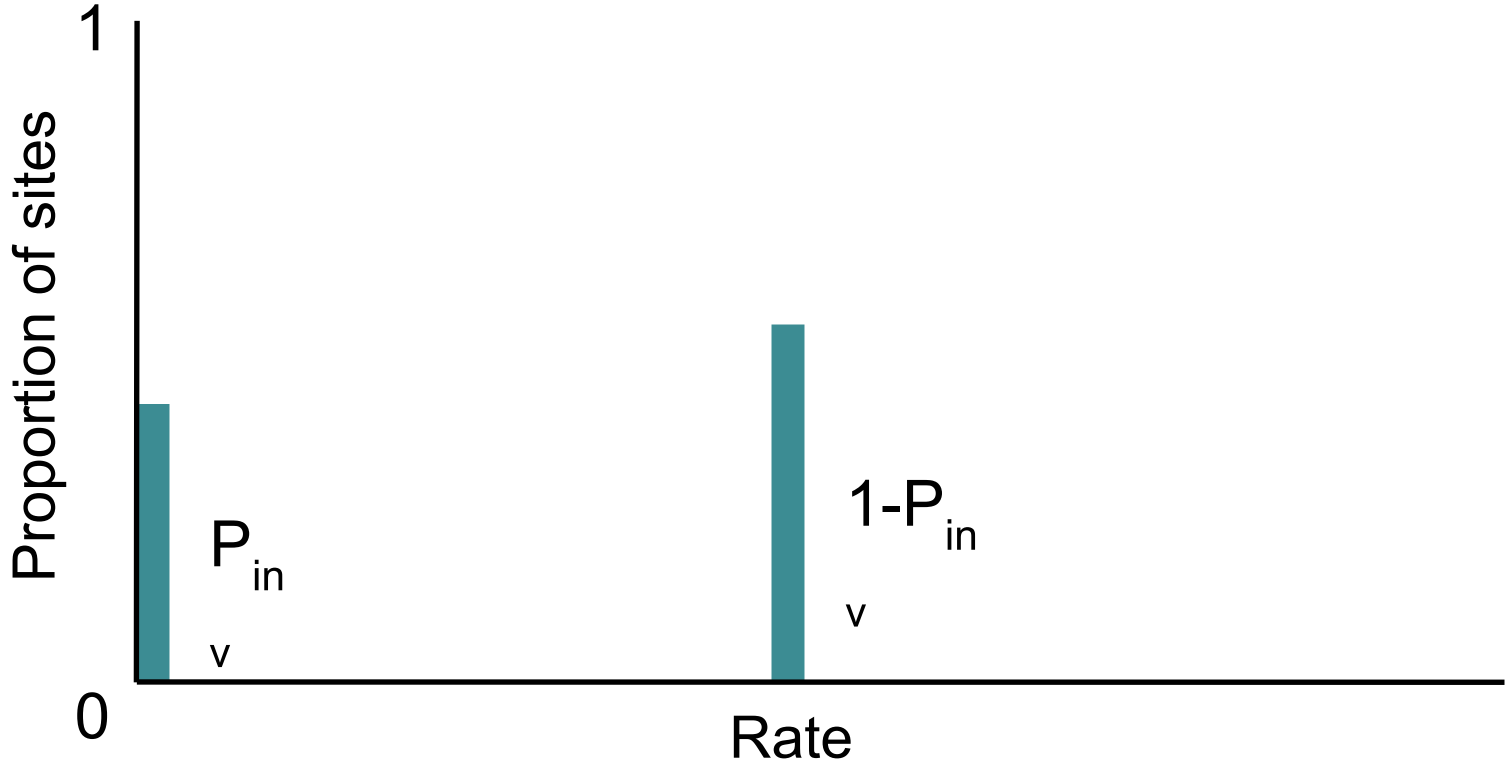
Rate variation among sites

1. Equal rates among sites (e.g., **JC**, **GTR**, **HKY** models)



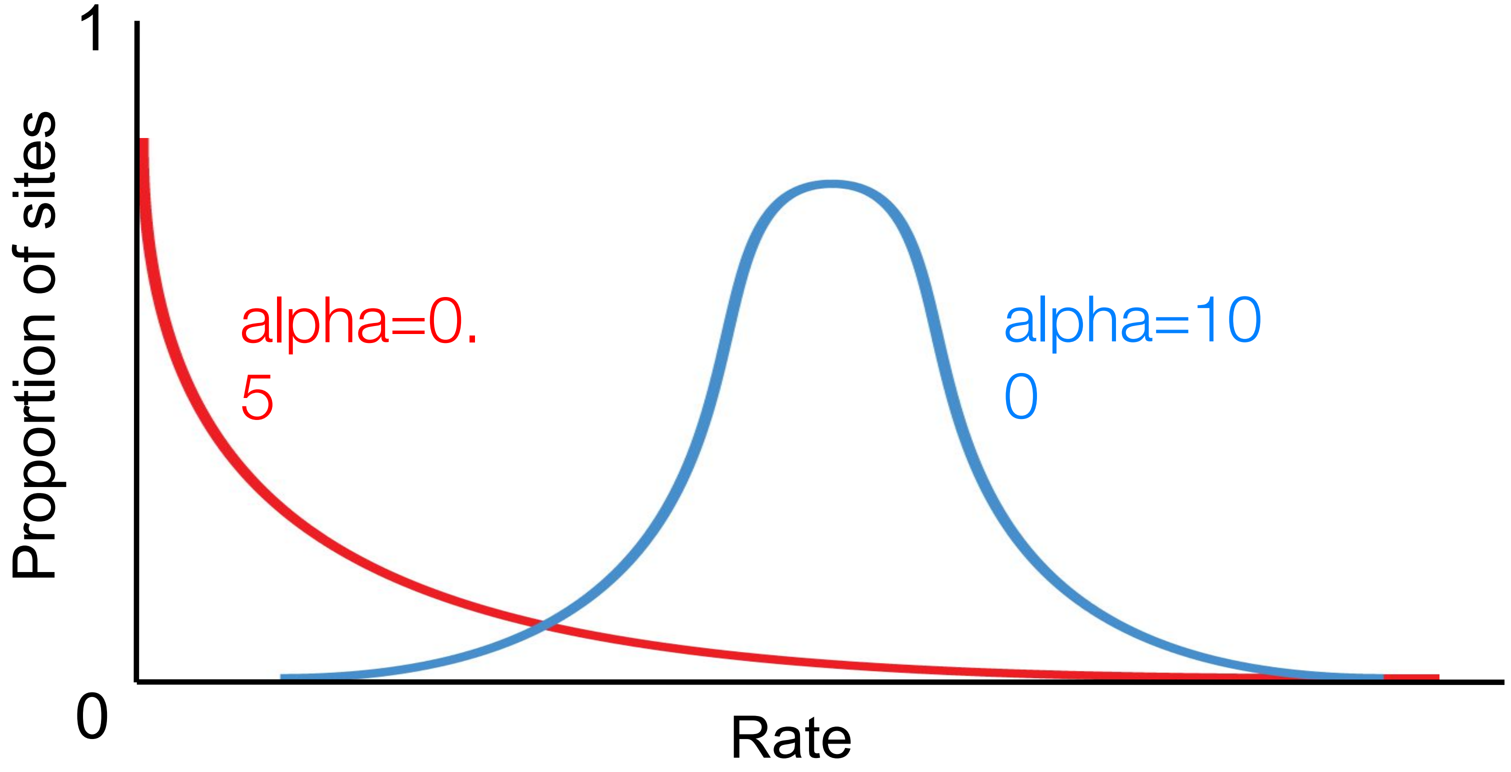
Rate variation among sites

- Proportion of invariable sites (e.g., **JC+I**, **GTR+I**, **HKY+I** models)



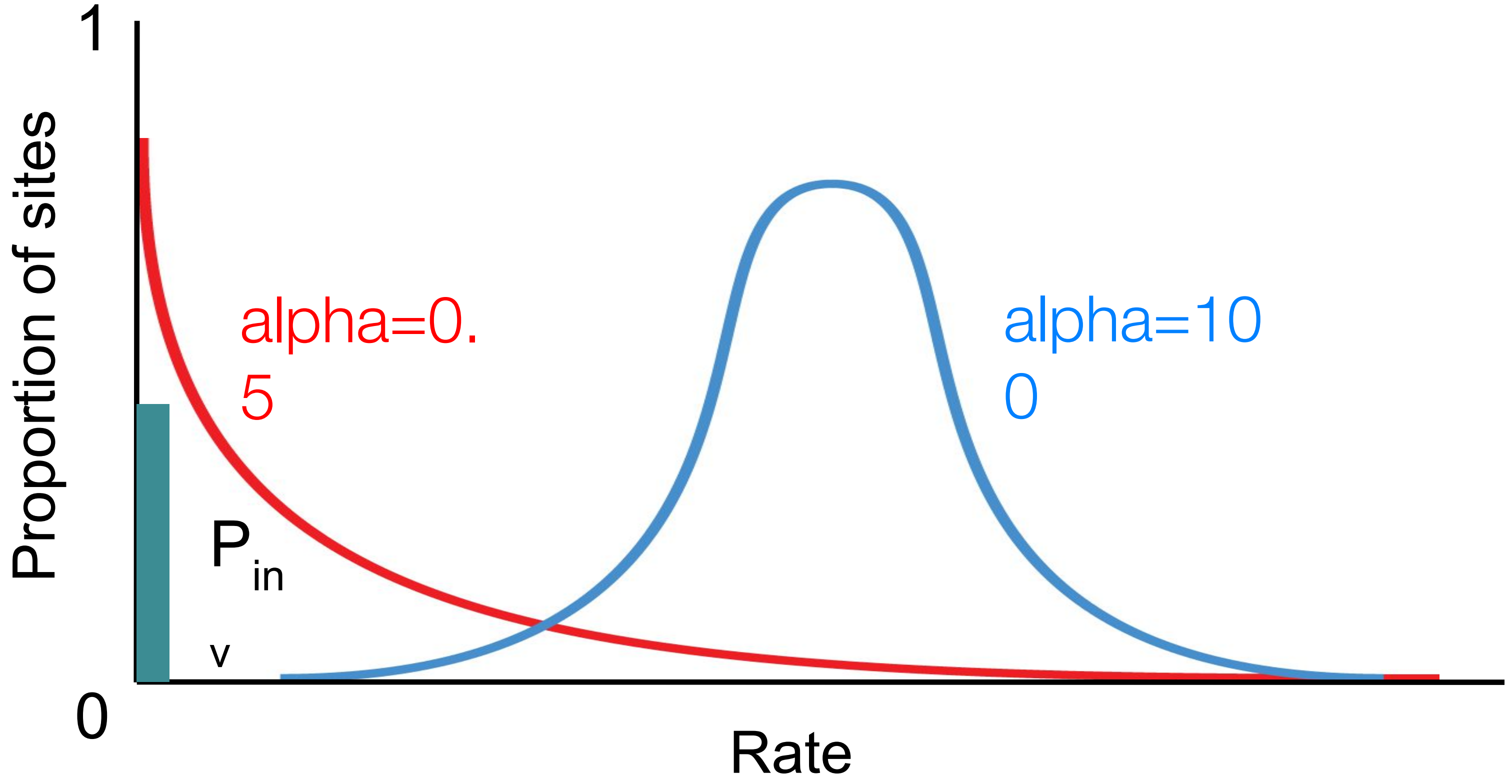
Rate variation among sites

- Gamma-distributed rate variation among sites (e.g., **JC+G**, **GTR+G**, **HKY+G** models)



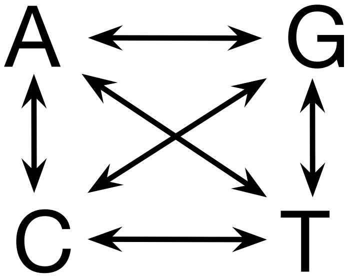
Rate variation among sites

- **JC+G+I**, **GTR+G+I**, **HKY+G+I** models



Evolutionary models

Rate Matrix



Base Frequencies

$$\pi_A + \pi_C + \pi_G + \pi_T = 1$$

Site Rates

+ I + G

#Models **203 X 15 X 14 = 12,180**

In phylogenetics, we typically consider a small subset of these

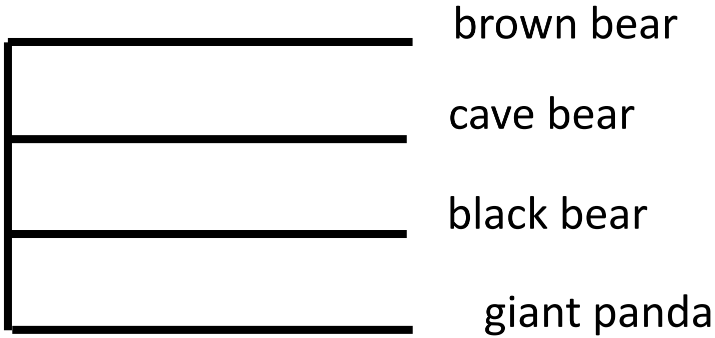
Neighbour-joining

CLUSTERING ALGORITHM

brown bear	CGTTAGTACACT
cave bear	CGATAGTTCACT
black bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

MODEL

	brown bear	cave bear	black bear	giant panda
brown bear	–			
cave bear	.1	–		
black bear	.3	.3	–	
giant panda	.4	.5	.4	–



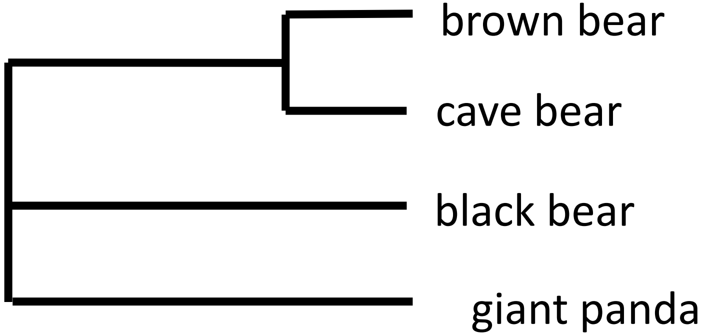
Neighbour-joining

brown bear	CGTTAGTACACT
cave bear	CGATAGTTCACT
black bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

MODEL

	brown bear	cave bear	black bear	giant panda
brown bear	–			
cave bear	.1	–		
black bear	.3	.3	–	
giant panda	.4	.5	.4	–

CLUSTERING ALGORITHM



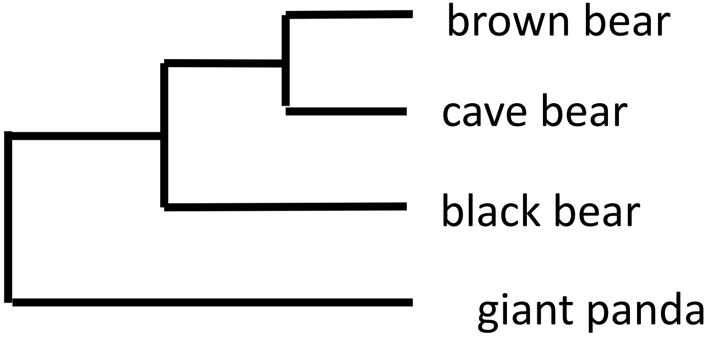
Neighbour-joining

brown bear	CGTTAGTACACT
cave bear	CGATAGTTCACT
black bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

MODEL

	brown bear	cave bear	black bear	giant panda
brown bear	–			
cave bear	.1	–		
black bear	.3	.3	–	
giant panda	.4	.5	.4	–

CLUSTERING ALGORITHM

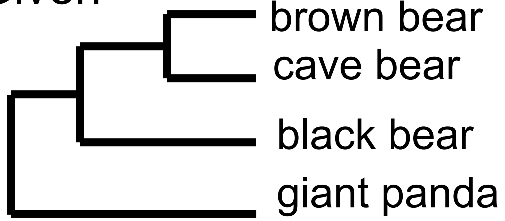


Maximum likelihood

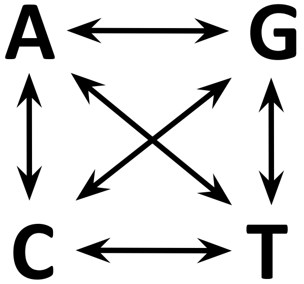
Likelihood of hypothesis *H* =

The probability of the data, given the hypothesis

Given



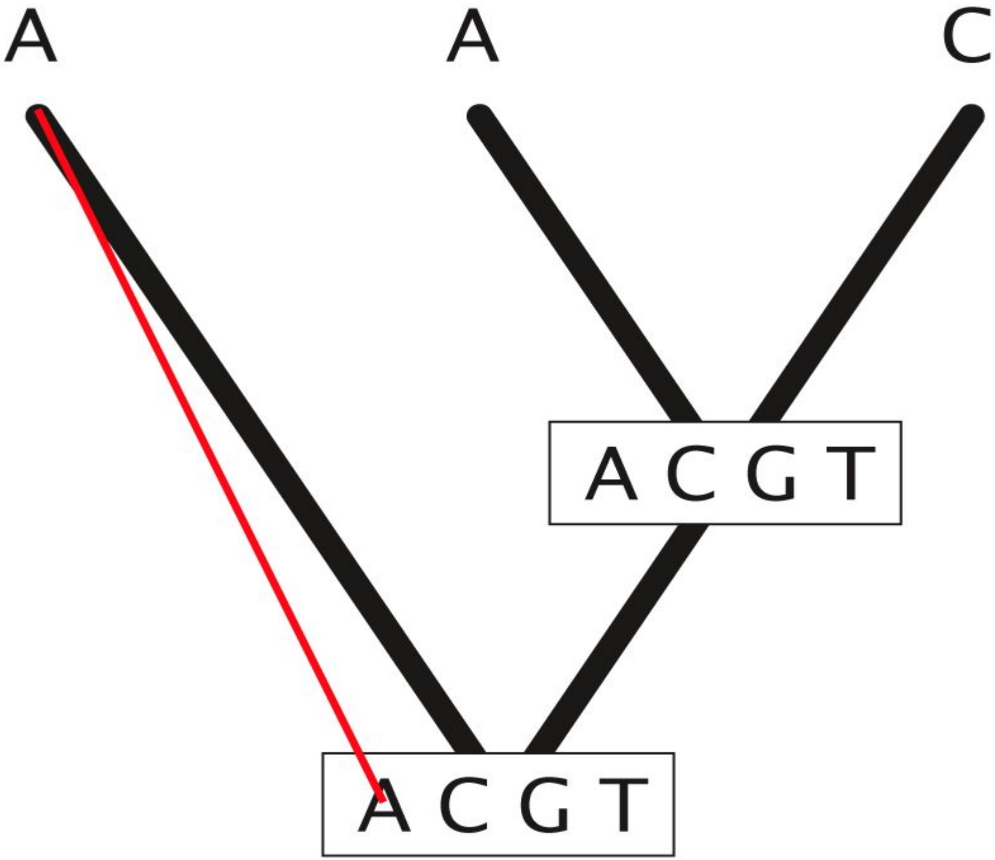
+



→

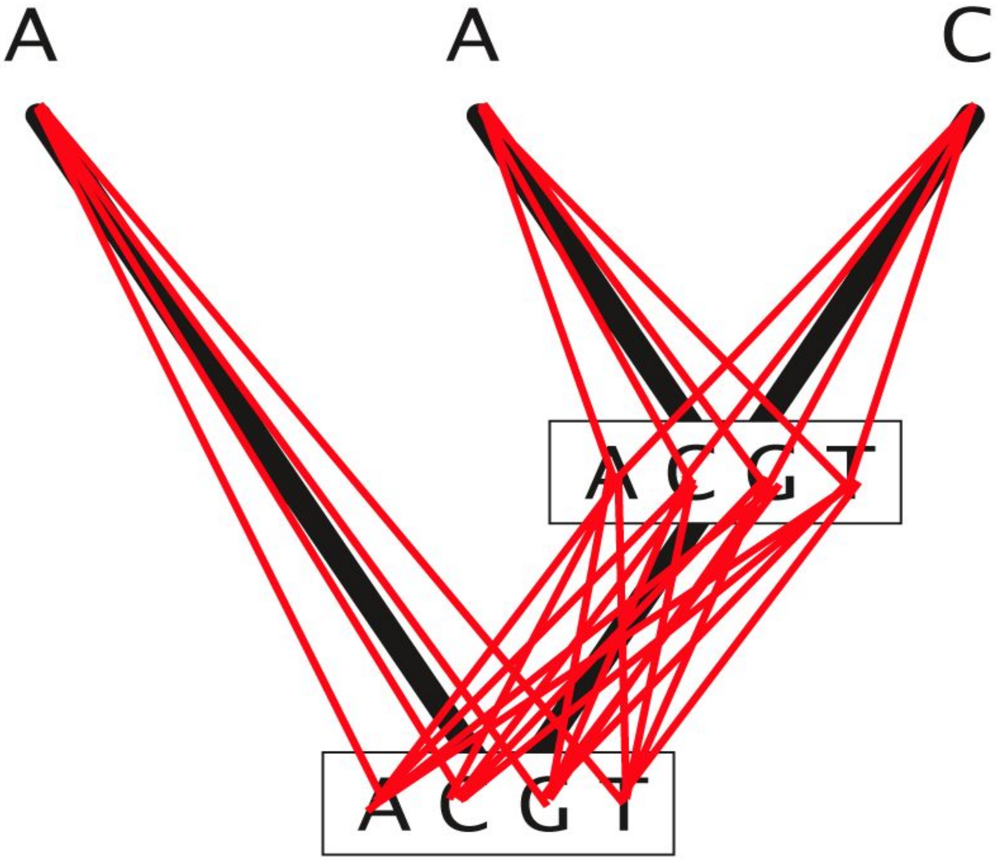
Probability of?

Brown bear	CGTTAGTACACT
Cave bear	CGATAGTTCACT
Black bear	CGTTAGTTTACC
Giant panda	CATTGGTTTACT



to

		A	C	G	T
from	A	–	a	b	c
	C	a	–	d	e
	G	b	d	–	f
	T	c	e	f	–



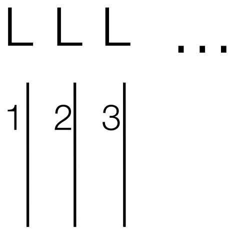
to

from

	A	C	G	T
A	-	a	b	c
C	a	-	d	e
G	b	d	-	f
T	c	e	f	-

Likelihood = all possible scenarios

Likelihood is multiplied across sites



Sample 1

CGTTAGTACACT

Sample 2

CGATAGTTCACT

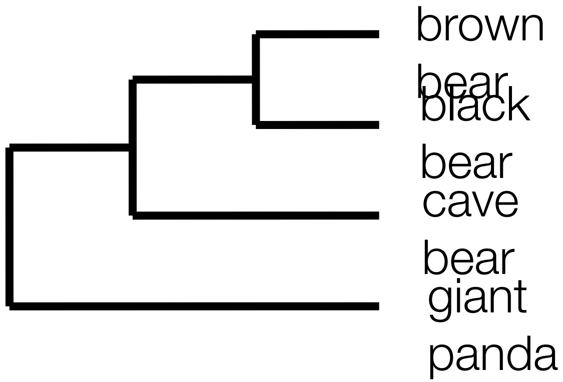
Sample 3

CGTTAGTTTACC

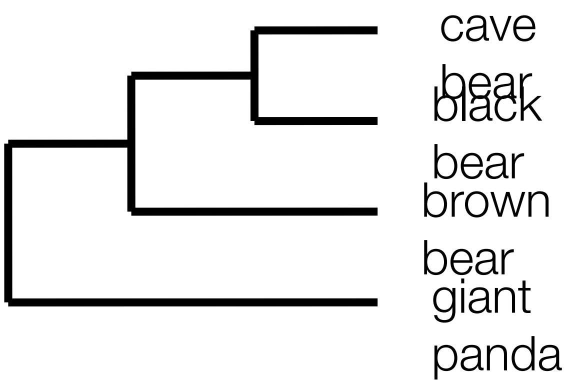
Likelihood values are very small!
(use log scale)

Sample 4

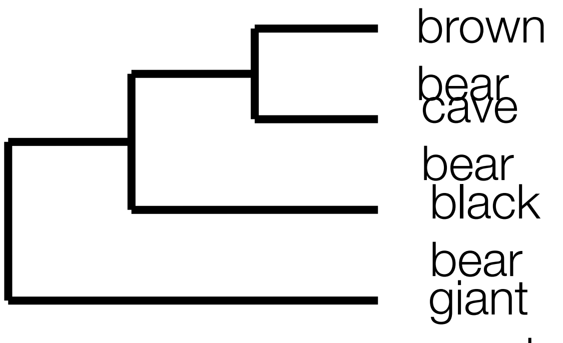
brown bear CGTTAGTACACT
cave bear CGATAGTTCACT
black bear CGTTAGTTTACC
giant panda CATTGGTTTACT



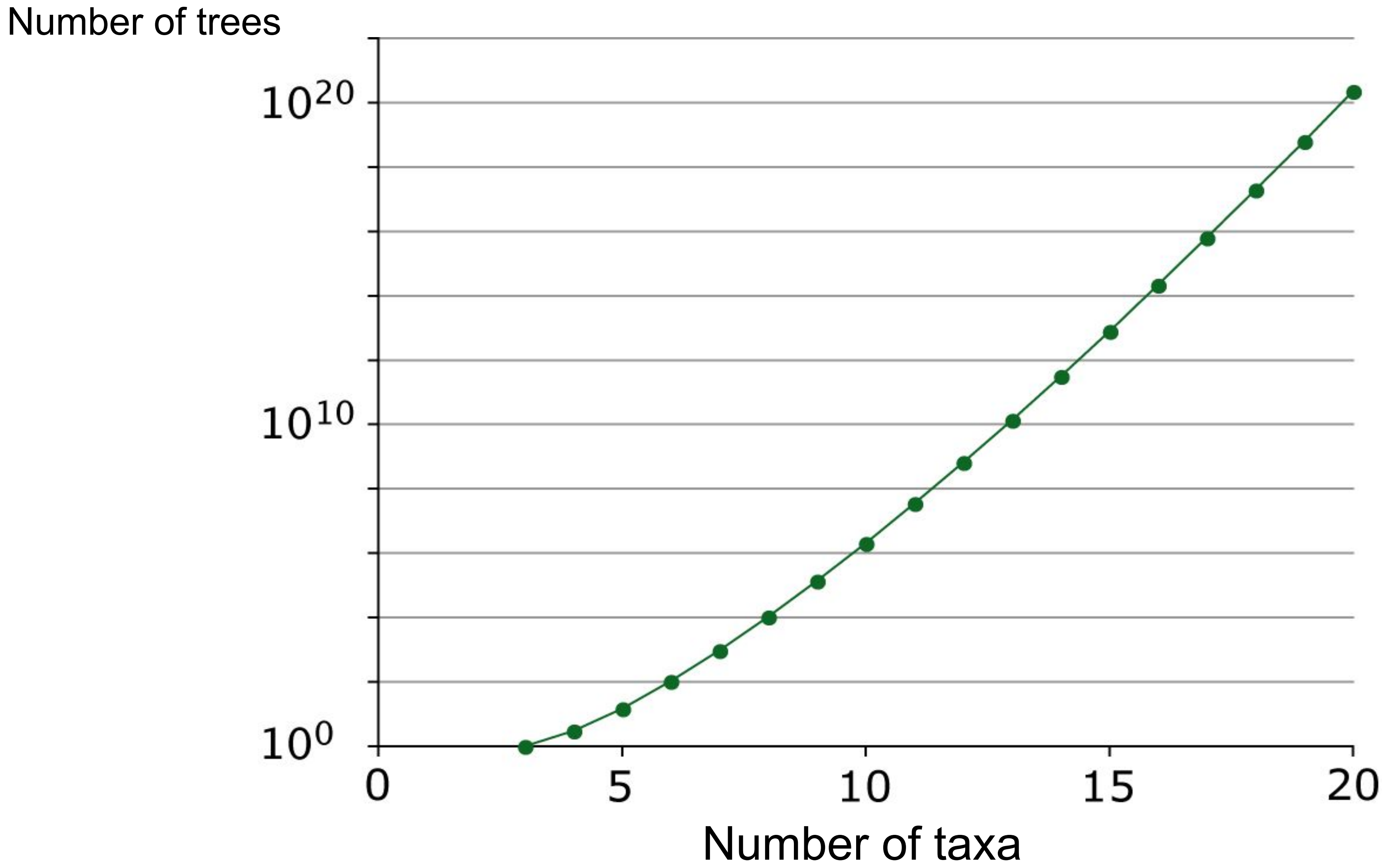
lnL = -1203.83



lnL = -1241.47



*lnL = -908.58


Number of trees
10^{20}
10^{10}
10^{0}
0
5
10
15
20
Number of taxa

Maximum likelihood

- Heuristic approaches to search tree space
- Single estimate of phylogenetic tree and parameters (MLE)
- Uncertainty via bootstrapping
- *We typically need additional methods to estimate rates, times, or demographic parameters

Bootstrapping

brown bear	CGTTAGTACACT
cave bear	CGATAGTTCACT
black bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

Bootstrapping

brown bear	CGTTAGTACACT
cave bear	CGATAGTTCACT
black bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

brown bear	T
cave bear	A
black bear	T
giant panda	T

Bootstrapping

brown bear	CGTTAGTACACT
cave bear	CGATAGTTCACT
black bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

brown bear	TT
cave bear	AT
black bear	TT
giant panda	TT

Bootstrapping

brown bear	CGTTAGTACACT
cave bear	CGATAGTTCACT
black bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

brown bear	TTC
cave bear	ATC
black bear	TTT
giant panda	TTT

Bootstrapping

brown bear	CGTTAGTACACT
cave bear	CGATAGTTCACT
black bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

brown bear	TTCT
cave bear	ATCT
black bear	TTTT
giant panda	TTTT

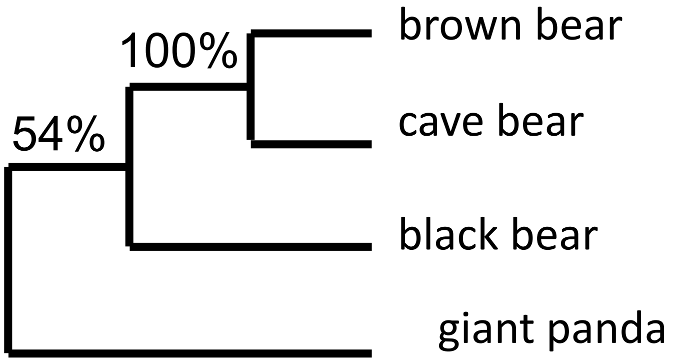
Bootstrapping

brown bear	CGTTAGTACACT
cave bear	CGATAGTTCACT
black bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

Pseudoreplication

Repeat 1,000 times

brown bear	TTCTAGTACACT
cave bear	ATCTAGTTCACT
black bear	TTTTAGTTTACC
giant panda	TTTTGGTTTACT



Concept review

- Maximum parsimony does not assume an explicit substitution model

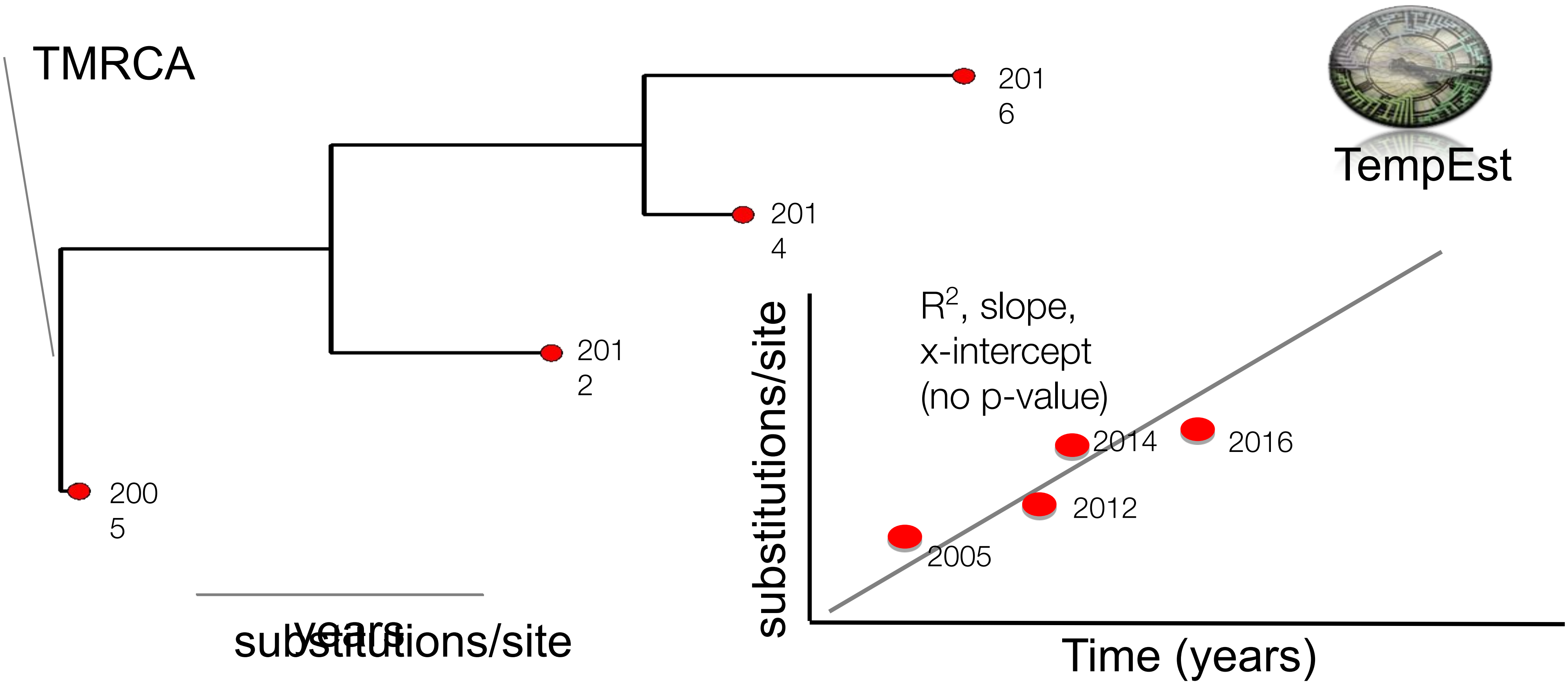
- Distance methods are very fast, but do not use all of the information
- for tree building

- Maximum likelihood is a true statistical approach. Obtaining uncertainty
- often requires additional approaches
  - (boostrapping, concordance factors)

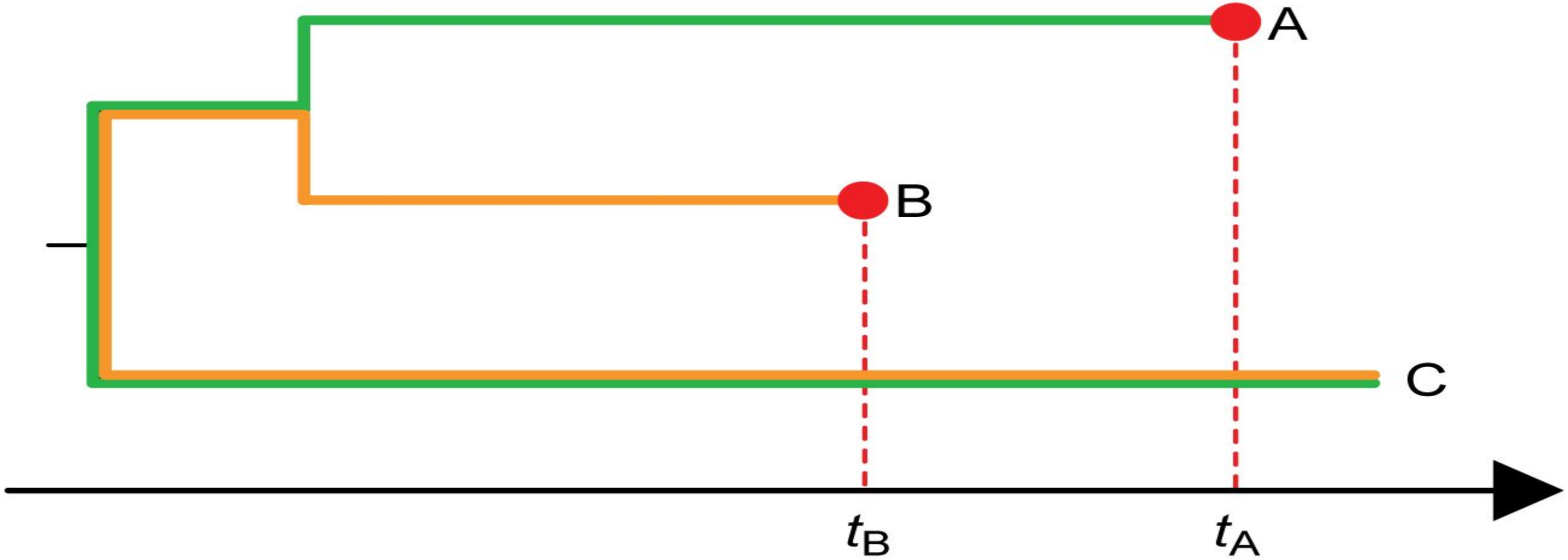
Recommended reading: Bromham, L. (2016). An introduction to molecular evolution and phylo.

The molecular clock

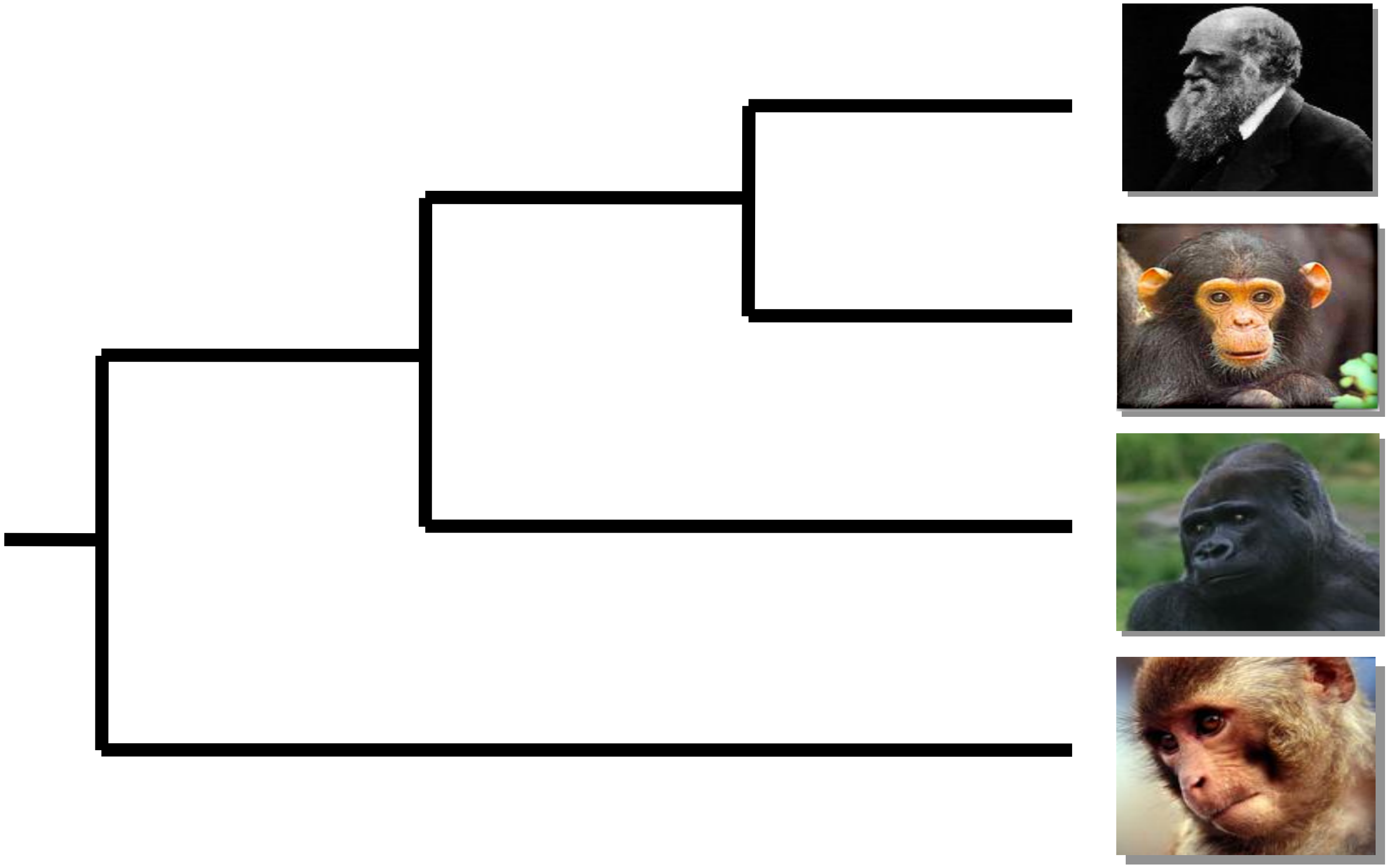
The molecular clock



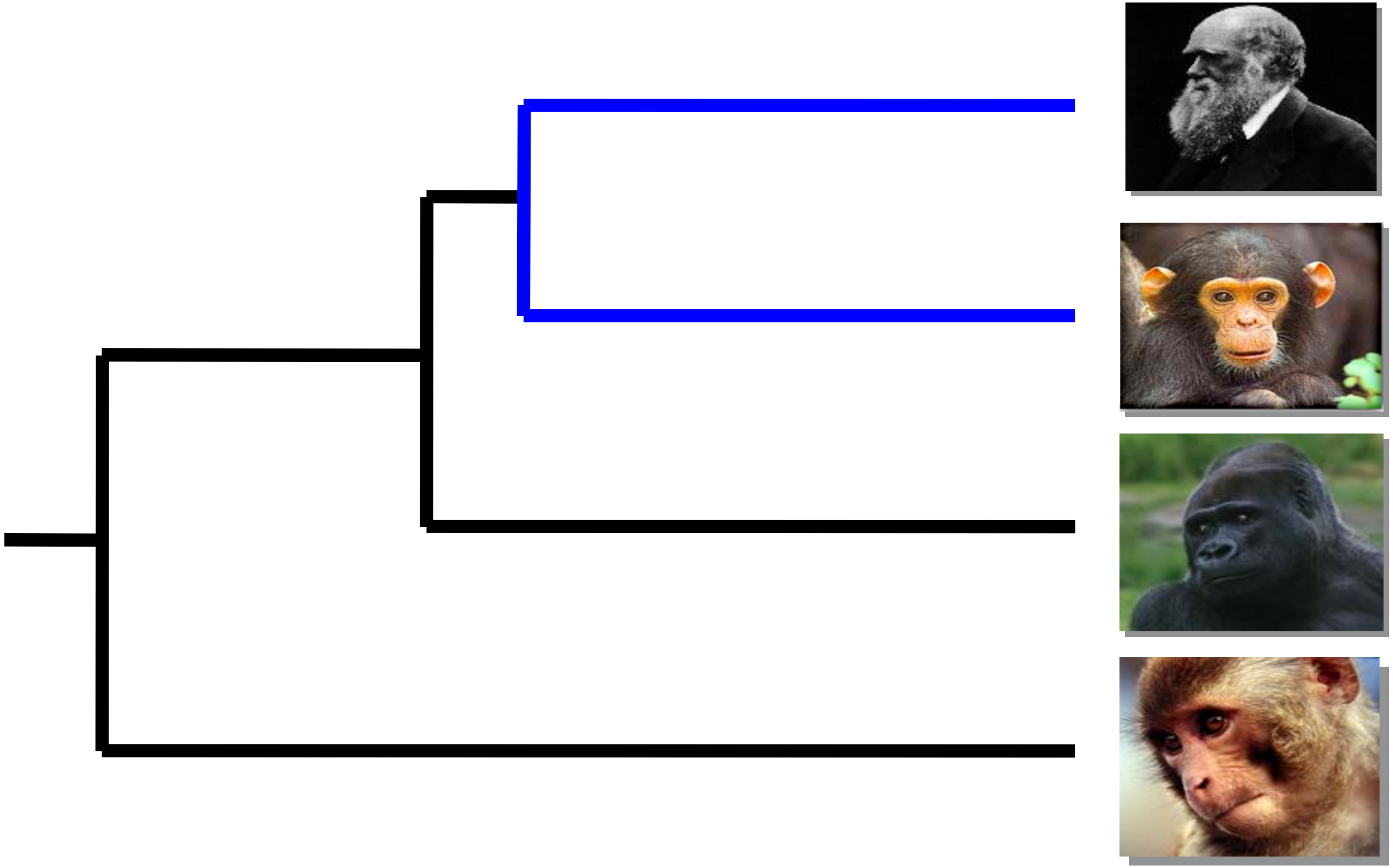
R^2, slope,
x-intercept
(no p-value)



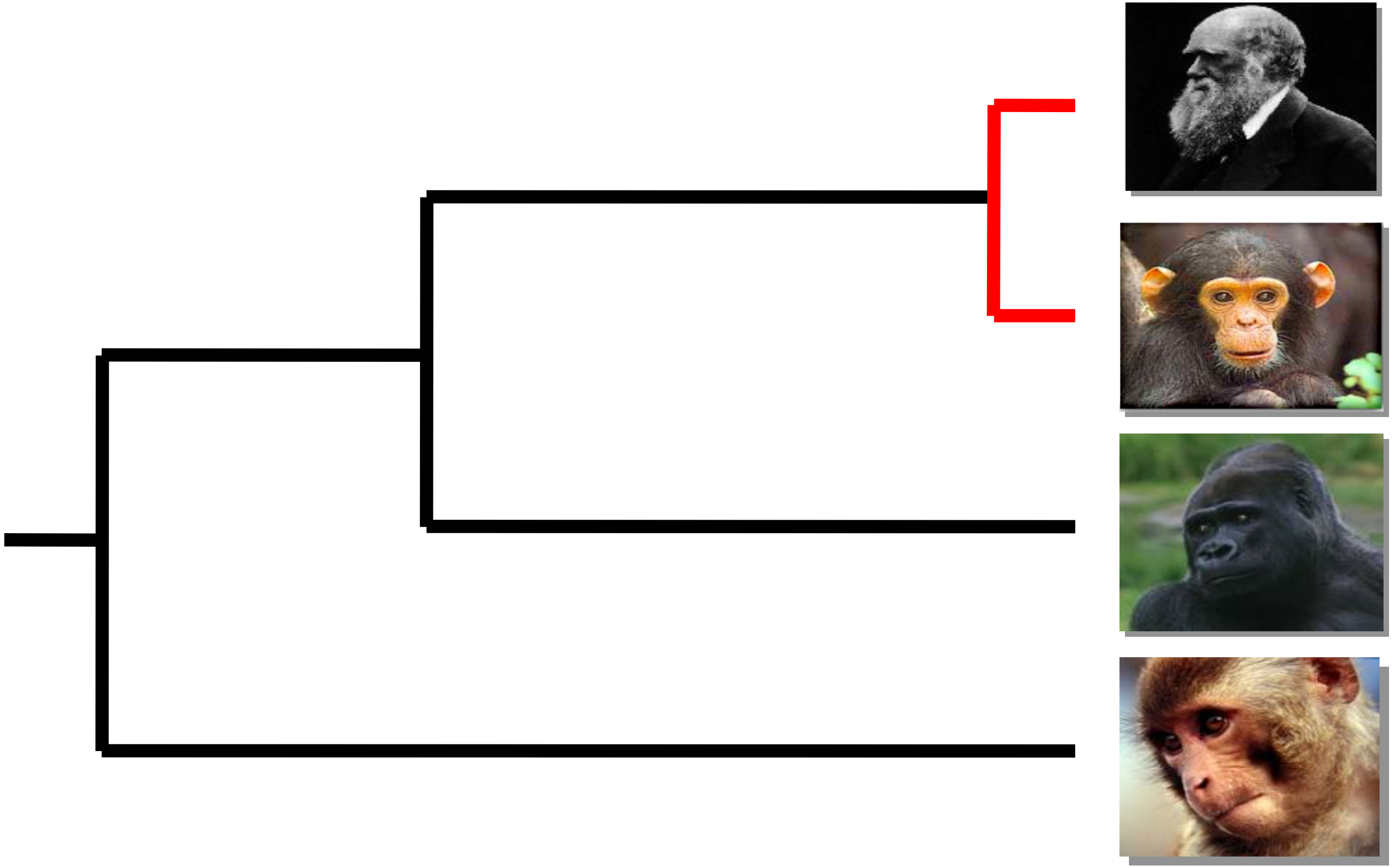
See: Rambaut (2000) Bioinformatics



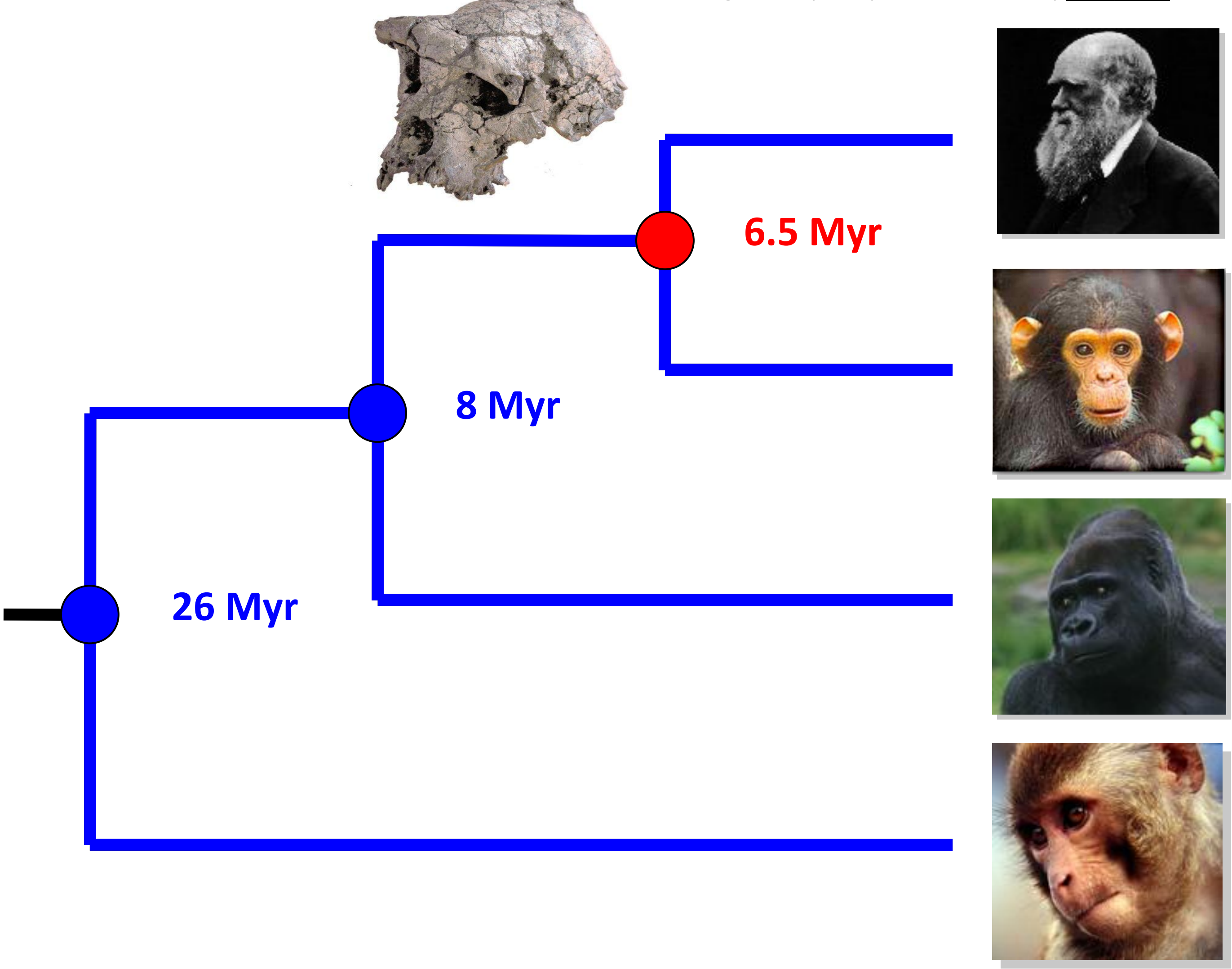
5% genetic difference



5% genetic difference



5% genetic difference



5% genetic difference

Concept review

- The molecular clock is the assumption that substitutions accumulate
- at a roughly constant rate over time

- Additional information, such as sequence sampling times or fossil
- data are required for **calibration** – rates and time are **unidentifiable**

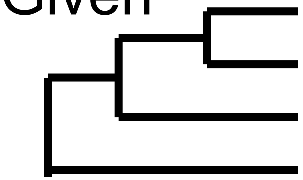
- The root-to-tip regression is a useful visual inspection, but it has major
- statistical limitations (do not interpret p-values)

Recommended reading:
Ho & Duchene (2014) Molecular-clock methods... Molecular
Ecology

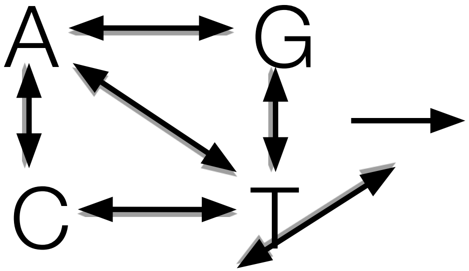
Bayesian phylogenetics: key concepts

Maximum likelihood

Given



+



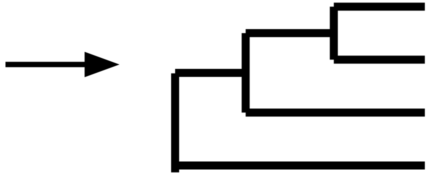
Probability of?

CGTTAGTACACT
CGATAGTTCACT
CGTTAGTTTACC
CATTGGTTTACT

Bayesian inference

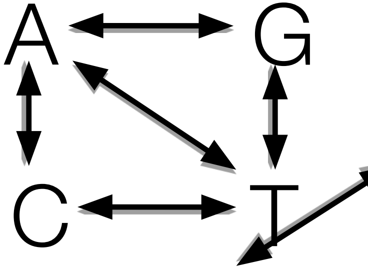
Given

CGTTAGTACACT
CGATAGTTCACT
CGTTAGTTTACC
CATTGGTTTACT

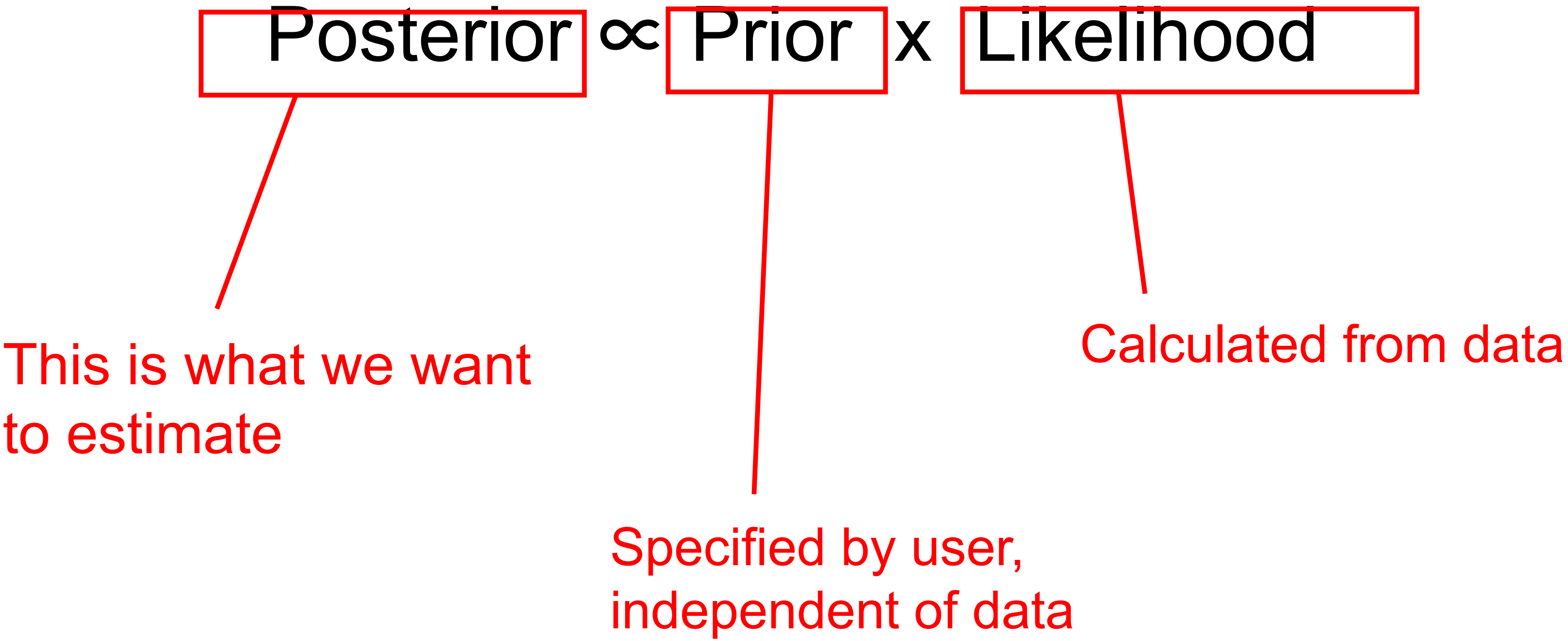


Probability of?

+



- Parameters have distributions
- Before the data are observed, each parameter has a prior distribution
- The likelihood of the data is computed, but not maximised
- The **prior** distribution is combined with the **likelihood** to yield the posterior distribution



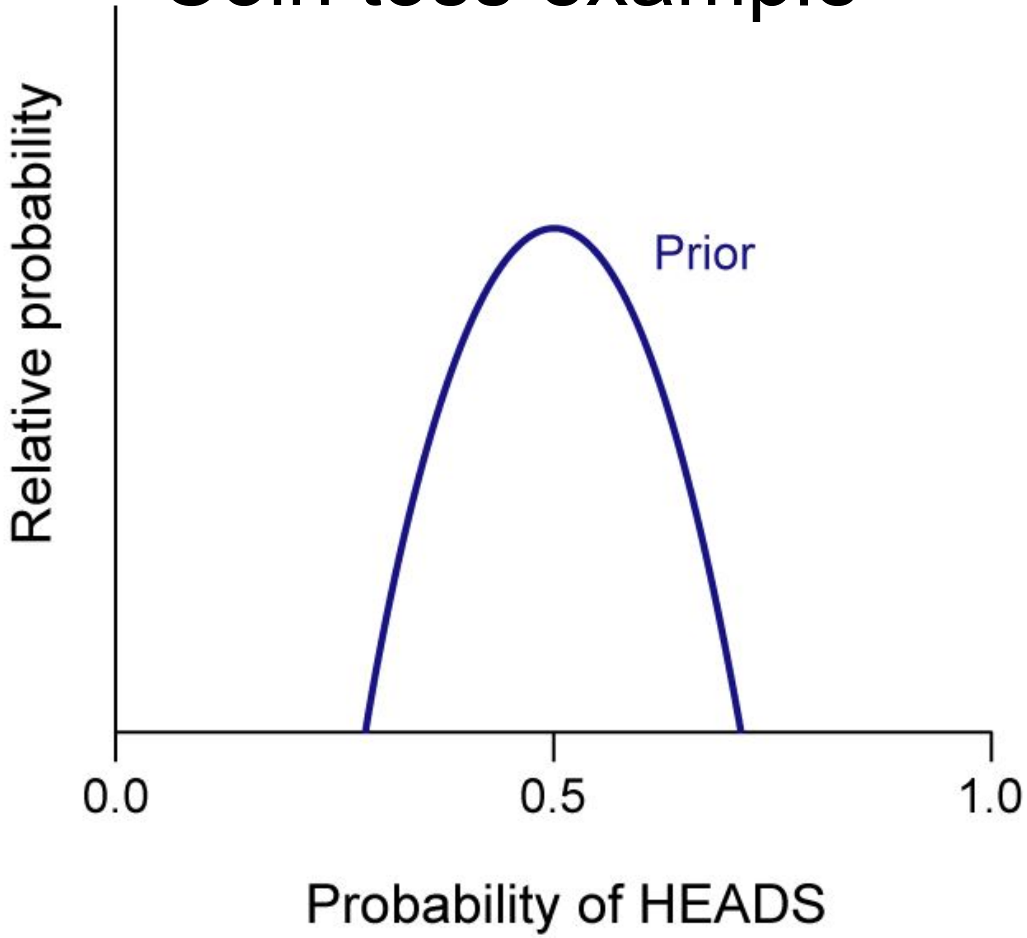
Posterior ∝ Prior x Likelihood

This is what we want to estimate

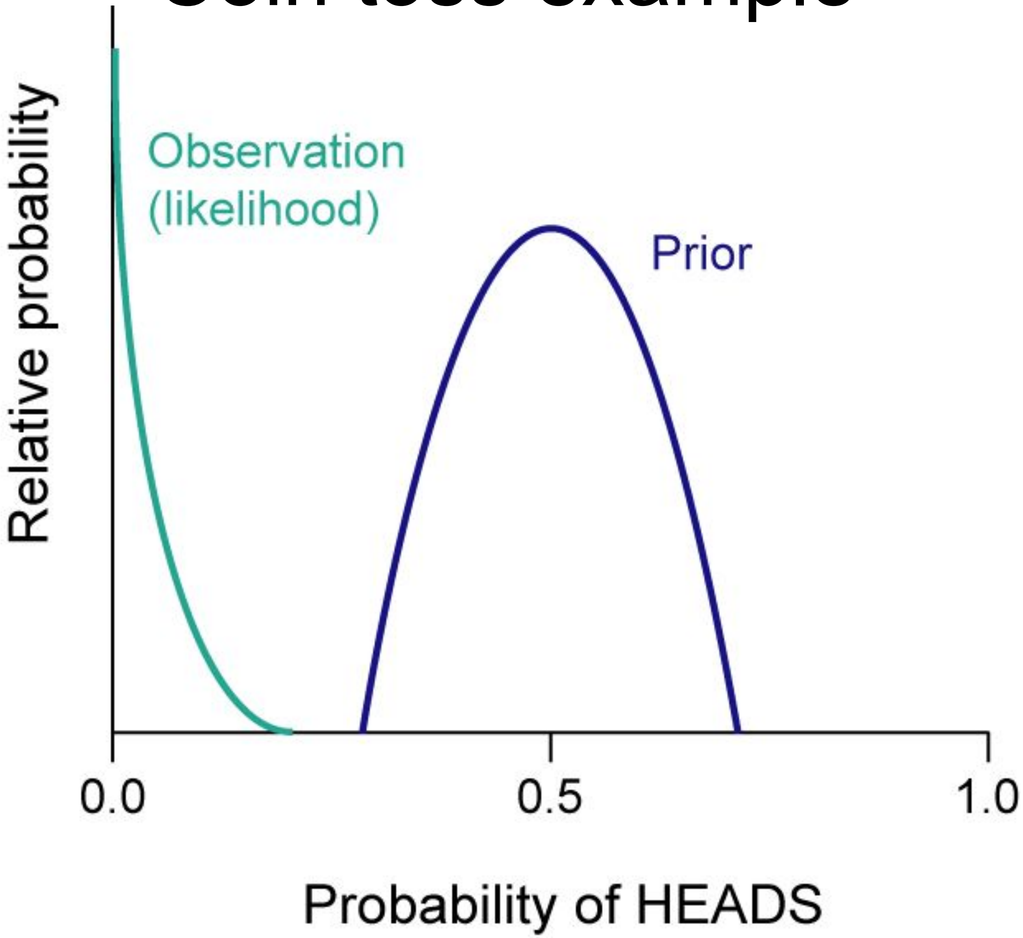
Specified by user, independent of data

Calculated from data

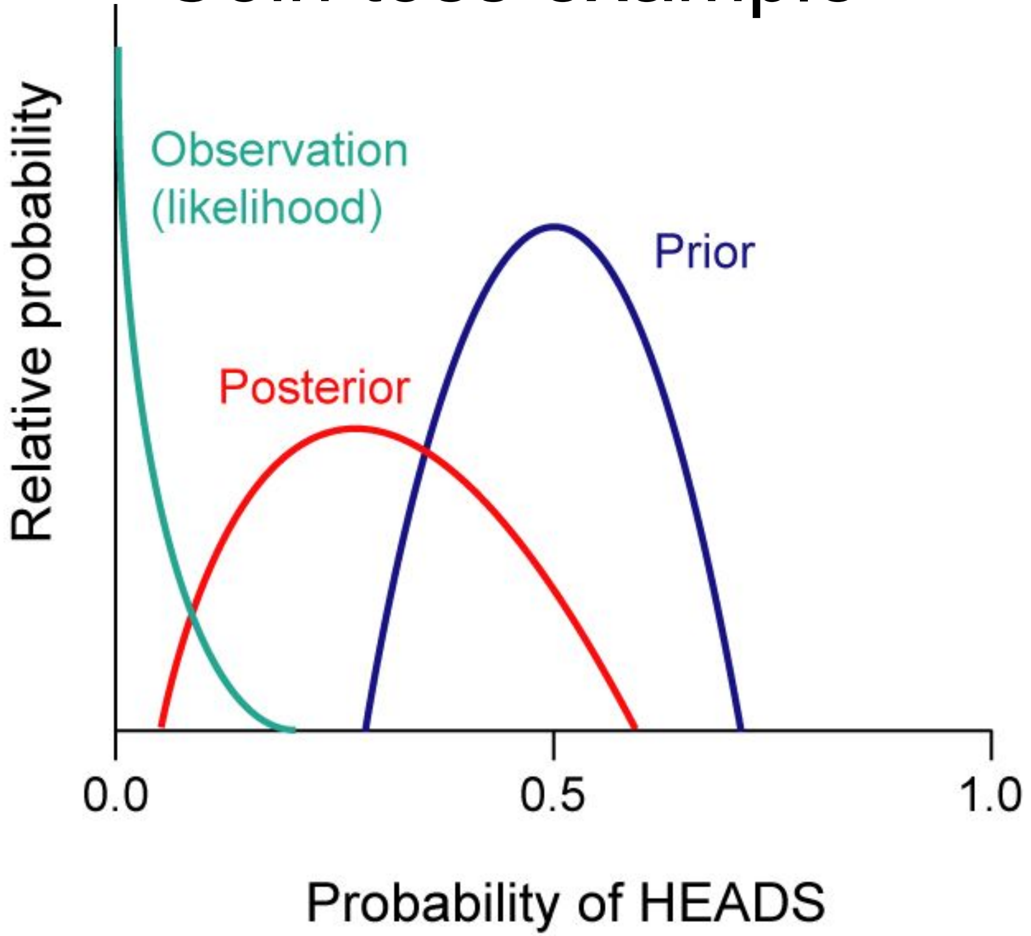
Coin toss example



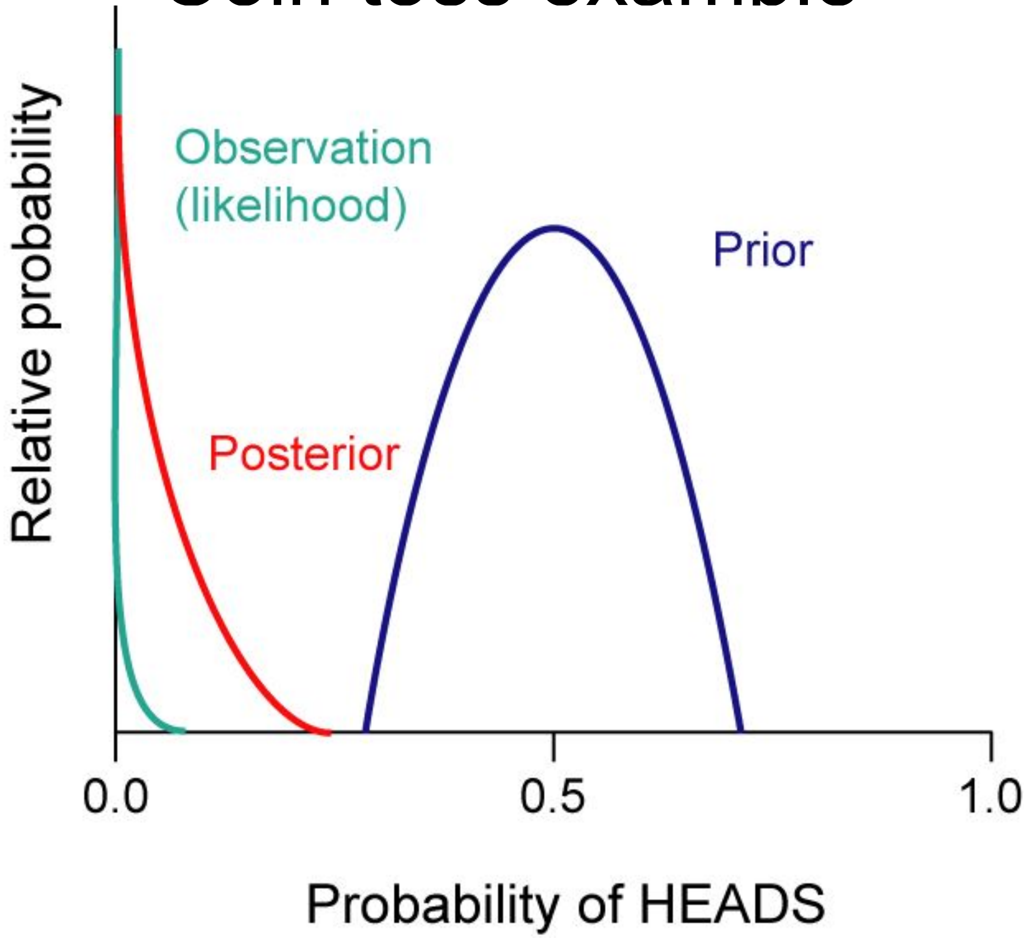
Coin toss example



Coin toss example



Coin toss example



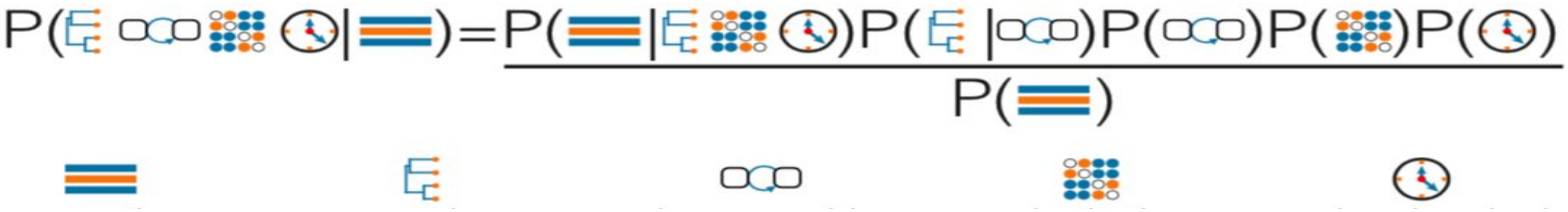
In phylogenetic models….

Phylogenetic tree (chronogram or phylogram)

Substitution model parameters

Evolutionary rates and time

Illustrations from du Plessis and Stadler 2015



Alignment

Chronogram

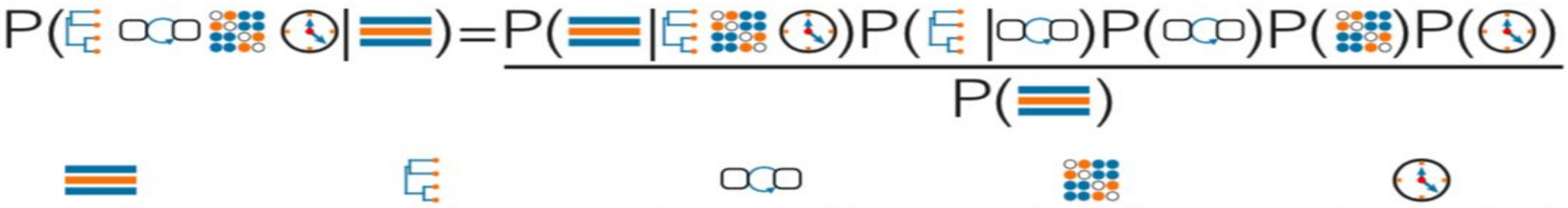
Branching model
(can be an epi model)

Substitution
model

Clock
model

For the tree prior we can use an epidemiological process to generate **chronograms**.

The phylogenetic likelihood is obtained by multiplying branching times by rates to obtain a **phylogram**



Alignment

Chronogram

Branching model
(can be an epi model)

Substitution
model

Clock
model

Note that the normalising constant, P(alignment), known as the marginal likelihood, is useful for model selection, but not usually computed.

Concept review

- Bayesian analyses also require computing a **likelihood**

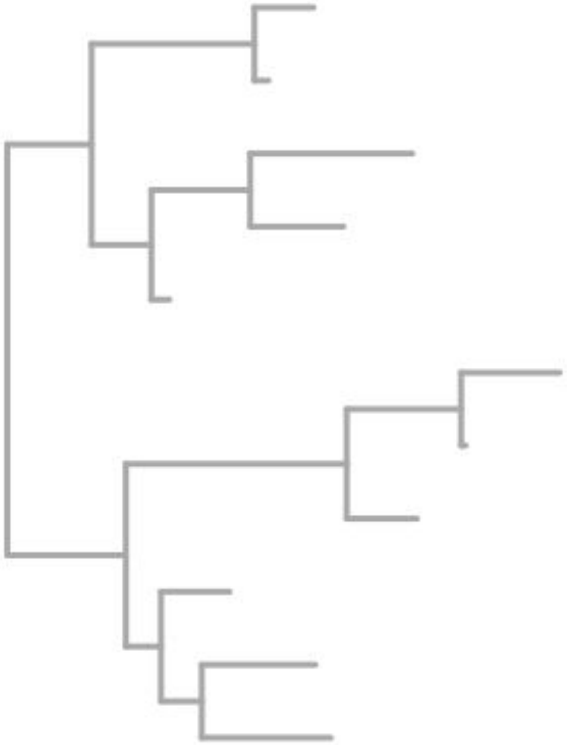
- The prior is essential for Bayesian analyses and usually obtained
- independently of the data

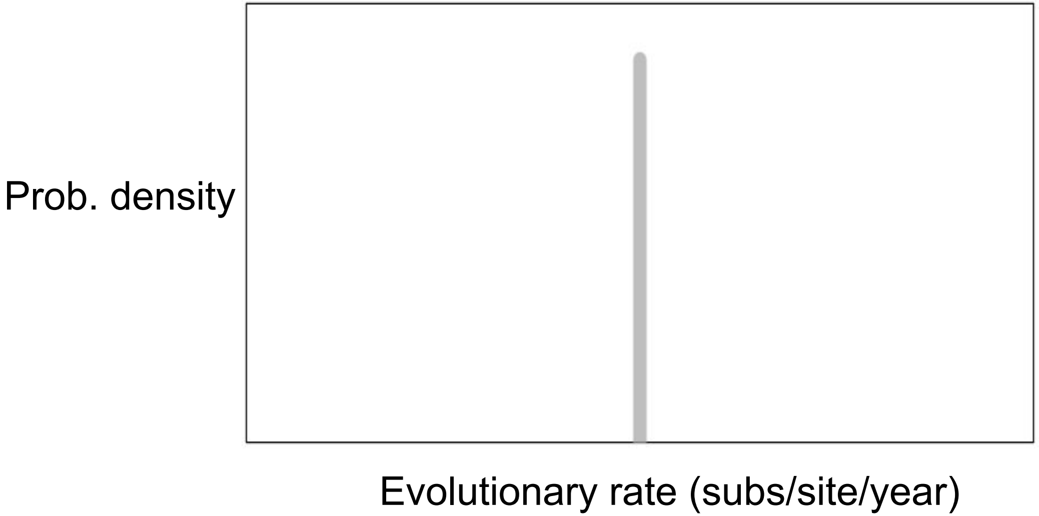
- We can specify more sophisticated models via the tree prior and
- the molecular clock model

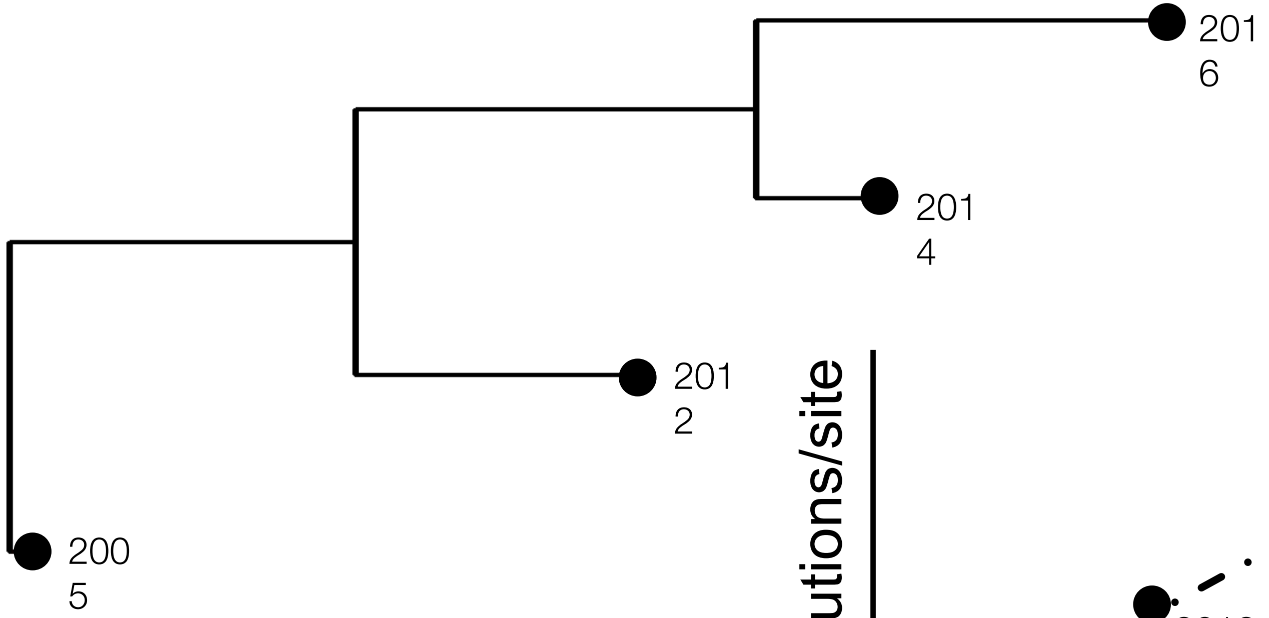
Recommended reading:
Bromham et al. (2018) Bayesian molecular dating... Biological

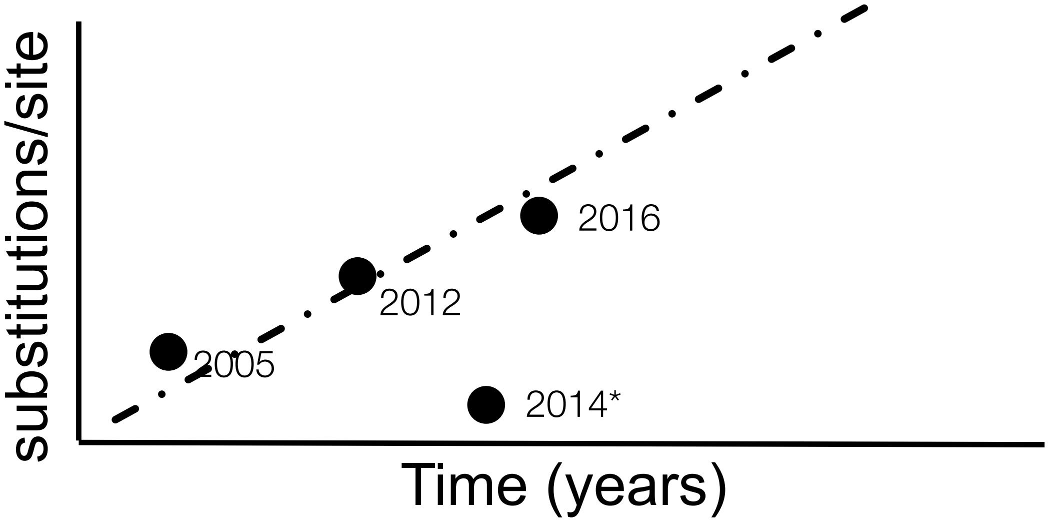
Bayesian phylogenetics: incorporating time and demography

A **strict** molecular clock

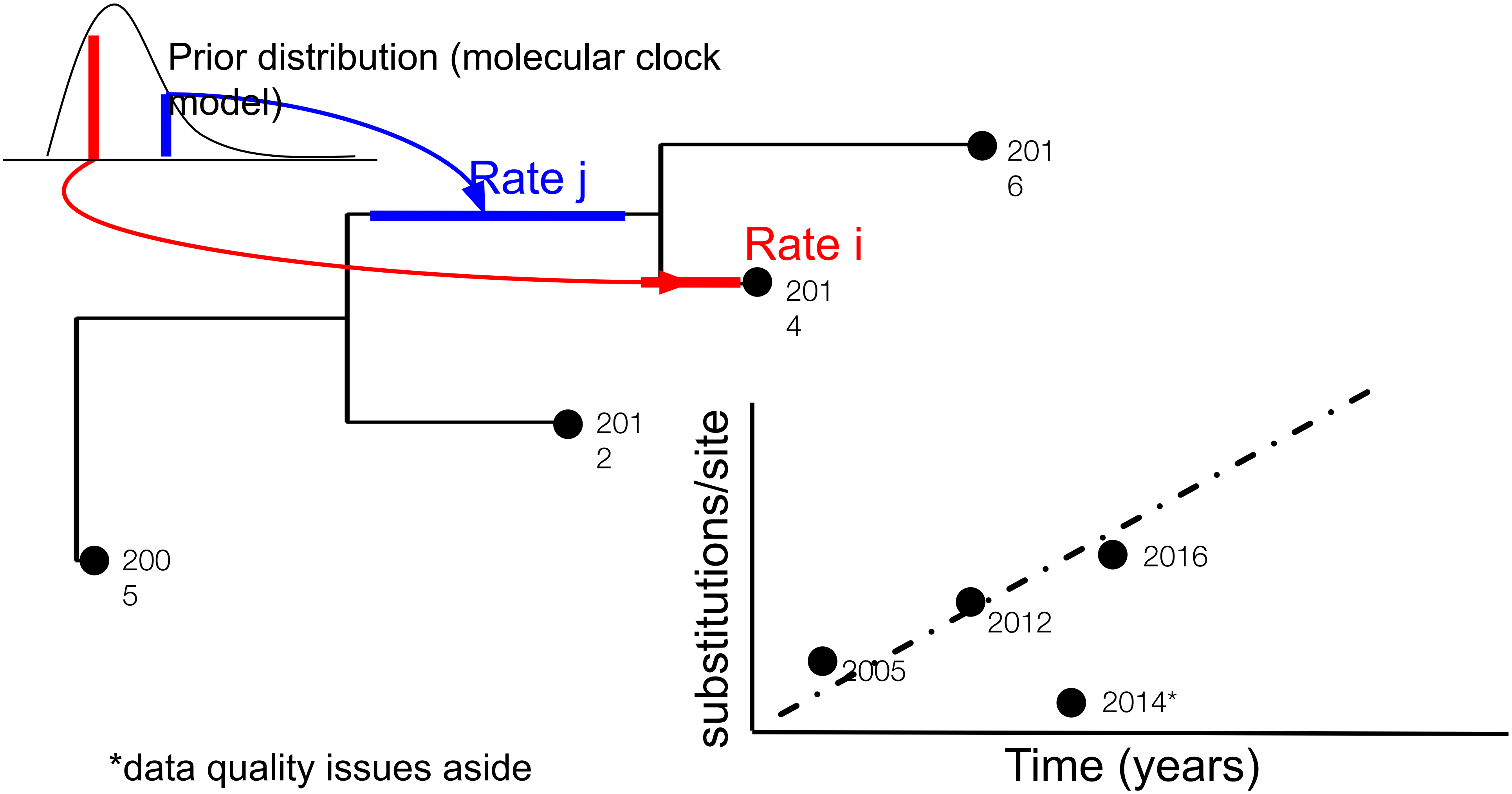






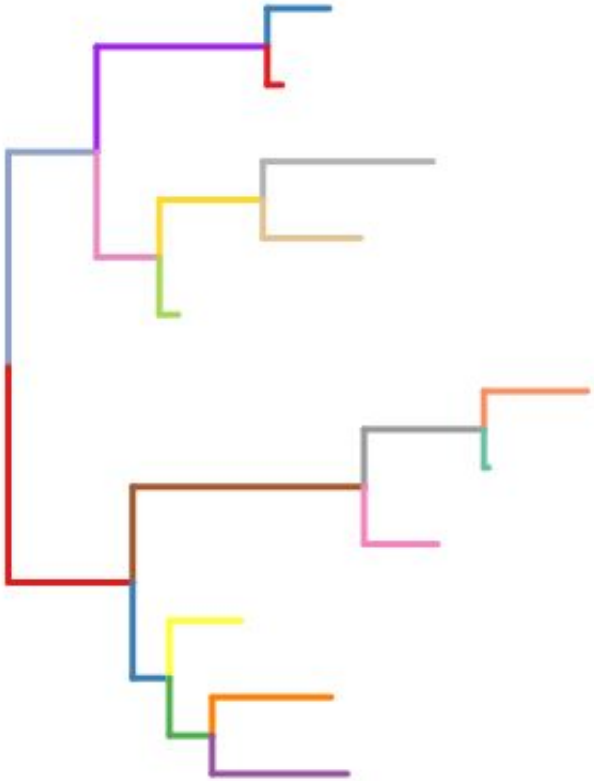


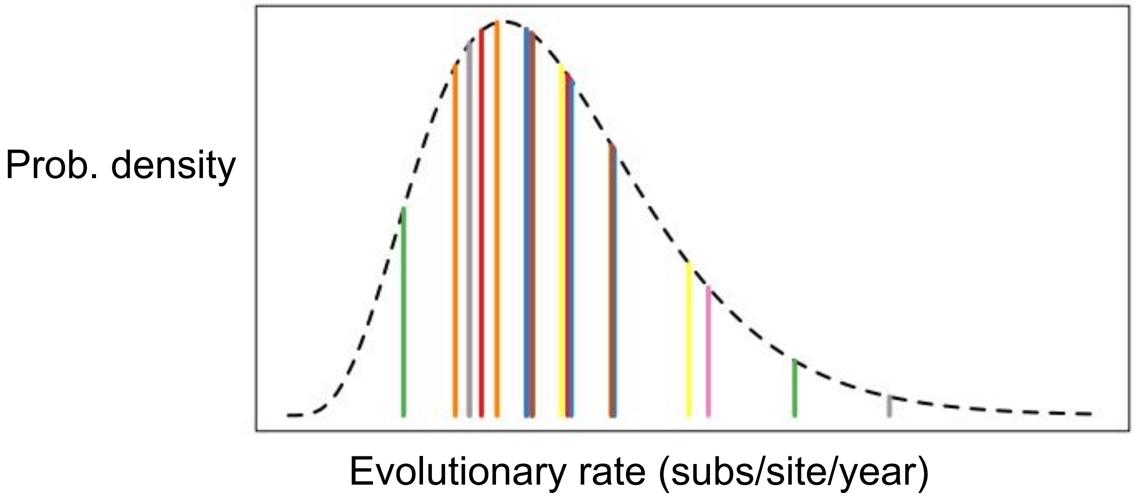
*data quality issues aside



*data quality issues aside

A **relaxed** molecular clock

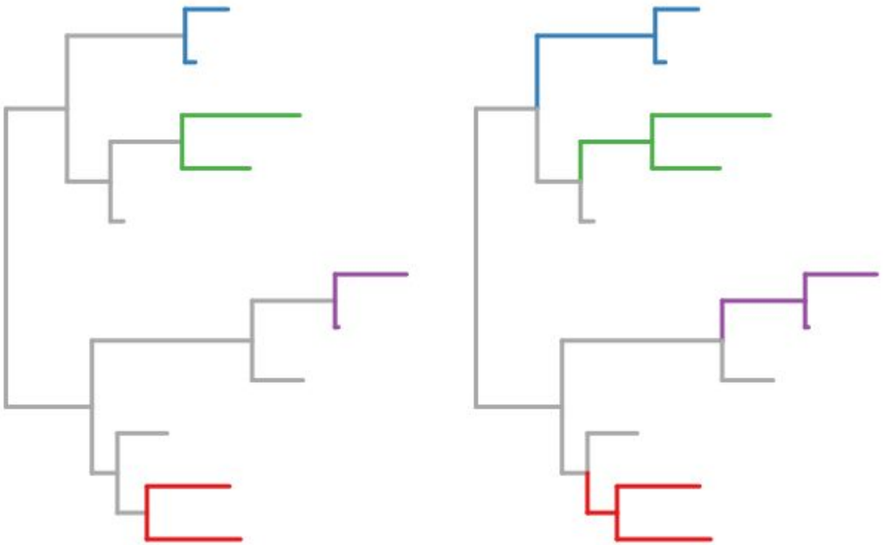


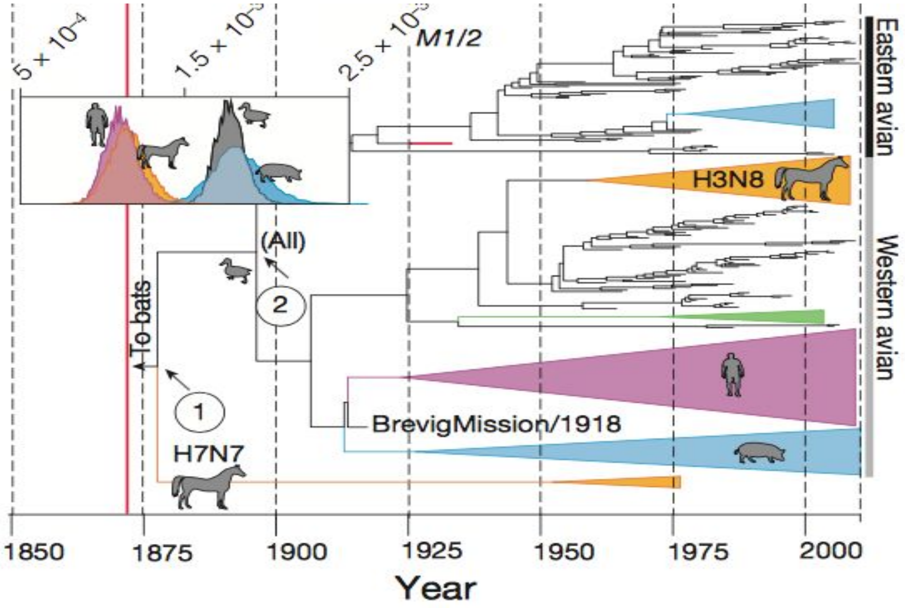


Gamma distro:
$\Gamma\ [\alpha, \beta]$
Mean = α / β

Lognormal distro:
LNorm $[\mu, \sigma^2]$
Mean = $e^{\mu + \sigma 2 / 2}$

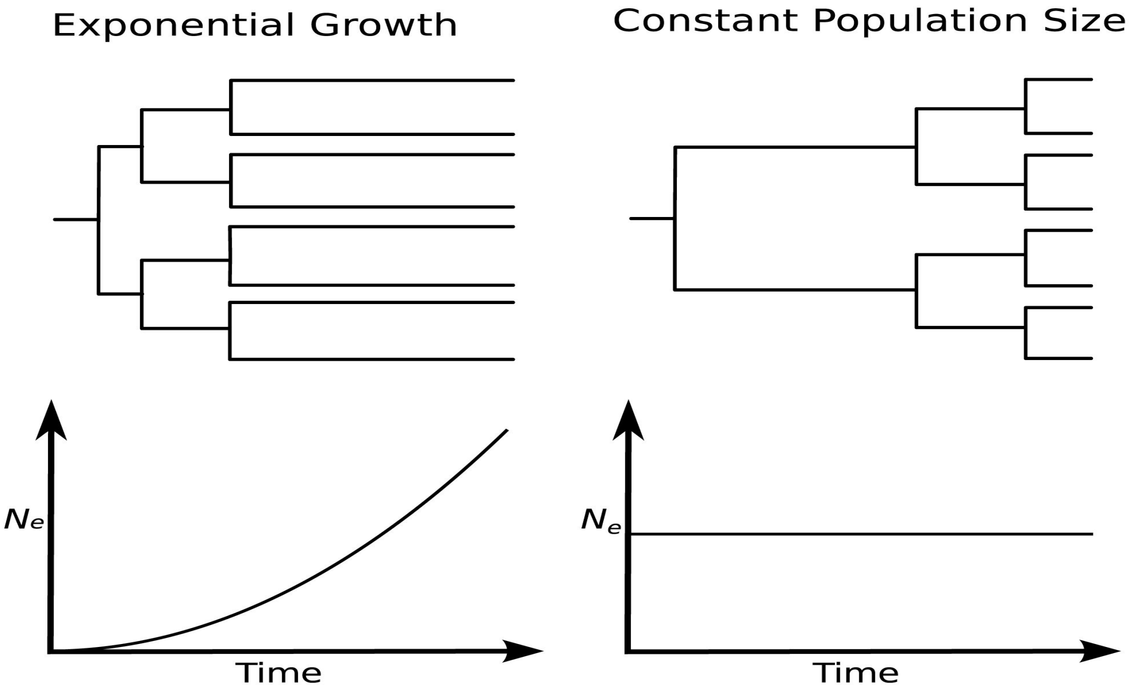
Local molecular clocks





From Worobey *et al.* (2014) Nature

Tree priors



From Volz et al. 2013

Recommended reading:
Featherstone et al. (2022). Epidemiological Inference From Pathogen Genomes... Virus Evolution

Concept review

- Bayesian molecular clock models can be proposed based on
- statistical convenience or biological motivation

- Demographic/epi models can inform us about changes in population
- size and genetic diversity – they are incorporated via the tree prior

- See how to sample the posterior distribution and summarise parameters
- and trees in the prac later today

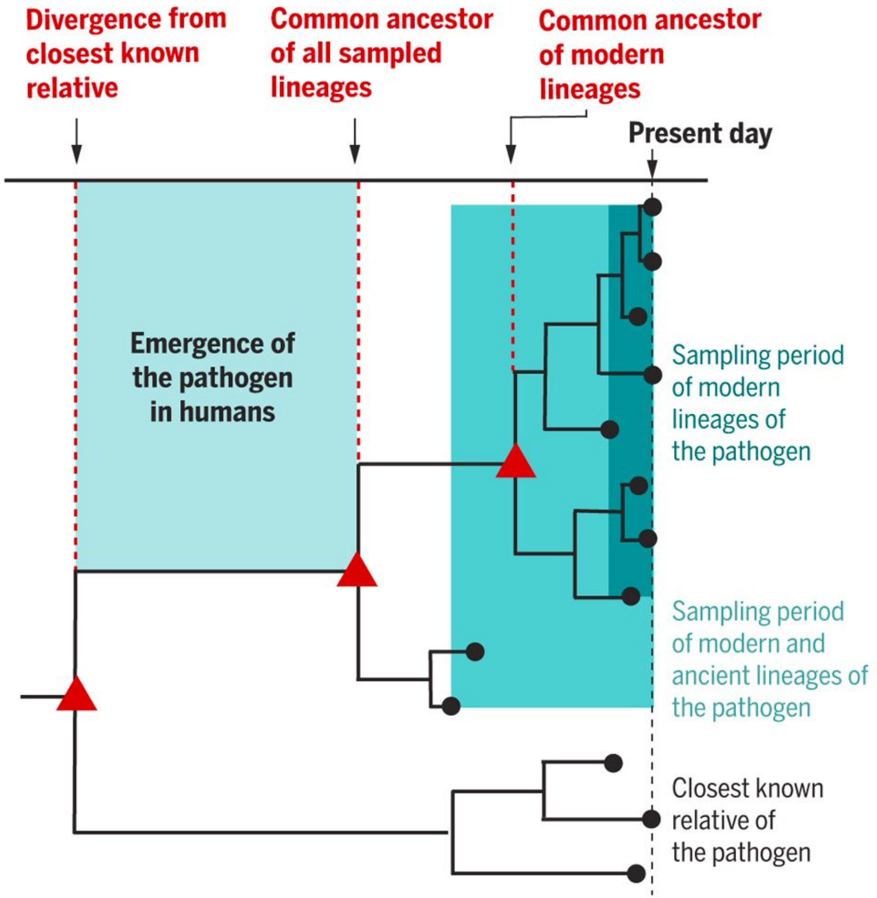
Molecular clock in ancient DNA data

Considerations for ancient DNA data

- Highly informative because the sampling window can be very wide
- (temporal signal)

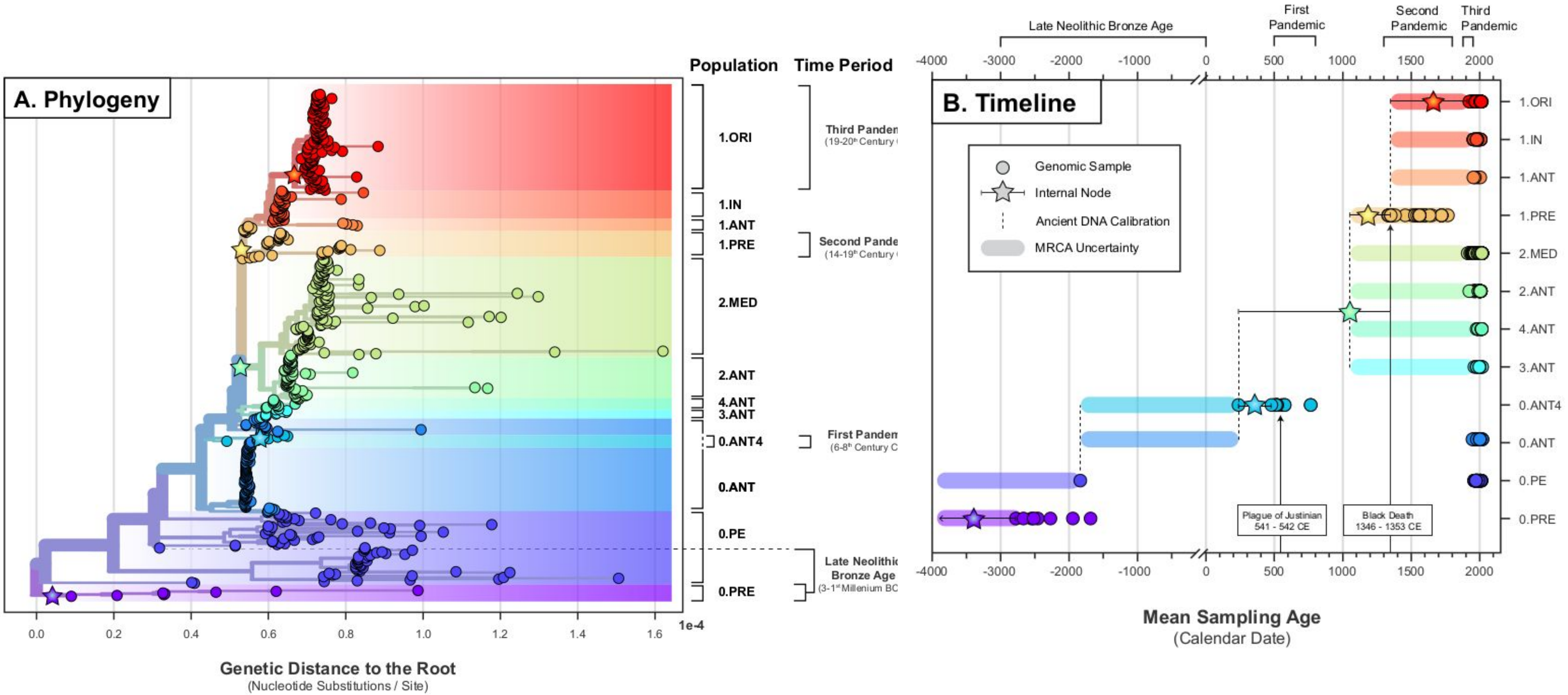
- Often many variable sites → lots of information, but also lots of computing

- The molecular clock rarely holds


Divergence from closest known relative
Common ancestor of all sampled lineages
Common ancestor of modern lineages
Present day
Emergence of the pathogen in humans
Sampling period of modern lineages of the pathogen
Sampling period of modern and ancient lineages of the pathogen
Closest known relative of the pathogen

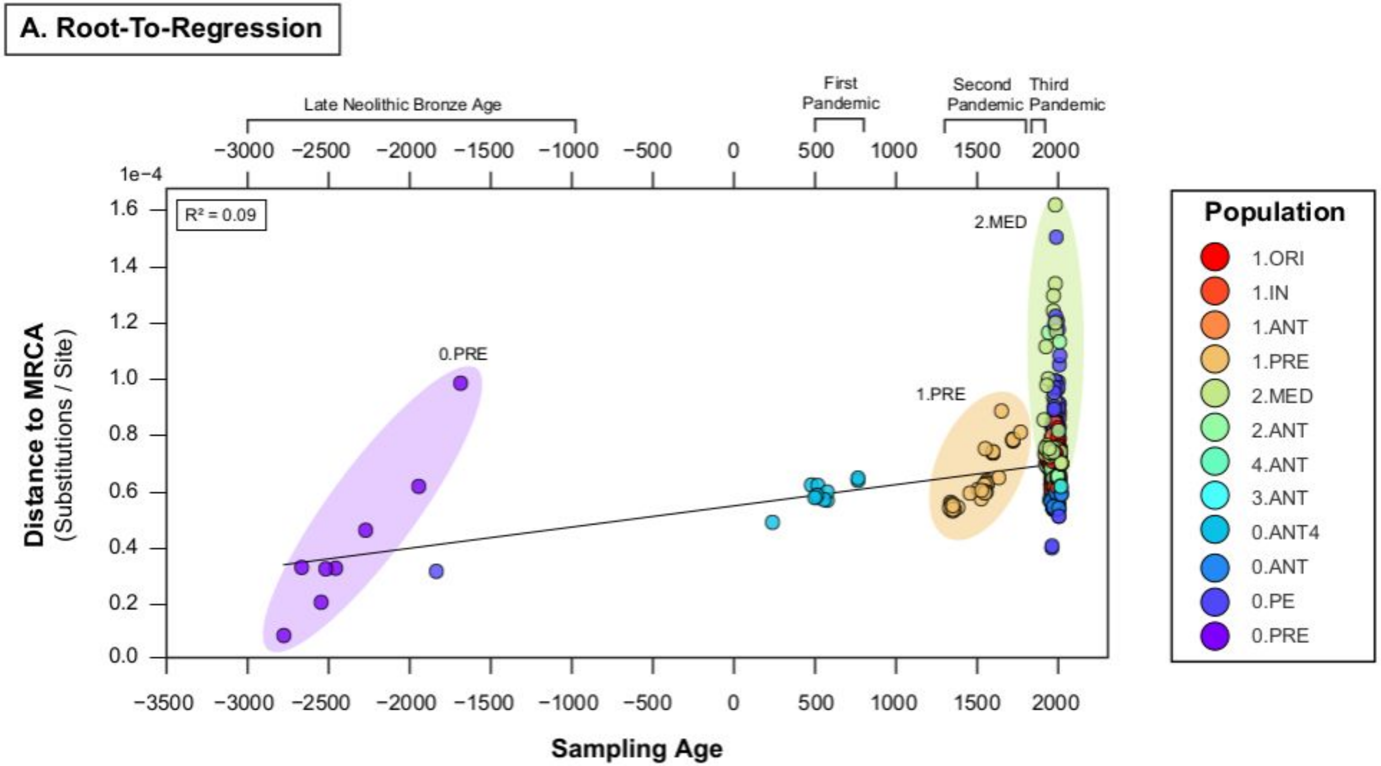
From Ho and Duchene 2020

The molecular clock of *Yersinia pestis*



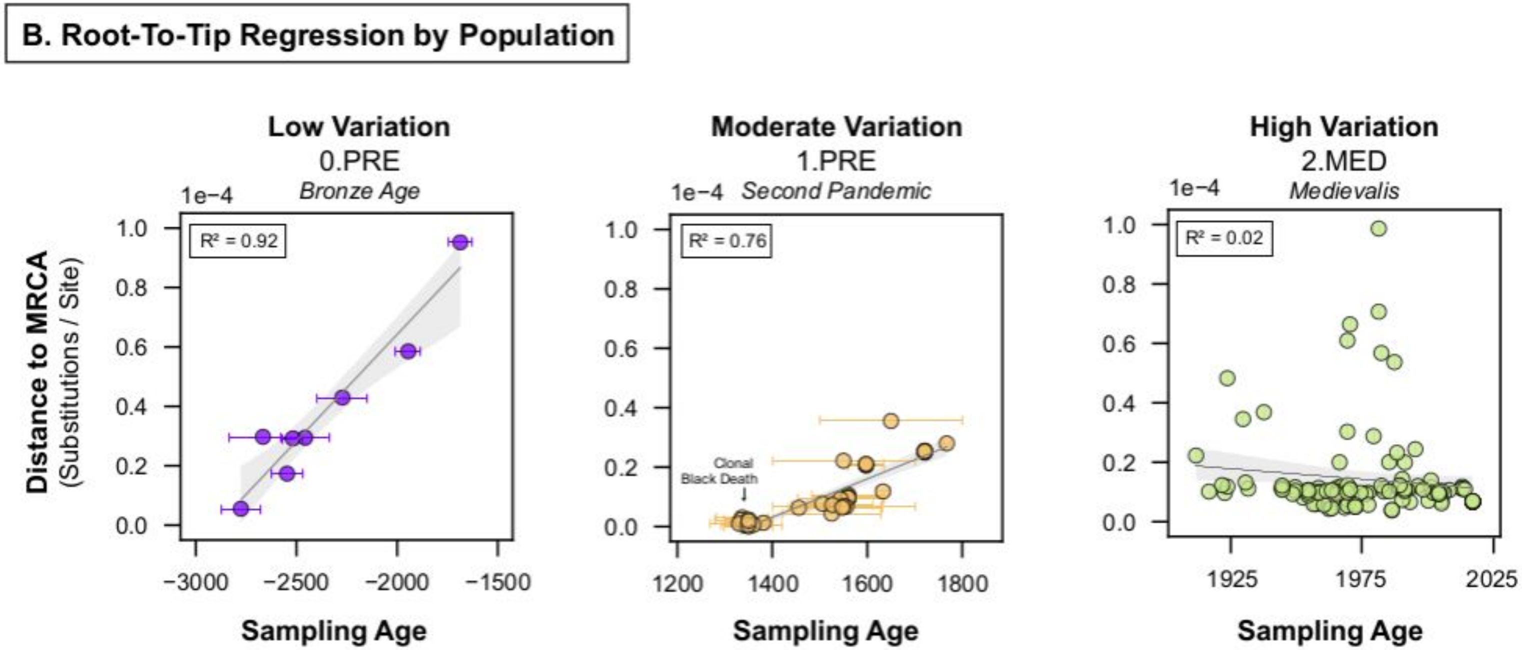
From Eaton et al 2022

The molecular clock of *Yersinia pestis*



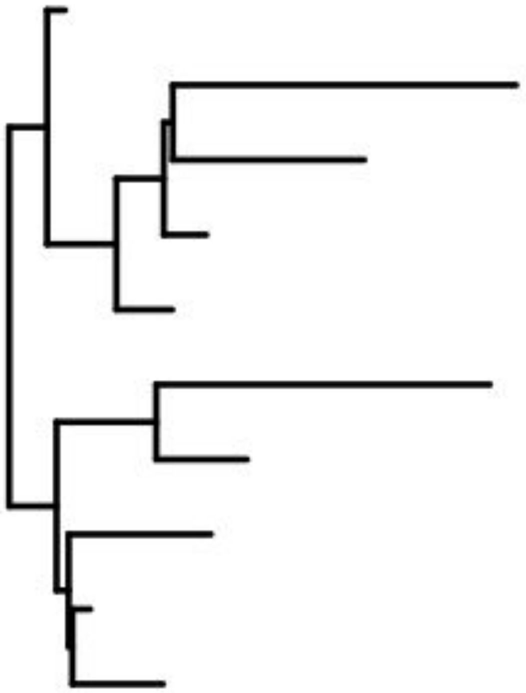
From Eaton et al 2022

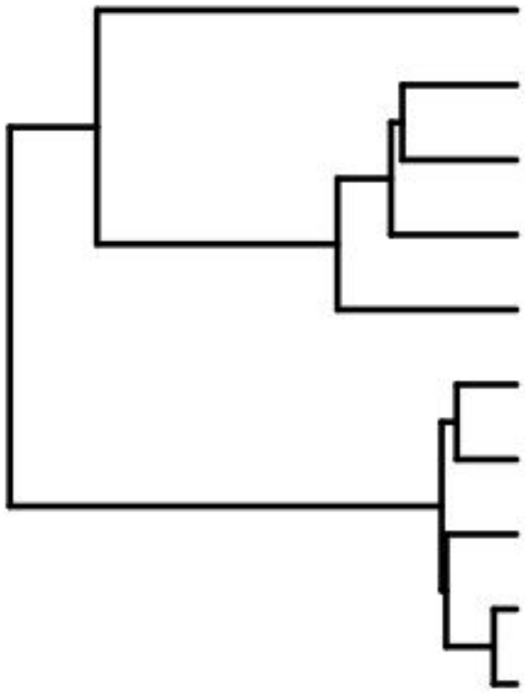
The molecular clock of *Yersinia pestis*



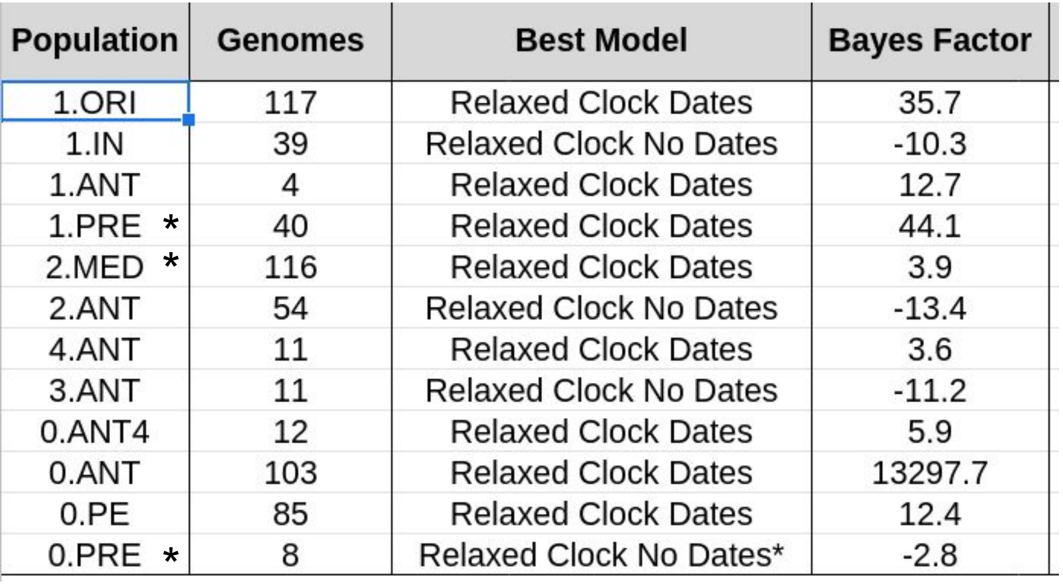
From Eaton et al 2022

Bayesian Evaluation of Temporal Signal (BETS)

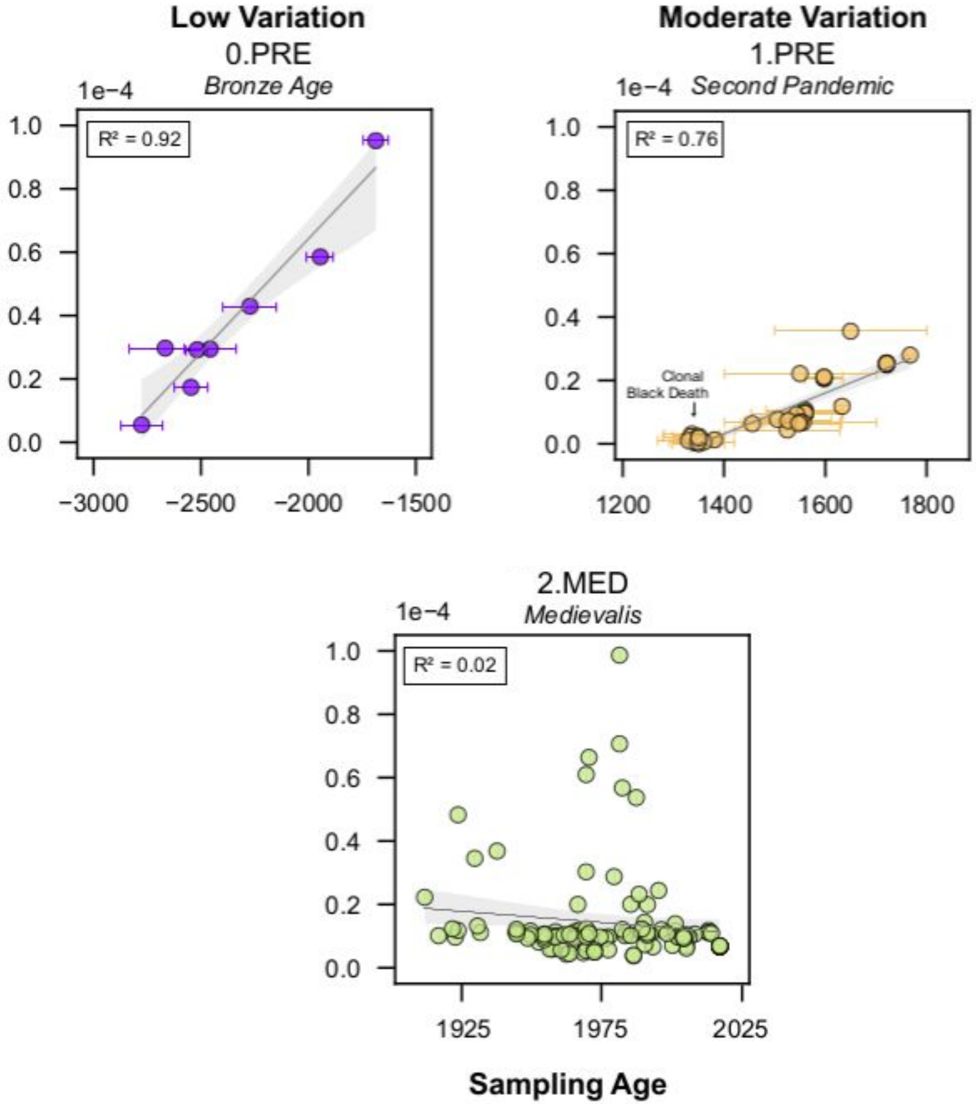




See:
Duchene et al. (2020). Bayesian evaluation of temporal signal... Molecular Biology and Evolution

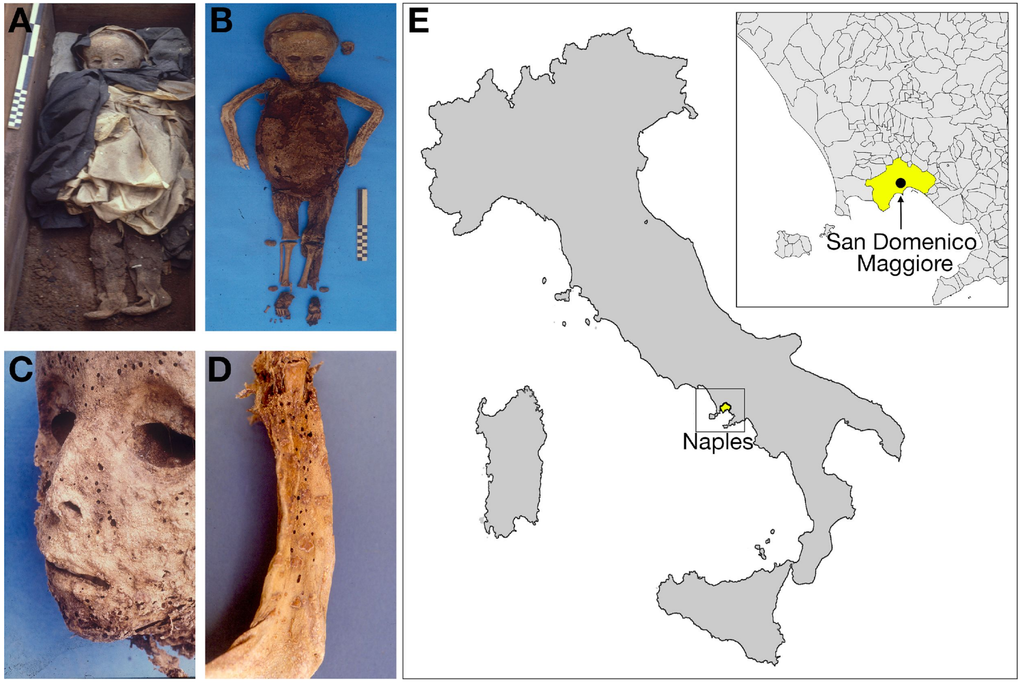


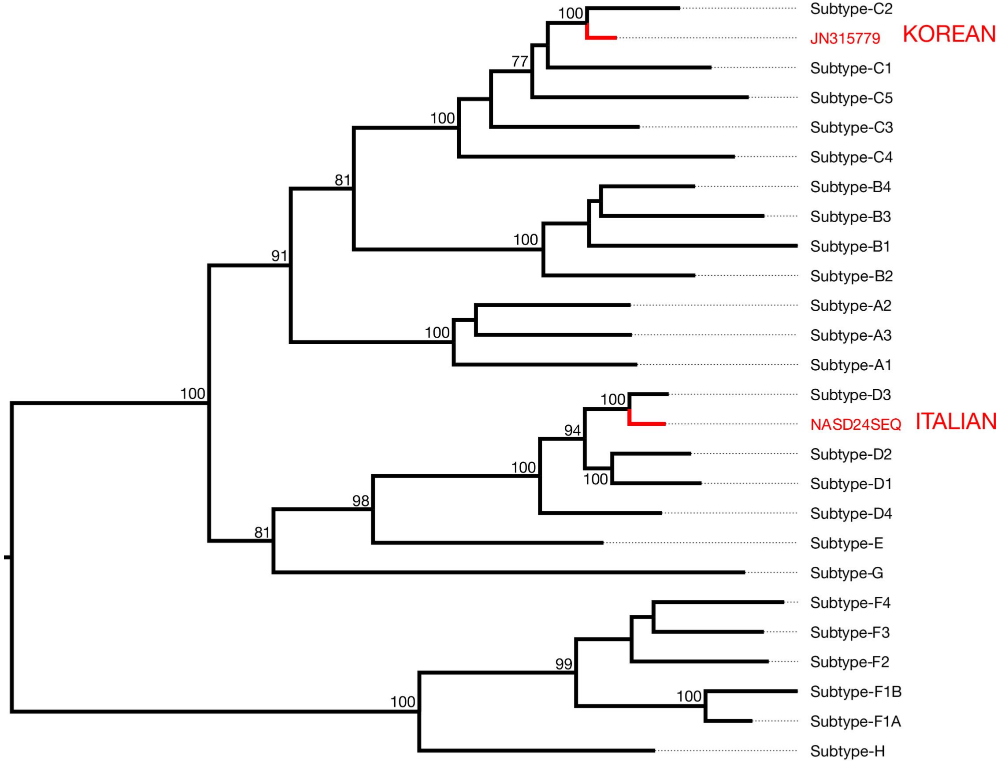
Population	Genomes	Best Model	Bayes Factor
1.ORI	117	Relaxed Clock Dates	35.7
1.IN	39	Relaxed Clock No Dates	-10.3
1.ANT	4	Relaxed Clock Dates	12.7
1.PRE *	40	Relaxed Clock Dates	44.1
2.MED *	116	Relaxed Clock Dates	3.9
2.ANT	54	Relaxed Clock No Dates	-13.4
4.ANT	11	Relaxed Clock Dates	3.6
3.ANT	11	Relaxed Clock No Dates	-11.2
0.ANT4	12	Relaxed Clock Dates	5.9
0.ANT	103	Relaxed Clock Dates	13297.7
0.PE	85	Relaxed Clock Dates	12.4
0.PRE *	8	Relaxed Clock No Dates*	-2.8



From Eaton et al 2022

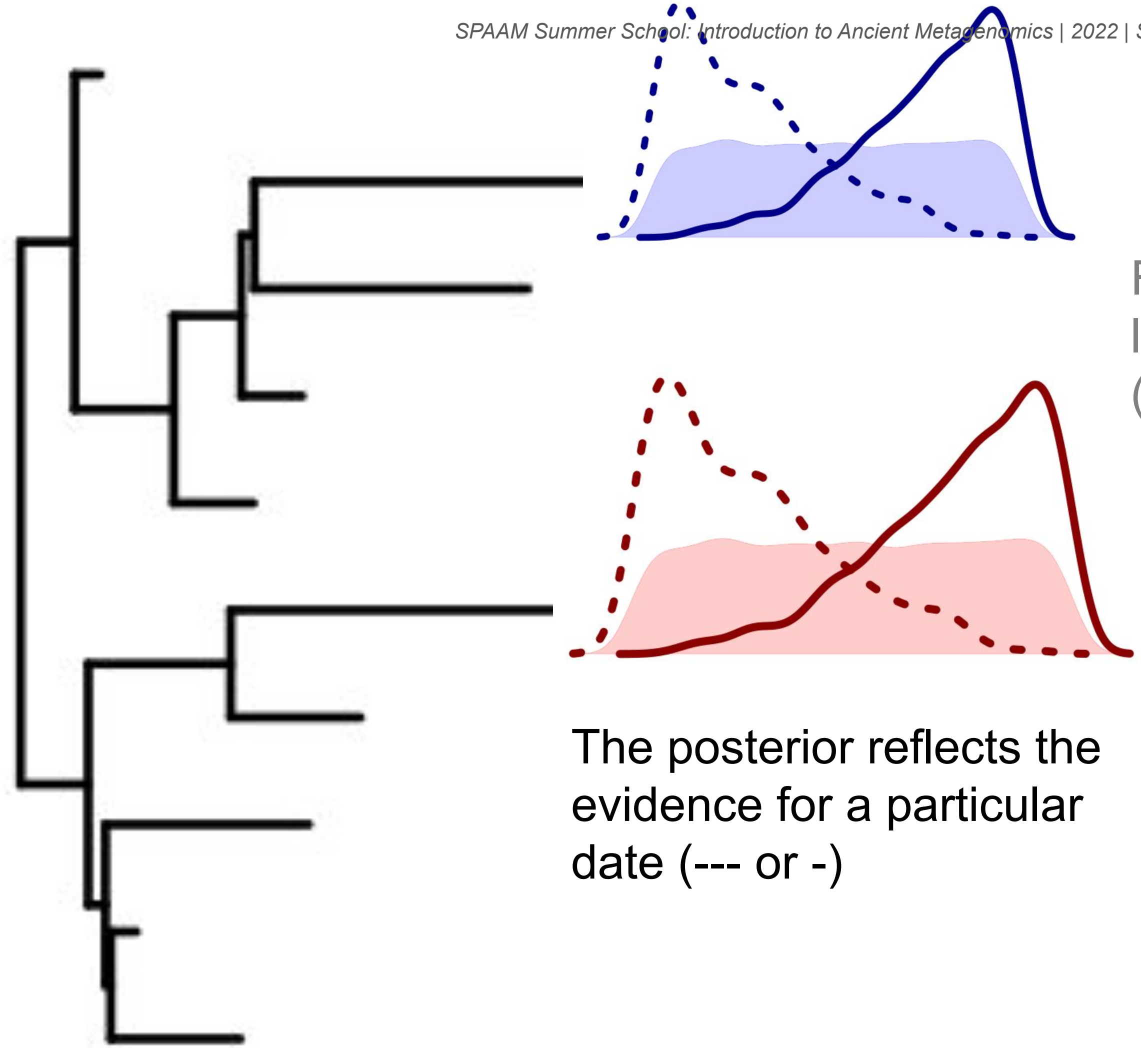
The molecular clock of *Hepatitis B Virus*





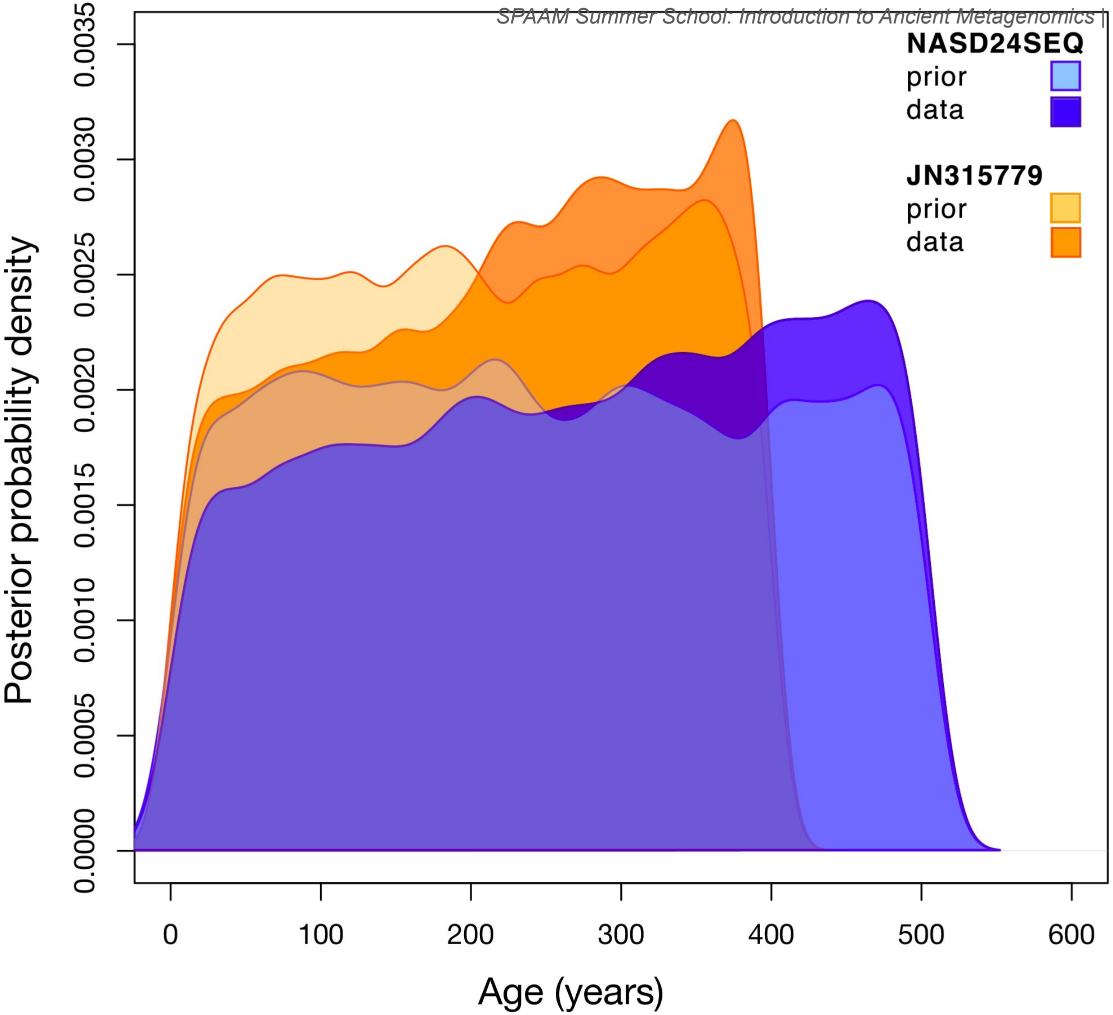
Modern or ancient?

From Patterson Ross et al 2018



For ancient samples use prior with low information content
(e.g. U(collection date, present).

The posterior reflects the evidence for a particular date (--- or -)



The data are not sufficiently informative to override the prior!

From Patterson Ross et al 2018

Thank you!

sduchene@unimelb.edu.au

NHMRC

Australian Government
Australian Research Council

Doherty Institute
A joint venture between the University of Melbourne and Melbourne Health

POSTERA CRESCAM LAUDE
THE UNIVERSITY OF MELBOURNE