Phylogenetics @ the Doherty 2021

Contributors

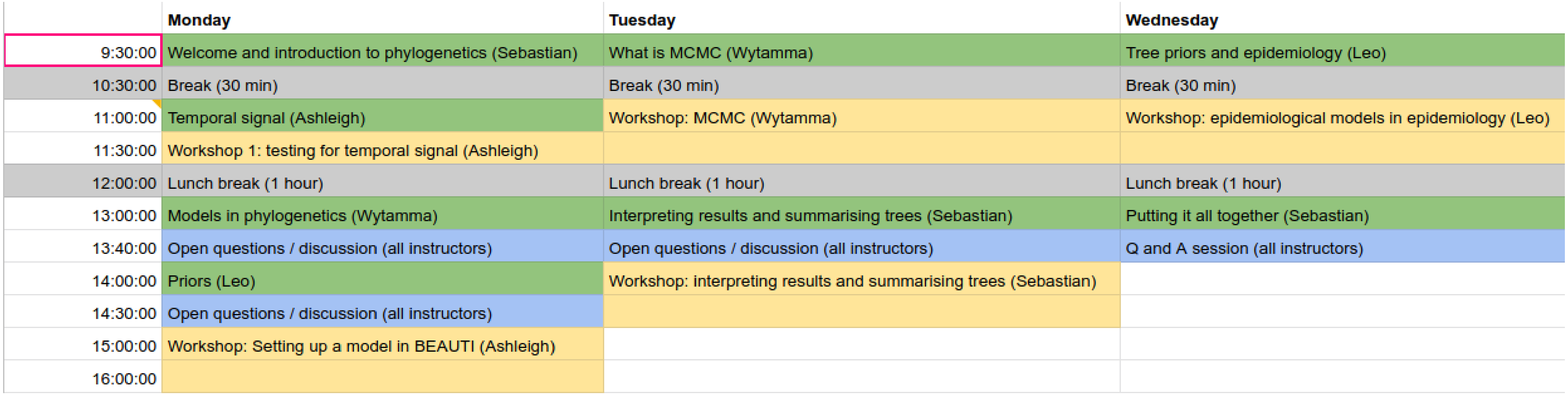
- ~26 workshops in 6 locations
- Several contributors over the years:

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

Jane Hawkey (Monash)
John-Sebastian Eden (Sydney Uni)
Remco Bouckaert (Melbourne Uni)

2021
Ashleigh Porter
Wytamma Wirth
Leo Featherstone
Sebastian Duchene

Workshop schedule



	Monday	Tuesday	Wednesday
9:30:00	Welcome and introduction to phylogenetics (Sebastian)	What is MCMC (Wytamma)	Tree priors and epidemiology (Leo)
10:30:00	Break (30 min)	Break (30 min)	Break (30 min)
11:00:00	Temporal signal (Ashleigh)	Workshop: MCMC (Wytamma)	Workshop: epidemiological models in epidemiology (Leo)
11:30:00	Workshop 1: testing for temporal signal (Ashleigh)		
12:00:00	Lunch break (1 hour)	Lunch break (1 hour)	Lunch break (1 hour)
13:00:00	Models in phylogenetics (Wytamma)	Interpreting results and summarising trees (Sebastian)	Putting it all together (Sebastian)
13:40:00	Open questions / discussion (all instructors)	Open questions / discussion (all instructors)	Q and A session (all instructors)
14:00:00	Priors (Leo)	Workshop: interpreting results and summarising trees (Sebastian)	
14:30:00	Open questions / discussion (all instructors)		
15:00:00	Workshop: Setting up a model in BEAUTI (Ashleigh)		
16:00:00			

Activities and learning

- Break out rooms
- Questions in chat
- Video recordings

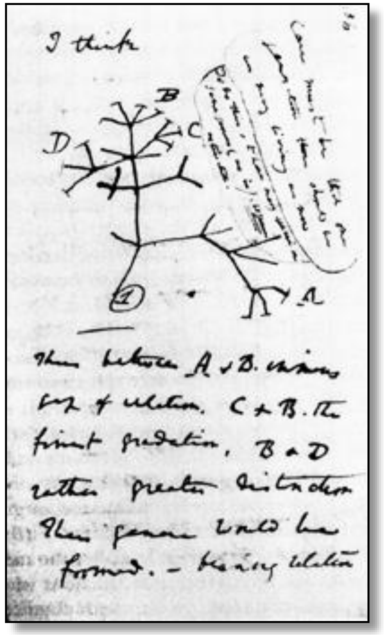
- Prizes:
  - Best question
  - Most questions
  - Best zoom animal
  - Spot the frog

Vouchers for Readings

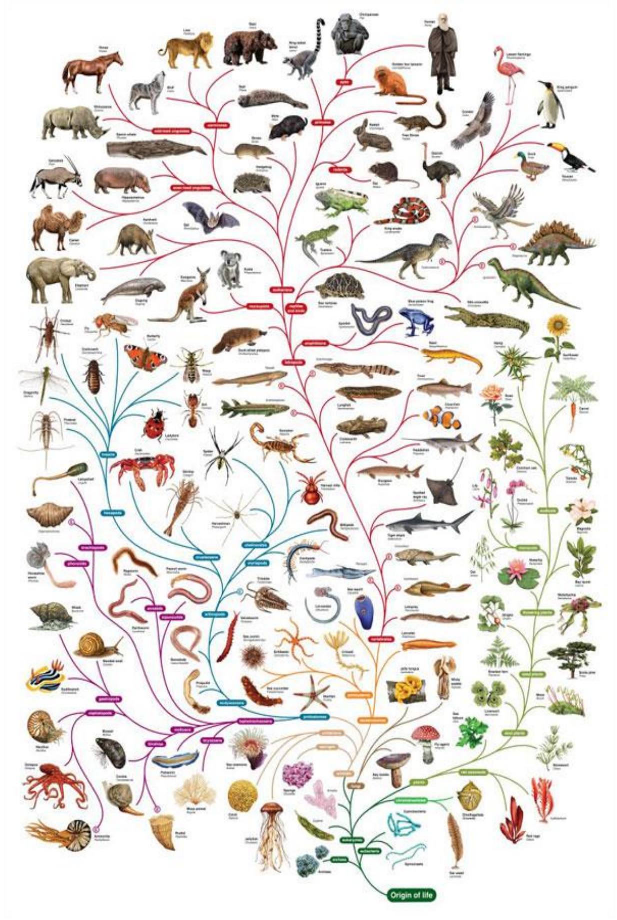
Lecture 1: Introduction

What is a phylogenetic tree?

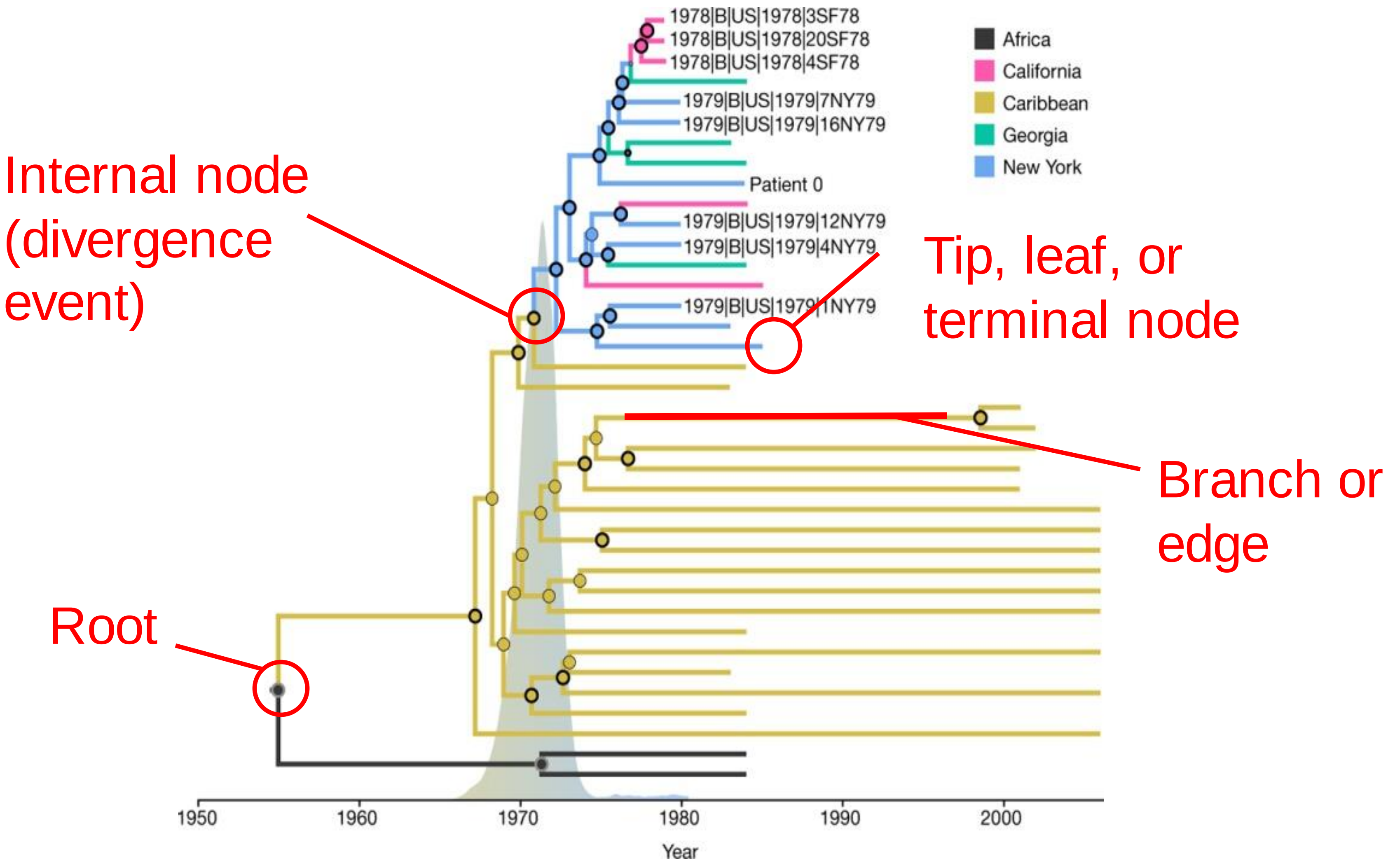
The phylogeny refers to the true evolutionary relationships among a set of organisms



Then 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



Internal node (divergence event)

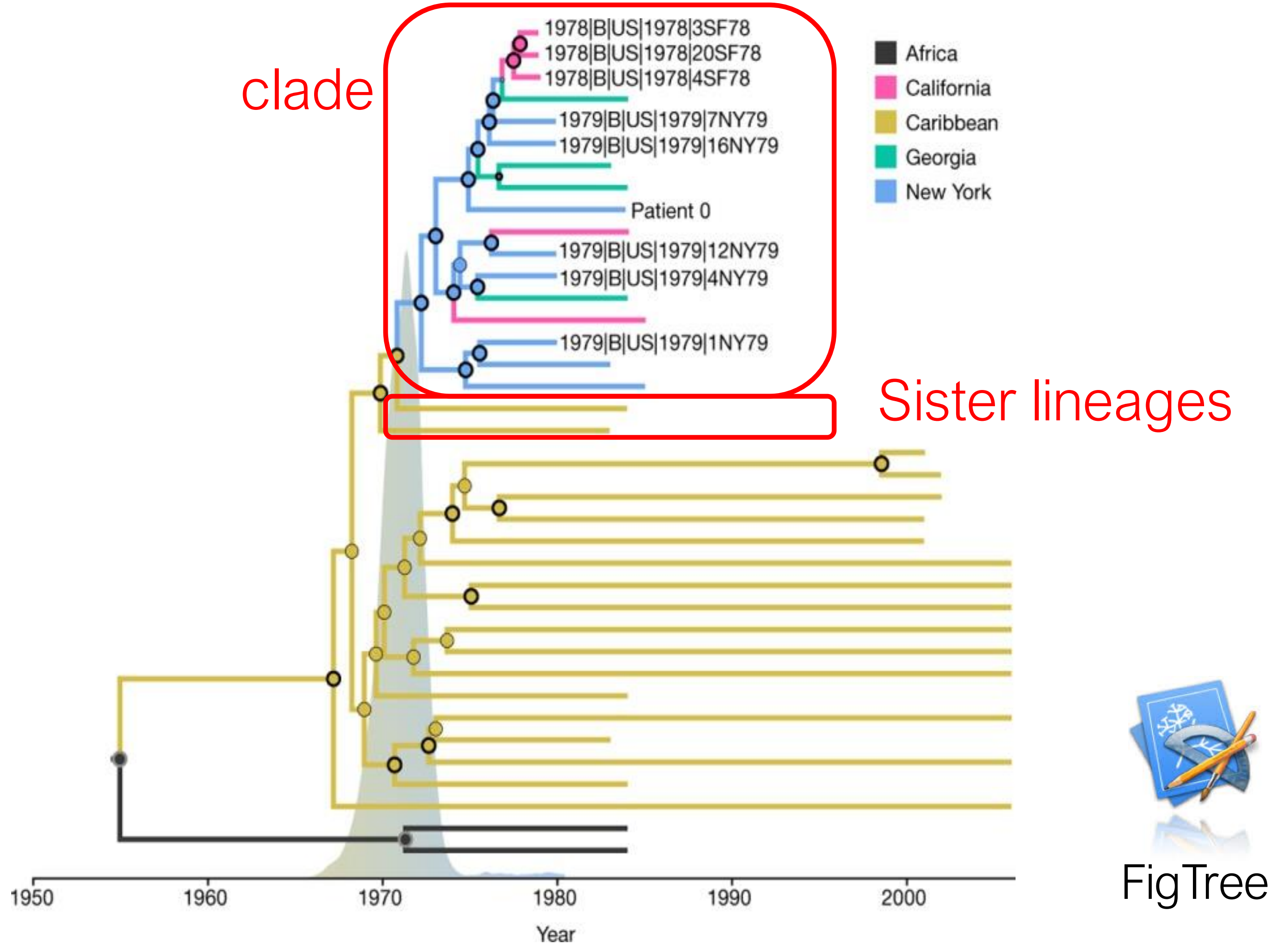
Tip, leaf, or terminal node

Branch or edge

Root

From Worobey et al. 2016 *Nature*

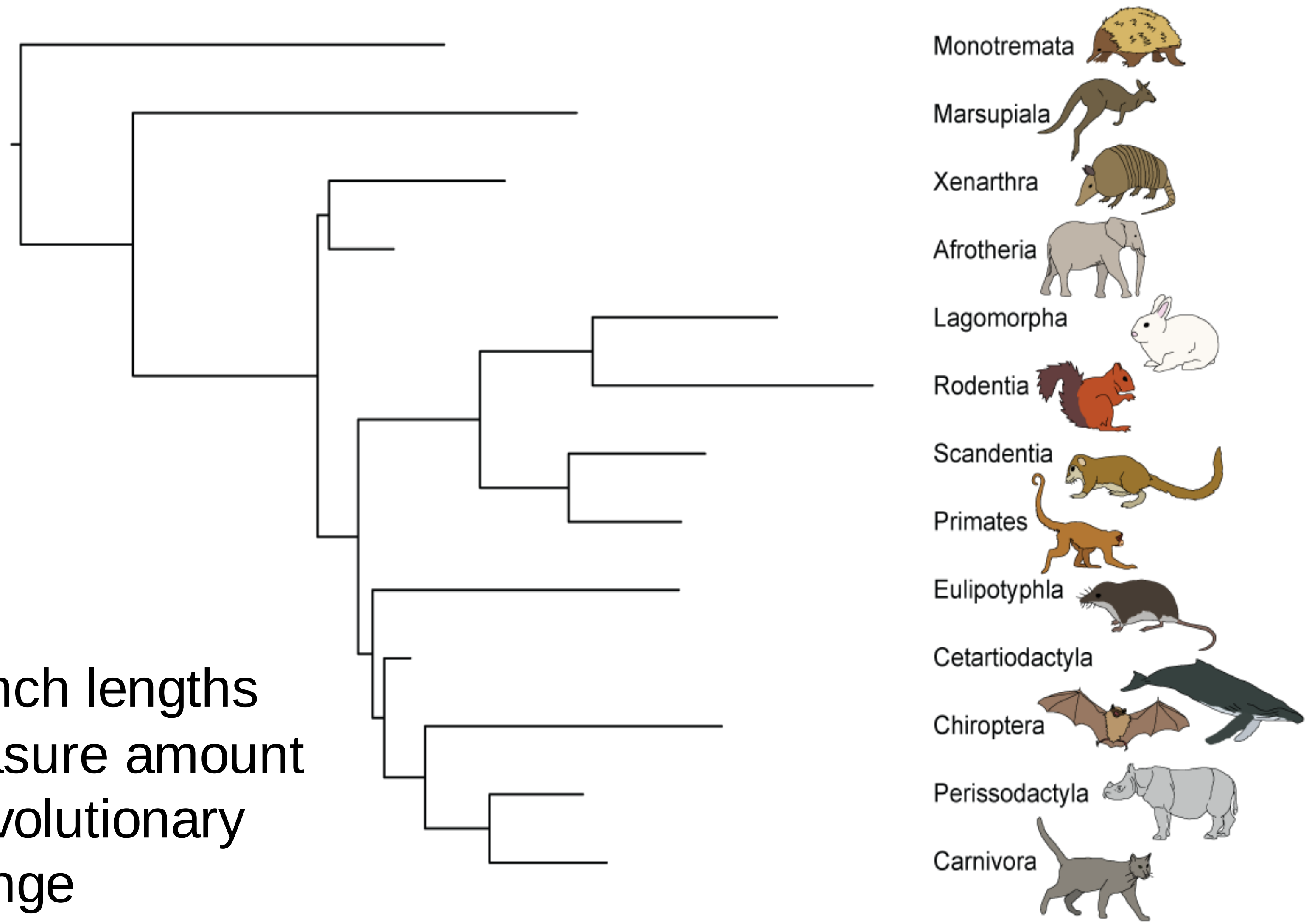
Phylogenetic trees



FigTree

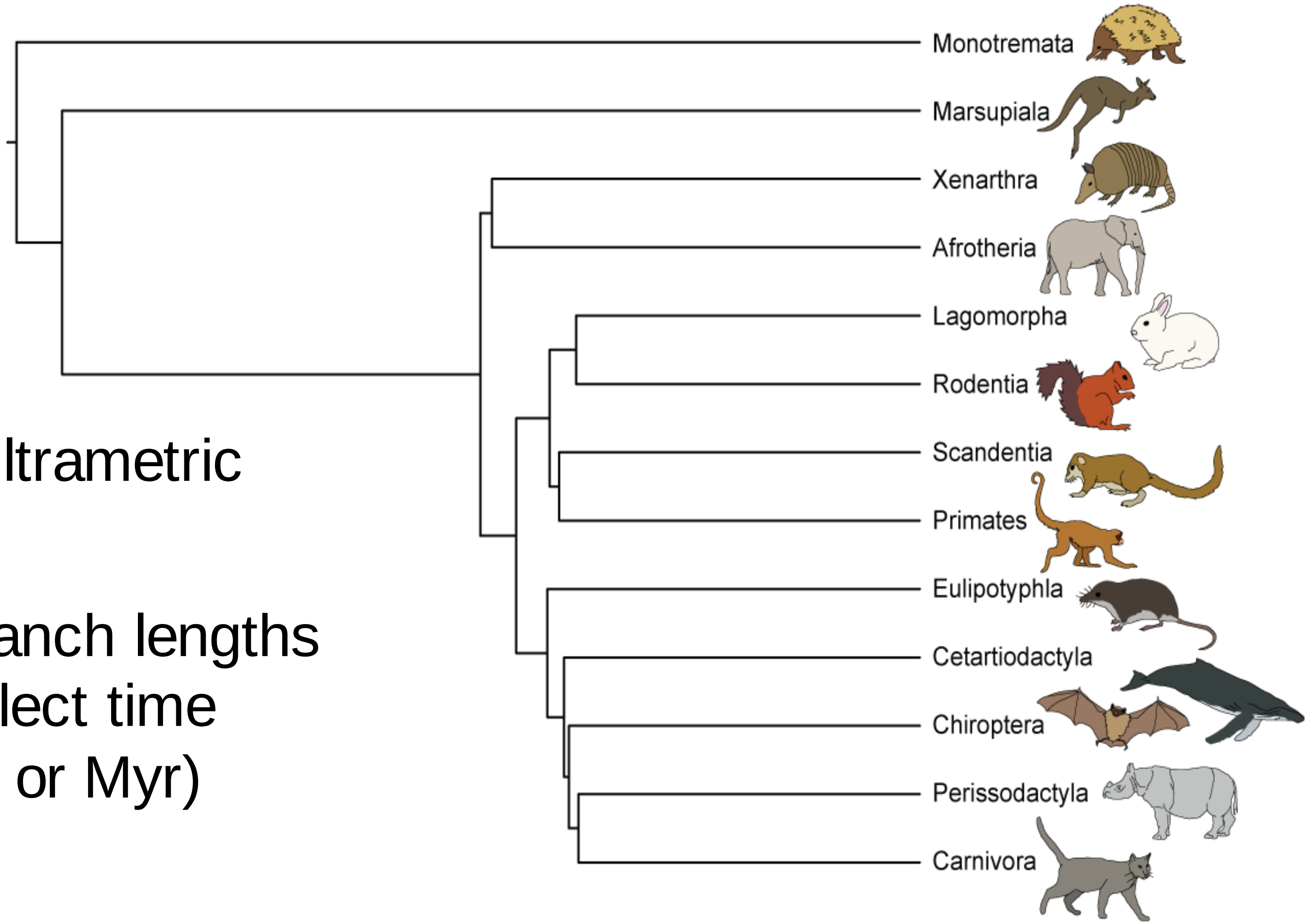
From Worobey et al. 2016 *Nature*

Phylogenetic trees: Phylogram



Branch lengths measure amount of evolutionary change

Phylogenetic trees: Chronogram



Ultrametric

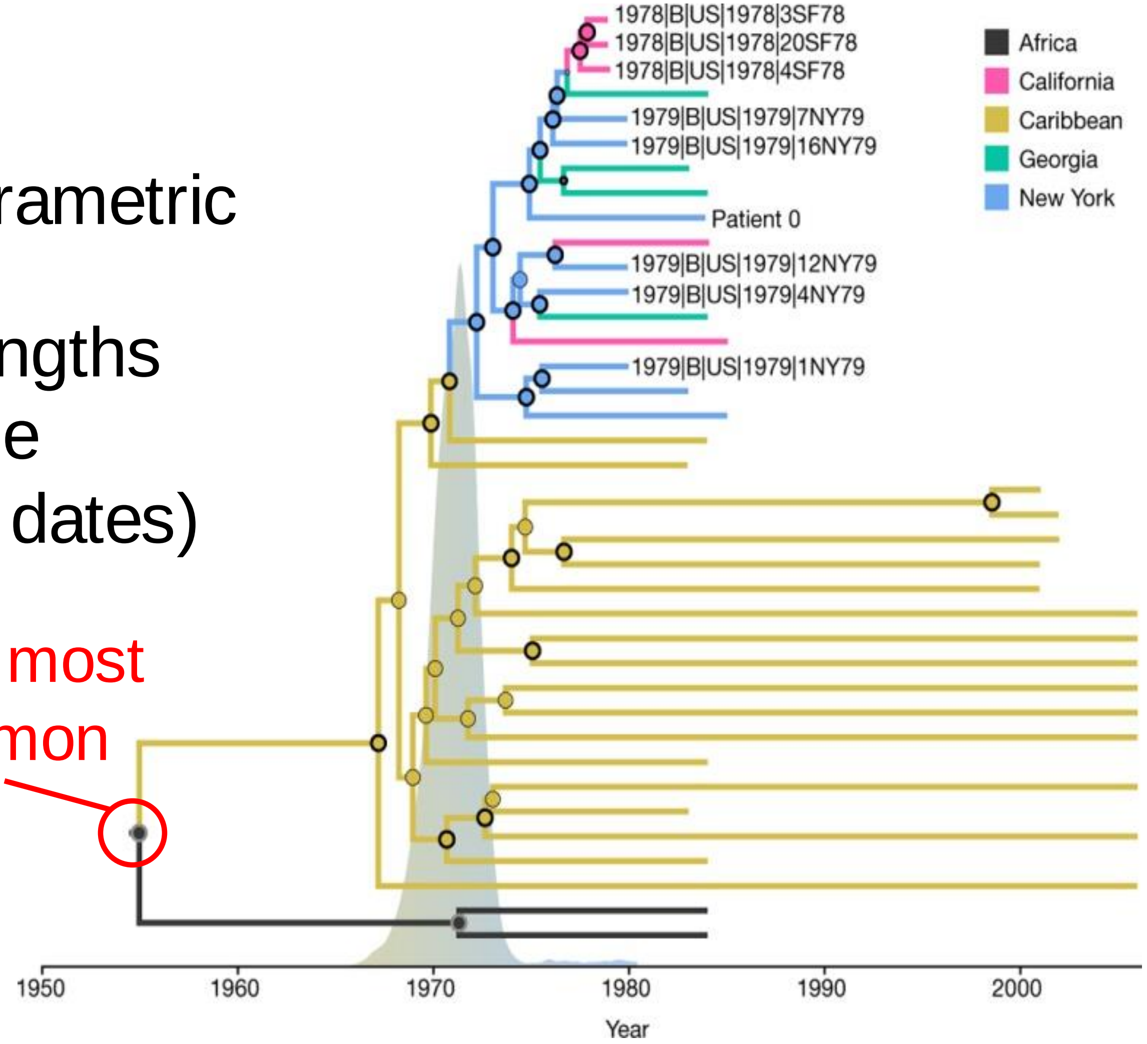
Branch lengths
reflect time
(yr or Myr)

Phylogenetic trees: Chronograms

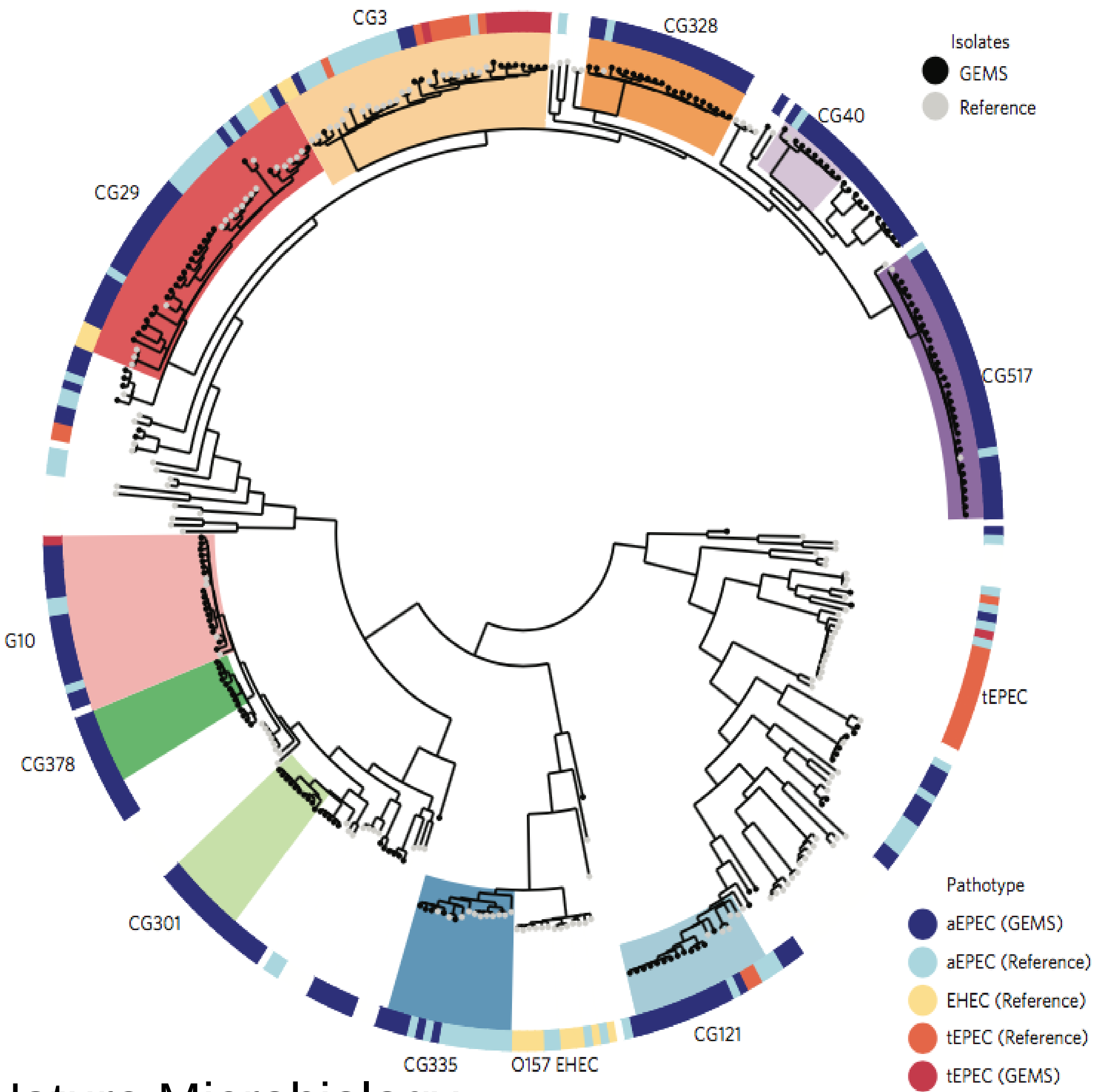
Non-ultrametric

Branch lengths reflect time (calendar dates)

Time to the most recent common ancestor



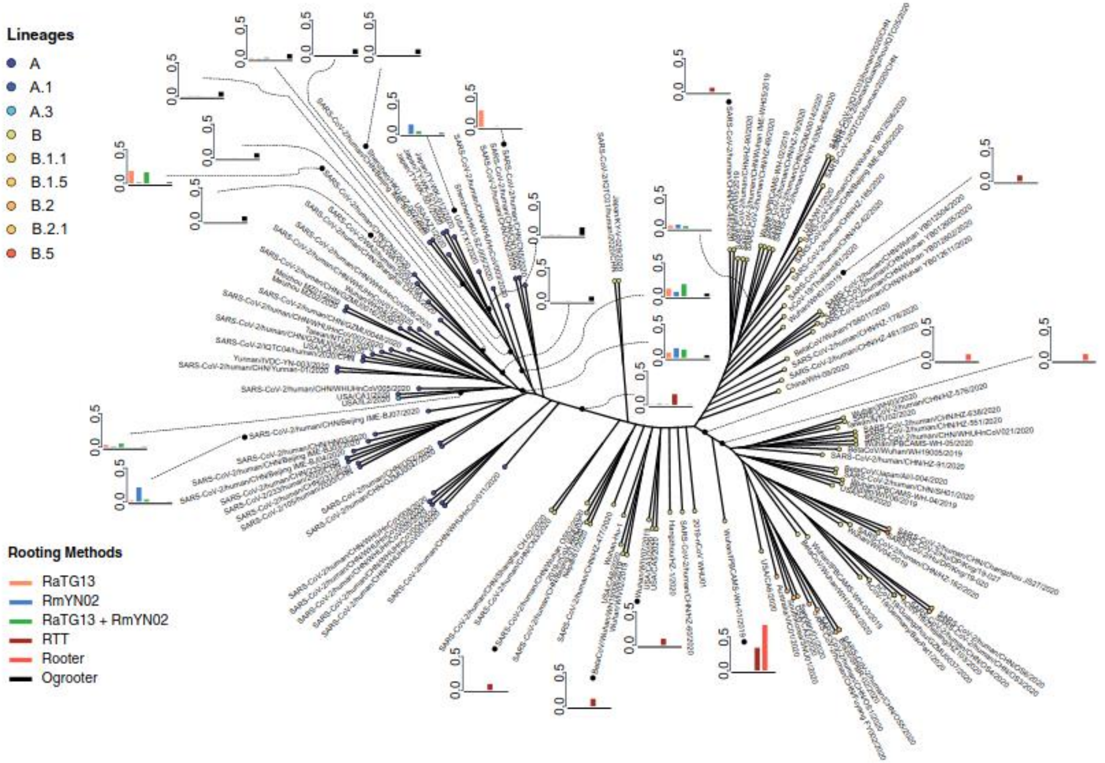
Phylogenetic trees: Circular



From Ingle et al. 2016 Nature Microbiology

Phylogenetic trees: Unrooted

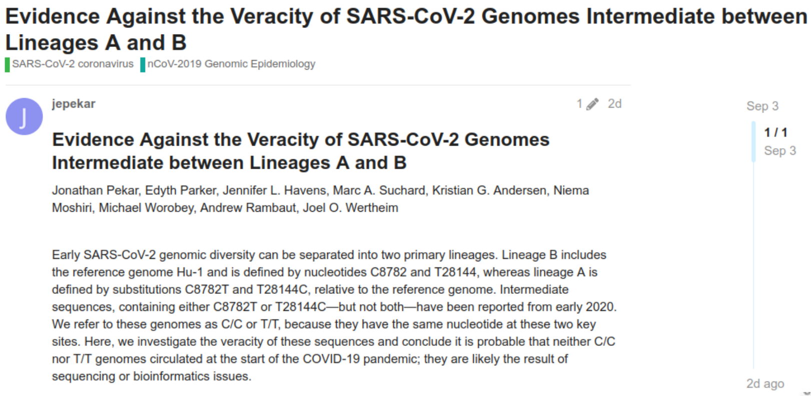
- Position of root is unknown
- Branch lengths usually represent amount of genetic change (substitutions/site)



From Pipes et al. MBE

Phylogenetic trees: Unrooted

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



Evidence Against the Veracity of SARS-CoV-2 Genomes Intermediate between Lineages A and B

SARS-CoV-2 coronavirus | nCoV-2019 Genomic Epidemiology

J jepekar 1 ✎ 2d

Evidence Against the Veracity of SARS-CoV-2 Genomes Intermediate between Lineages A and B

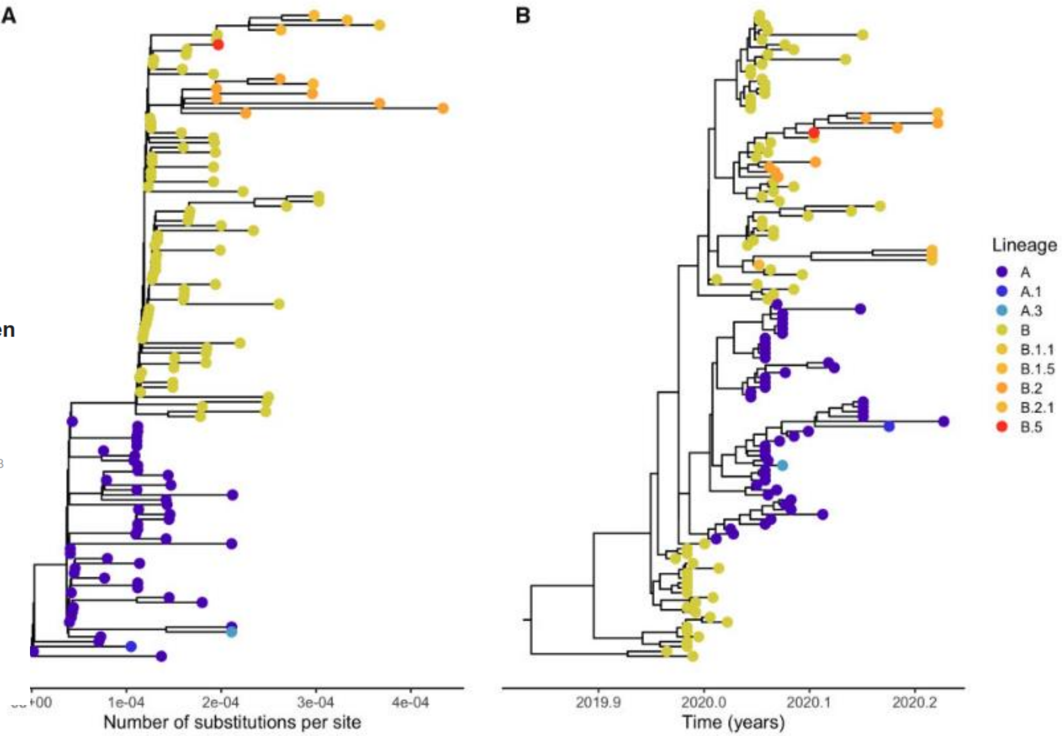
Jonathan Pekar, Edyth Parker, Jennifer L. Havens, Marc A. Suchard, Kristian G. Andersen, Niema Moshiri, Michael Worobey, Andrew Rambaut, Joel O. Wertheim

Early SARS-CoV-2 genomic diversity can be separated into two primary lineages. Lineage B includes the reference genome Hu-1 and is defined by nucleotides C8782 and T28144, whereas lineage A is defined by substitutions C8782T and T28144C, relative to the reference genome. Intermediate sequences, containing either C8782T or T28144C—but not both—have been reported from early 2020. We refer to these genomes as C/C or T/T, because they have the same nucleotide at these two key sites. Here, we investigate the veracity of these sequences and conclude it is probable that neither C/C nor T/T genomes circulated at the start of the COVID-19 pandemic; they are likely the result of sequencing or bioinformatics issues.

Sep 3

1 / 1
Sep 3

2d ago



From Pipes et al. MBE

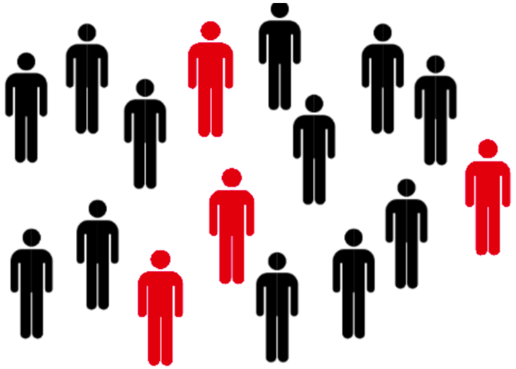
Concept summary

- Phylogenetic trees have parts (e.g. tip, node, root).
- Chronogram vs phylogram.
- Trees must be rooted to interpret time.

Inferring phylogenetic trees:

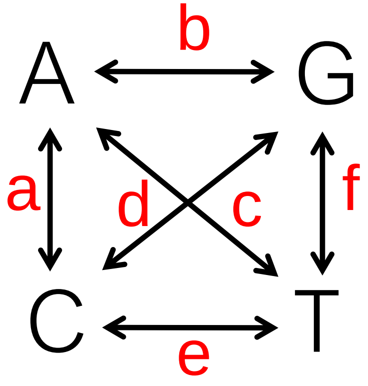
Maximum likelihood and Bayesian inference

Estimating phylogenetic trees from molecular data



```
Sample 1 - AAAATCGCG
Sample 2 - AAAGATGCG
Sample 3 - AAAACCGCG
Sample 4 - AAAACCGTG
```

Rate Matrix



Base Frequencies

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

Site Rates

+ I + G

Some common substitution models: JC, GTR+I+G, HKY

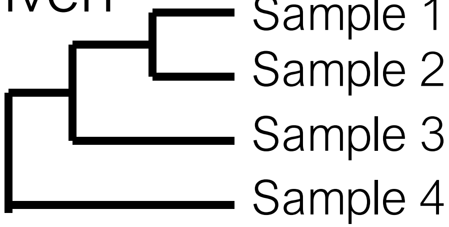
Maximum likelihood

Likelihood of hypothesis H =

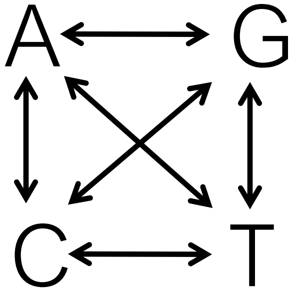
$$\boxed{P(D|H)}$$

the probability of the data, given the hypothesis

Given



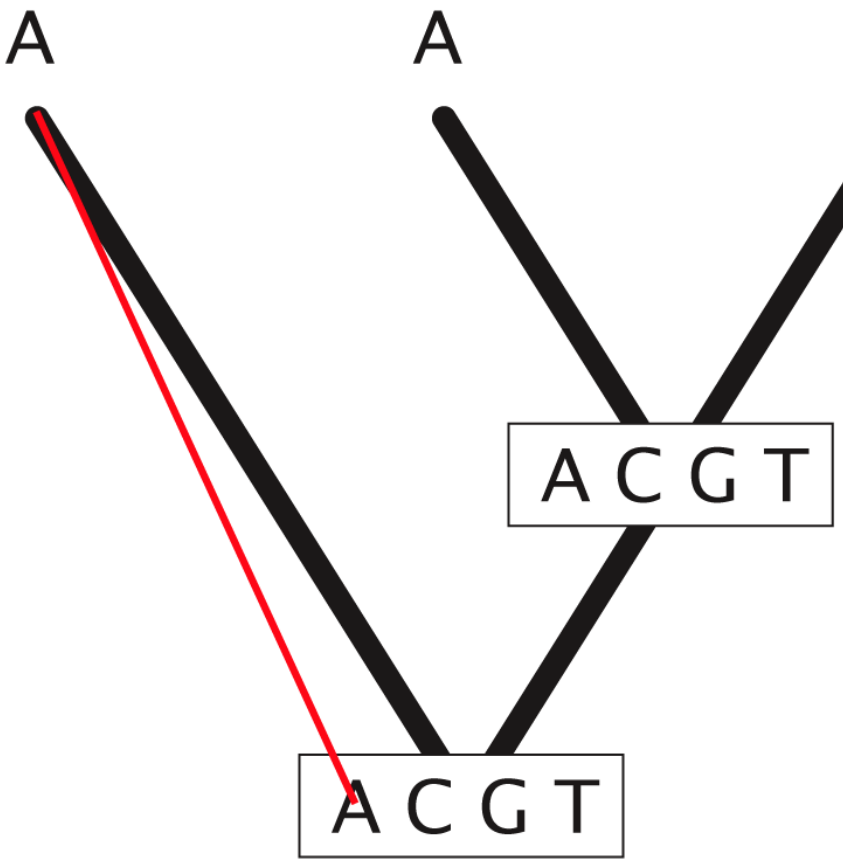
+



Probability of?

```
Sample1  CGTTAGTACACT
Sample2  CGATAGTTCACT
Sample3  CGTTAGTTTACC
Sample4  CATTGGTTTACT
```

Maximum likelihood

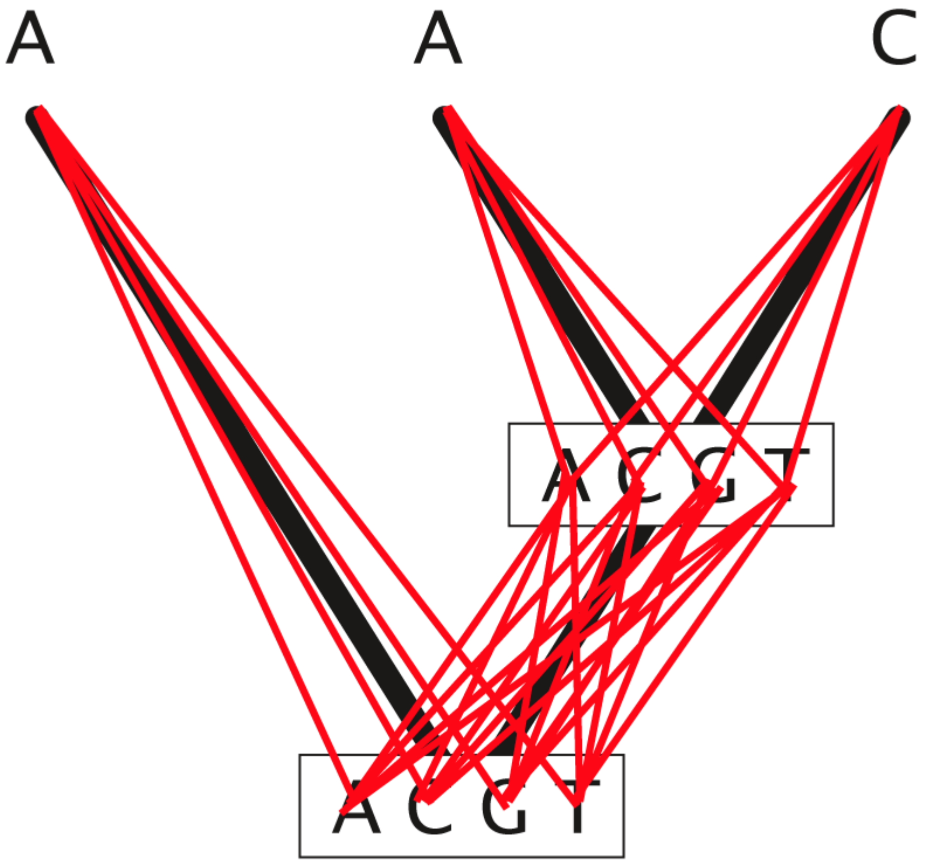


to

from

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

Maximum likelihood



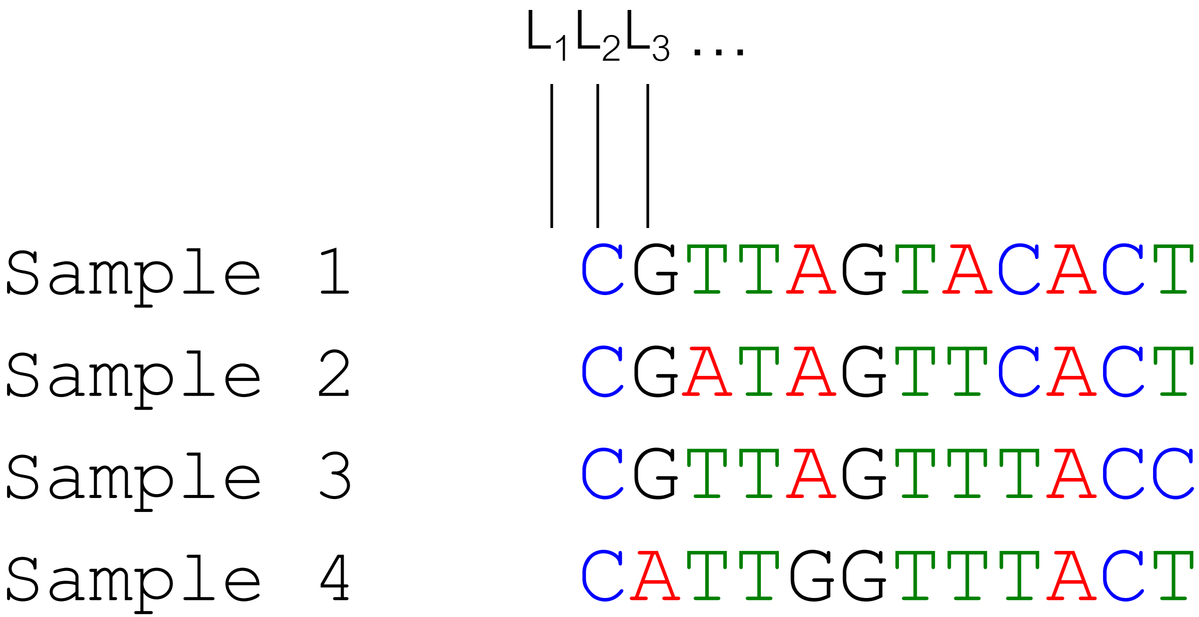
to

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

Likelihood = sum of all possible scenarios

Maximum likelihood

Likelihood is multiplied across sites



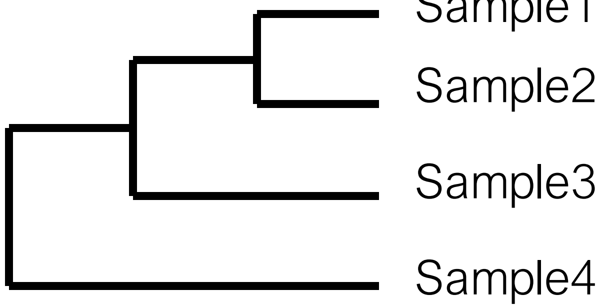
$L_1 L_2 L_3$ …

```
Sample 1     CGTTAGTACACT
Sample 2     CGATAGTTCACT
Sample 3     CGTTAGTTTACC
Sample 4     CATTGGTTTACT
```

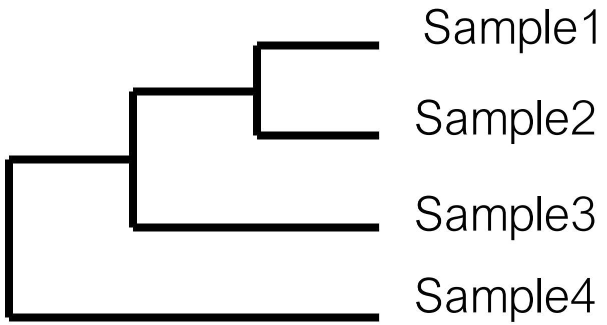
Likelihood values are very small!

Maximum likelihood

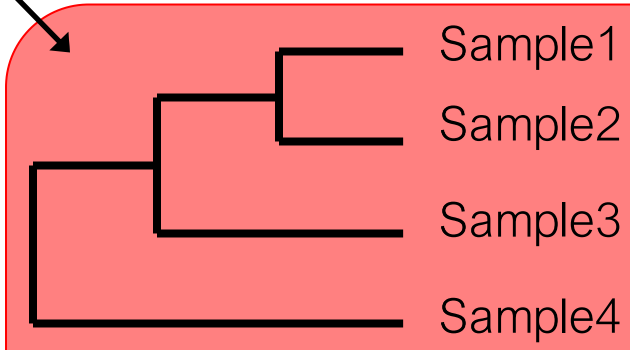
Sample 1	CGTTAGTACACT
Sample 2	CGATAGTTCACT
Sample 3	CGTTAGTTTACC
Sample 4	CATTGGTTTACT



lnL = -1203.83



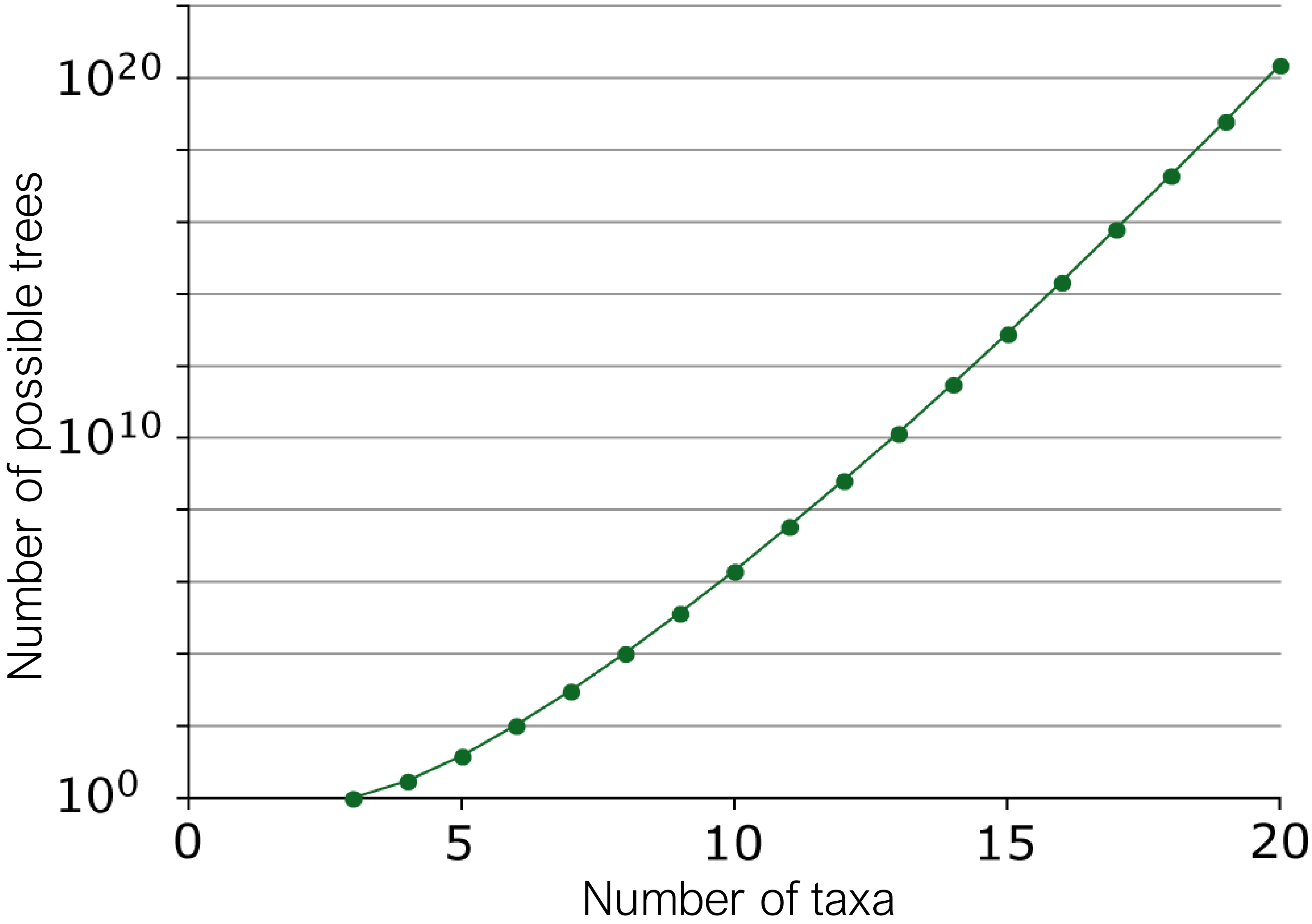
lnL = -1241.47



lnL = -908.58

Software: PhyML, RAxML, IQtree

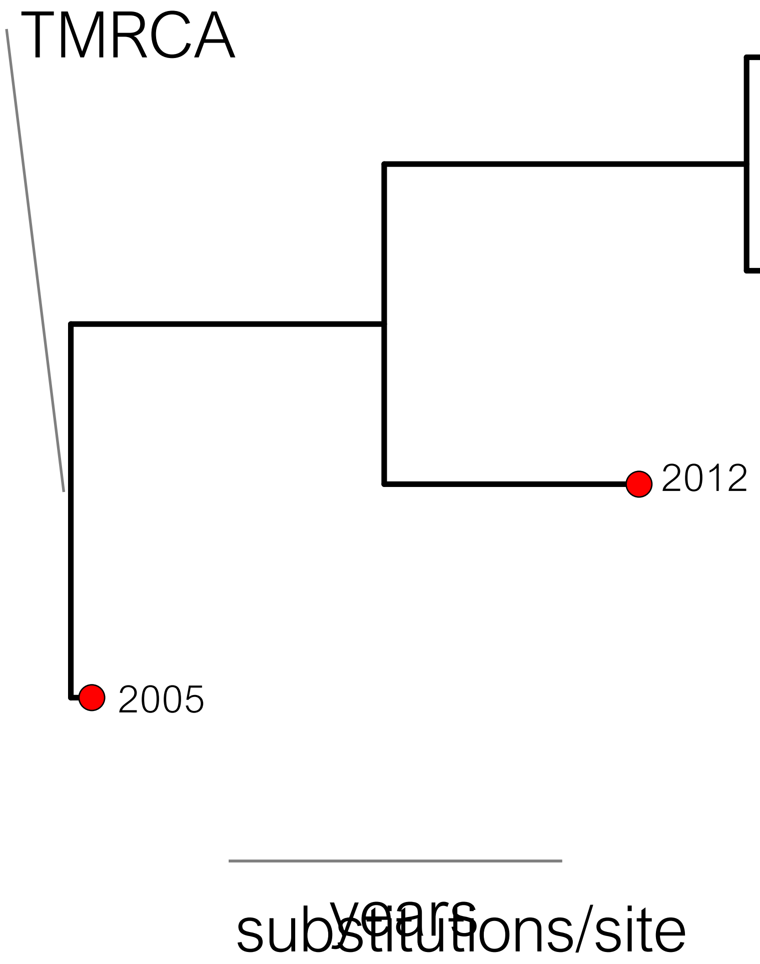
Searching tree space



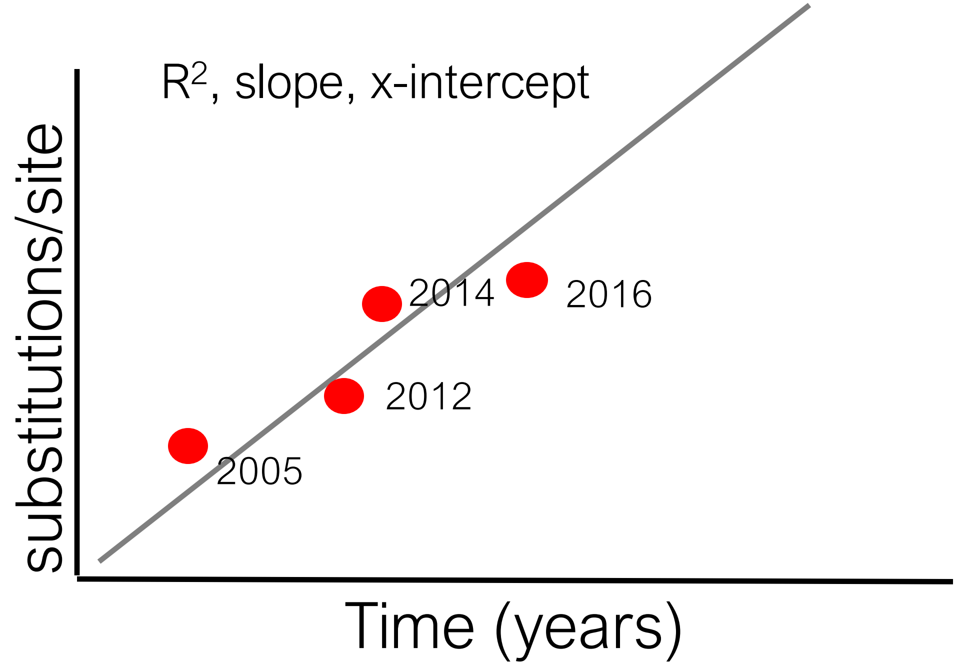
Maximum likelihood

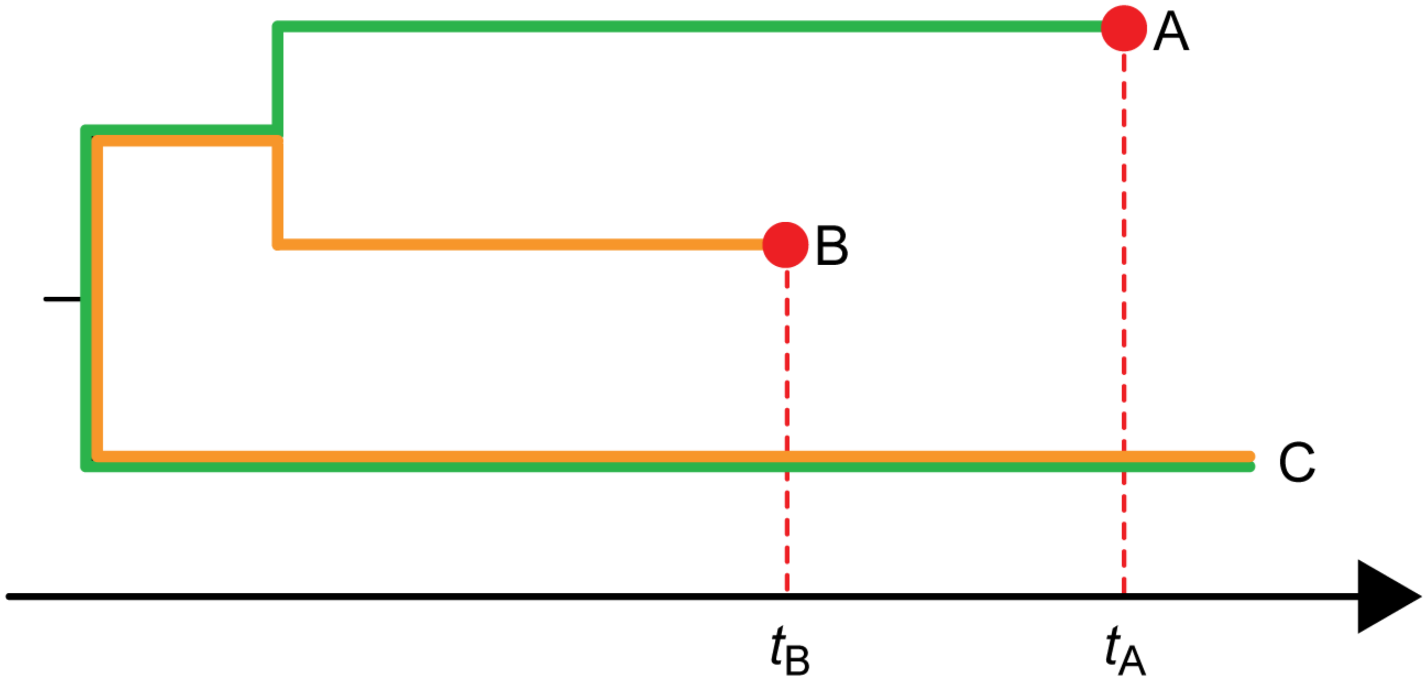
- Single estimate of phylogenetic tree and parameters (MLE).
- Use heuristics to search tree space.
- Use indirect methods to obtain uncertainty (e.g. bootstrapping).
- Additional steps for estimating rates and times.

The molecular clock



TempEst





Rambaut (2000) Bioinformatics

Concept summary

- Likelihood based inference (ML and BI) require a subst. model.
- The likelihood is the probability of observing a data (sites) under a model and tree.
- Our goal is to find the *best* tree and parameters.
- ML typically returns a *phylogram*.
- We need a molecular clock to infer a *chronogram*.

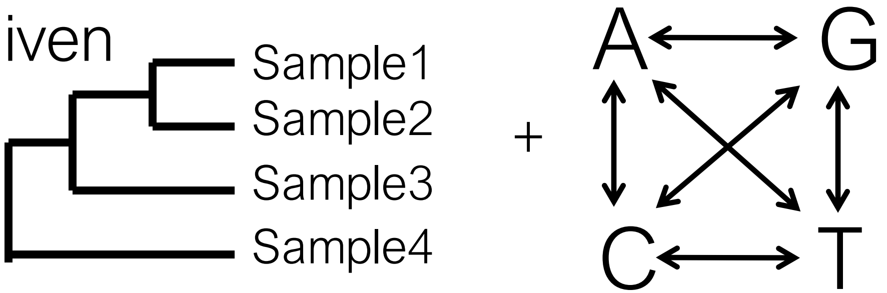
Inferring phylogenetic trees:

Maximum likelihood and Bayesian inference

Bayesian versus likelihood

Maximum likelihood

Given



→

Probability of?

Sample1	CGTTAGTACACT
Sample2	CGATAGTTCACT
Sample3	CGTTAGTTTACC
Sample4	CATTGGTTTACT

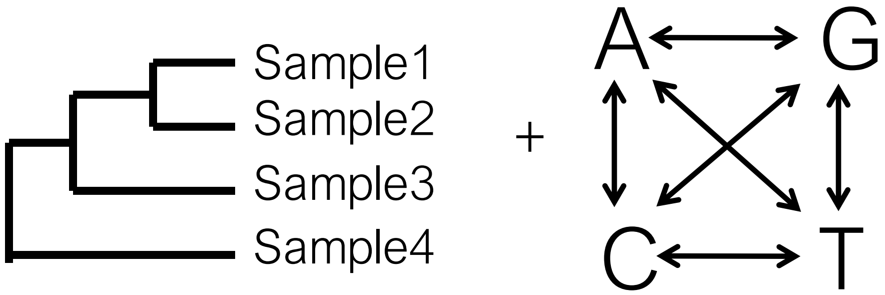
Bayesian inference

Given

Sample1	CGTTAGTACACT
Sample2	CGATAGTTCACT
Sample3	CGTTAGTTTACC
Sample4	CATTGGTTTACT

→

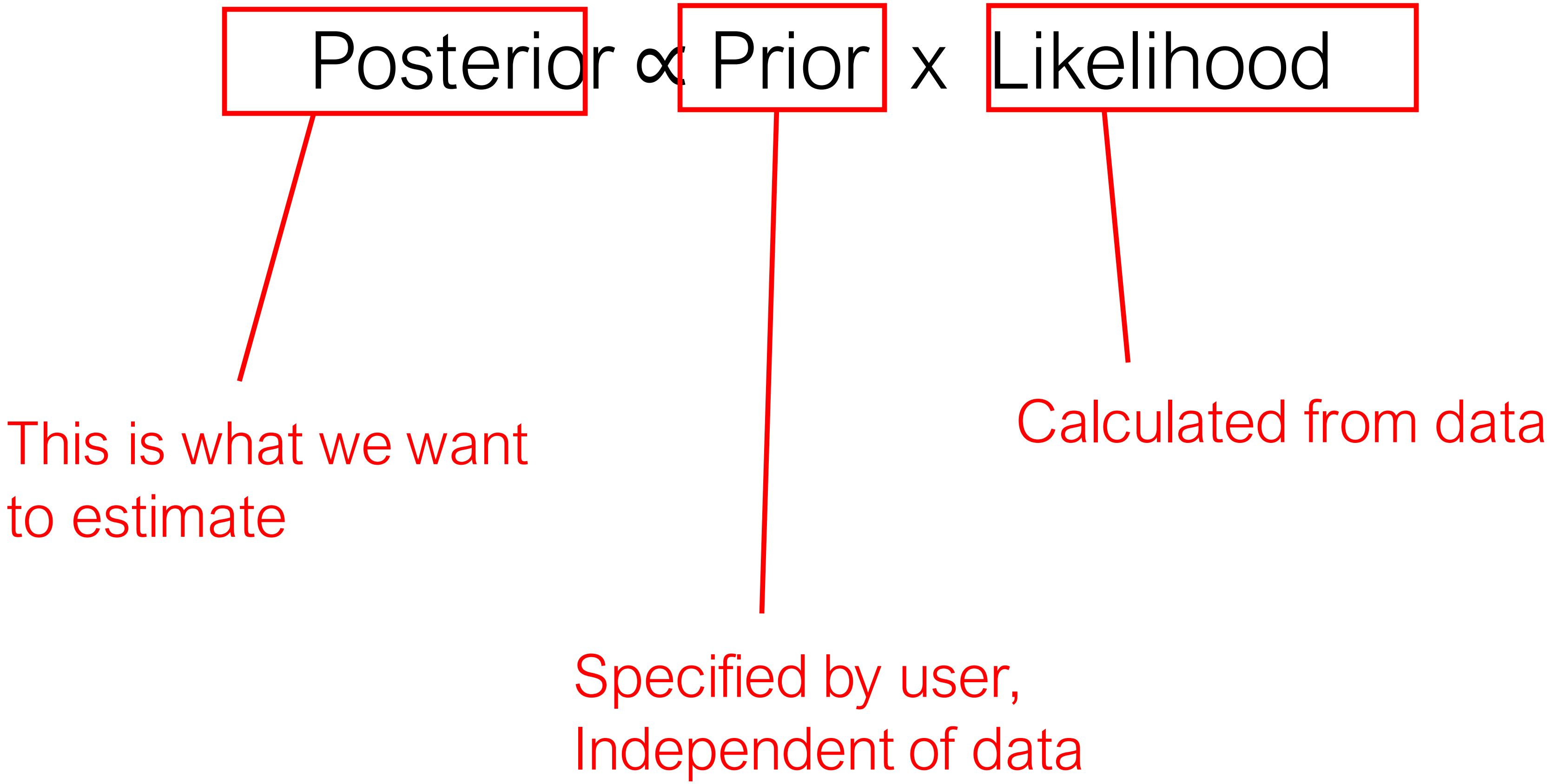
Probability of?



The Bayesian paradigm

- Parameters have distributions
- Before the data are observed, each parameter has a prior distribution
- The likelihood of the data is computed
- The prior distribution is combined with the likelihood to yield the posterior distribution

Bayesian inference



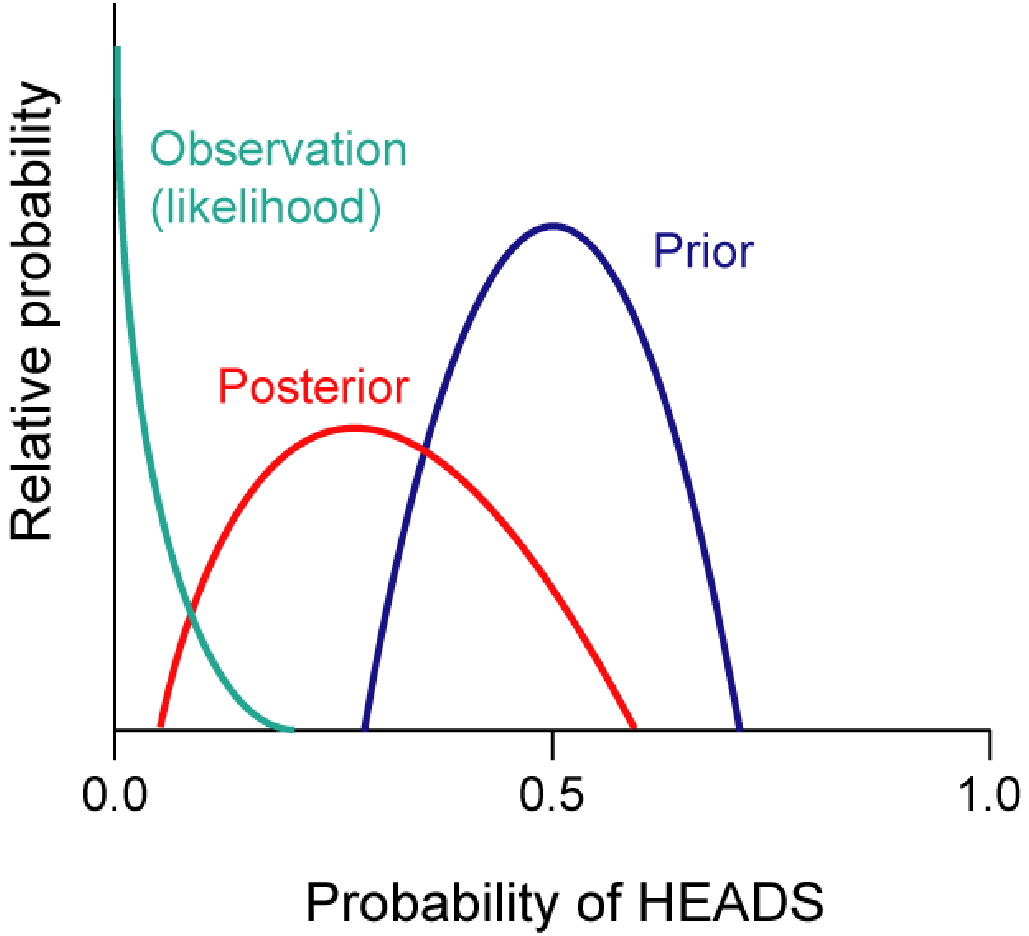
Posterior $\propto$ Prior x Likelihood

This is what we want to estimate

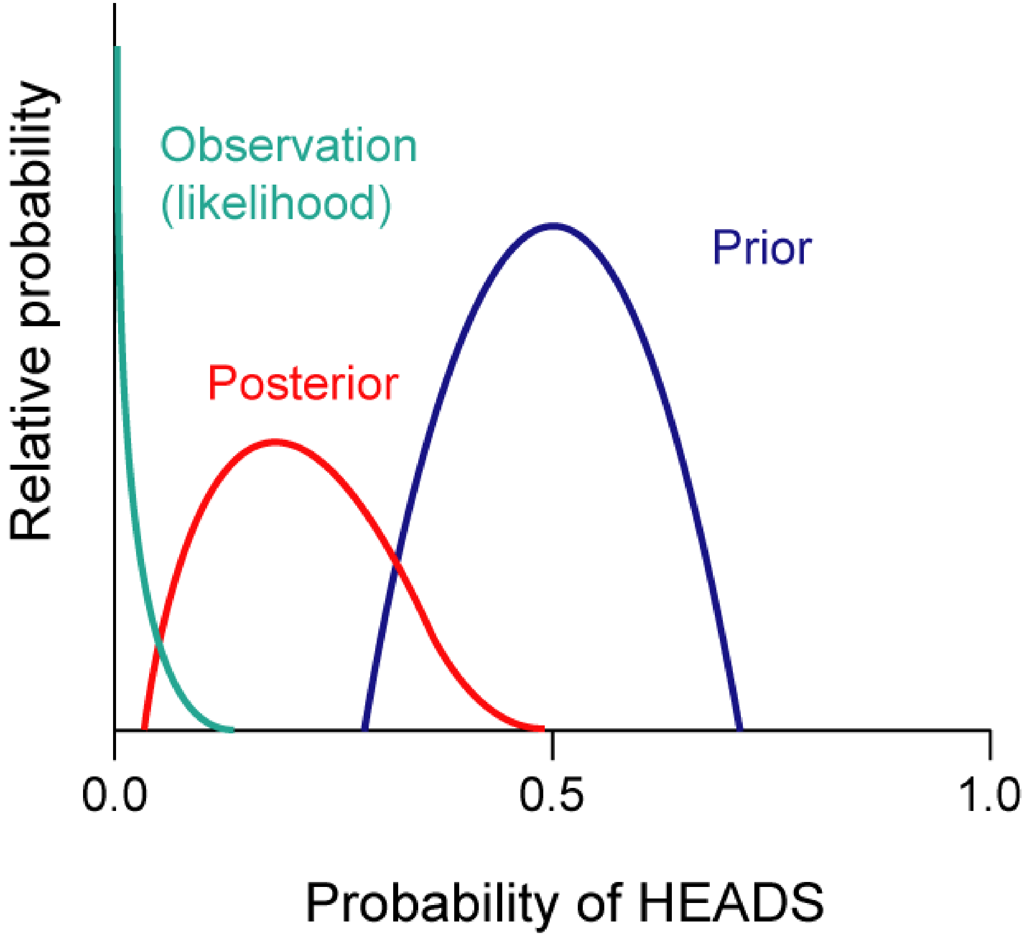
Specified by user, Independent of data

Calculated from data

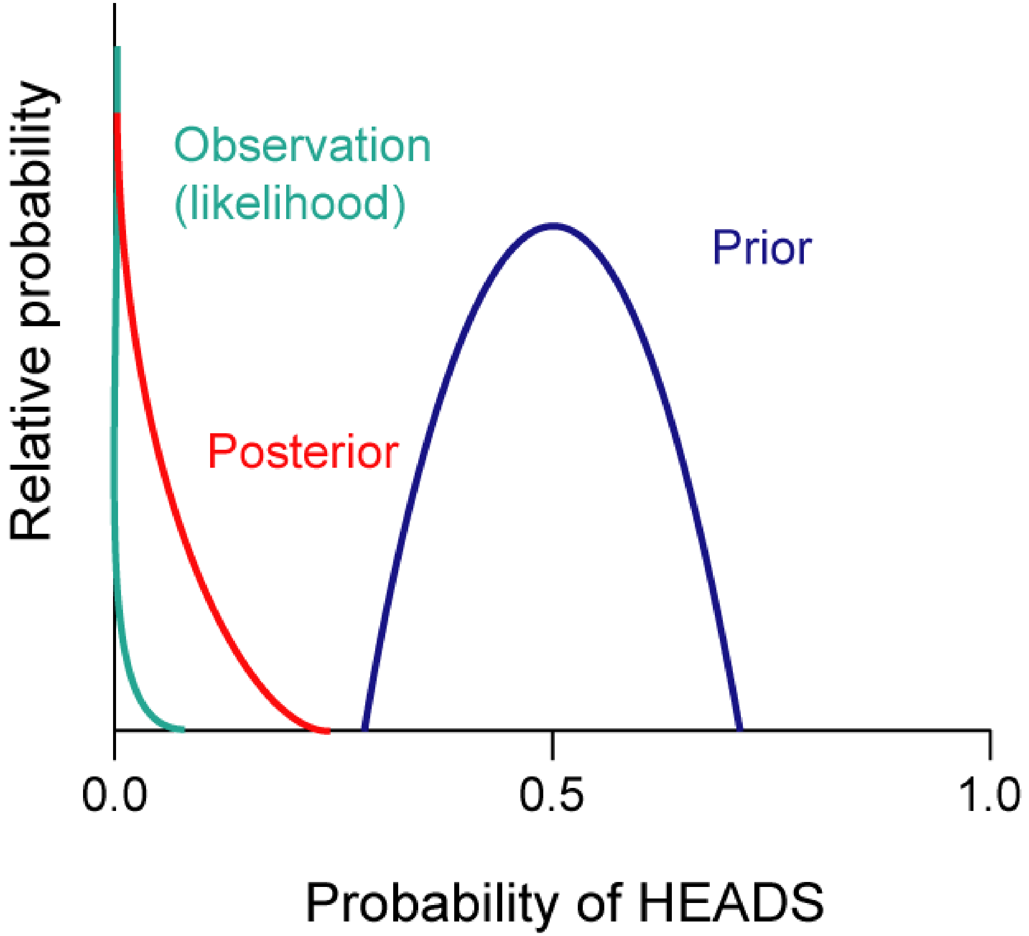
Coin toss example



Coin toss example



Coin toss example

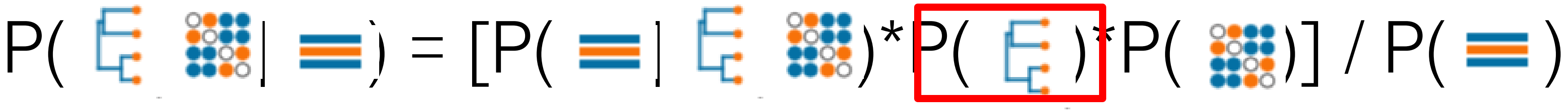


Parameters

Phylogenetic tree
(chronogram or phylogram)

Substitution model
parameters

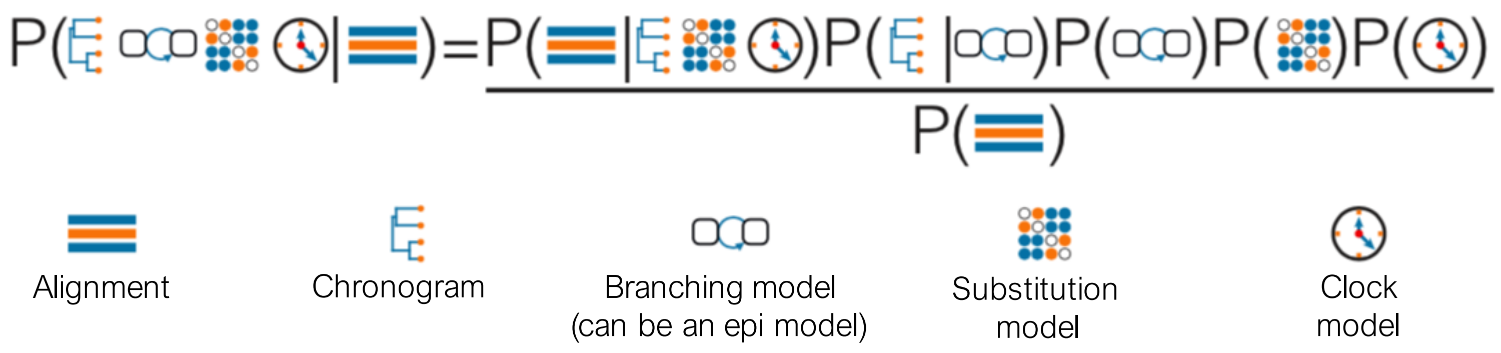
Evolutionary rates and time



Posterior = (Likelihood * prior) / marginal likelihood

For the tree prior we can use an epidemiological process → chronograms

We then need to multiply branch lengths by a clock rate to generate phylograms to obtain the likelihood –we treat branch lengths as the product of rates and times



Alignment

Chronogram

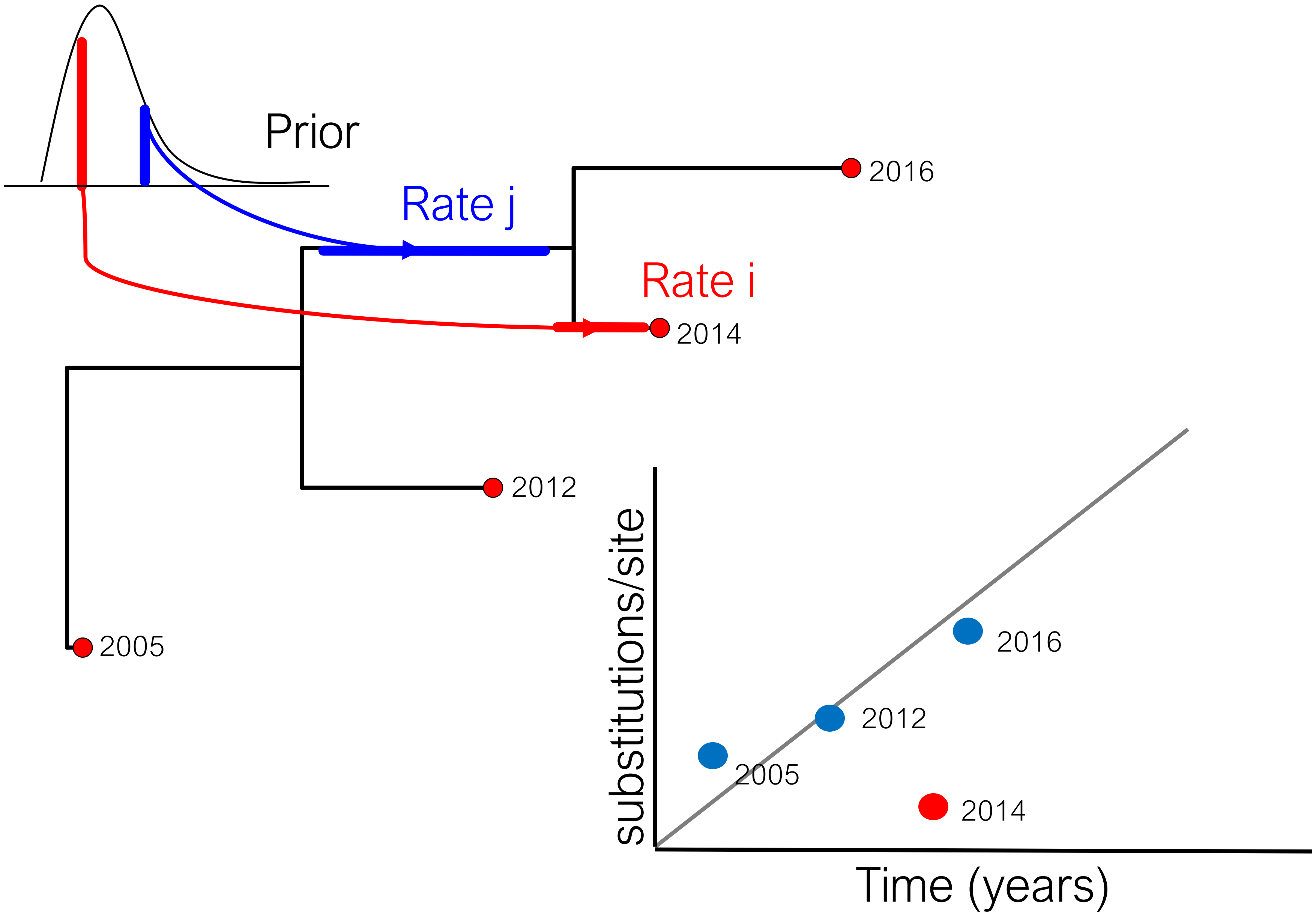
Branching model
(can be an epi model)

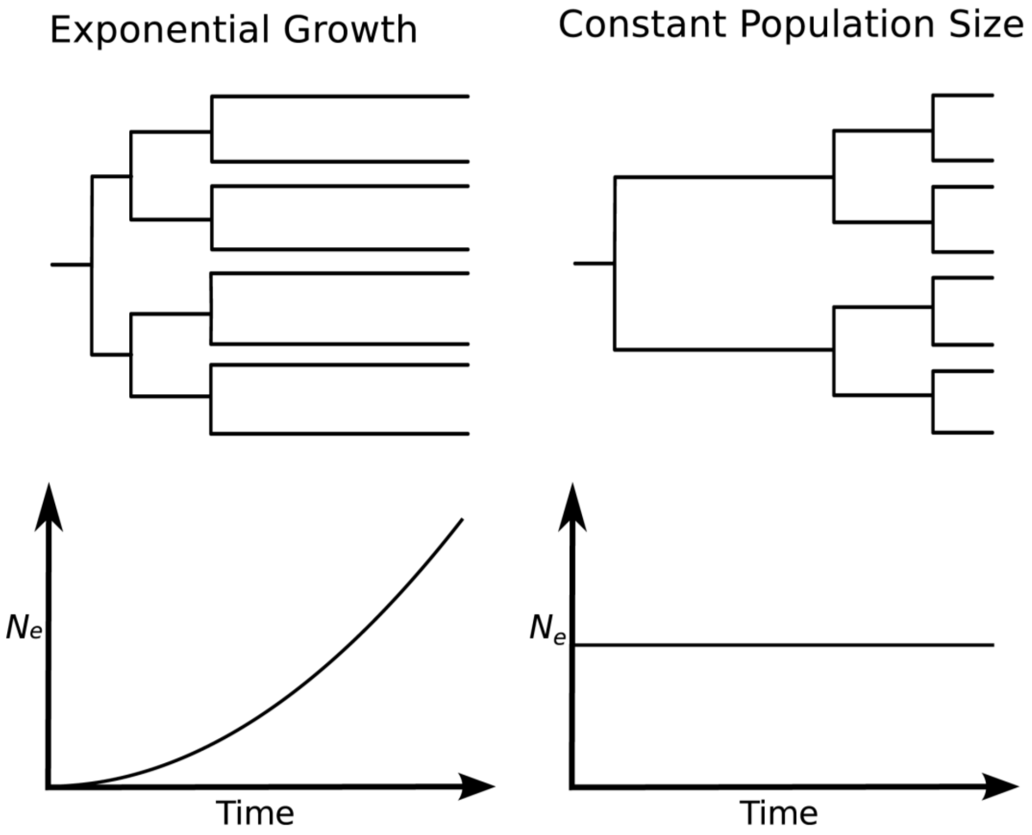
Substitution
model

Clock
model

The posterior now has two more terms (a clock model and a branching model) to specify more sophisticated models.

From: Du Plessis and Stadler 2015


Prior
Rate j
Rate i
2016
2014
2012
2005
substitutions/site
Time (years)
2016
2012
2005
2014



From: Volz et al. 2013

Concept summary

- BI requires prior information on all parameters. (we can use less informative priors)
- The goal is to obtain uncertainty in all estimates, not the single best tree and parameters (natural product of BI).
- We can specify more complex models than in ML.