Relaxed Bayesian phylogenetics

Alexei Drummond, alexei@cs.auckland.ac.nz
University of Auckland

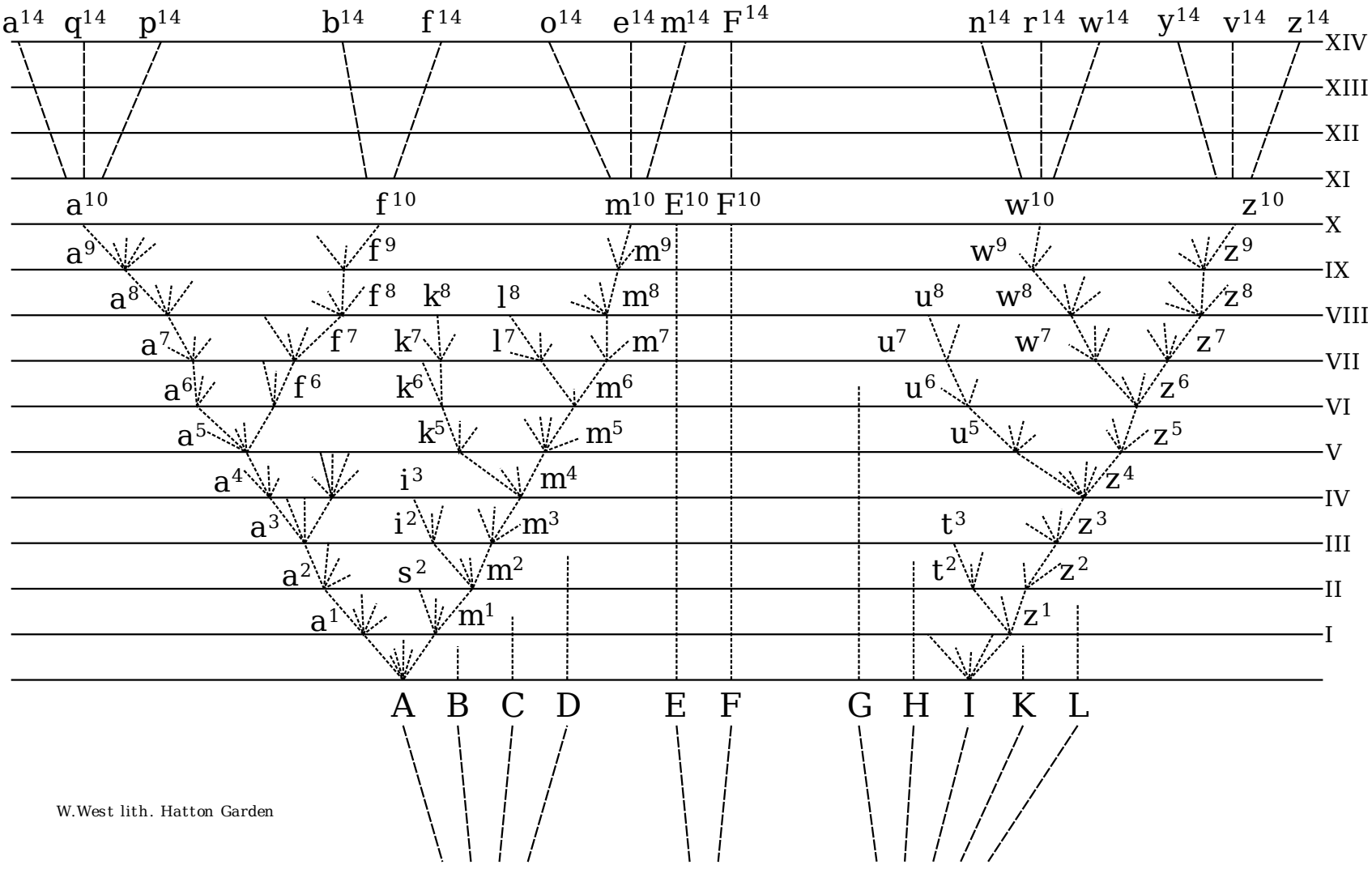
25th July 2017

1. Tree Space
2. Bayesian phylogenetics
3. Clocks and calibrations
4. Evolution now
5. Relaxed phylogenetics

Tree Space

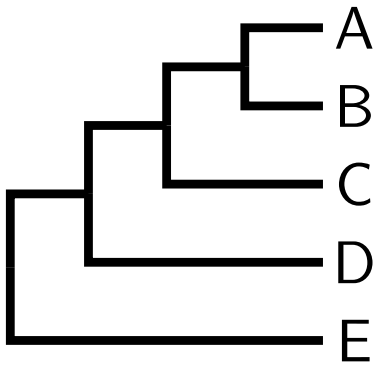
Darwin's Tree of Life

The only illustration in the *Origin of Species* (Darwin, 1859)

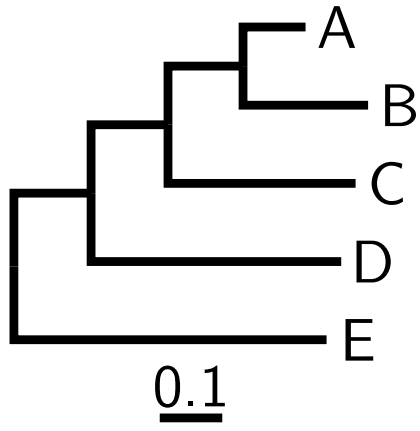


Types of phylogenies and representations

rooted trees

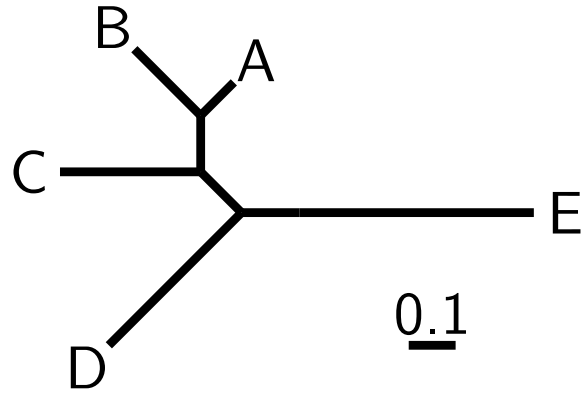


(a) cladogram



(b) phylogram

unrooted tree



(c) unrooted tree

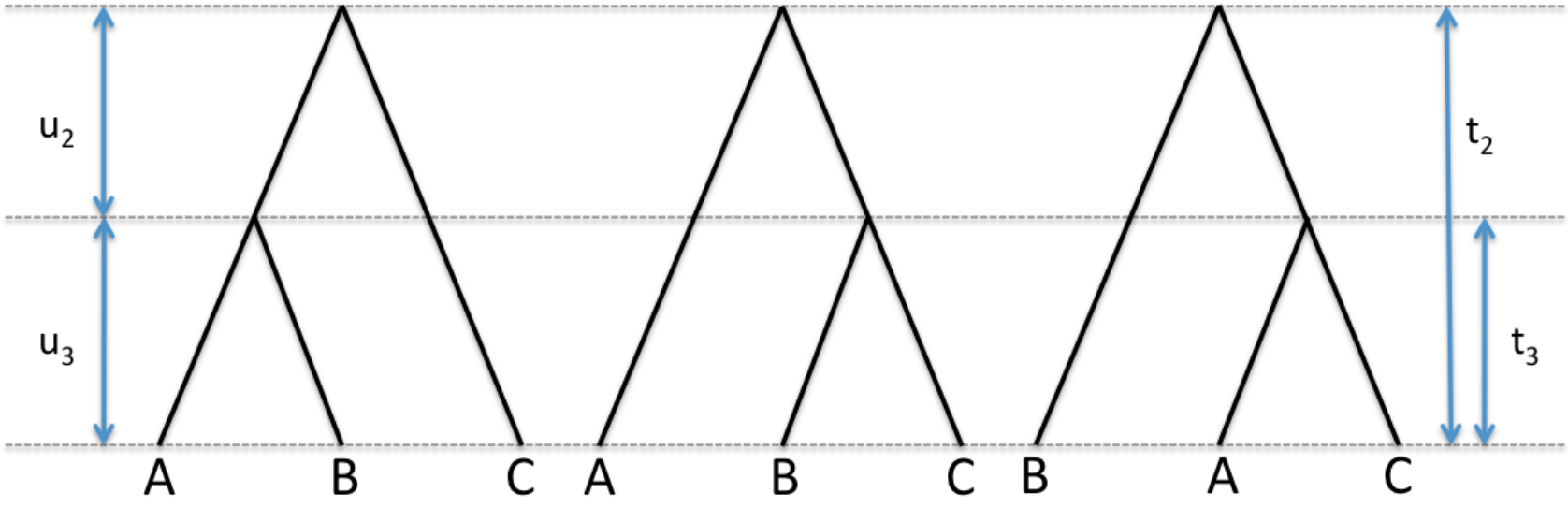
((((A, B), C), D), E);

((((A:0.1, B:0.2):0.12, C:0.3):0.123, D:0.4):0.1234, E:0.5);

branches (edges) and their lengths, nodes, tips (leaves)

The tip-labeled time-tree

A tip-labeled time-tree is described by a *tip-labeled ranked topology* of size k and *coalescent times*, $\mathbf{u} = \{u_2, \ldots, u_k\}$.



These time-trees of size 3 can be interpreted as describing the possible alternative evolutionary histories for three species or (uniparental) ancestries of the three individuals represented by the labeled tips.

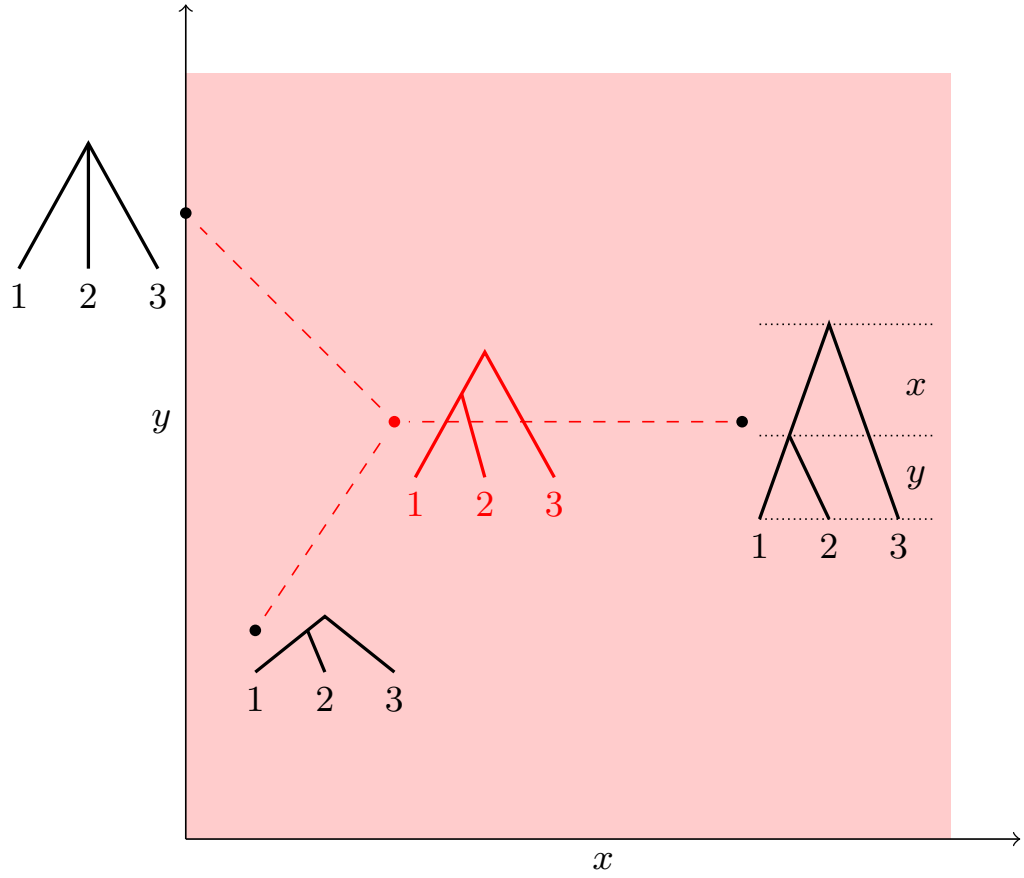


Figure: A Euclidean two-dimensional space representing the space of all possible time-trees for the topology ((1,2),3). There are two parameters, x and y, one for each of the two inter-coalescent intervals, the sum of which is the age of the root ($t_{root} = x + y$). Three trees are displayed, along with their arithmetic mean tree, also called the *centroid*. The dashed lines show the path connecting each of the three trees to the mean tree by the shortest distance (i.e. their deviations from the mean).

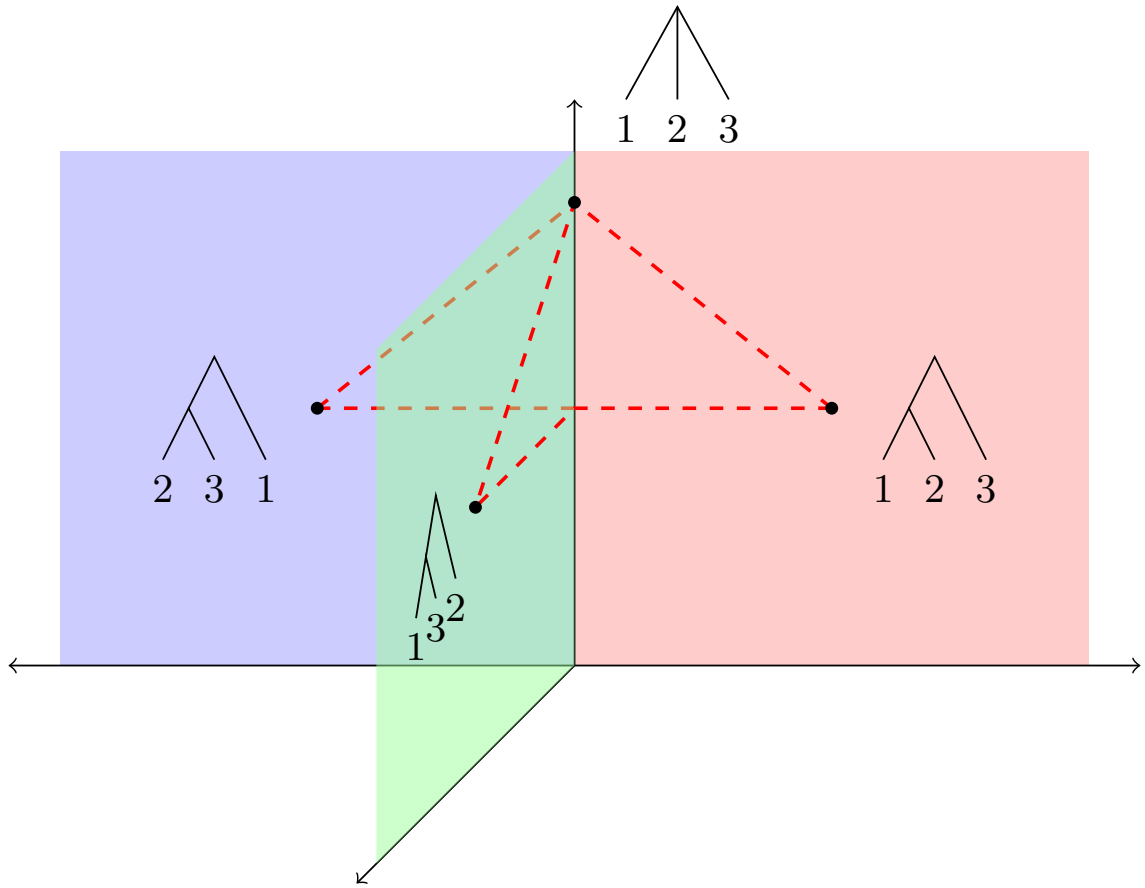


Figure: τ_3, the simplest non-trivial tree space (for time-trees), representing the space of time-trees for $n = 3$ taxa sampled contemporaneously. Each of the three non-degenerate tree topologies is represented by a two-dimensional Euclidean space (as illustrated in Figure 1) and these subspaces meet at a single shared edge representing the star tree, which is a one-dimensional subspace and thus has a single parameter (the age of the root). The dashed lines shows the paths of shortest distance between the four displayed trees.

Another space of tip-labeled time-trees of size 3

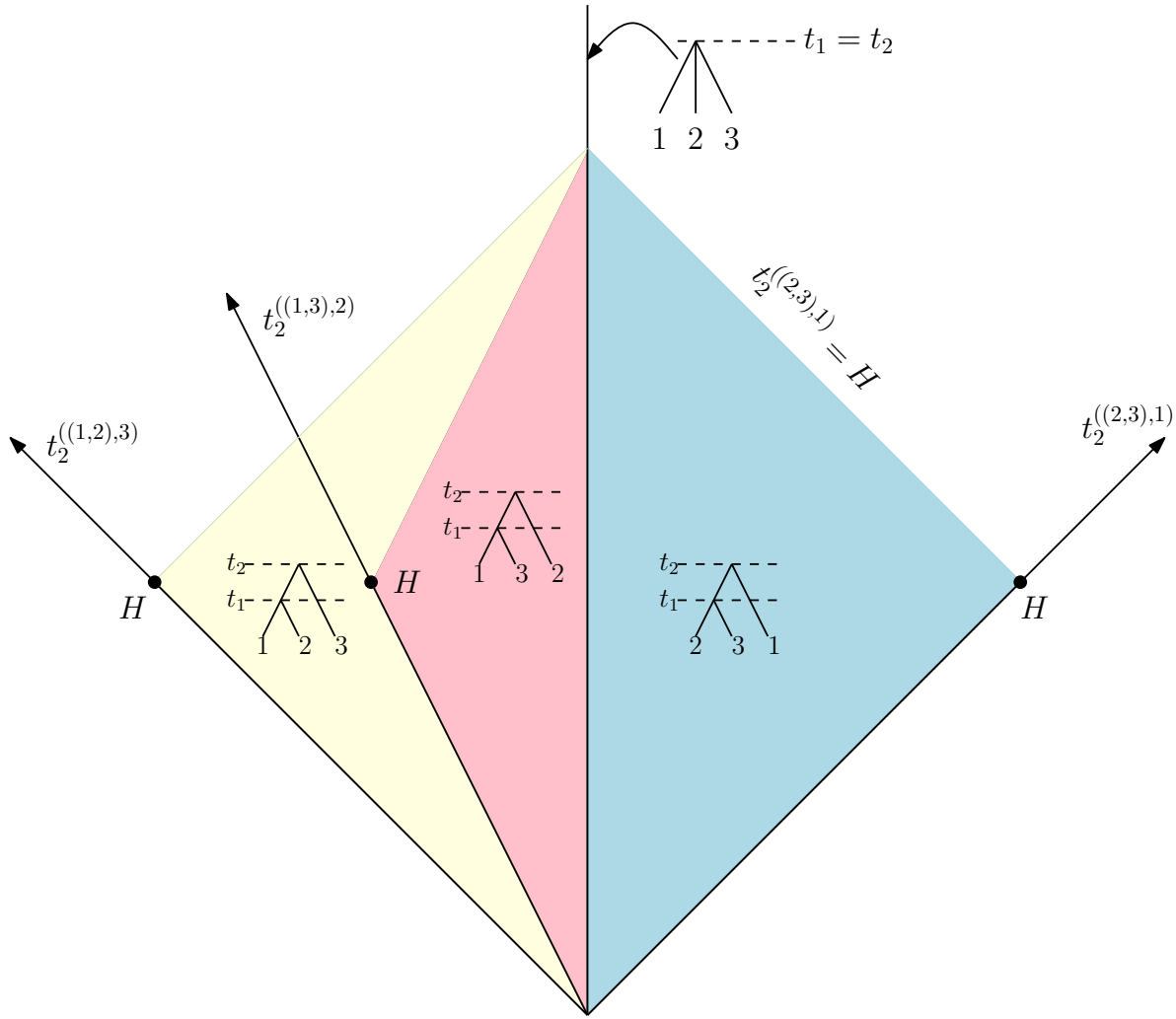


Figure: Space $\mathbb{T}_3$.

Figure from Gavryushkin and Drummond (2016)

A space of tip-labeled time-trees of size 4

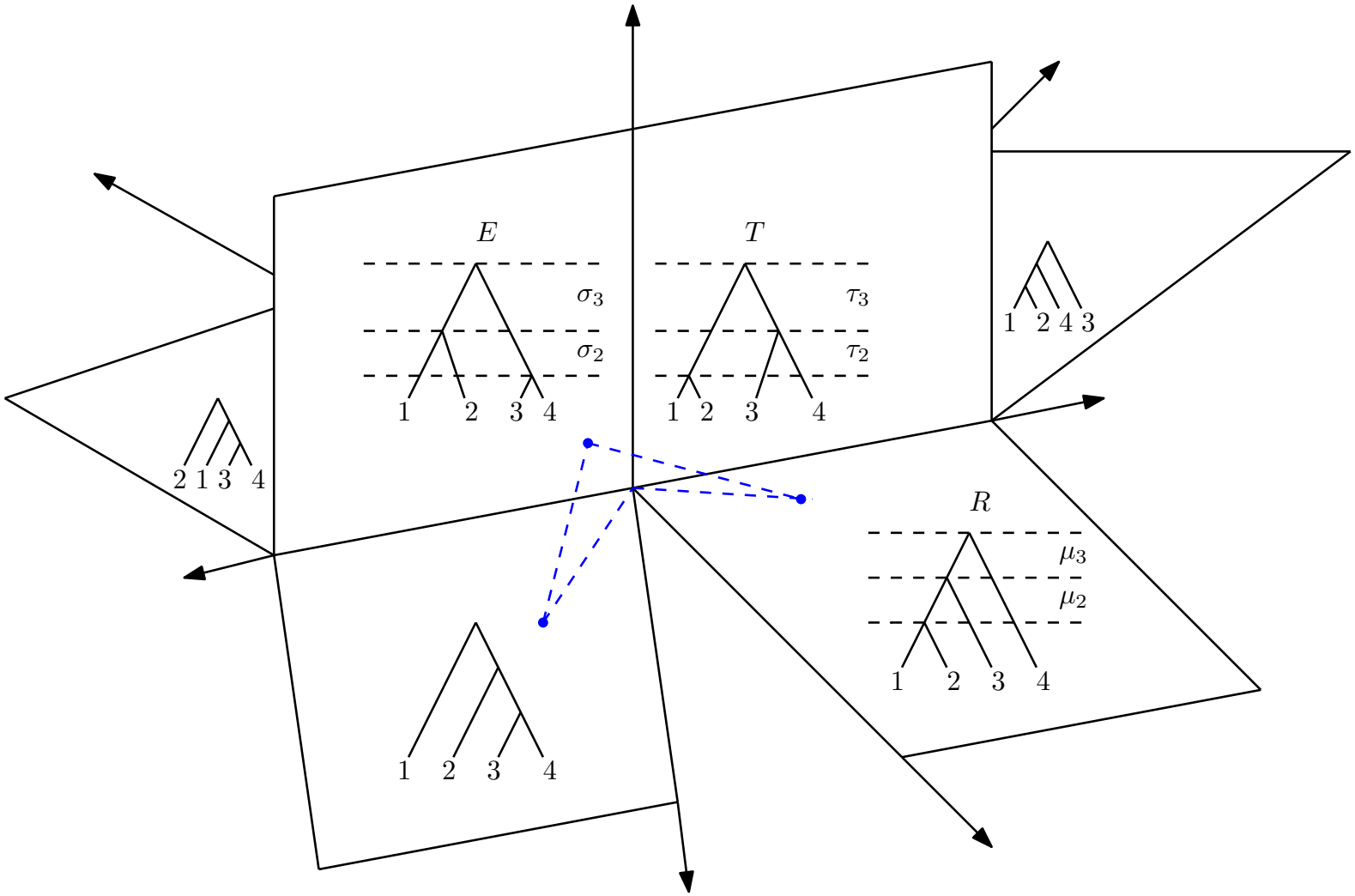
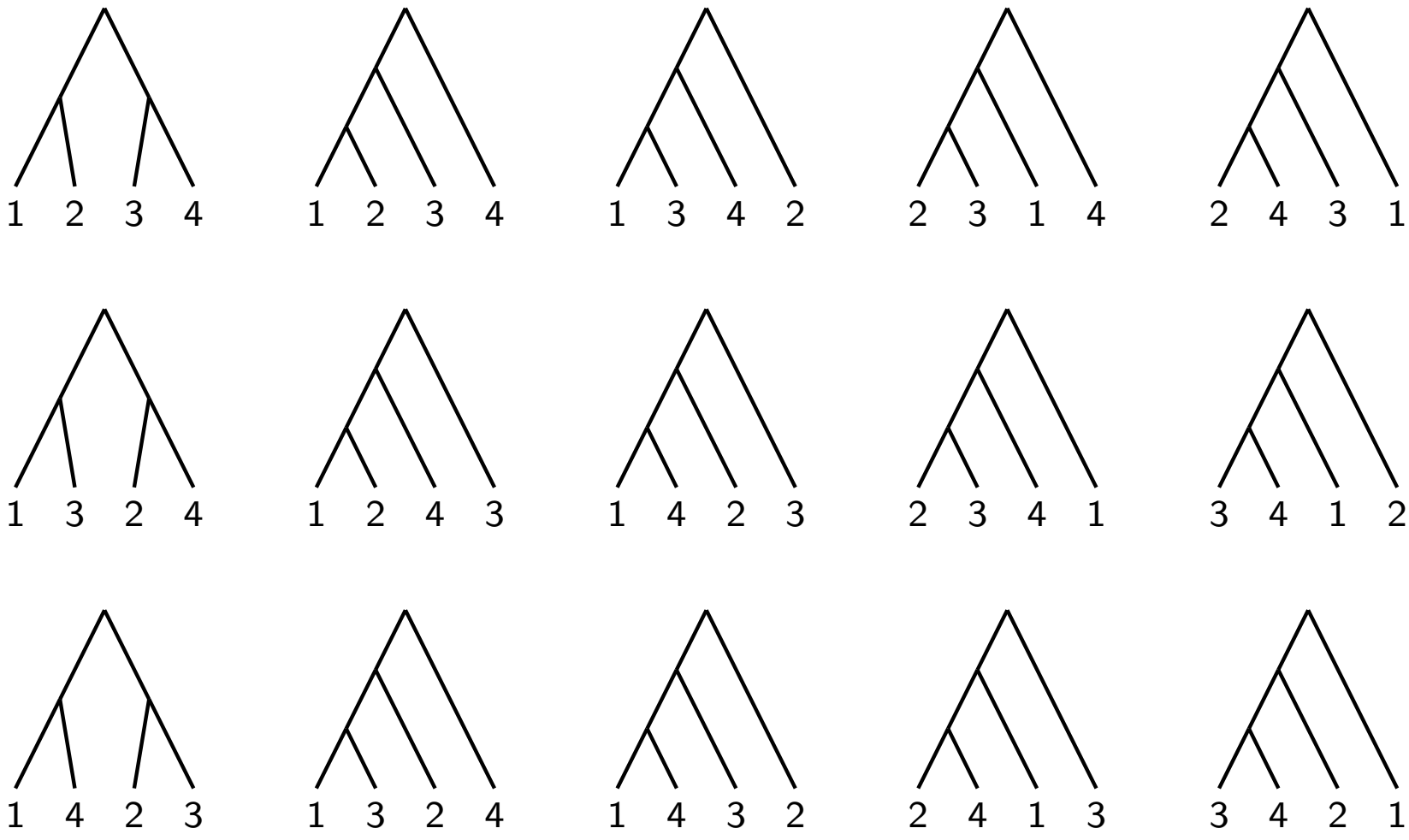


Figure: Three-dimensional projection of 4-dimensional τ-space T_4.

Figure from Gavryushkin and Drummond (2016)

Unranked tree topologies of size 4



How many trees are there?

For n species there are

$$T_n = 1 \times 3 \times 5 \times \cdots \times (2n-3) = \frac{(2n-3)!}{(n-2)!2^{n-2}}$$

rooted, tip-labelled binary trees:

n	#trees	
4	15	enumerable by hand
5	105	enumerable by hand on a rainy day
6	945	enumerable by computer
7	10395	still searchable very quickly on computer
8	135135	about the number of hairs on your head
9	2027025	greater than the population of Auckland
10	34459425	$\approx$ upper limit for exhaustive search
20	8.20×10^{21}	$\approx$ upper limit of branch-and-bound searching
48	3.21×10^{70}	$\approx$ the number of particles in the Universe
136	2.11×10^{267}	number of trees to choose from in the "Out of Africa" data (Vigilant *et al.* 1991)

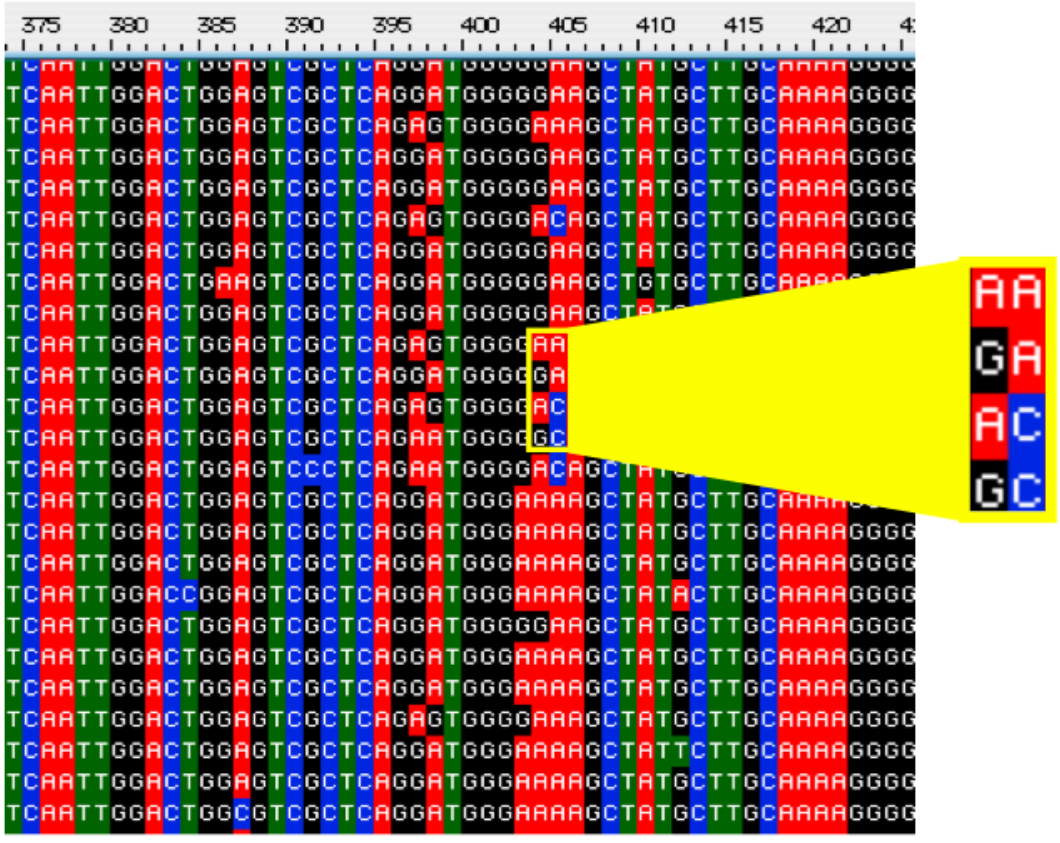
Counting different types of rooted trees

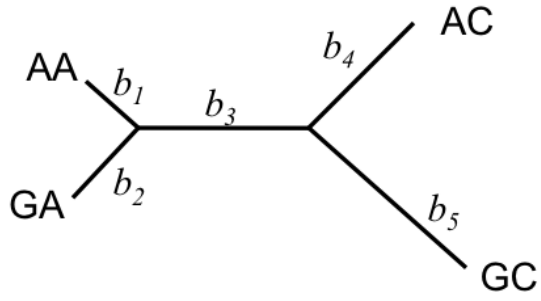
n	#shapes	#trees, $\lvert\mathcal{T}_n\rvert$	#ranked trees	#fully ranked trees
2	1	1	1	1
3	1	3	3	4
4	2	15	18	34
5	3	105	180	496
6	6	945	2700	11056
7	11	10395	56700	349504
8	23	135135	1587600	14873104
9	46	2027025	57153600	819786496
10	98	34459425	2571912000	56814228736

Table: The number of unlabeled rooted tree shapes, the number of labelled rooted trees, the number of labelled ranked trees (on contemporaneous tips), and the number of fully-ranked trees (on distinctly-timed tips) as a function of the number of taxa, n.

Bayesian phylogenetics

Felsenstein's likelihood (1981)





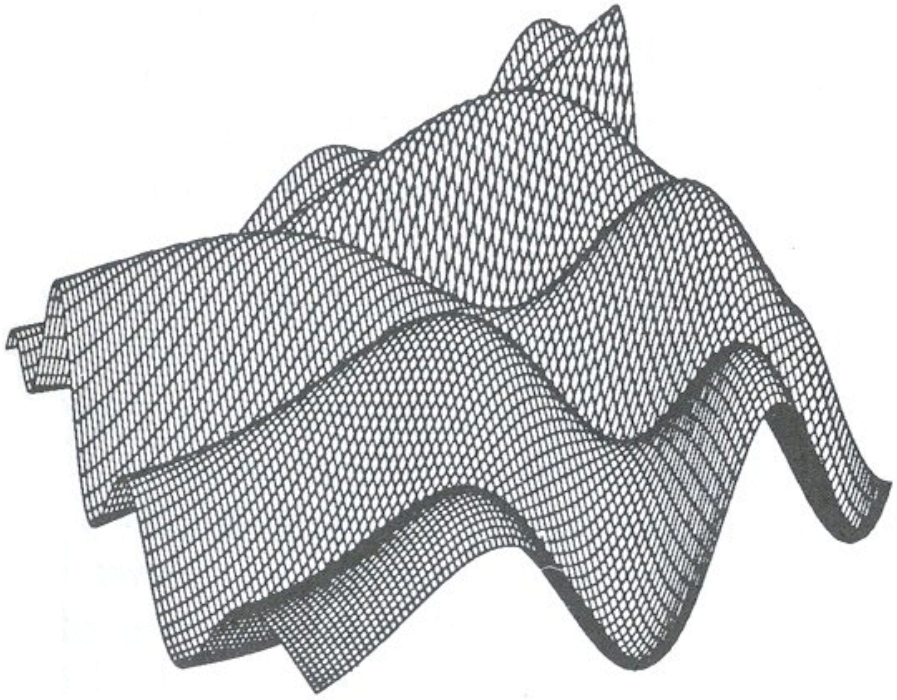
$$L(T) = Pr\{D|T, Q\}$$

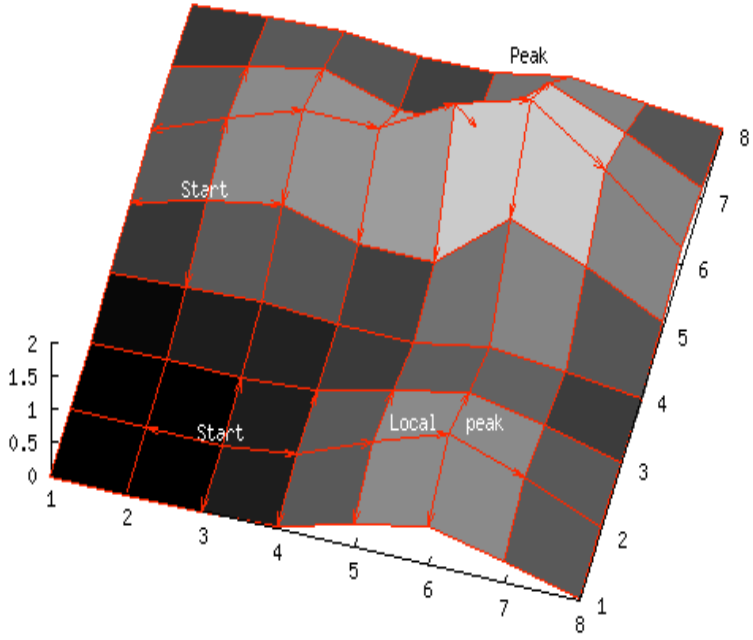
The probability of the data, $Pr\{D|T, Q\}$ can be efficiently calculated given a phylogenetic tree (T), and a **probabilistic model** of molecular evolution (Q).

In statistical phylogenetics, branch lengths are traditionally unconstrained.

Tree space as a hilly landscape

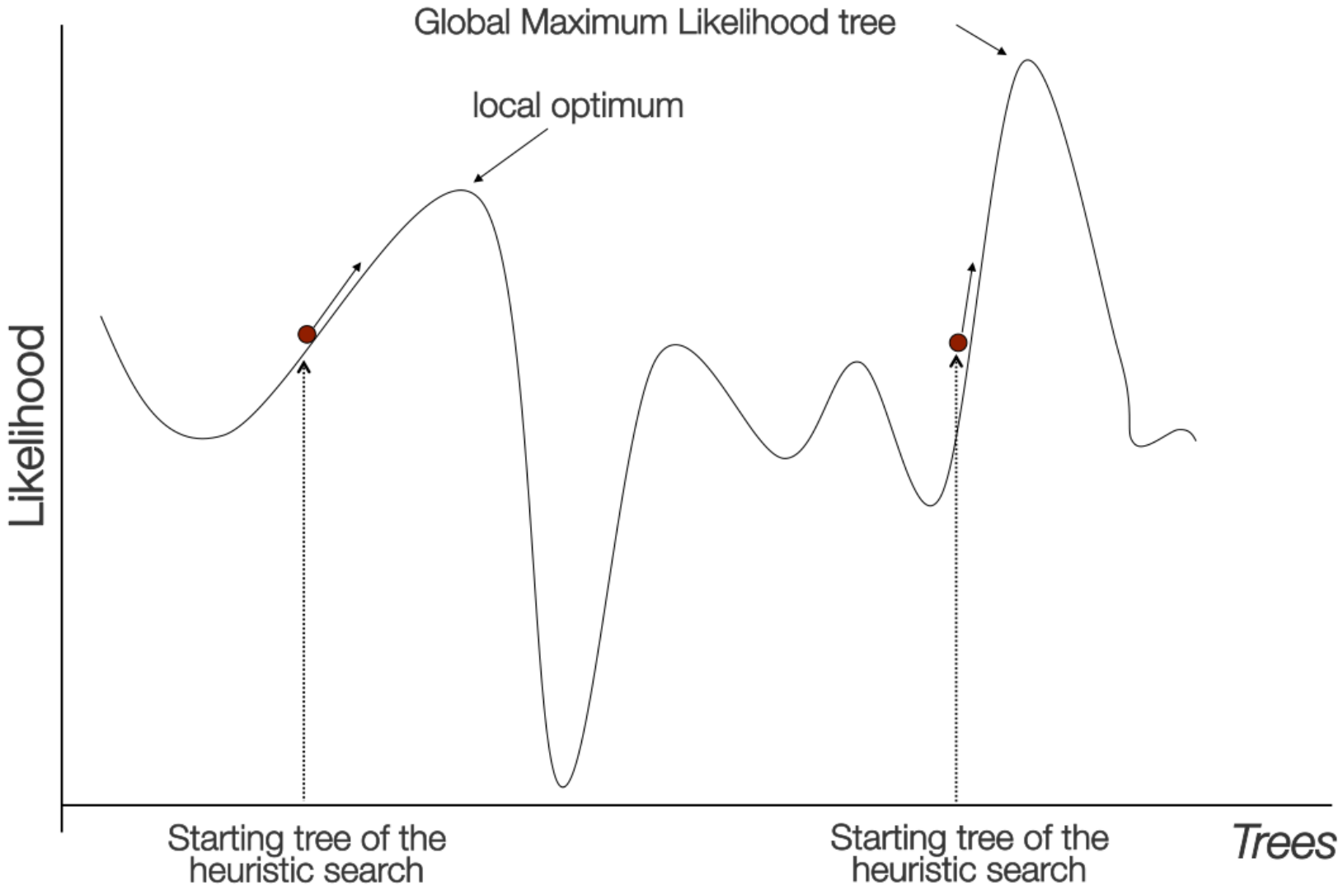
The space of all possible trees can be visualized as a hilly landscape. Nearby points in this landscape represent similar trees, and the height of the landscape is the probability of the tree at that point.





- This space can be **sampled** in a Bayesian analysis with MCMC
- The peak can be identified by a **search algorithm** in the context of maximum likelihoods

Local tree search and multiple optima



Bayes rule in statistics

$$Pr(\theta|D) = \frac{Pr(D|\theta)Pr(\theta)}{Pr(D)}$$

where

- $P(D|\theta)$ is the likelihood,
- $Pr(\theta)$ is the prior distribution and
- $Pr(\theta|D)$ is the posterior distribution.
- $Pr(D)$ is the marginal likelihood of the data.

Bayes rule in phylogenetics

$$p(T, Q|D) = \frac{Pr\{D|T, Q\}p(T)p(Q)}{Pr\{D\}}$$

where

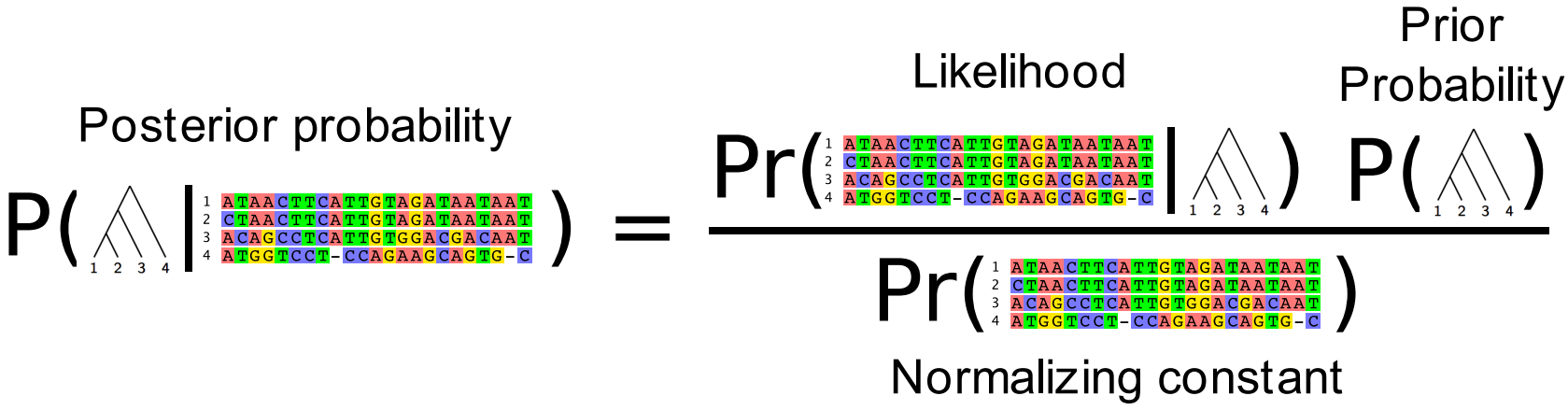
- $Pr(D|T, Q)$ is Felsenstein's likelihood,
- $p(T)$ is the prior distribution on phylogenetic trees,
- $p(Q)$ is the prior distribution on the model of evolution and
- $p(T, Q|D)$ is the posterior distribution
- $Pr(D)$ is the marginal likelihood of the data.

Bayesian reconstruction of phylogenetic trees

Yang & Rannala (1997), Mau, Newton & Larget (1998)

In the context of Bayesian phylogenetics, what we want to compute is the **probability of the tree** given the data.

We can compute that from the **likelihood** using **Bayes Theorem**:



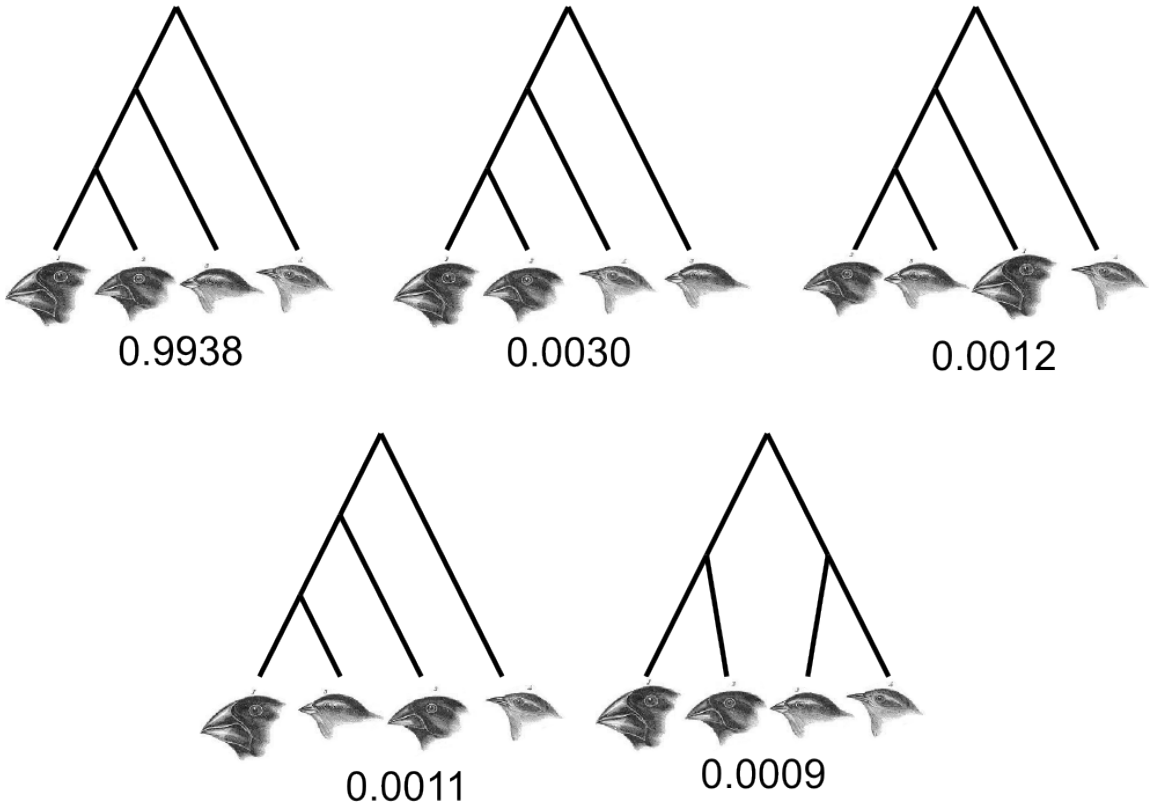
This is known as the **Posterior probability** of the tree. Another method of reconstructing the evolutionary history is then to find the tree that has the **Maximum Posterior probability**.

Bayesian Phylogenetics

- The output of a Bayesian evolutionary analysis is a probability distribution on trees and parameter values.
- For phylogenetics the tree topology is the object of interest. The substitution parameters and tree prior parameters are a nuisance that we average over using MCMC and then ignore.
- For population genetics the tree and substitution parameters are a nuisance that we average over and then ignore, focusing instead on the population parameters.
- Often a more specific hypothesis is of interest (like "Did this adaptive radiation predate the Miocene?") and then the result of the analysis should be the testing of this hypothesis, averaged over all trees and parameter values, weighted by their probability given the data.

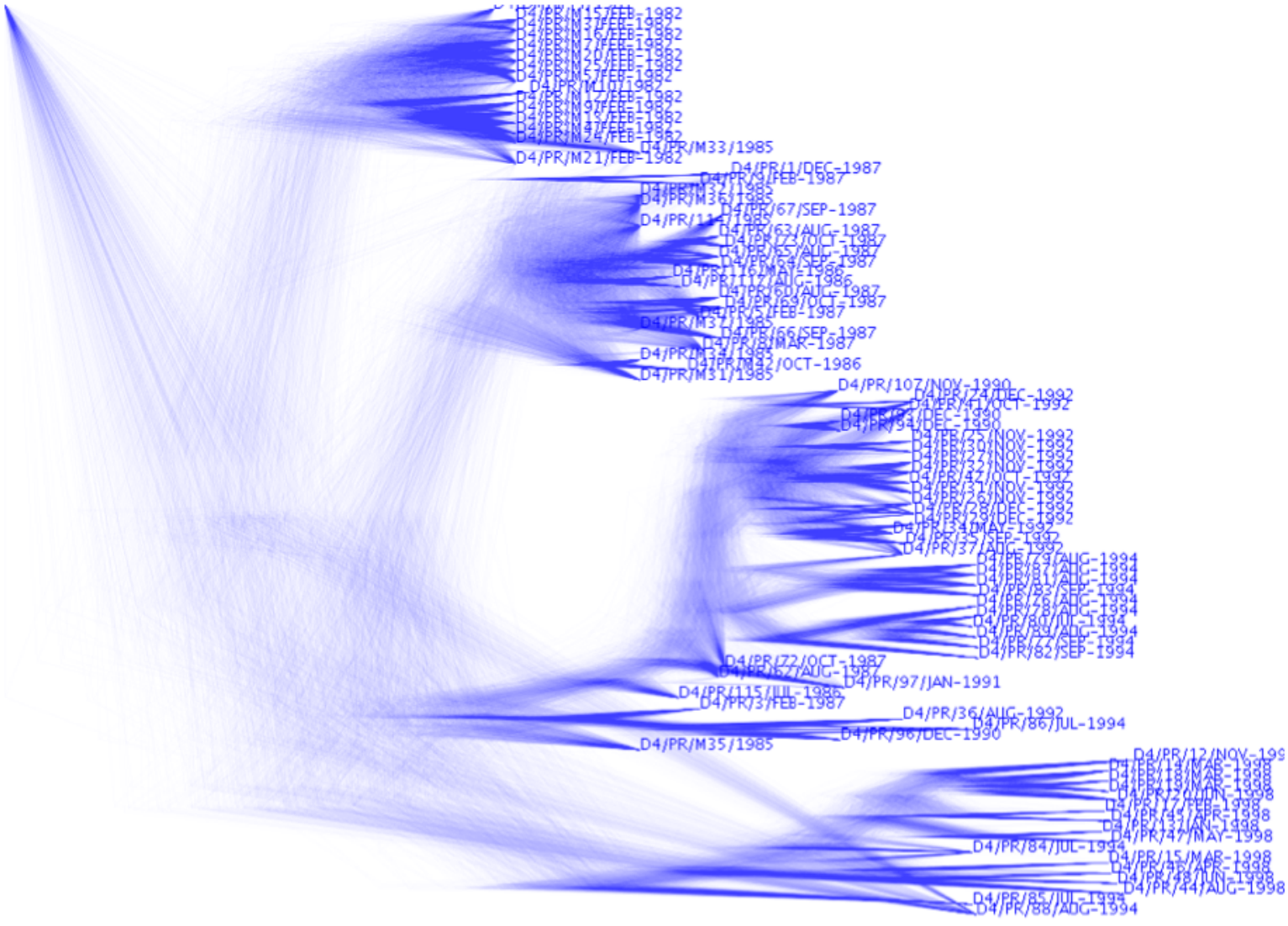
The Posterior Distribution on Darwin's Finches

```
1  ATAACTTCATTGTAGATAATAAT
2  CTAACTTCATTGTAGATAATAAT
3  ACAGCCTCATTGTGGACGACAAT
4  ATGGTCCT-CCAGAAGCAGTG-C
```



This posterior probability distribution was computed using an algorithm called **Markov chain Monte Carlo** implemented in the BEAST software package (Drummond & Rambaut, 2007).

The posterior distribution for larger trees



Elaborating the model

Basic model: (posterior proportional to likelihood $\times$ prior)

$$p(T|D) \propto \Pr\{D|T\}p(T)$$

Substitution model estimation:

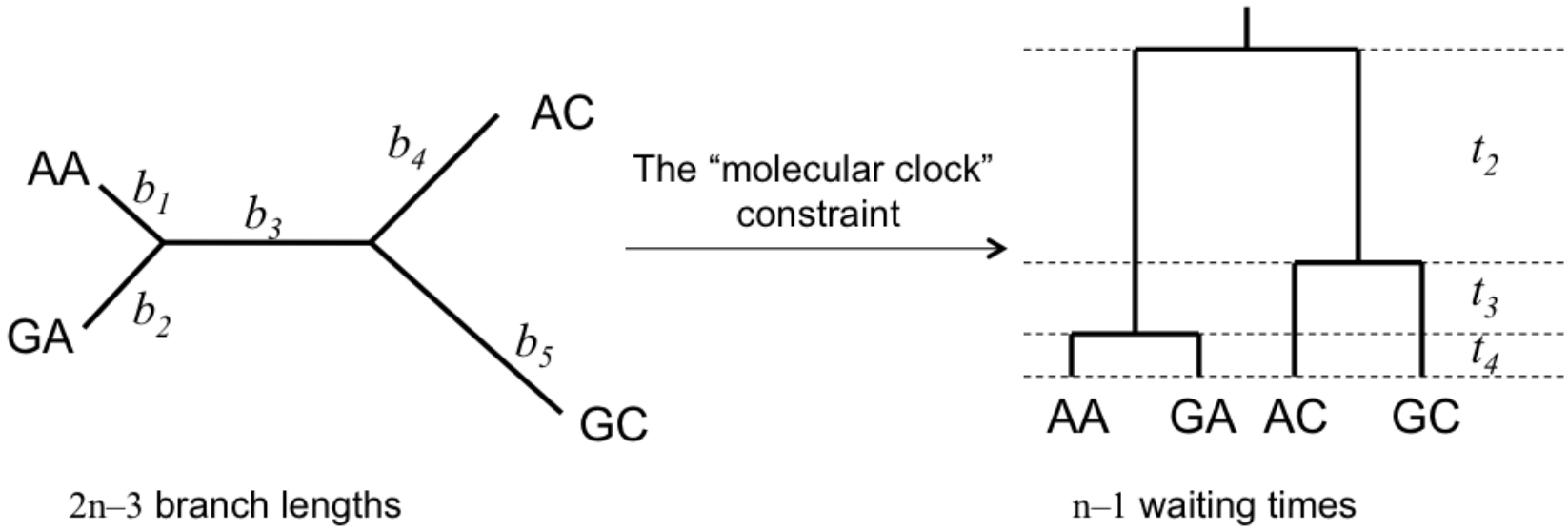
$$p(T, Q|D) \propto \Pr\{D|T, Q\}p(T)p(Q)$$

Substitution model and parametric tree prior:

$$p(T, Q, \theta|D) \propto \Pr\{D|T, Q\}p(T|\theta)p(Q)p(\theta)$$

Clocks and calibrations

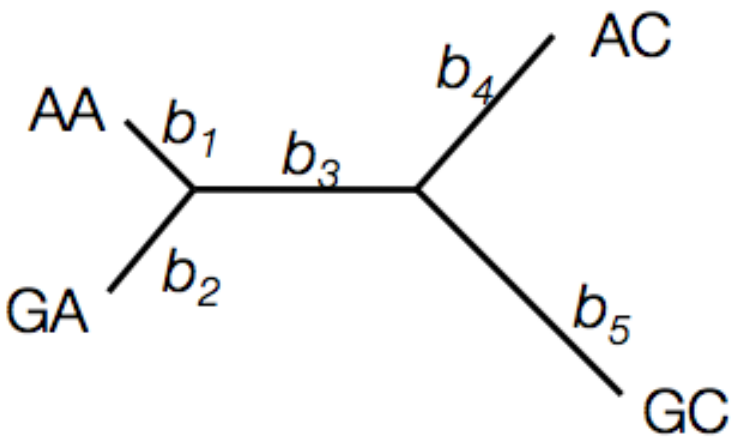
The molecular clock constraint

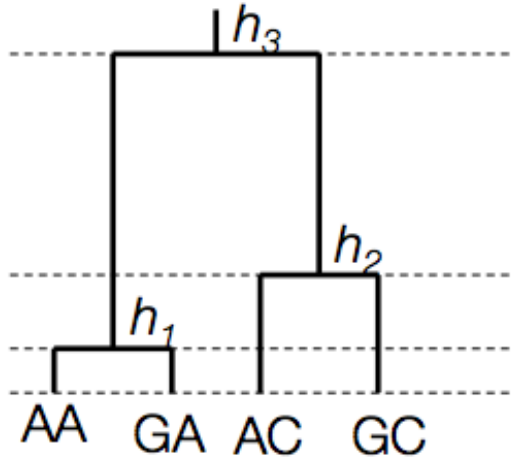


$$h(g, Q|D) \propto Pr\{D|\vec{\mu}g, Q\} f_G(g) f_Q(Q)$$

The joint posterior probability of the **rooted** time-tree (g) and the substitution matrix (Q) are estimated using Markov chain Monte Carlo (Drummond *et al*, 2002; 2006)

Model assumptions





- Product of rate and time (branch length) is independent and identically distributed among branches.
- The root of the tree could be anywhere with equal probability.
- Topology implies nothing about individual branch lengths.

- Rate of evolution is the same on all branches.
- The root of the tree is equidistant from all tips.
- Topology constrains branch lengths (e.g. two branches in a cherry must be of equal length)

Calibration via a global molecular clock

Basic model: (Tree in expected substitutions per site)

$$p(g, \theta|D) \propto \Pr\{D|g\}p(g|\theta)p(\theta)$$

Fix (i.e. condition on) the global rate to μ:

$$p(g, \theta|D) \propto \Pr\{D|\mu \times g\}p(g|\theta)p(\theta)$$

Estimate the global rate:

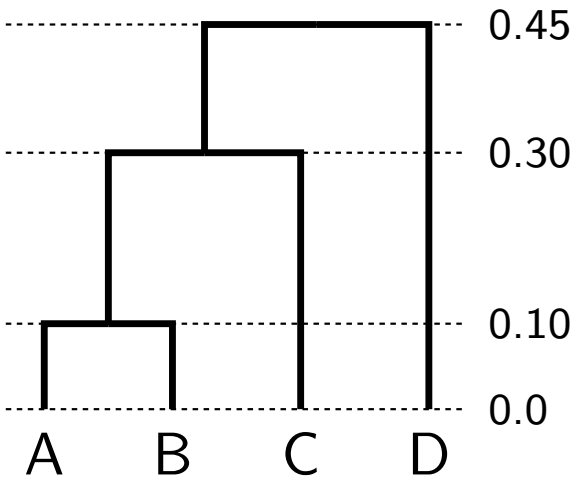
$$p(g, \mu, \theta|D) \propto \Pr\{D|\mu \times g\}p(g|\theta)p(\theta)p(\mu)$$

In the models above the parameters related to the details of the substitution process (Q) have been suppressed for simplicity.

Genetic distance = rate × time

Strict molecular clock

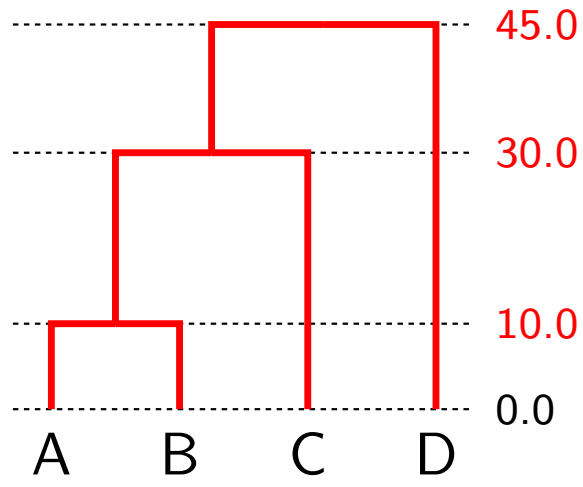
$$T = \mu \times g$$



"substitution tree"

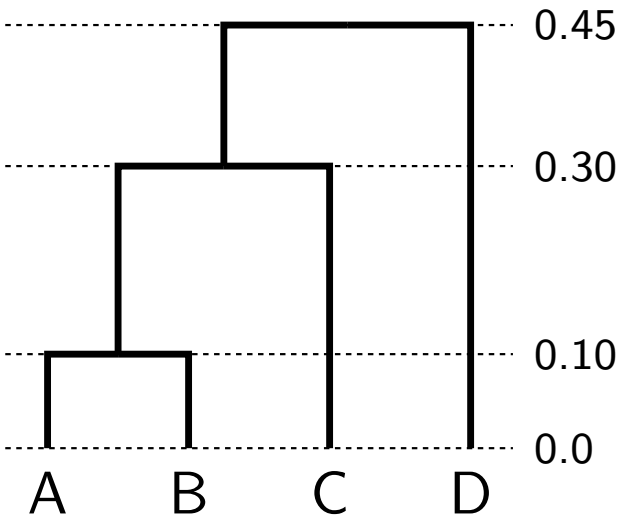
$= \quad 0.01 \quad \times$

evolutionary rate
substitutions / site / unit time

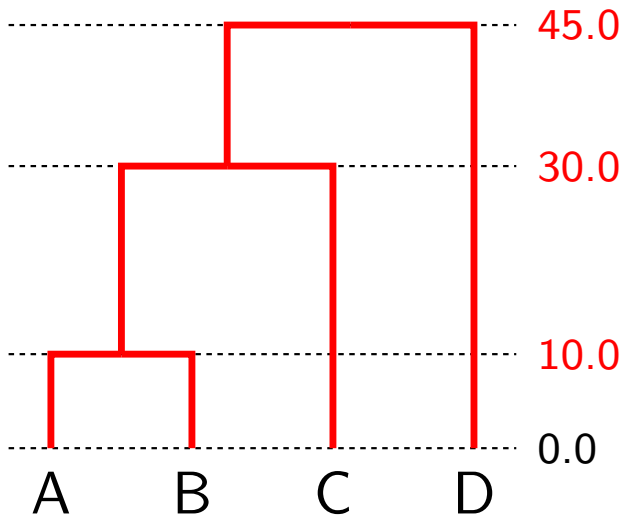


time tree

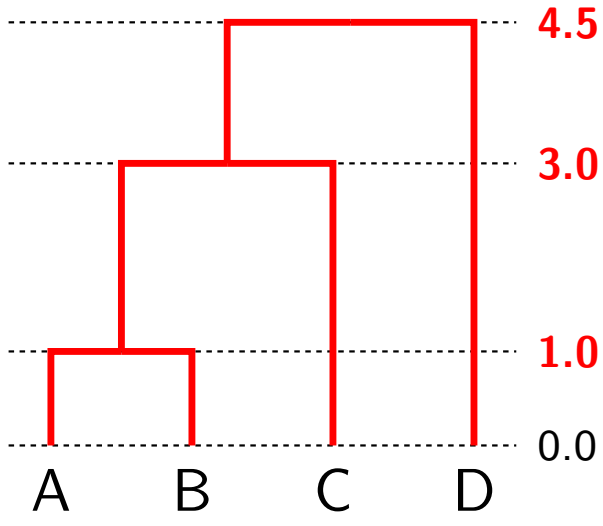
Non-identifiability of rate and times



$=$ 0.01 $\times$



$=$ **0.1** $\times$

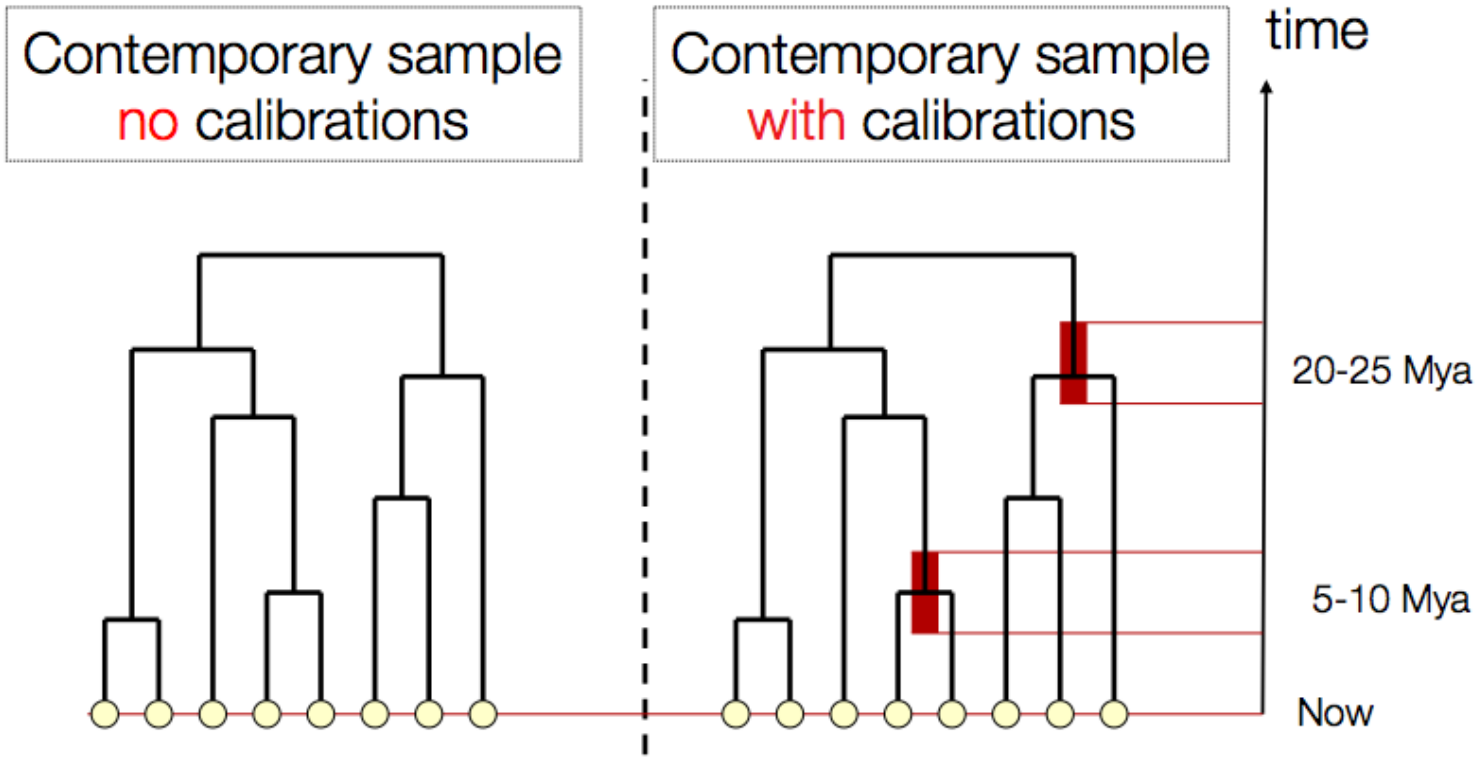


"substitution tree"

evolutionary rate
substitutions / site / unit time

time tree

Absolute time via calibrations

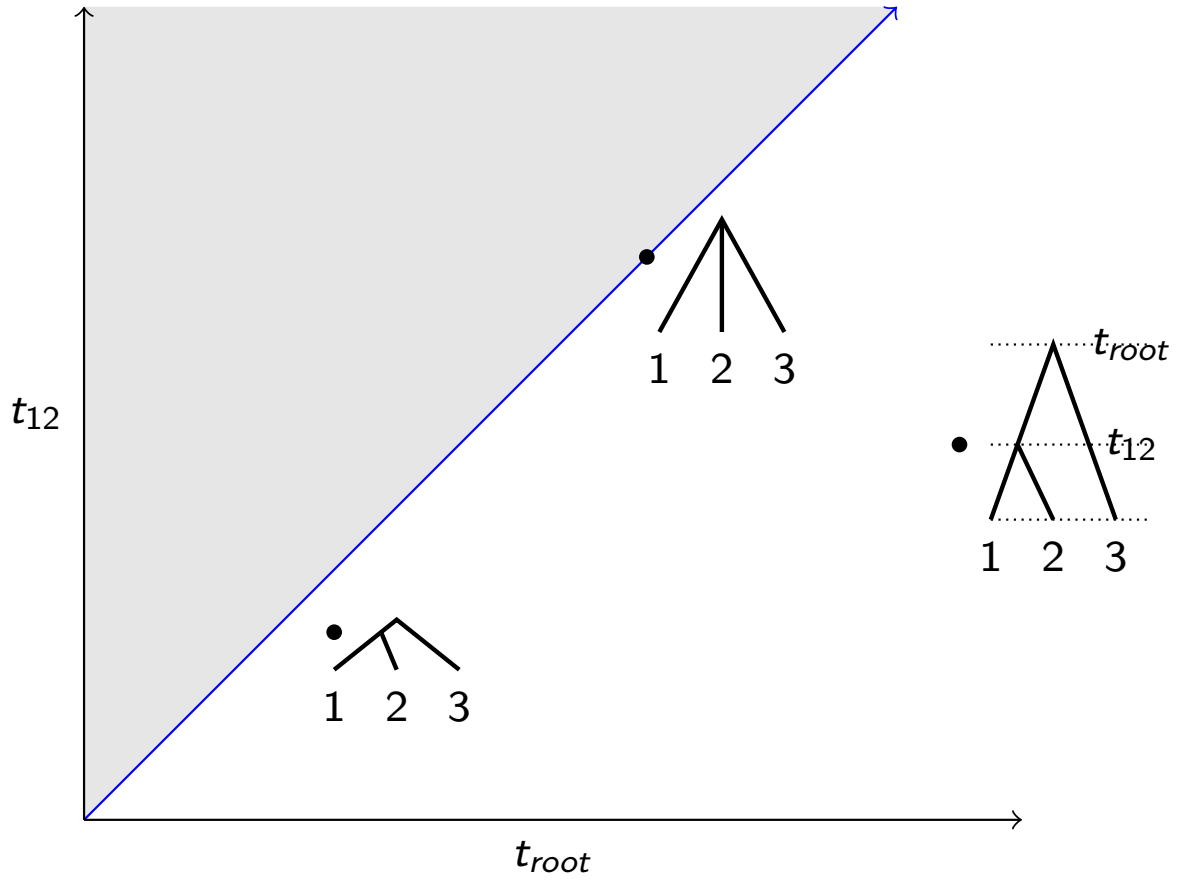


Let $\rho_G(g)$ be "calibrated" $f_G(g)$ and allow the rate(s), $\vec{\mu}$, to be estimated:

$$p(\vec{\mu}, g, Q|D) \propto Pr\{D|\vec{\mu}g, Q\}\rho_G(g)f_Q(Q)f_M(\vec{\mu})$$

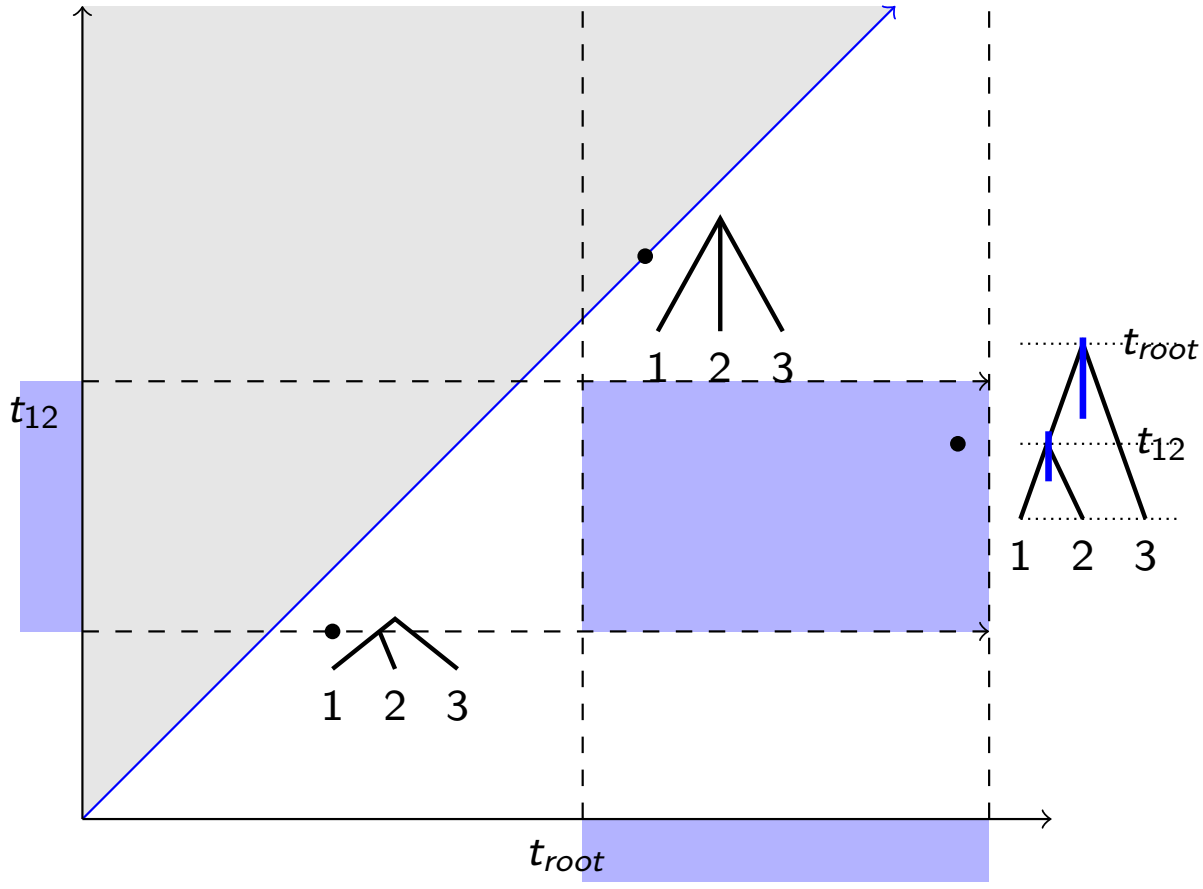
Calibrating tree space

What do calibrations look like in a small tree space?



Calibrating tree space

Two calibrations seems like it might be okay?

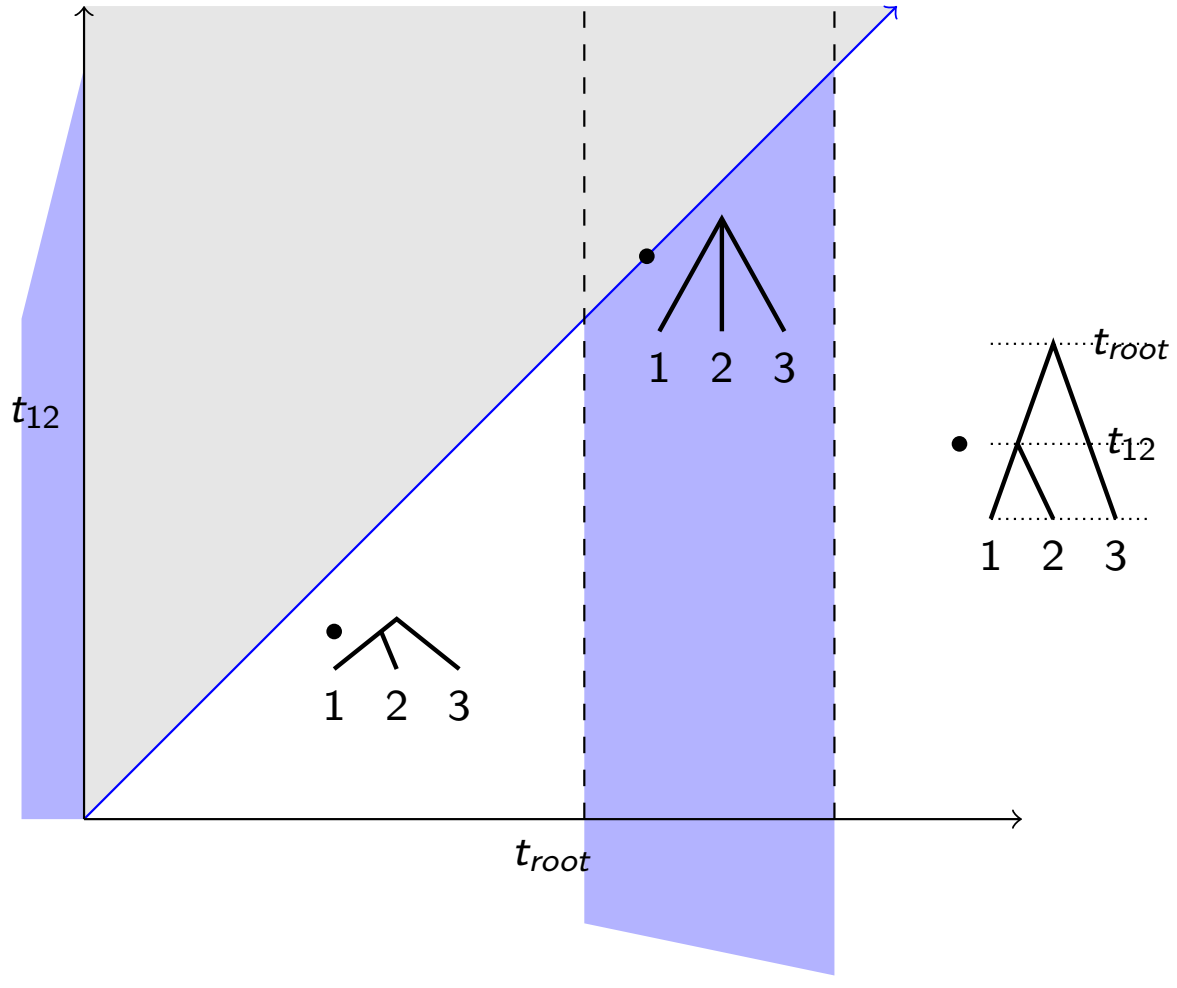


First calibration: $8 < t_{root} < 14.5$

Second calibration: $3 < t_{12} < 7$

Calibrating tree space

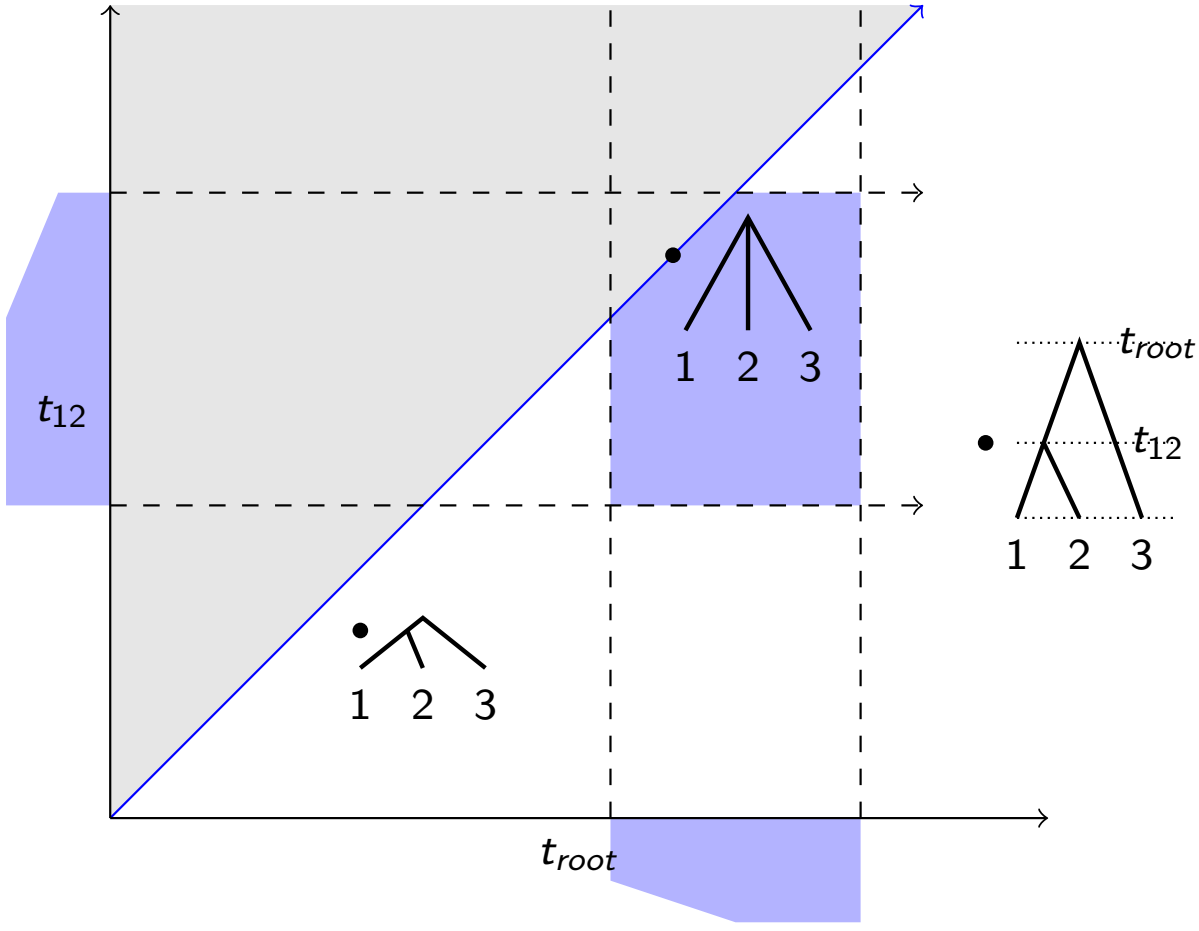
One calibration is not simple!



Calibration: $8 < t_{root} < 12$

Calibrating tree space

Two calibrations is even less simple!



First calibration: $8 < t_{root} < 12$

Second calibration: $5 < t_{12} < 10$

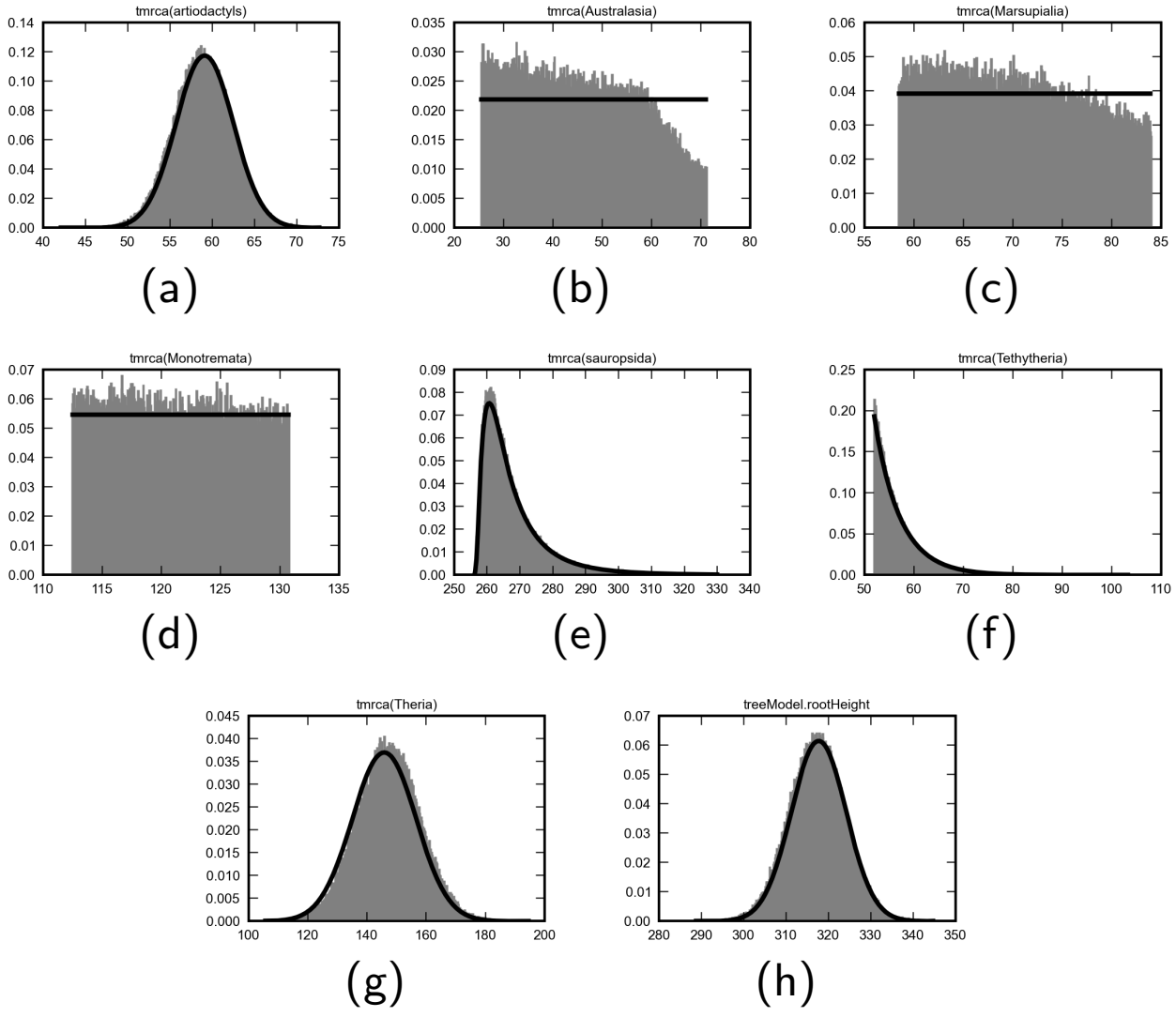


Figure: A simple construction of calibrated tree prior: $\rho_G(g) \propto f_G(g) \times \prod_{i=1}^{k} f_i(s_i)$. Where $f_i()$ is the univariate "calibration density" for the divergence time of the i'th calibrated node in the tree. Monophyly is enforced for each calibrated node.

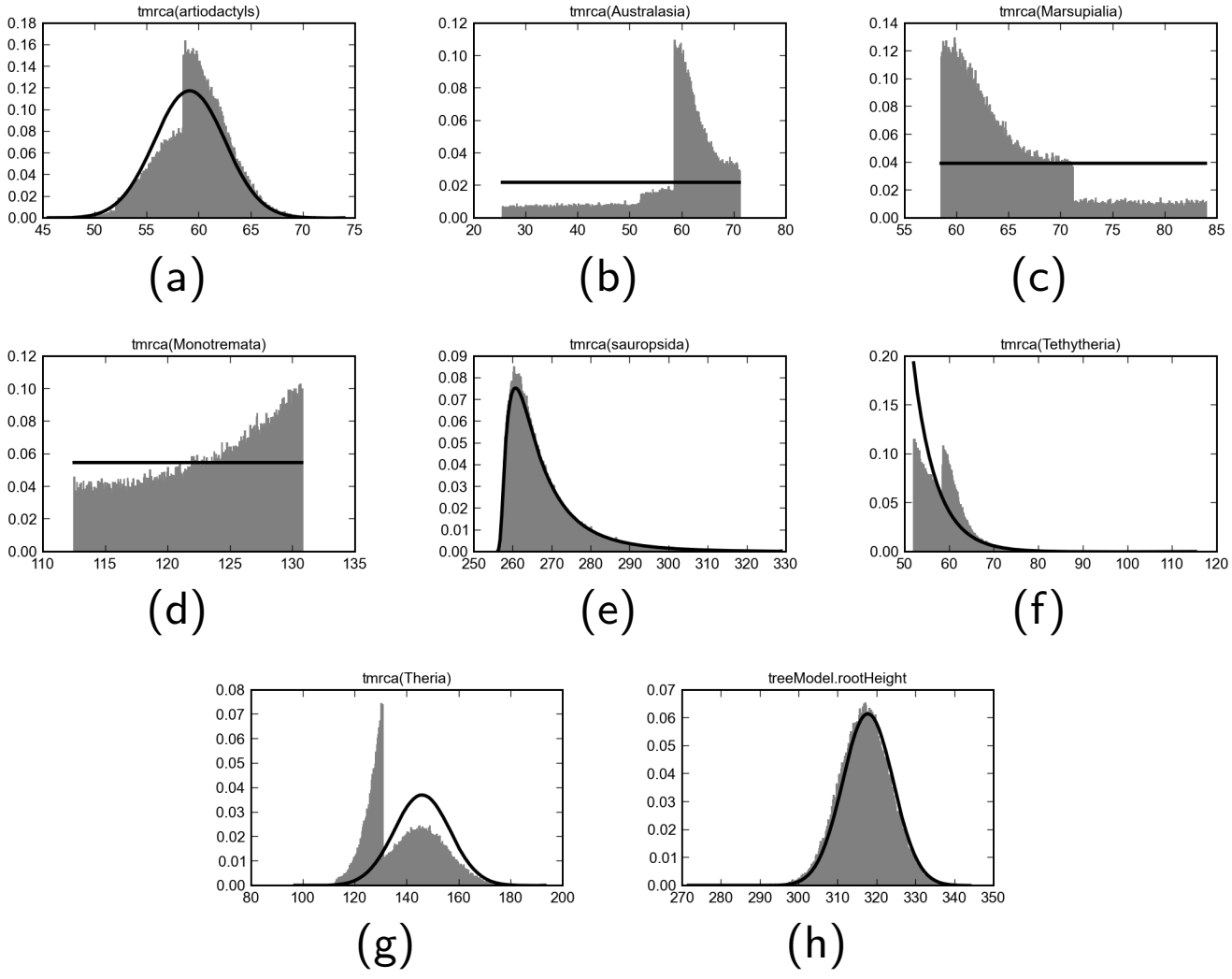
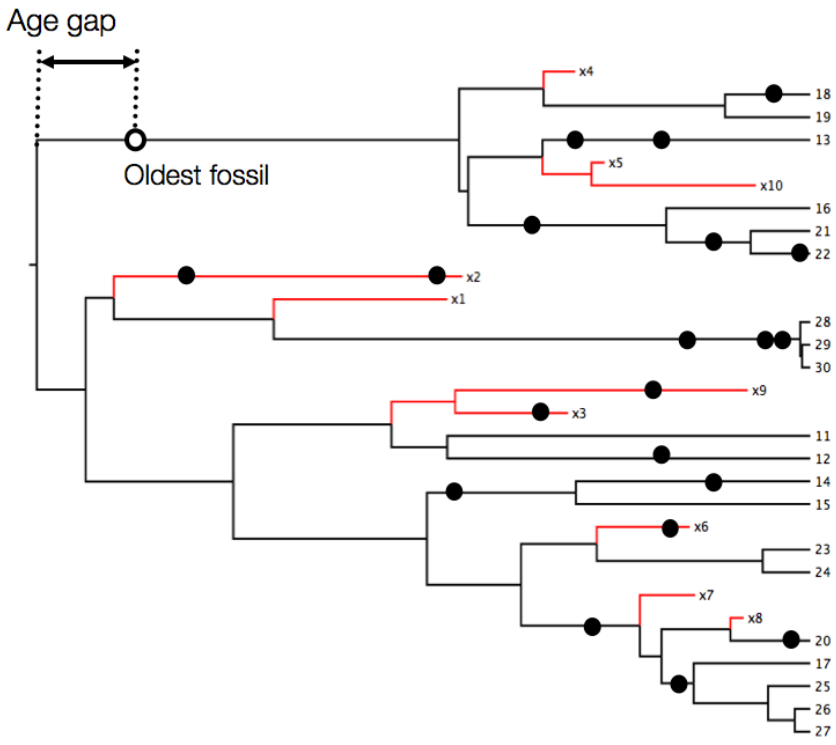


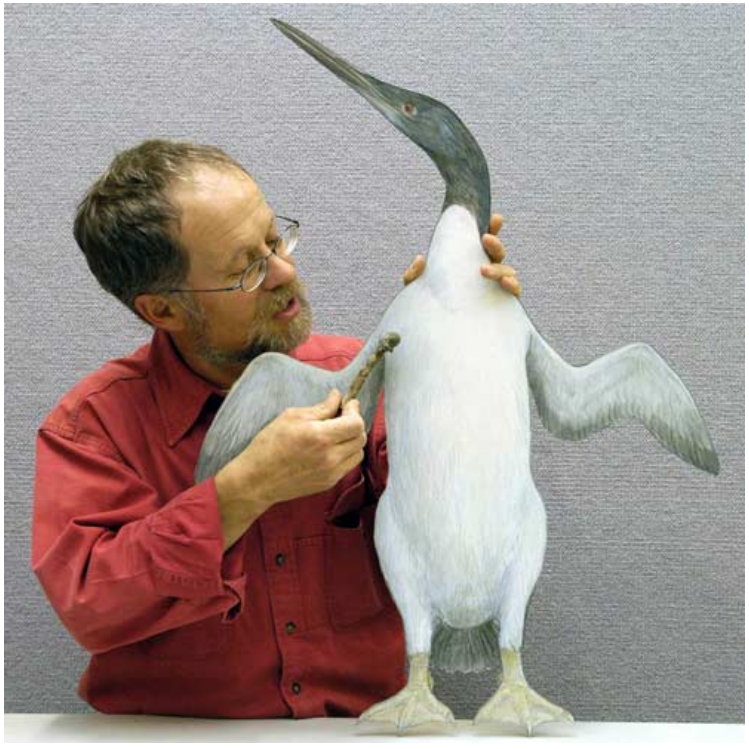
Figure: The marginal prior distributions that result from BEAST (gray) versus calibration densities (black) specified for the calibrated nodes from [**?**]. The marginal prior distributions were obtained from a MCMC run using the prior only.

How do I pick the calibration density?

Modeling the Fossil Age Gap



What is the probability distribution of the age gap?

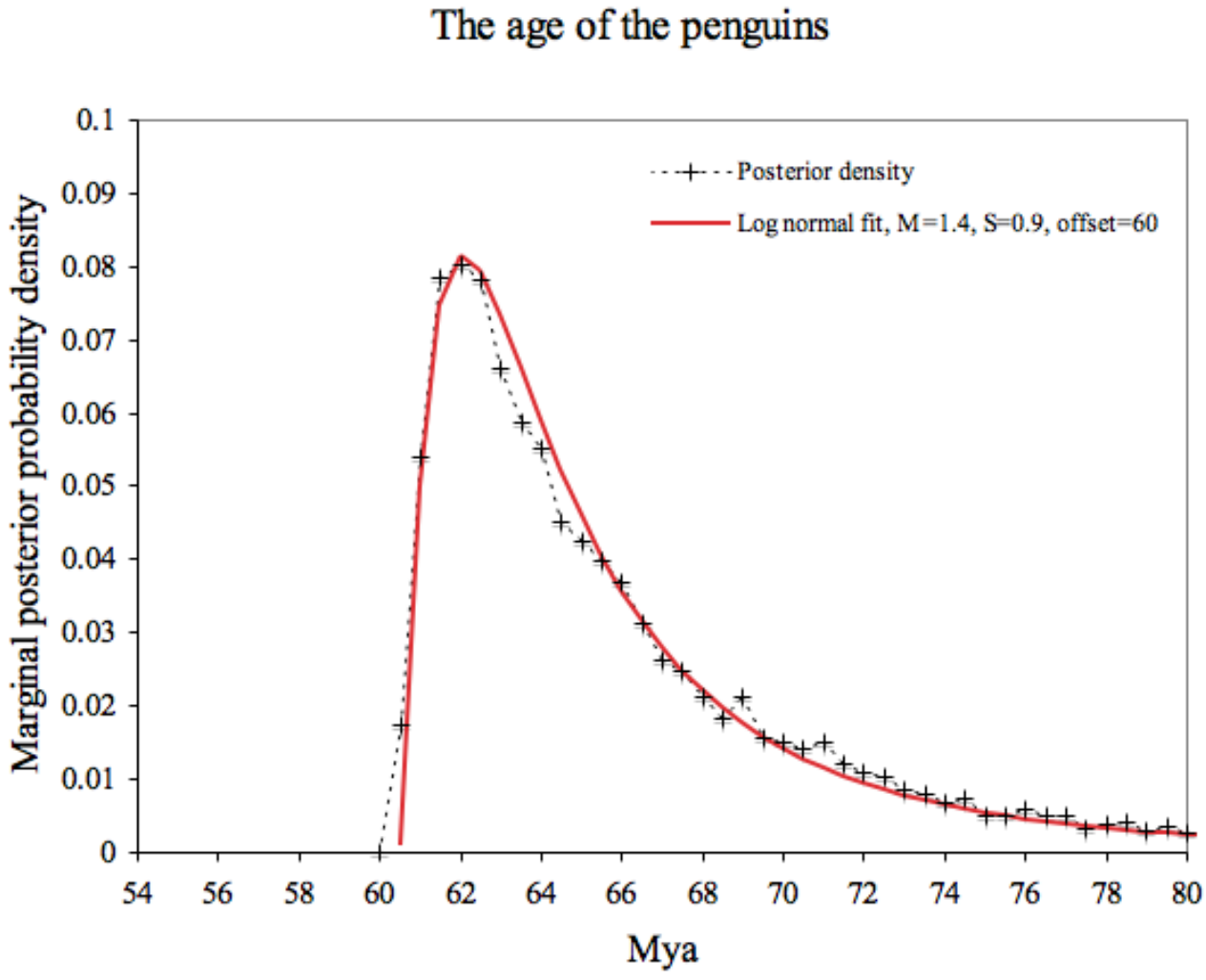


60-61.5 Myr penguin

Current day penguin species: 20

Number of independent penguin fossils with good geological age from all ages: 20-60

The posterior estimate of the age of penguins

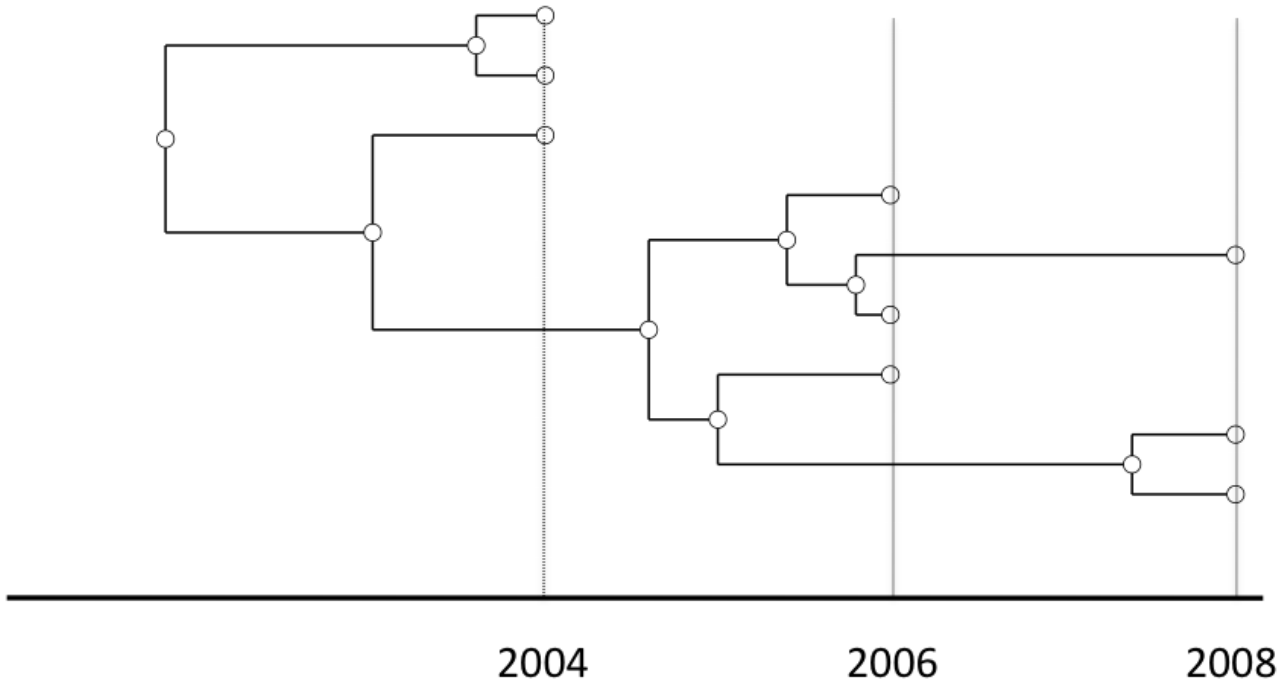


Evolution now

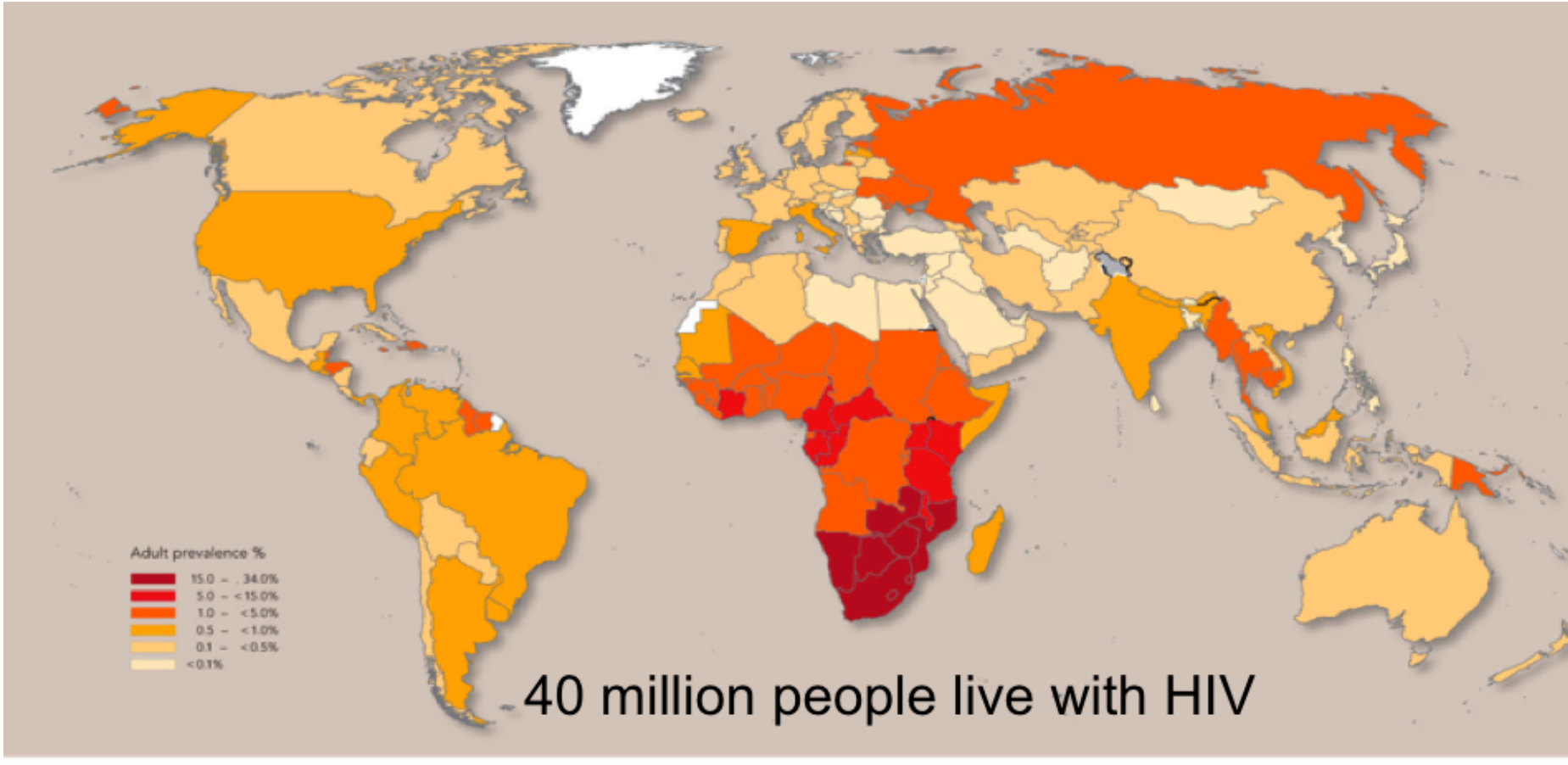
Evolution is happening right now!

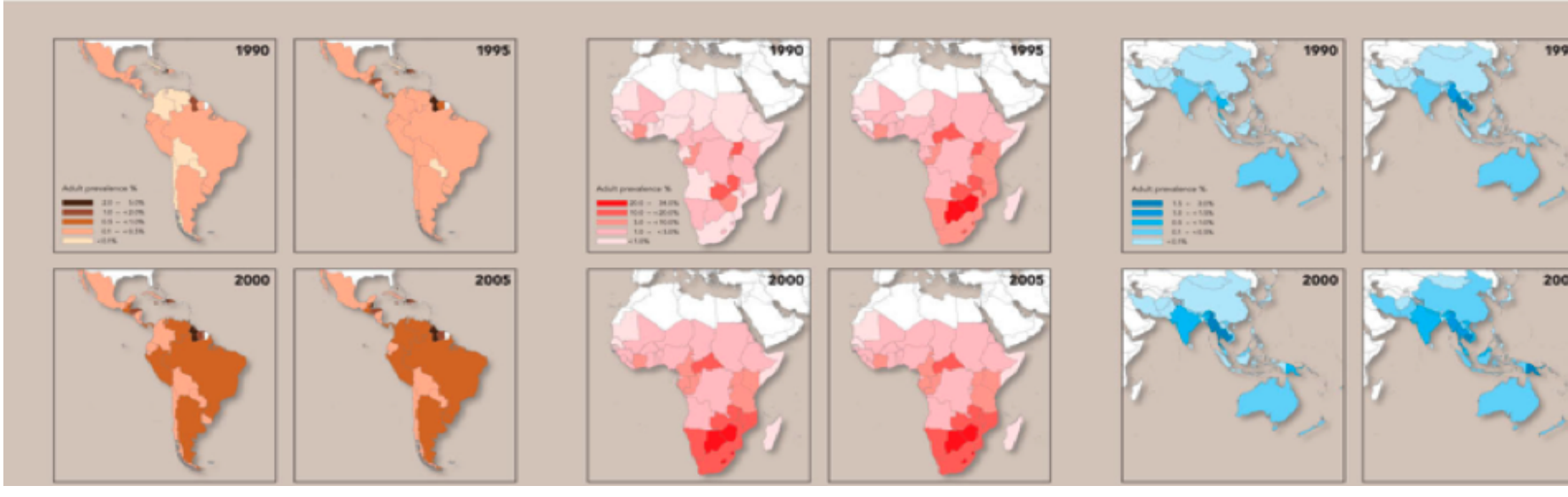
Rodrigo and Felsenstein, 1999; Drummond *et al*, 2002

Many pathogens, such as HIV, Hepatitis C and Influenza A, evolve very rapidly, so that samples of the virus population from different times directly reveal evolutionary change.



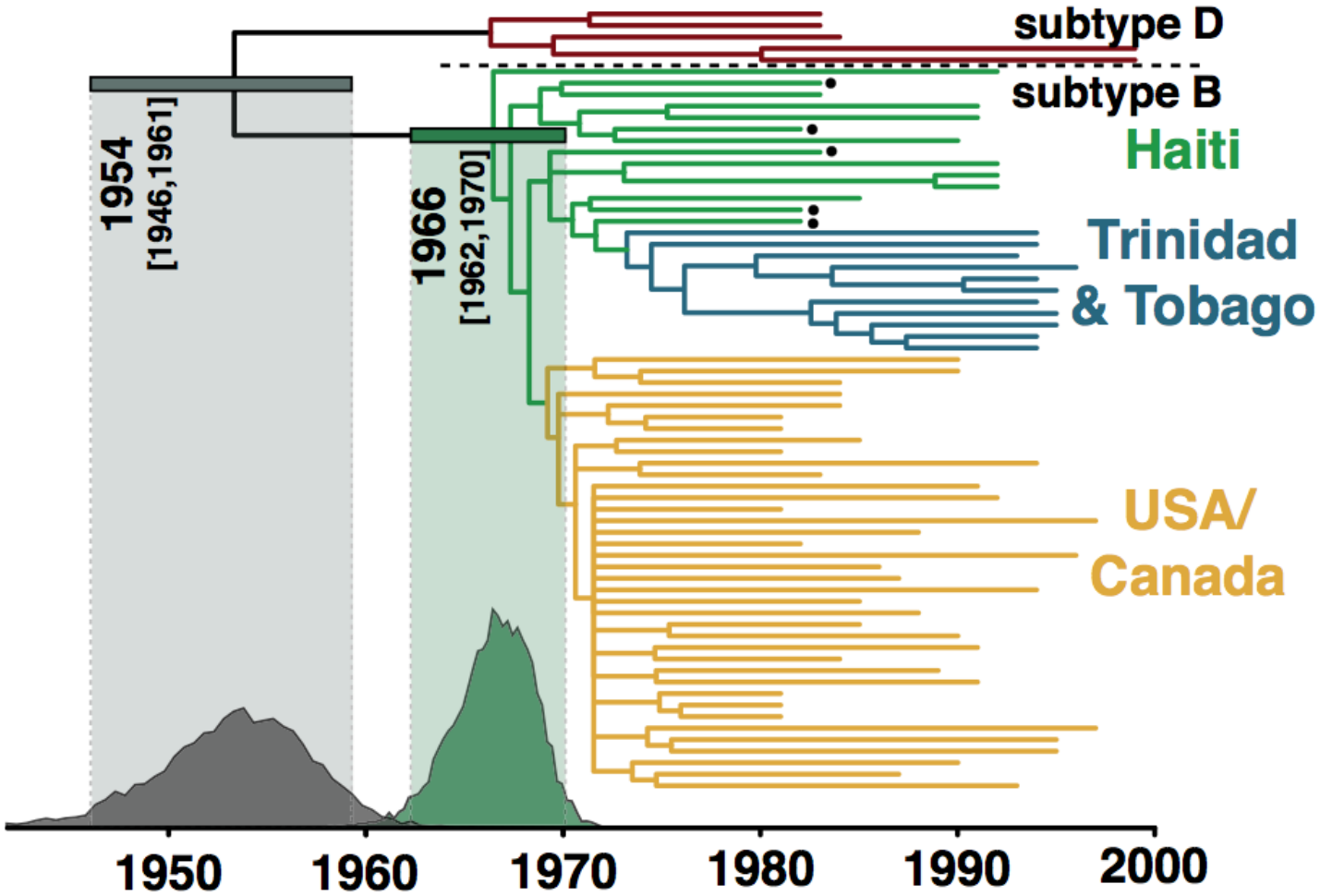
In fact it becomes possible to **calibrate** the tree and thus place the tree on a time scale - by constraining the tips to known sampling times


Adult prevalence %
15.0 – 34.0%
5.0 – <15.0%
1.0 – <5.0%
0.5 – <1.0%
0.1 – <0.5%
<0.1%
40 million people live with HIV


1990
1995
2000
2005
1990
1995
2000
2005
1990
1995
2000
2005

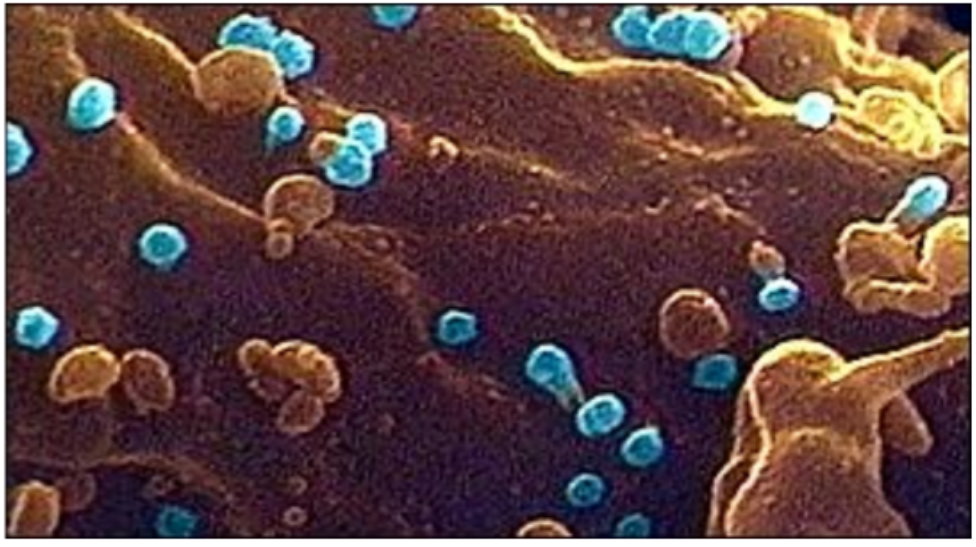
A calibrated phylogenetic inference

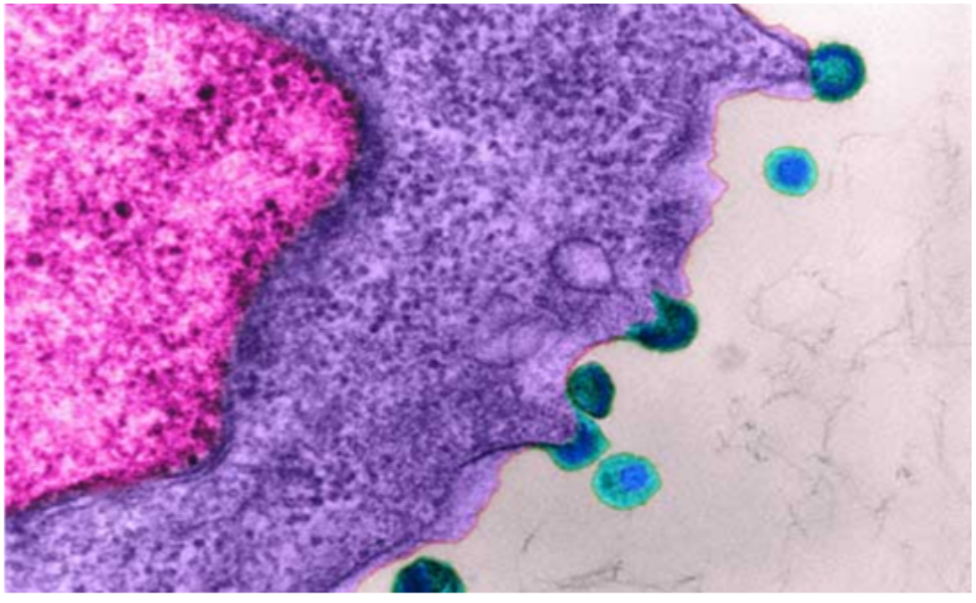
Origin of HIV Epidemic in the Americas, Gilbert *et al* (2007)



A phylogenetic reconstruction of samples of HIV-1 virus. Each degree one node represents a single infected individual from whom a blood sample has been taken.

Human immunodeficiency virus type 1(HIV-1)

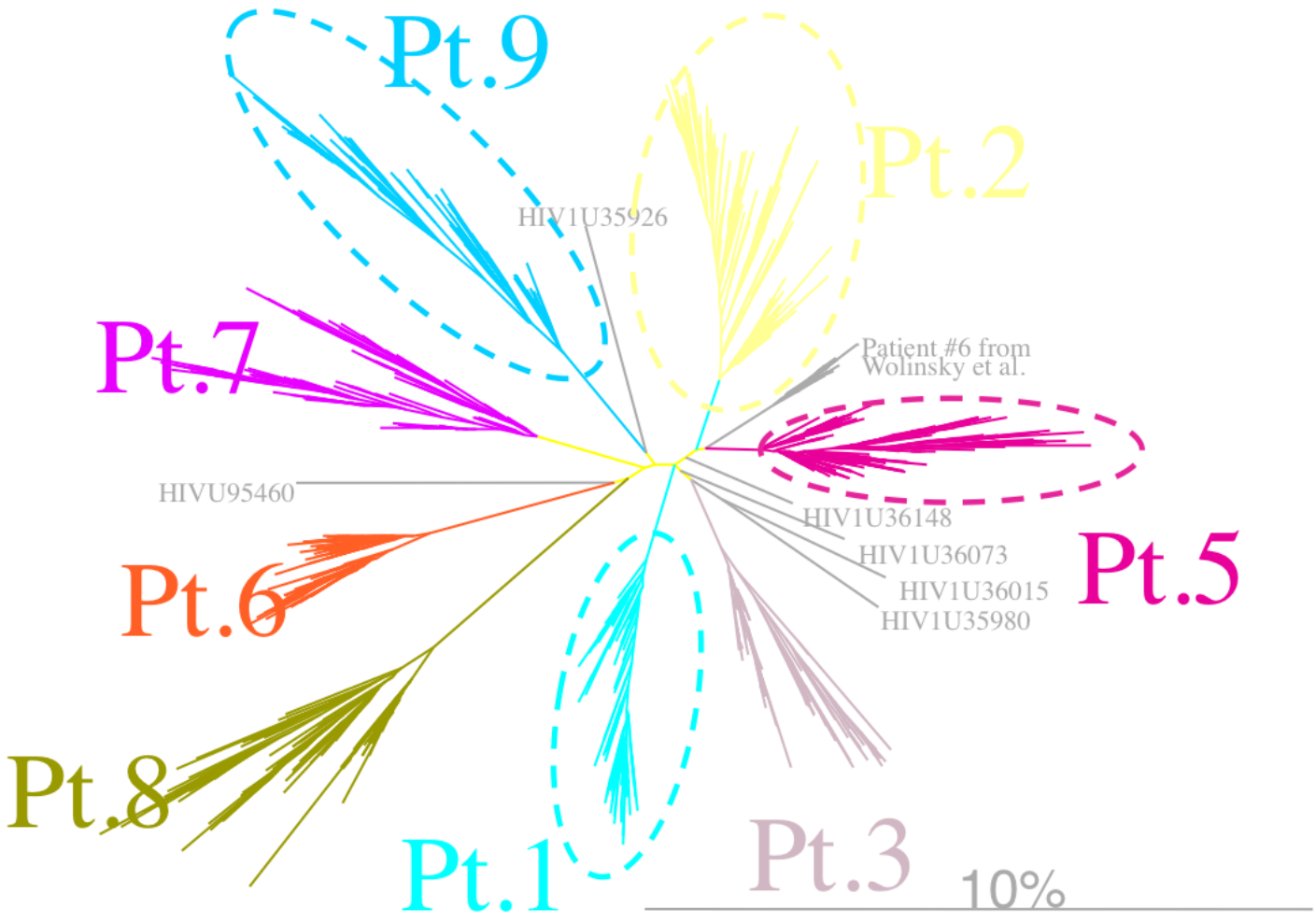




A single HIV-1 infected person has at least $10^7 - 10^8$ infected cells, with each infected cell producing $\sim 10^3$ viral particles during its life time.

A tree of HIV sequences from 9 infected patients

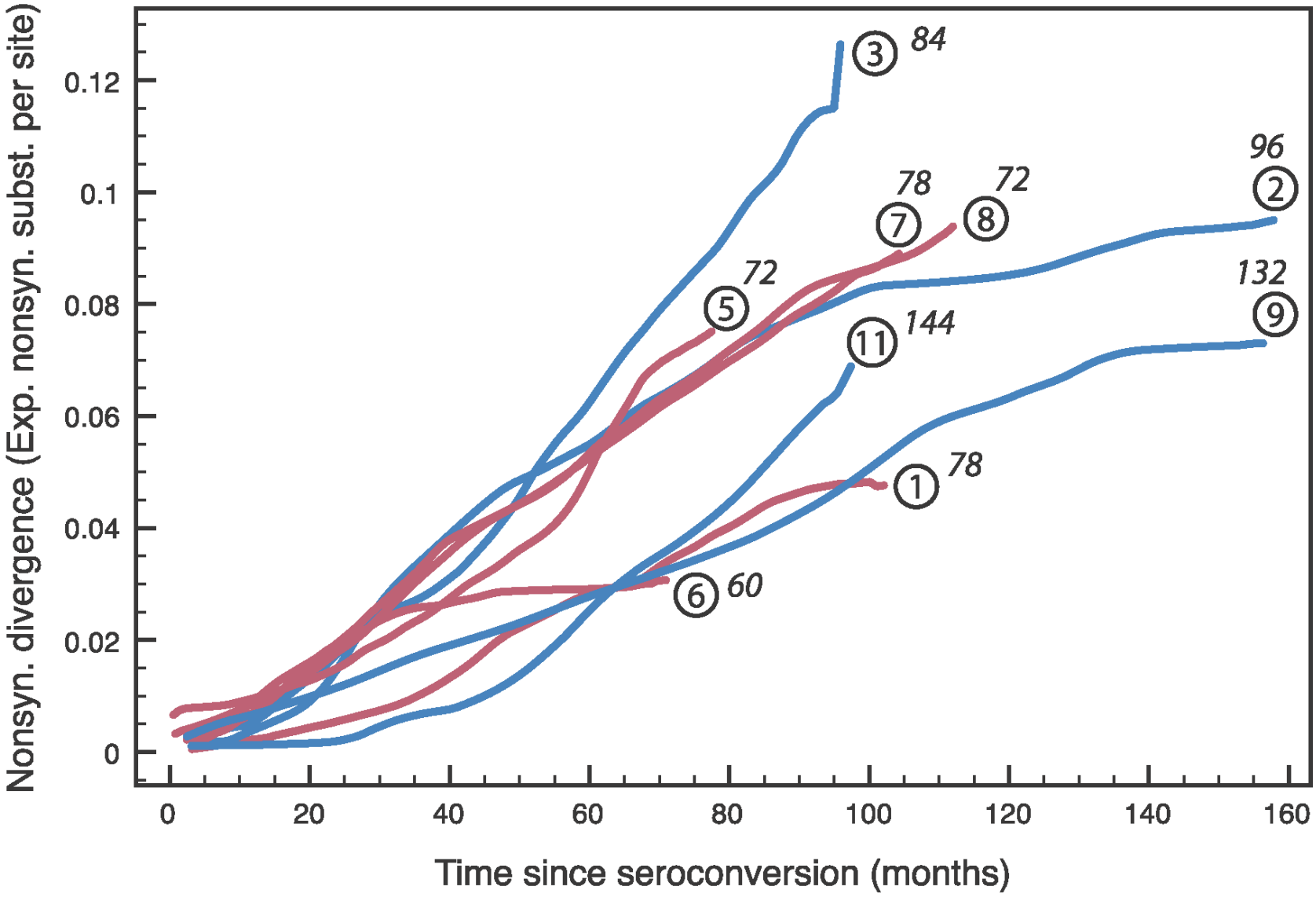
Shankarappa *et al* (1999)



A phylogenetic reconstruction of samples of HIV-1 virus. Each degree one node represents a single virus particle isolated from a blood sample of one of 9 patients.

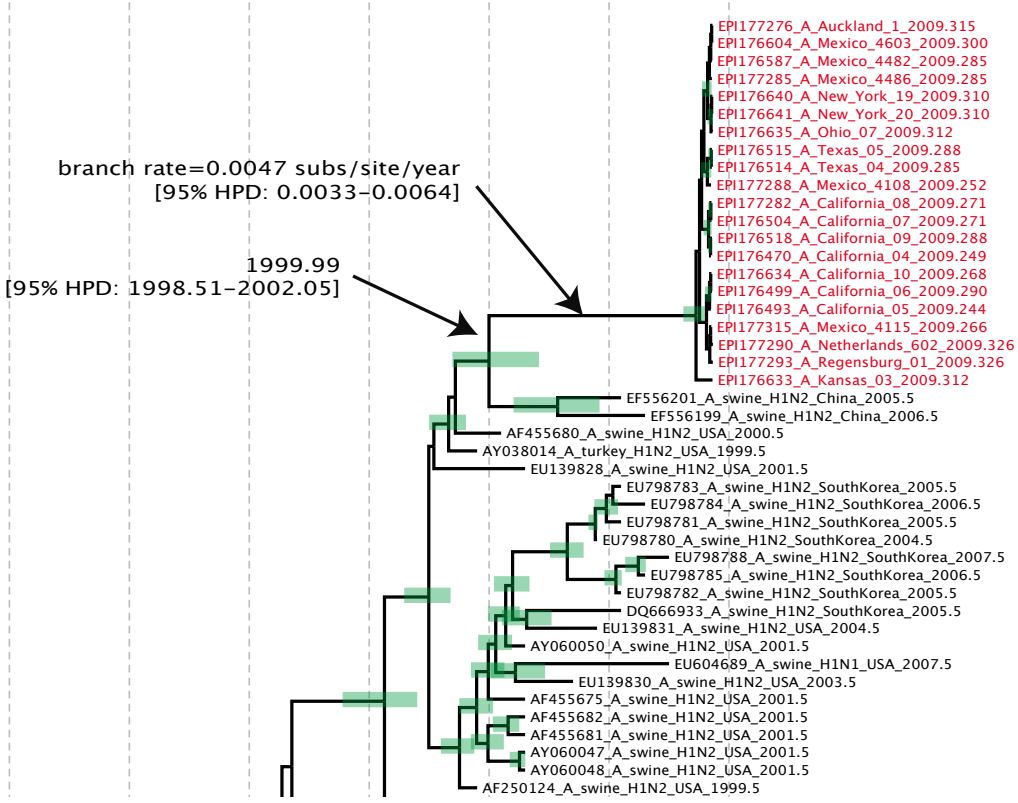
Estimated accumulation of evolutionary change

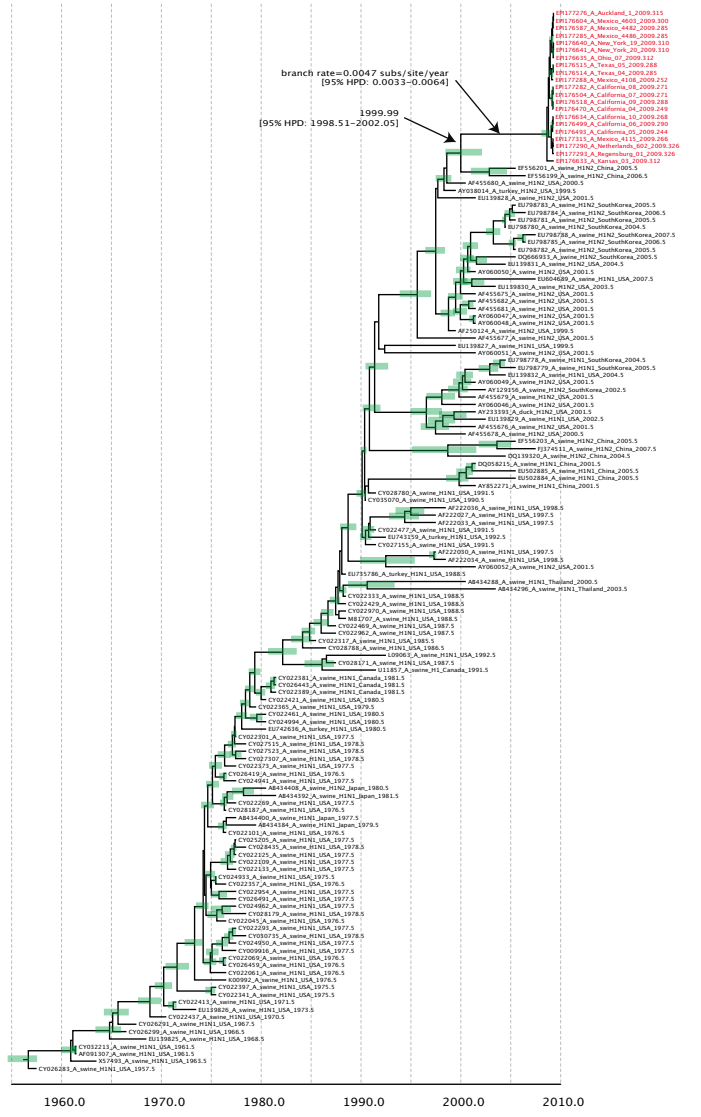
Lemey *et al* (2008)



On the Origin of 2009 H1N1 Swine Flu outbreak

http://tree.bio.ed.ac.uk/groups/influenza/



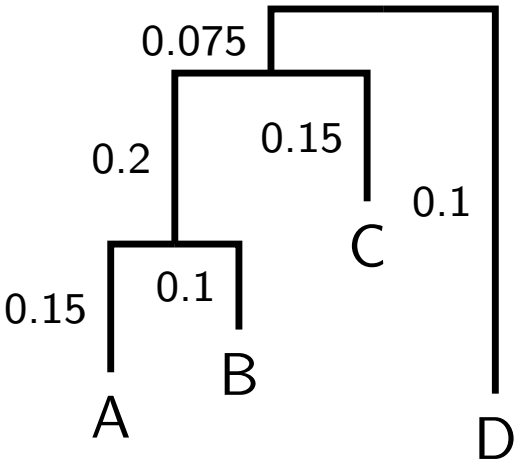


Relaxed phylogenetics

Genetic distance = rate × time

Relaxed molecular clock

$$T = \vec{\mu} \star g$$

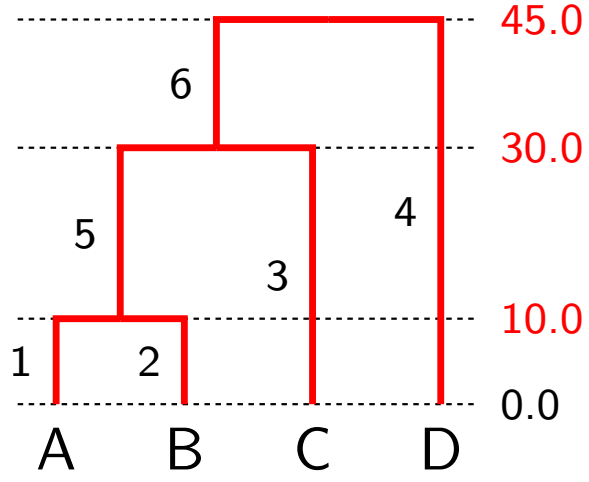


"substitution tree"

$$= \begin{pmatrix} 0.015 \\ 0.01 \\ 0.005 \\ 0.01 \\ 0.01 \\ 0.005 \end{pmatrix} \star$$

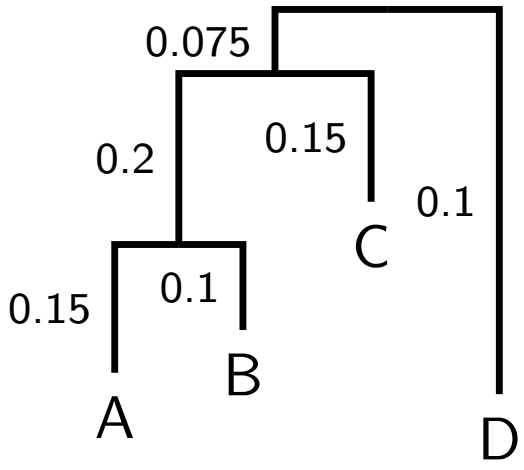
evolutionary rates

substitutions / site / unit time

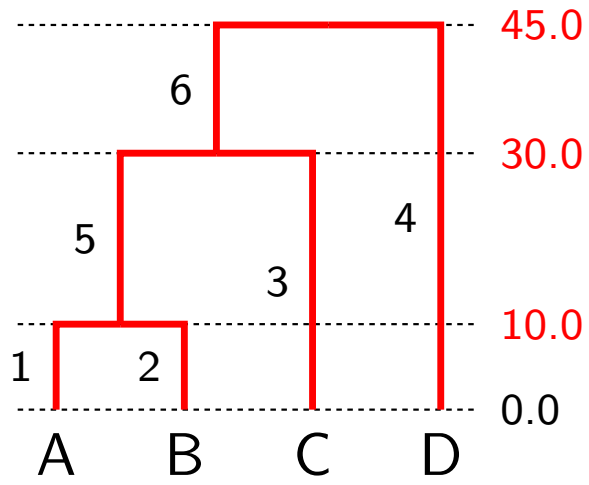


time tree

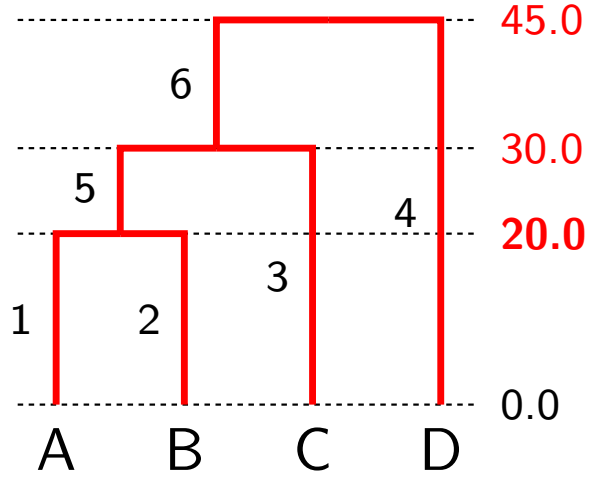
Nonidentifiability in the relaxed clock



$$= \begin{pmatrix} 0.015 \\ 0.01 \\ 0.005 \\ 0.01 \\ 0.01 \\ 0.005 \end{pmatrix} \star$$



$$= \begin{pmatrix} \mathbf{0.0075} \\ \mathbf{0.005} \\ 0.005 \\ 0.01 \\ \mathbf{0.02} \\ 0.005 \end{pmatrix} \star$$

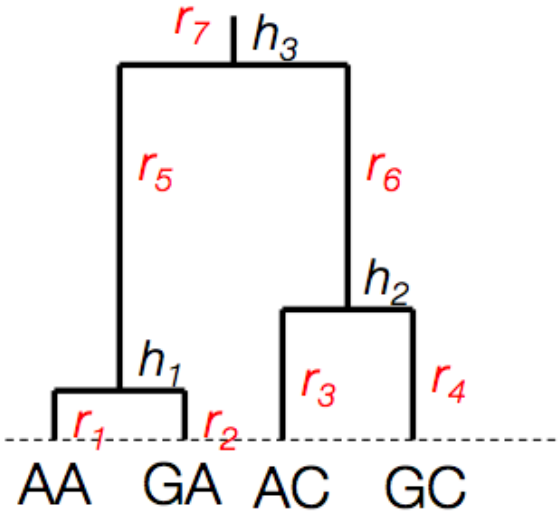


"substitution tree"

evolutionary rates
substitutions / site / unit time

time tree

Relaxing the molecular clock



In the field of divergence time estimation auto-correlated relaxed clocks have been considered.

e.g. Thorne et al, 1998:

$$r_i \sim LogNormal(r_{A(i)}, \sigma^2 \Delta t_i)$$

AC

$r \sim Exp(\lambda)$

$r \sim LogNormal(\mu, \sigma^2)$

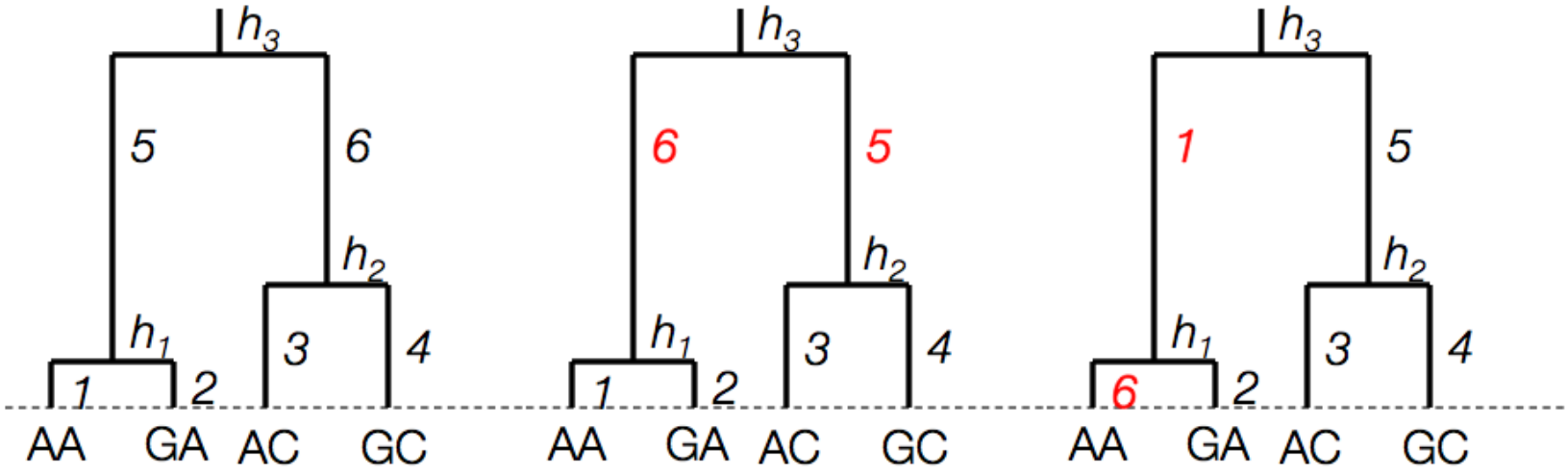
$r \sim Gamma(\alpha, \beta)$

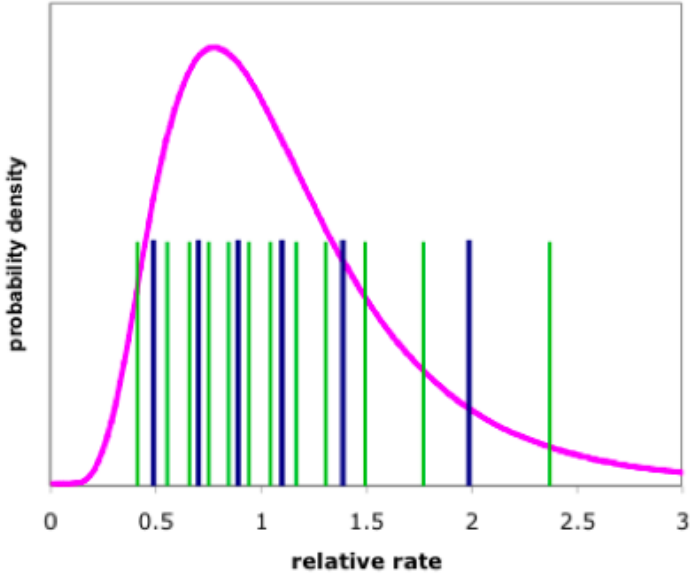
We introduce a relaxed clock model in which there is no prior correlation between child and parent rates

"Un-correlated" or "memory-less" relaxed clocks

ML

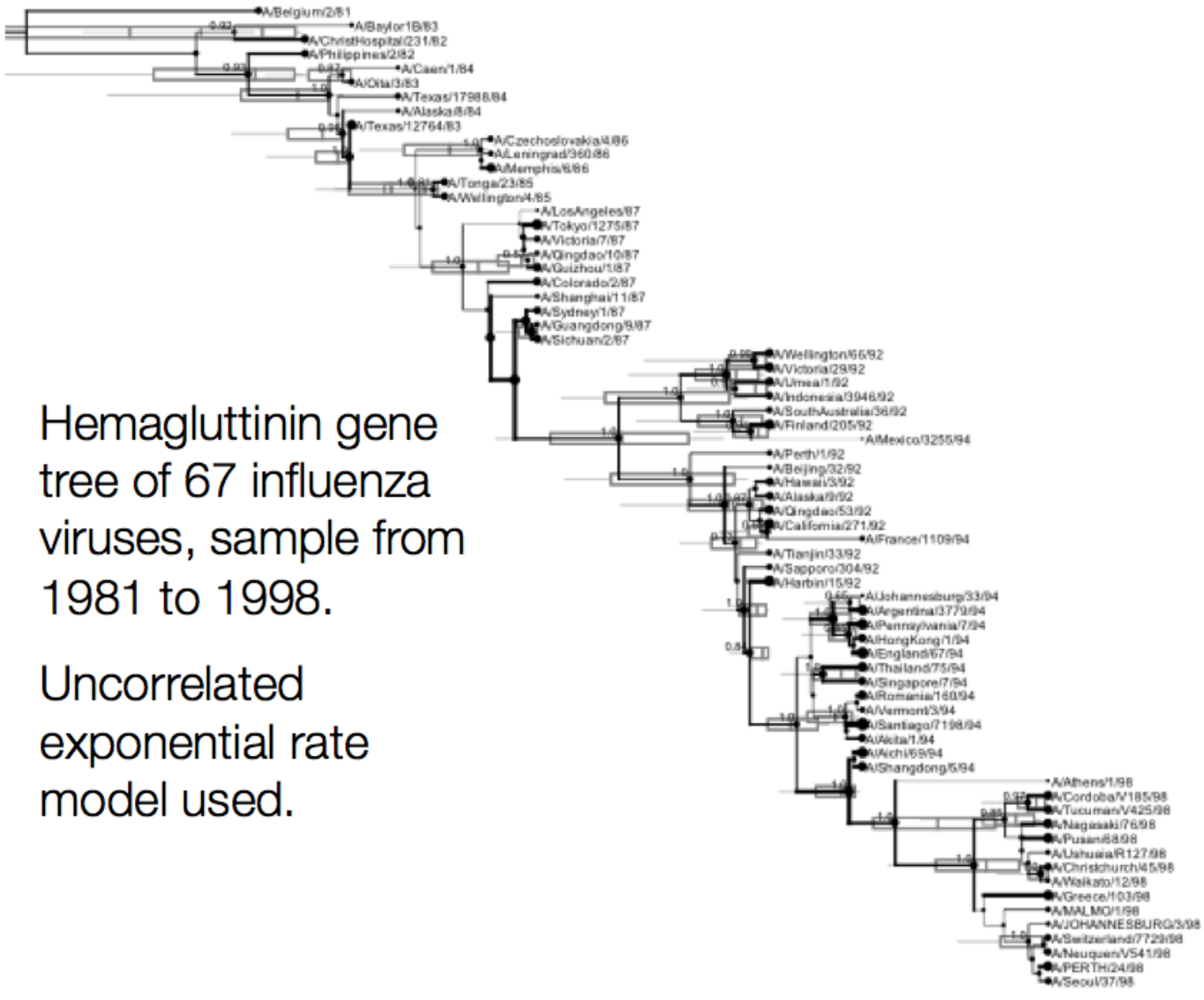
Sampling branch rates using MCMC





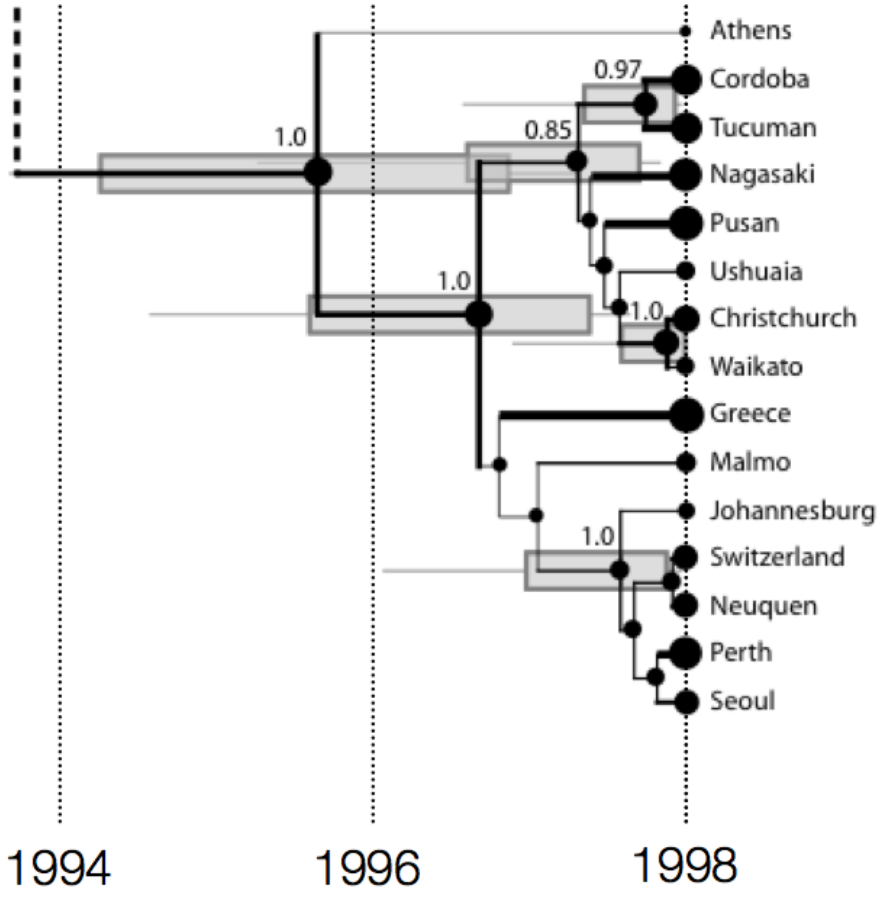
1. Rates are summarized into 2n-2 rate categories (e.g. blue is 6 categories; green is 12 categories).
2. Rates categories are sampled during MCMC by two operators:
    1. Random walk operator
    2. Swap operator
3. For purposes of topology changes, rate categories are associated with child node.

Influenza A gene tree estimated by relaxed molecular clock



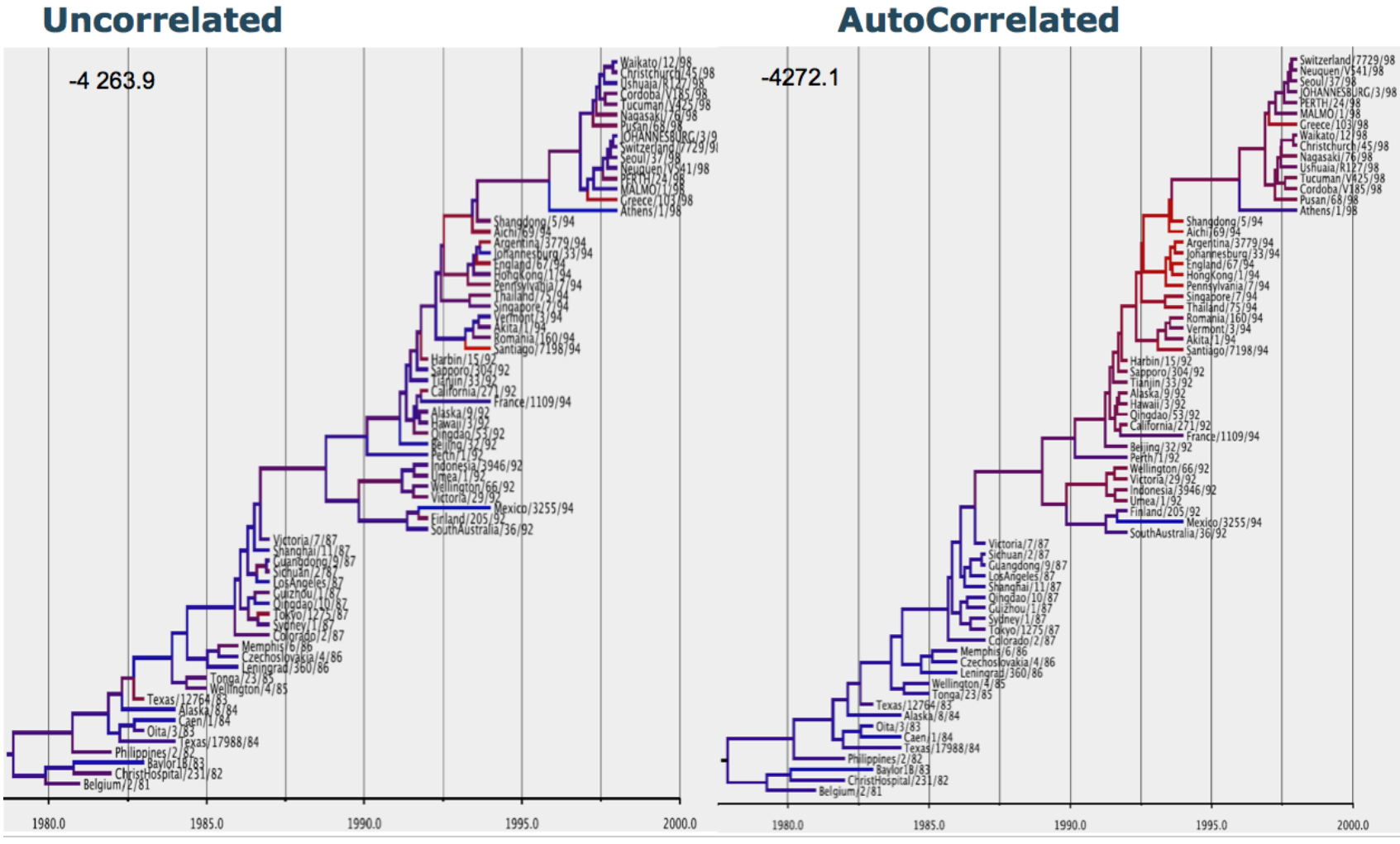
1. Hemagluttinin gene tree of 67 influenza viruses, sample from 1981 to 1998.
2. Uncorrelated exponential rate model used.

Influenza A gene tree estimated by relaxed molecular clock

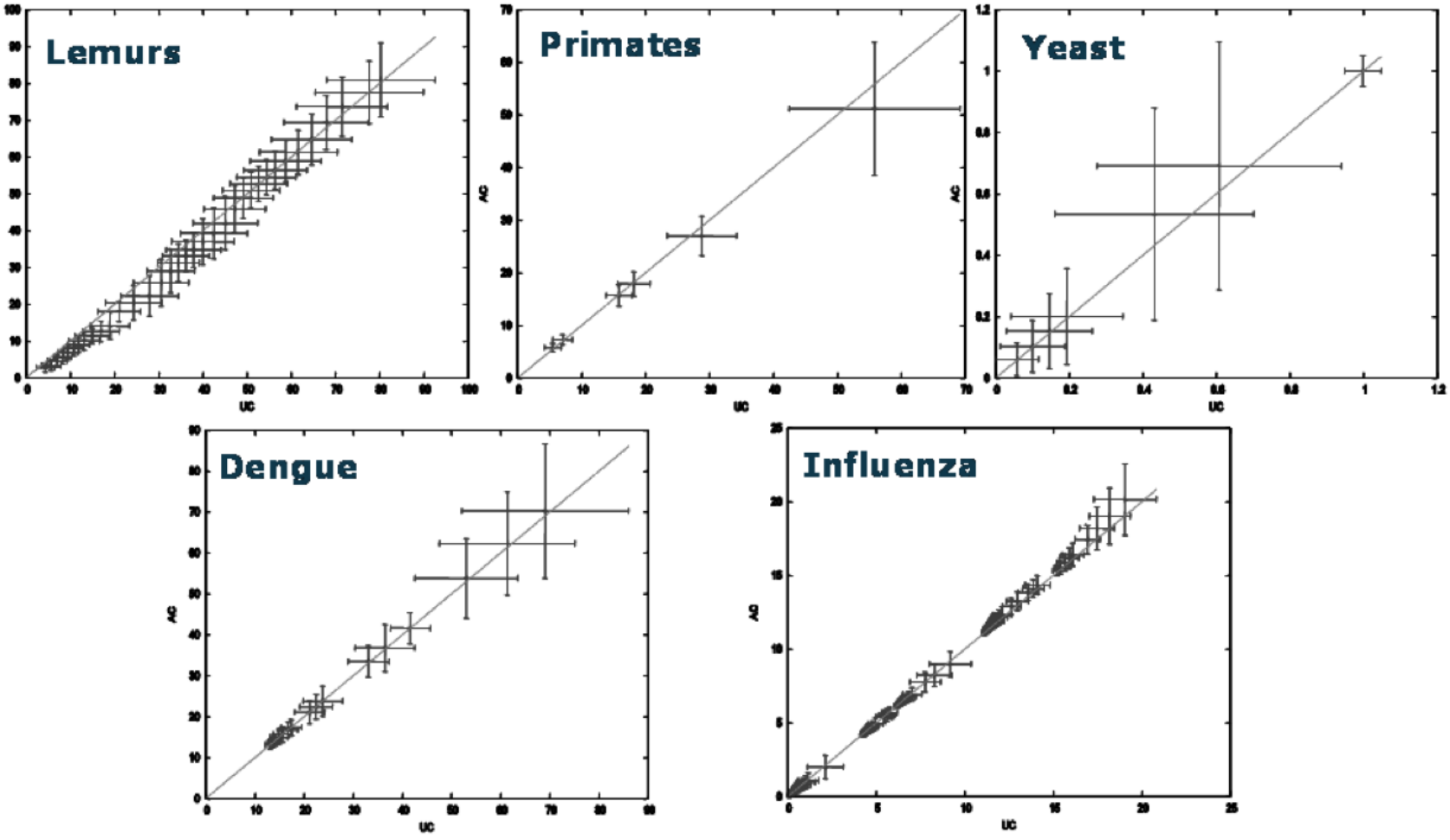


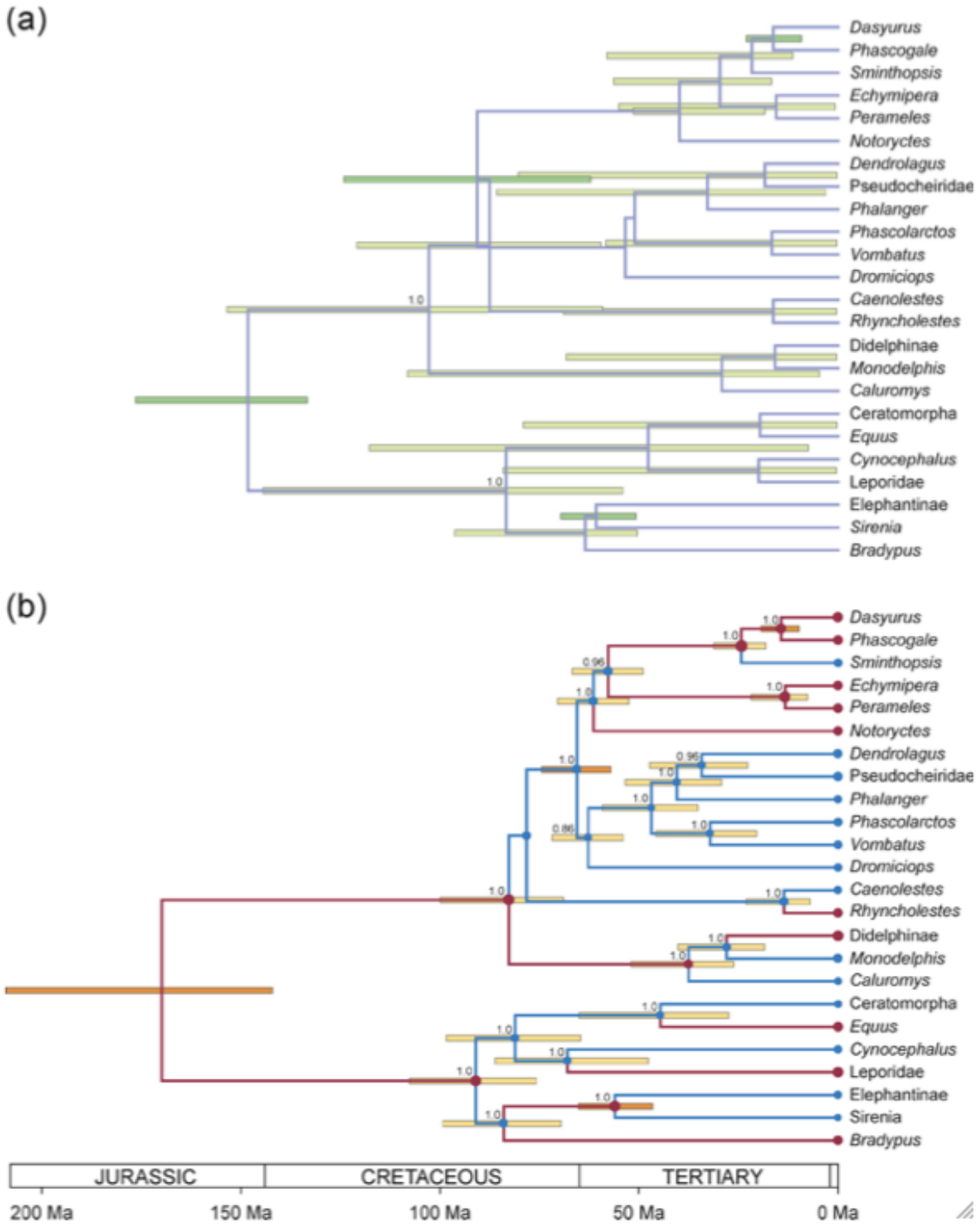
- Box-and-whisker plots show uncertainty in divergence times (only for splits with posterior probability > 0.5)
- Node size and branch thickness proportional to evolutionary rate.

Influenza trees under different relaxed clock models



UC versus AC on five data sets



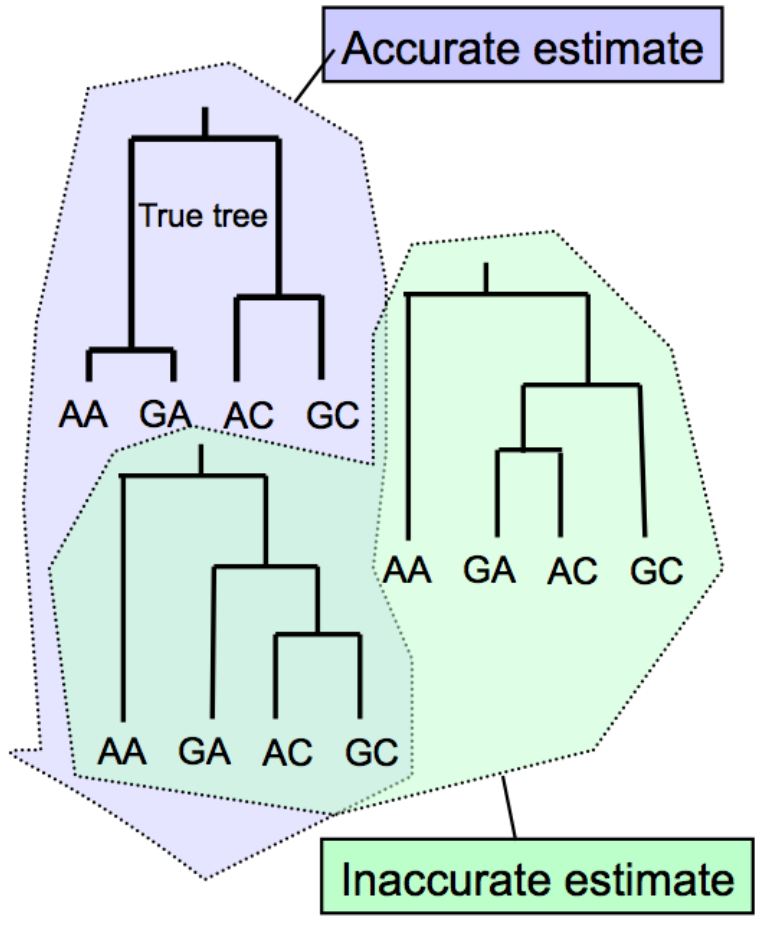


Prior versus Posterior

Marsupials example (24 taxa, 5658 nucleotides)

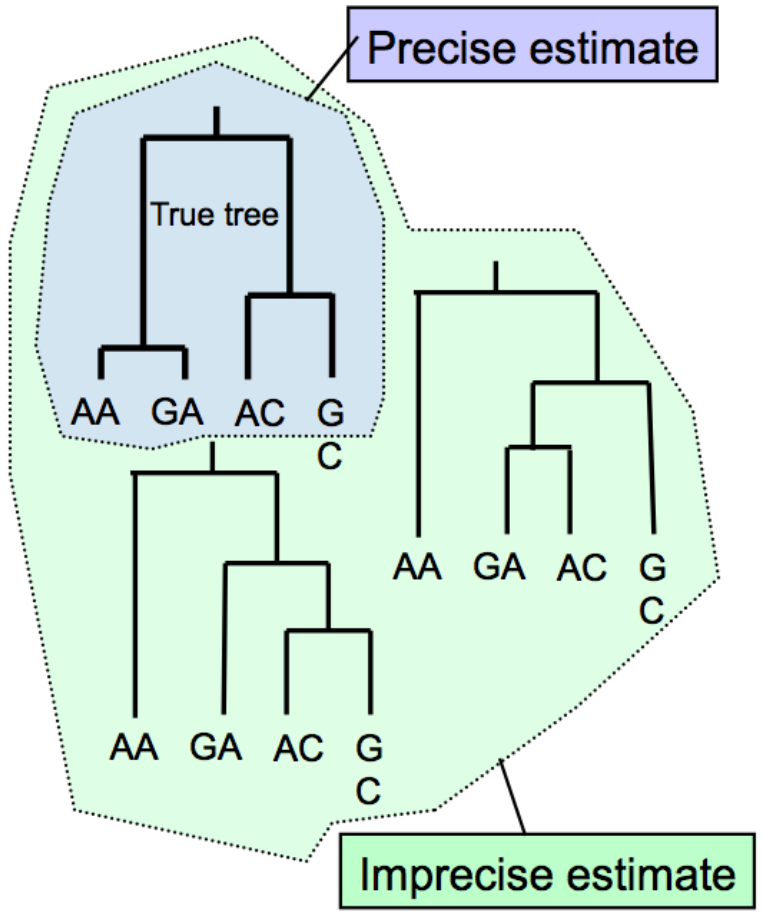
Accuracy in Bayesian Phylogenetics

- Phylogenetics is an estimation problem, in which the phylogenetic tree topology is the object we wish to estimate.
- The error associated with this estimation can be described by the 95% credible set of trees: the smallest set of trees including 95% of the posterior probability.
- A standard measure of accuracy is the false positive rate. How often do we exclude the true tree from the 95% credible set?



Precision in Bayesian Phylogenetics

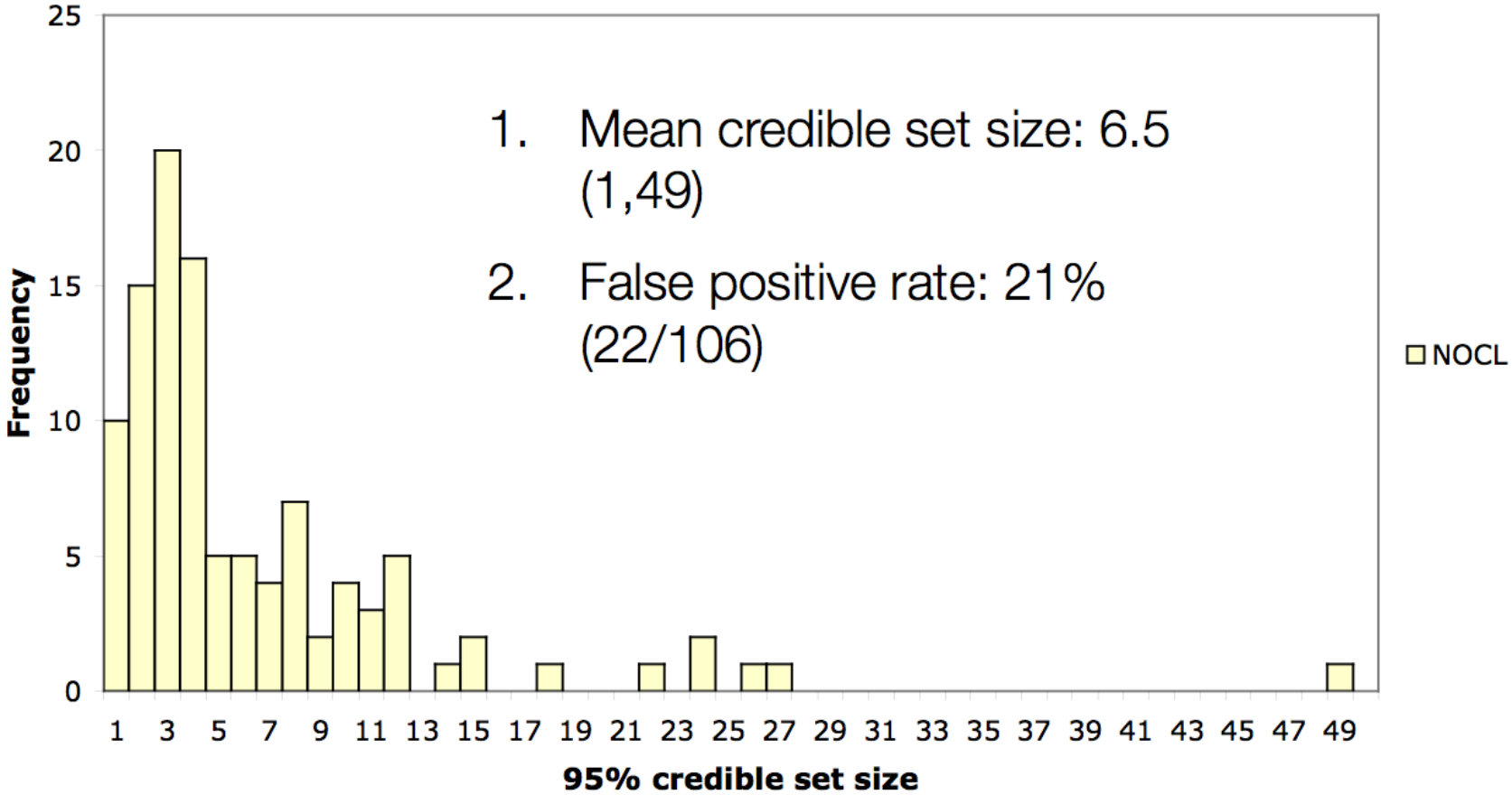
- The precision of an estimate can be described by how much is excluded.
- How small is the 95% credible set of trees?



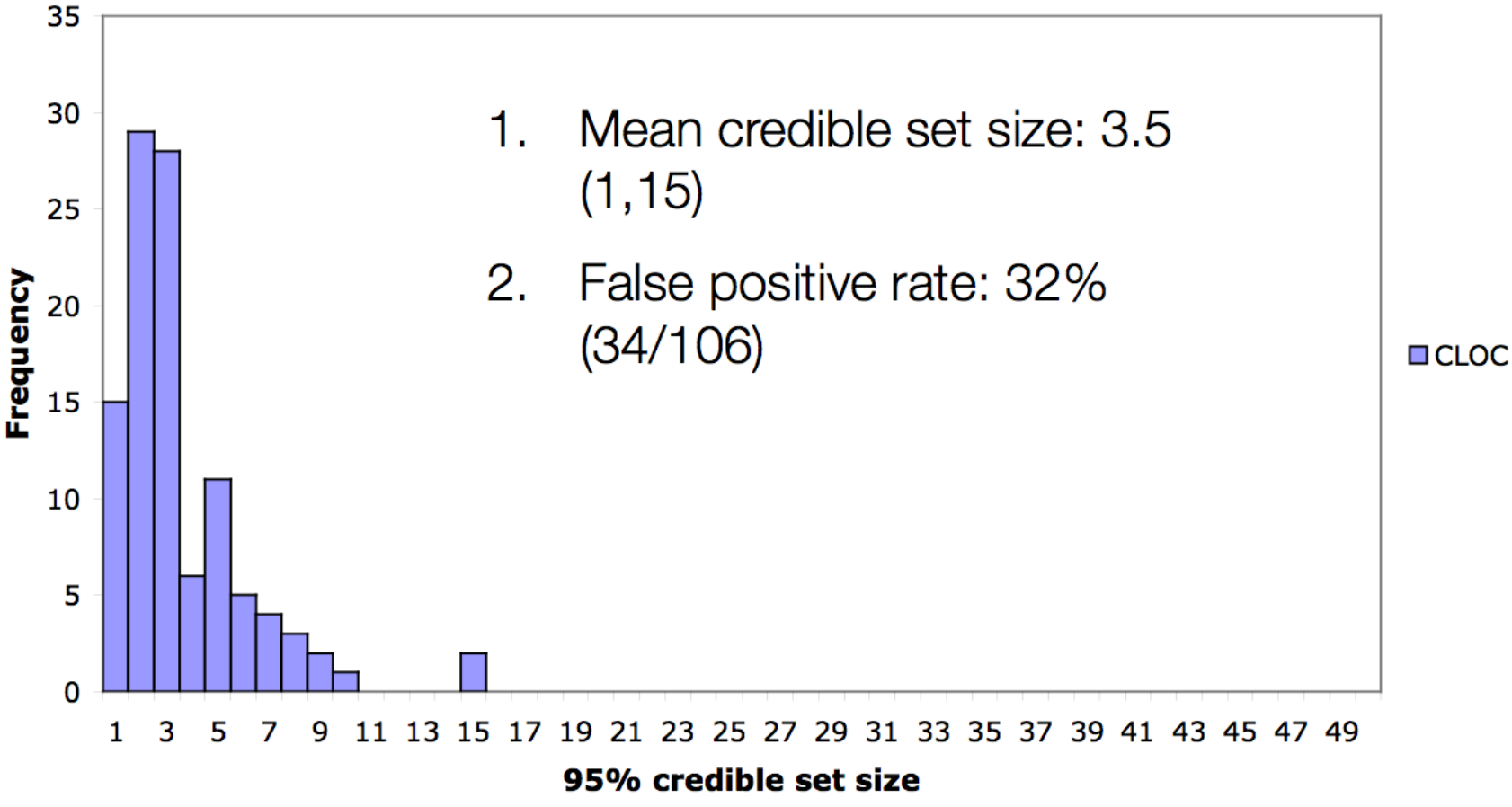
Testing Accuracy and Precision with **real data**

- Used 106 genes from 8 species of yeast (Rokas *et al*, 2003) and 4 other "phylogenomic" data sets
- For each gene used both MrBayes and BEAST to estimate phylogeny and 95% credible set
- Assumed true tree is the tree estimated using all the concatenated data set.
- Tabulated number of trees in credible set and whether the true tree was in credible set for MrBayes (unconstrained) and BEAST (MLLN and CLOC models)

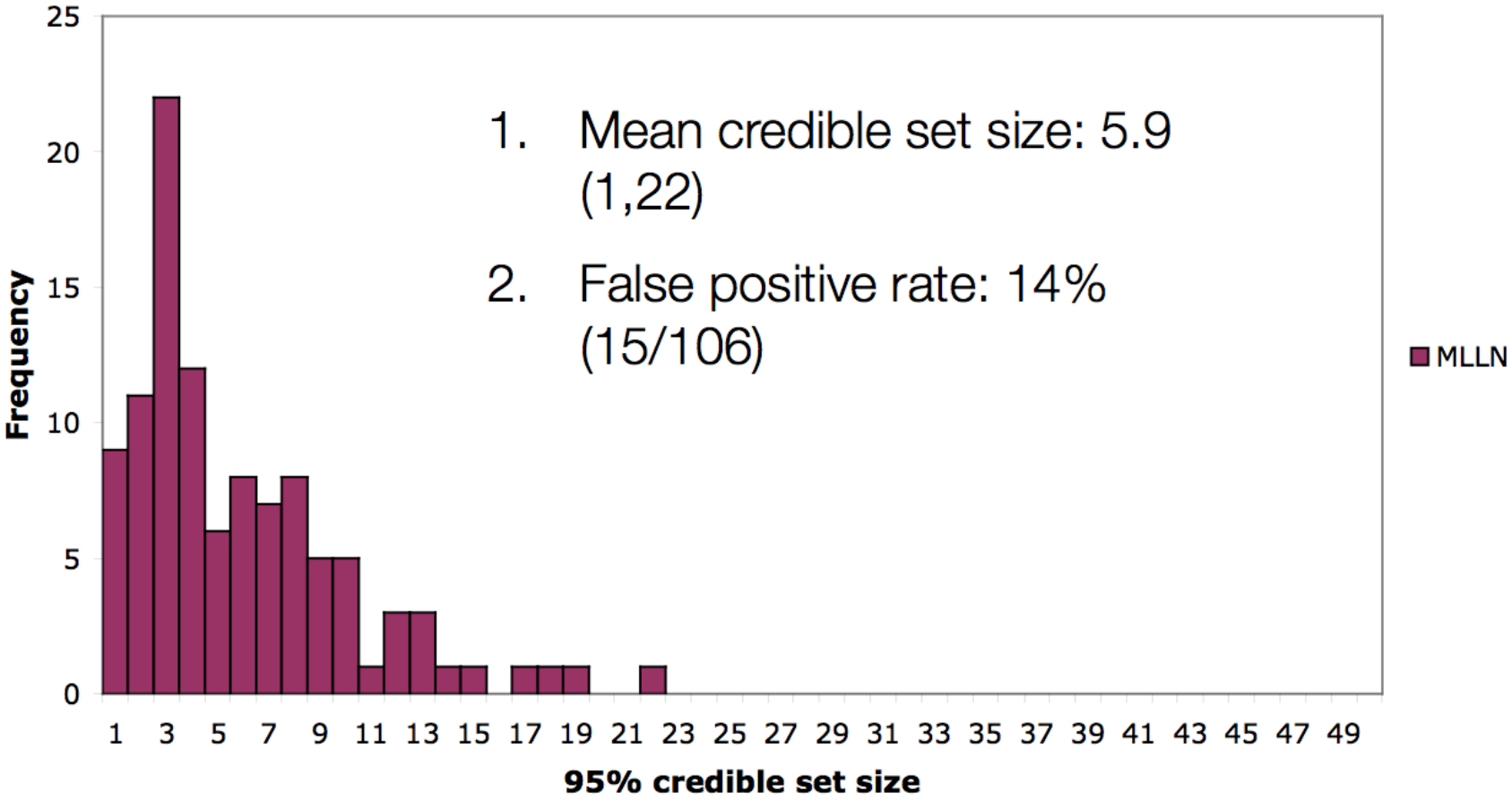
Rokas data: MrBayes tree estimates



Rokas data: Strict clock tree estimates from BEAST



Rokas data: Relaxed clock tree estimates from BEAST



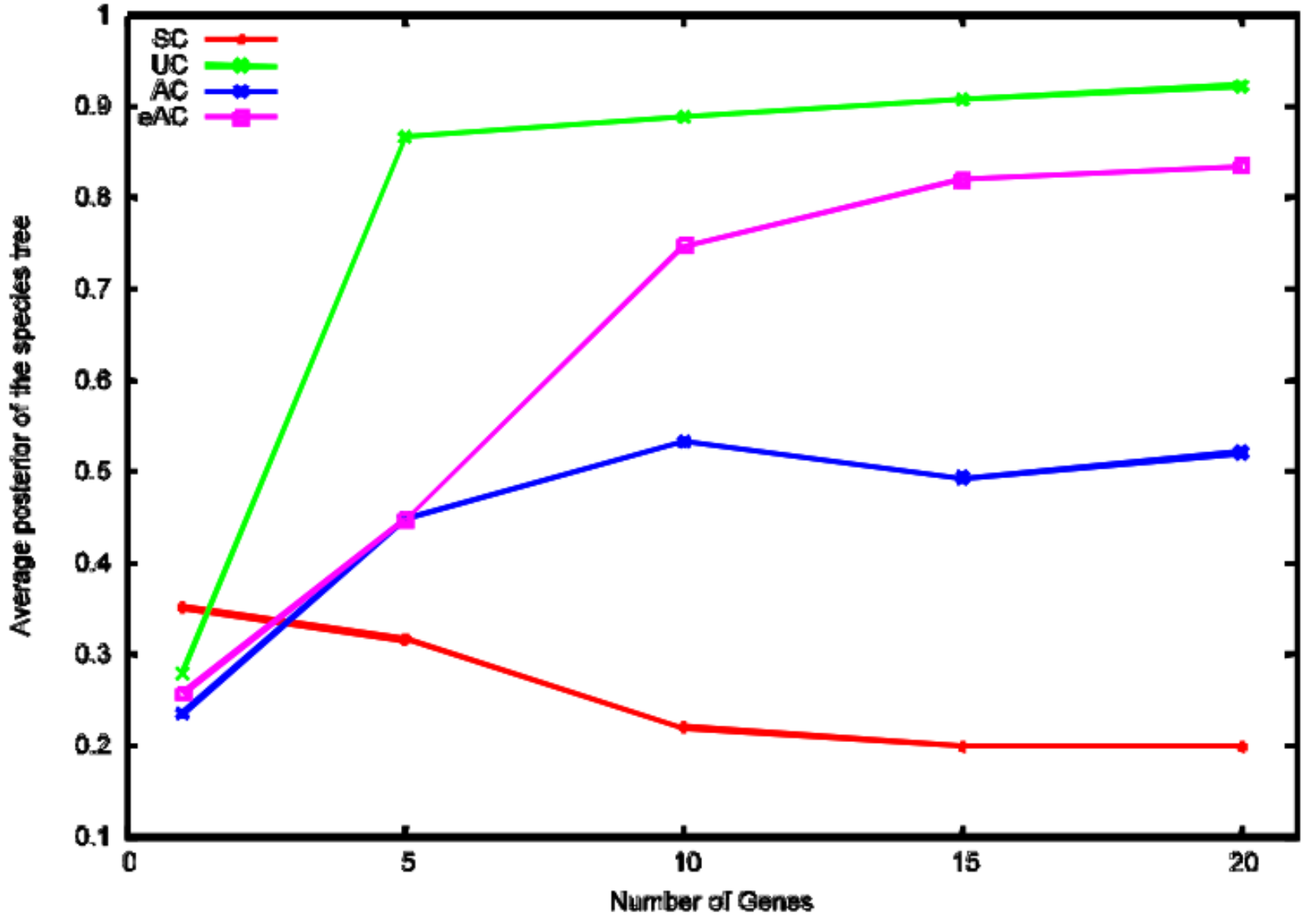
Summary of Bayesian Accuracy on five large data sets

Dataset	Sample Size	Average Length	Clock Rejected by LRT	Accuracy (%) (True Tree in 95% Credible Set)[a]		
				CLOC	UCLN	UF
Bacteria	102	170 aa	26%	46.1	**48.0**	42.2
Yeast	106	1,198 bp	76%	67.0	**84.9**	79.2
Plants	61	647 bp	67%	**91.8**	88.5	83.6
Animals	99	197 aa	59%	64.6	**69.7**	57.6
Primates	500	632 bp	13%	88.8	**89.0**	88.8

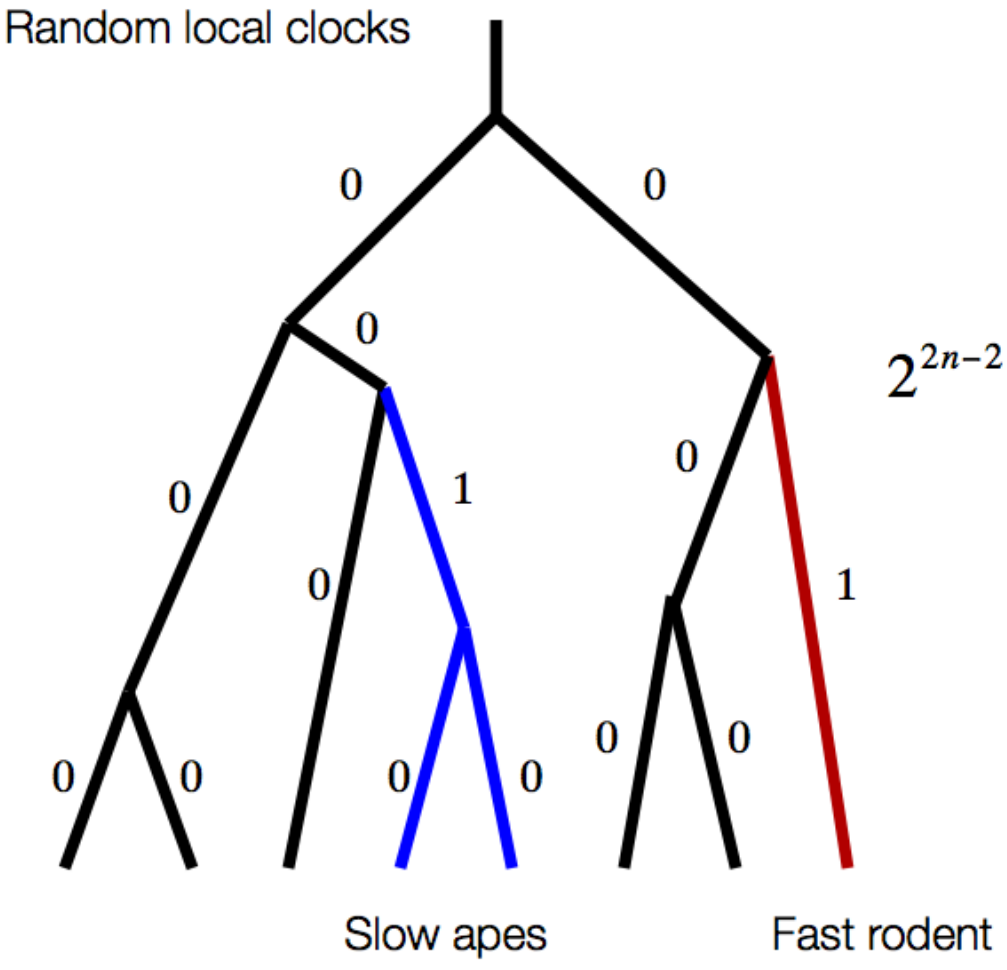
Summary of Bayesian Precision on five large data sets

Dataset	Sample Size	Average Length	Clock Rejected by LRT	Precision (Number of Trees in 95% Credible Set)[b]		
				CLOC	UCLN	UF
Bacteria	102	170 aa	26%	5.7	10.3	11.3
Yeast	106	1,198 bp	76%	3.5	5.9	6.5
Plants	61	647 bp	67%	7.5	15.4	9.2
Animals	99	197 aa	59%	5.7	10.2	14.2
Primates	500	632 bp	13%	3.1	3.4	5.1

Increasing the length of the sequence

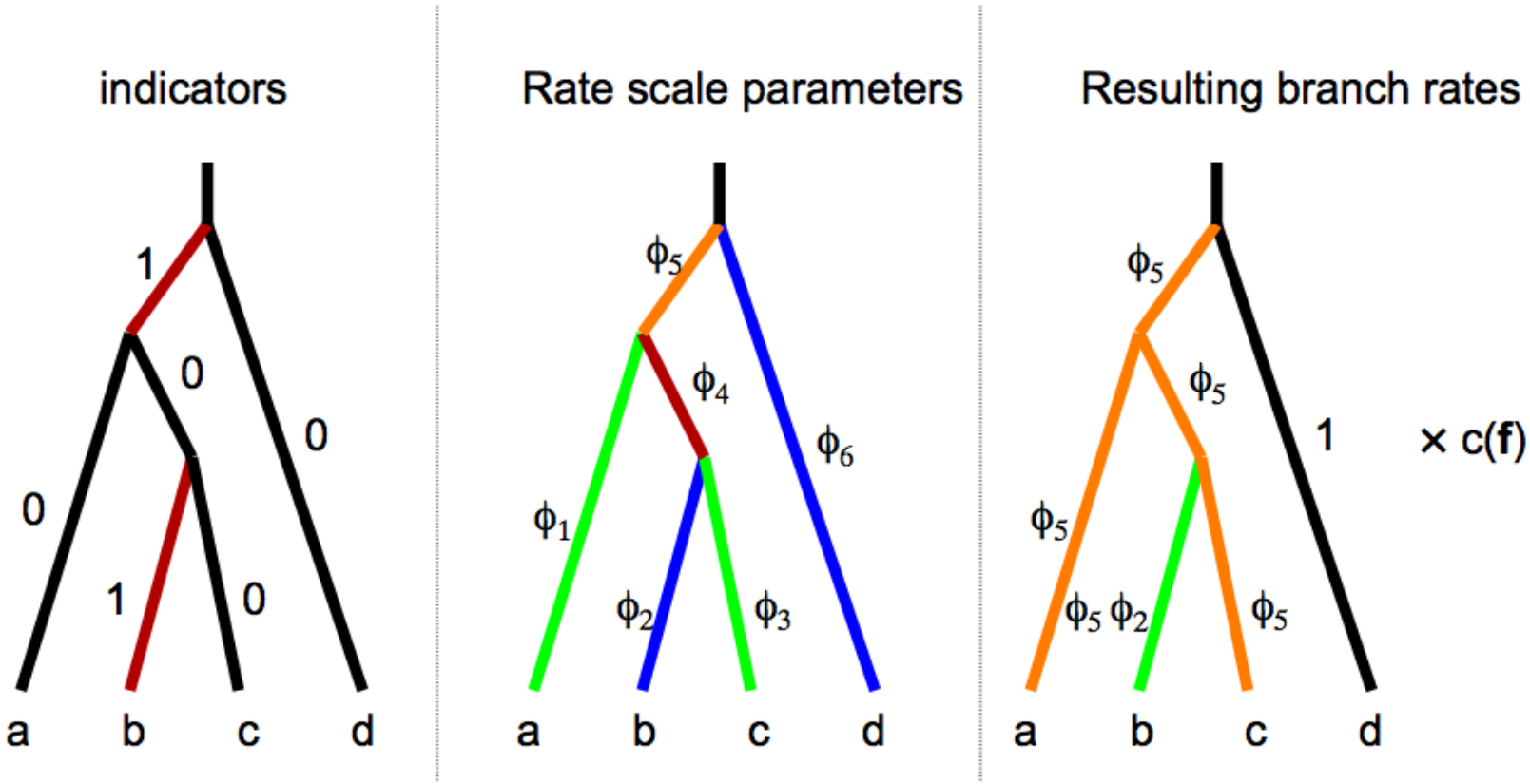


Random local molecular clocks



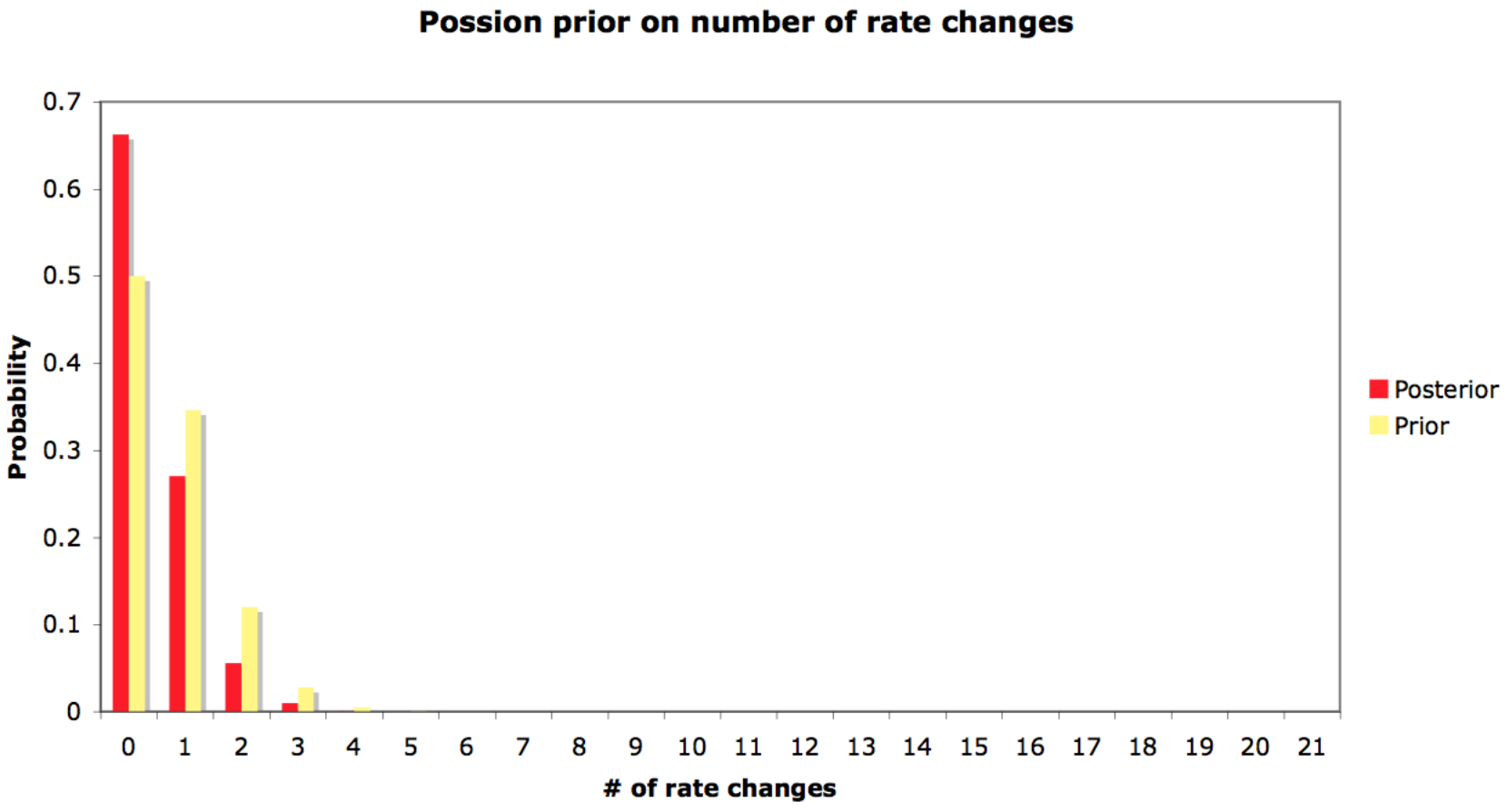
2^{2n-2} local molecular clock models

Random local molecular clocks

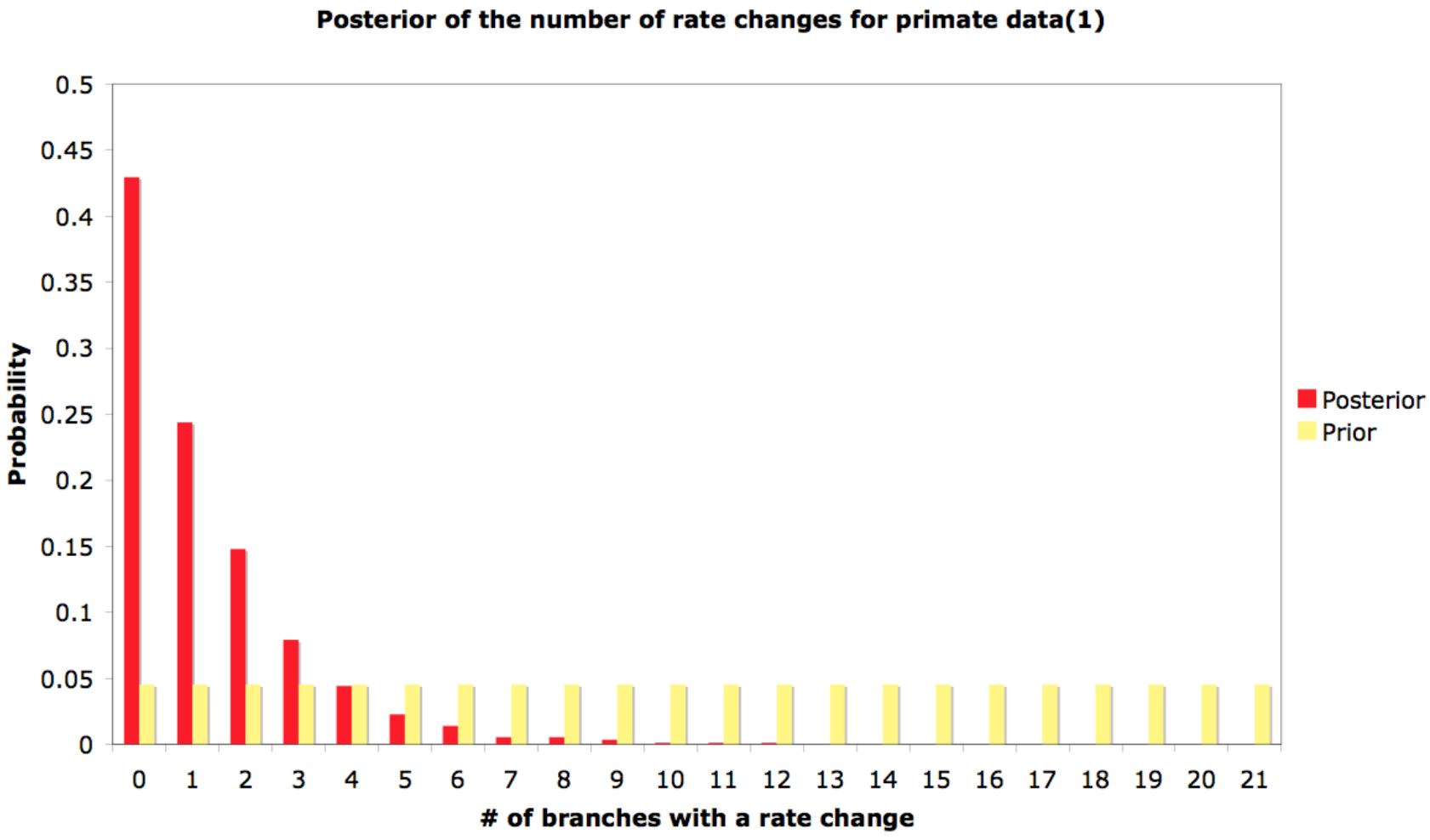


Red/Orange fast, Green/Blue slow

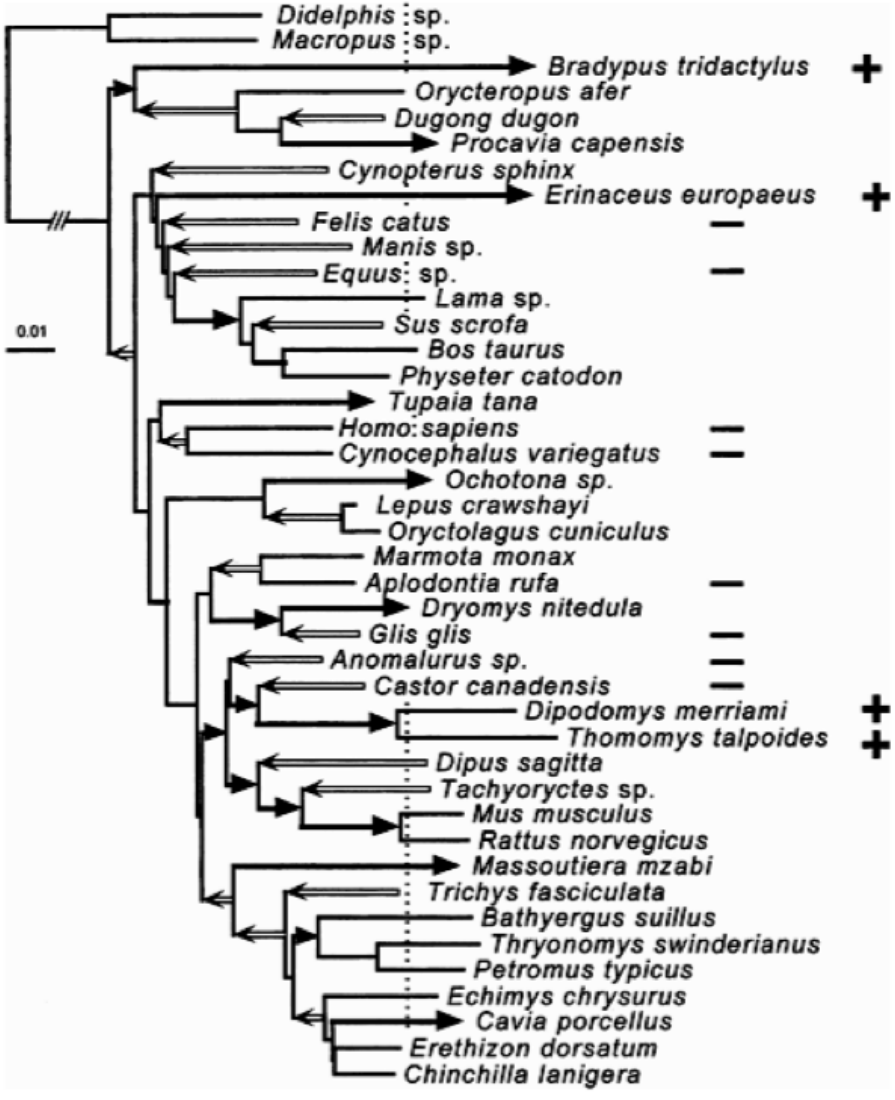
Primate data set (Poisson prior on # rate changes)



Primate data set (Uniform prior on # rate changes)



Rodents (1+2 codon positions from 3 nuclear genes)

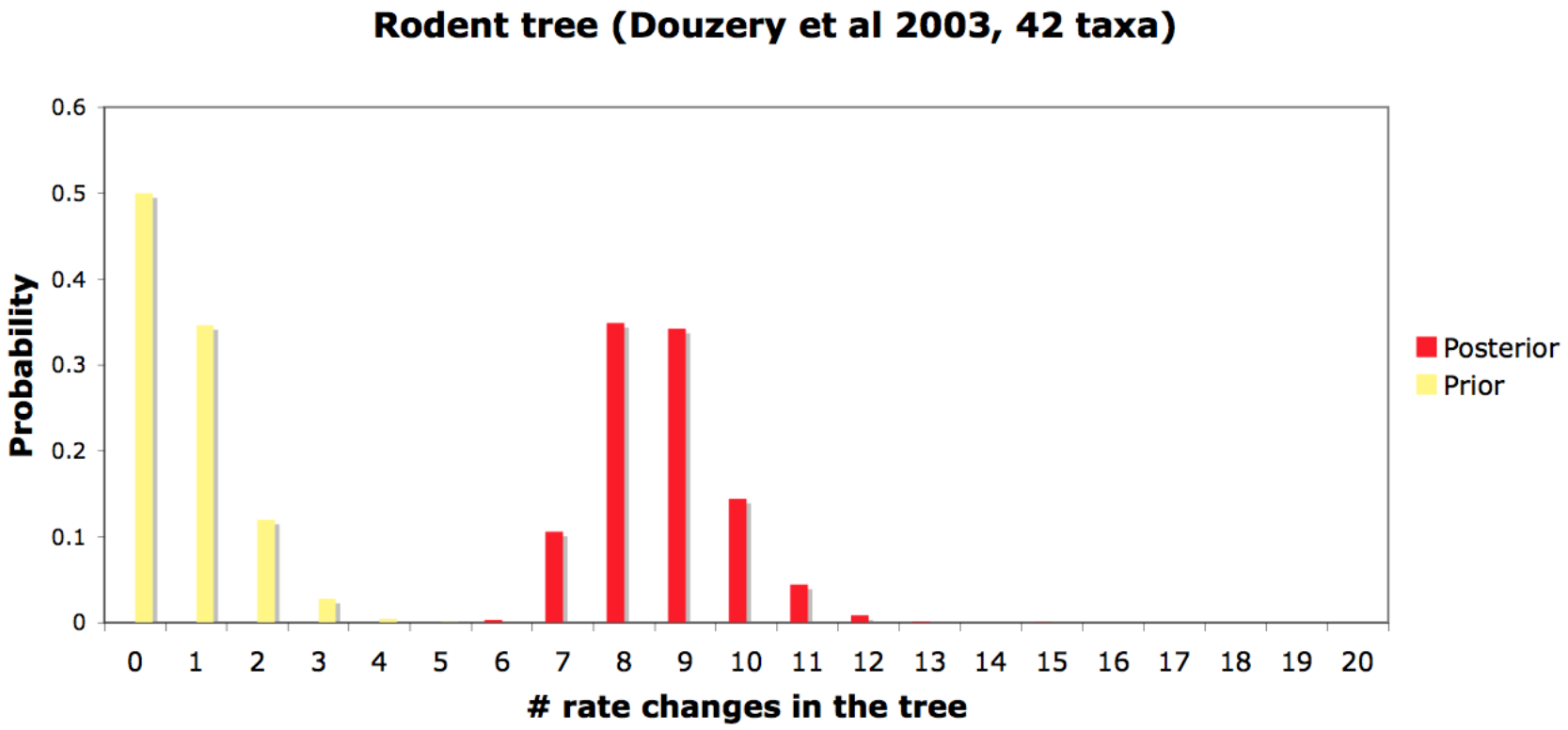


82 branches

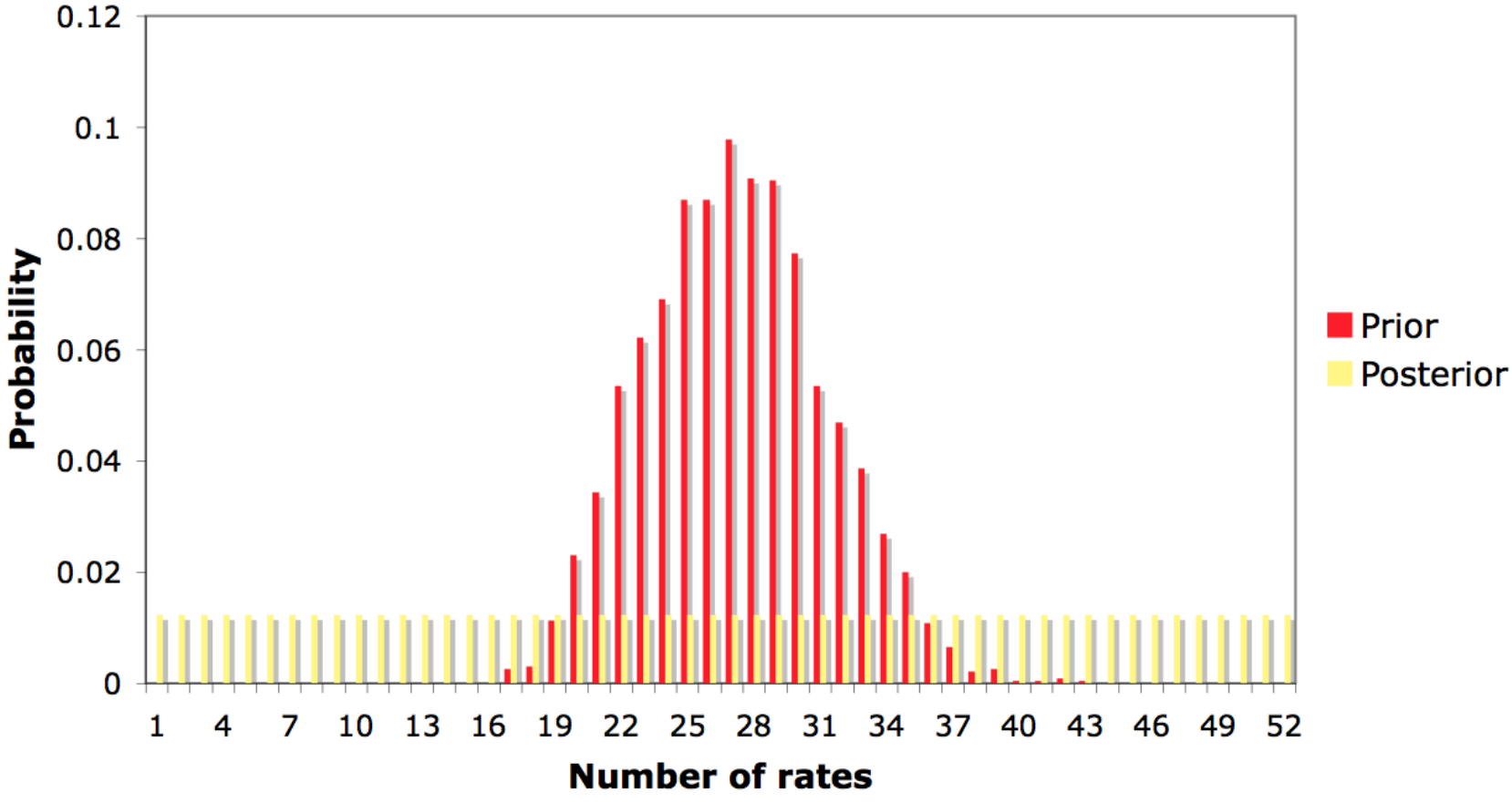
38 rate changes according to Douzery et al 2003

Fig. 1. Extensive nucleotide substitution rate variations in the first two codon positions of the ADRA2B + IRBP + vWF nuclear genes between placental mammals. The *vertical dashed line* indicates the mean value of the root-to-tip distance of the 40 placental taxa. Significantly faster- or slower-evolving species are indicated, respectively, by a + or a − as evidenced by the branch-length test. Significantly faster- and slower-evolving branches as evidenced by the two-cluster test are indicated, respectively, by *filled arrows* pointing right and *open arrows* pointing left. The scale unit corresponds to the expected number of nucleotide substitutions per site. The log-likelihood of this tree is $\ln L = -26{,}054.36$, and its AIC is 52282.78. In the clock-like constrained model—with a single global clock—a significant loss of log-likelihood is observed ($\ln L = -26{,}222.37$, AIC = 52,538.74).

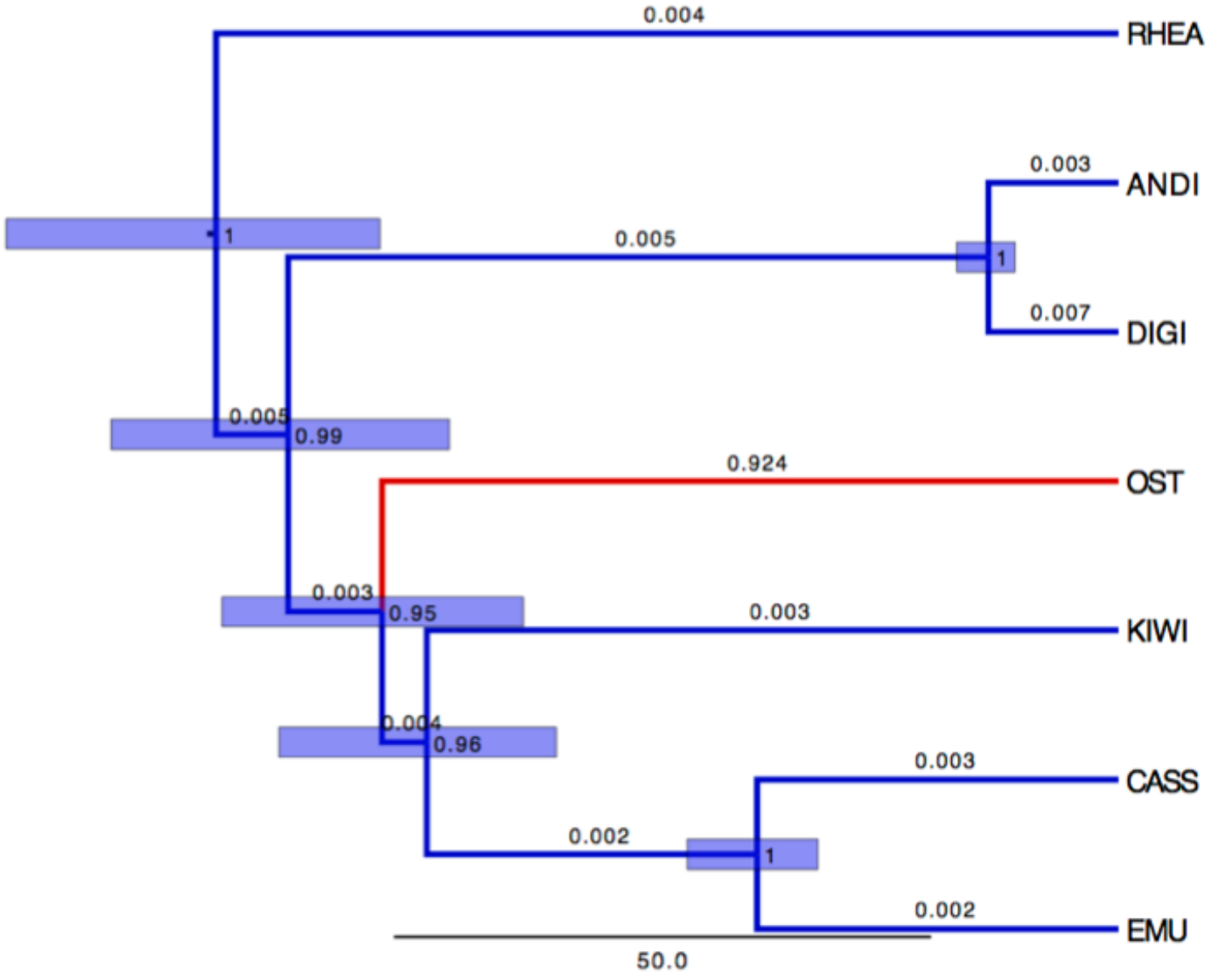
Rodent data set (Poisson prior on # rate changes)



Rodents data set (Uniform prior on # rate changes)



Ratite relaxed clock on full mitochondrial sequences



Conclusions

- Relaxed molecular clocks have many benefits over unconstrained models for phylogenetic inference
  - They appear to estimate the phylogenetic tree more accurately on real data sets
  - They automatically provide estimates of a root position, without the need for an outgroup
  - They automatically provide estimates of relative divergence dates, or absolute divergence dates when calibration information is available
- Calibration is hard and interesting
  - Specifying natural means of calibrating phylogenies is subtle
  - Recent methods for including fossil evidence include new tree priors, and opportunities for total evidence dating.
- The geometry of (time) is understudied and its study could lead to new methods for doing phylogenetic inference and posterior post-processing and summary.