Bayesian relaxed phylogenetics with BEAST

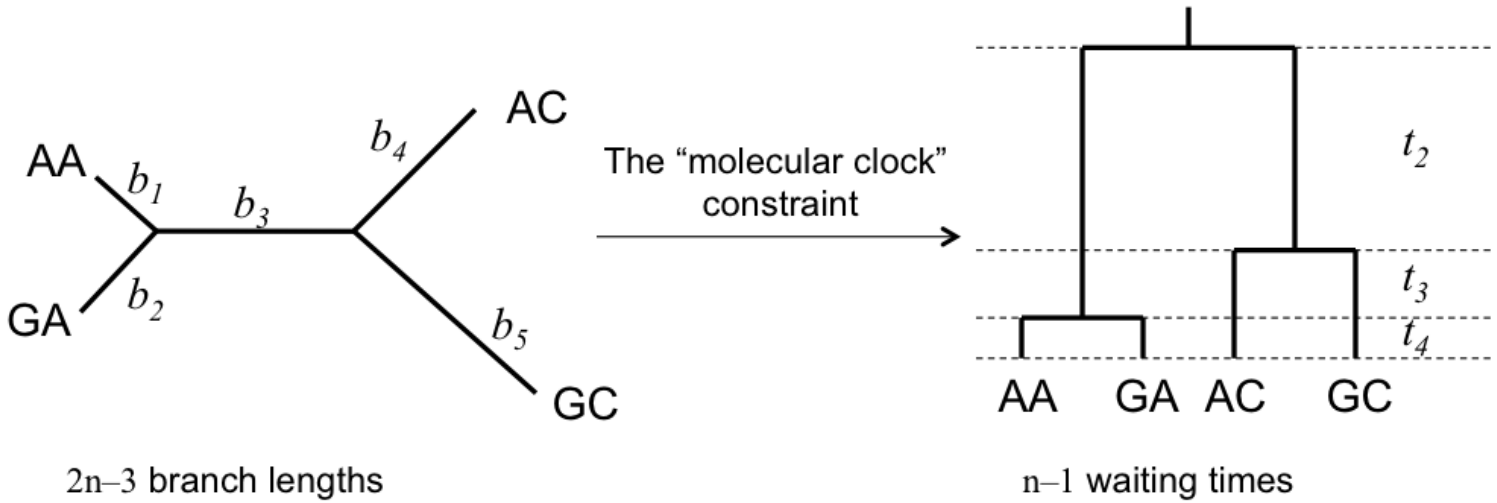
Alexei Drummond
Center for Computational Evolution
University of Auckland

Taming the BEAST, Engelberg, Switzerland, 2016

28th June 2016

Clocks and calibrations

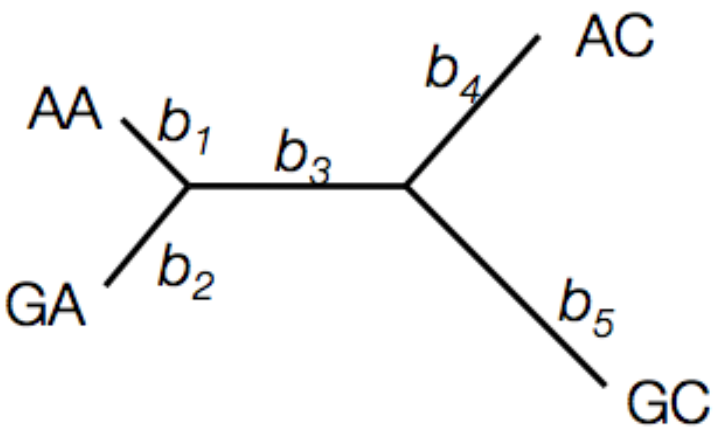
The molecular clock constraint



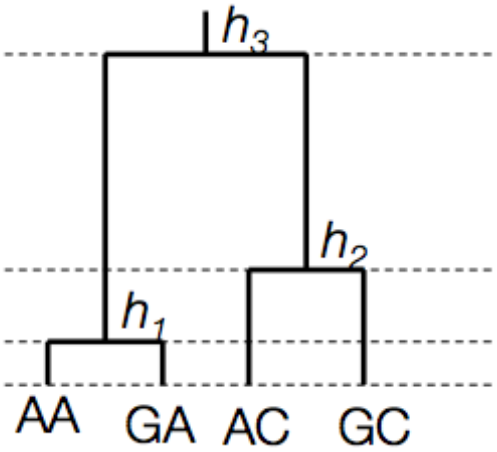
$$h(g, Q|D) \propto Pr\{D|\vec{\mu}g, Q\}f_G(g)f_Q(Q)$$

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:

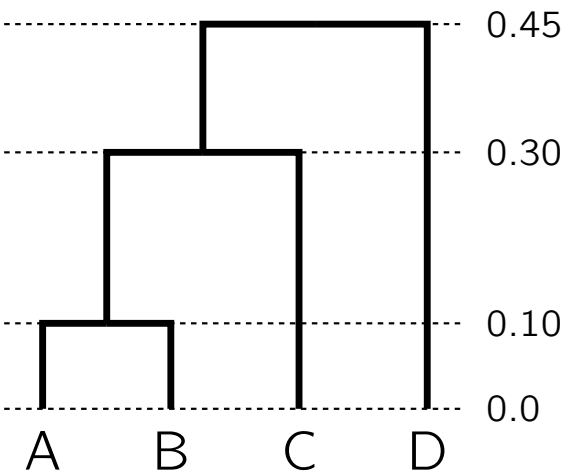
$$p(g, \mu, \theta|D) \propto Pr\{D|\mu \times g\}p(g|\theta)p(\theta)p(\mu)$$

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

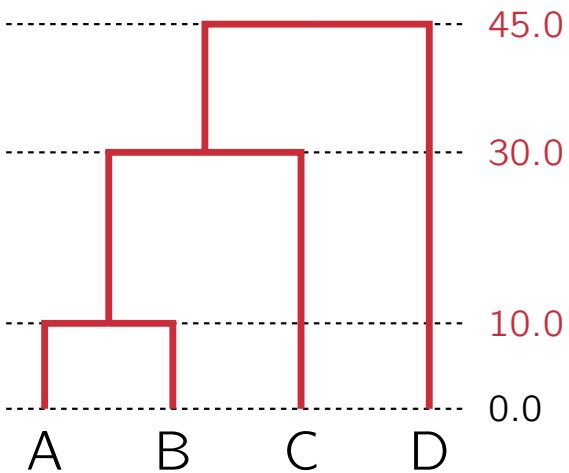
Genetic distance = rate × time

Strict molecular clock

$$T = \mu \times g$$



$= \quad 0.01 \quad \times$

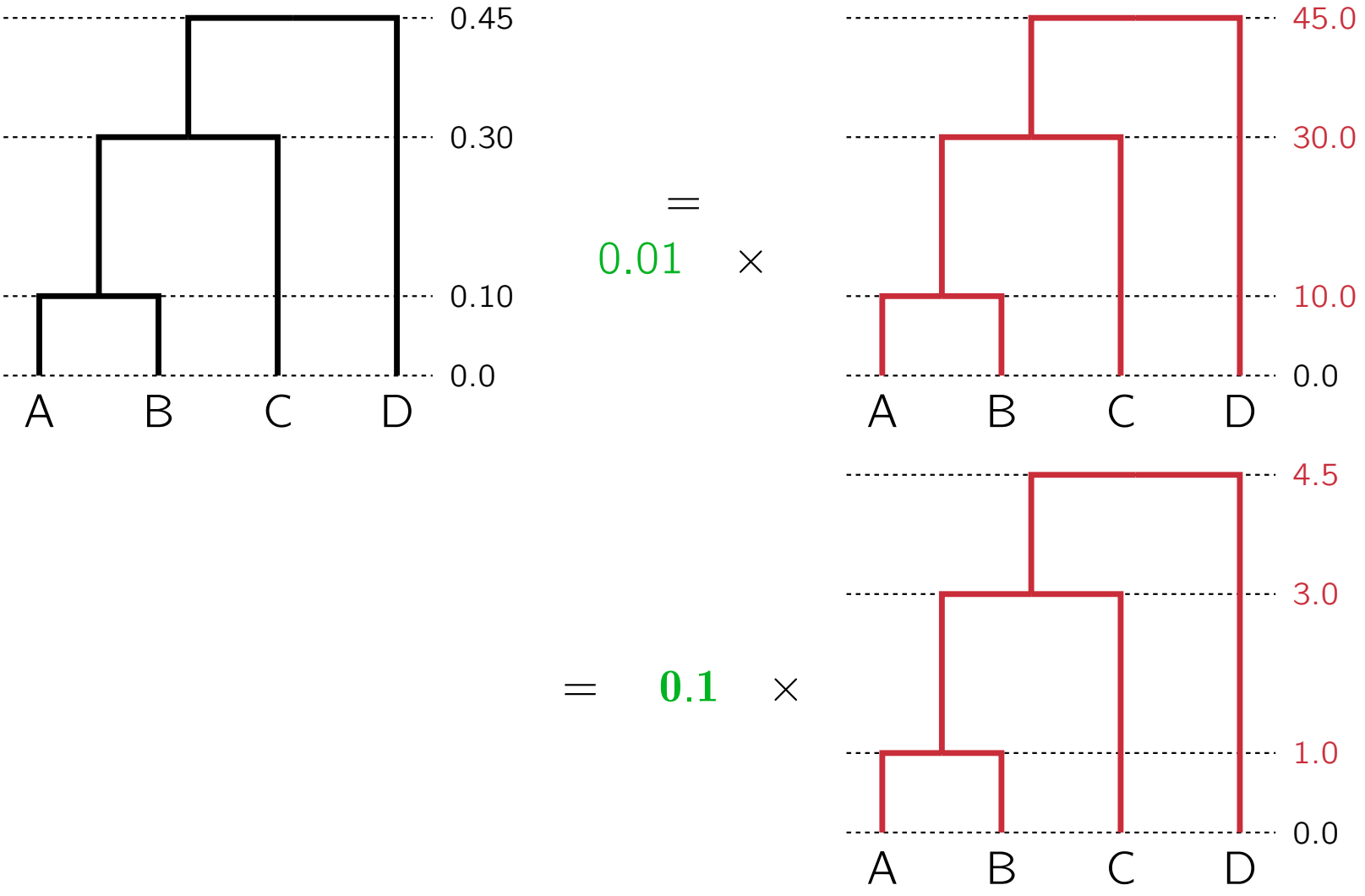


"substitution tree"

evolutionary rate
substitutions / site / unit time

time tree

Non-identifiability of rate and times



"substitution tree"

evolutionary rate
substitutions / site / unit time

time tree

A simple calibration is not simple

Consider the simplest type of calibration to admit uncertainty: the placement of an upper and a lower limit on the age of a single calibrated divergence (h_C) in the tree:

$$f(h_C) = \begin{cases} 1/(u-l) & l \leqslant h_C \leqslant u \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

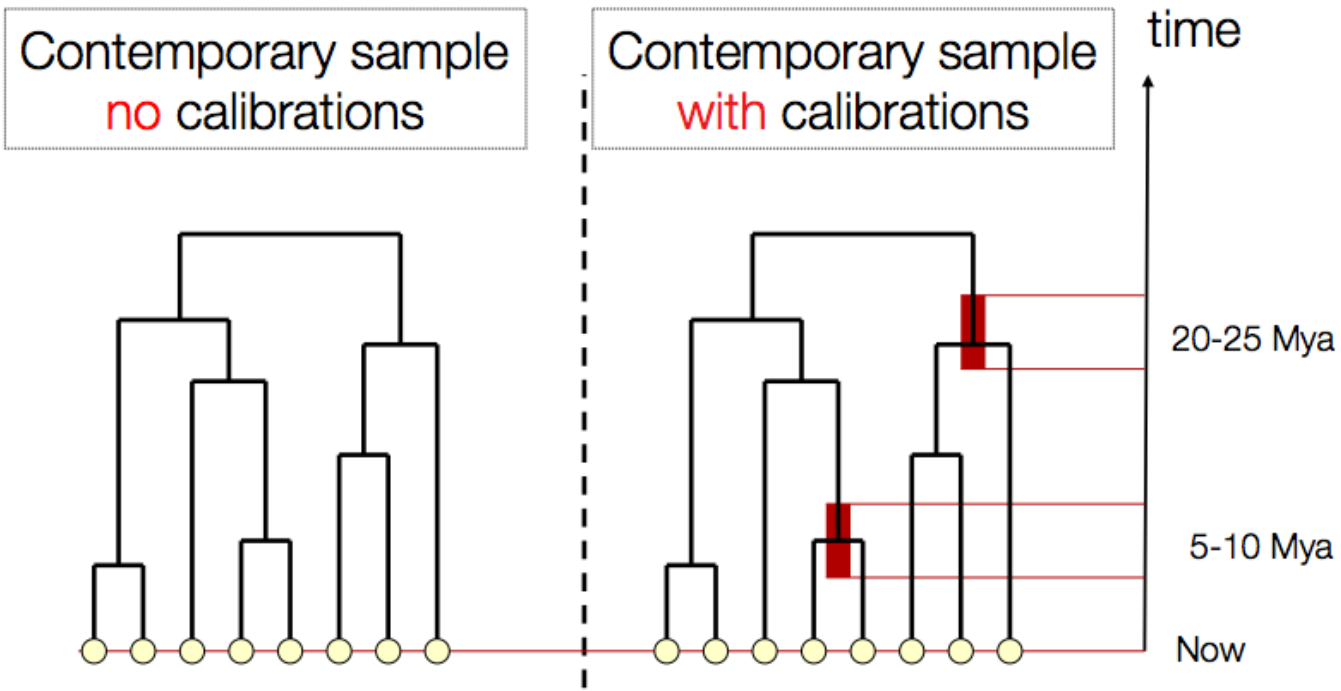
This calibration already has two quite distinct interpretations. One interpretation is that the resulting marginal prior distribution on the calibrated divergence should obey the tree process prior (f_G, e.g. Yule or Birth-death) but be **constrained** to be within the upper and lower bounds:

$$\rho_G(g|\theta) \propto f_G(g|\theta) f(h_C), \quad (2)$$

Alternatively, the marginal prior of h_C is uniform and conditioned on:

$$\rho_G(g|\theta) \propto f_G(g_{-h_C}|\theta, h_C) f(h_C), \quad (3)$$

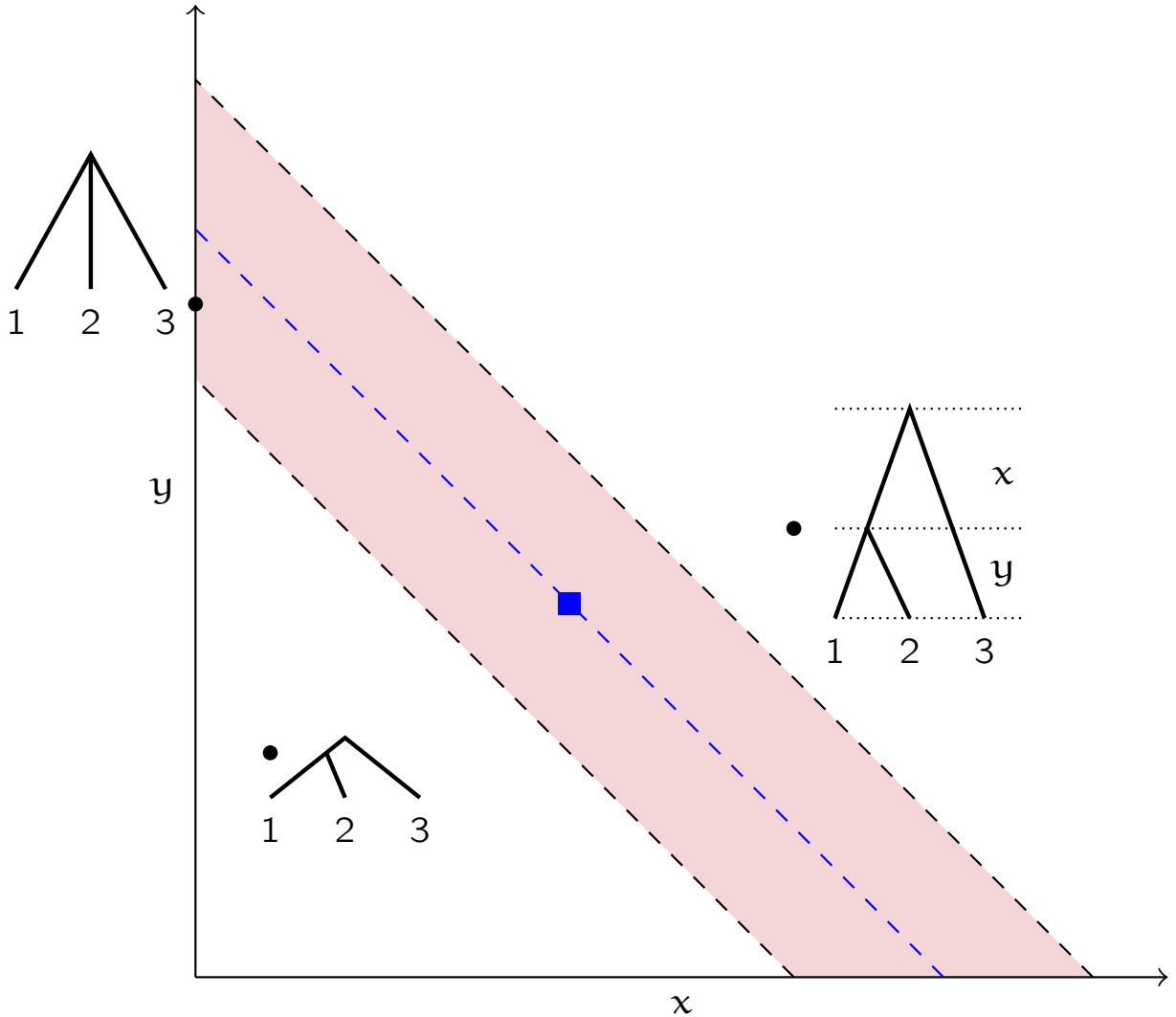
Absolute time via calibrations



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

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

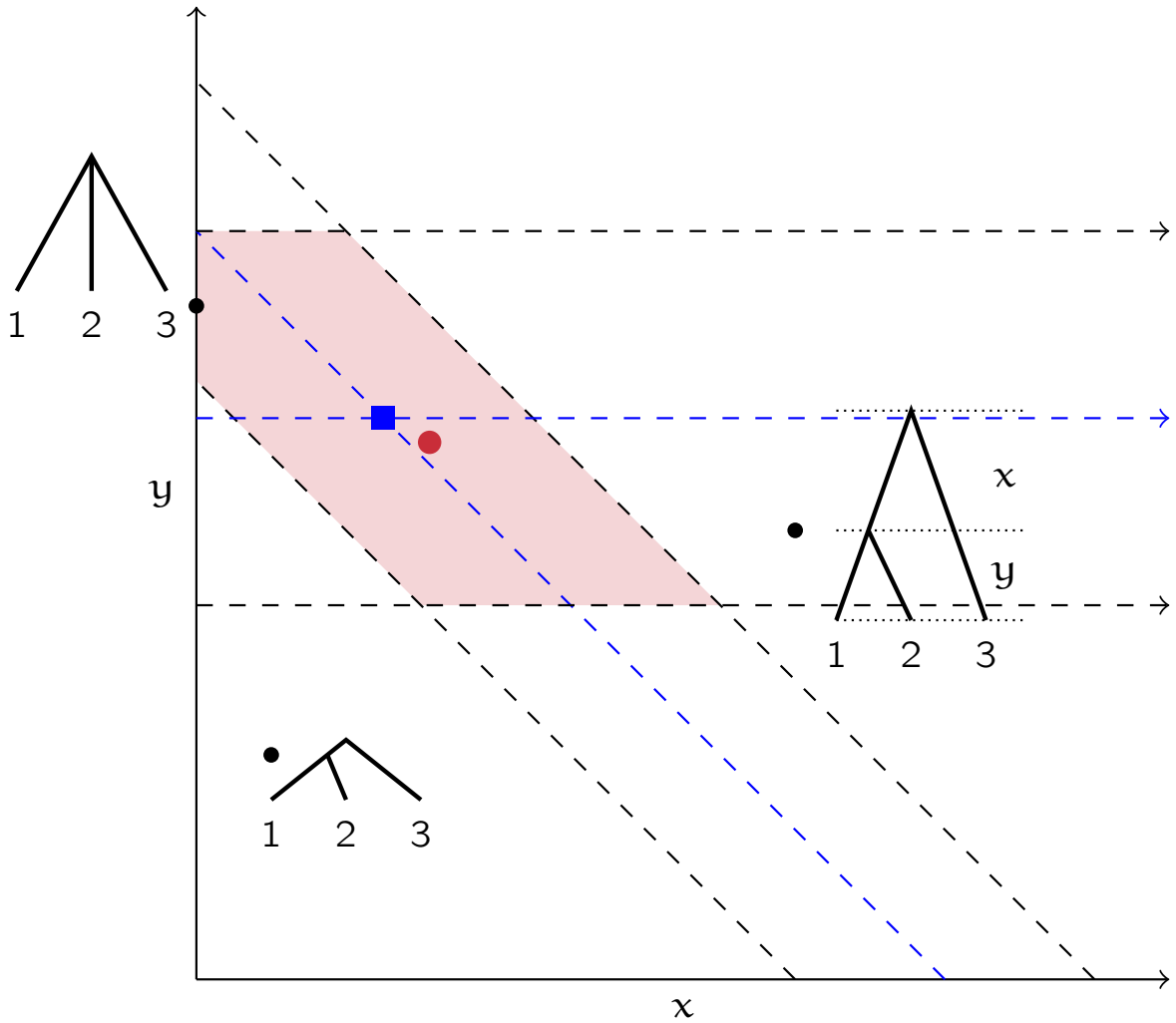
Calibrating tree space



Single calibration on the root height: $8 < x + y < 12$

Calibrating tree space

Two calibrations is even less simple!



First calibration: $8 < x + y < 12$

Second calibration: $5 < y < 10$

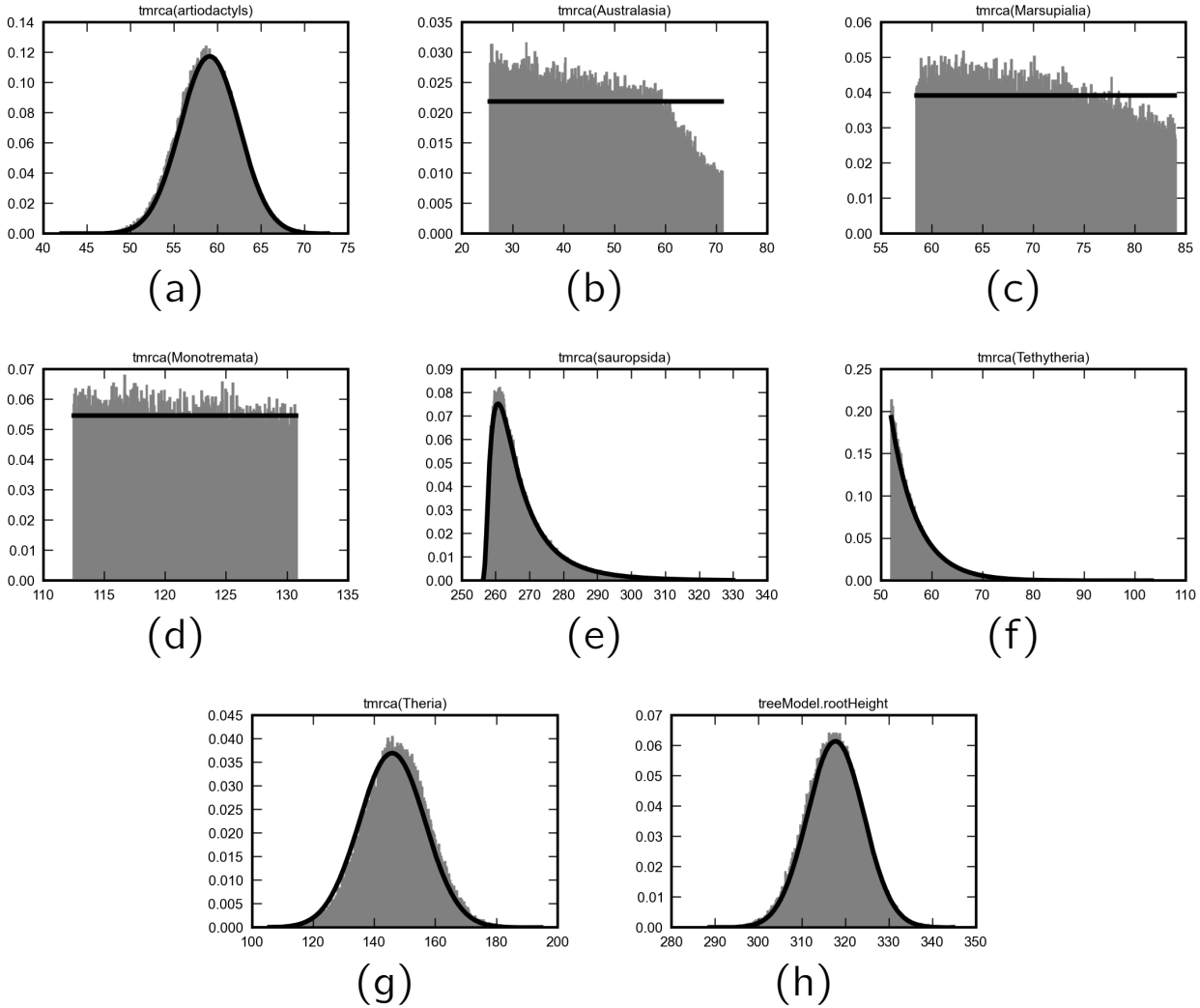


Figure adapted from A simple construction of calibrated tree prior: $\rho_G(g) \propto f_G(g) \times \prod_{i=1}^{k} f_i(s_i)$. Where $f_i()$ is the univariate "calibration density" for the divergence time of the i'th calibrated node in the tree. Monophyly is enforced for each calibrated node.

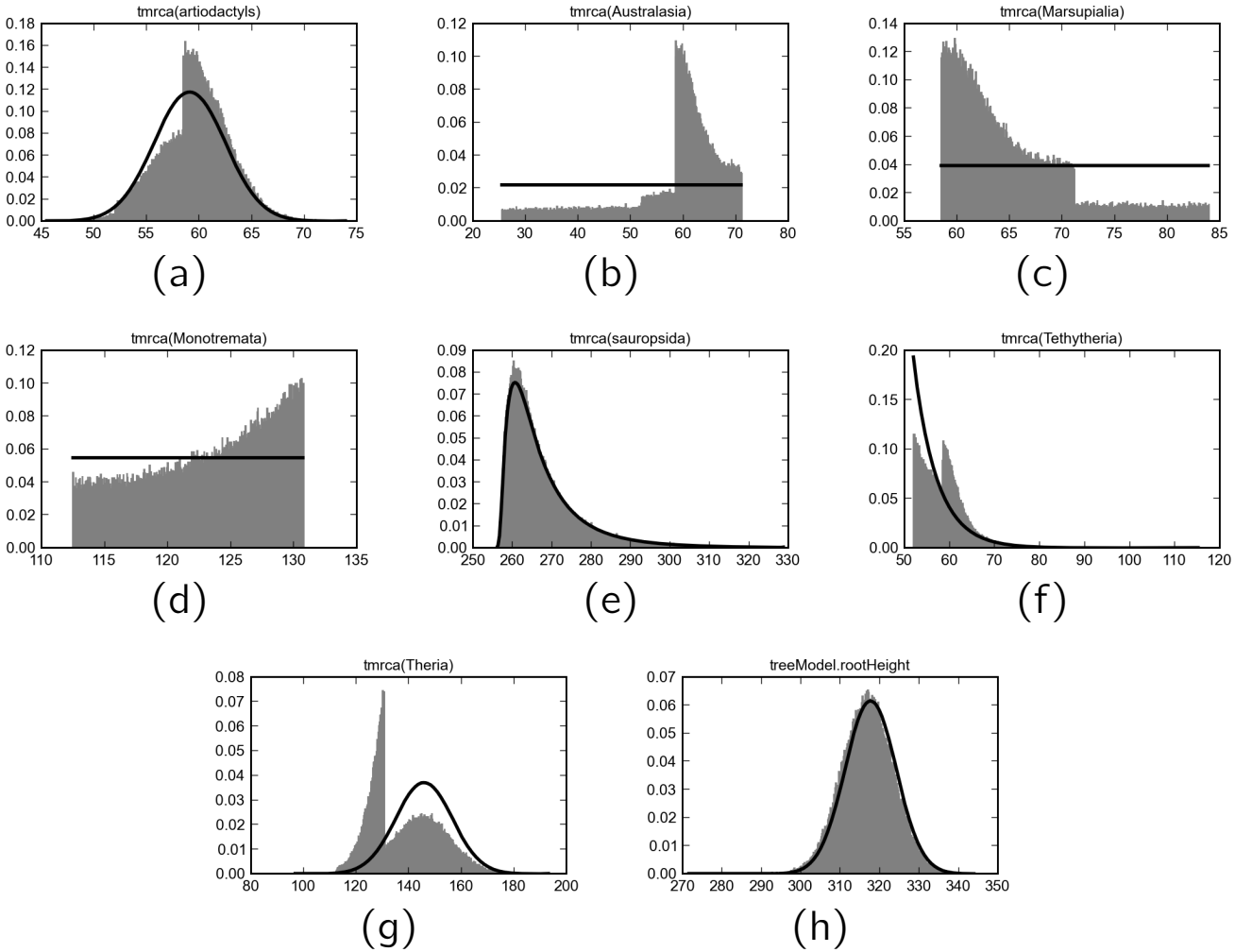
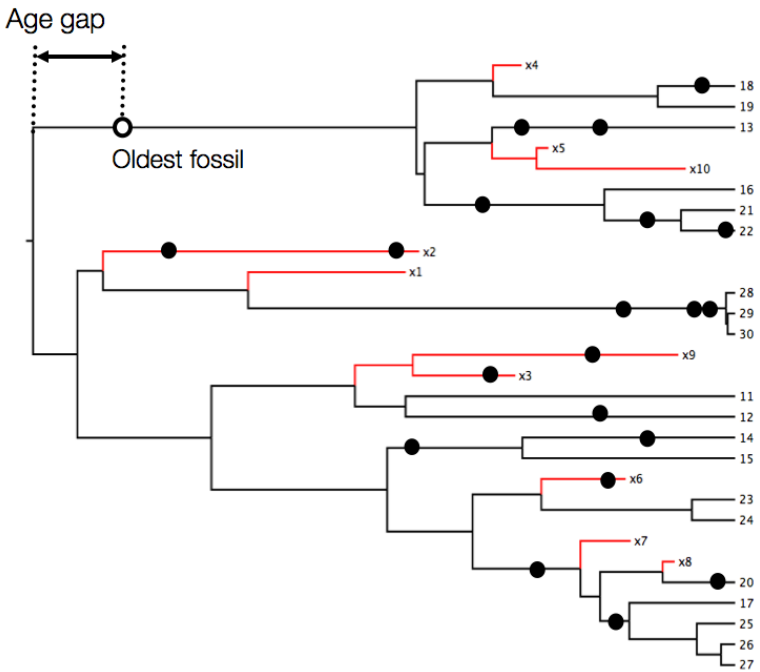


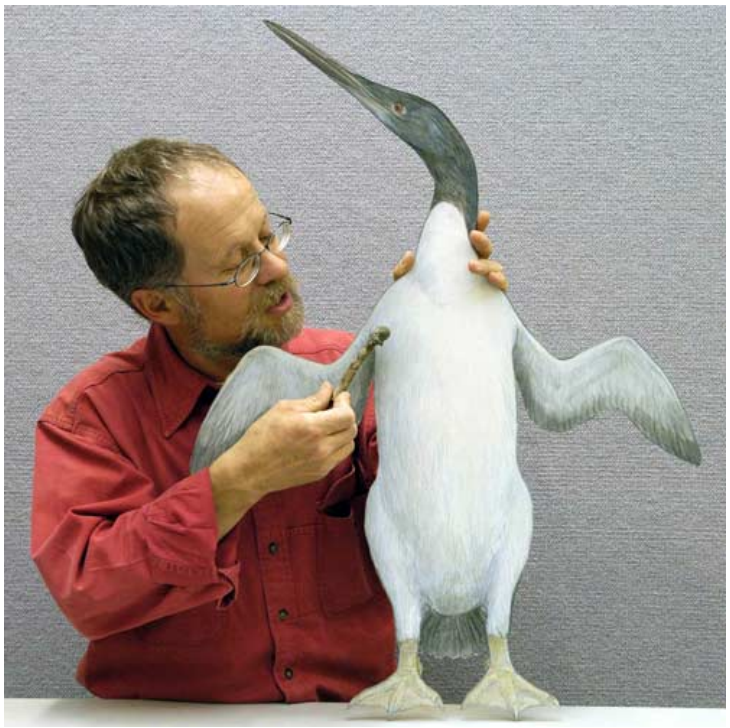
Figure adapted from The marginal prior distributions that result from BEAST (gray) versus calibration densities (black) specified for the calibrated nodes. The marginal prior distributions were obtained from a MCMC run using the prior only.

How do I pick the calibration density?

Modeling the Fossil Age Gap



What is the probability distribution of the age gap?

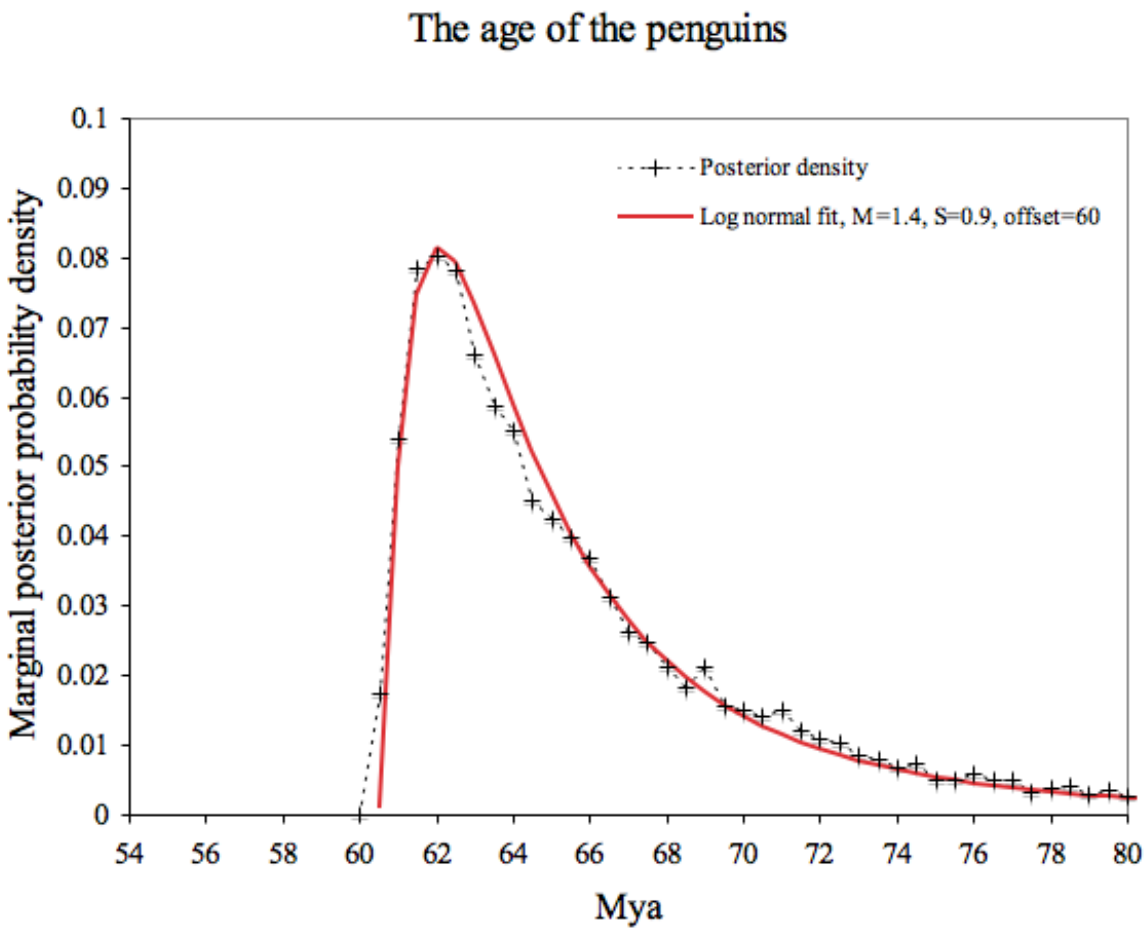


60-61.5 Myr penguin

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

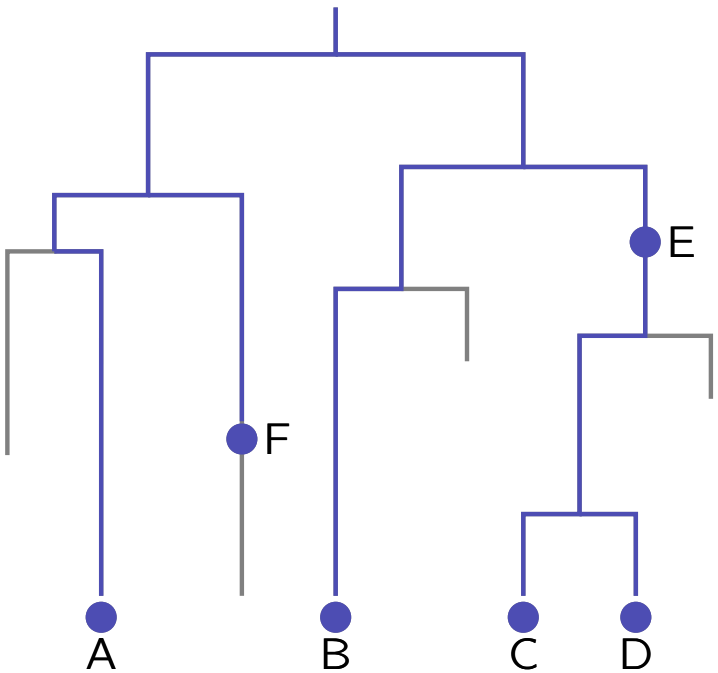


Total evidence dating

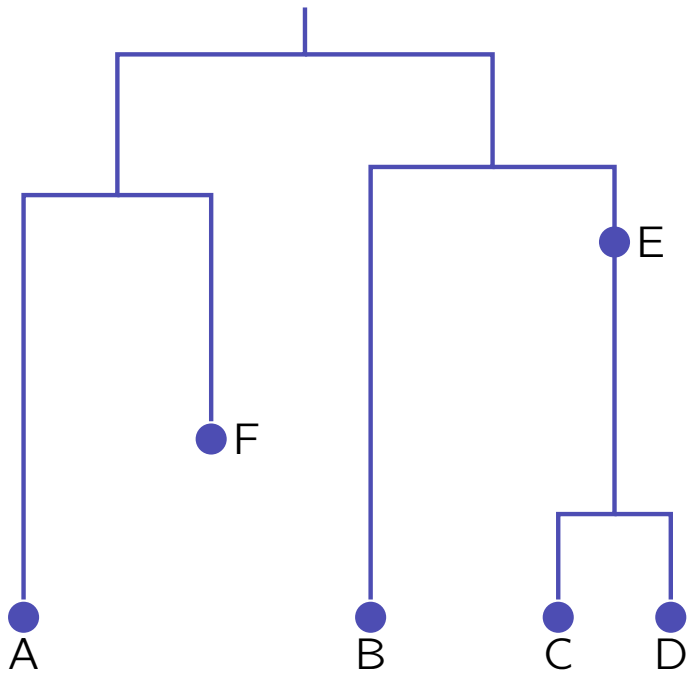
Total evidence phylogenetic dating

- We have
  - **molecular data** from extant species,
  - **morphological data** from both extant and extinct species and
  - **A fossil age** (or geological time interval) for each extinct species.
- We want to utilise all this data to learn about **evolutionary history** and **macroevolutionary processes**.
- **Bayesian statistical inference** is our preferred method of learning about the world.

Sampled-ancestor time-trees



Full tree



Sampled tree

Analysis of penguin morphological data

Penguin dataset (Ksepka et al 2012) consisting of morphological data of:

- all extant penguins (19 species)
- 36 fossil species assigned to stratigraphic intervals

Models of morphological evolution and species diversification:

- Lewis MK vs MKv (Lewis 2001)
- Partitions vs single alignment
- Rate variation across sites and across partitions
- Tree prior: FBD and Skyline FBD
- Different tree model parameterisations

Described in Gavryushkina et al (2015) on the ArXiv.

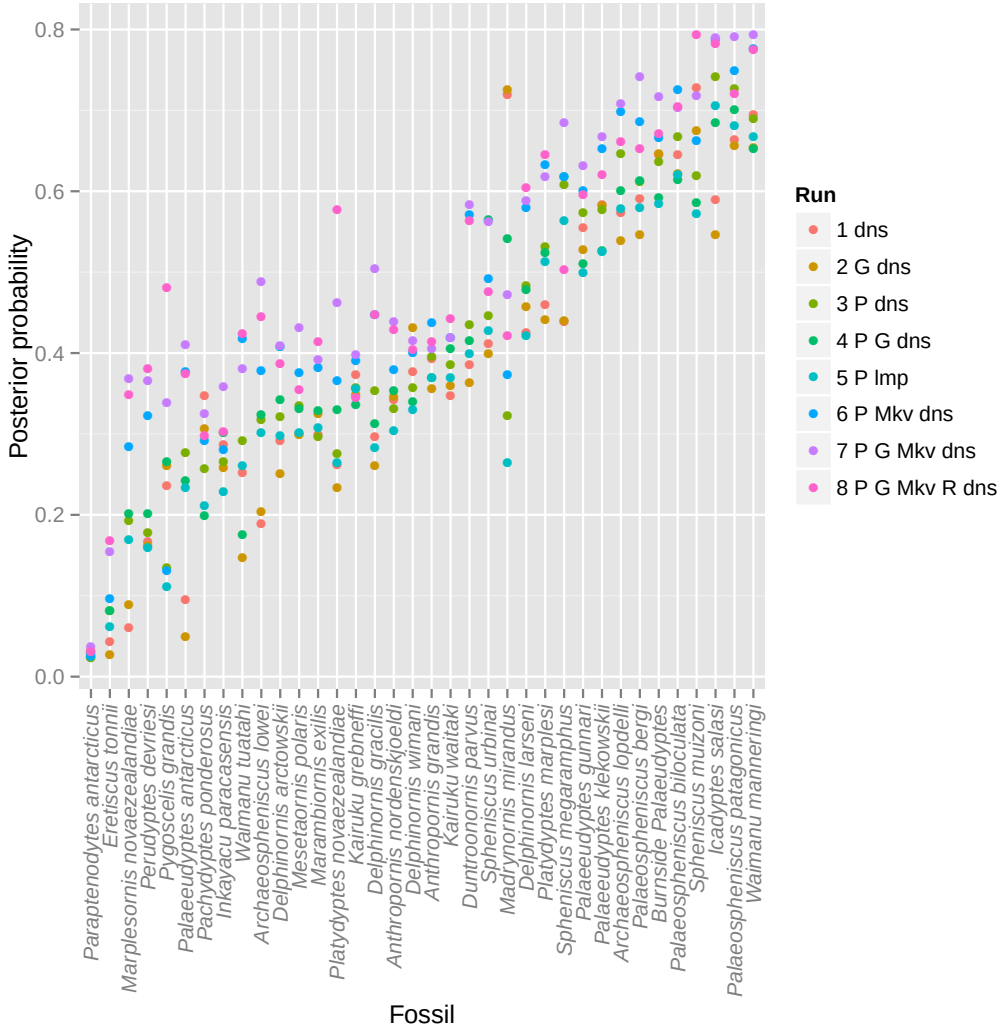
Parameterisations

lmp:

birth rate	λ
death rate	μ
sampling rate	ψ
other parameters	t_{or}, ρ

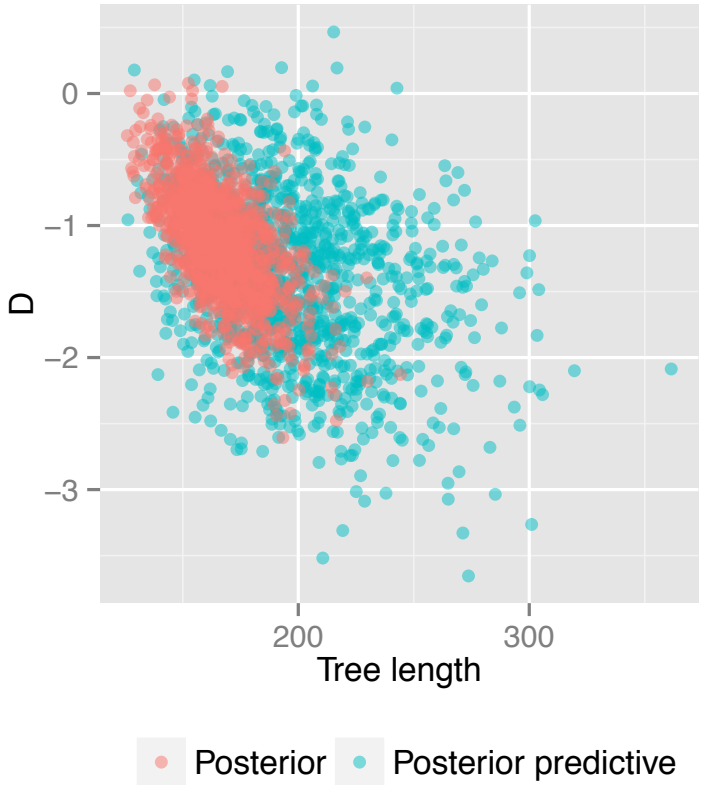
dvs:

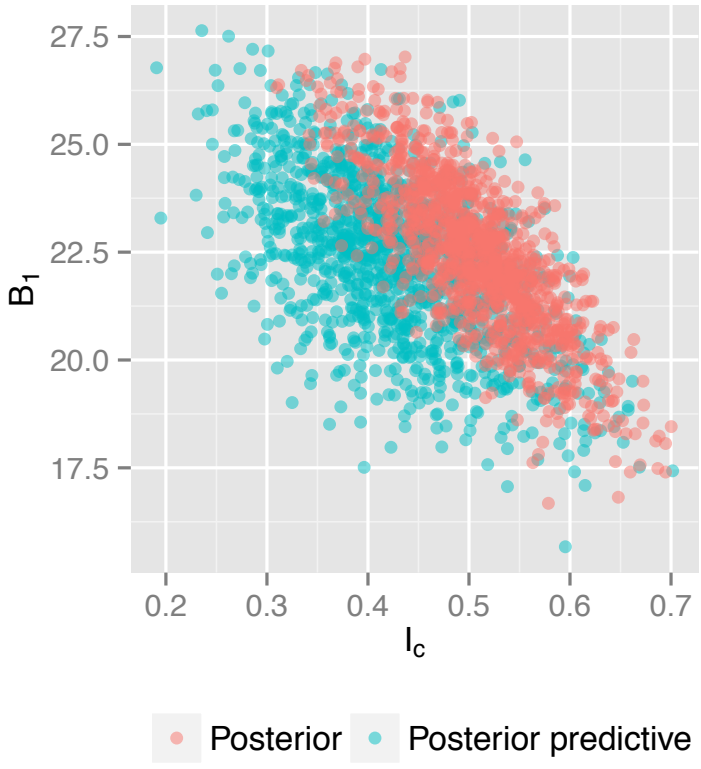
net diversification rate	$d = \lambda - \mu$
turnover rate	$\nu = \frac{\mu}{\lambda}$
sampling proportion	$s = \frac{\psi}{\mu+\psi}$
other parameters	t_{or}, ρ



Probabilities of each fossil to be a sampled ancestor across models.

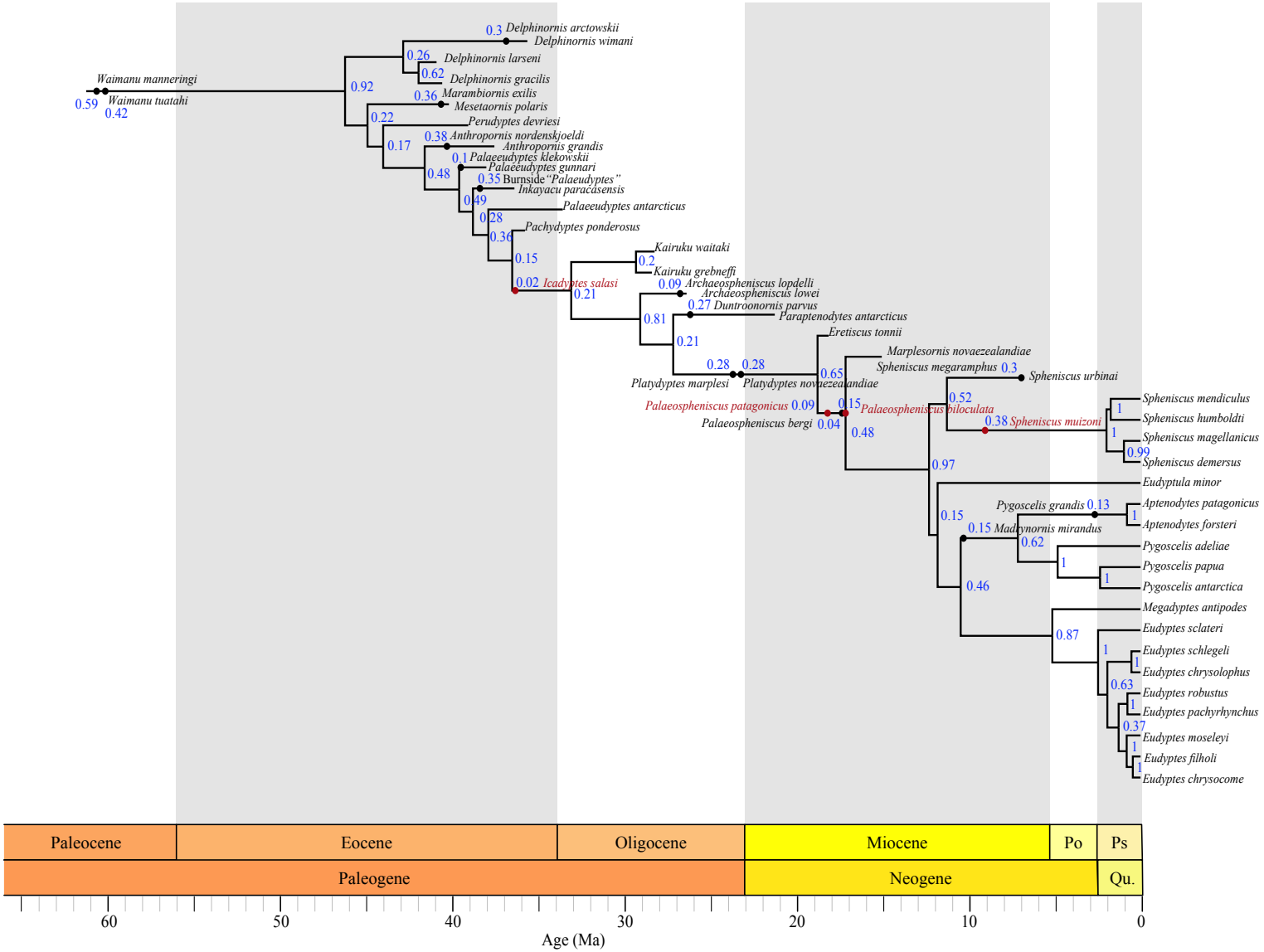
P stands for the partitioned model, *G* for gamma variation across sites, *Mkv* for conditioning on variable characters, *R* for relaxed clock model, *dns* for $\mathfrak{d}$, ν, and s tree prior parameterisation, *lmp* for λ, μ, and ψ tree prior parameterisation.



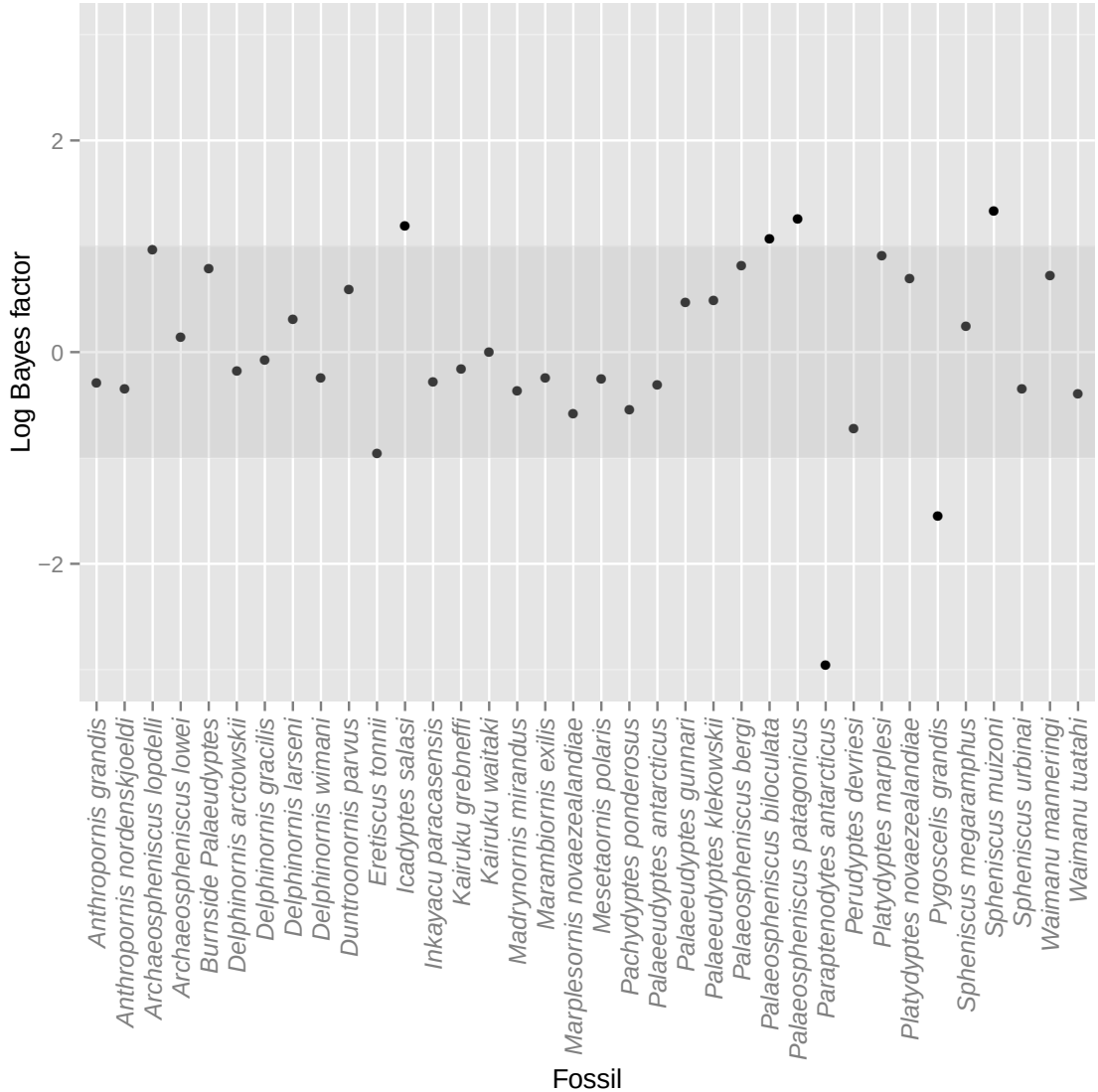


The posterior and posterior predictive distributions for (left) tree length, T, and genealogical D_F statistics and (right) B_1 tree imbalance statistic and Colless's tree imbalance index, I_c, for model 8.

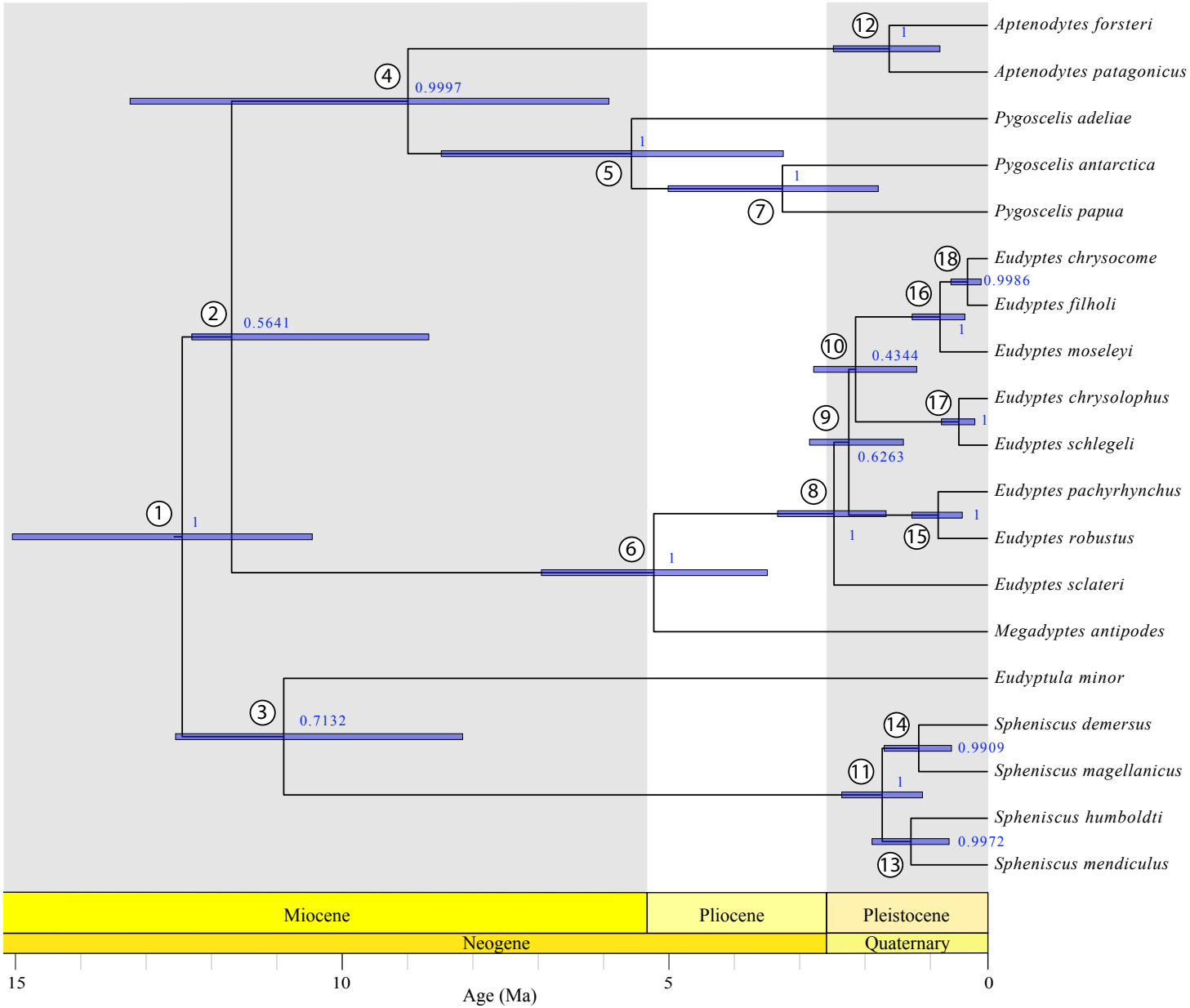
The posterior trees are on the unbalanced end of the predictive distribution.



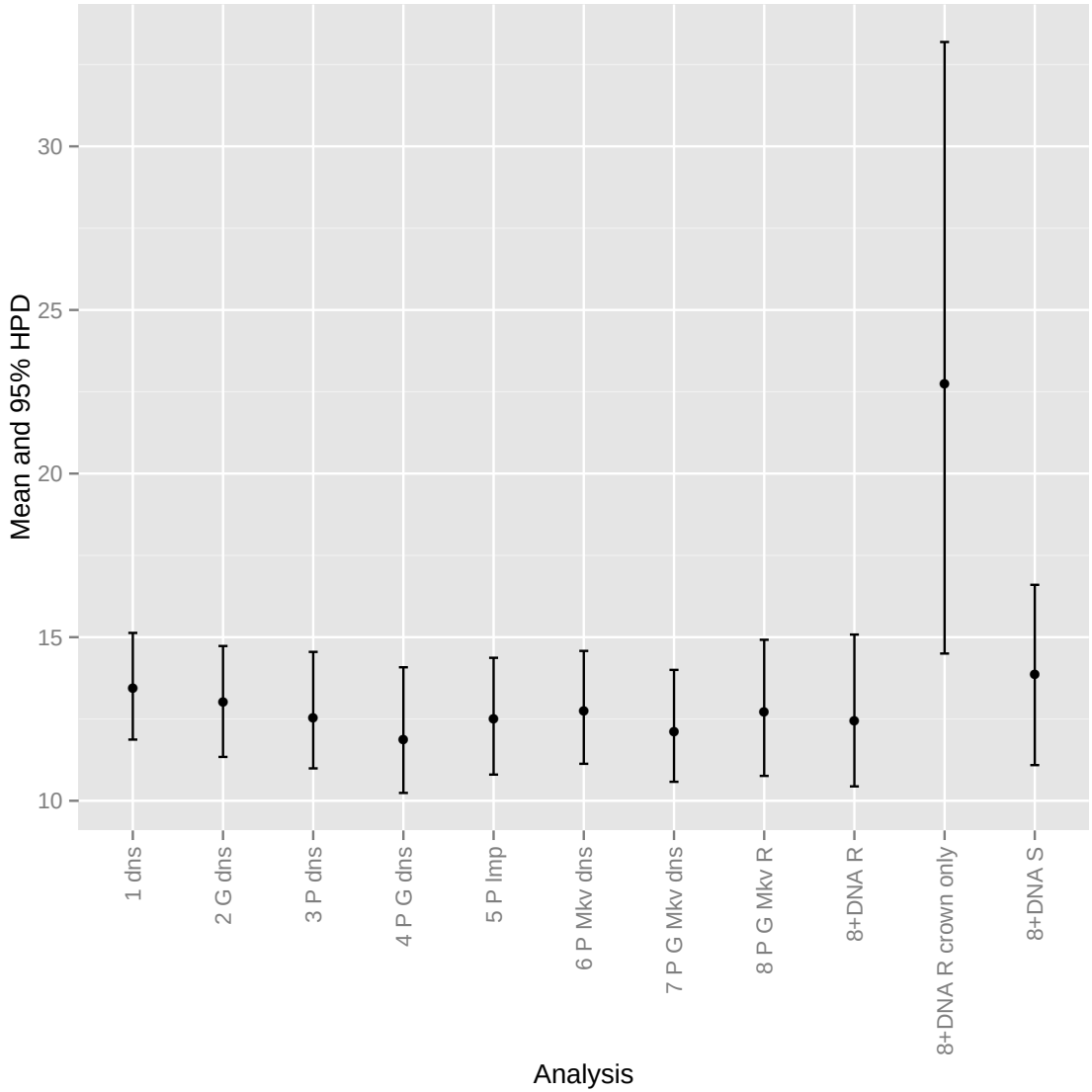
A maximum sampled ancestor clade credibility tree for the total-evidence analysis. The numbers in blue show the posterior supports of clades. The filled red and black circles represent sampled ancestors. Fossils with positive evidence of being sampled ancestors are shown in red.



The evidence for fossil sampled ancestors. The samples above the shaded area have positive evidence to be sampled ancestors and below the shaded area have positive evidence to be terminal nodes.

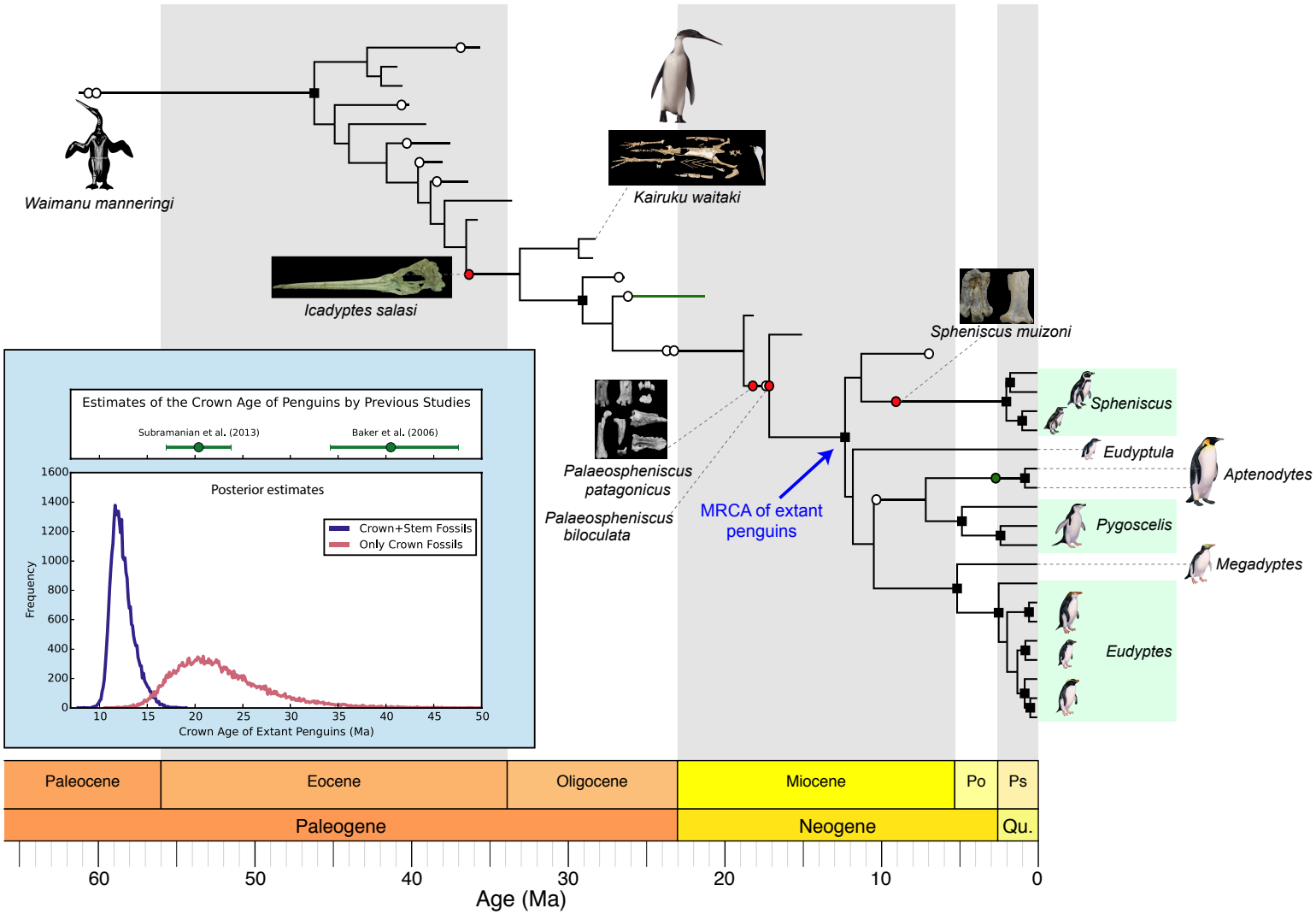


The posterior summary tree of extant penguins with 95% HPD intervals and posterior supports of clades after removing fossils.

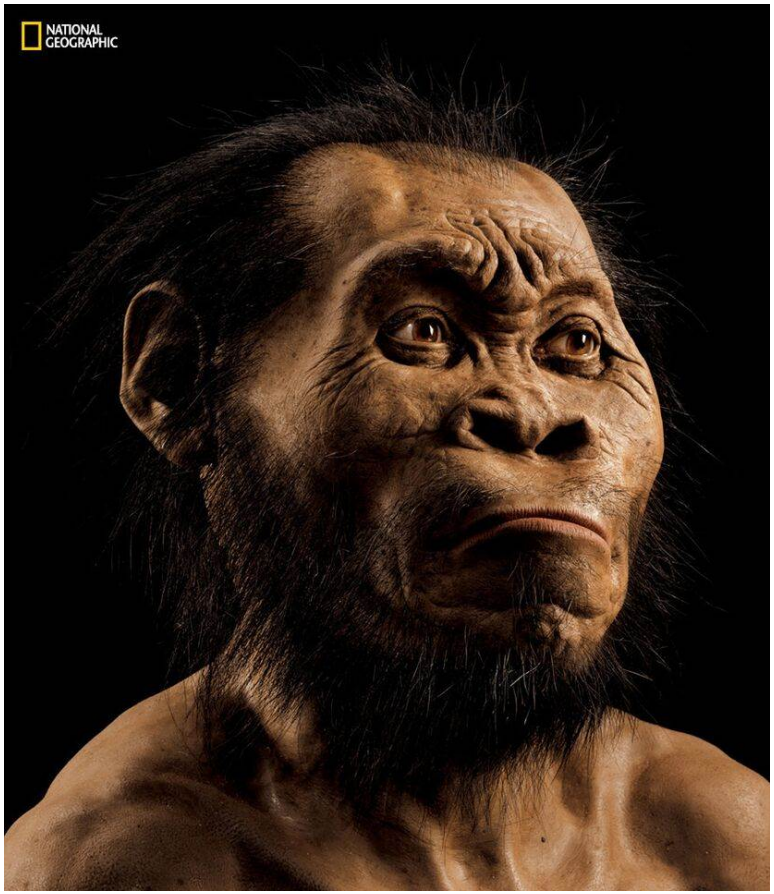


The age of root of extant penguins across models including total-evidence analysis under relaxed (8+DNA R) and strict (8+DND S) molecular clocks and total-evidence analysis with crown fossils only.

Dating the evolutionary history of penguins



Fossils without geological dates



SCIENCE

Why Don't We Know the Age of the New Ancient Human?

Because dating fossils is hard, and it's not always possible.

ED YONG | SEP 14, 2015

Last Thursday, the world said hello to *Homo naledi*, a new species of ancient human discovered in South Africa's Rising Star cave. As I reported at the time, scientists extracted 1,550 fossil fragments from the cave, which were then assembled into at least 15 individual skeletons—one of the richest hauls of hominid fossils ever uncovered.

But one significant problem clouded the excitement over the discovery: The team doesn't know how old the fossils are. And without that age, it's hard to know how *Homo naledi* fits into the story of human evolution, or how to interpret its apparent habit of deliberately burying its own kind. Everyone from professional paleontologists to interested members of the public raised the same question: Why hadn't the team dated the fossils yet?

The simple answer is: Because dating fossils is really difficult. Scientific papers

Can we date fossils using Bayesian phylogenetics?

Given:

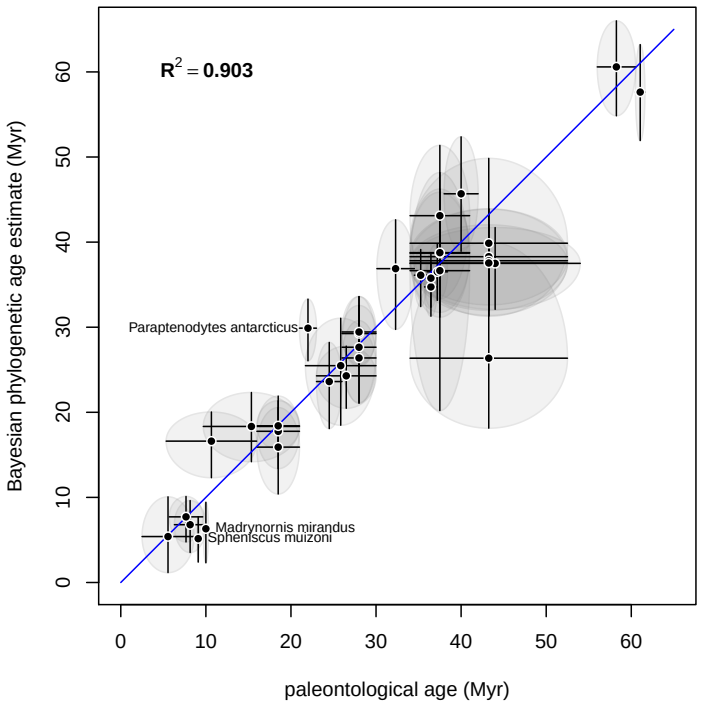
- A reference set of fossils of known age, and
- A morphological characterization for reference fossils, and
- The fossilized birth-death model of sampled ancestor trees
- A "clock-like" model of morphological evolution
- A morphological characterization of a related focal fossil of unknown age

Can the age of the focal fossil be estimated using Bayesian phylogenetic inference?

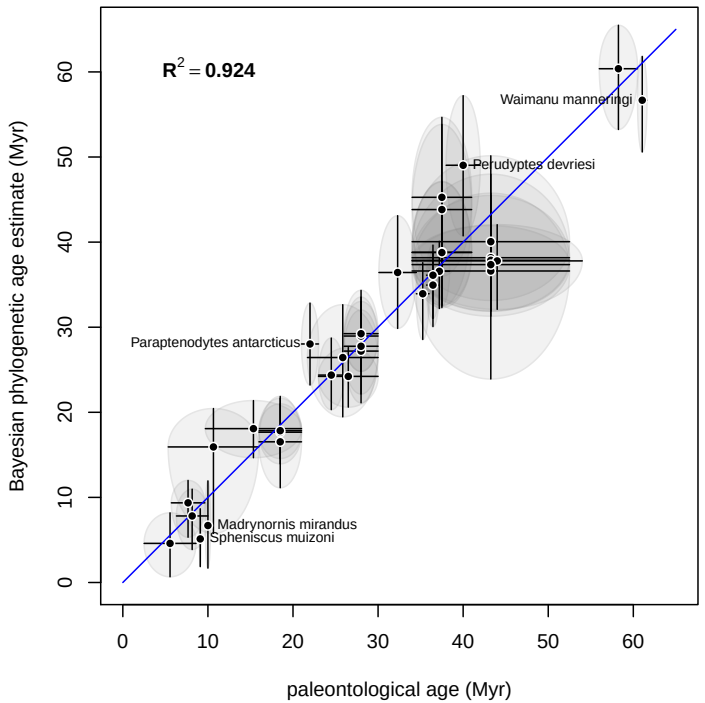
We tested this hypothesis by using the penguin data set. For each of the 36 fossils in turn, we discarded the age information of the focal fossil to mimic the scenario where the age was unknown, and sought to estimate its age using the reference 35 fossils as the reference set.

Estimating the phylogenetic age of a fossil

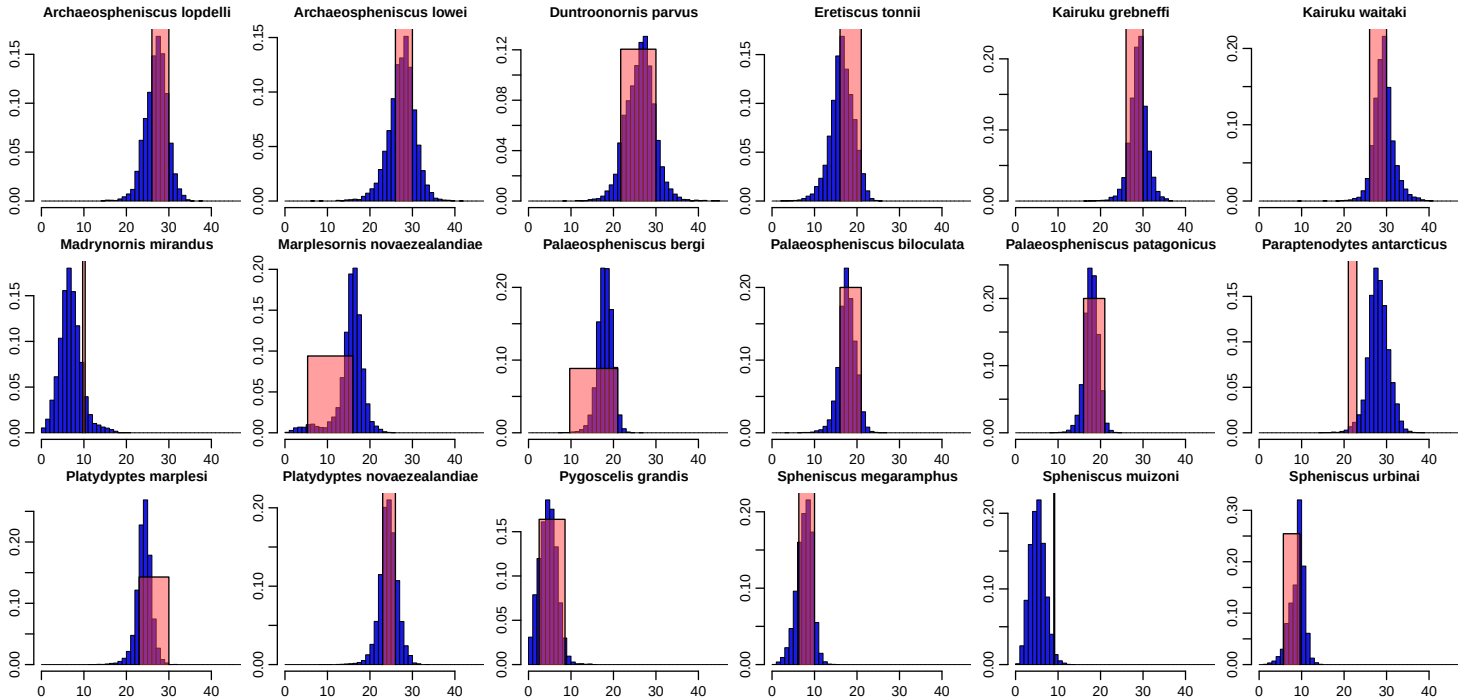
(a) Model 1



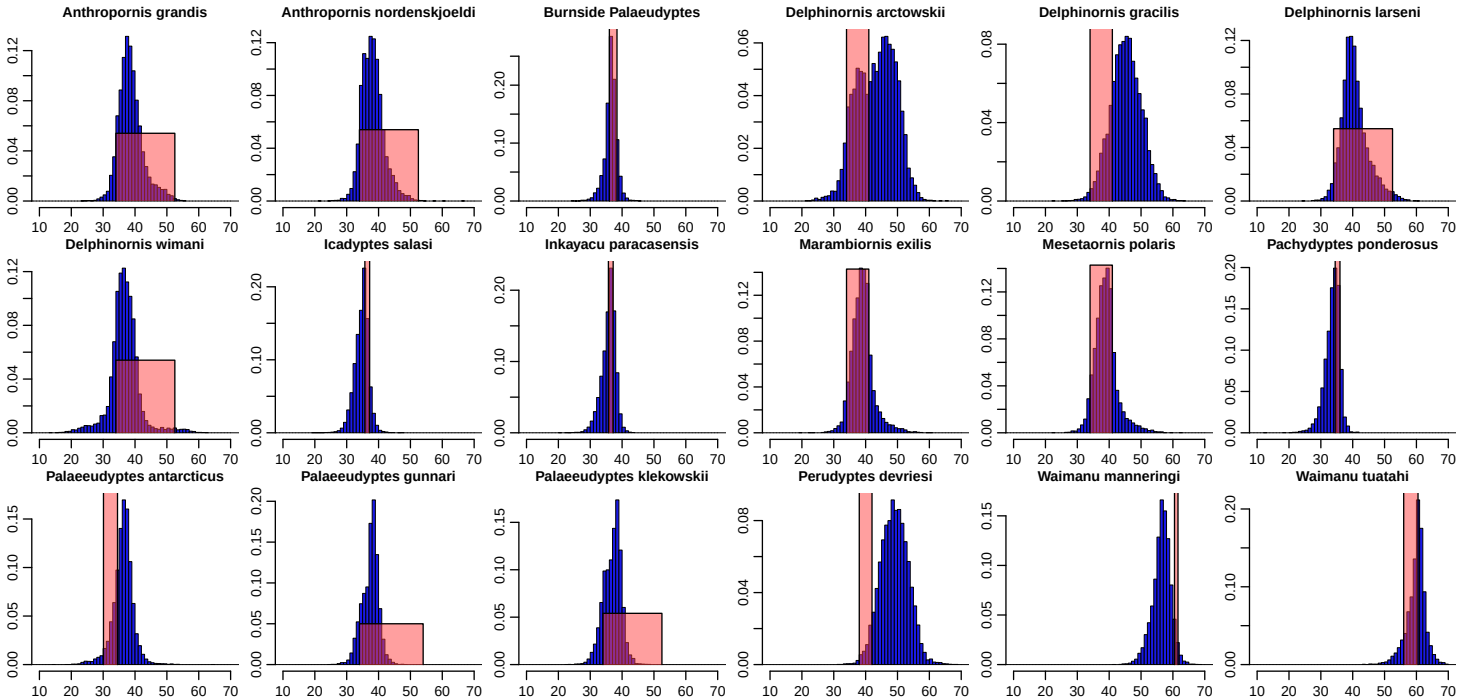
(b) Model 8



Bayesian phylogenetic age estimates for each of 36 penguin fossils plotted against their palaeontological age estimates, under two alternative evolutionary models. The blue line shows the $x = y$. If the vertical line doesn't cross $x = y$, then the midpoint of the geological range is not in the phylogenetic 95% HPD.

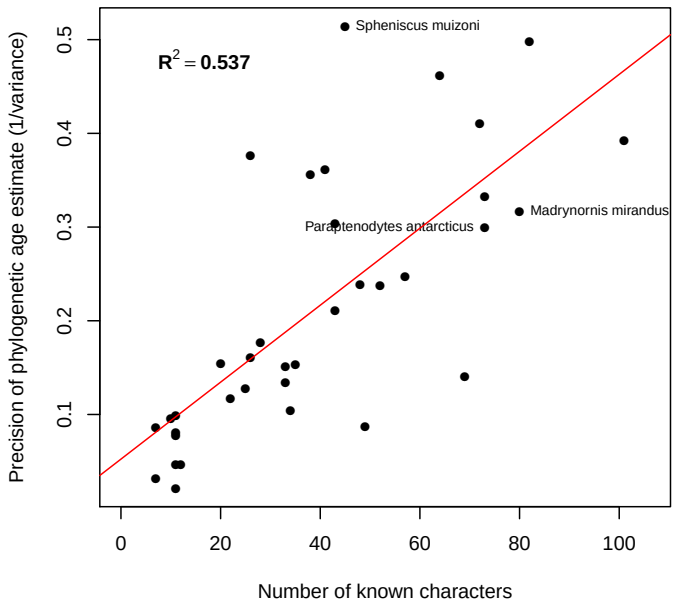


Marginal posterior density plots for the phylogenetic age estimate of each of the 18 penguin fossils younger than 30 Myr using M8. Red boxes are the superimposed age ranges derived from geological data.

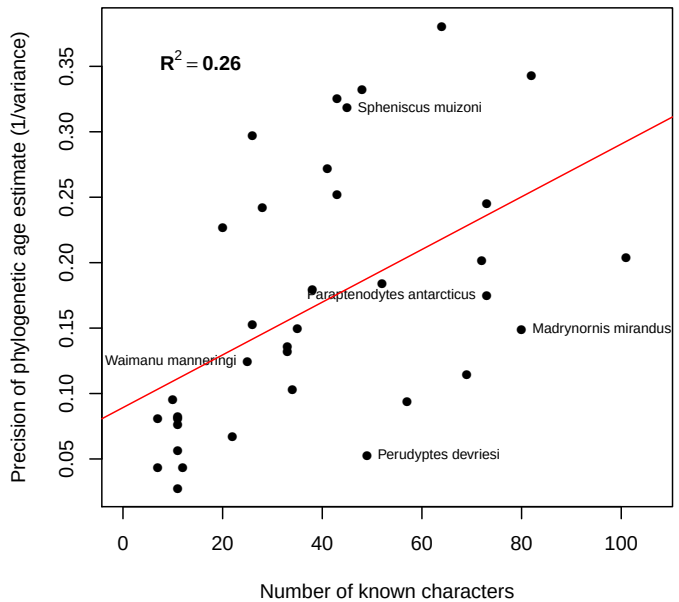


Marginal posterior density plots for the phylogenetic age estimate of each of the 18 penguin fossils older than 30 Myr using M8. Red boxes are the superimposed age ranges derived from geological data.

(a) Model 1



(b) Model 8



A plot of the number of non-ambiguous morphological sites for the taxon against the precision of the phylogenetic age for (a) M1 and (b) M8 (i.e. the precision is 1/variance in the marginal posterior distribution of the age).

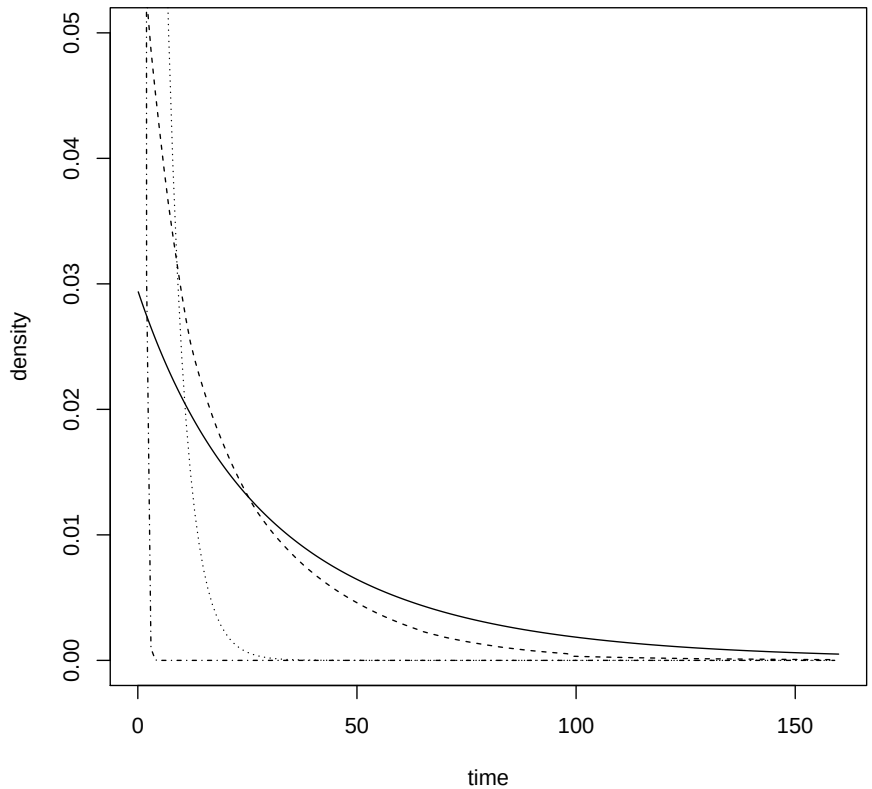
Computing the phylogenetic evidence for an age range

The Bayes factor (BF) computes the evidence for one hypothesis (H_1) over another (H_2) as the ratio of the marginal probability of the data under each of the two hypotheses and a model M,

$$BF = \frac{p(D|H_1, M)}{p(D|H_2, M)} = \frac{p(H_1|D, M)}{p(H_2|D, M)} \frac{p(H_2|M)}{p(H_1|M)}. \quad (4)$$

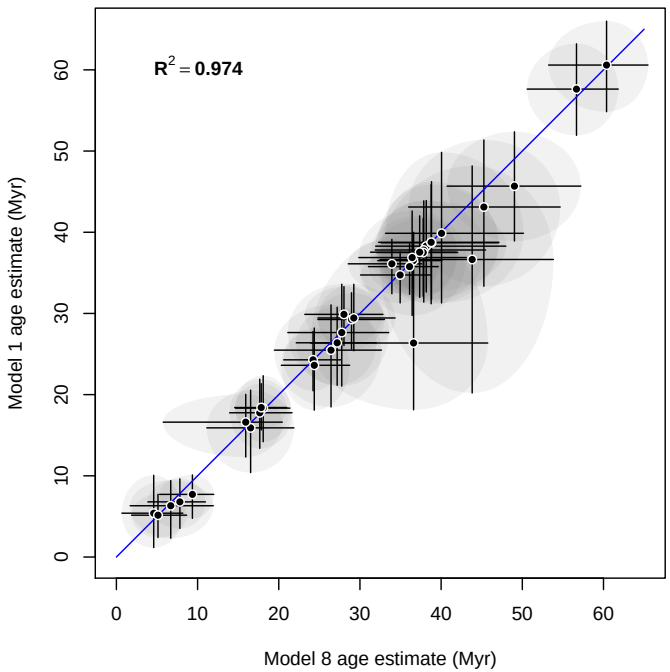
We are interested in computing the Bayes factor that quantifies the amount of phylogenetic evidence in support of the palaeontological age range for each fossil. In this case H_1 is the hypothesis that the true fossil age is within the given paleontological age range, and H_2 is the alternative hypothesis that the true fossil age is outside the palaeontological range.

So what do $p(H_1|M)$ and $p(H_2|M)$ look like?

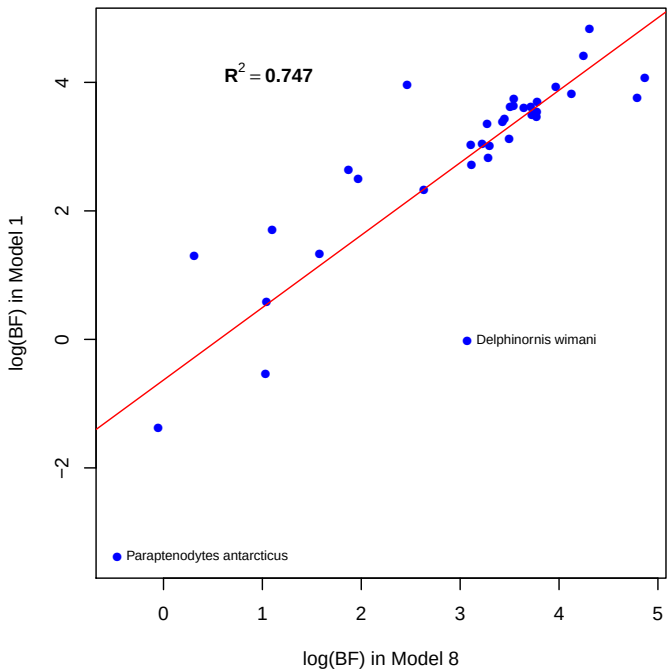


Prior density for sampling times under fossilized birth-death process. Dot-dashed line uses priors from Gavryushkina *et al* (2015). Solid line is new prior with implicit assumptions on T and s, dashed line results from only assuming implicit prior on T, dotted line results from only assuming implicit prior on s.

Comparison of age & BF estimates between models

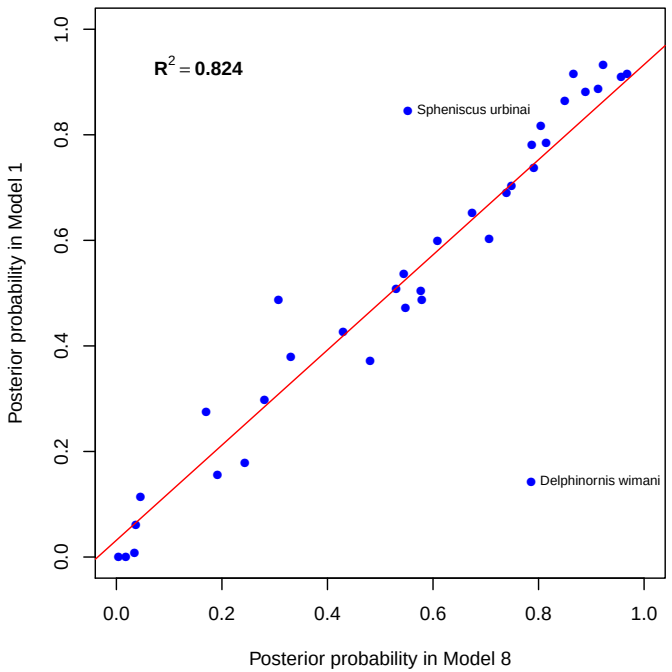


(a) Estimated phylogenetic age of M1 against M8.

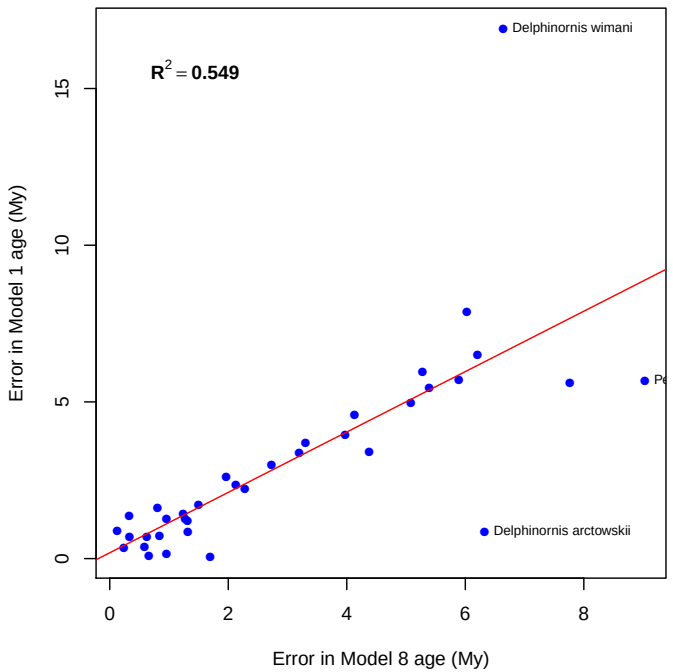


(b) Regression of Bayes factor (BF) for palaeontological range of M1 against M8.

Comparison of probability & error between models

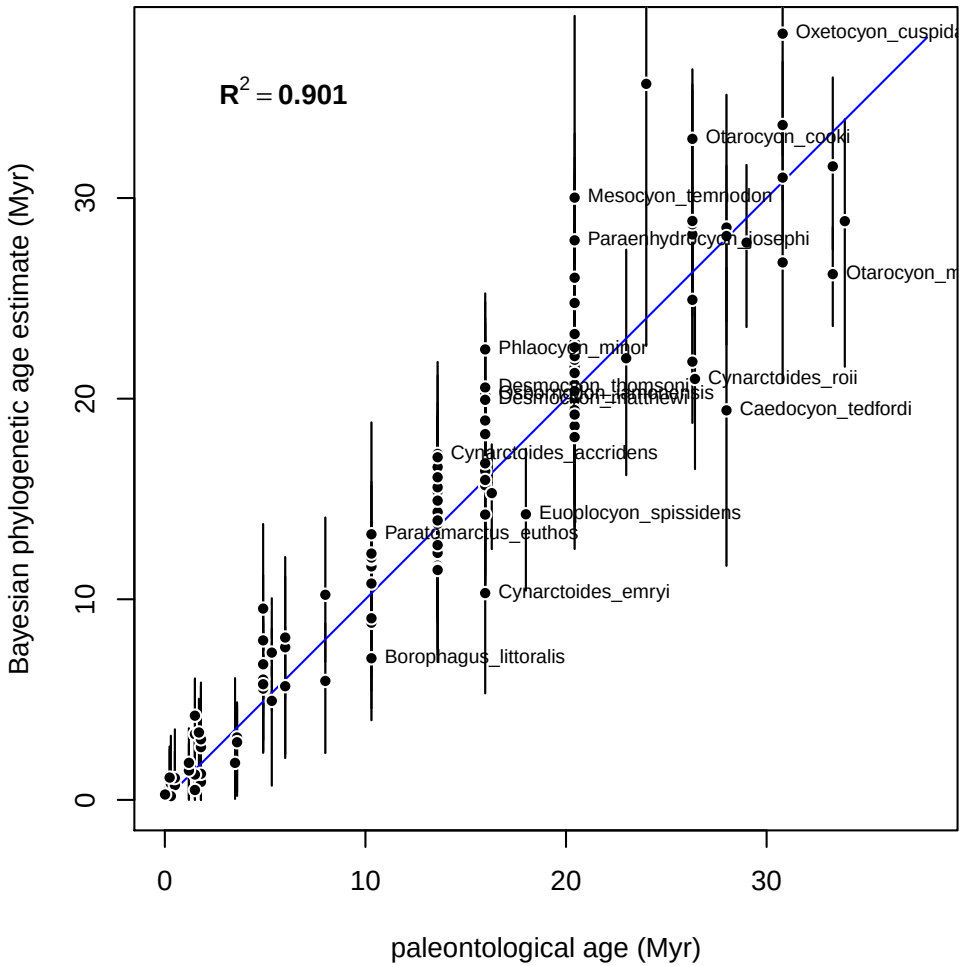


(c) Regression of posterior probability of palaeontological range of M1 against M8.



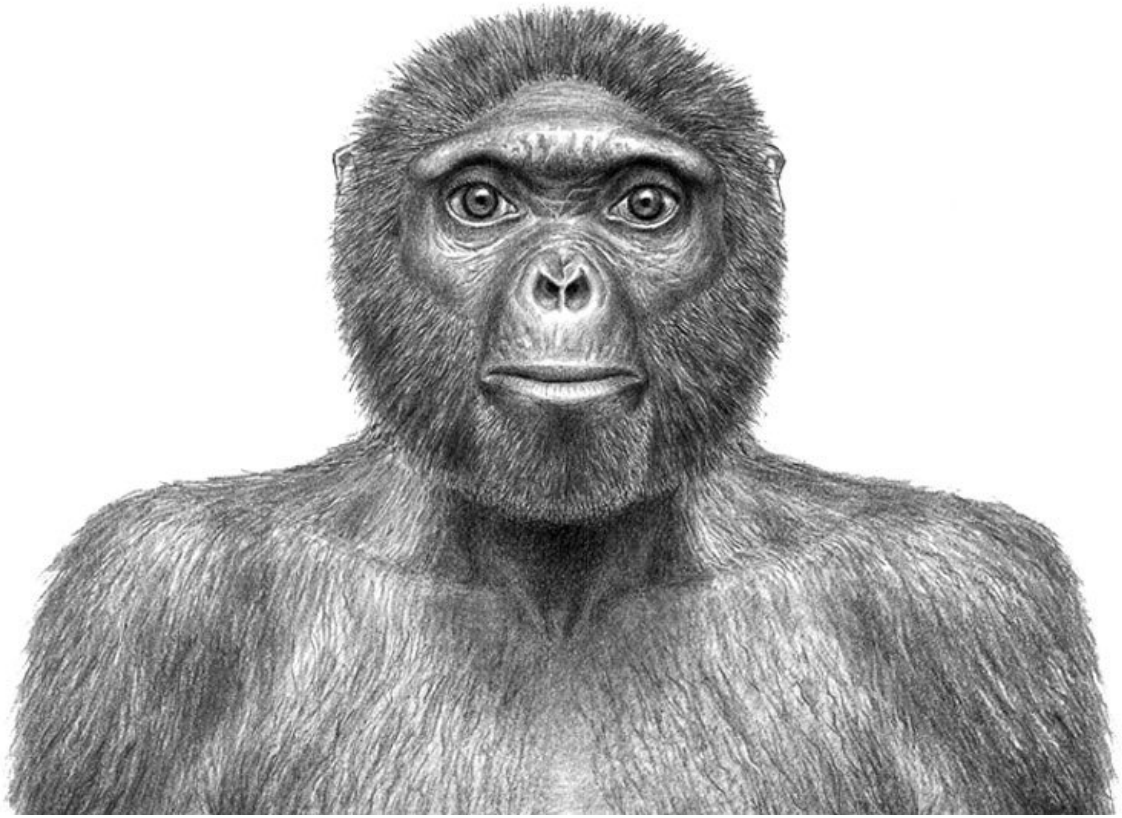
(d) Regression of error in estimated phylogenetic age of M1 against M8.

Performance is similar in a canid data set



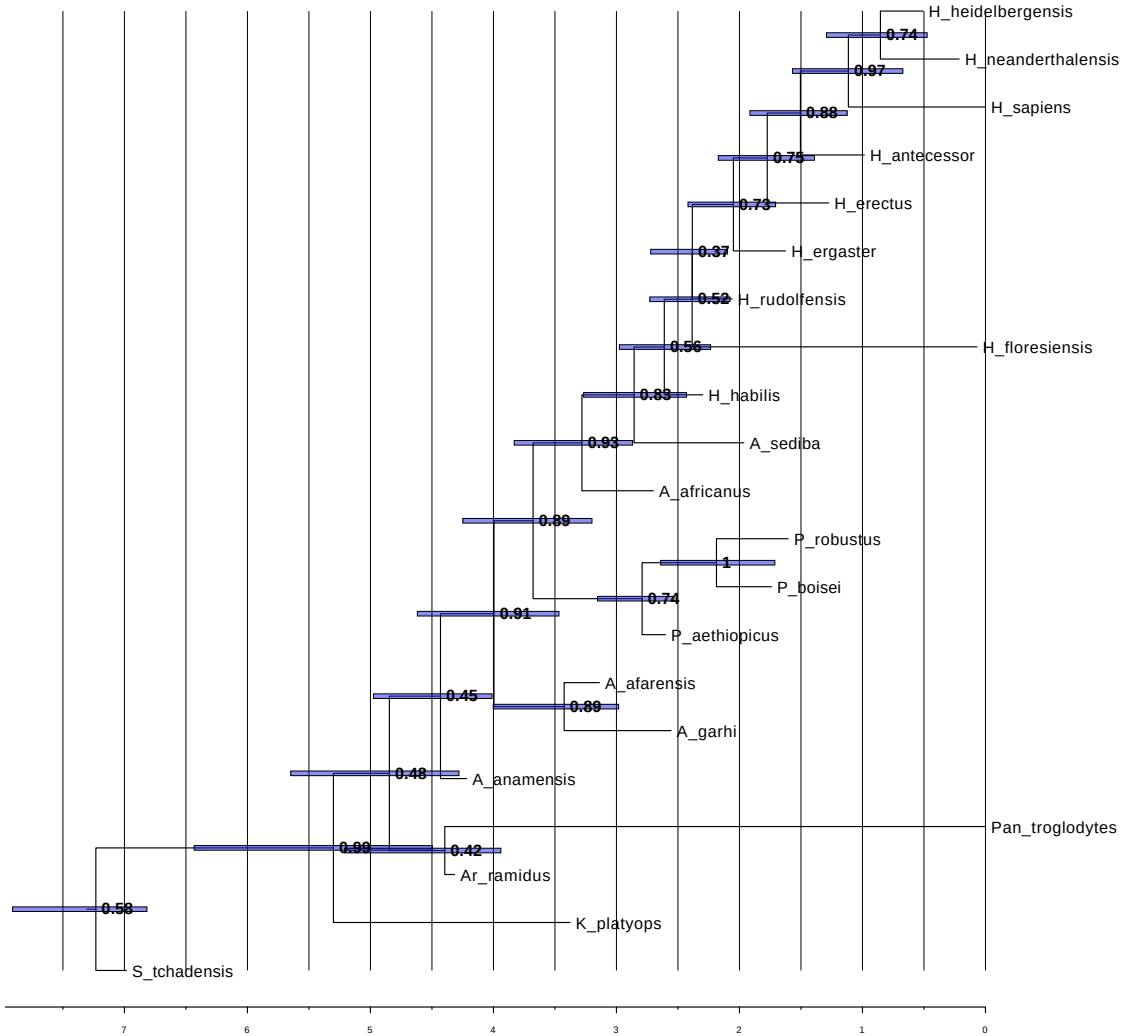
Bayesian phylogenetic ages for 123 canid fossils plotted against palaeontological ages, under M1. 16 fossils with inconsistent estimates are labelled.

Controversies in human evolution: *Ardipithecus ramidus*



'Ardi', 4.4M years old was described as a the oldest fossil relative on the human lineage, a *hominin*, and thus more closely related to the human than the chimpanzee.

Hominin phylogeny inferred using FBD tree prior & morphological clock.

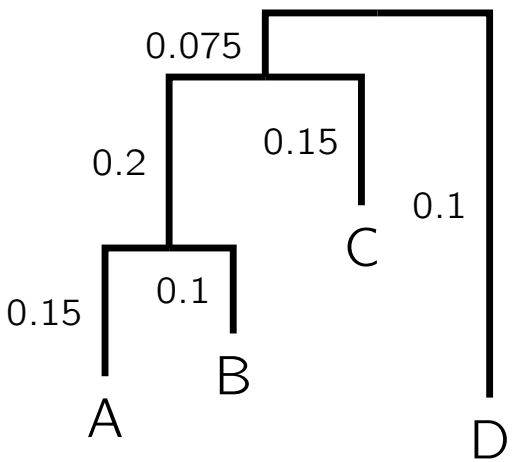


Relaxed phylogenetics

Genetic distance = rate × time

Relaxed molecular clock

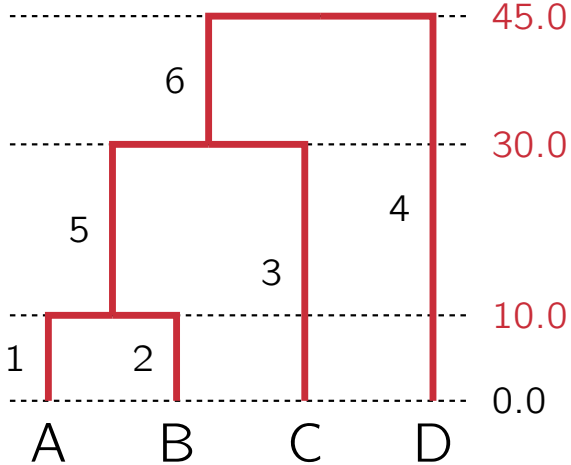
$$T = \vec{\mu} \star g$$



"substitution tree"

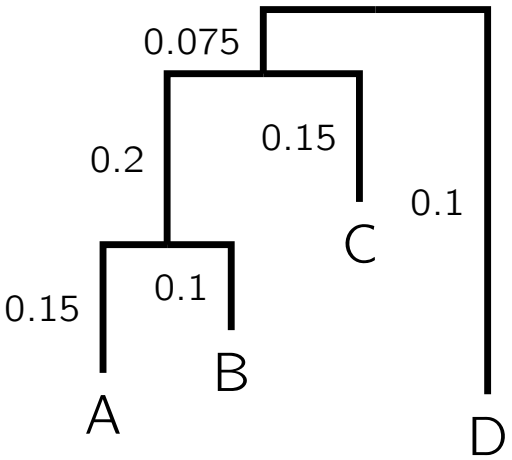
$$= \begin{pmatrix} 0.015 \\ 0.01 \\ 0.005 \\ 0.01 \\ 0.01 \\ 0.005 \end{pmatrix} \star$$

evolutionary rates
substitutions / site / unit time

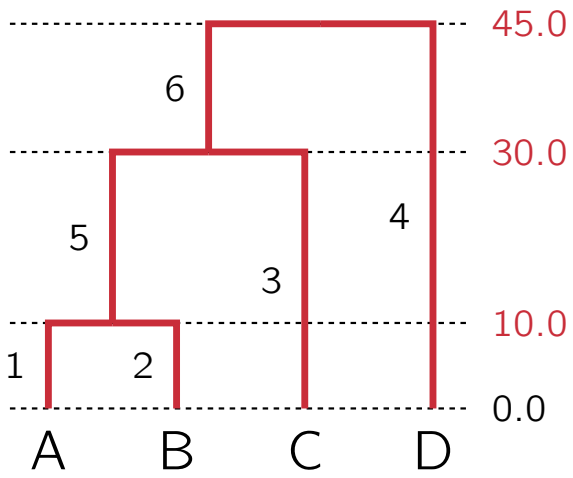


time tree

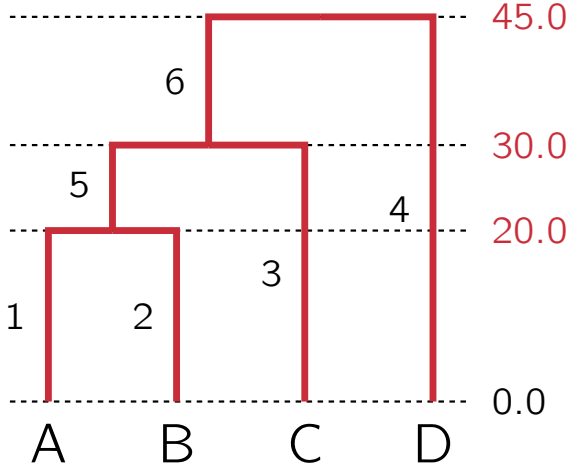
Nonidentifiability in the relaxed clock



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



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

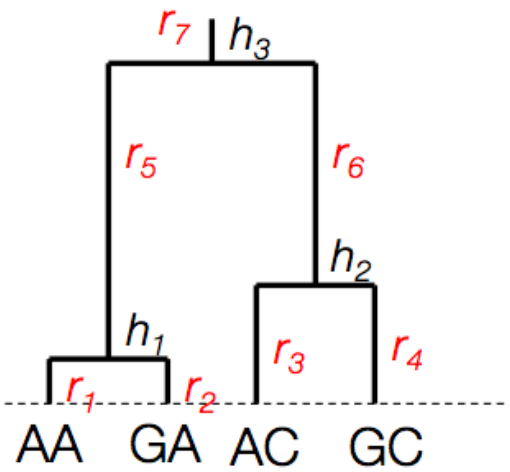


"substitution tree"

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

$$r \sim Exp(\lambda)$$

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

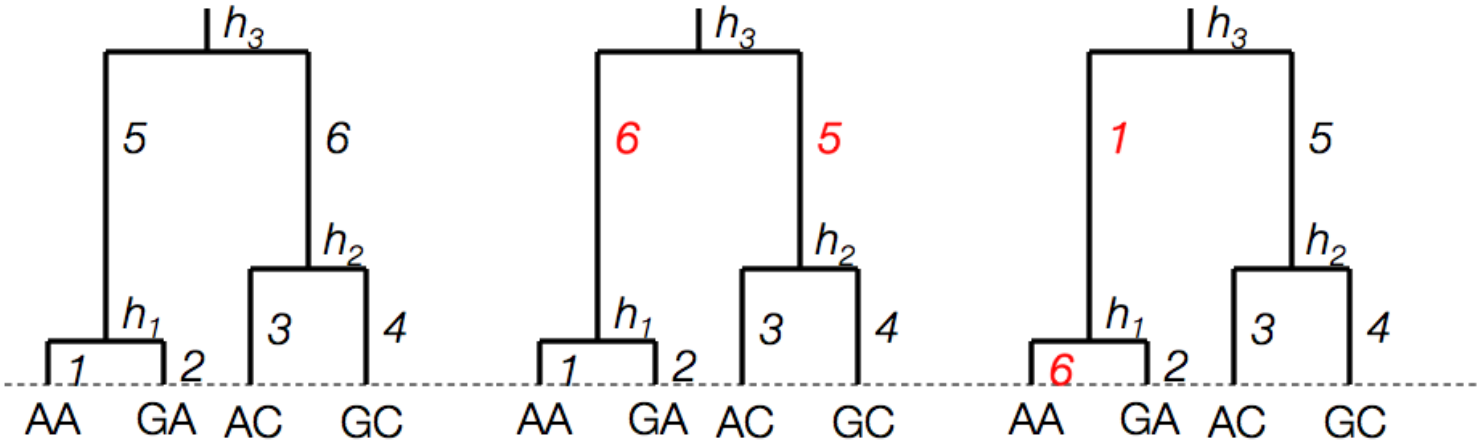
$$r \sim Gamma(\alpha, \beta)$$

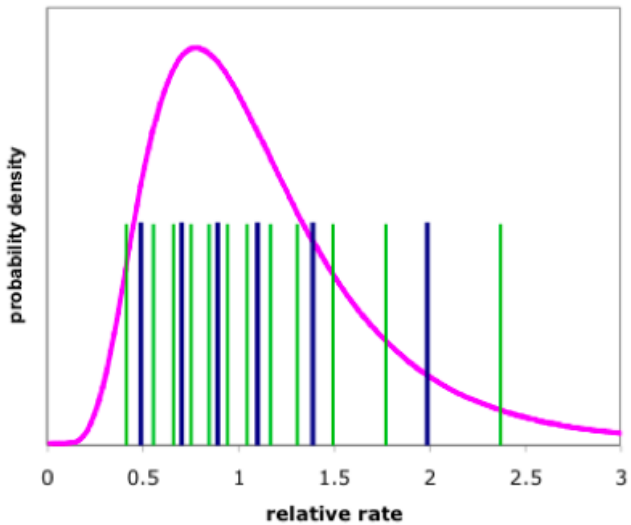
We introduce a relaxed clock model in which there is no prior correlation between child and parent rates

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

ML

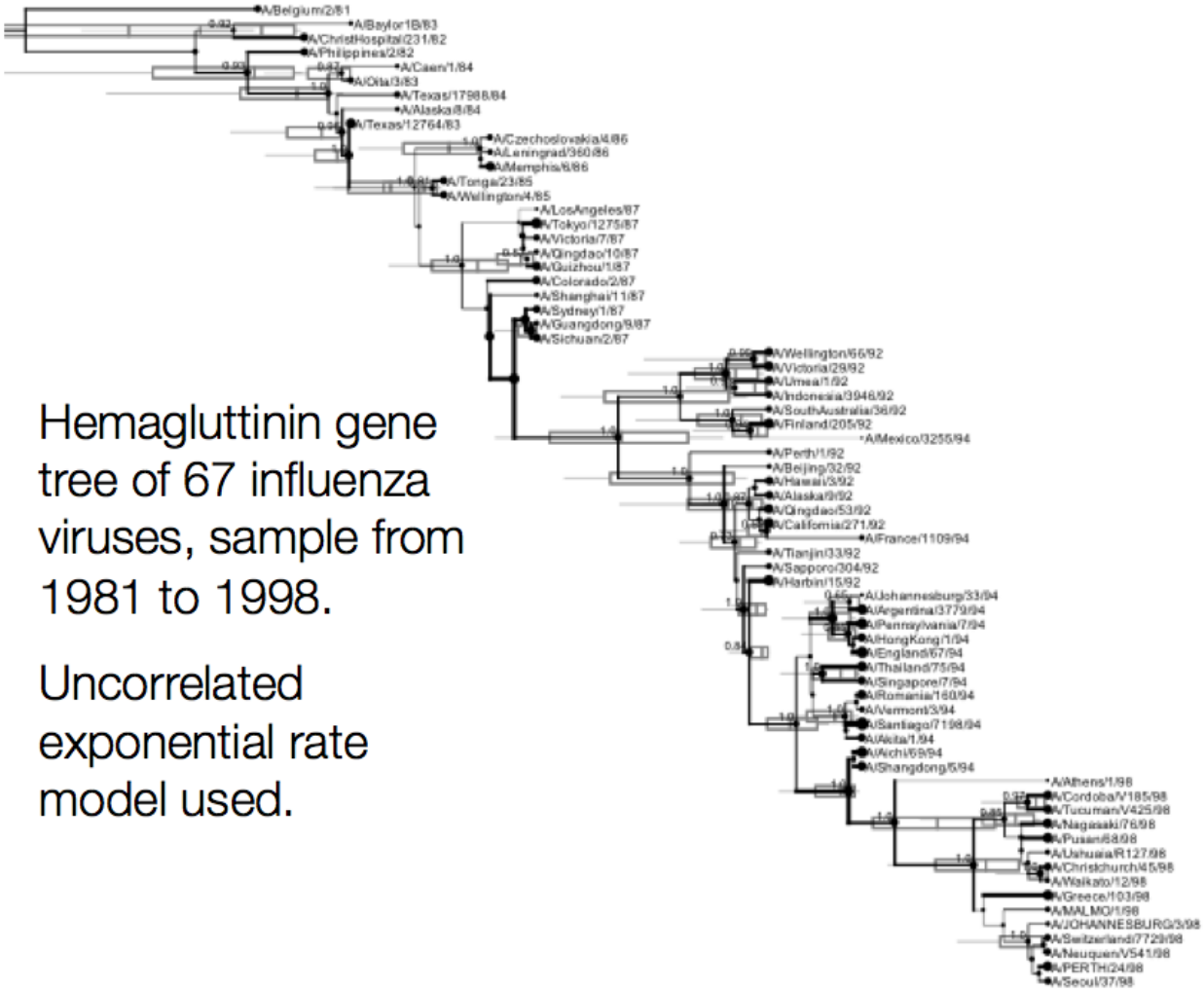
Sampling branch rates using MCMC





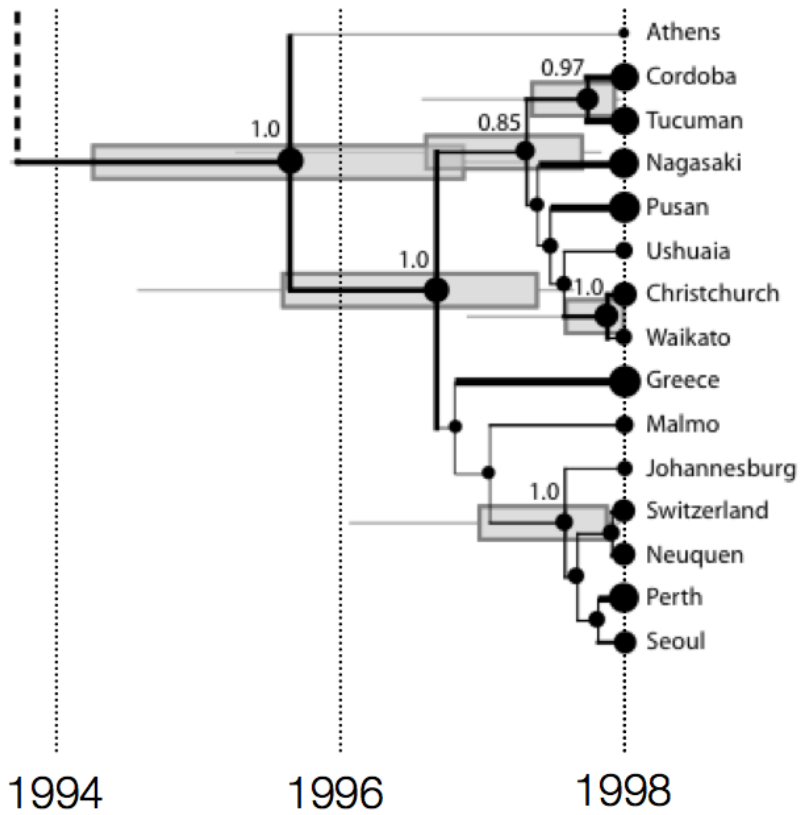
1. Rates are summarized into 2n-2 rate categories (e.g. blue is 6 categories; green is 12 categories).
2. Rates categories are sampled during MCMC by two operators:
    1. Random walk operator
    2. Swap operator
3. For purposes of topology changes, rate categories are associated with child node.

Influenza A gene tree estimated by relaxed molecular clock



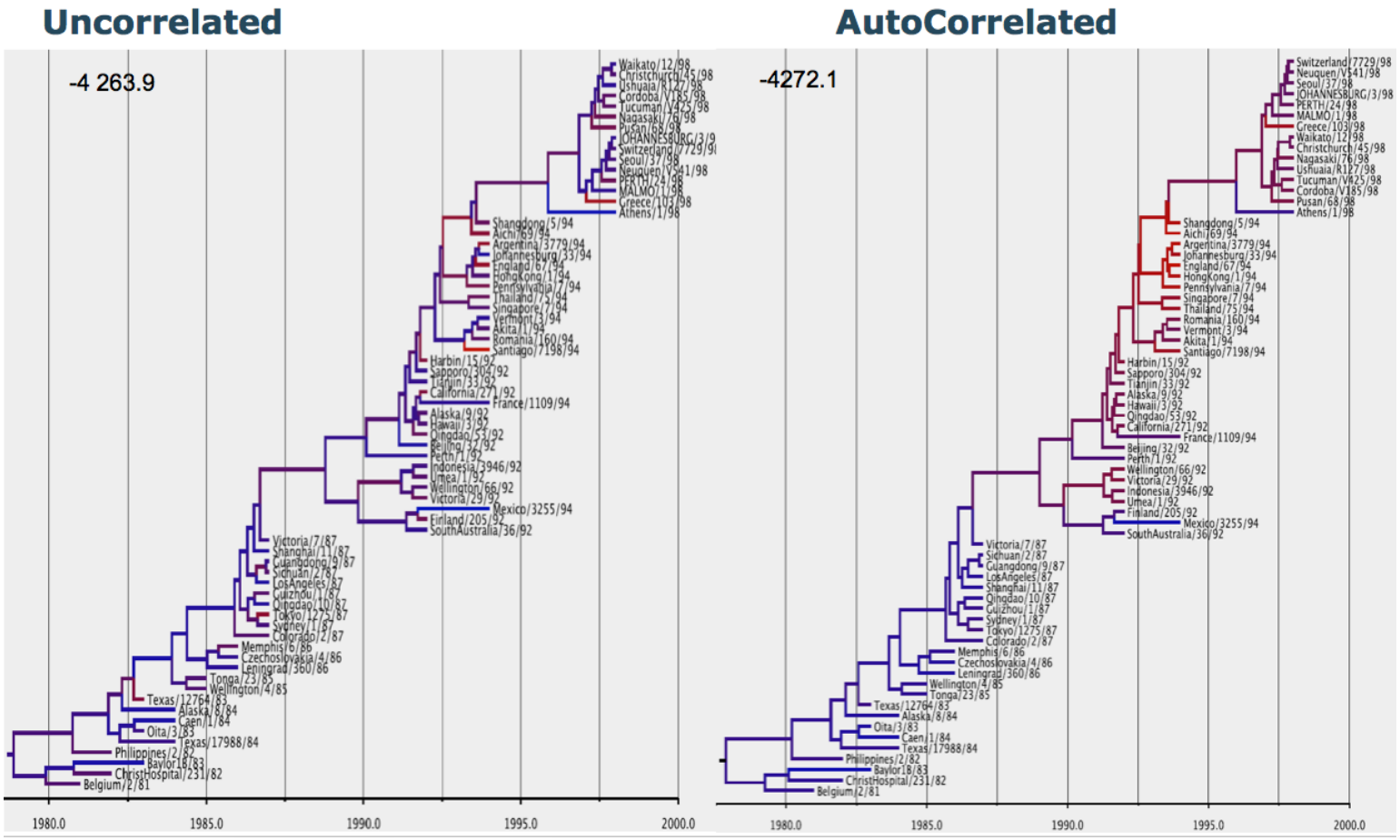
1. Hemagluttinin gene tree of 67 influenza viruses, sample from 1981 to 1998.
2. Uncorrelated exponential rate model used.

Influenza A gene tree estimated by relaxed molecular clock

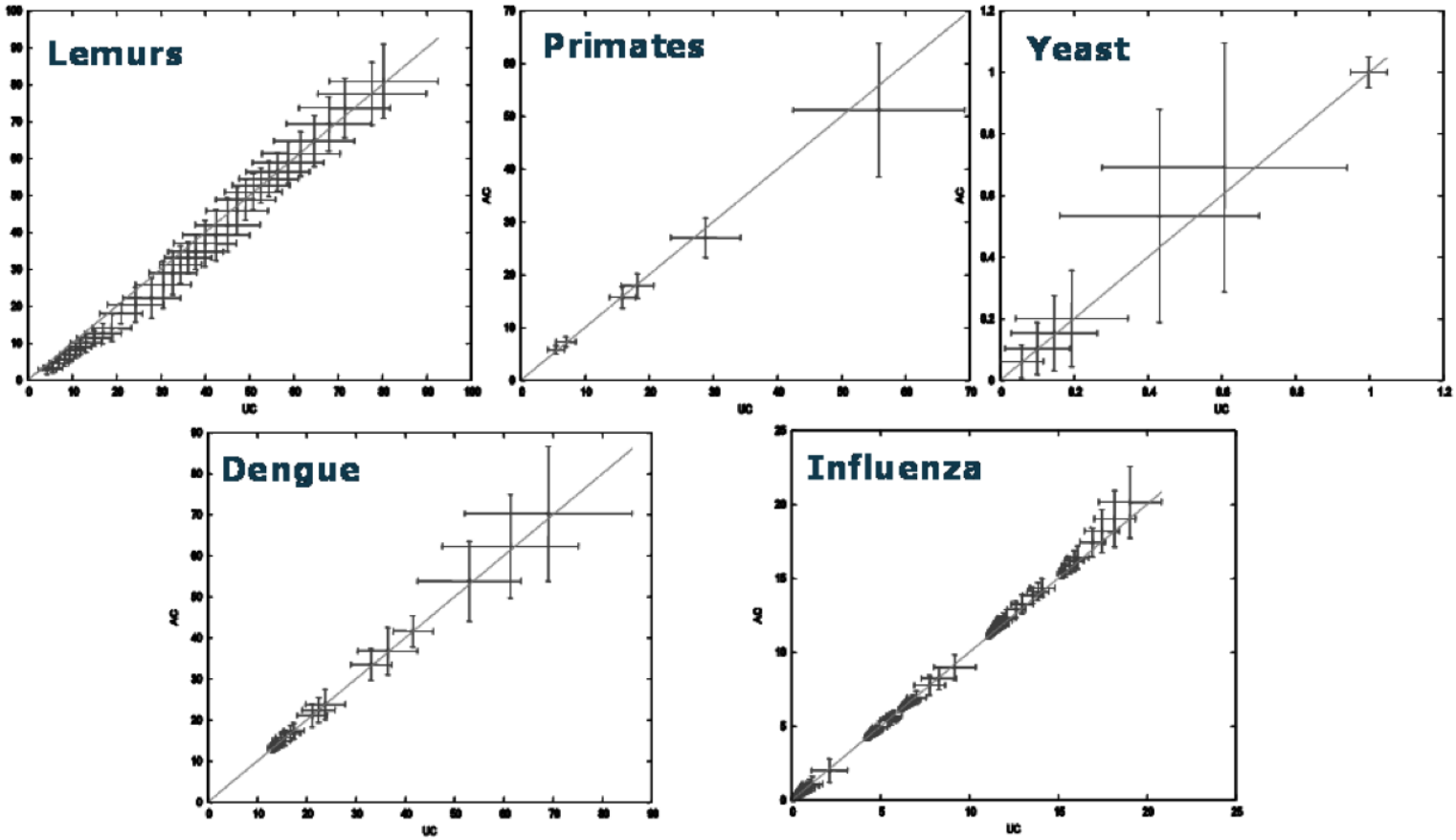


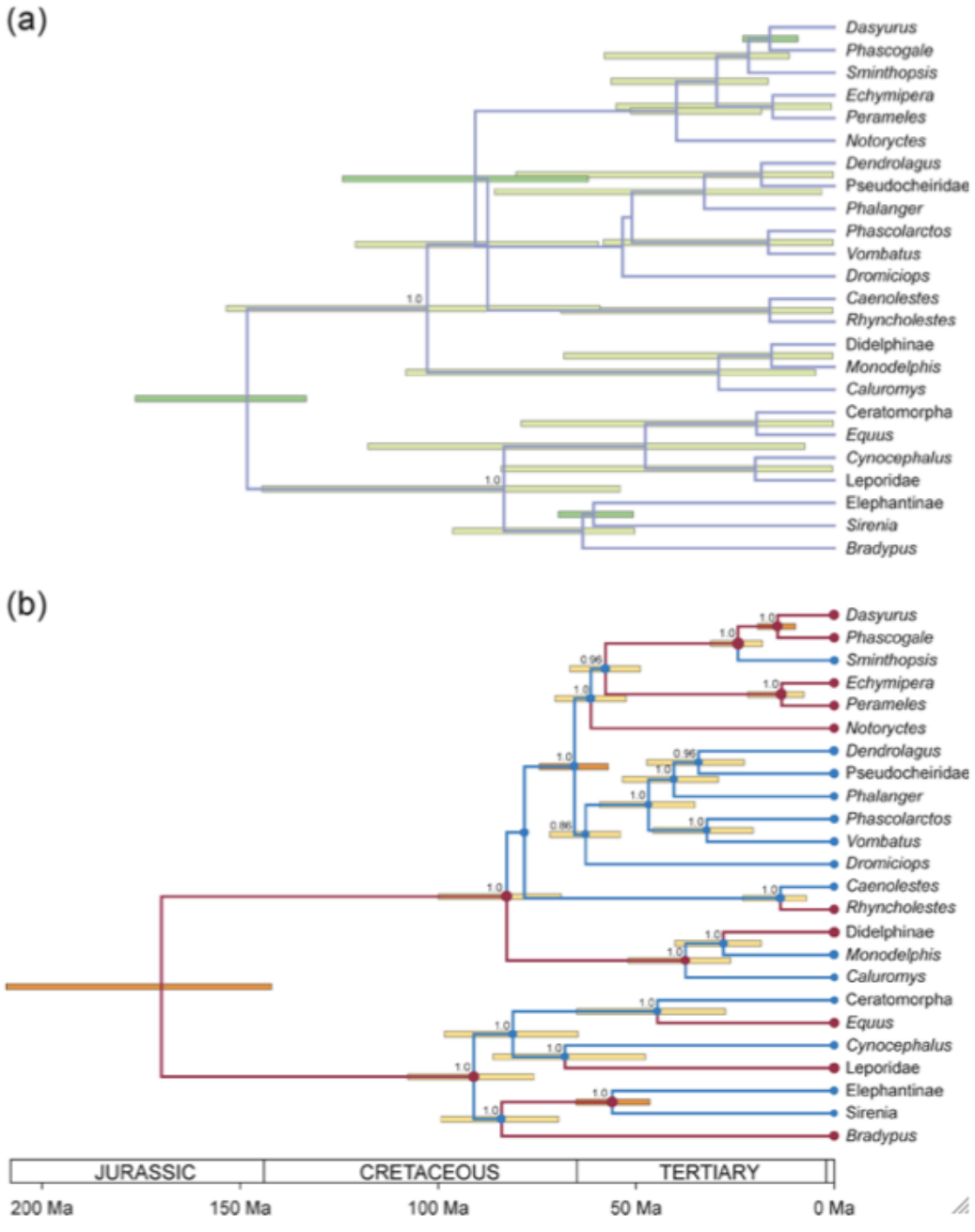
- Box-and-whisker plots show uncertainty in divergence times (only for splits with posterior probability > 0.5)
- Node size and branch thickness proportional to evolutionary rate.

Influenza trees under different relaxed clock models



UC versus AC on five data sets



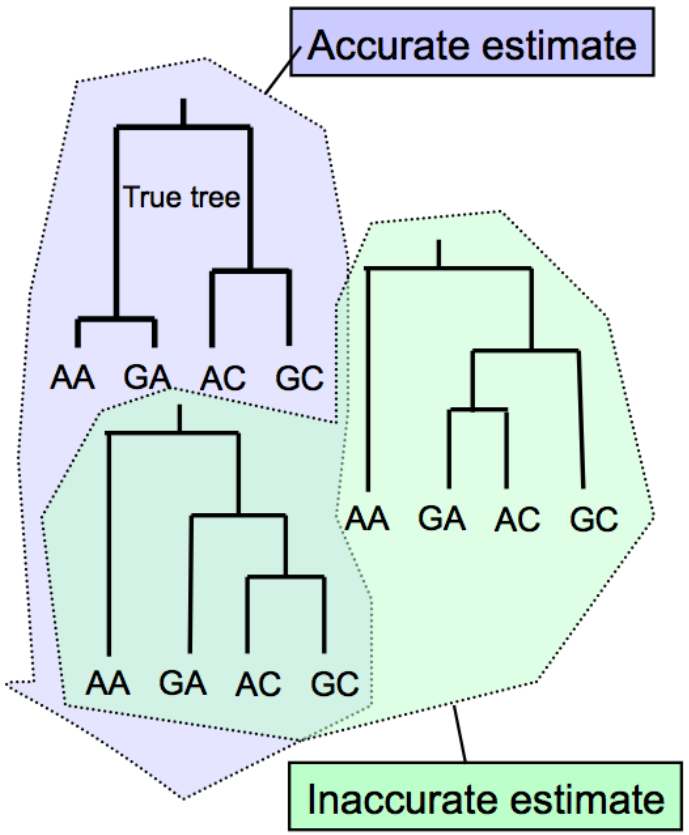


Prior versus Posterior

Marsupials example (24 taxa, 5658 nucleotides)

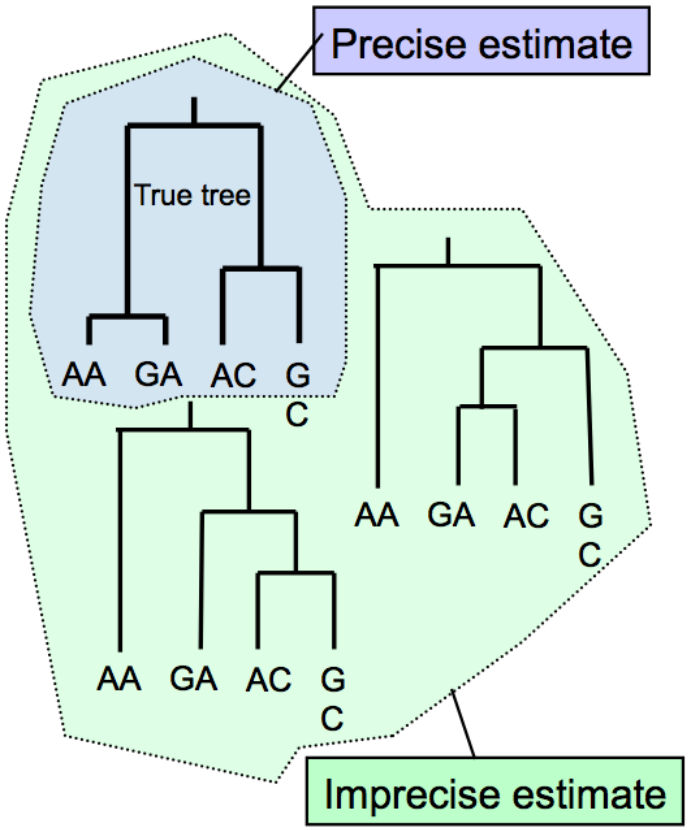
Accuracy in Bayesian Phylogenetics

- 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

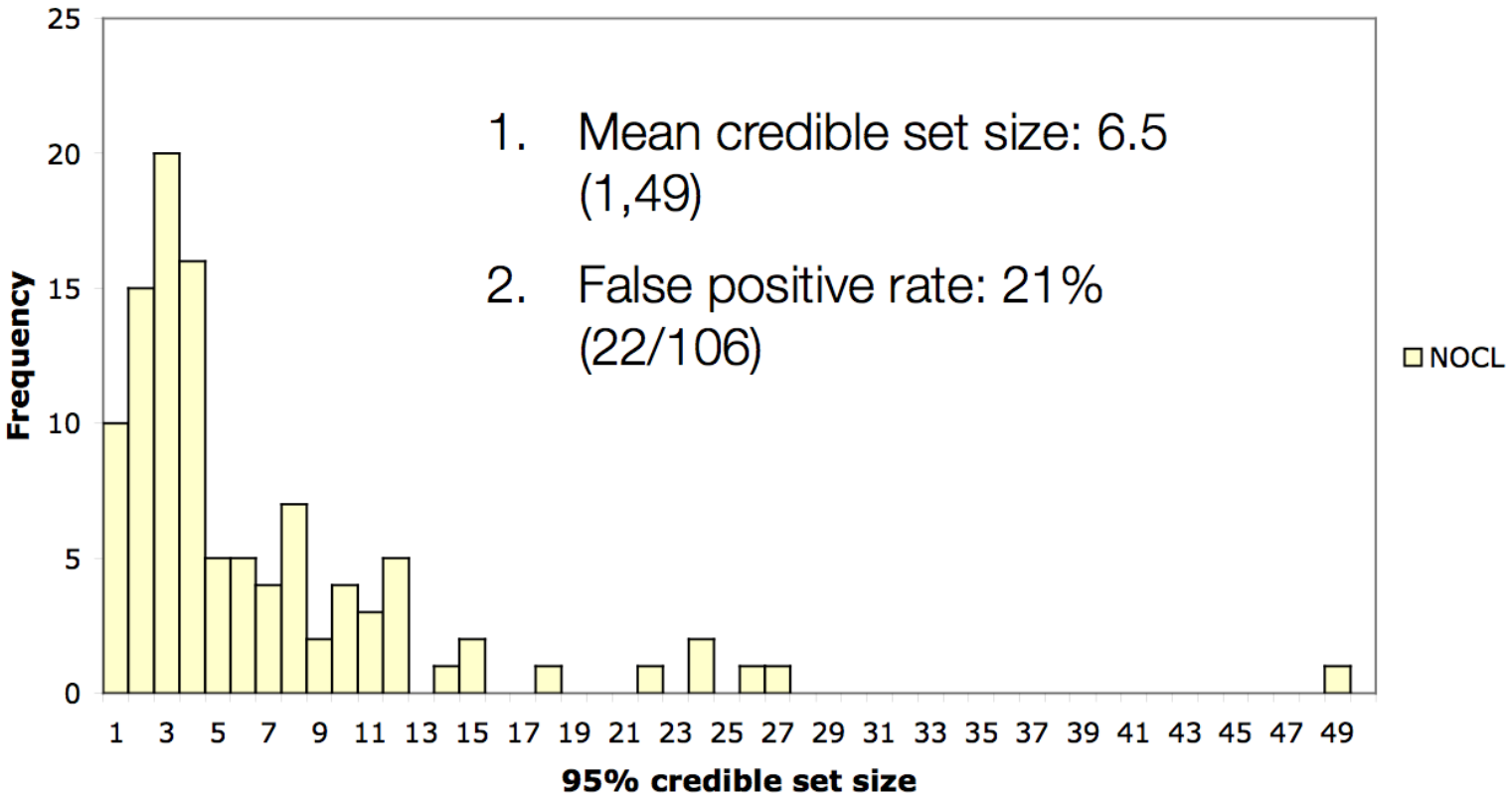
- The precision of an estimate can be described by how much is excluded.
- How small is the 95% credible set of trees?



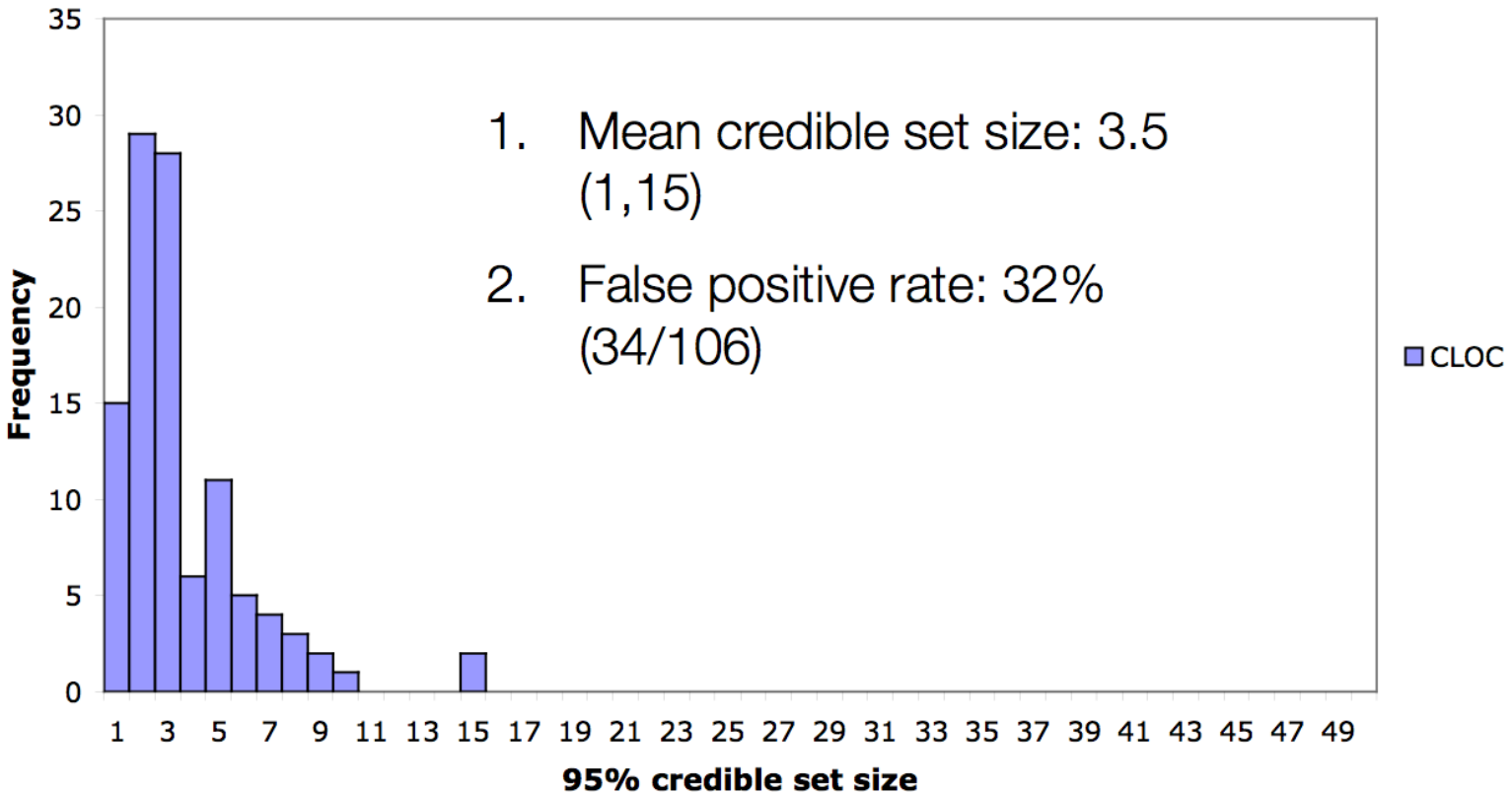
Testing Accuracy and Precision with **real data**

- 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)

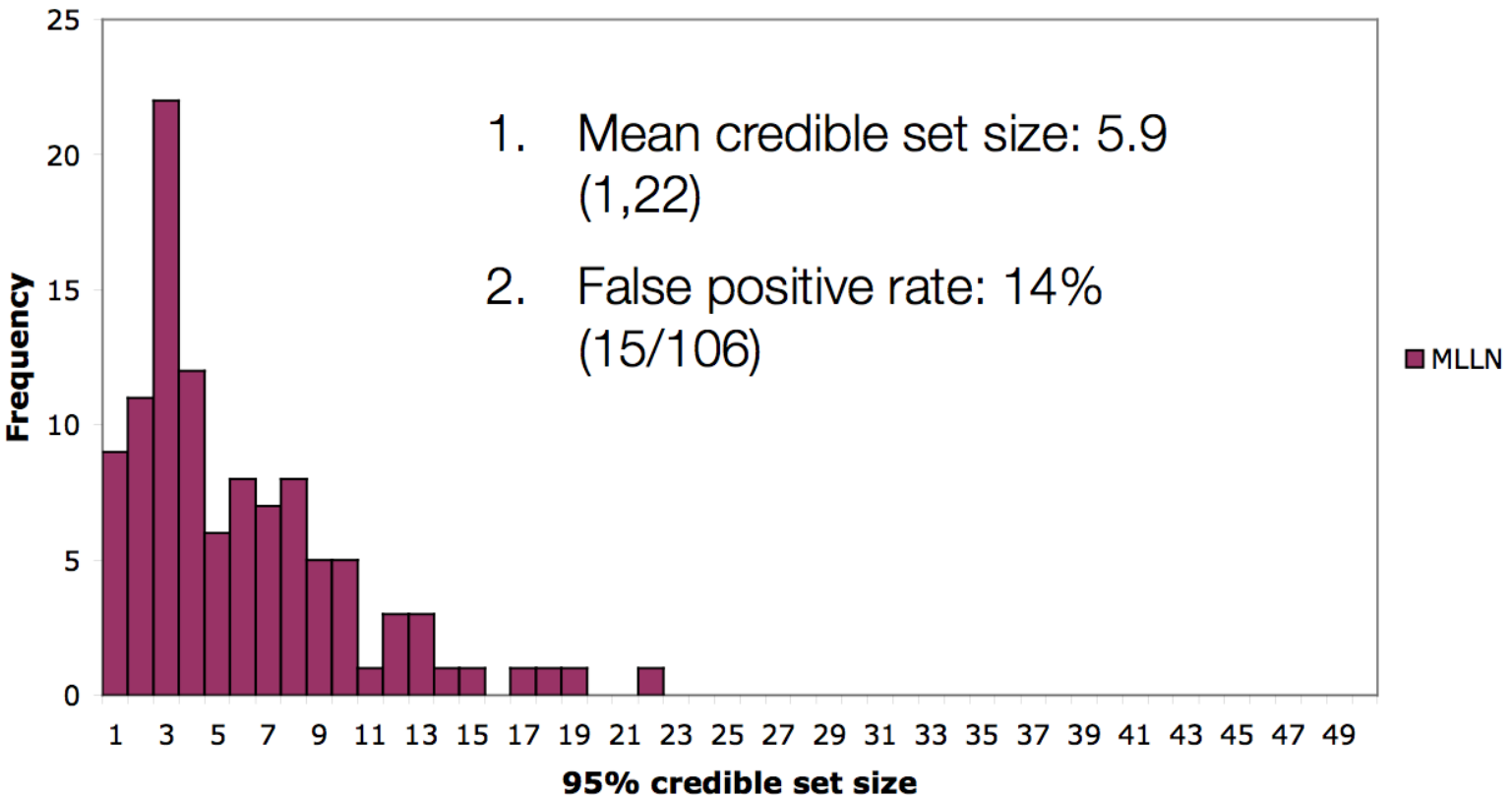
Rokas data: MrBayes tree estimates



Rokas data: Strict clock tree estimates from BEAST



Rokas data: Relaxed clock tree estimates from BEAST



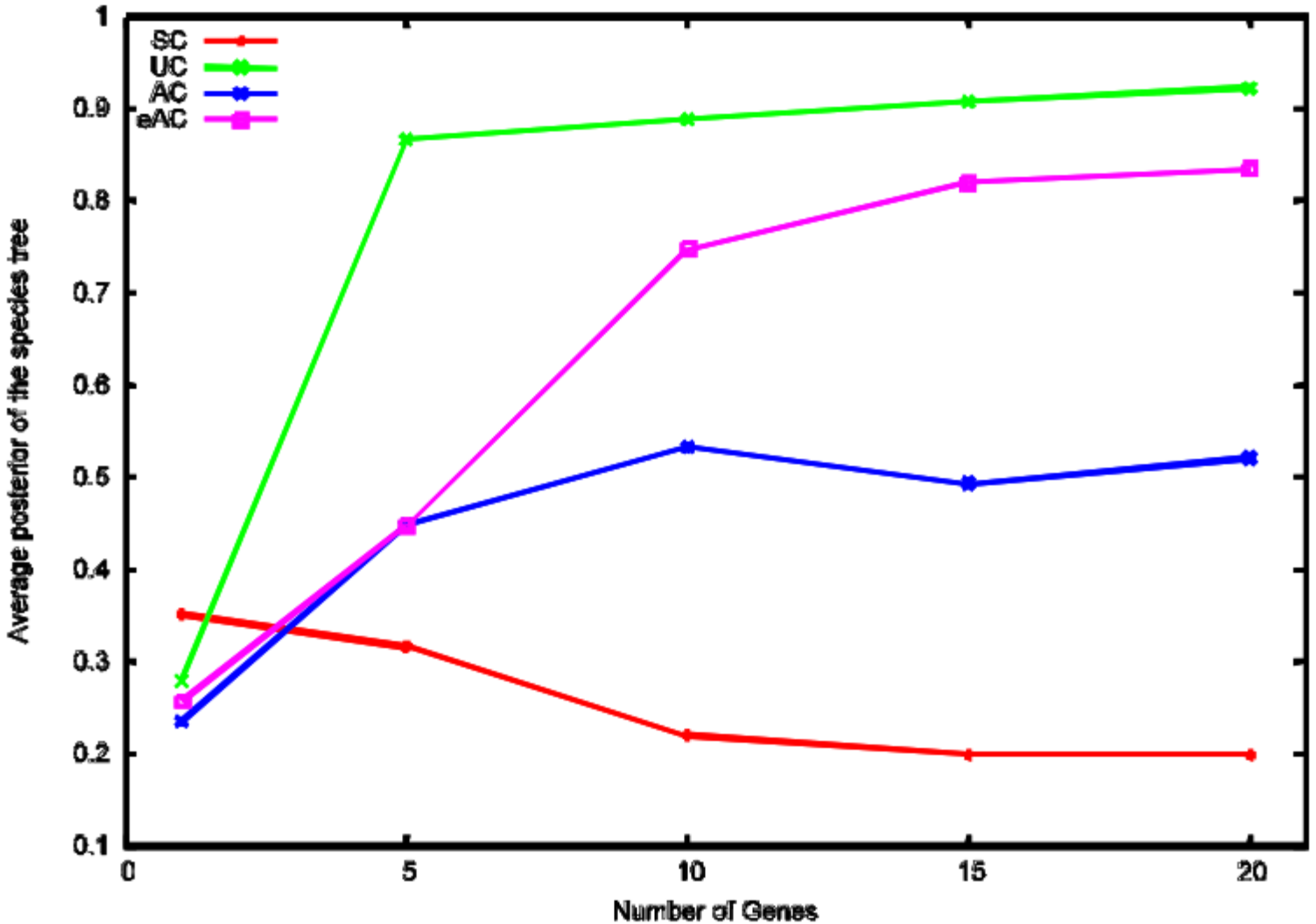
Summary of Bayesian Accuracy on five large data sets

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

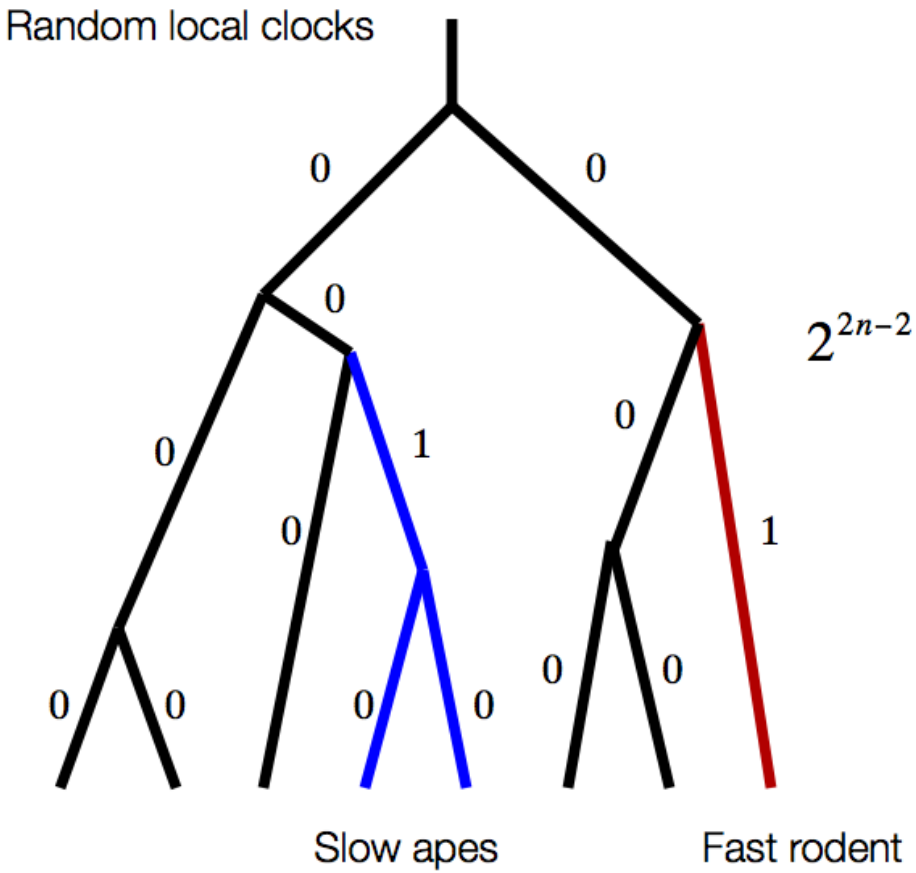
Summary of Bayesian Precision on five large data sets

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

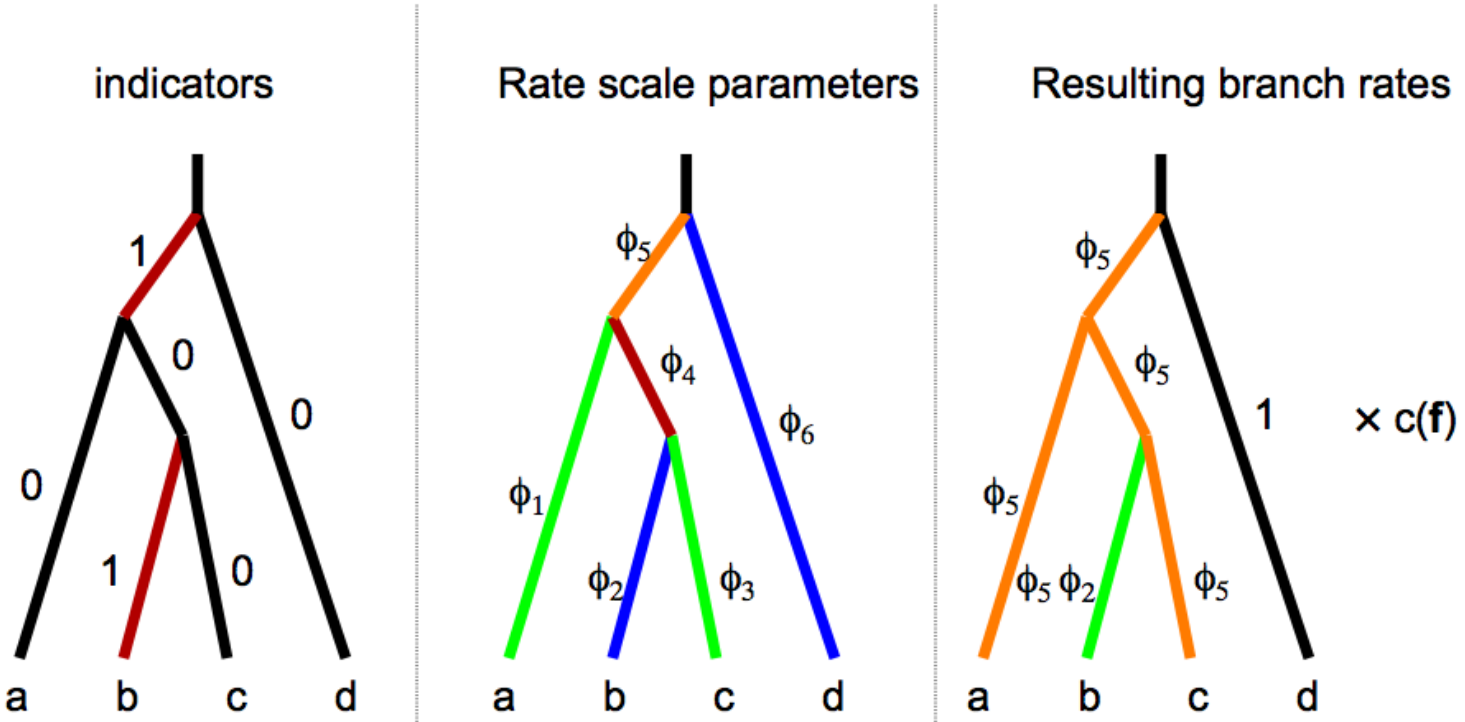


Random local molecular clocks



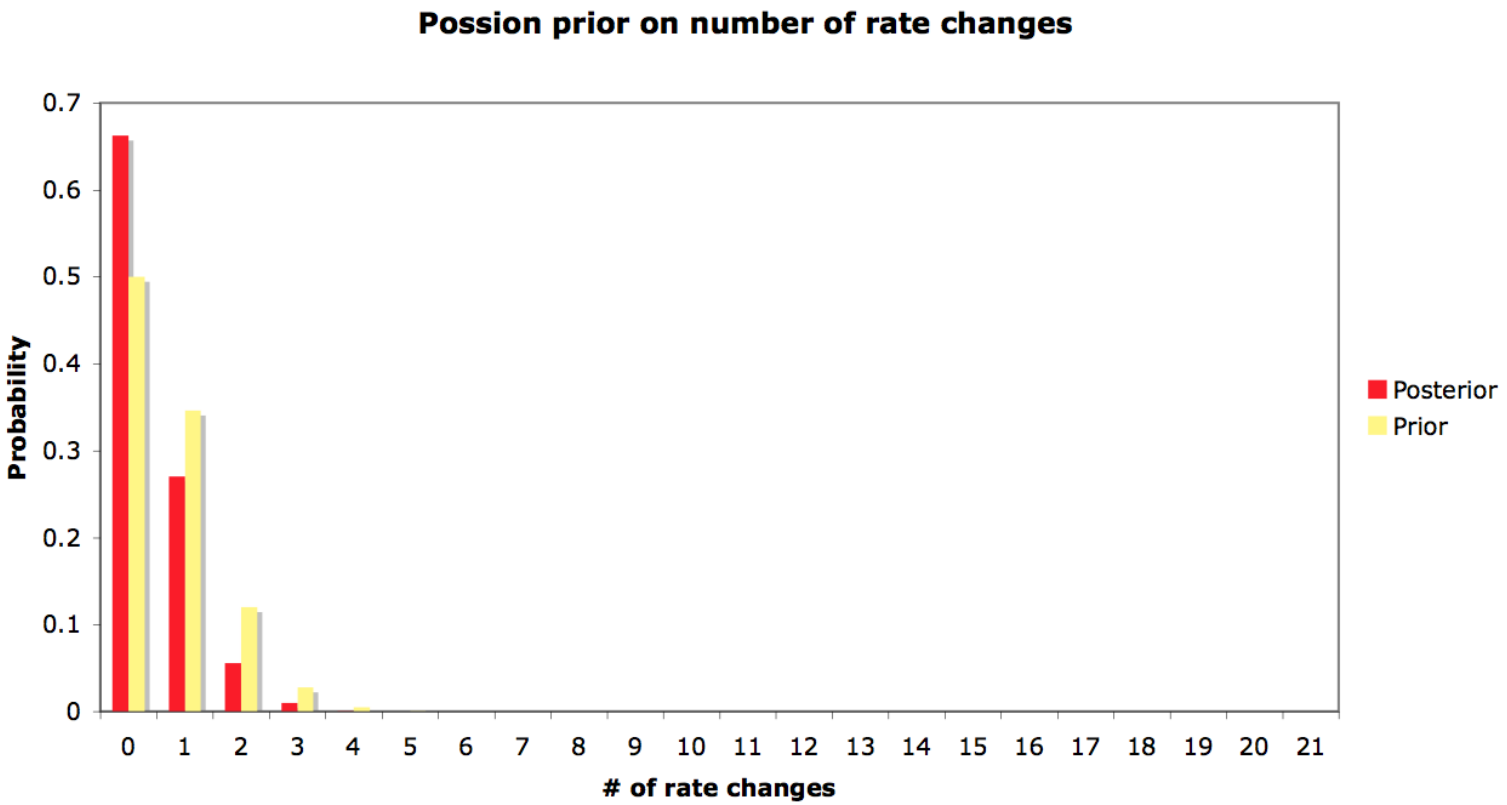
2^{2n-2} local molecular clock models

Random local molecular clocks

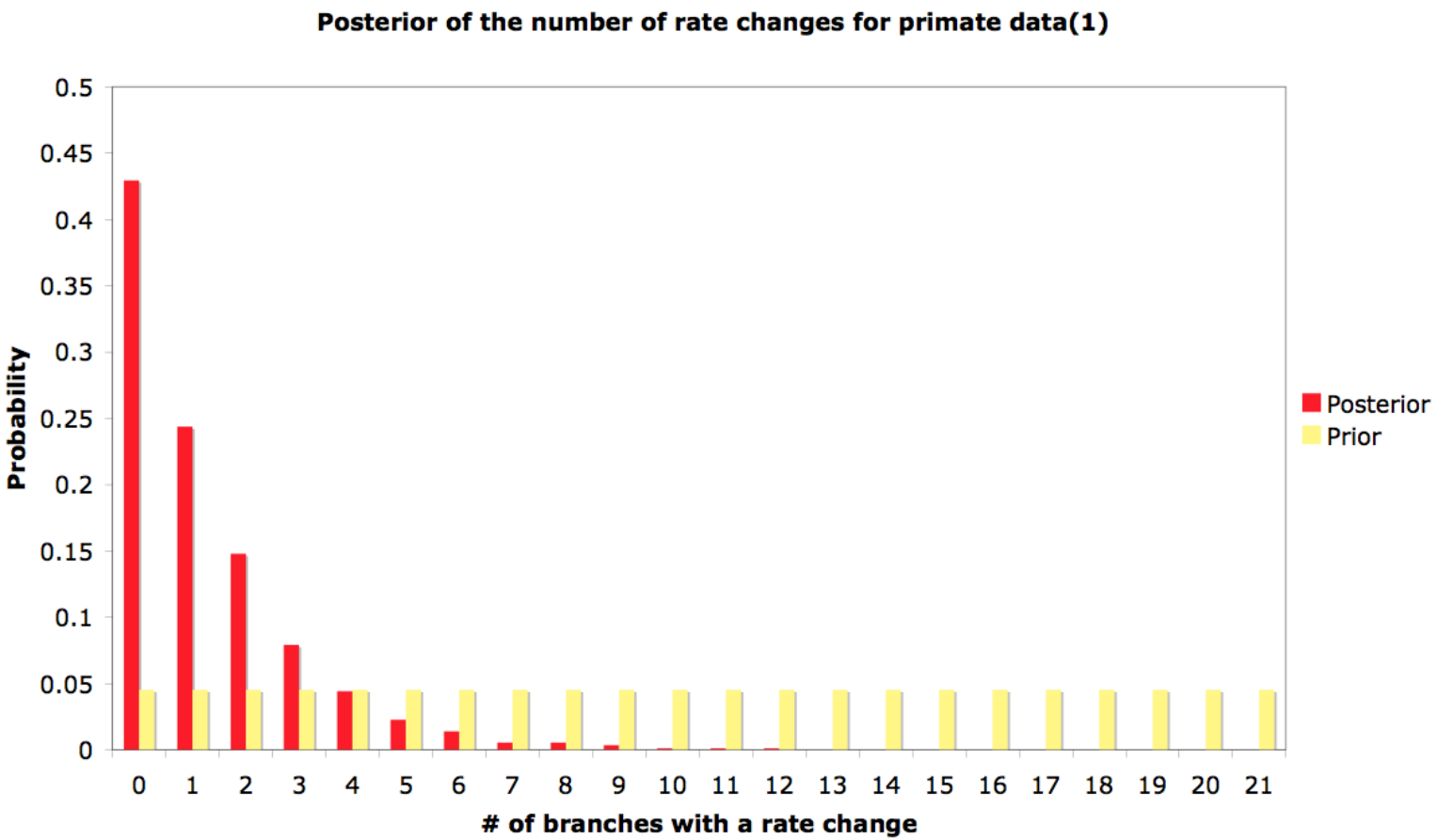


Red/Orange fast, Green/Blue slow

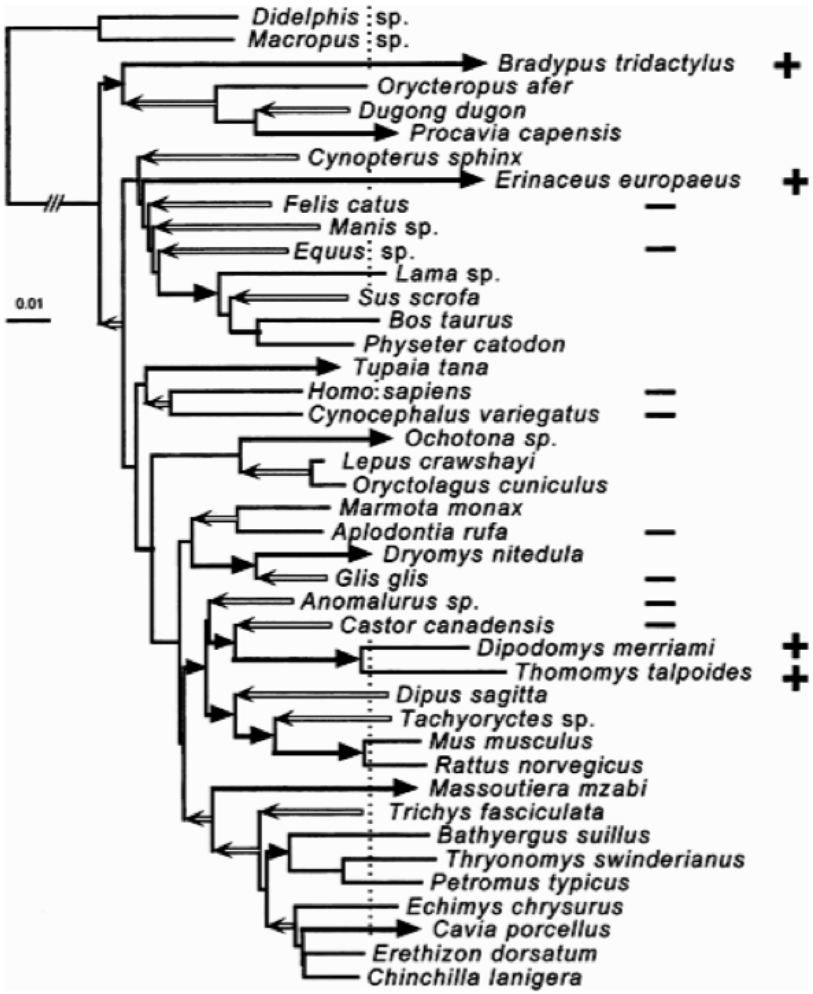
Primate data set (Poisson prior on # rate changes)



Primate data set (Uniform prior on # rate changes)



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

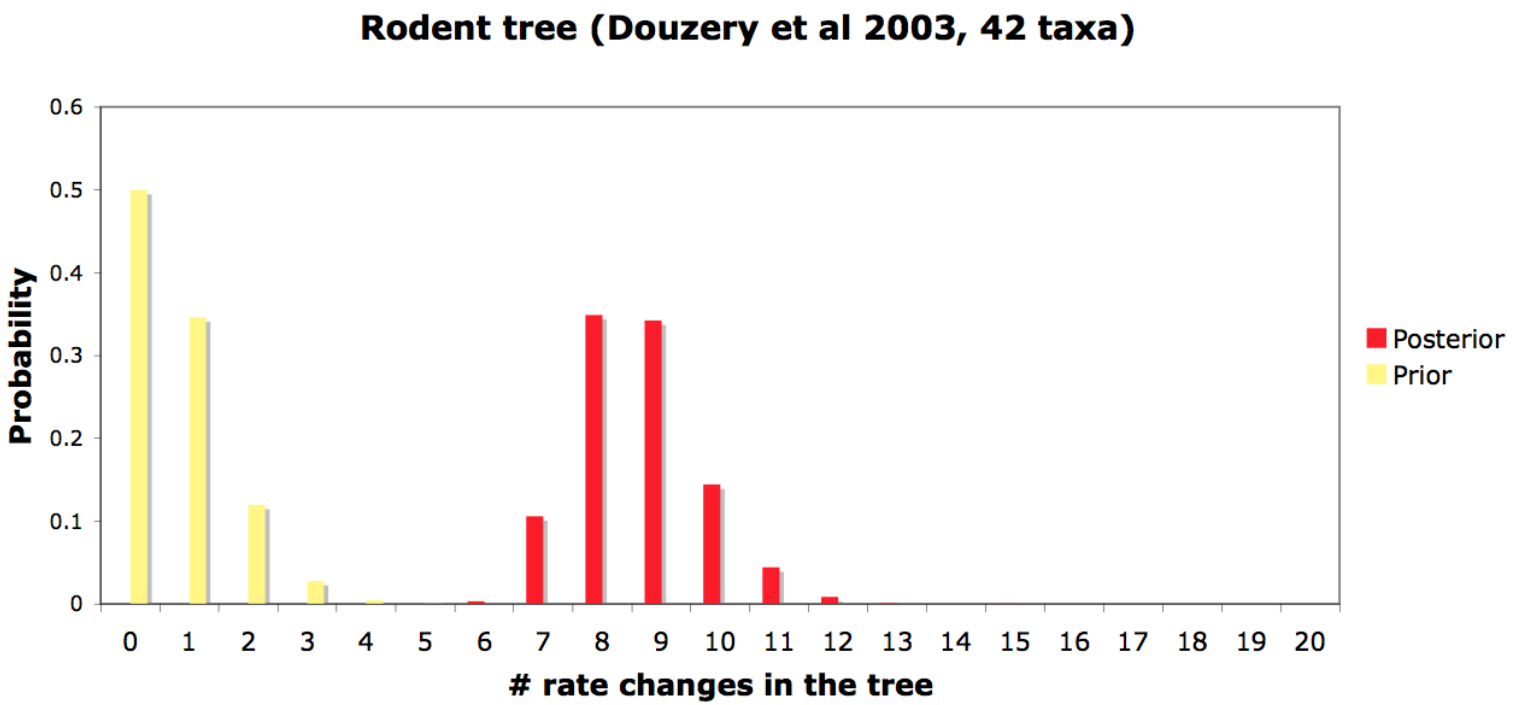


82 branches

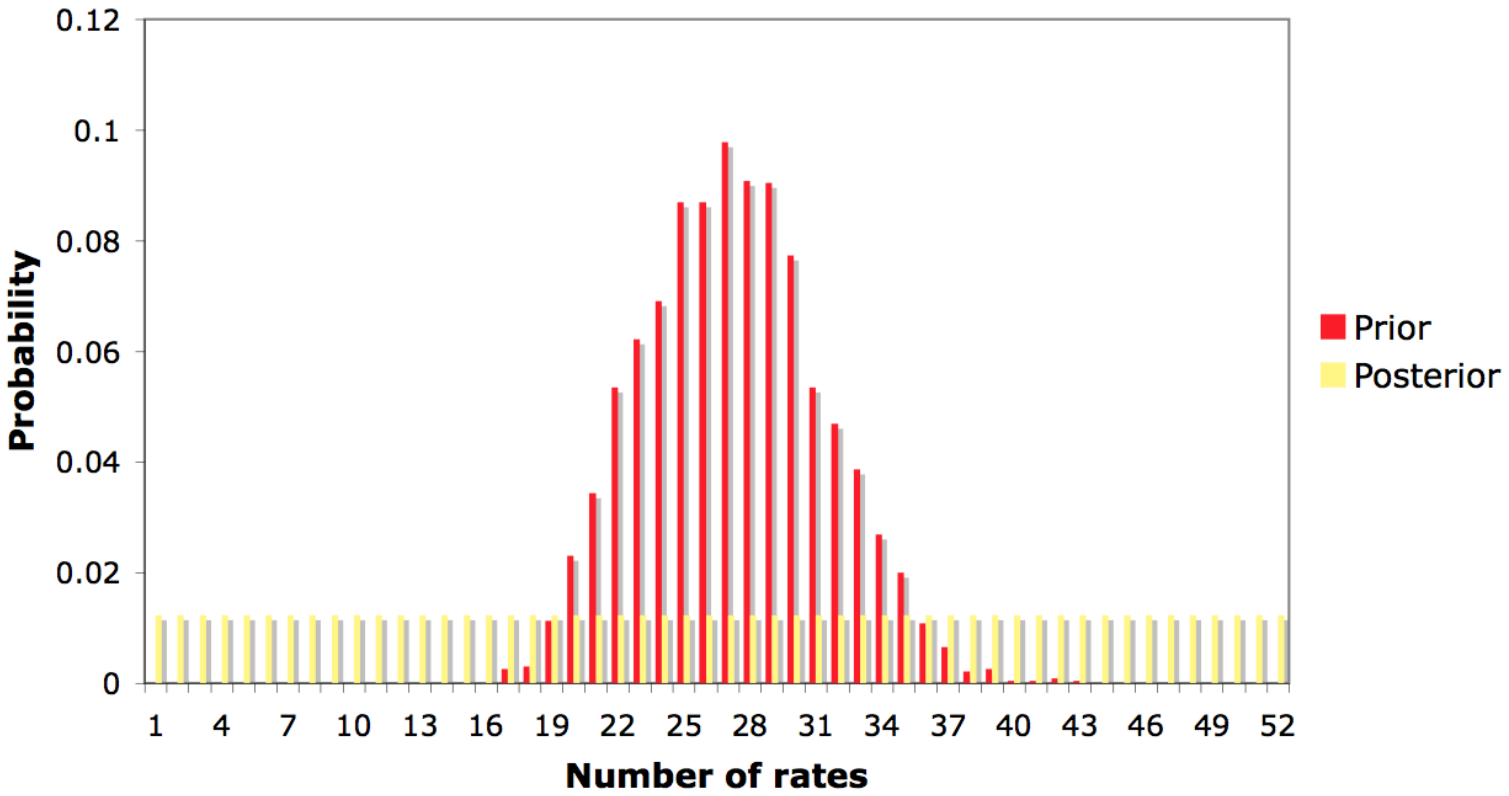
38 rate changes according to Douzery et al 2003

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).

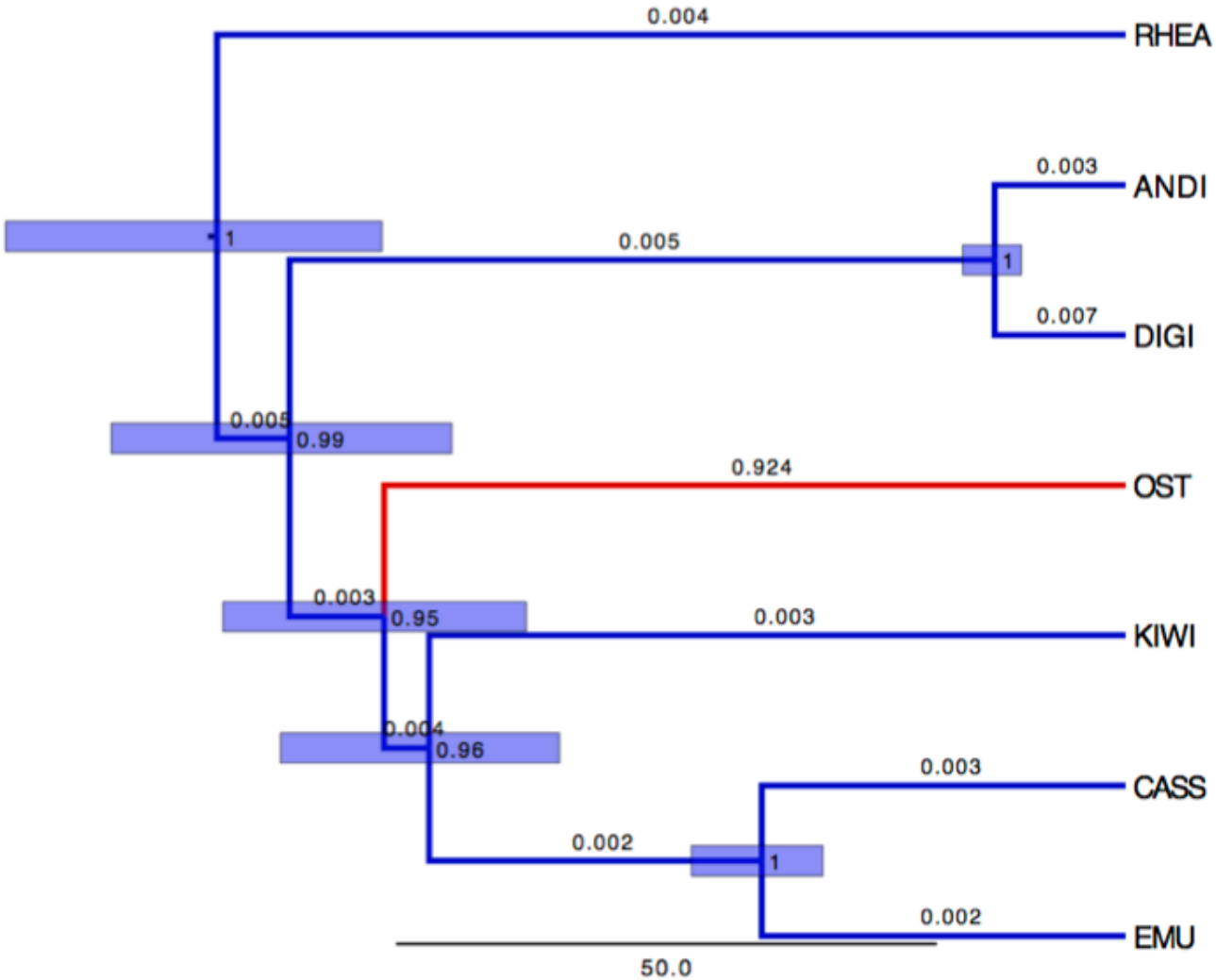
Rodent data set (Poisson prior on # rate changes)



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.

References I