Clocks, calibrations + tree priors

Professor Alexei Drummond

Director of Centre for Computational Evolution
Department of Computer Science
University of Auckland

13th August 2019, Taming the BEAST Eh!, Squamish, Canada

Outline

- **Clocks and calibrations** - a modelling framework for modelling rates of evolution

- **Relaxed clocks** - modelling variation of rates across branches.

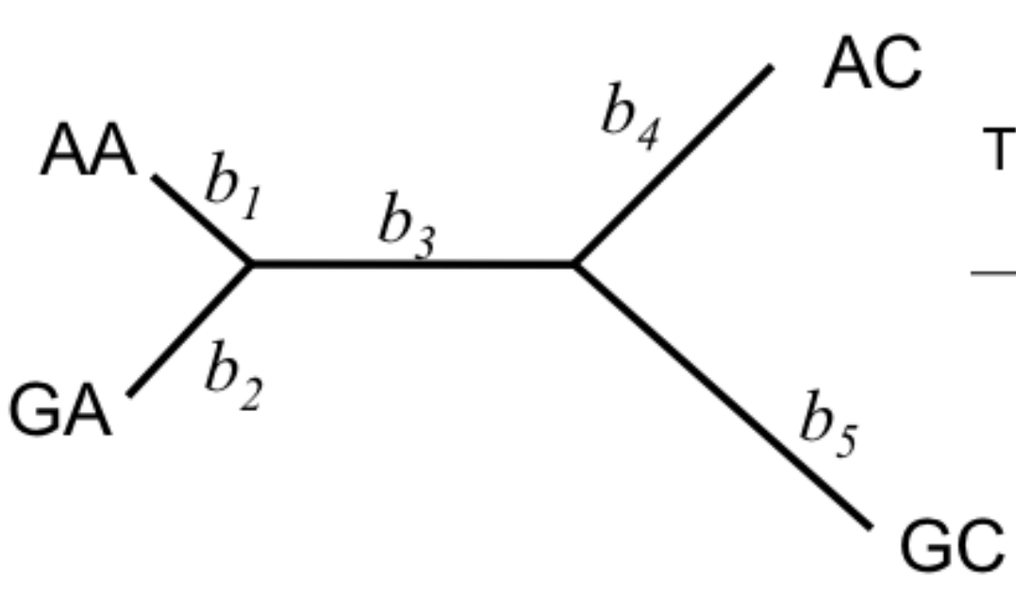
- **Tree priors** - modelling phylogenetic dynamics of infectious disease (and macroevolution).

- **Integrative models**

Clocks and calibrations

The molecular clock constraint

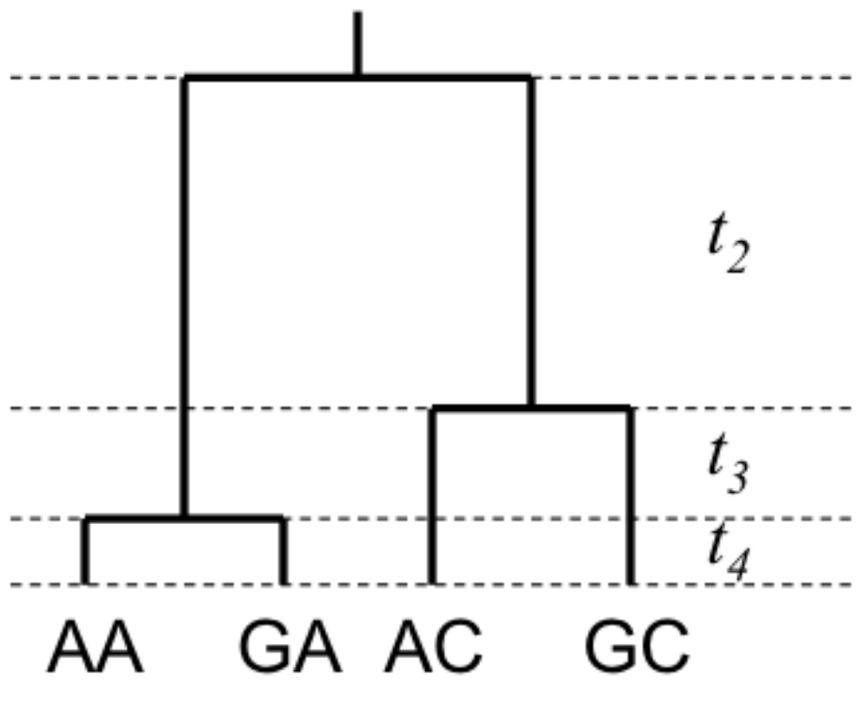
T



2n–3 branch lengths

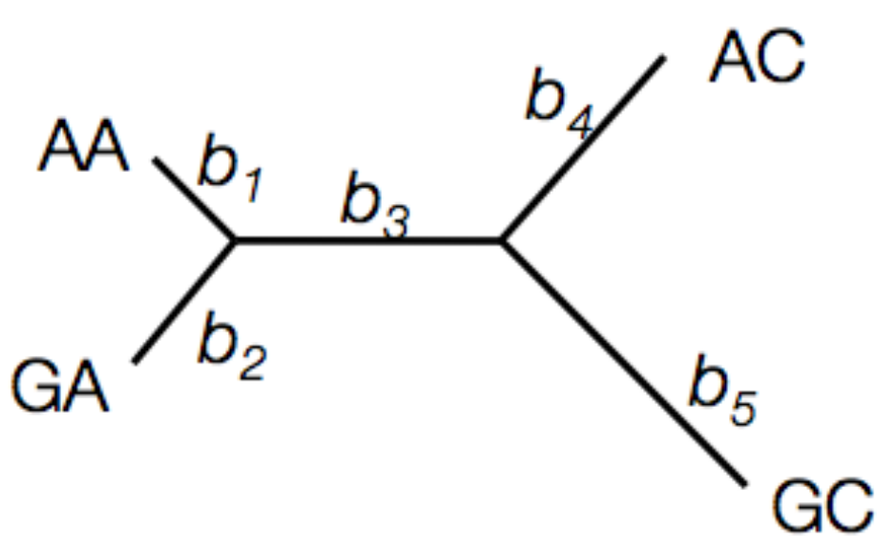
The "molecular clock" constraint

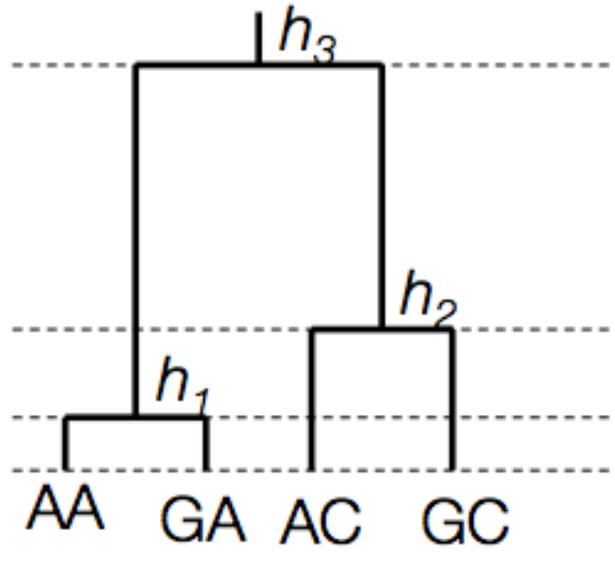
g



n–1 waiting times

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:

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

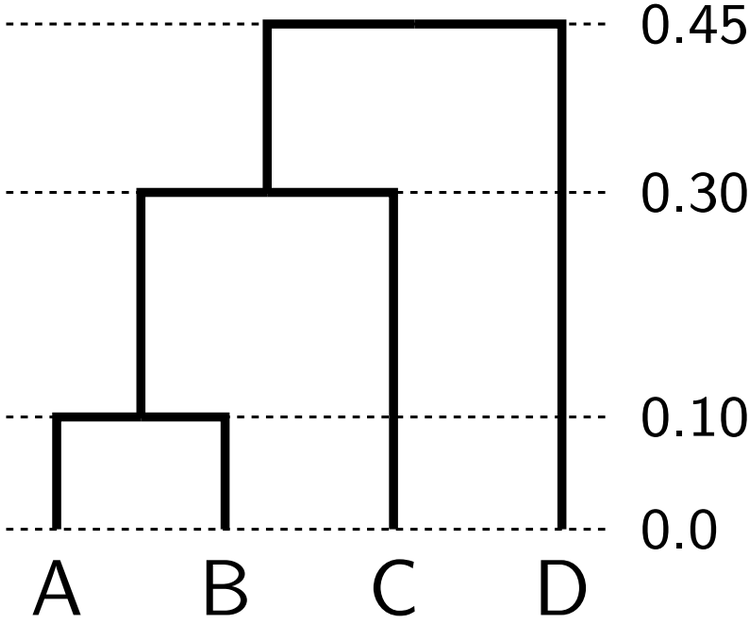
Estimate the global rate:

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

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

Genetic distance = rate x time

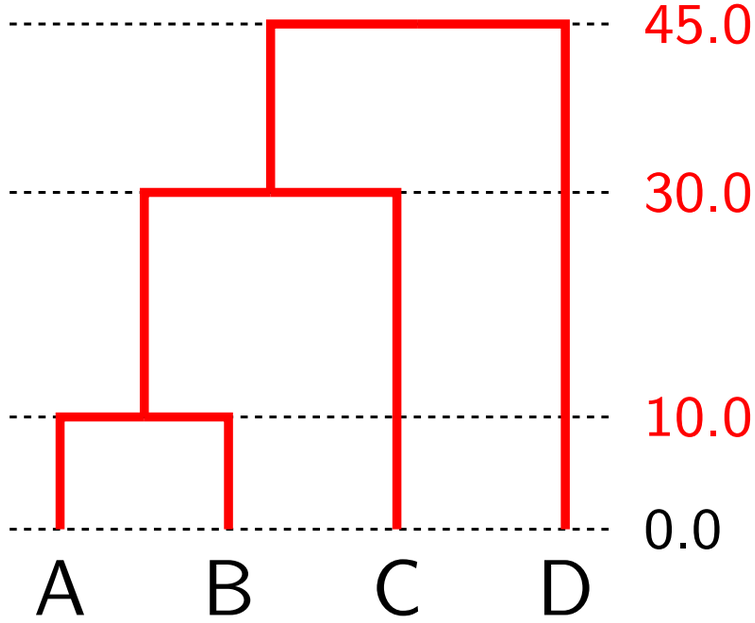
$$T = \mu \times g$$



"substitution tree"

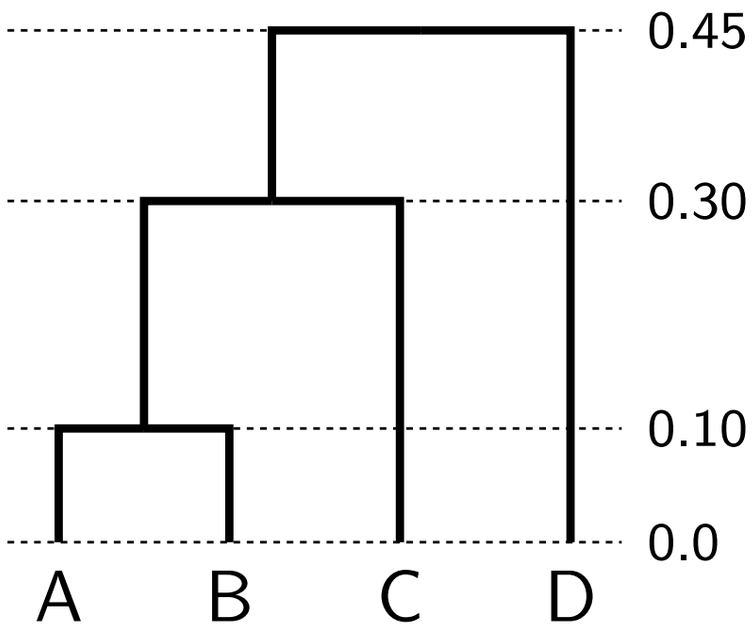
$=$ 0.01 $\times$

evolutionary rate
substitutions / site / unit time

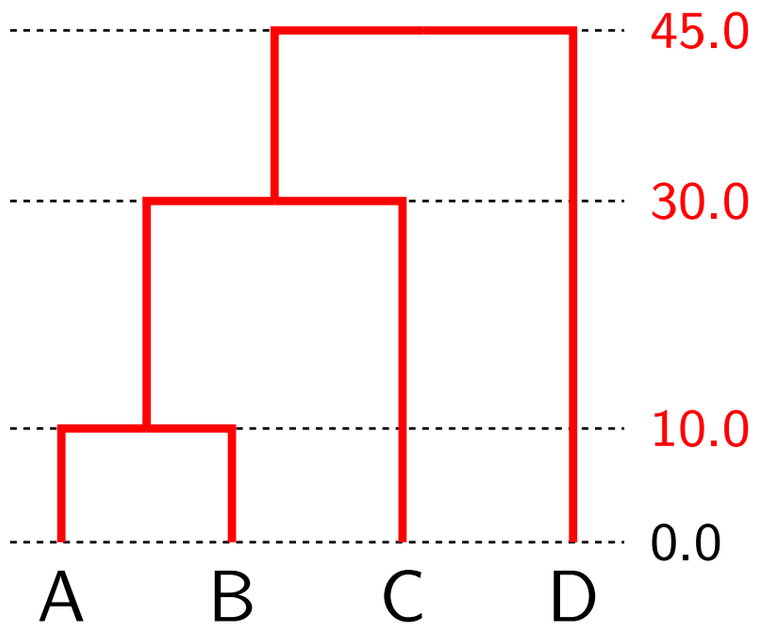


time tree

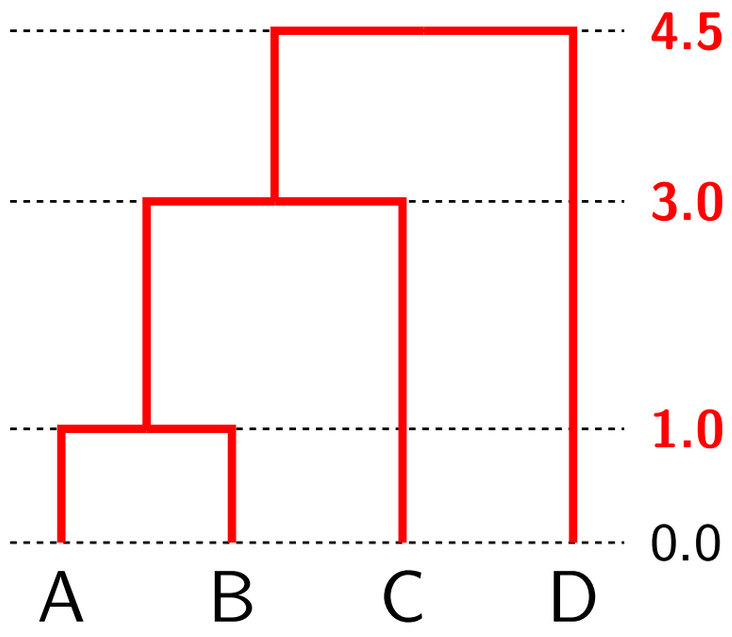
Non-identifiability of rates and times



$=$ 0.01 $\times$



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



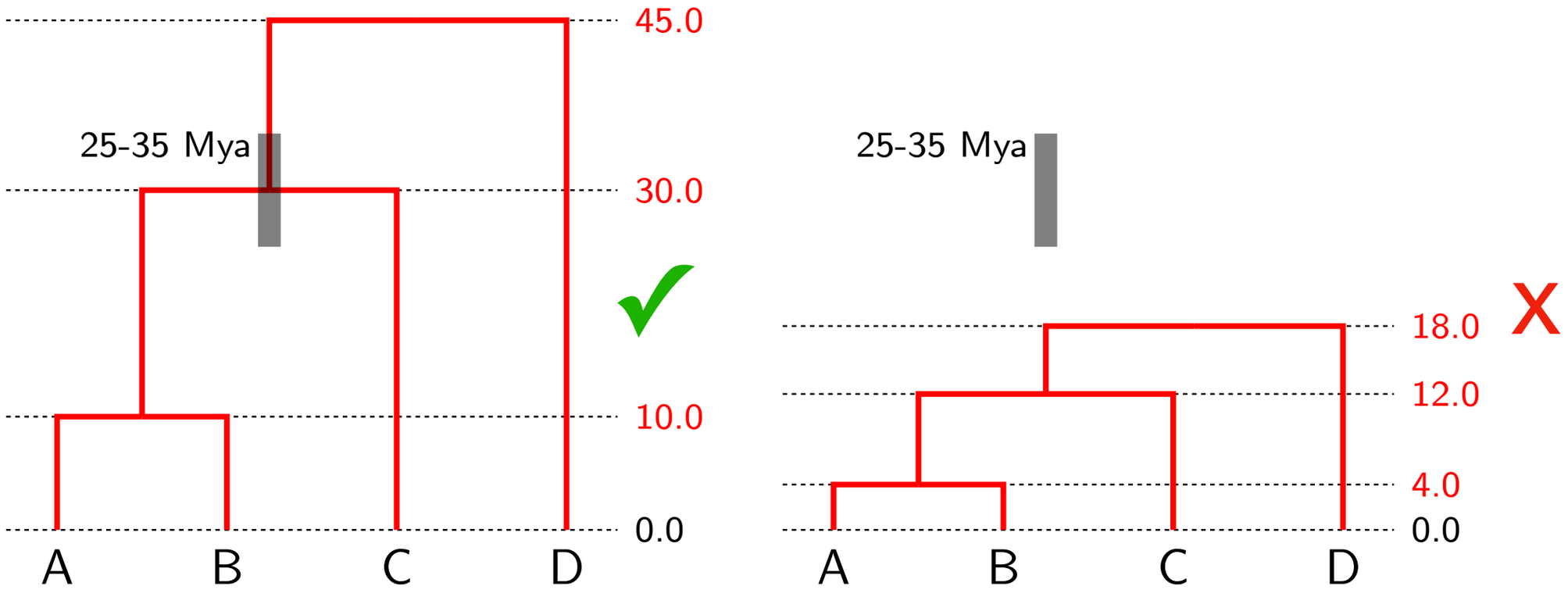
"substitution tree"

evolutionary rate
substitutions / site / unit time

time tree

Node calibration

Suppose fossil evidence shows the common ancestor of species A, B and C lived 25-35 Mya. The **left tree is consistent**, the **right tree is not consistent**.

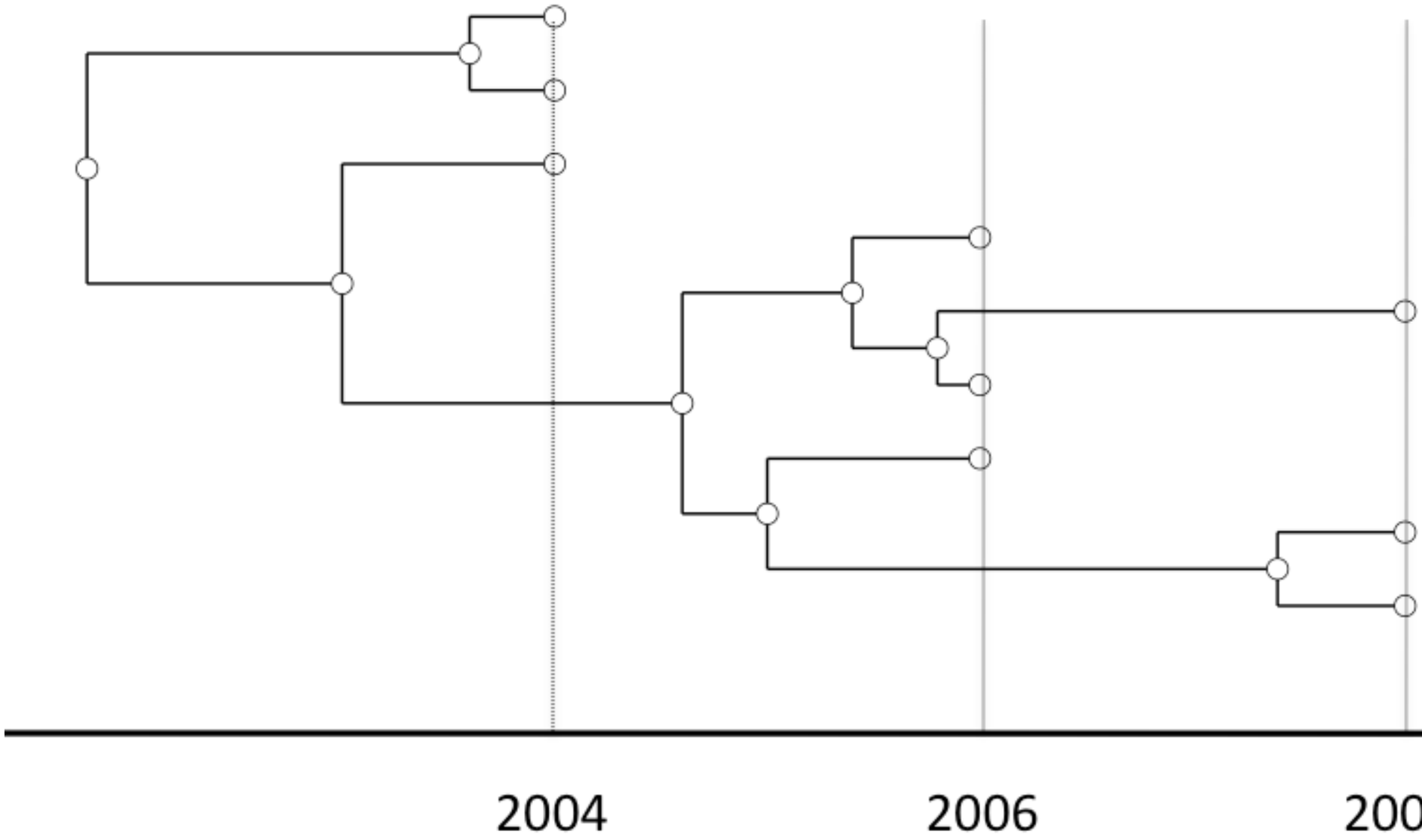


With a strict molecular clock, only the age (range) of a single node in the tree is needed in order to interpolate and extrapolate the ages of all other divergence times.

Once a known node age like this "calibrates" the tree, the genetic distances can be separated into an absolute rate and divergence times.

Bayesian evolutionary analysis of time-stamped data

Drummond *et al* (2002)

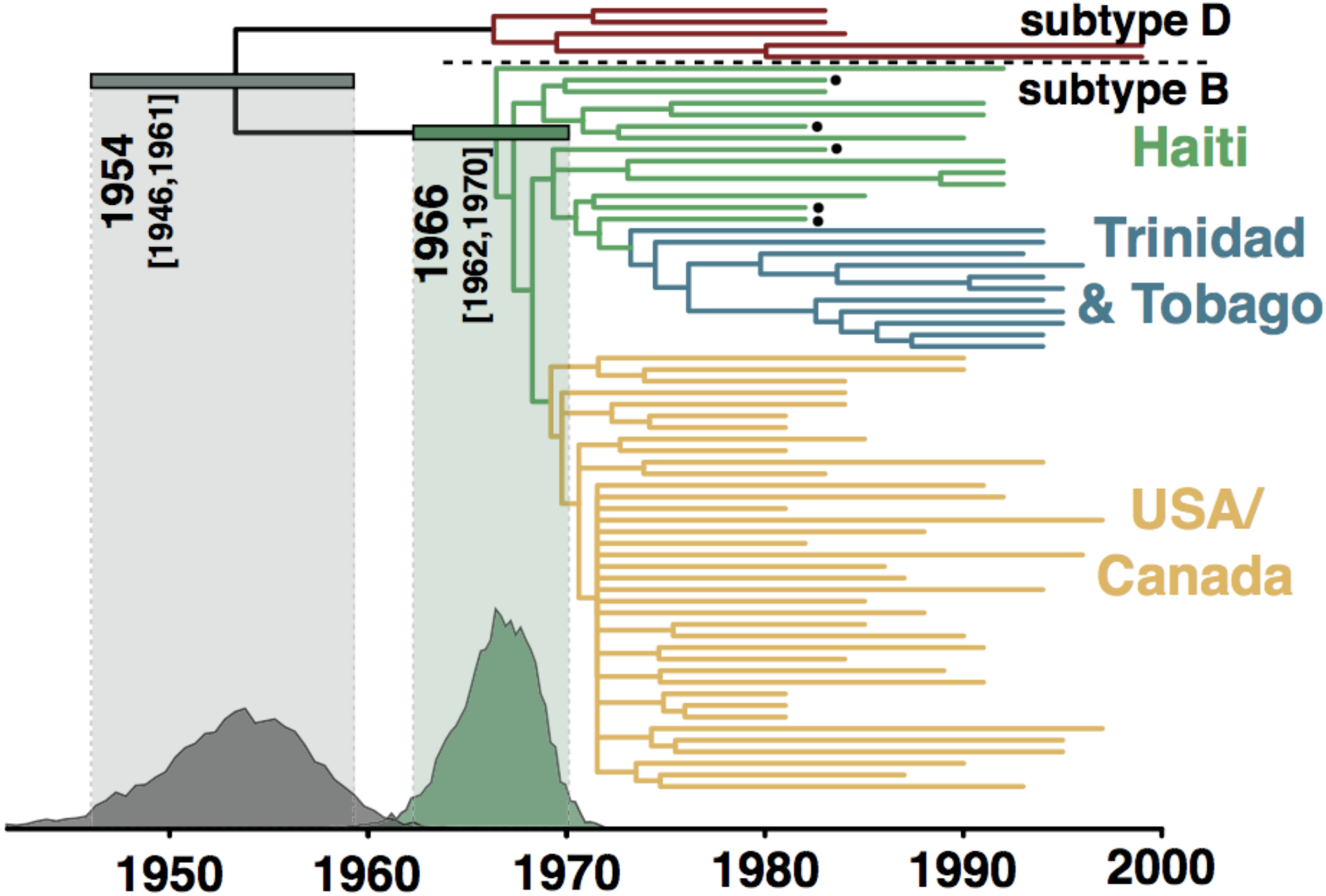


- Rapidly evolving microbes
- Ancient DNA
- Cancer
- Somatic evolution
- Languages
- *et cetera*

$$P(g, \mu, Q, \theta | D) \propto \Pr(D | g \times \mu, Q) P(g | \theta) P(\theta) P(Q) p(\mu)$$

A calibrated phylogenetic inference

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



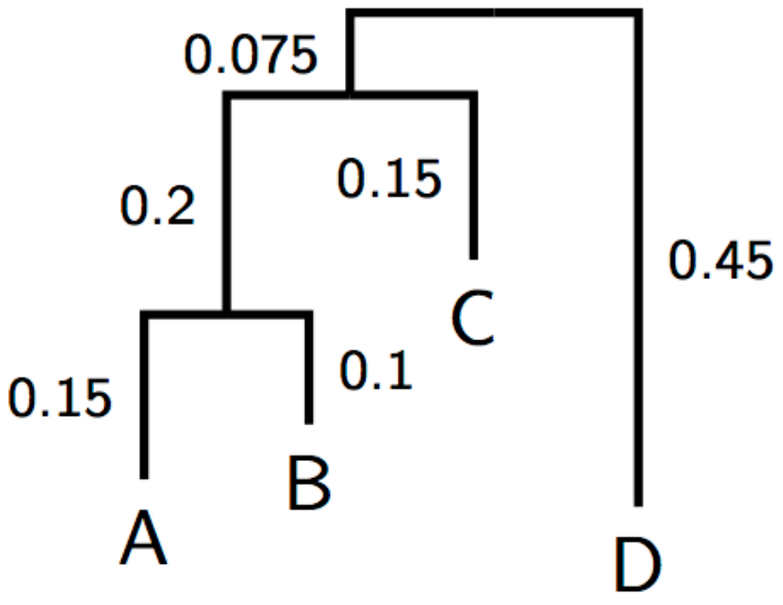
A phylogenetic reconstruction of samples of HIV-1 virus. Each tip represents a single infected individual from whom a blood sample has been taken.

Relaxed phylogenetics

Genetic distance = rate x time

Relaxed molecular clock

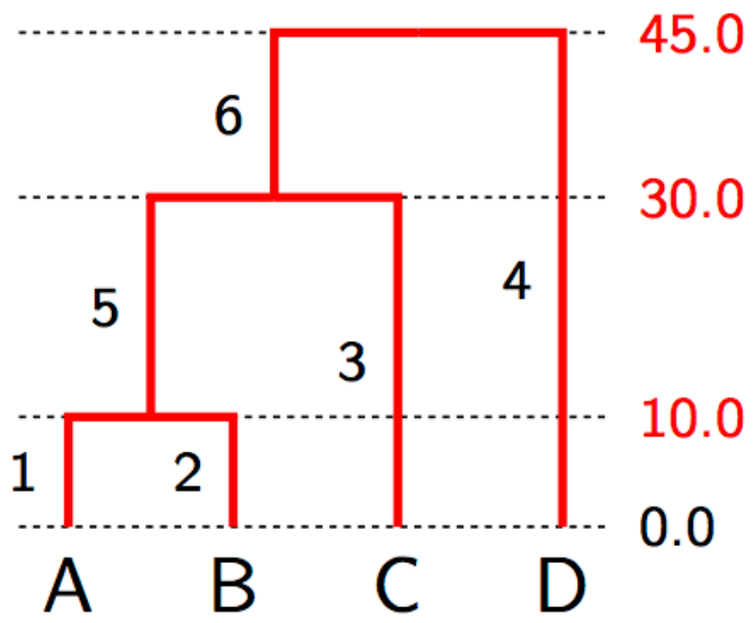
$$T = \vec{\mu} \star g$$



"substitution tree"

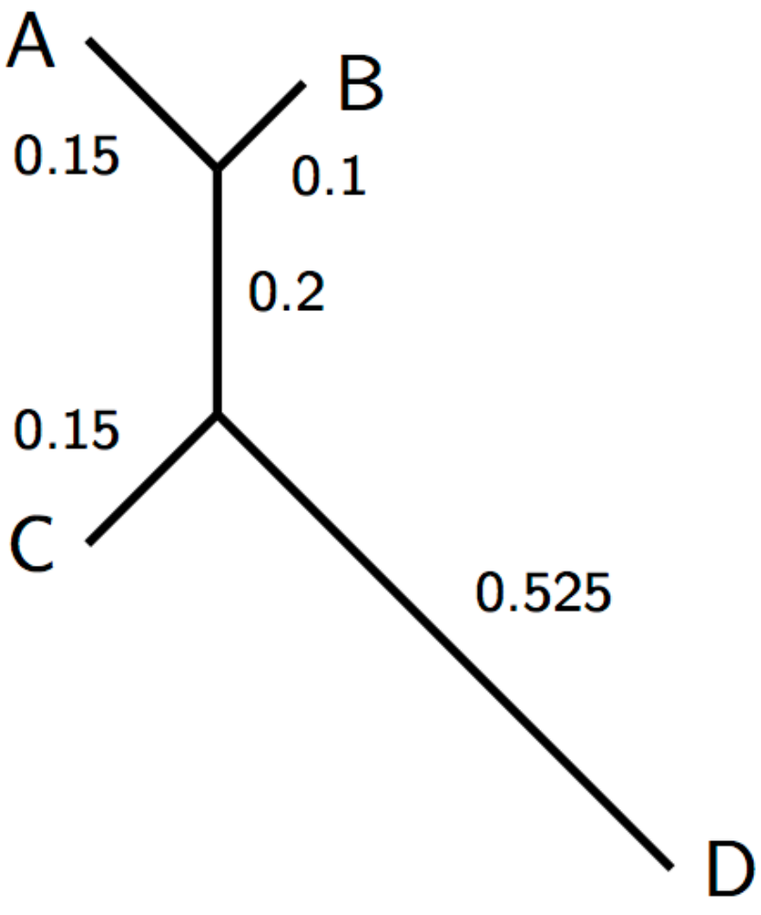
$$= \begin{pmatrix} 0.015 \\ 0.01 \\ 0.005 \\ 0.01 \\ 0.01 \\ 0.005 \end{pmatrix} \star$$

evolutionary rates
substitutions / site / unit time



time tree

Genetic distance = rate x time

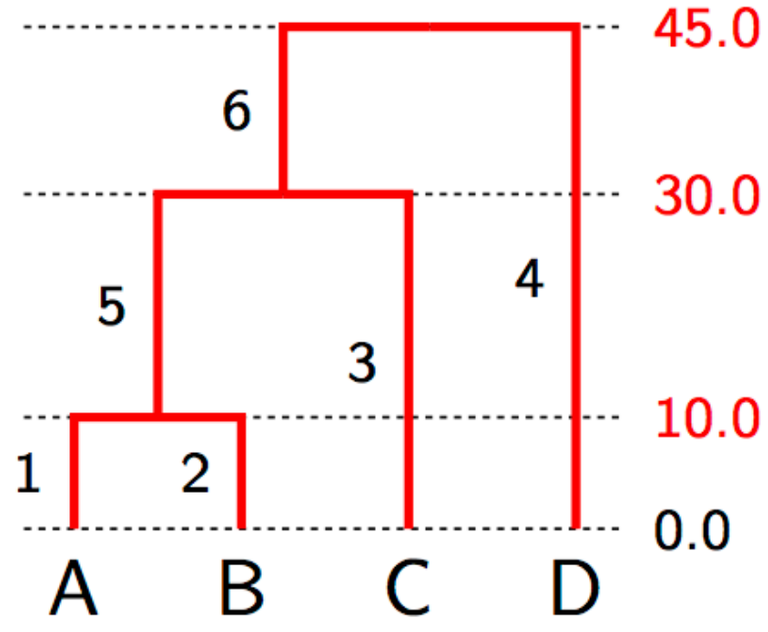


"substitution tree"

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

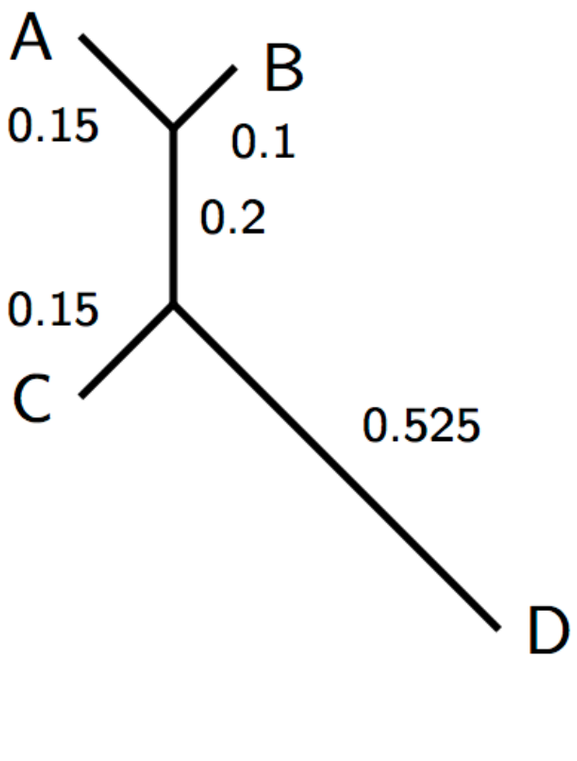
$$= \begin{pmatrix} 0.015 \\ 0.01 \\ 0.005 \\ 0.01 \\ 0.01 \\ 0.005 \end{pmatrix} \star$$

evolutionary rates
substitutions / site / unit time

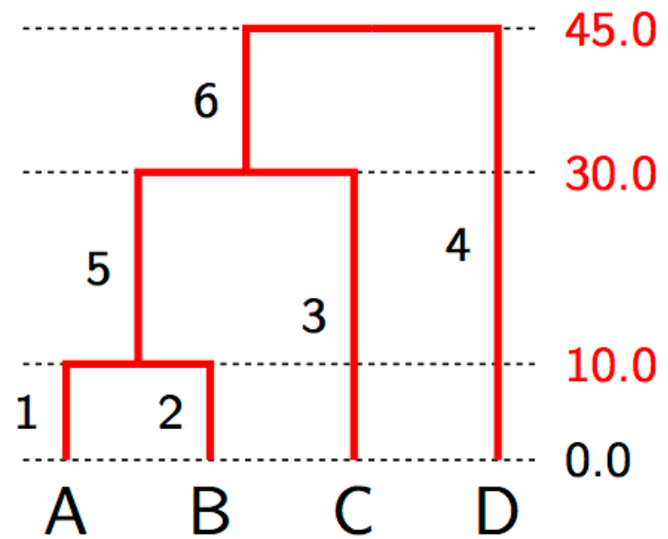


time tree

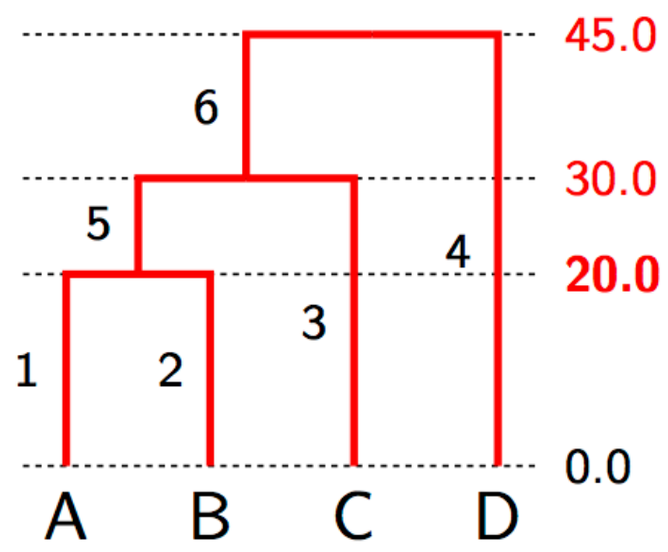
Non-identifiability of rates and times



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



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

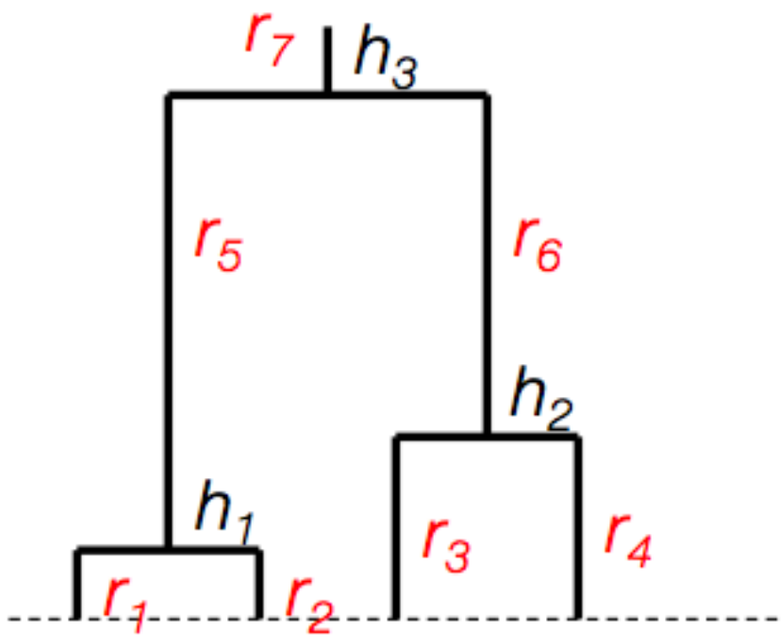


"substitution tree"

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

$r \sim Exp(\lambda)$

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

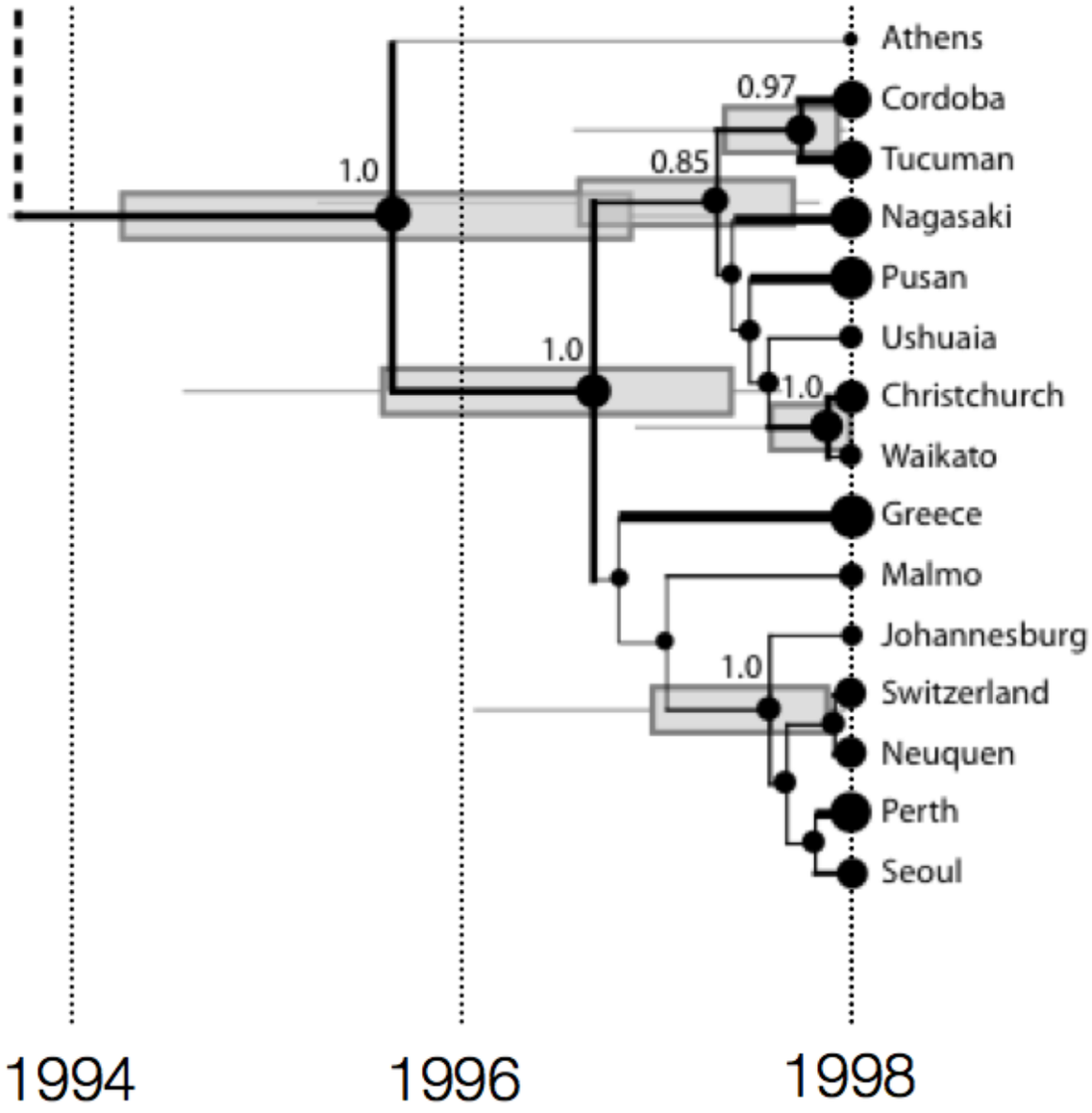
$r \sim Gamma(\alpha, \beta)$

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

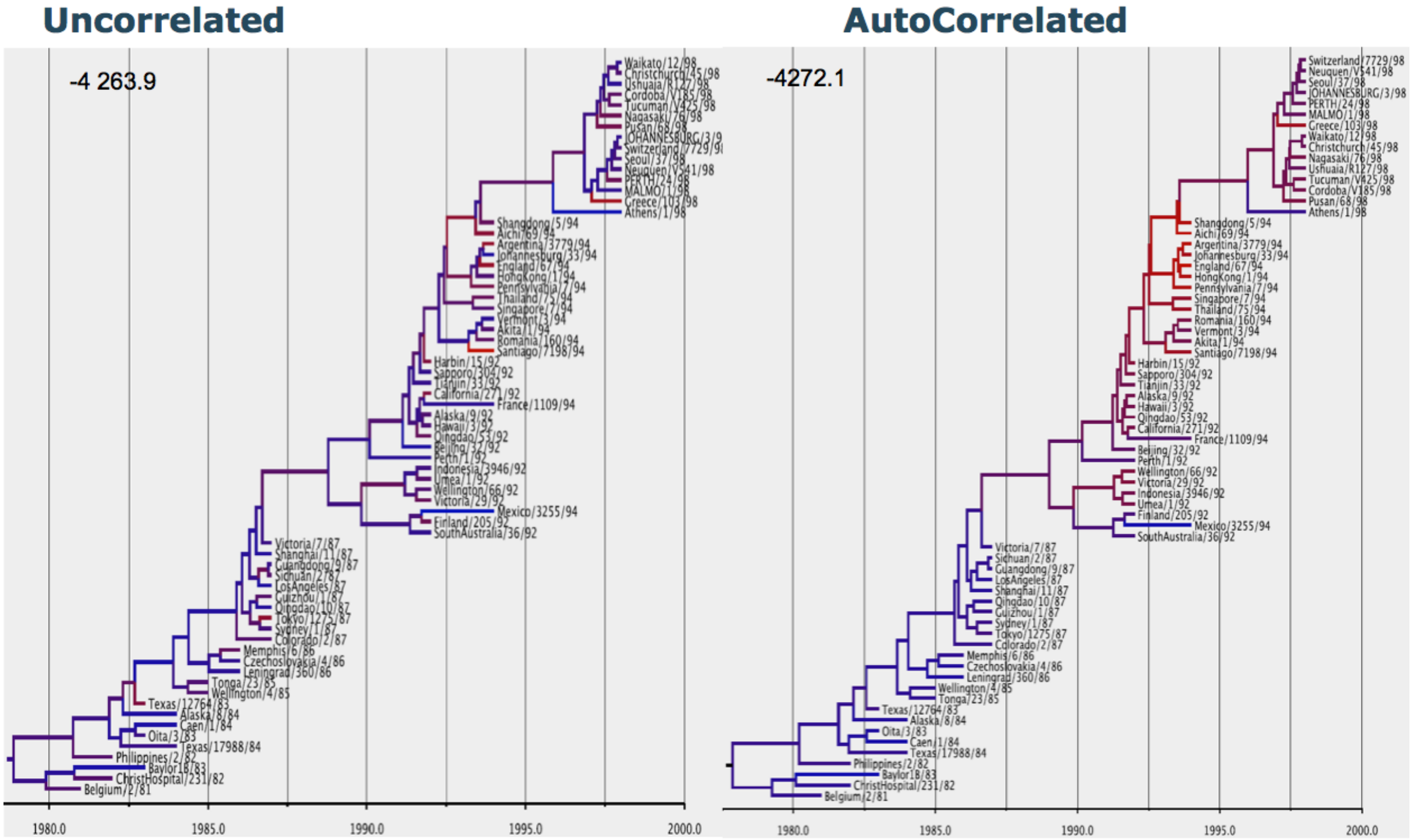
ML

Influenza A gene tree estimating using relaxed molecular clock

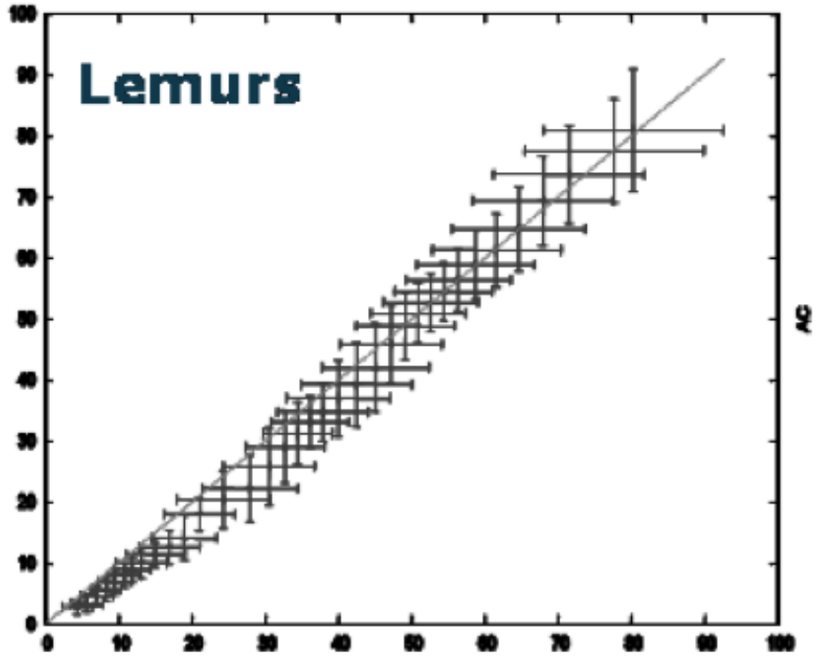


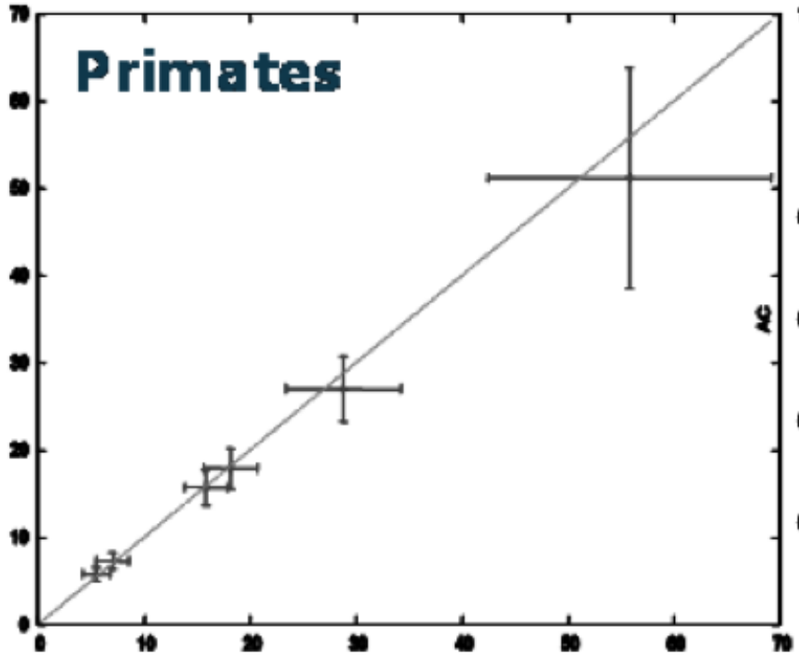
- Box-and-whisker plots show uncertainty in divergence times (only for splits with posterior probability > 0.5)
- Node size and branch thickness proportional to evolutionary rate.

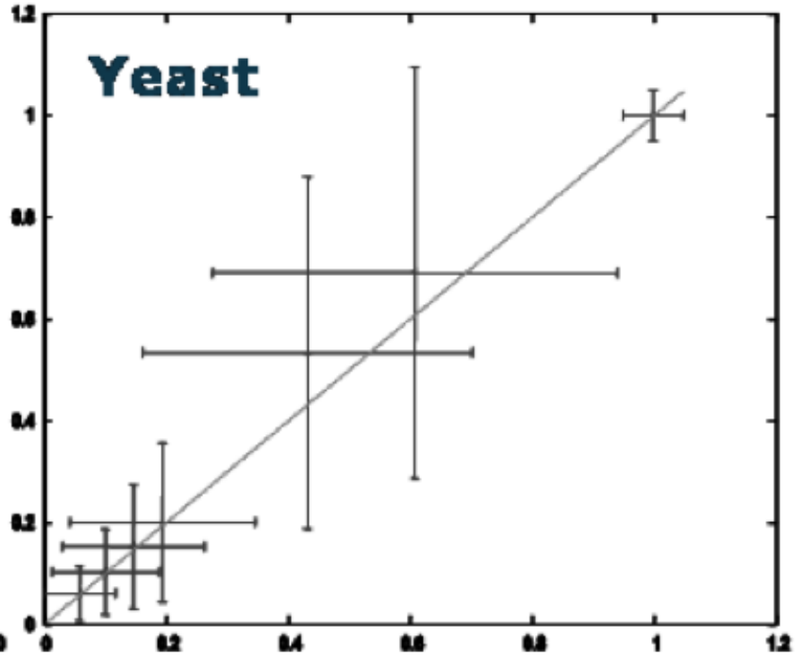
Influenza trees under different relaxed molecular clocks

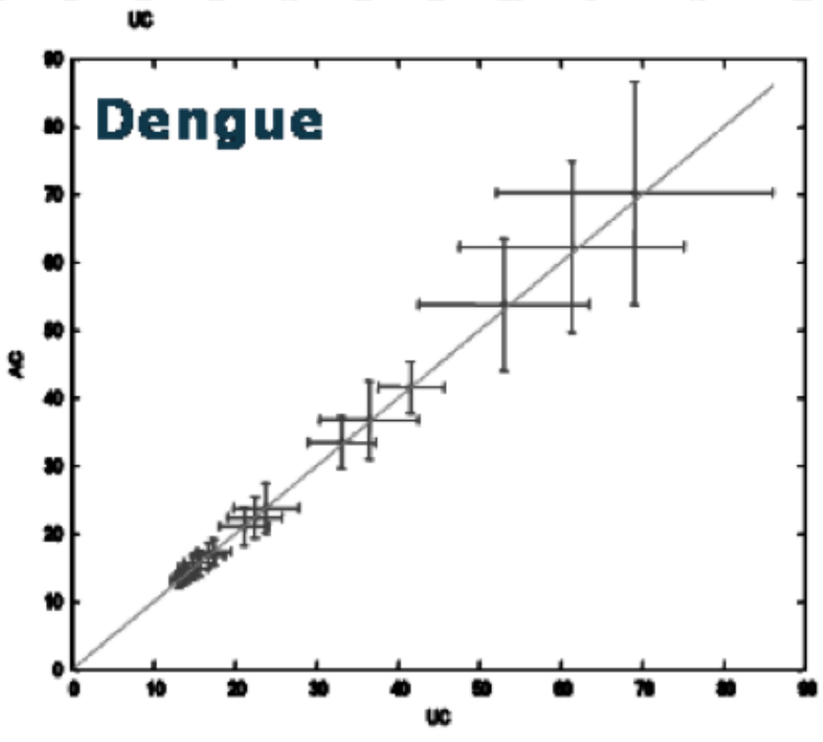


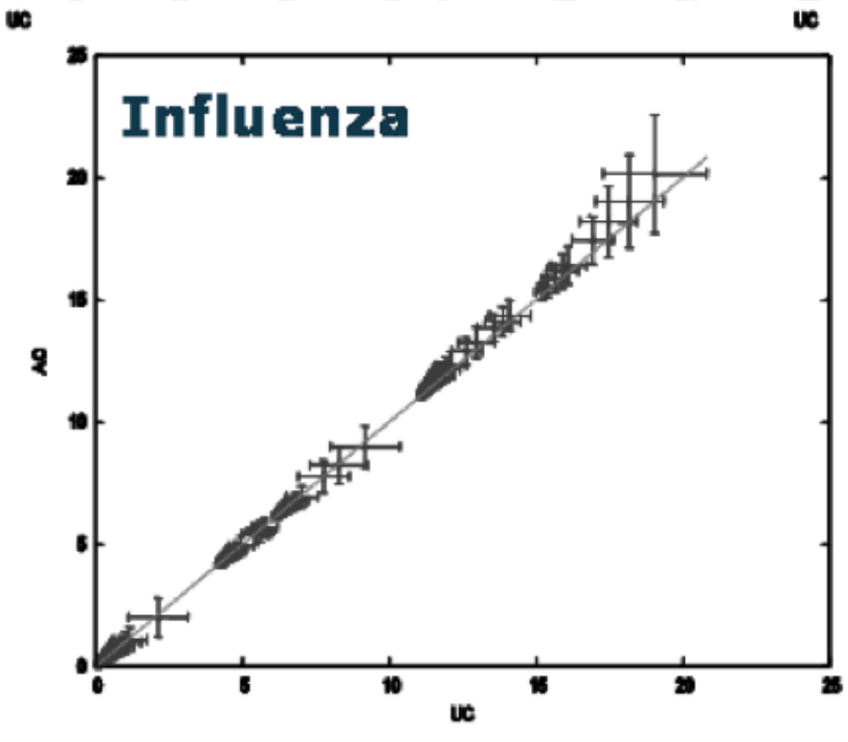
UC versus AC on five data sets



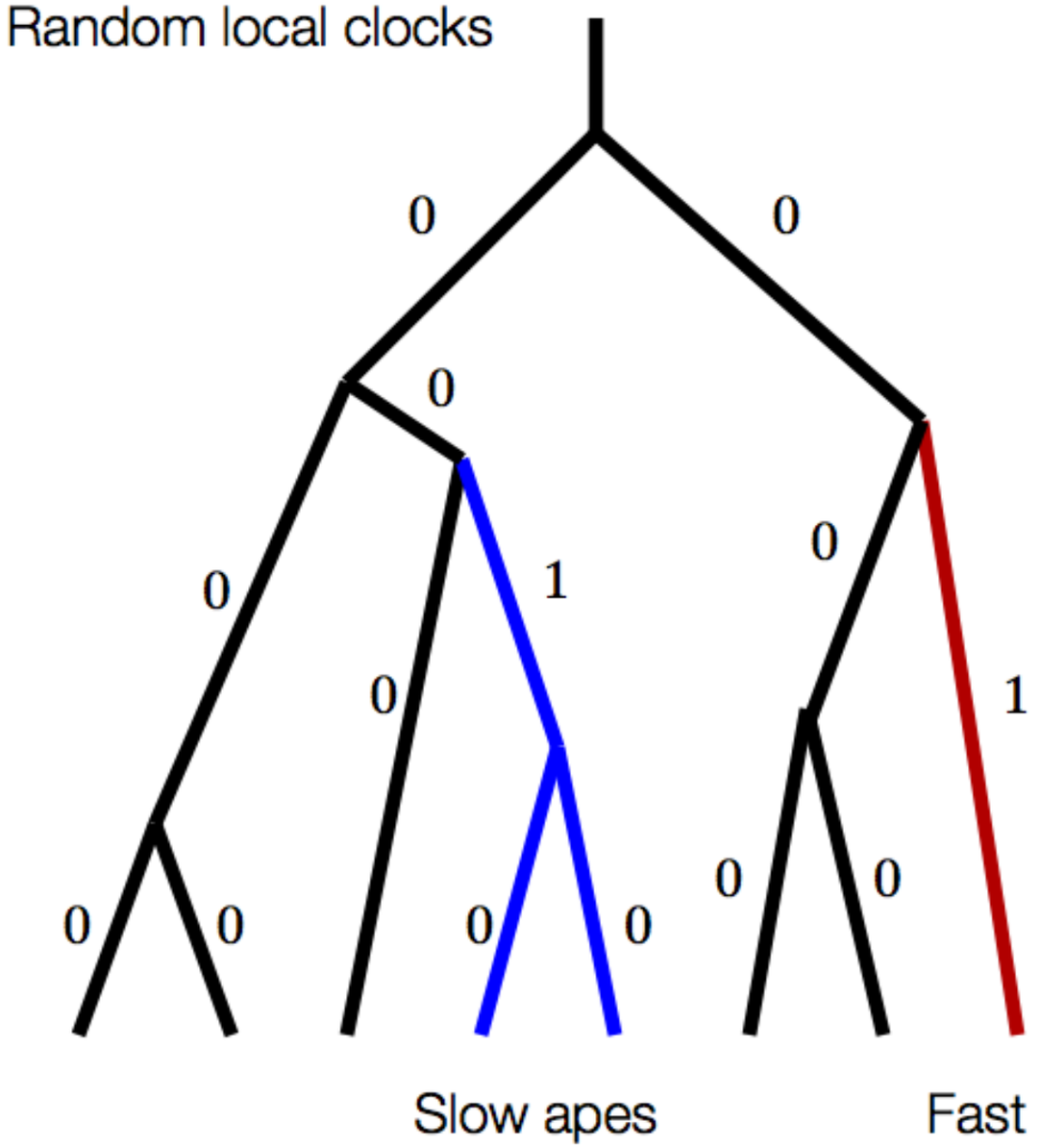






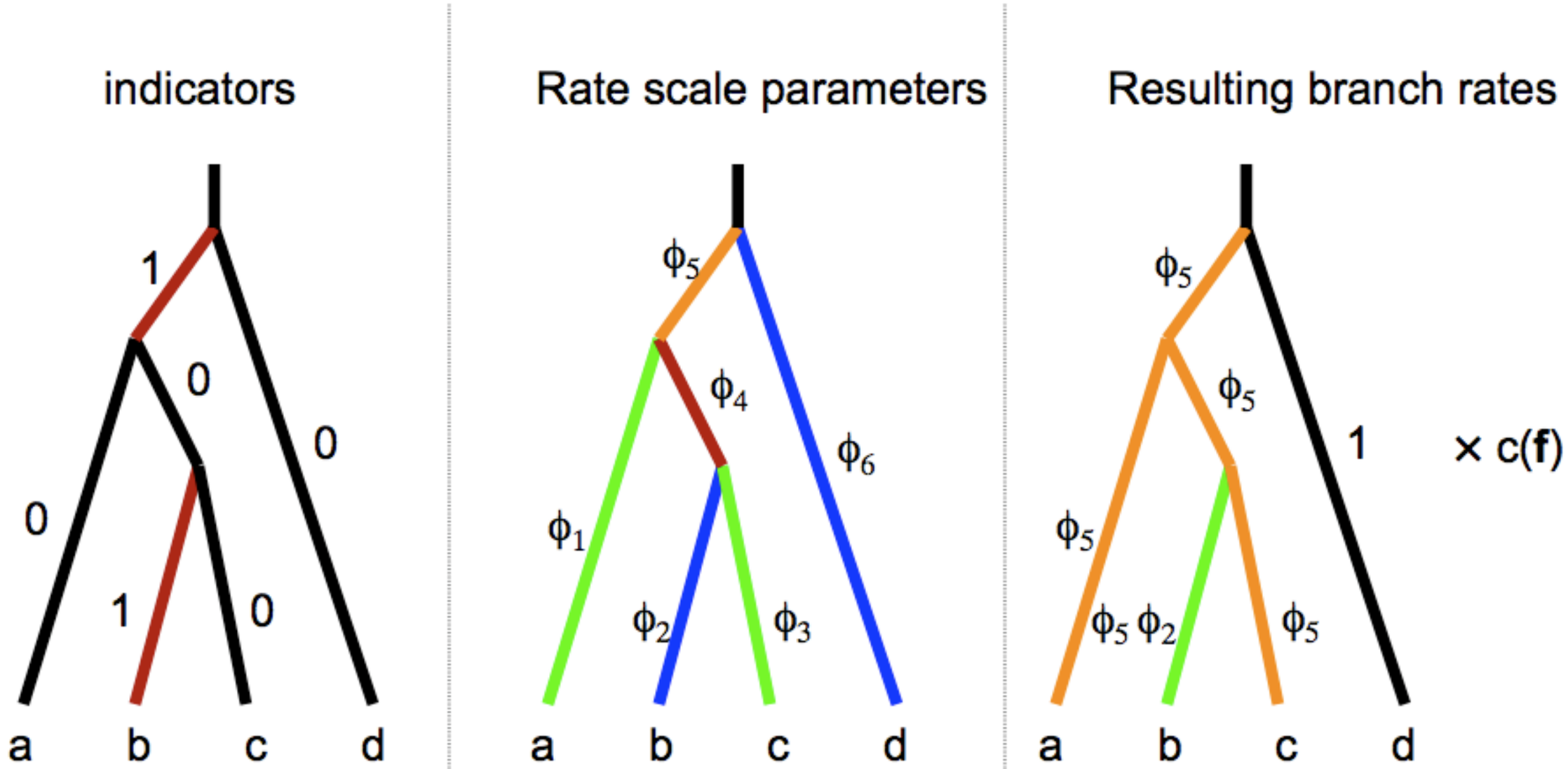


Random local molecular clocks



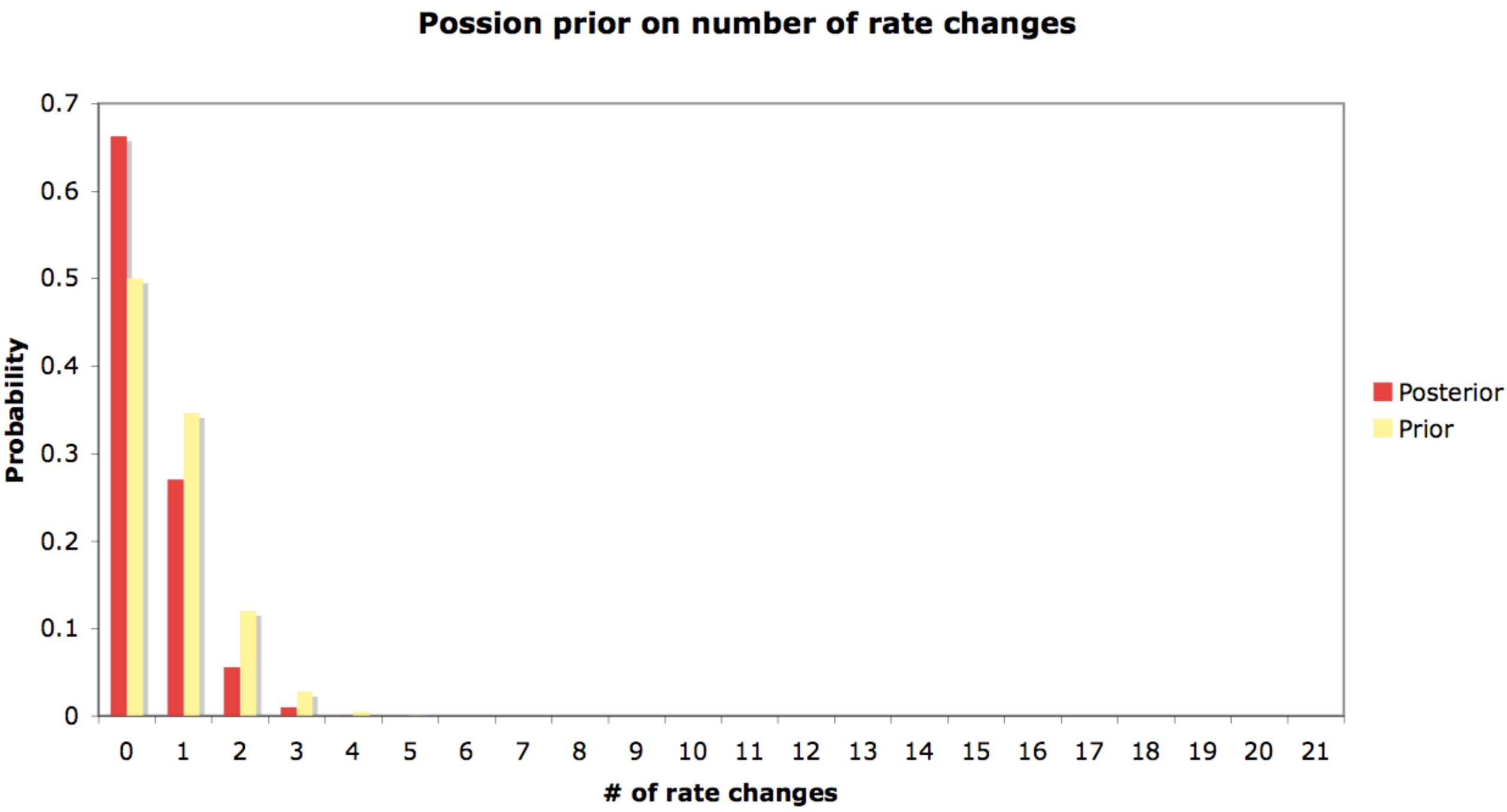
2^{2n-2} local molecular clock models

Random local molecular clocks

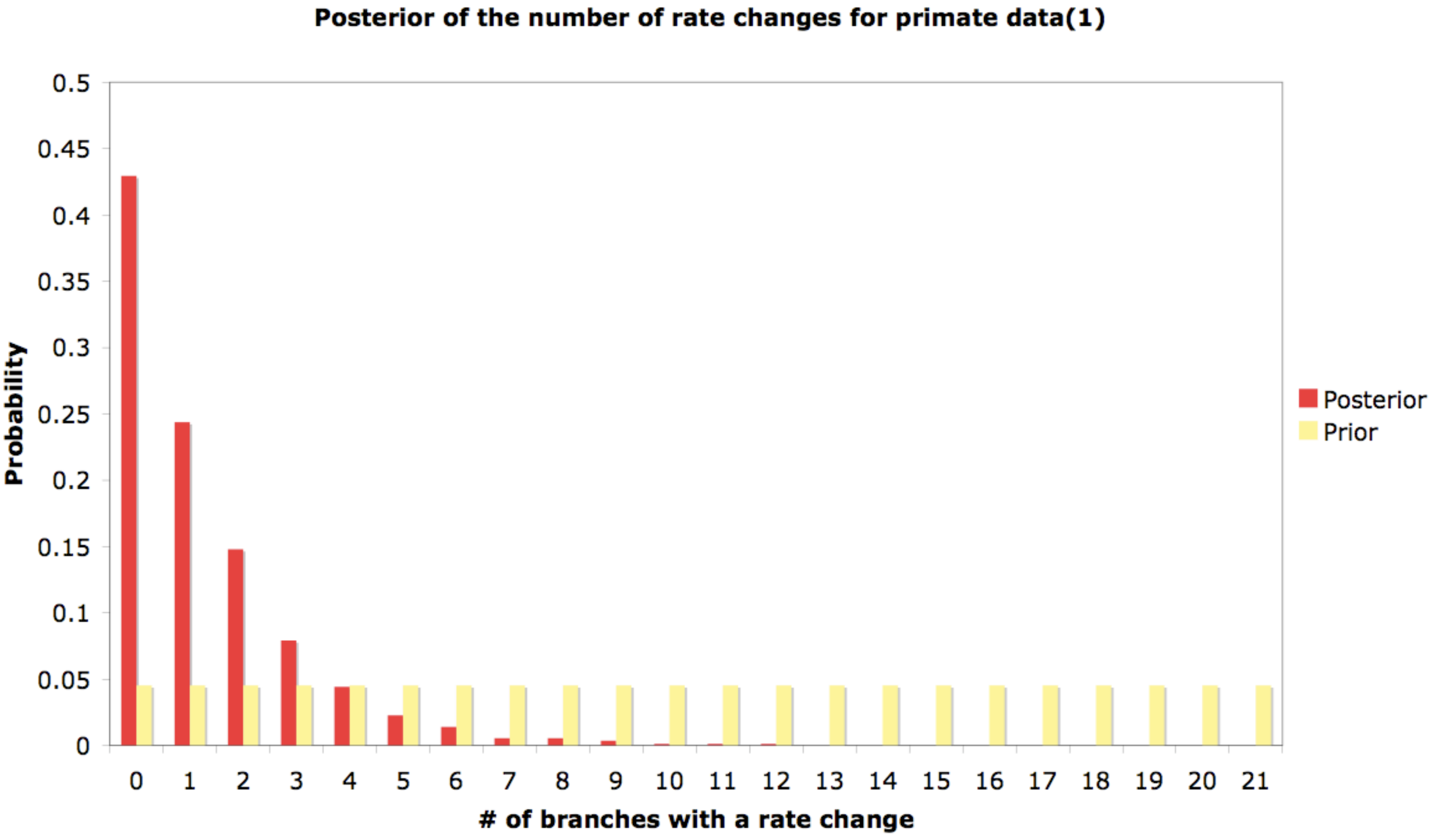


Red/Orange fast, Green/Blue slow

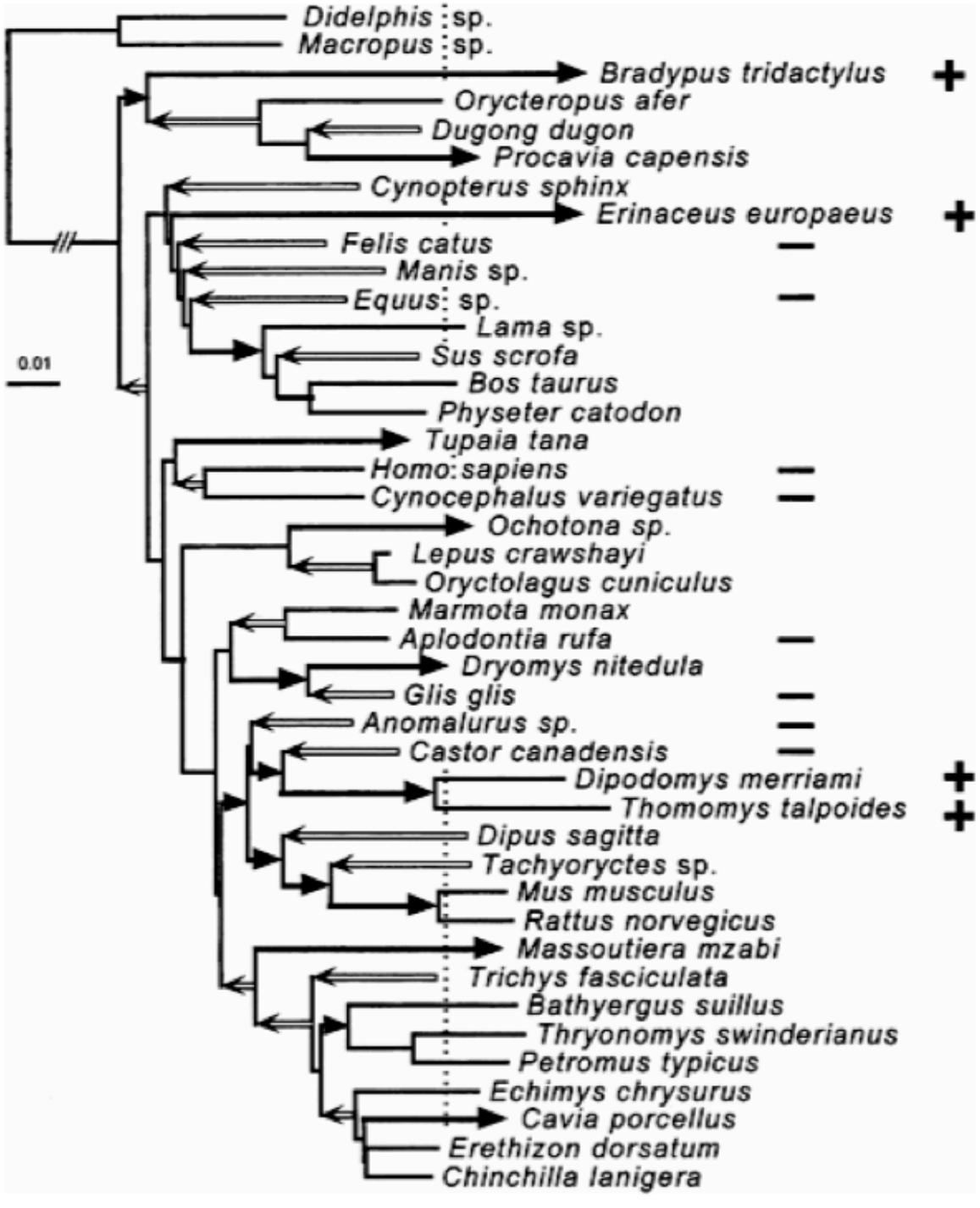
Primate data set (Poisson prior on # changes)



Primate data set (uniform prior on # changes)



Rodents (1+2 codon positions from 3 genes)

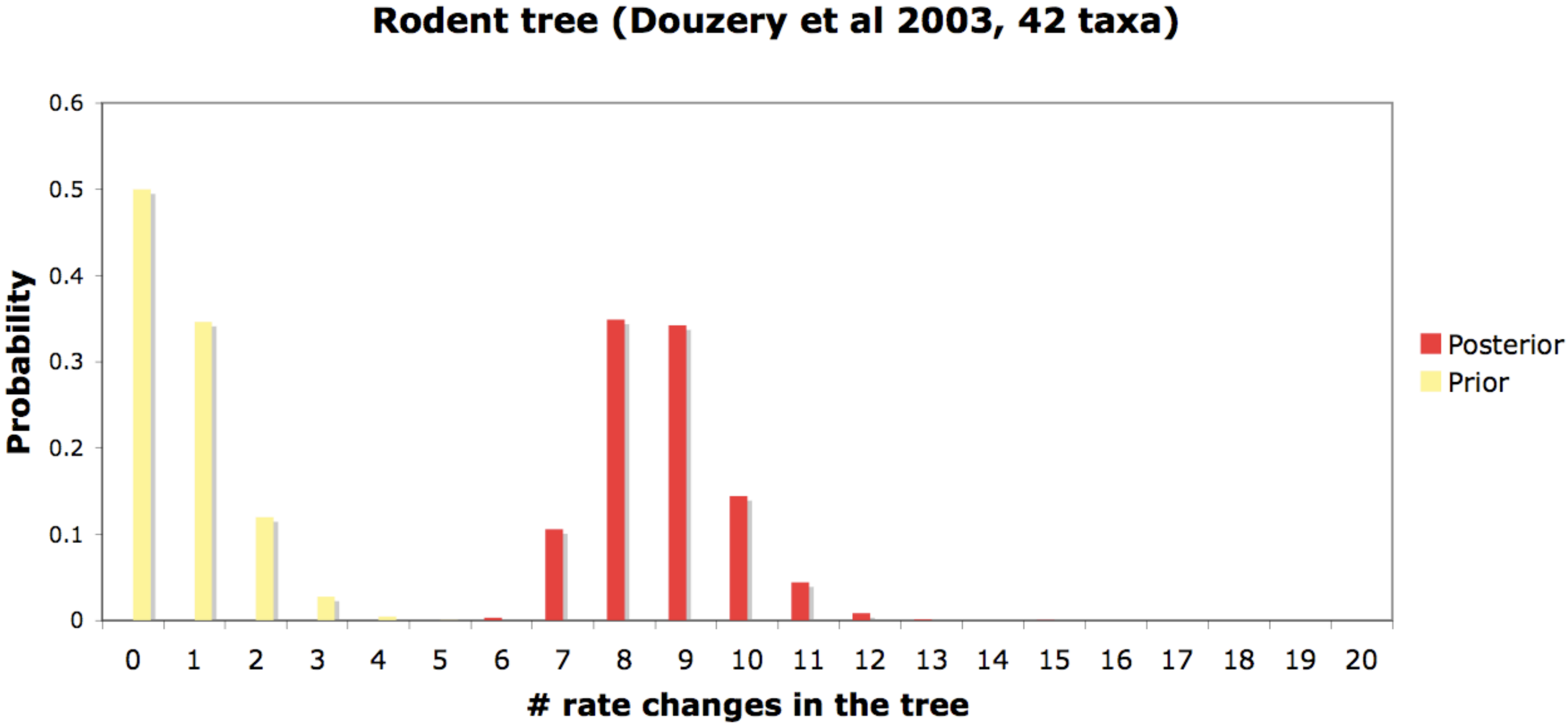


82 branches

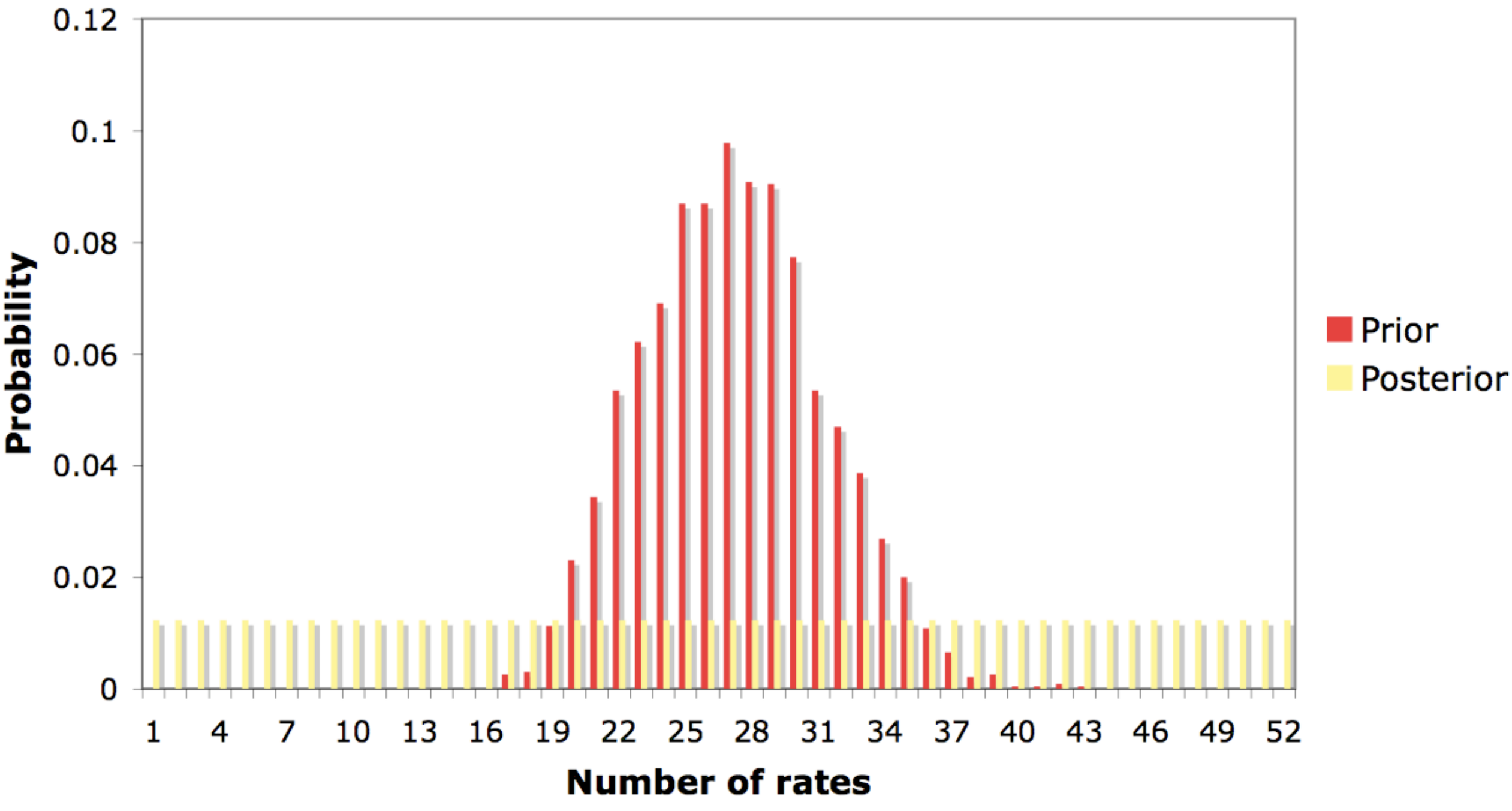
38 rate changes according to Douzery et al 2003

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).

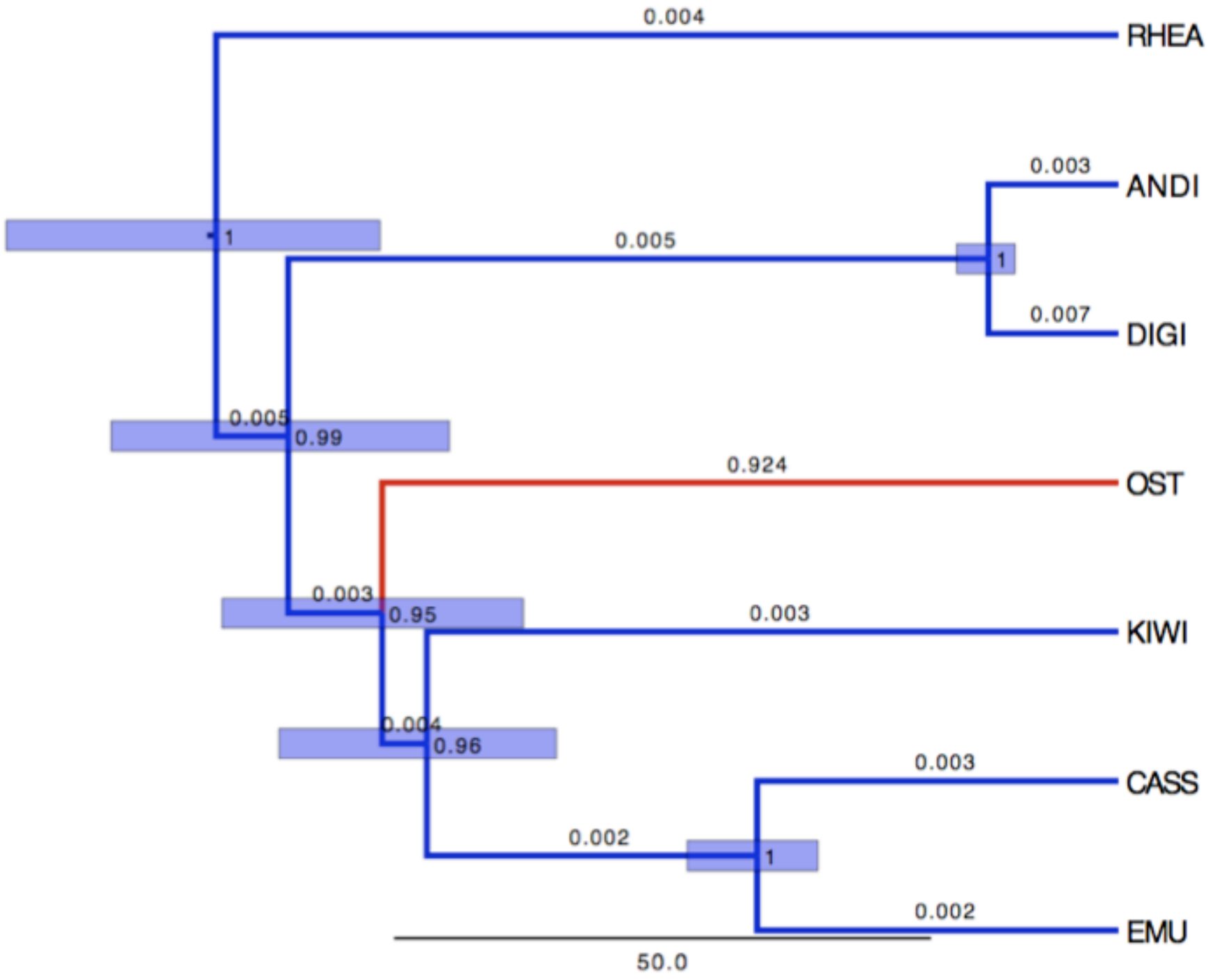
Rodent data set (Poisson prior on # changes)



Rodent data set (uniform prior on # changes)



Ratite relaxed clock on full mitochondrial sequences



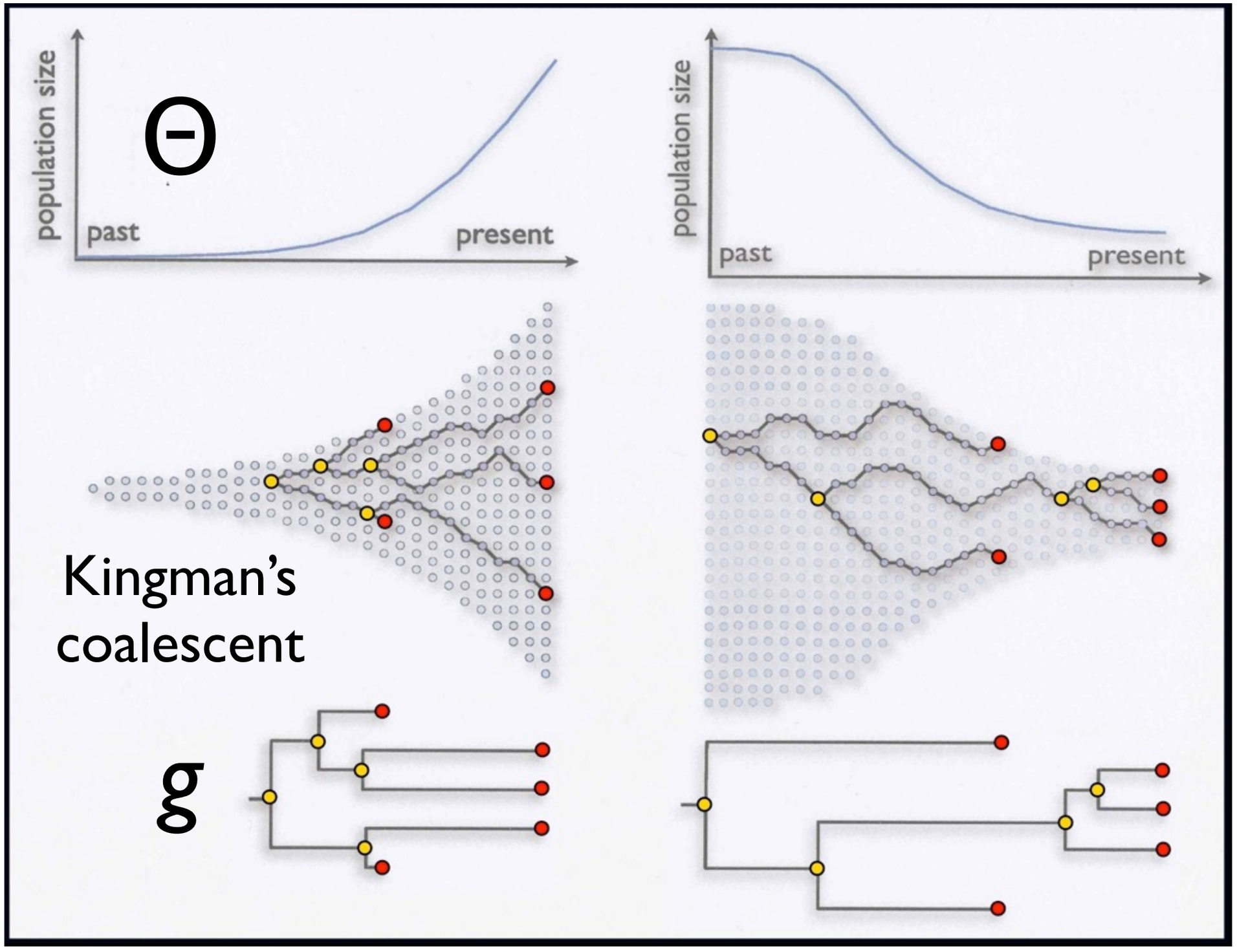
Coalescent tree priors

Questions:

What should we expect a tree to look like for different types of phylodynamics?

What does the shape of a tree tell us about the underlying phylodynamic process?

Coalescent theory: p(g|Θ)



$$P(g, \mu, Q, \theta \mid D) \propto \Pr(D \mid g \times \mu, Q) P(g \mid \theta) P(\theta) P(Q) p(\mu)$$

Bayesian skyline plots of influenza

Rambaut et al (2008)

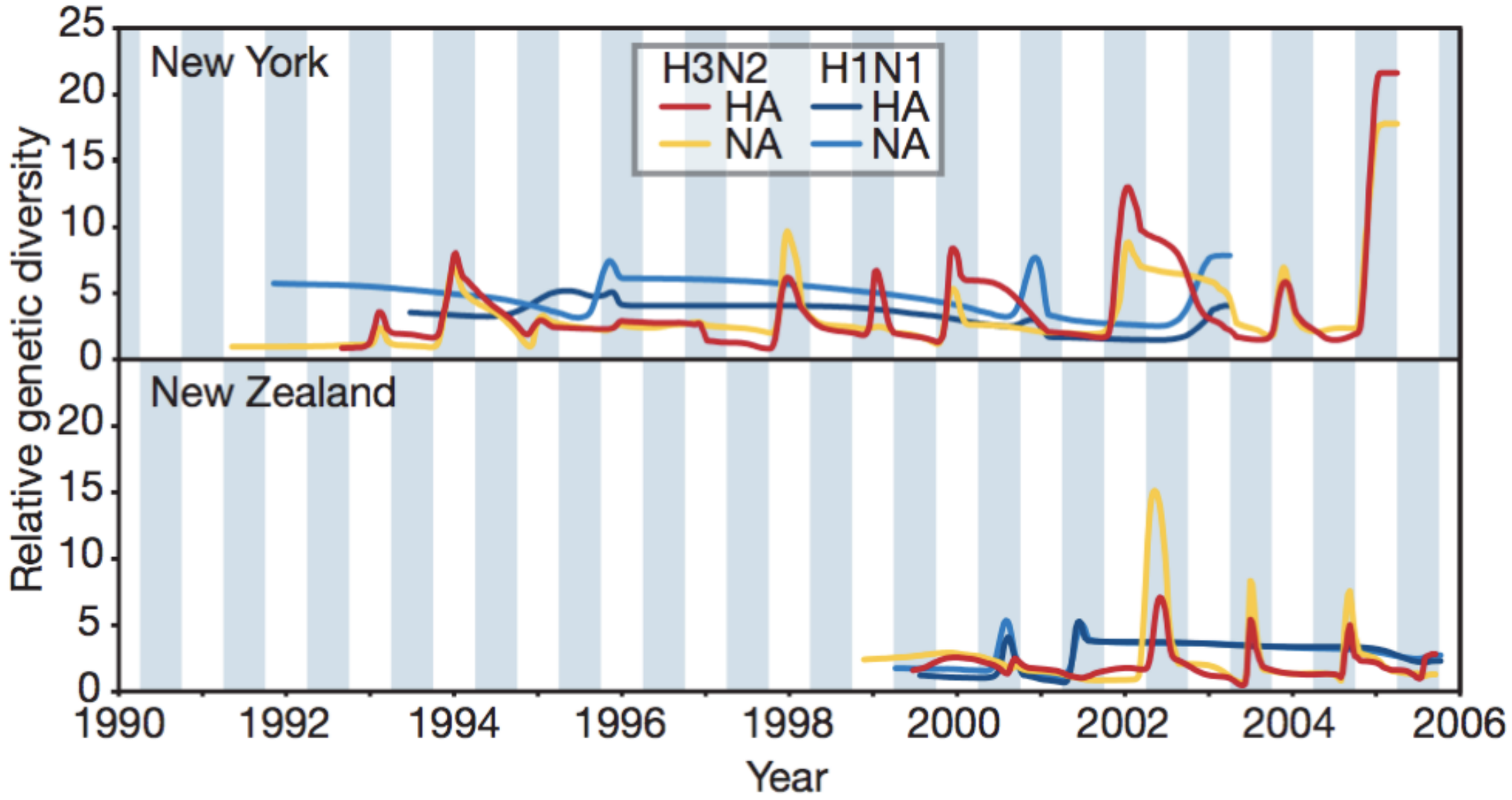
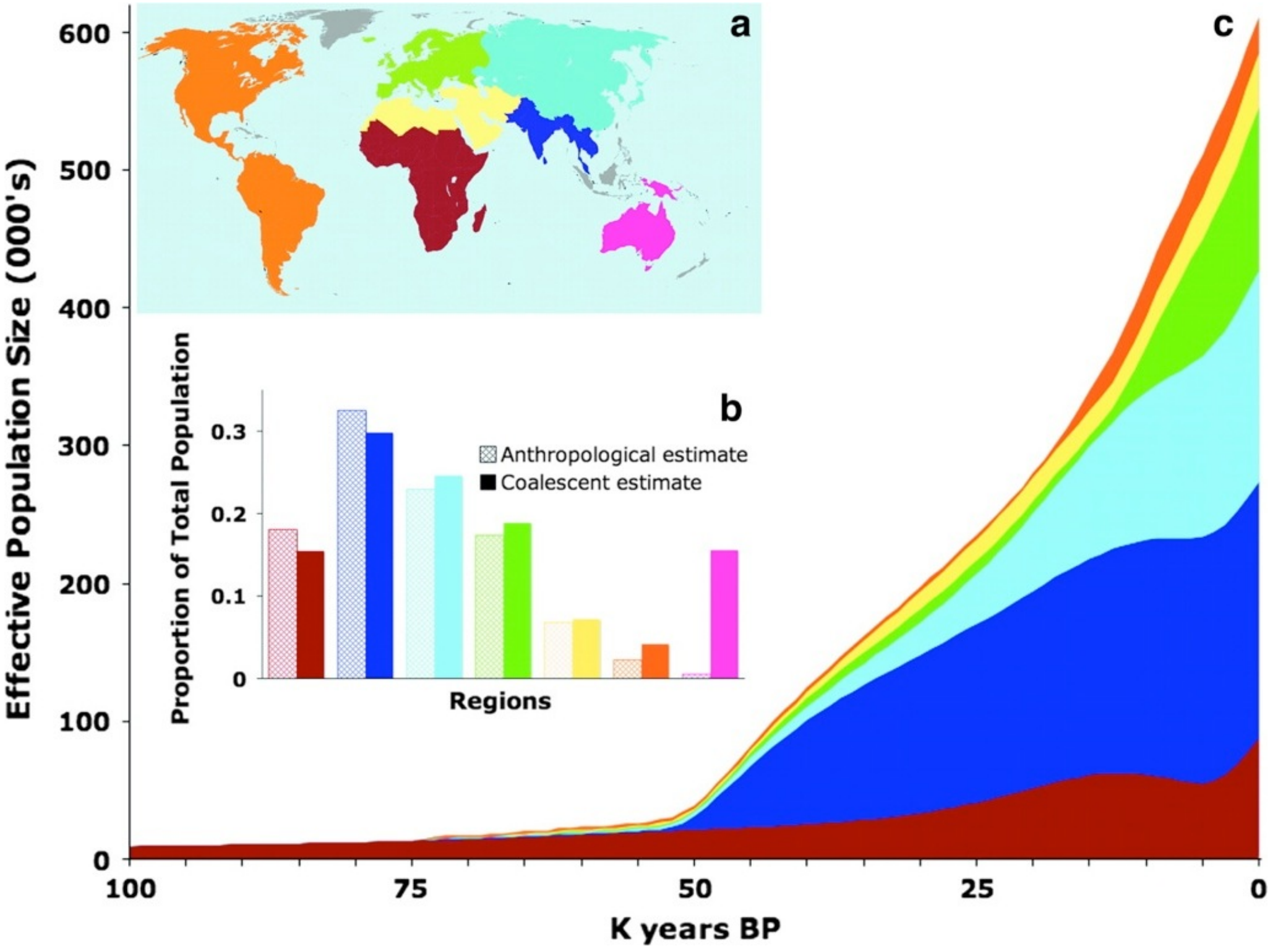


Figure 1 | Population dynamics of genetic diversity in influenza A virus. Bayesian skyline plots of the HA and NA segments for the A/H3N2 and A/H1N1 subtypes in New York state (top) and New Zealand (bottom). The horizontal shaded blocks represent the winter seasons. The *y*-axes represent a measure of relative genetic diversity (see Methods for details). The shorter timescale of New Zealand skyline plot is due to the shorter sampling period.

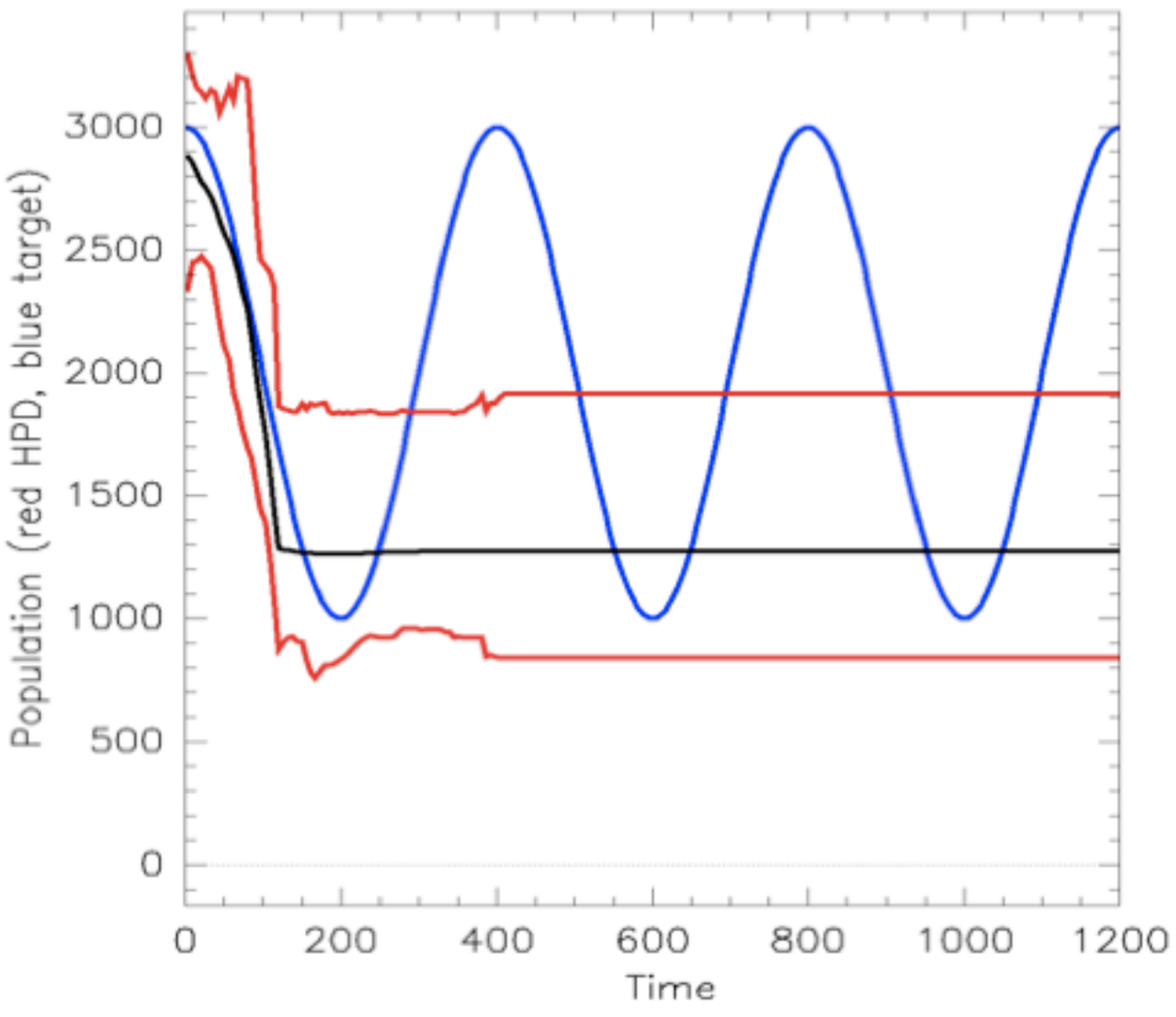
Mitochondrial DNA Variation Predicts Population Size in Humans and Reveals a Major Southern Asian Chapter in Human Prehistory



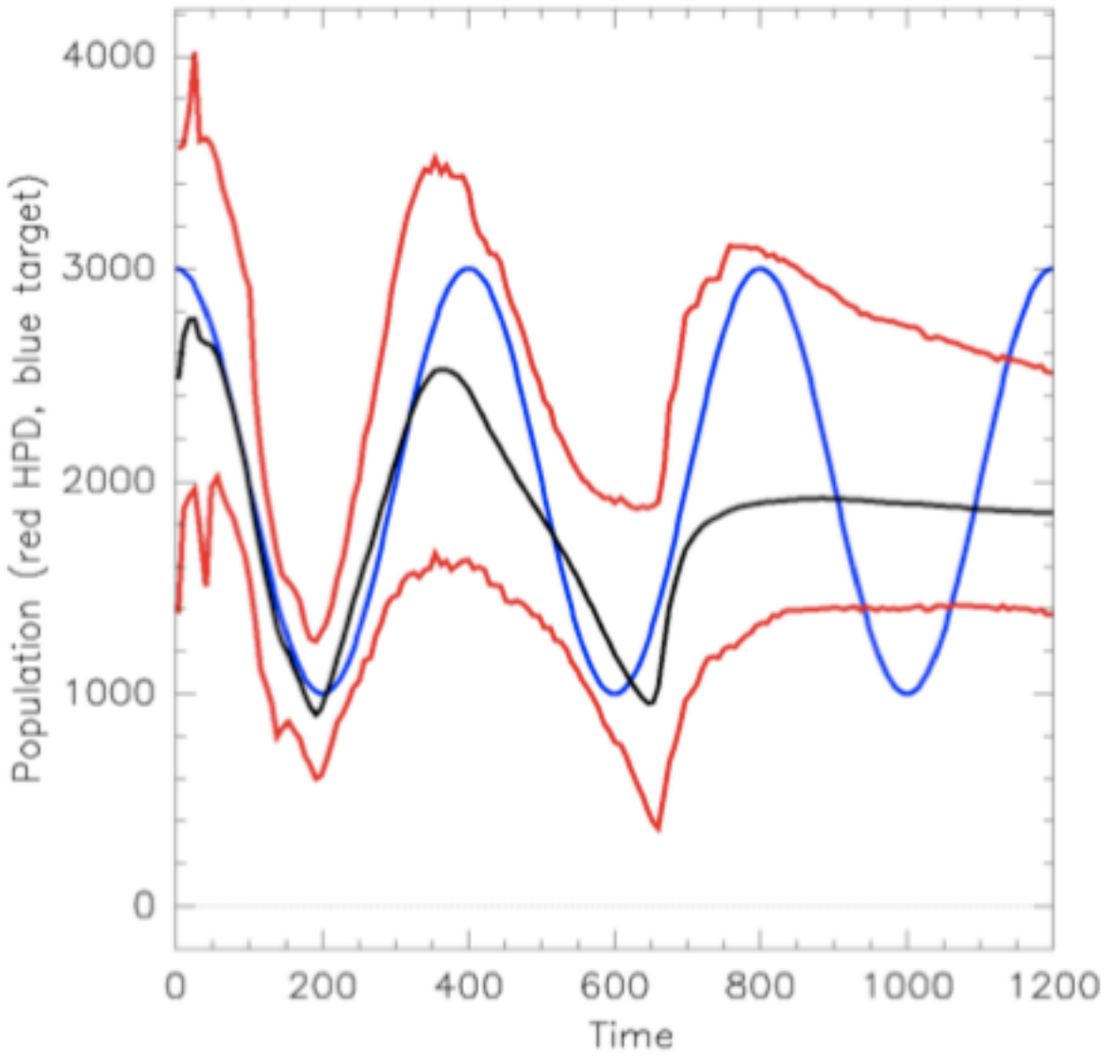
Atkinson, Gray and Drummond, 2008, Molecular Biology and Evolution

(Extended) Bayesian skyline plot

Drummond et al (2005); Heled & Drummond (2008)



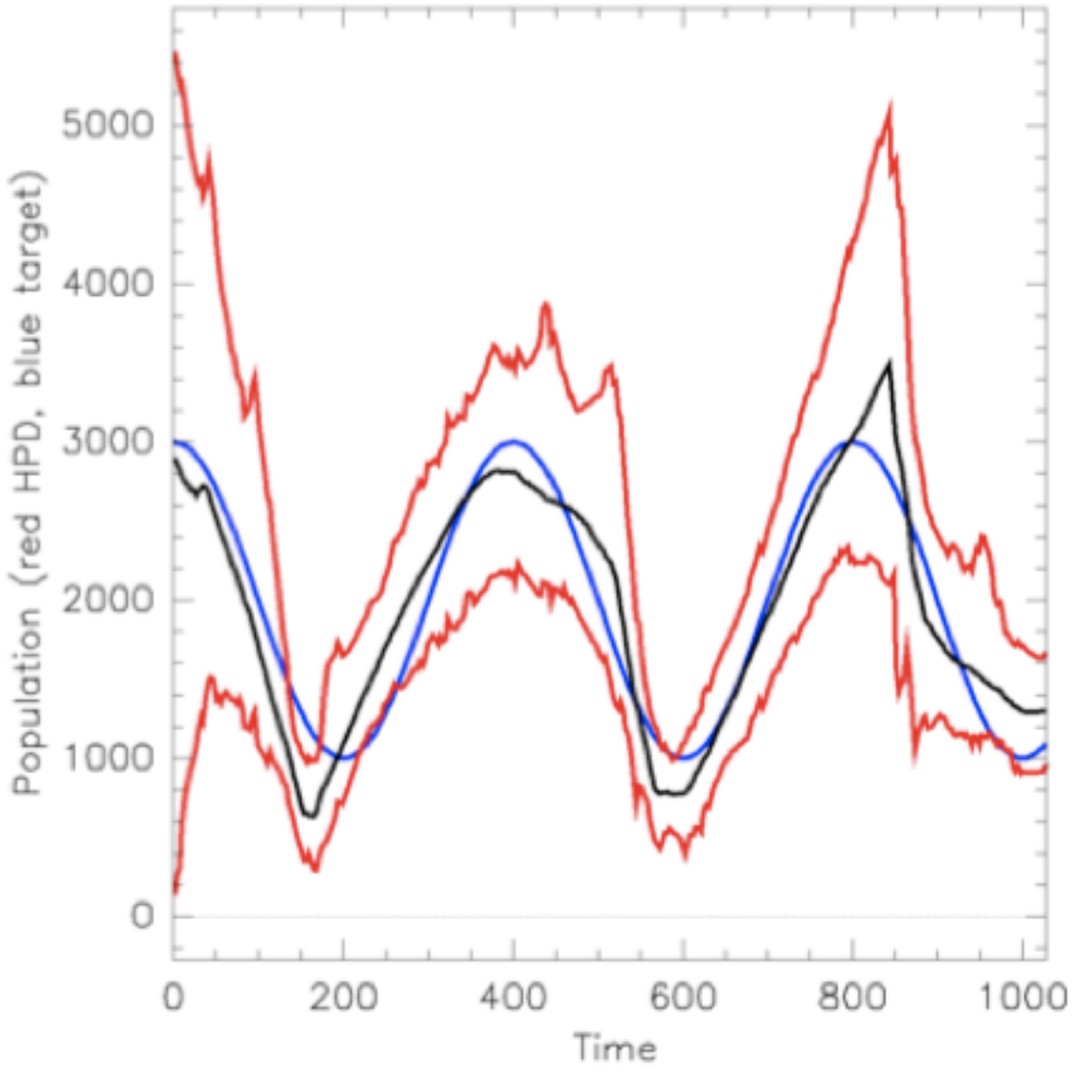
one gene sampled from 480 sampled individuals (480 gene sequences in total)



32 genes sampled from each of 16 sampled individuals (480 gene sequences in total)

(Extended) Bayesian skyline plot

Drummond et al (2005); Heled & Drummond (2008)



one gene from 480 individuals sampled through time (480 gene sequences in total)

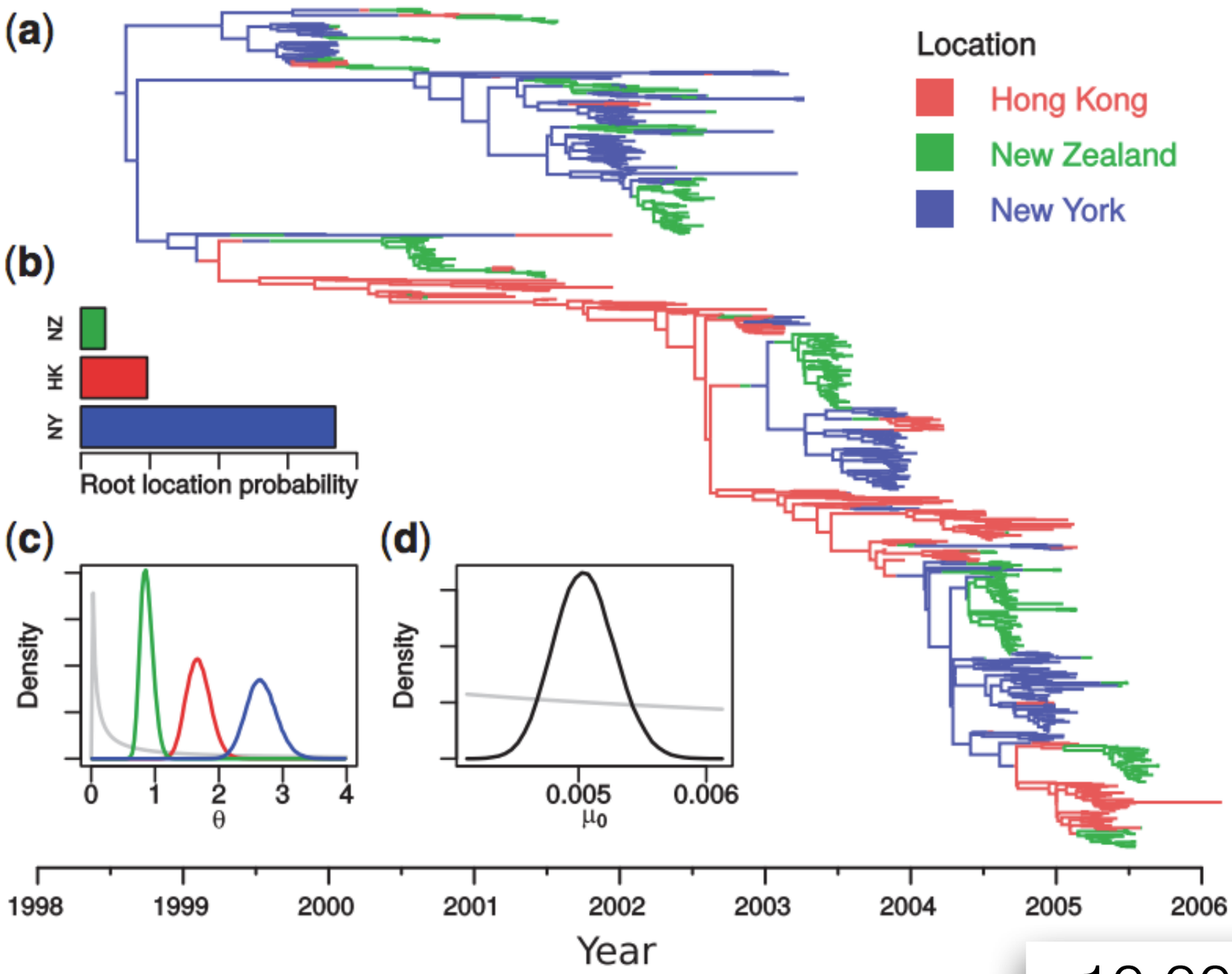
BIOINFORMATICS ORIGINAL PAPER 2014, pages 1–8
doi:10.1093/bioinformatics/btu201

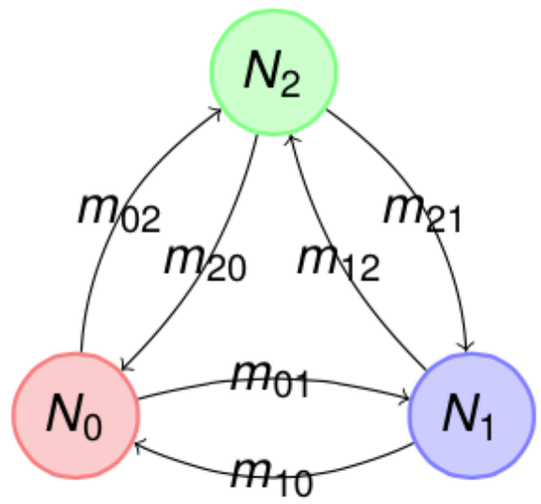
Phylogenetics Advance Access publication April 20, 2014

Efficient Bayesian inference under the structured coalescent

Timothy G. Vaughan[1,*], Denise Kühnert[1,2,3], Alex Popinga[1,3], David Welch[1,3] and Alexei J. Drummond[1,3]

Evolution provides a record of influenza's global dynamics

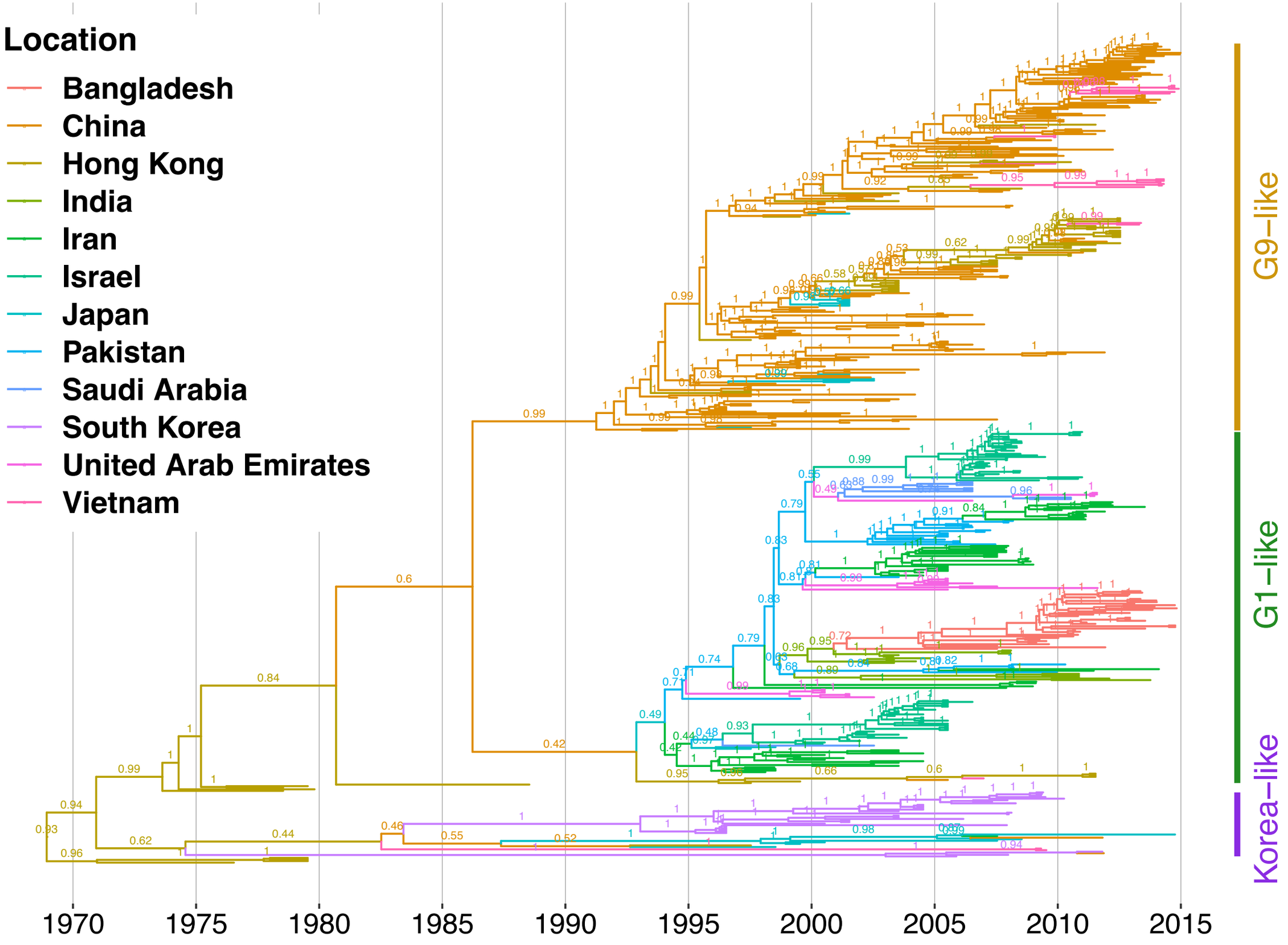




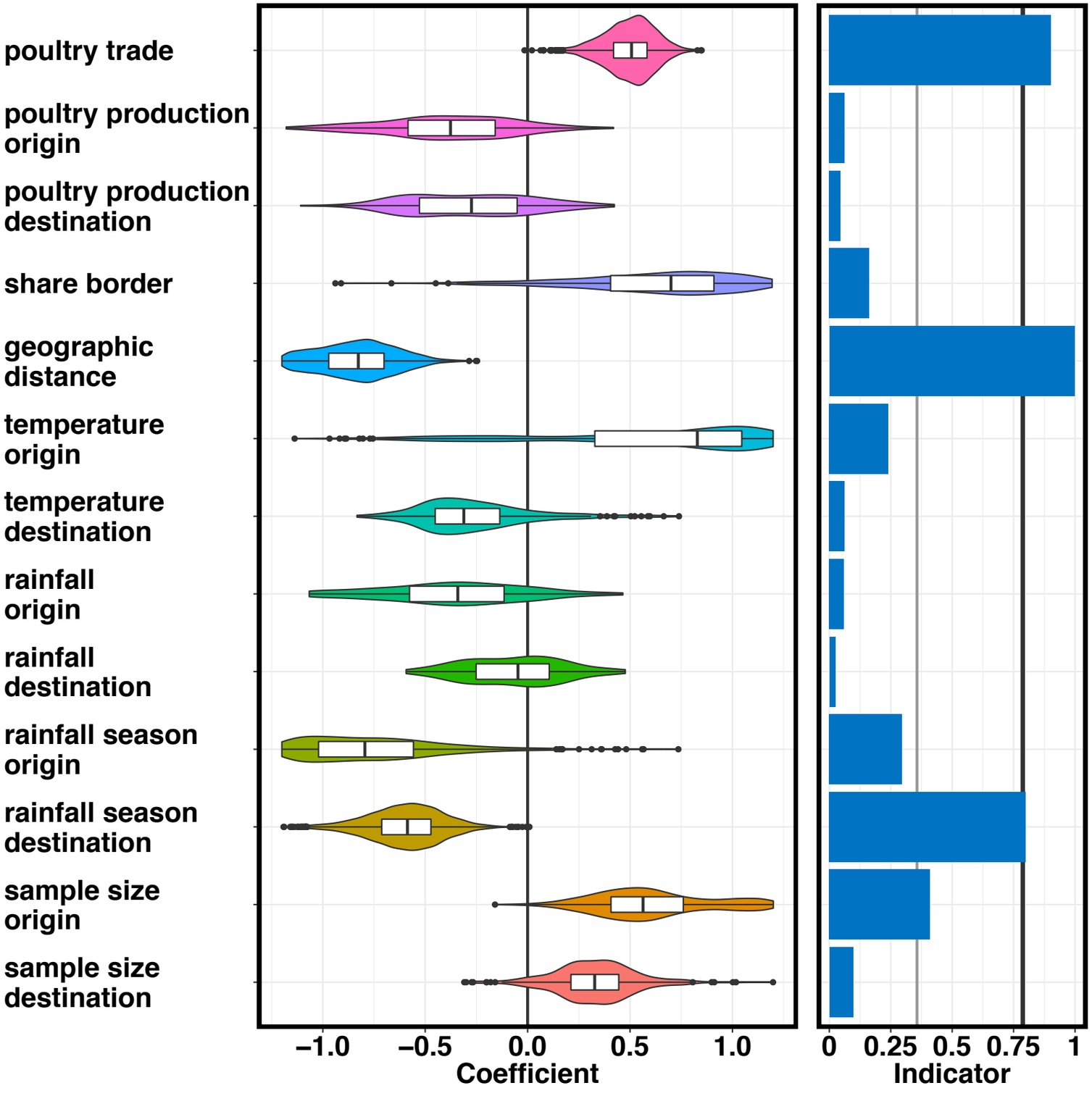
~10,000,000 nucleotides

Avian influenza H9N2 in asia

Yang, Mueller, Bouckaert, Xu, Drummond (https://doi.org/10.1371/journal.pcbi.1007189)



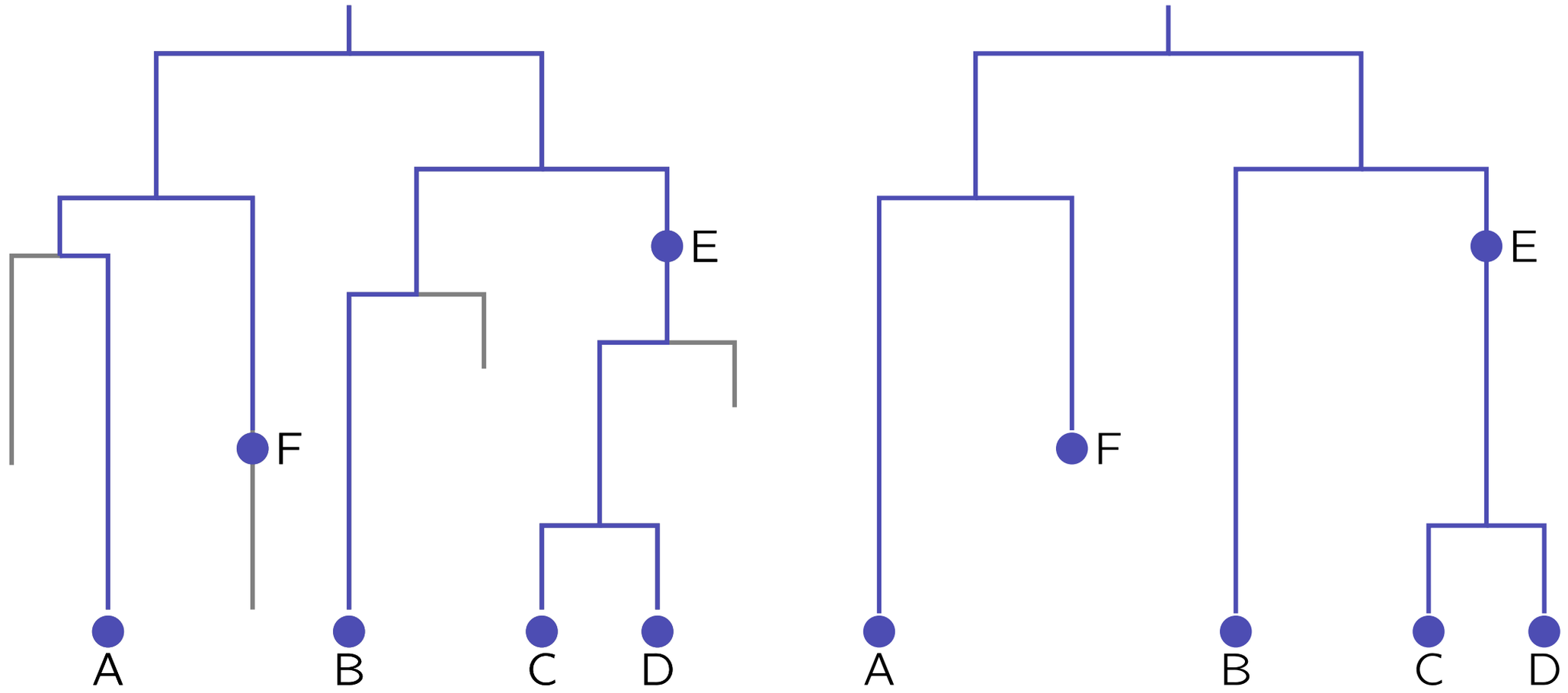
Phylogenetic GLM predictors of virus migration rate based on time-series data



Integrative phylogenomics

- The evidence for the evolutionary history of a clade of related species comes from a number of independent sources
  - **Genomic sequence data** from extant species
  - **Ancient DNA** from sub fossil species
  - **Phenotypic data** from extant and extinct species
  - **Fossil occurrence and age data** from extant and extinct species
  - **Biogeographic occurrence data** from extant and extinct species

Fossilized birth-death (FBD) model



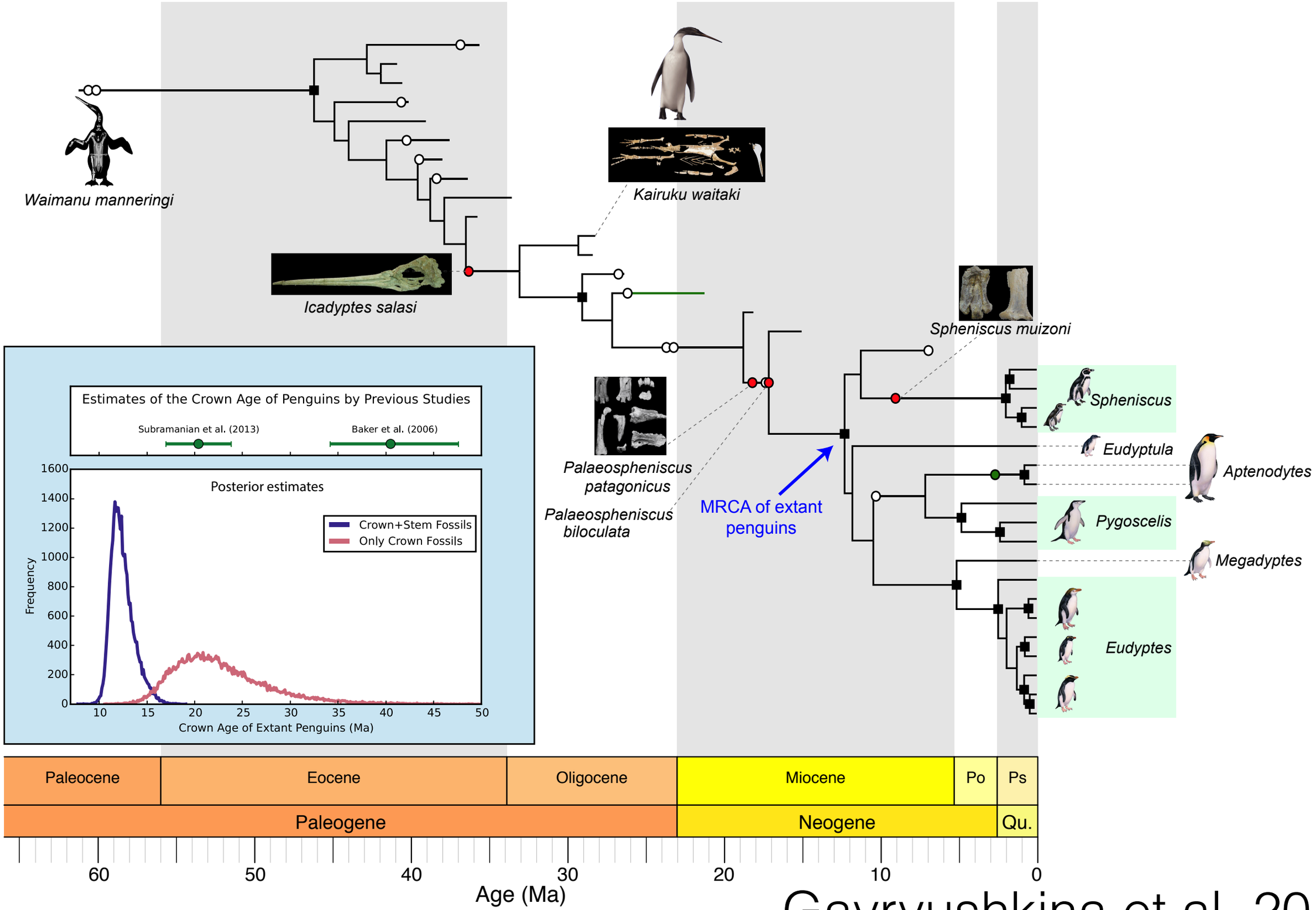
Full tree

Sampled tree

Bayesian phylogenetics

- **Bayesian phylogenetics** (Rannala & Yang,1996; Huelsenbeck & Ronquist, 2001)
- **Morphological substitution models** (Lewis, 2001)
- **Bayesian tip-dated phylogenetic models** (Drummond et al, 2002)
  - For ancient DNA and rapidly evolving viruses
- **Relaxed phylogenetics** (Drummond et al, 2006)
  - Reconciling branch-rate variation with time-trees by relaxing the strict molecular clock
- **Multispecies coalescent** inference (Liu, 2008; Heled and Drummond, 2010)
  - Gene tree / Species tree discordance (Pamilo & Nei, 1988; Maddison, 1997)
- **Fossilised birth-death process** (Heath et al, 2014, Gavryushkina et al, 2014)
  - Macroevolutionary inference of sampled ancestors

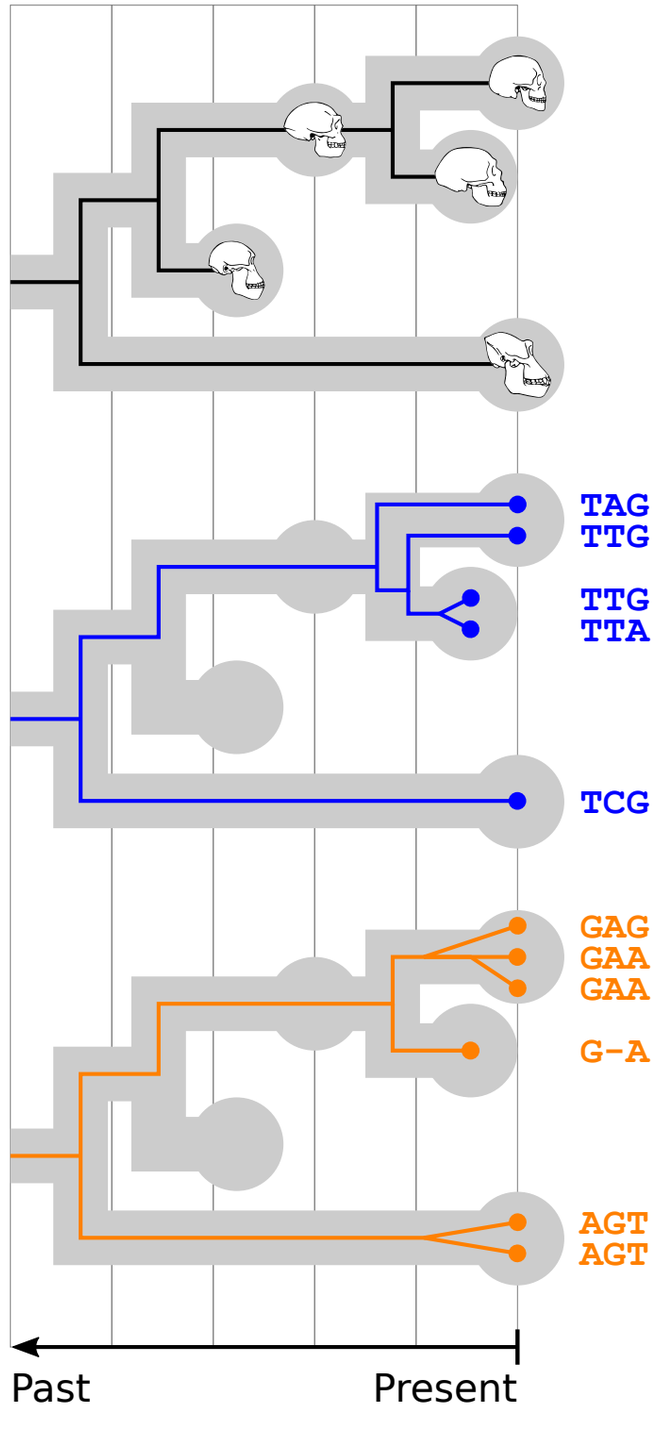
First step: Total-evidence of **DNA and fossils**



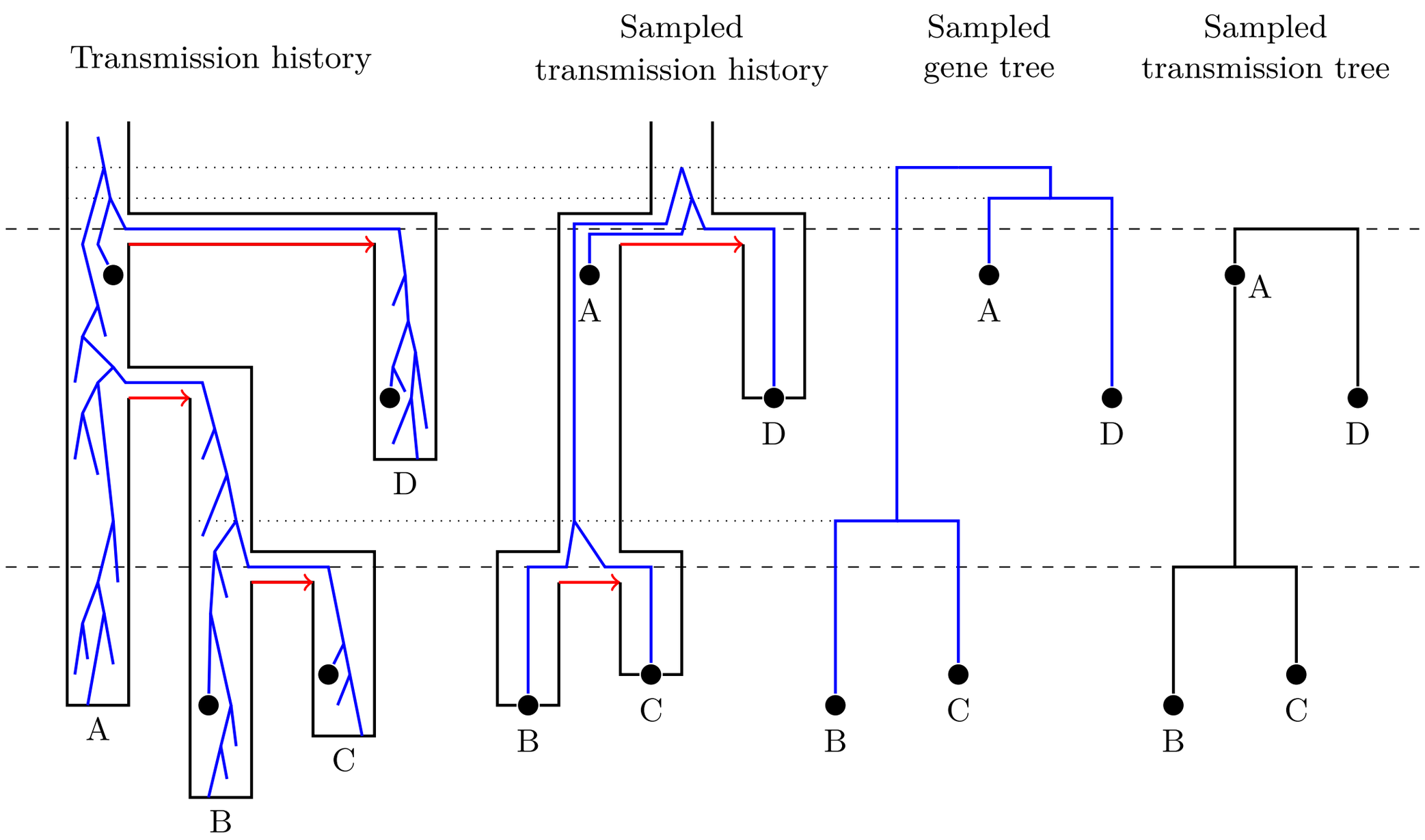
Gavryushkina et al, 2016

An integrative model

- Species tree modelled by the fossilised birth-death process
- Fossils may be samples from directly ancestral species
- Gene trees modelled by the multispecies coalescent process
  - No direct ancestors
- Genomic data evolves down gene trees
- Morphological fossil data evolves down the species tree

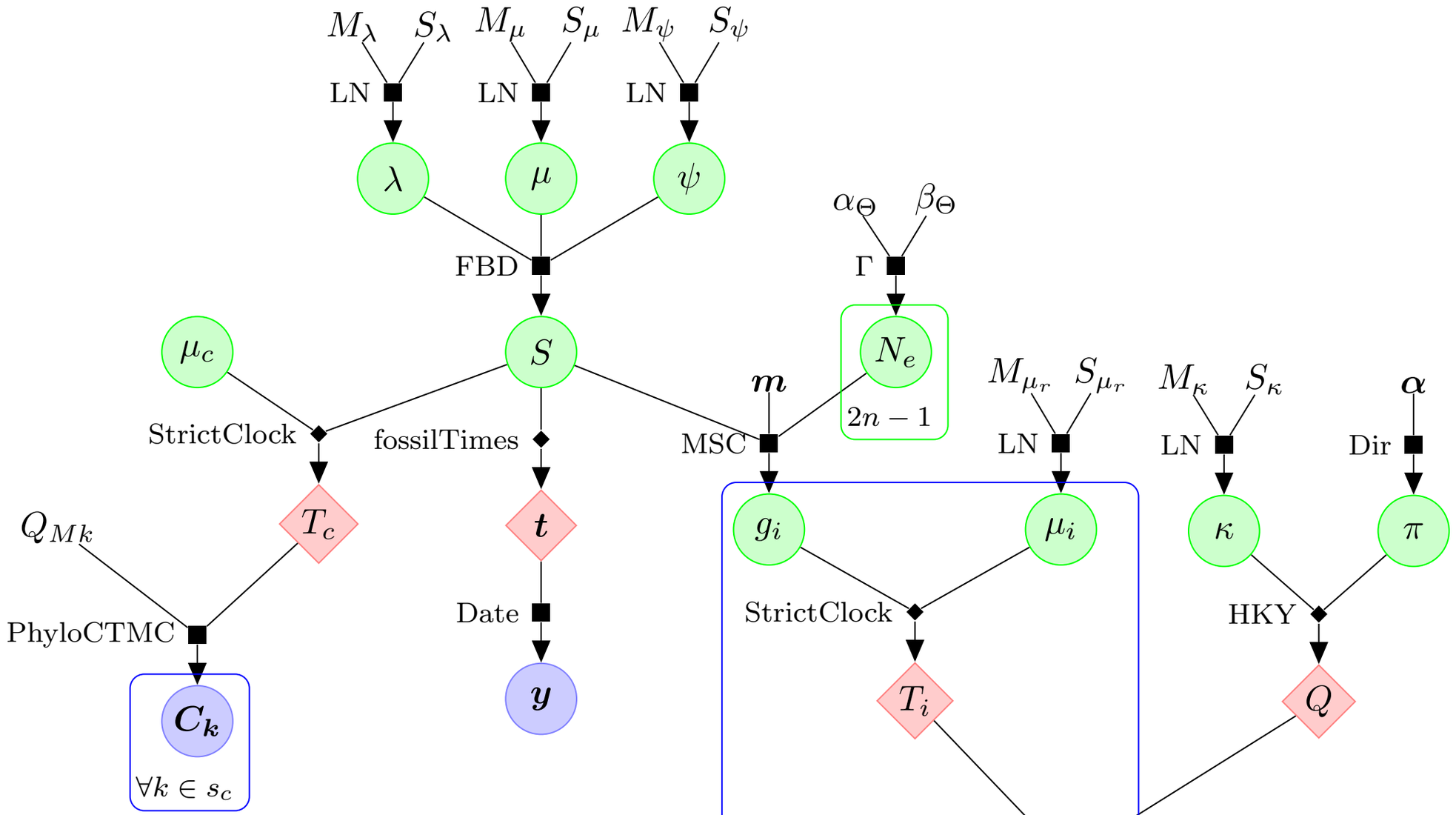


Another integrative model: pathogen evolution inside transmission trees



Goal: An integrated model of stochastic infection dynamics and viral genomic evolution. Integrating infection dynamics, genomic sequence evolution, (and phenotypic data?) into a single model.

Bayesian graphical model



$$P(S, \boldsymbol{N_e}, \boldsymbol{g}, \boldsymbol{\mu}, \lambda, Q | \boldsymbol{D}, \boldsymbol{C}, \boldsymbol{y}) \propto \prod_i \left[P(g_i | S, \boldsymbol{N_e}) \prod_j \Pr(\boldsymbol{D_{i,j}} | g_i \mu_i, Q) \right] \times$$
$$\prod_k \Pr(\boldsymbol{C_k} | \mu_c S, Q_{Mk}) \times$$
$$P(\boldsymbol{y}|S) P(S|\lambda) P(\lambda) P(\boldsymbol{N_e}) P(\boldsymbol{\mu}) P(\mu_c) P(Q)$$

Bayesian model

$$P(S, G, \theta | D, C) = \frac{1}{Z} \prod_i \left(\Pr(D_i | G_i) \cdot P(G_i | S) \right) \cdot \prod_j \Pr(C_j | S) \cdot P(S | \theta)$$

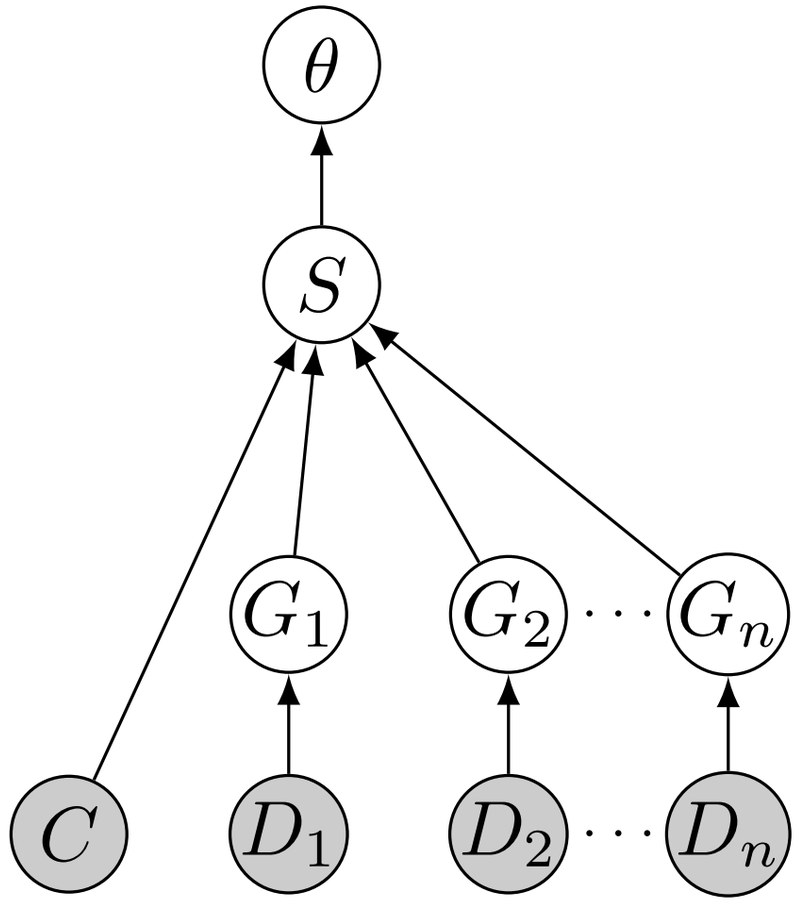
- Θ are the parameters of the birth-death tree generation process

- S is the species time-tree in calendar units (including effective population sizes in “calendar units”)

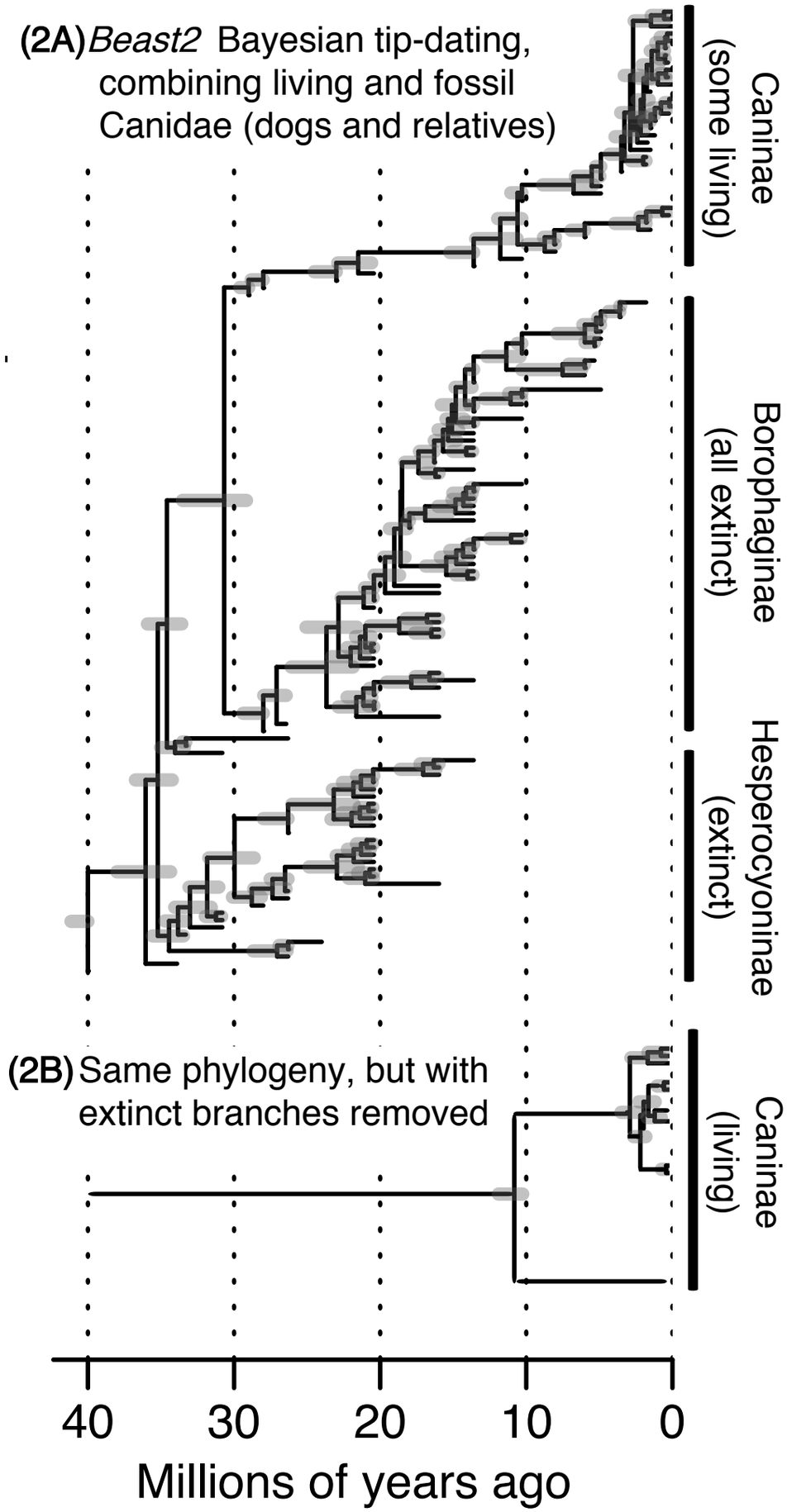
- G are the gene trees in calendar units

- C morphological character data, fossil occurrence and geological age data

- D gene sequence alignments (and radiocarbon sample ages in the case of ancient DNA)



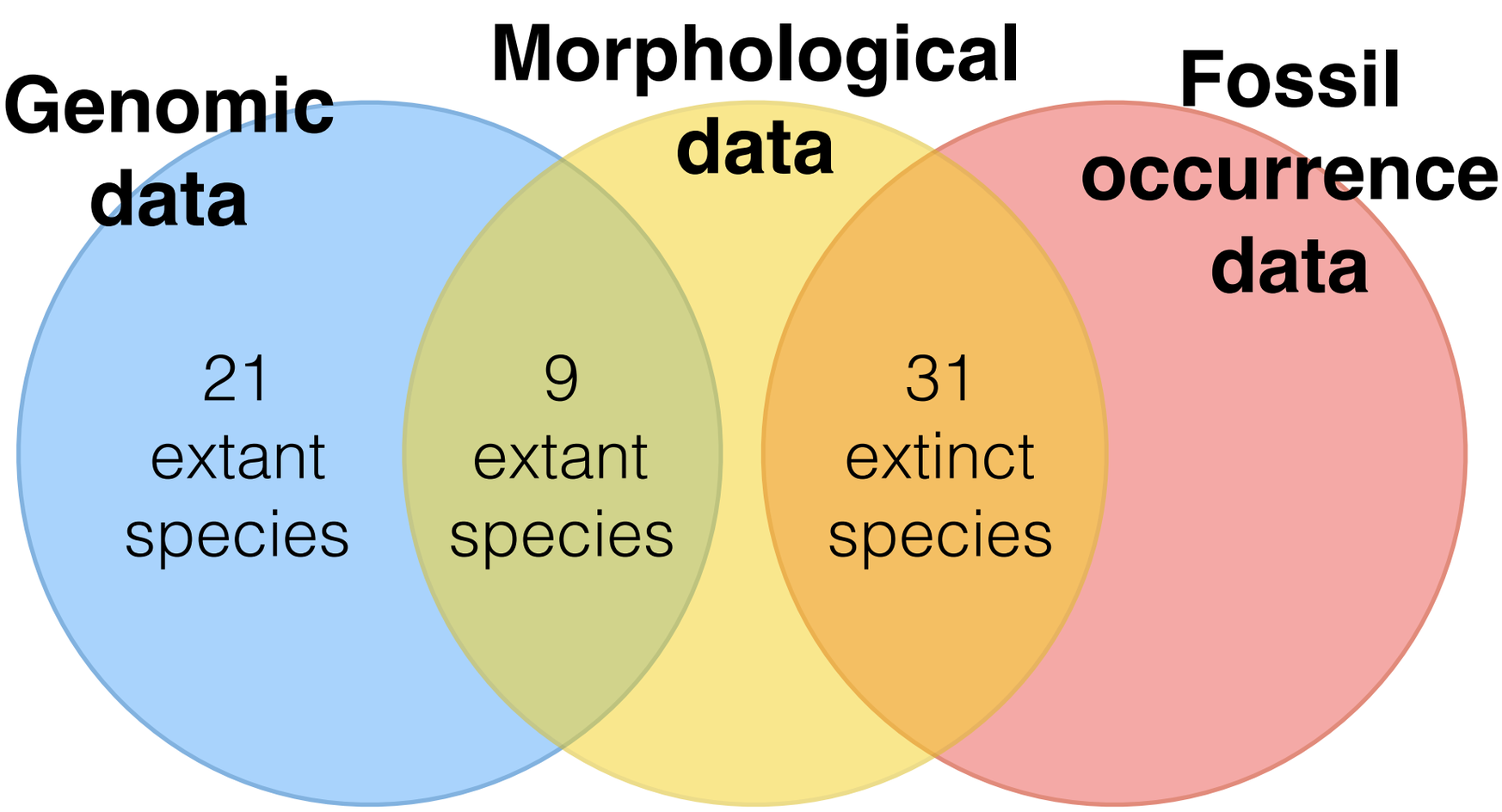
Caninae example



Caninae example

- 19 independent nuclear loci from 30 species (570 gene sequences; 490,760 nucleotide calls)

- 72 morphological characters from 40 species (2880 morphological character calls)



Caninae analyses

- Fixed molecular clock, 30 extant, no fossil species
    - Fixed molecular rate: $8x10^{-4}$ substitutions per site per million years
        - MSC and concatenated analyses
- molecular + morphological data, 30 extant, no fossil species
    - MSC and concatenated analyses
- Estimating molecular clock using molecular data, morphological data and fossil times, 30 extant, 31 extinct species
    - MSC and concatenated analyses

Fixed clock multispecies coalescent tree annotated with concatenation-analysis branch length differences

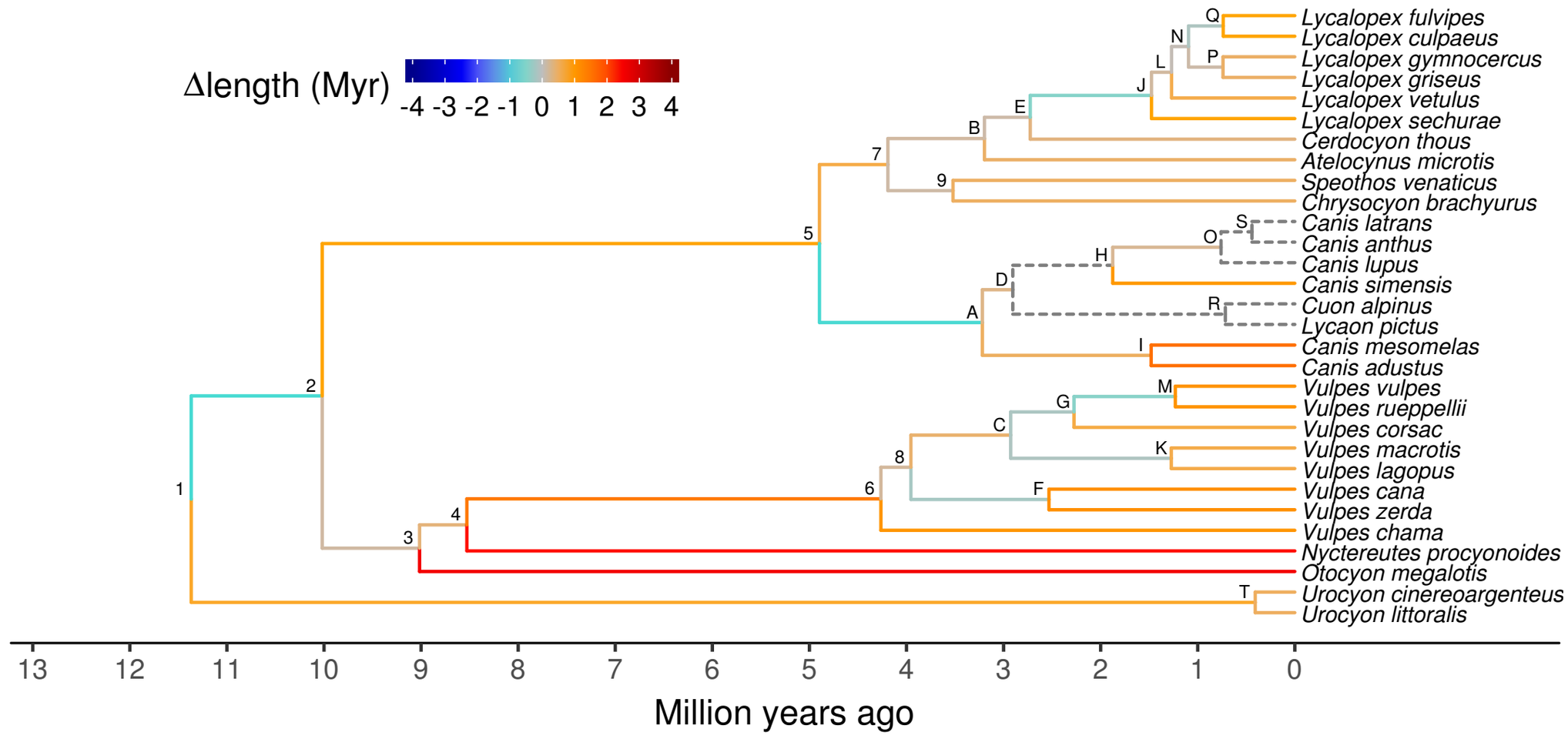
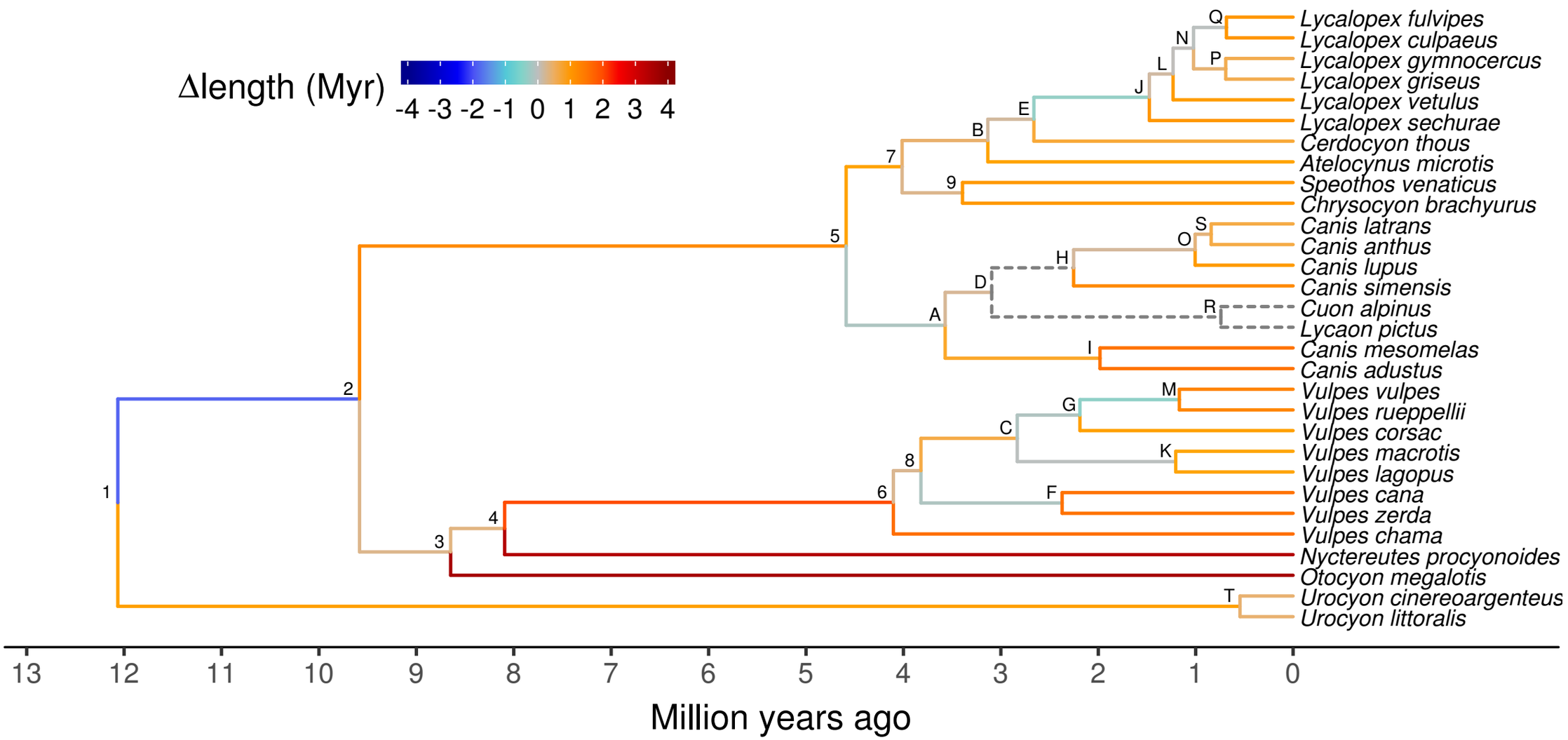
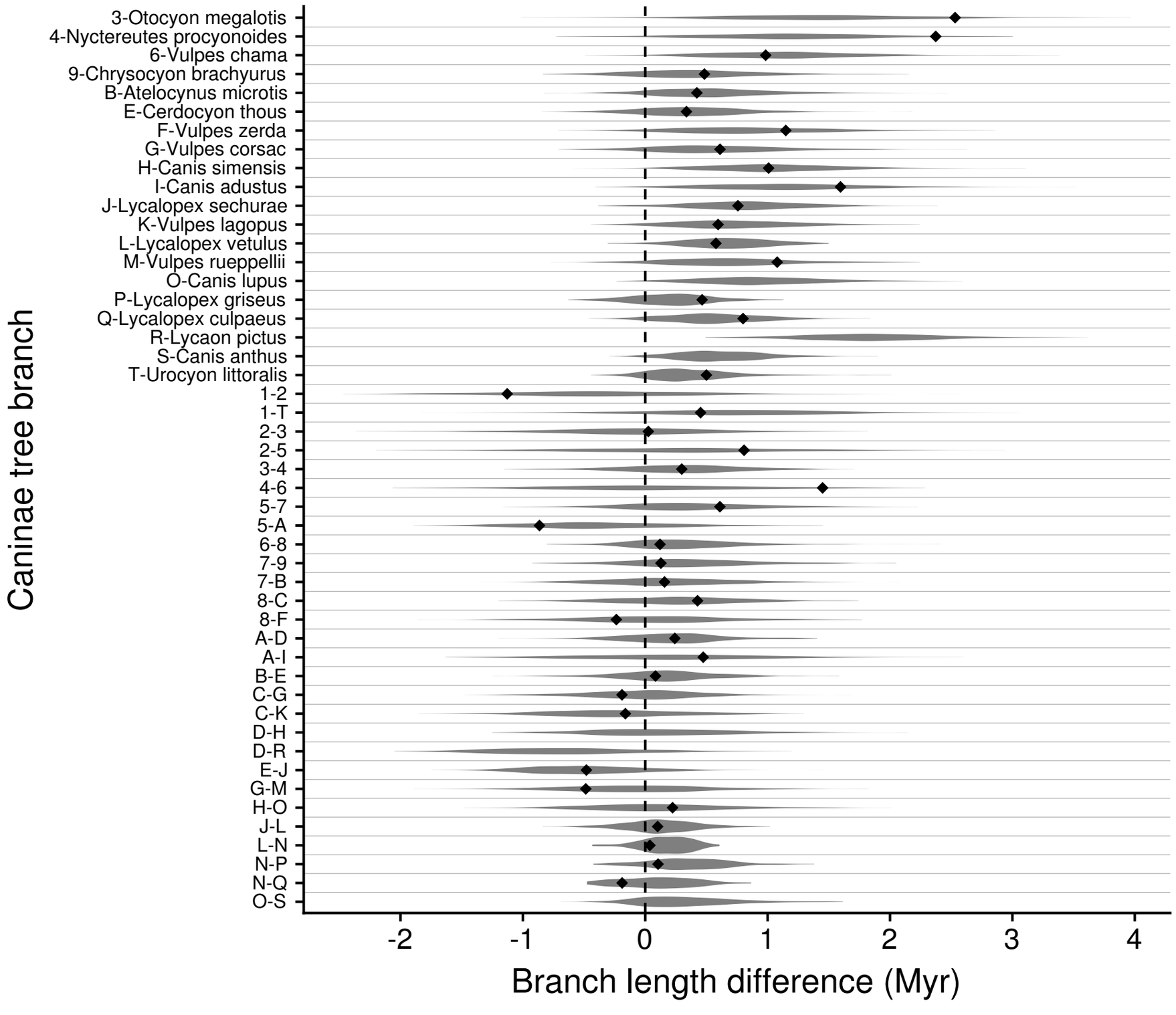


Figure 3: Branch length changes resulting from concatenation using a fixed clock. The color shows how branch length estimates differ when using concentration rather than the multispecies coalescent (MSC). The tree is a maximum clade credibility (MCC) summary tree with mean node ages, generated from the MSC posterior distribution of species trees, inferred using molecular and extant morphological data with a fixed clock. The difference in branch lengths is the mean among concatenation samples including that branch, less the MSC mean. Dashed lines represent branches with less than 0.5% support using concatenation.

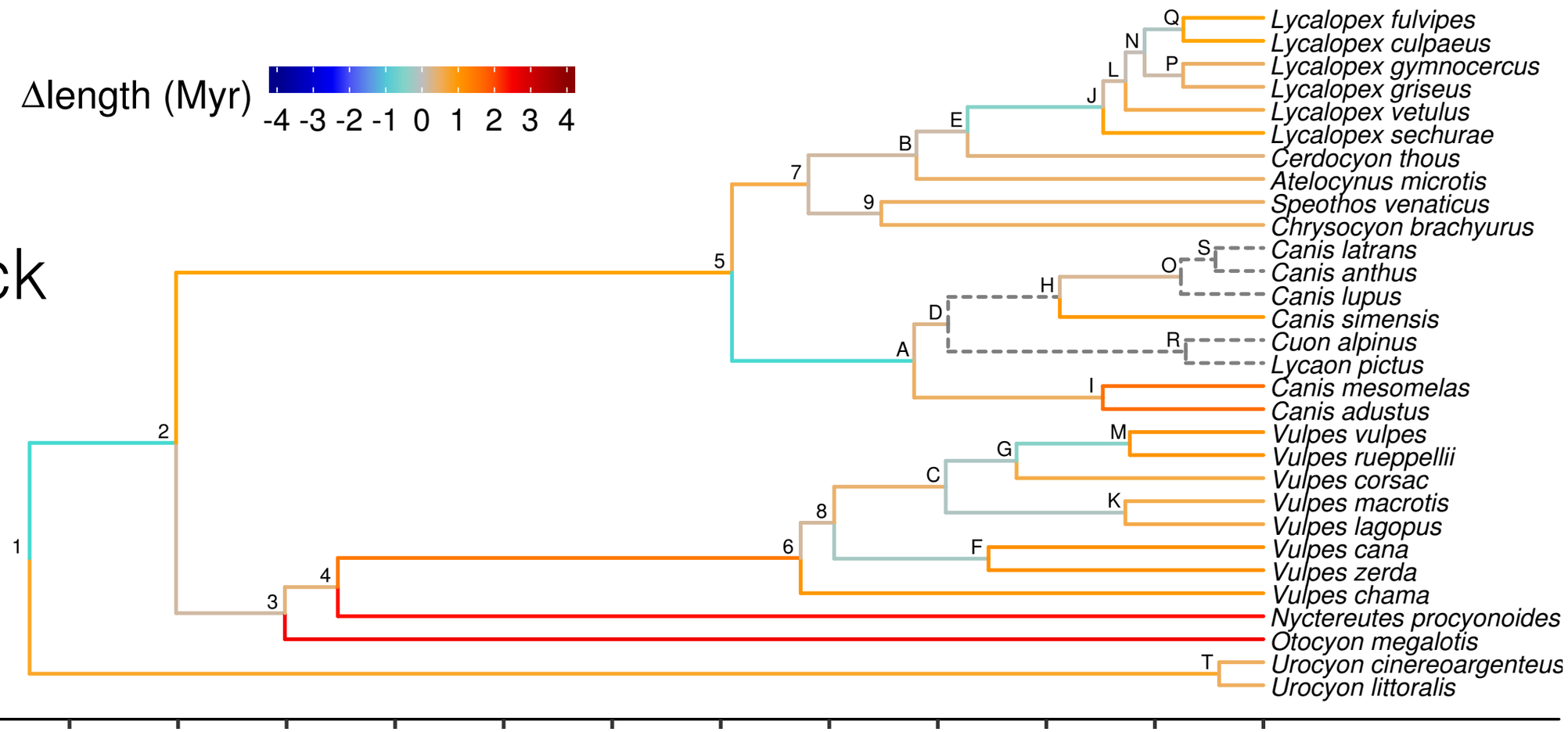
FBD multispecies coalescent tree trimmed to show only extant taxa annotated with concatenation-analysis branch length differences



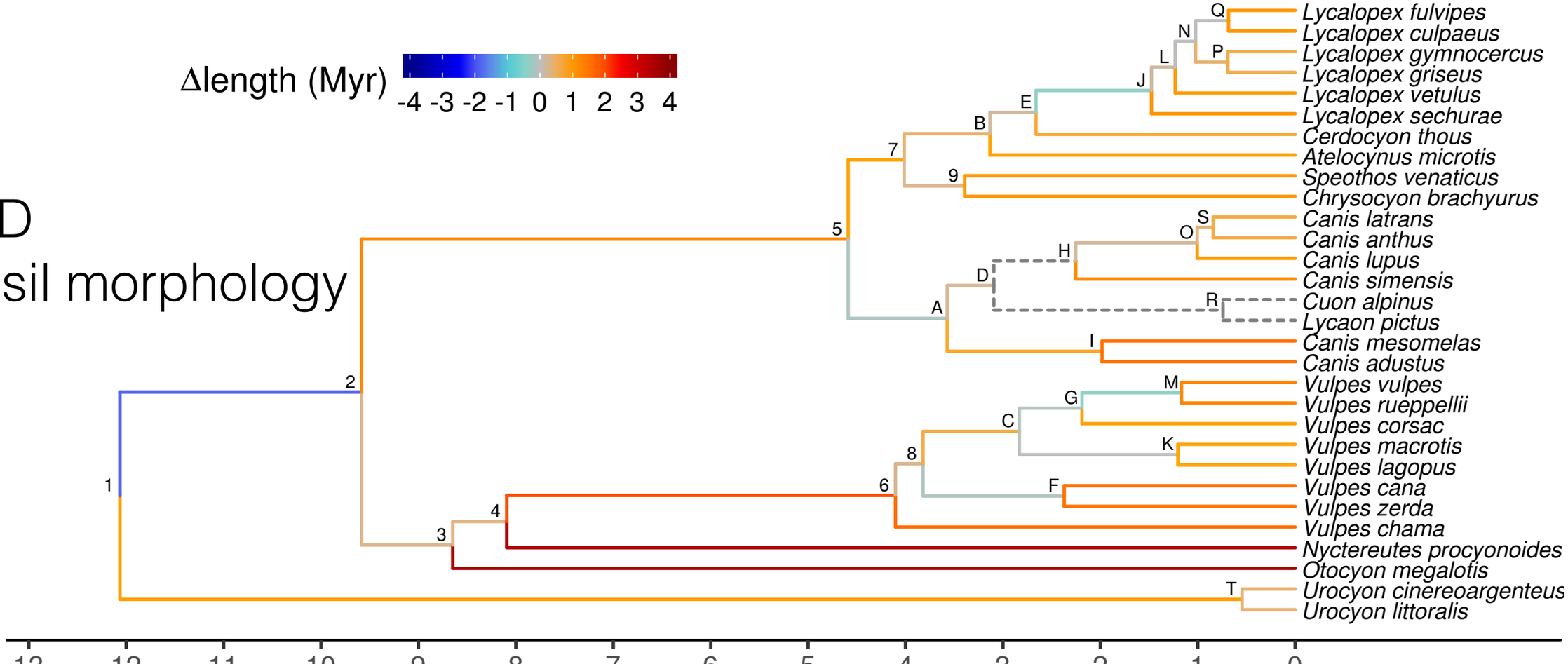
Posterior and predictive differences between MSC and concatenation



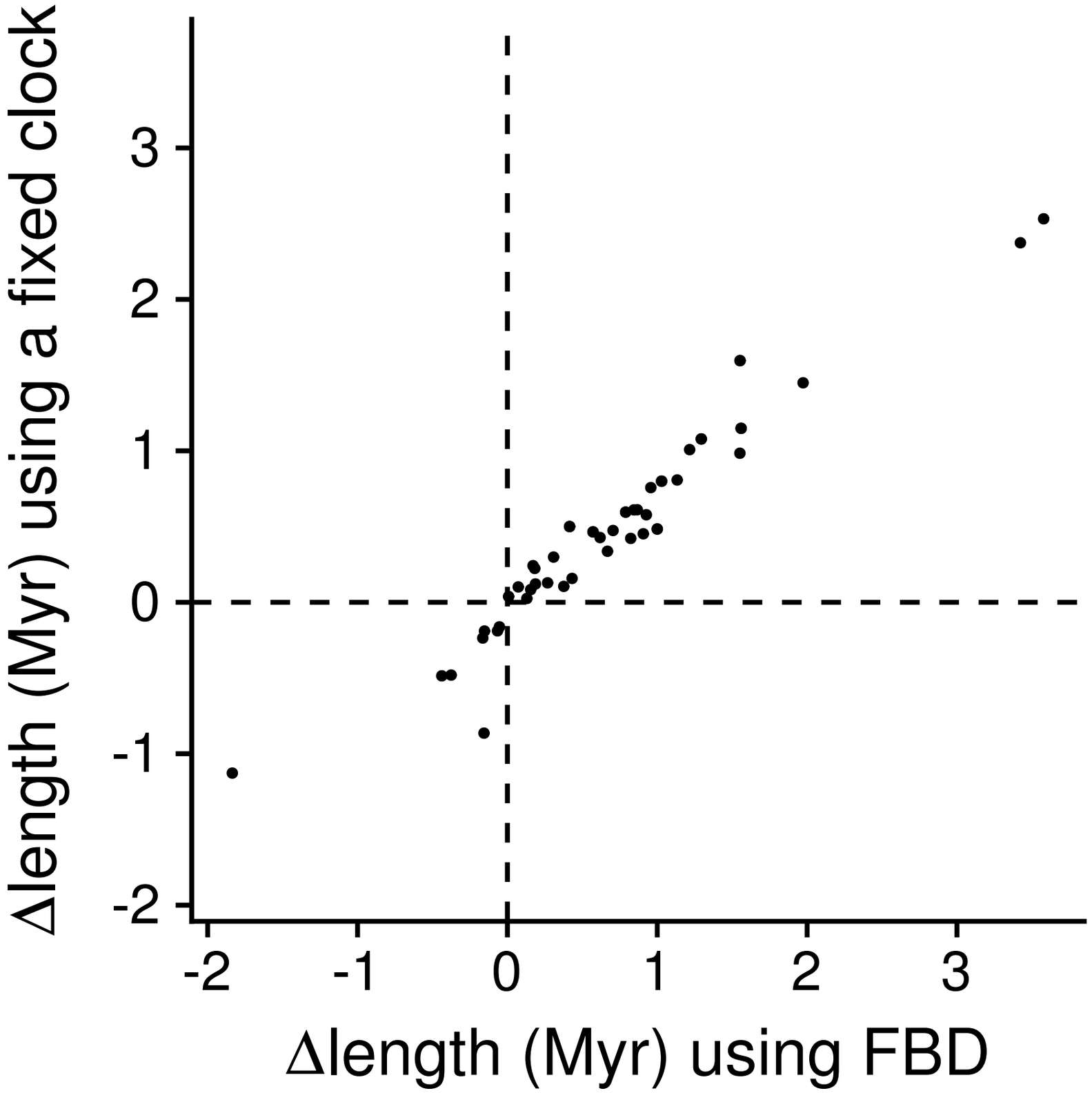
Fixed clock



+ FBD
+ fossil morphology



Consistency of concatenation distortions



Macroevolutionary parameters

	MO	FC	FBD
	Multispecies coalescent		
Molecular clock rate $\times 10^{-4}$	8	8	8.24 (6.02–9.78)
Morphological clock rate	NA	0.13 (0.09–0.17)	0.05 (0.04–0.06)
Mean $N_e g$	0.47 (0.37–0.58)	0.50 (0.40–0.62)	0.51 (0.36–0.66)
Diversification rate $\lambda - \mu$	0.23 (0.11–0.35)	0.23 (0.10–0.36)	0.14 (0.02–0.26)
Turnover $\mu \div \lambda$	0.26 (0.00–0.64)	0.29 (0.00–0.66)	0.71 (0.47–0.95)
Sampling proportion $\psi \div (\psi + \mu)$	NA	NA	0.31 (0.14–0.50)
	Concatenation		
Molecular clock rate $\times 10^{-4}$	8	8	7.36 (5.79–8.61)
Morphological clock rate	NA	0.09 (0.07–0.11)	0.05 (0.04–0.06)
Diversification rate $\lambda - \mu$	0.19 (0.10–0.28)	0.19 (0.10–0.29)	0.12 (0.03–0.20)
Turnover $\mu \div \lambda$	0.20 (0.00–0.52)	0.20 (0.00–0.51)	0.63 (0.33–0.91)
Sampling proportion $\psi \div (\psi + \mu)$	NA	NA	0.40 (0.18–0.65)

Parameters were estimated using a fixed clock and molecular only data (MO), a fixed clock with molecular and extant morphological data (FC), or a fossilized birth-death process with molecular, extant morphological and fossil data (FBD).

Values are posterior mean estimates followed in brackets by 95% highest posterior densities.

Clock rates are in units of per-site or per-character per million years. Diversification rate is in per million years.

Mean $N_e g$ refers to the mean of the effective population size N_e distribution, which is scaled by generation time g.

Divergence time estimation

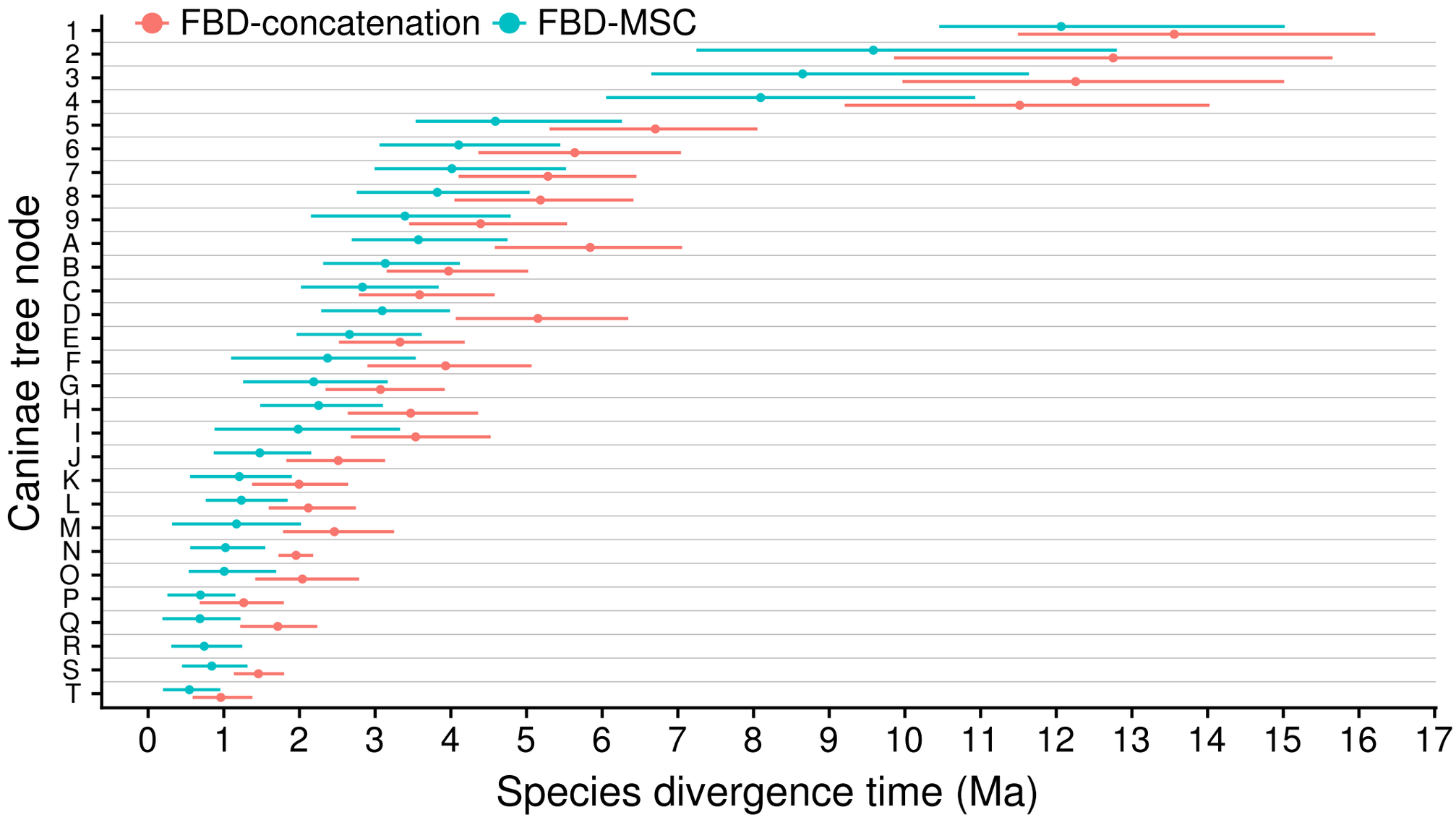


Figure 7: Speciation times estimated by fossilized birth-death with multispecies coalescent (FBD-MSC) and with concatenation (FBD-concatenation) models. Posterior mean FBD-MSC node ages (solid circles) and 95% highest posterior density (HPD) intervals (lines) are estimated from samples where that clade is present. FBD-concatenation ages and intervals are also conditioned on clade presence. Node labels correspond to those in Figure 3 and 5.

Lineages through time

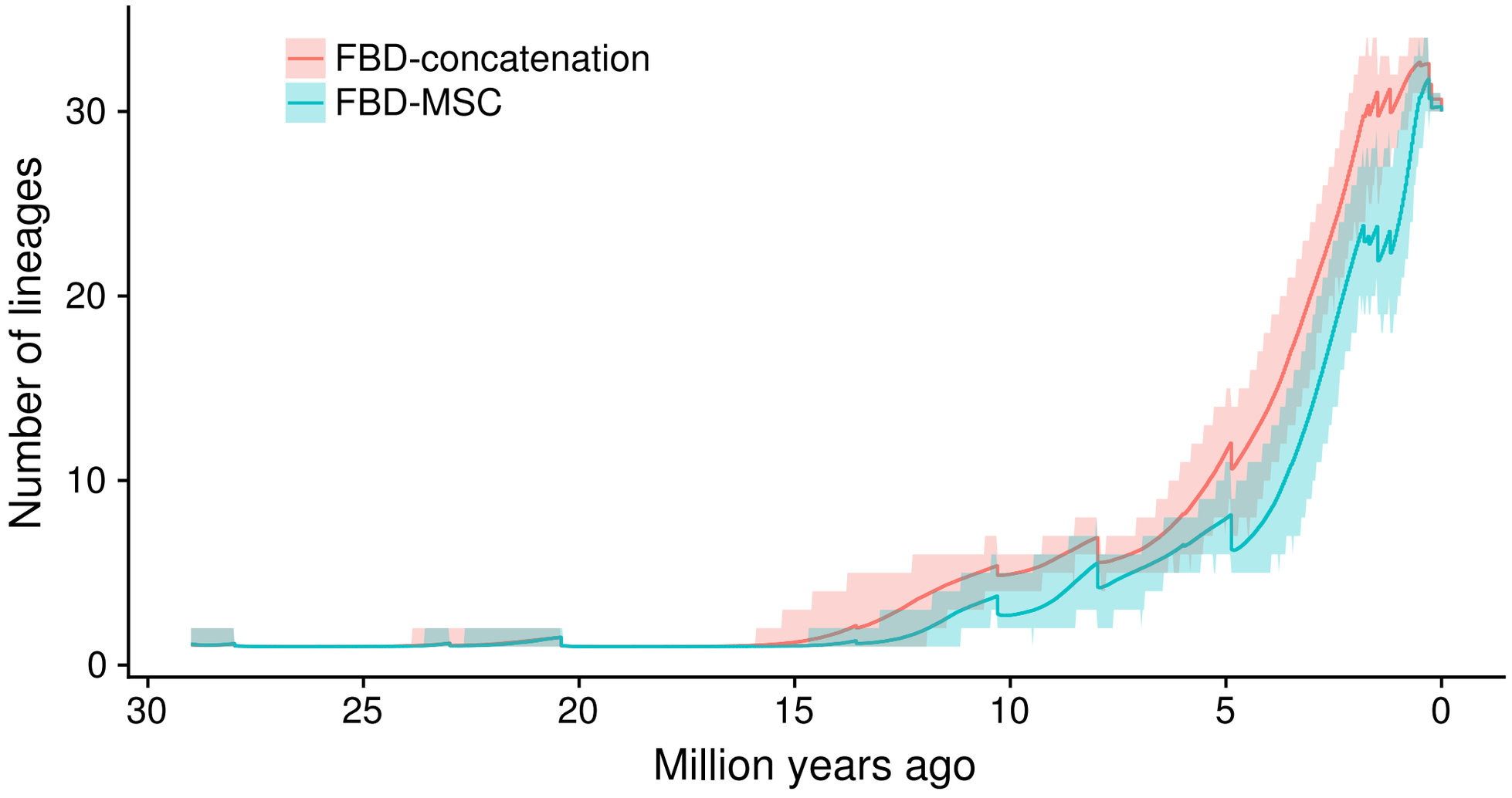


Figure 8: Lineages-through-time (LTT) plot of Caninae diversification. Posterior mean estimates (solid lines) of LTT are calculated for 1,001 evenly spaced time steps spanning 0 to 1, and include extant, fossil and ancestral taxa, and sampled ancestors (which are both fossil and ancestral). 95% highest posterior density (HPD) intervals were also calculated for each step, and are shown as translucent ribbons.

Conclusions

- Considering the totality of our results, the FBD-MSC results are more plausible than the FBD-concatenation results.

- The posterior predictive simulations show that the observed differences in branch lengths between the MSC and concatenation are expected due to a failure to account for coalescent processes.

- This has important implications for downstream analyses, as seen in the LTT plots where

  - the FBD-concatenation LTT curve suggests a slowdown in Caninae diversification during the past ~2 million years.

  - In contrast, the FBD-MSC LTT curve shows a burst of diversification in the same time frame.

Final Perspectives

- **Evolutionary biology has become a multidisciplinary analytical science**, with major input from computer scientists, statisticians, mathematicians and physicists.

- **Evolutionary biology is not just an historical science**. Rapidly evolving natural systems, low-cost high-throughput sequencing and high-throughput automated experimental evolution platforms, all add up to the potential to close the loop between experimental and theoretical evolutionary biology.

- **A common set of evolutionary modelling principles can inform** us on diverse questions spanning most forms of life and a vast range of evolutionary timescales.

Conclusions

- Bayesian statistical inference derives natural from the rules of probability, and is the only inferential method that provides a consistent way to build up knowledge as evidence accumulates, and to bridge differences in prior knowledge.

- The hypothesis space of phylogenetics (tree space + parameters) has a distinctive structure that frustrates attempts to use standard statistical inference software. Thus specialist inference software such as BEAST and MrBayes have been developed.

- Evolutionary biology and phylogenetics is a statistical science, in which mature statistical inference methods are now routinely used.

- Research in this field depends on continuous development and maintenance of large software packages, and this is currently still a challenge for science funding models.