Molecular clocks, calibrations and tree priors

Professor Alexei Drummond

School of Biological Sciences
School of Computer Science
University of Auckland

15th August 2023, Taming the BEAST eh!, Squamish, Canada

Outline

- **Molecular clock and calibrations** - a modelling framework for modelling rates of evolution

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

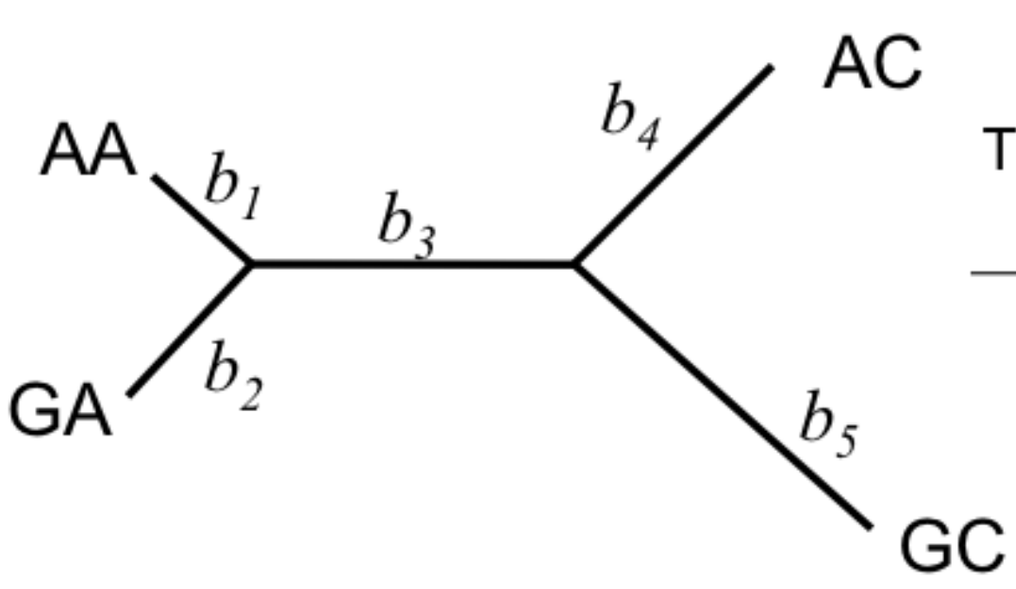
- **Tree priors** - an overview of coalescent and some key models

- **Integrative models**

Molecular clocks and calibrations

The molecular clock constraint

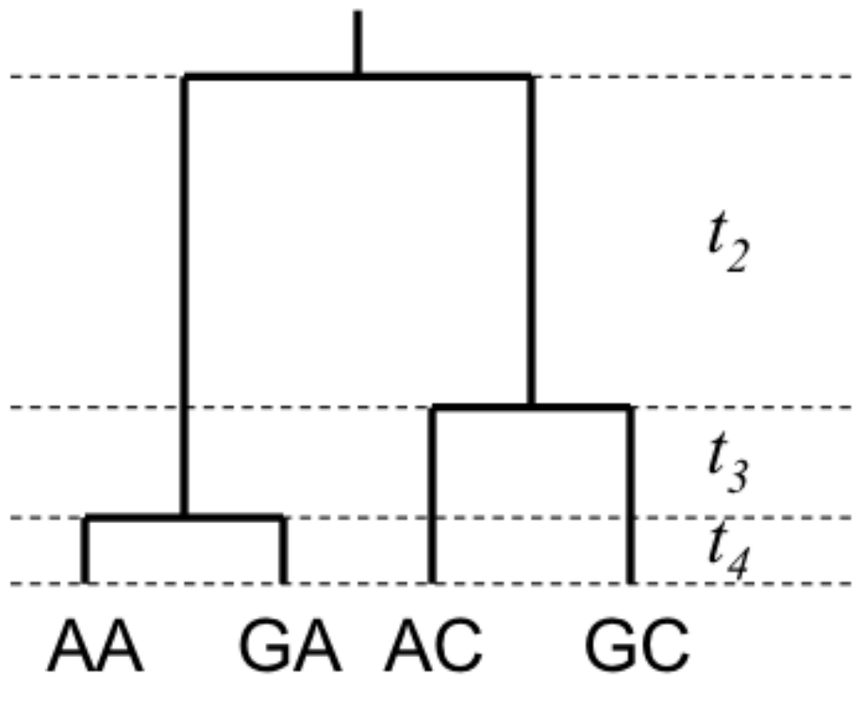
T



2n–3 branch lengths

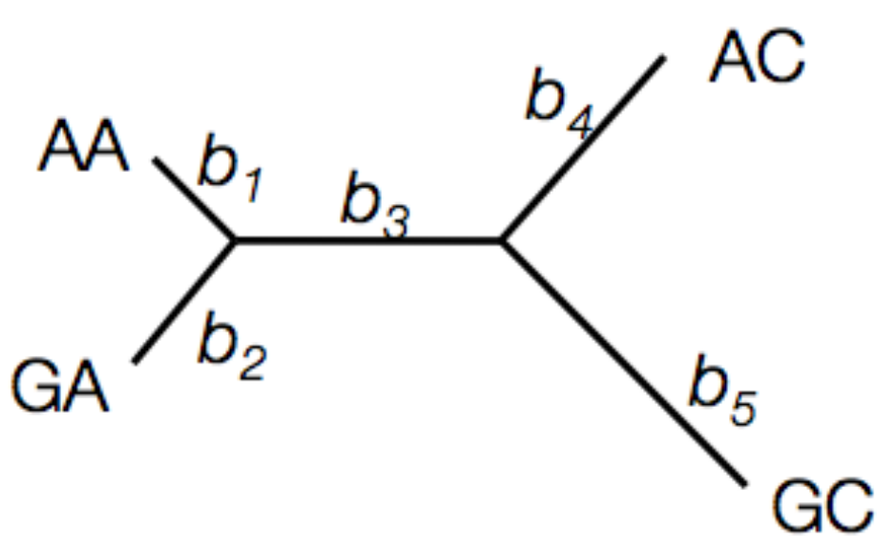
The "molecular clock" constraint

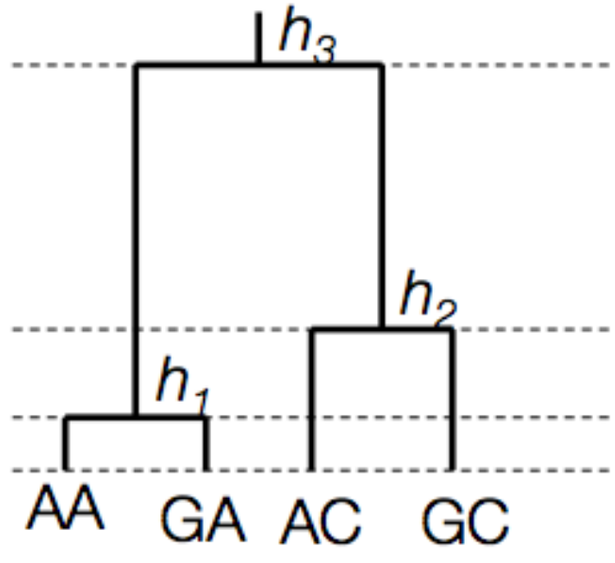
g



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:

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

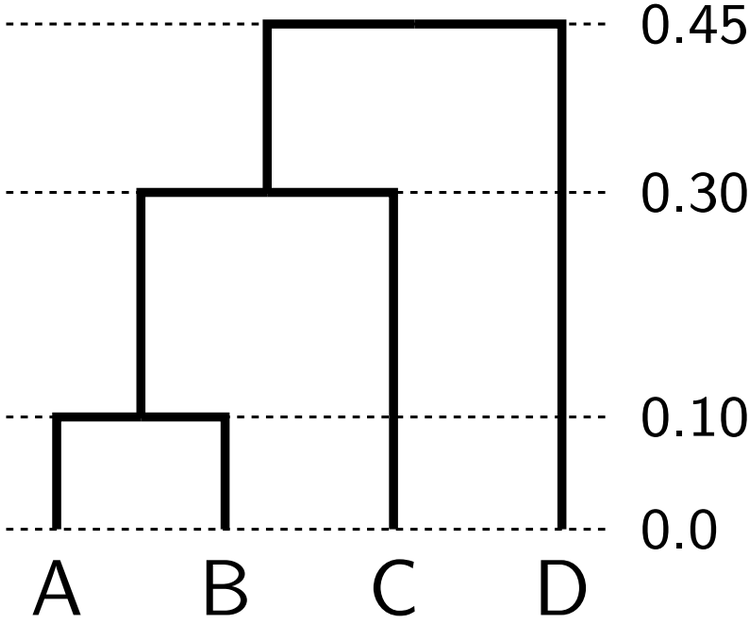
Estimate the global rate:

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

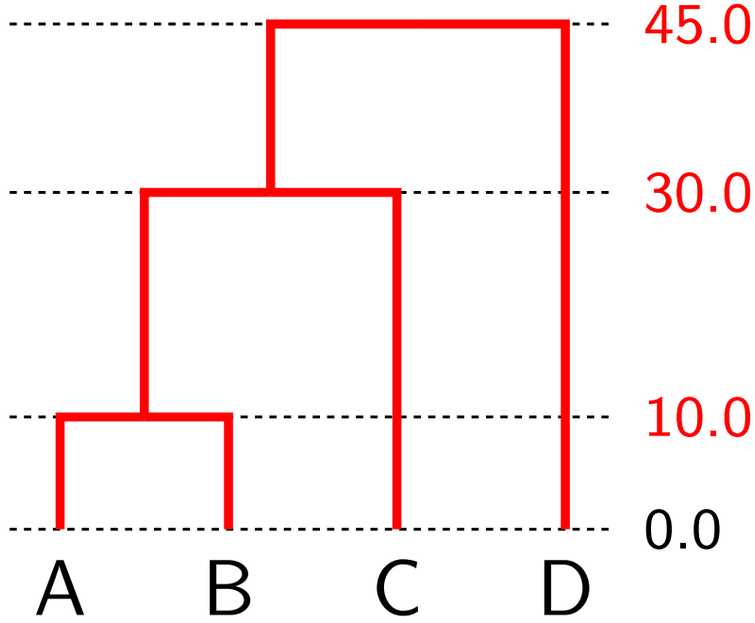
In the models above the parameters related to the details of the substitution process (Q) have been suppressed for simplicity.

Genetic distance = rate x time

$$T = \mu \times g$$



= 0.01 ×

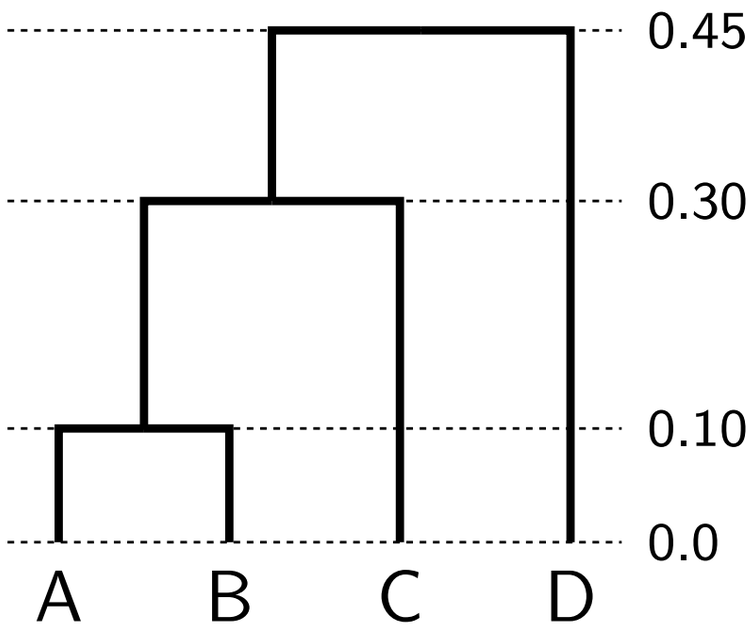


"substitution tree"

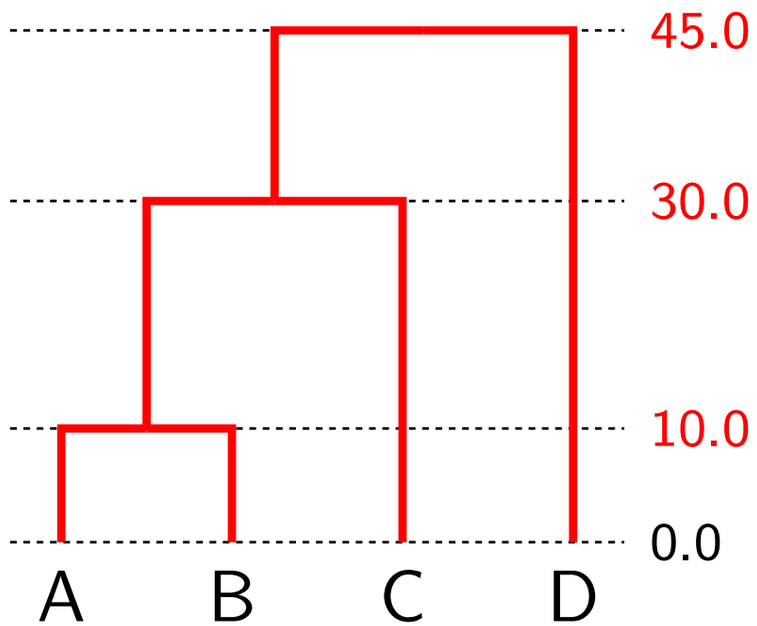
evolutionary rate
substitutions / site / unit time

time tree

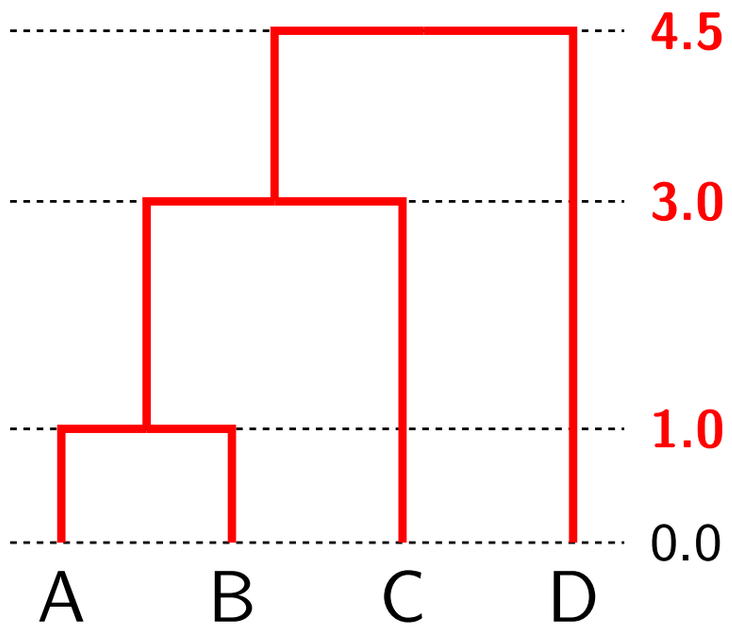
Non-identifiability of rates and times



$=$ 0.01 $\times$



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



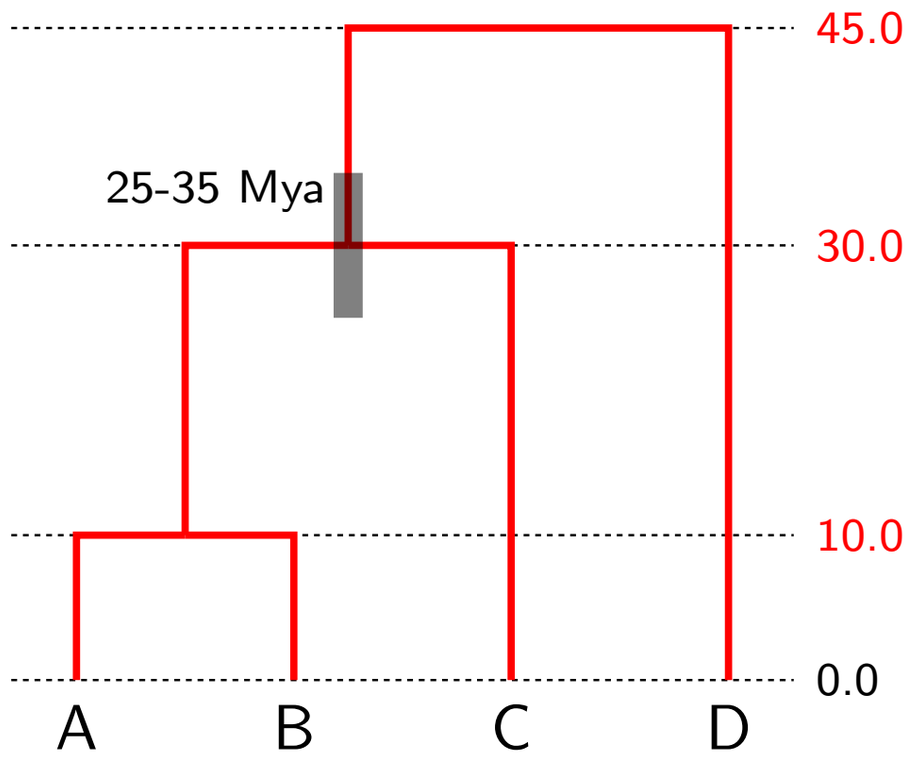
"substitution tree"

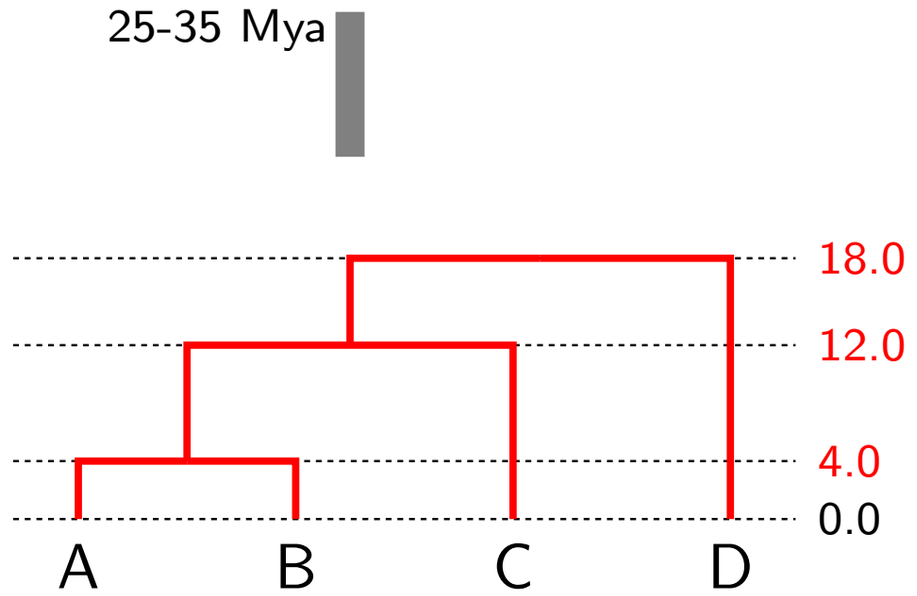
evolutionary rate
substitutions / site / unit time

time tree

Node calibration

Suppose fossil evidence shows the common ancestor of species A, B and C lived 25-35 Mya.





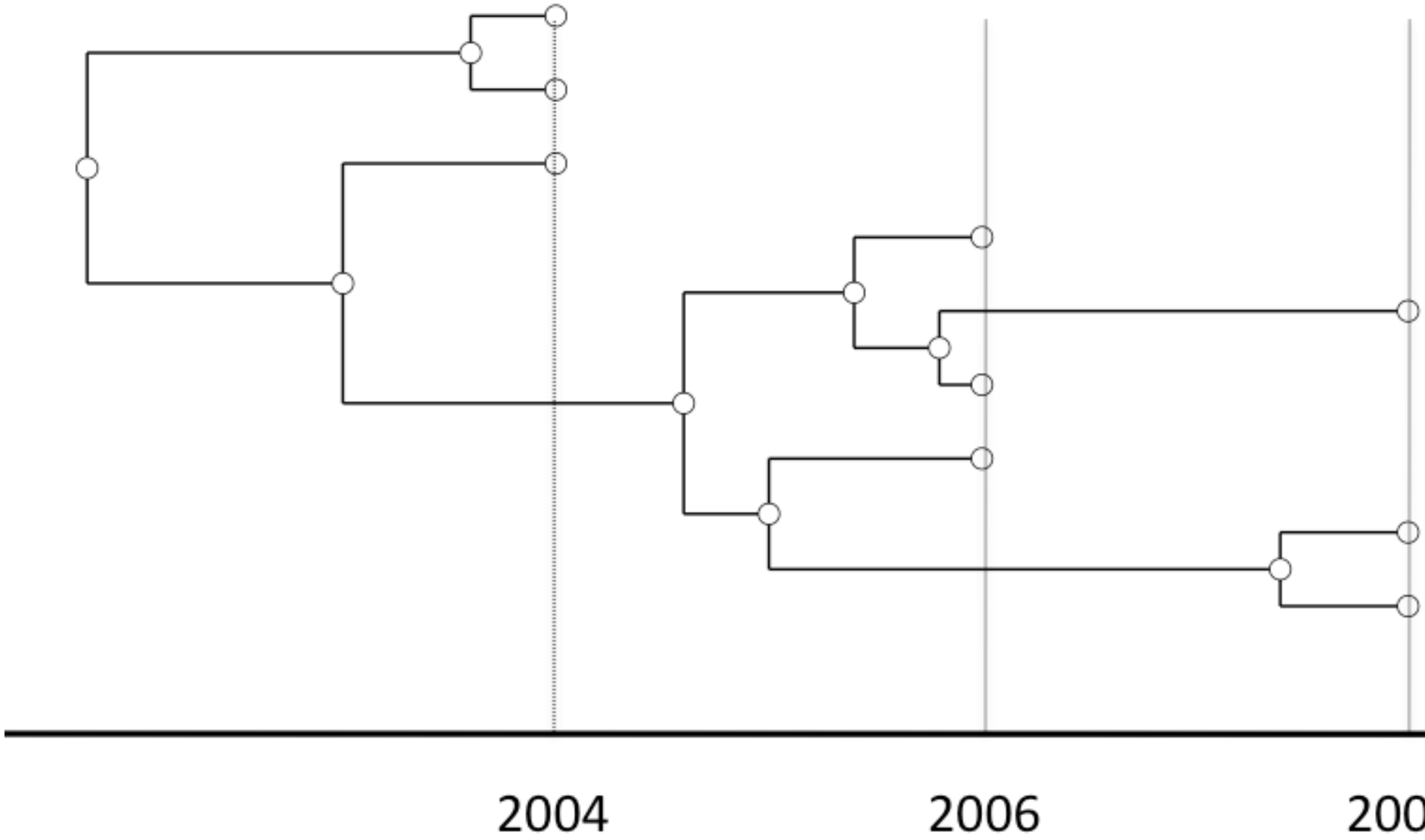
With a strict molecular clock, only the age (range) of a single node in the tree is needed in order to interpolate and extrapolate the ages of all other divergence times.

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

Tip calibration

Modelling phylogenetic data sampled through time

Drummond *et al* (2002)

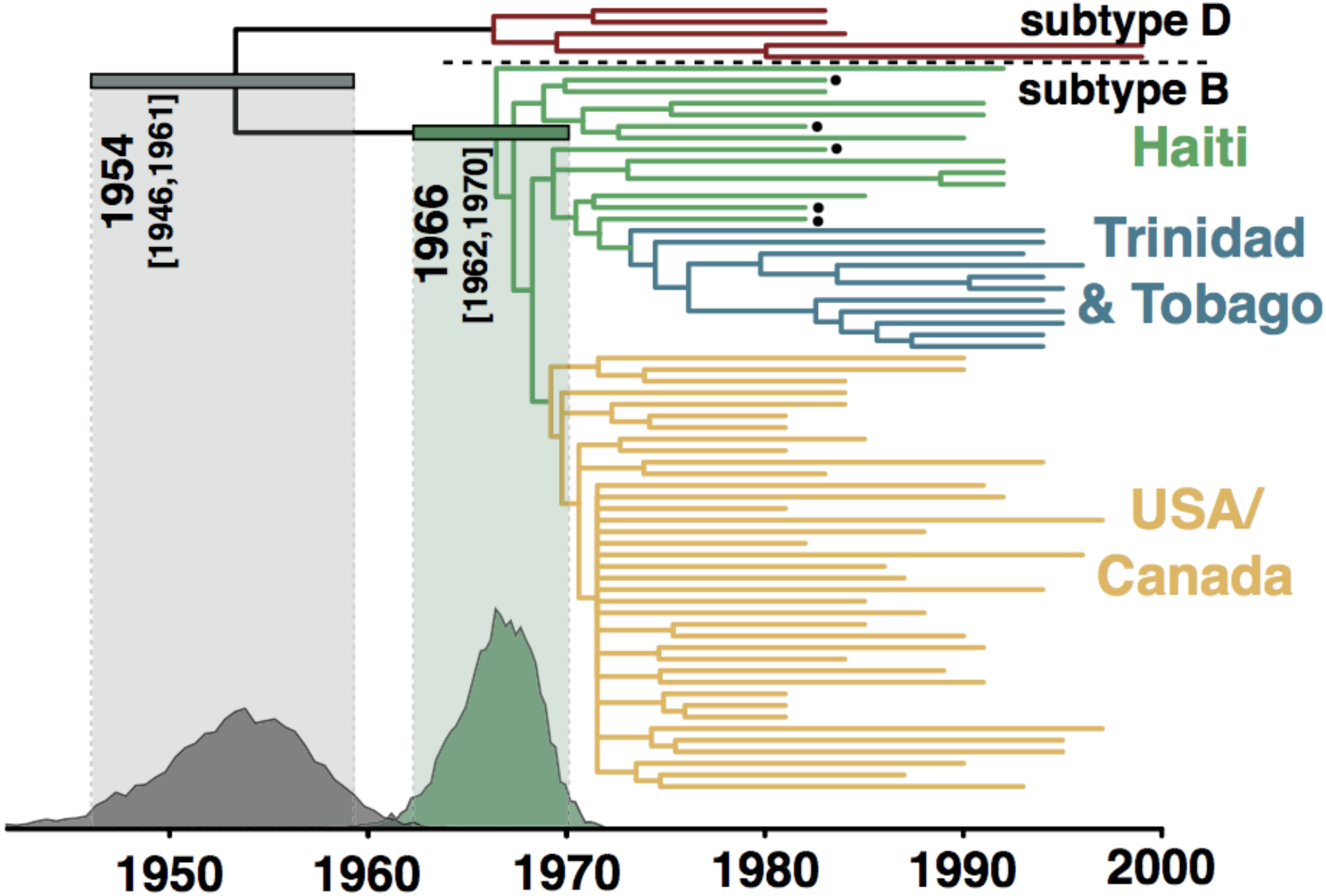


- Rapidly evolving microbes
- Ancient DNA
- Cancer
- Somatic evolution
- Languages
- *et cetera*

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

A calibrated phylogenetic inference

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



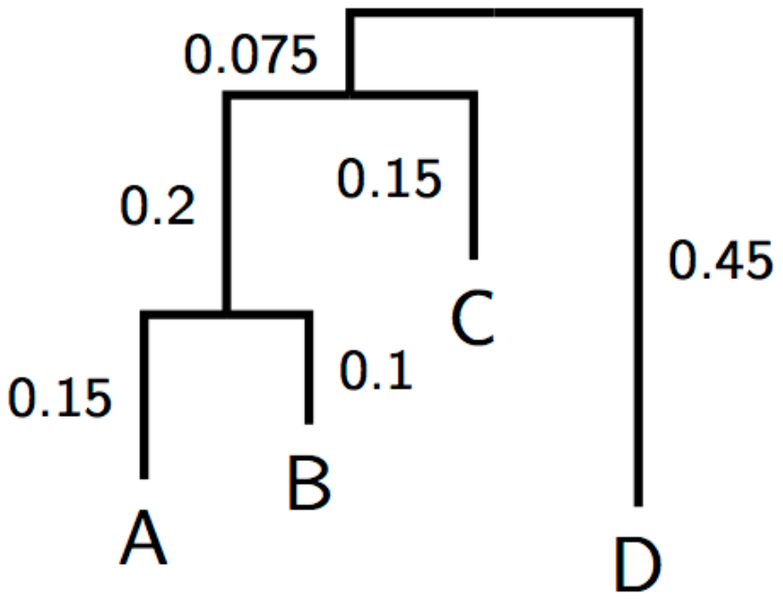
A phylogenetic reconstruction of samples of HIV-1 virus. Each tip represents a single infected individual from whom a blood sample has been taken.

Relaxed phylogenetics

Genetic distance = rate x time

Relaxed molecular clock

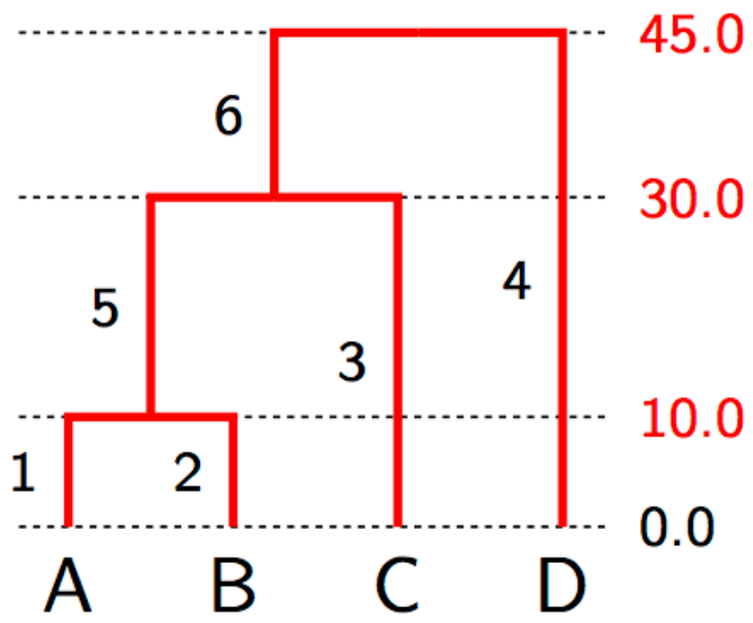
$$T = \vec{\mu} \star g$$



"substitution tree"

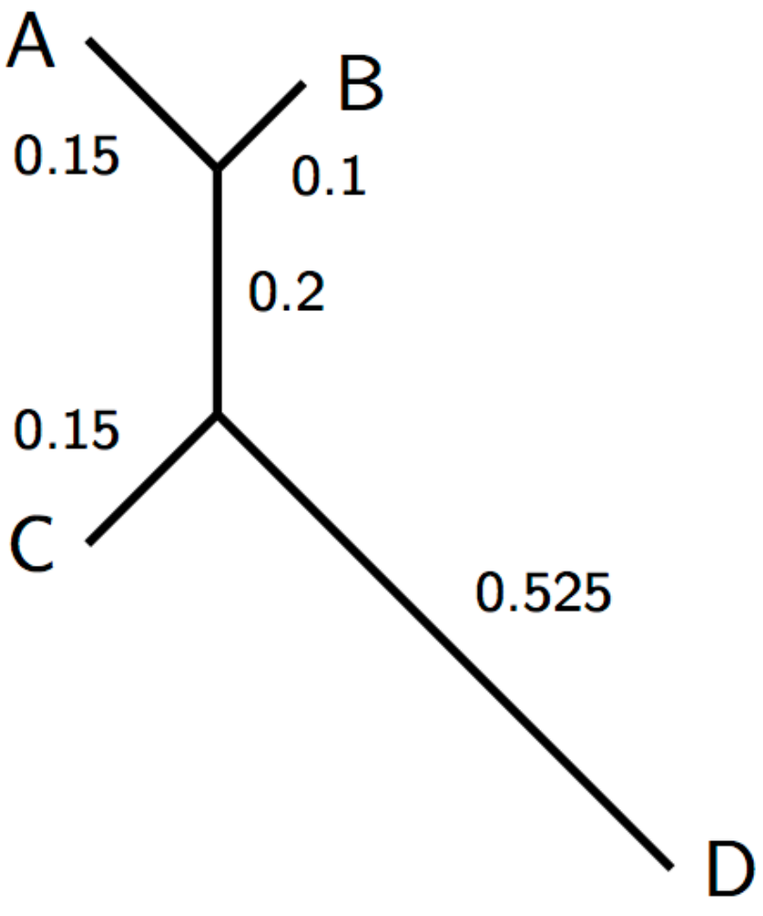
$$= \begin{pmatrix} 0.015 \\ 0.01 \\ 0.005 \\ 0.01 \\ 0.01 \\ 0.005 \end{pmatrix} \star$$

evolutionary rates
substitutions / site / unit time



time tree

Genetic distance = rate x time

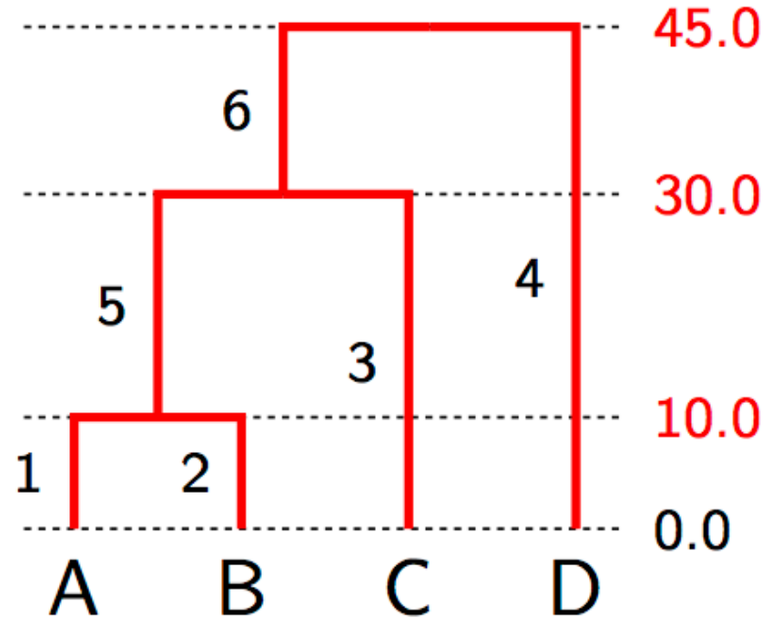


"substitution tree"

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

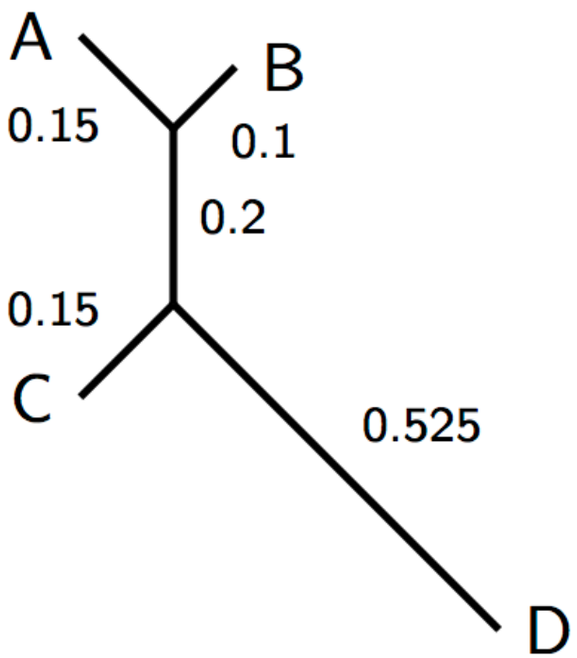
$$= \begin{pmatrix} 0.015 \\ 0.01 \\ 0.005 \\ 0.01 \\ 0.01 \\ 0.005 \end{pmatrix} \star$$

evolutionary rates
substitutions / site / unit time

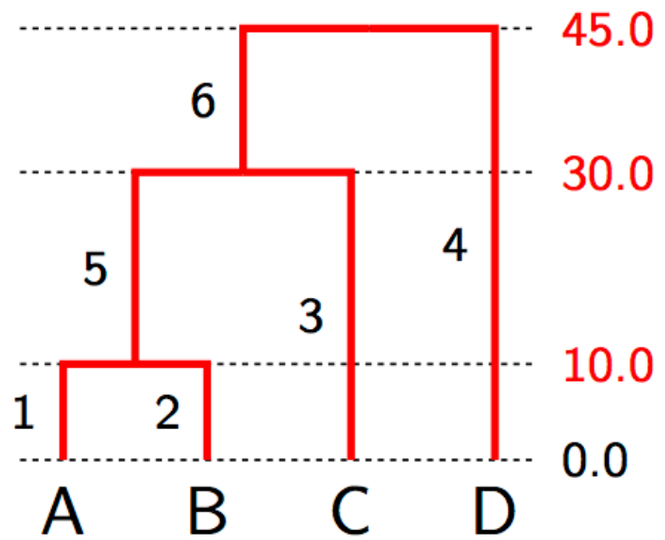


time tree

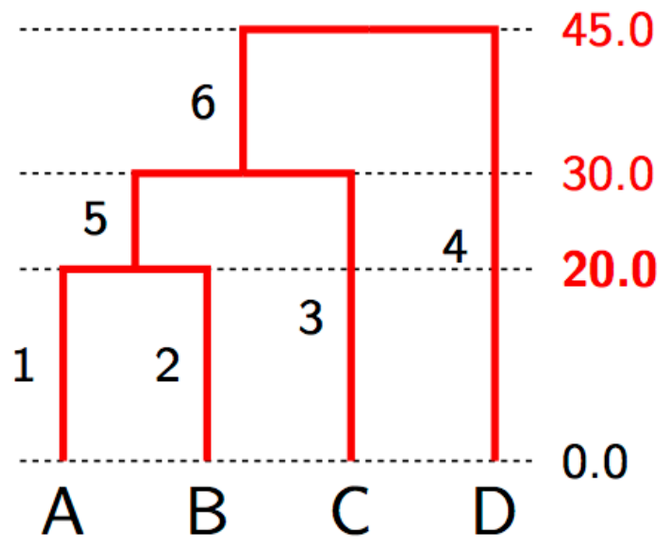
Non-identifiability of rates and times



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



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

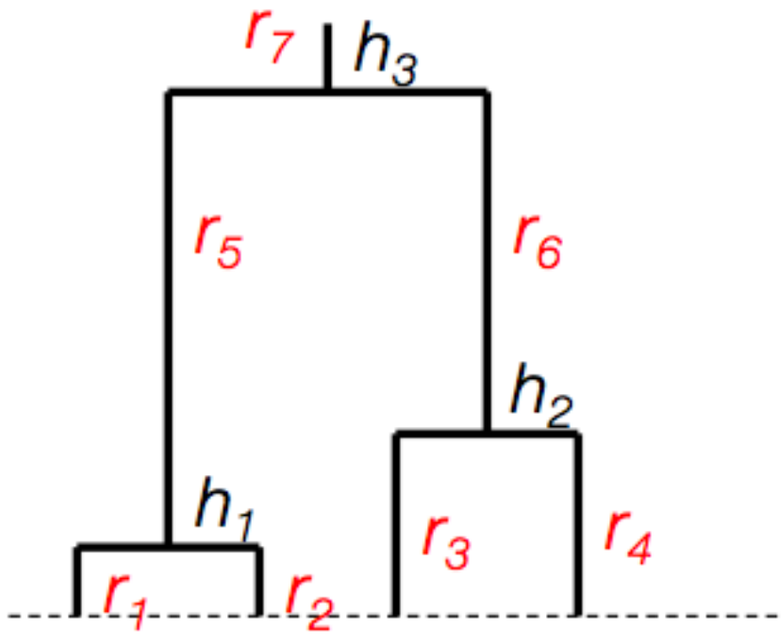


"substitution tree"

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

$r \sim Exp(\lambda)$

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

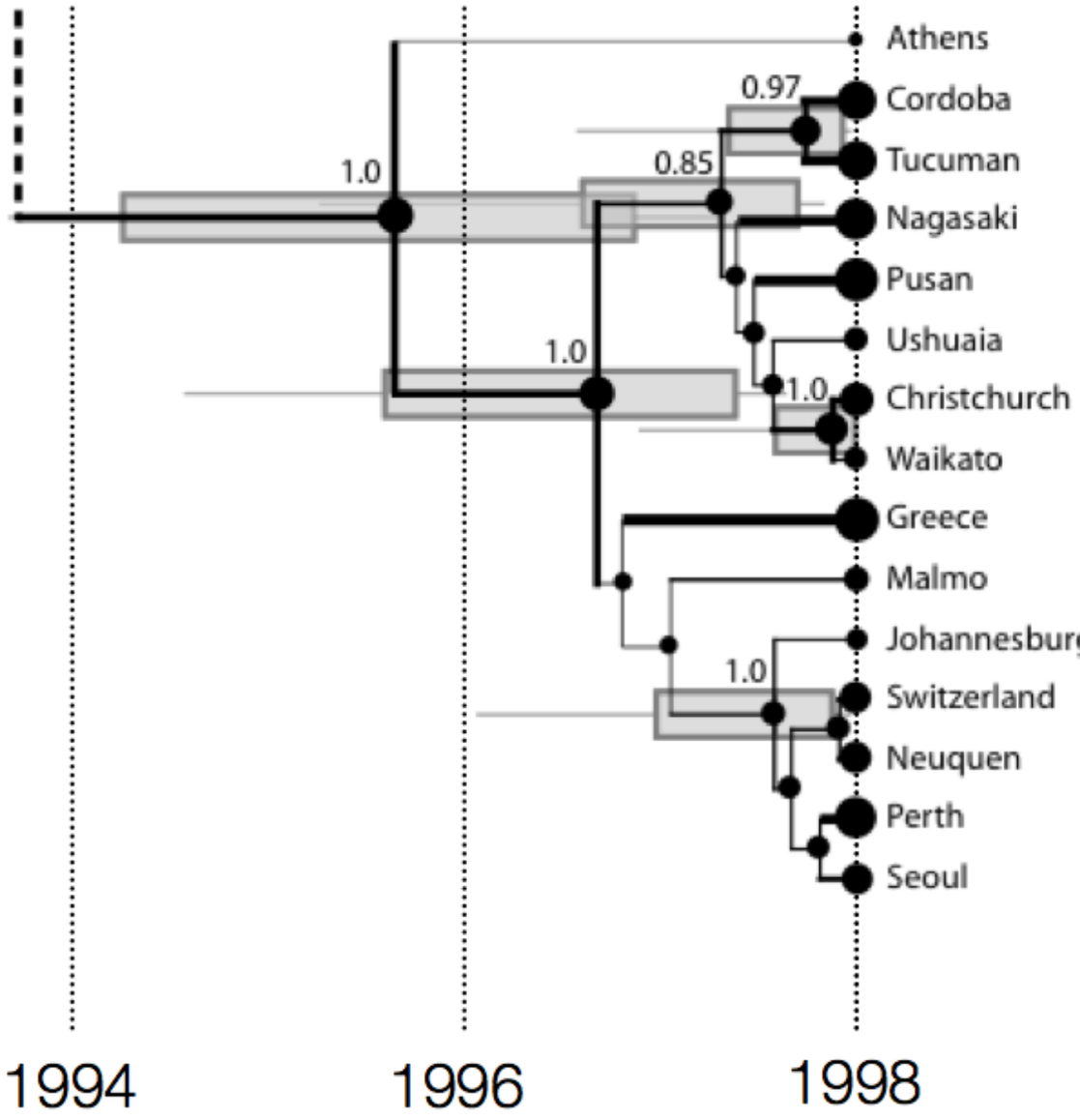
$r \sim Gamma(\alpha, \beta)$

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

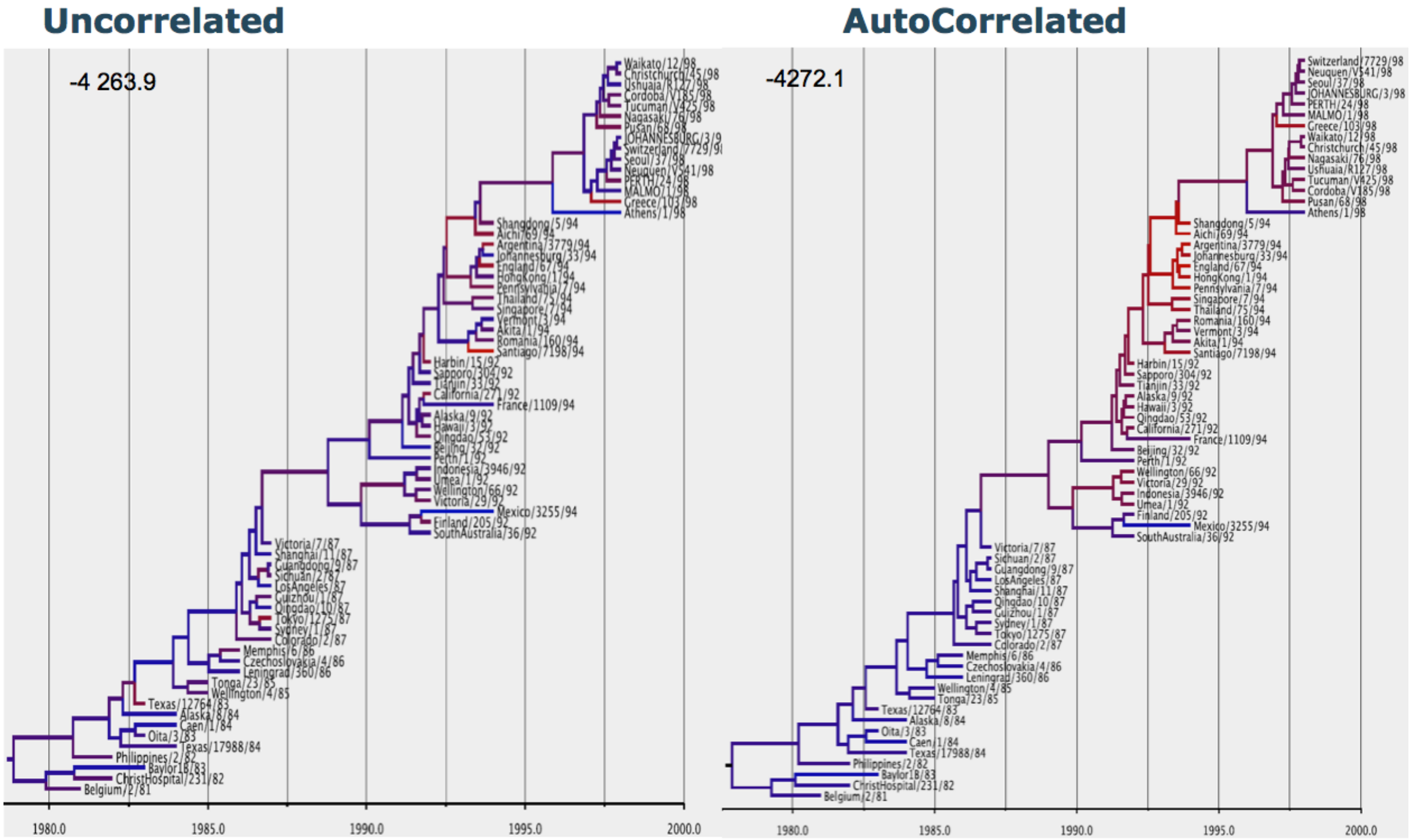
ML

Influenza A gene tree estimating using relaxed molecular clock

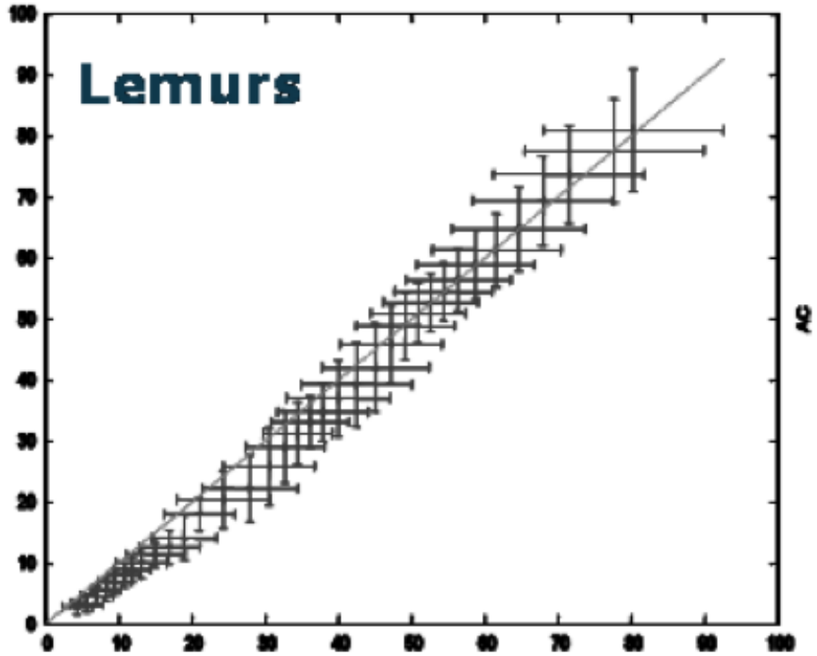


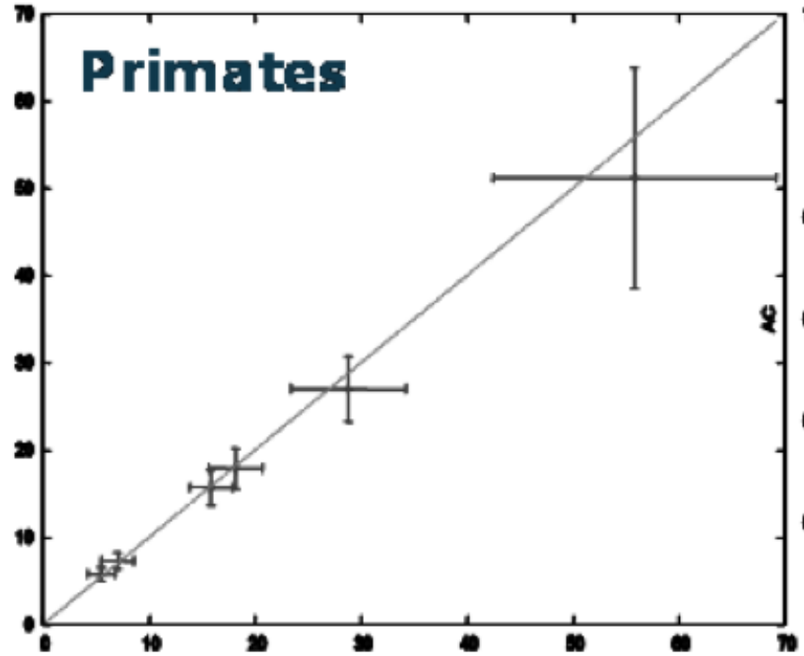
- Box-and-whisker plots show uncertainty in divergence times (only for splits with posterior probability > 0.5)
- Node size and branch thickness proportional to evolutionary rate.

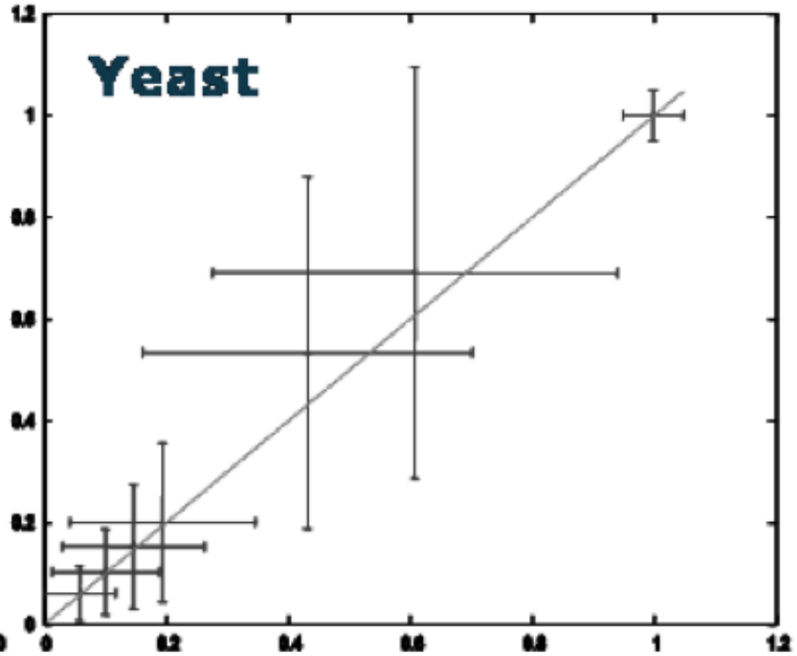
Influenza trees under different relaxed molecular clocks

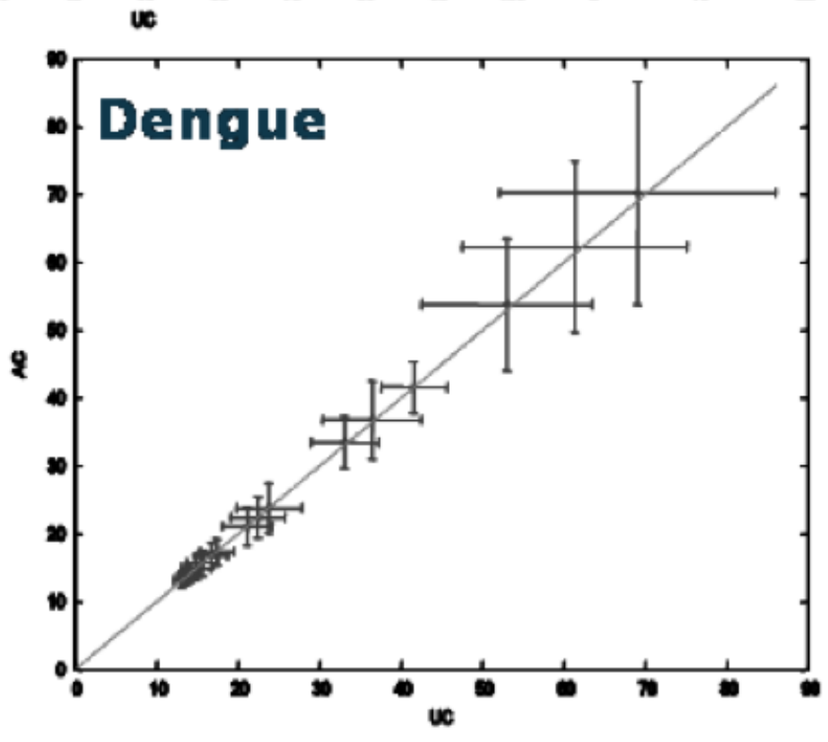


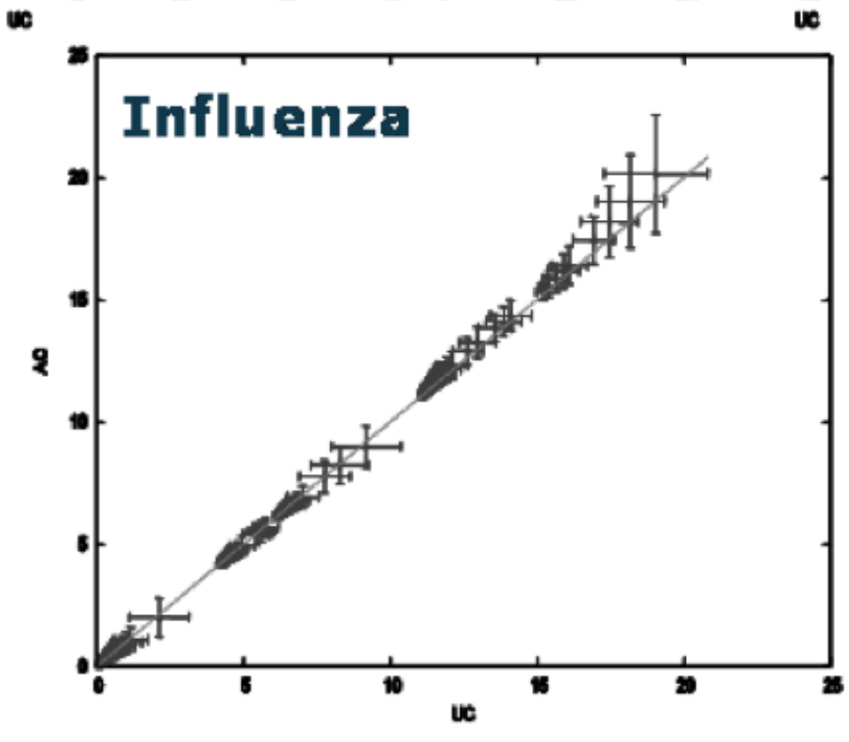
UC versus AC on five data sets



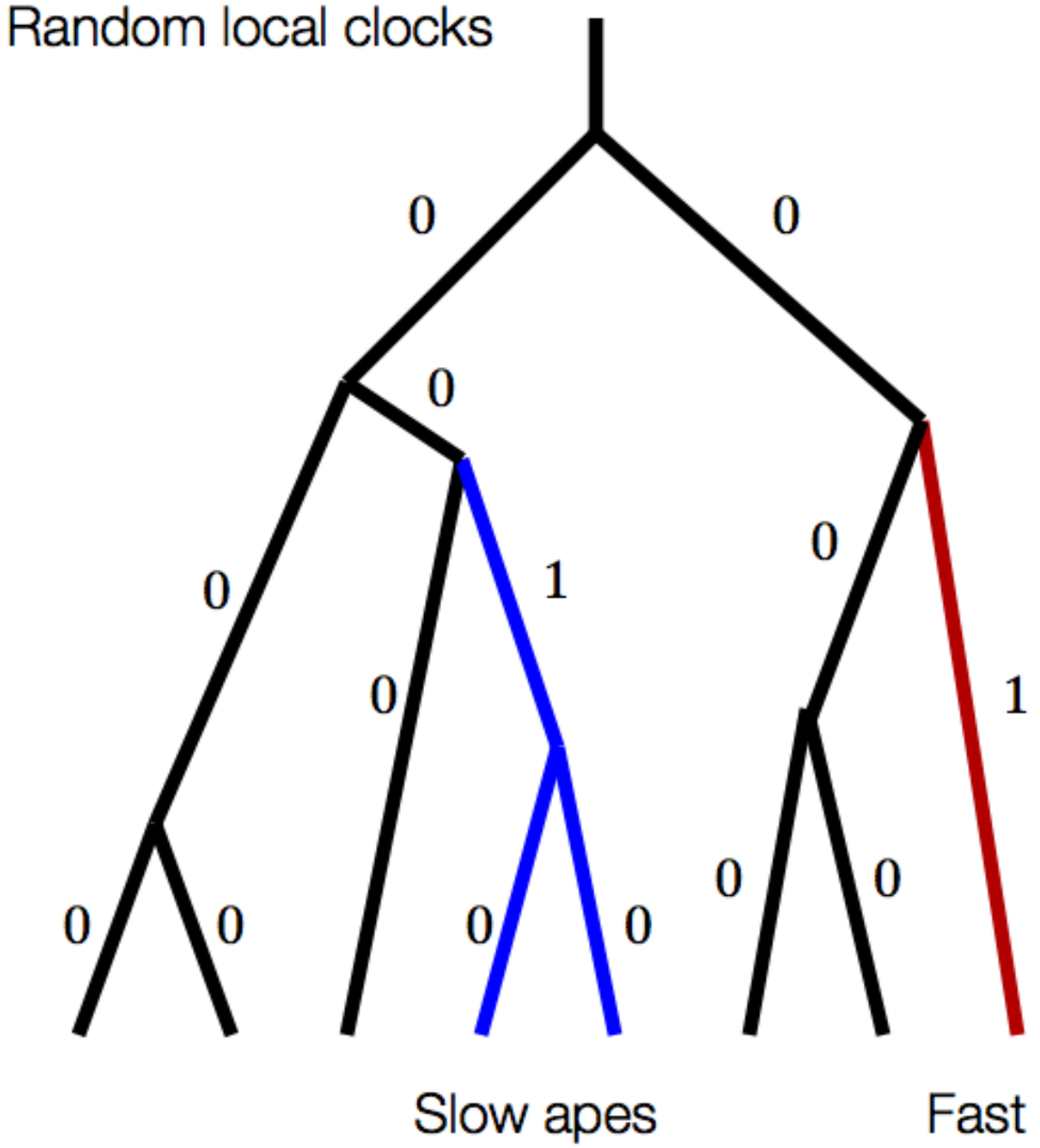






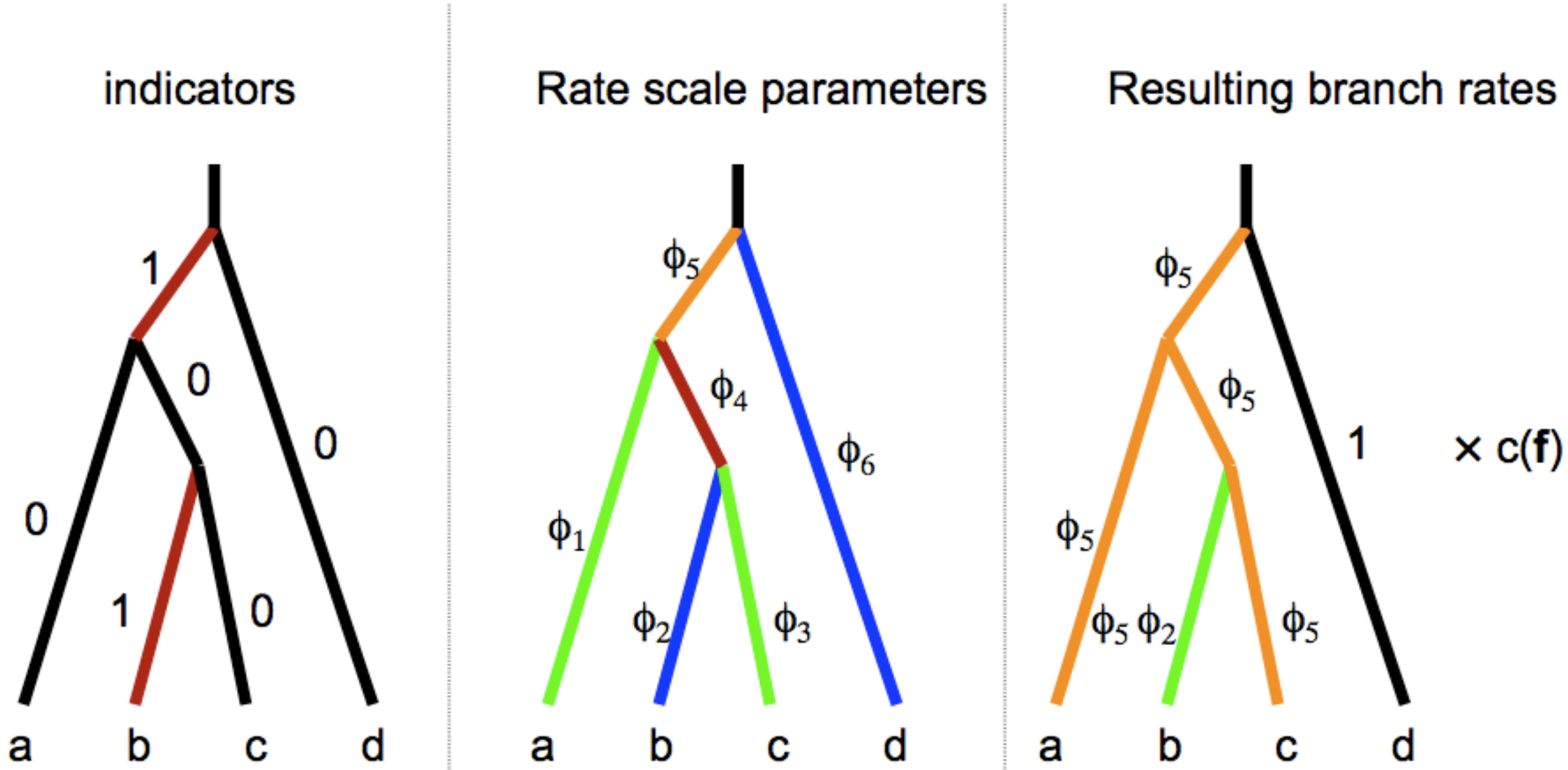


Random local molecular clocks



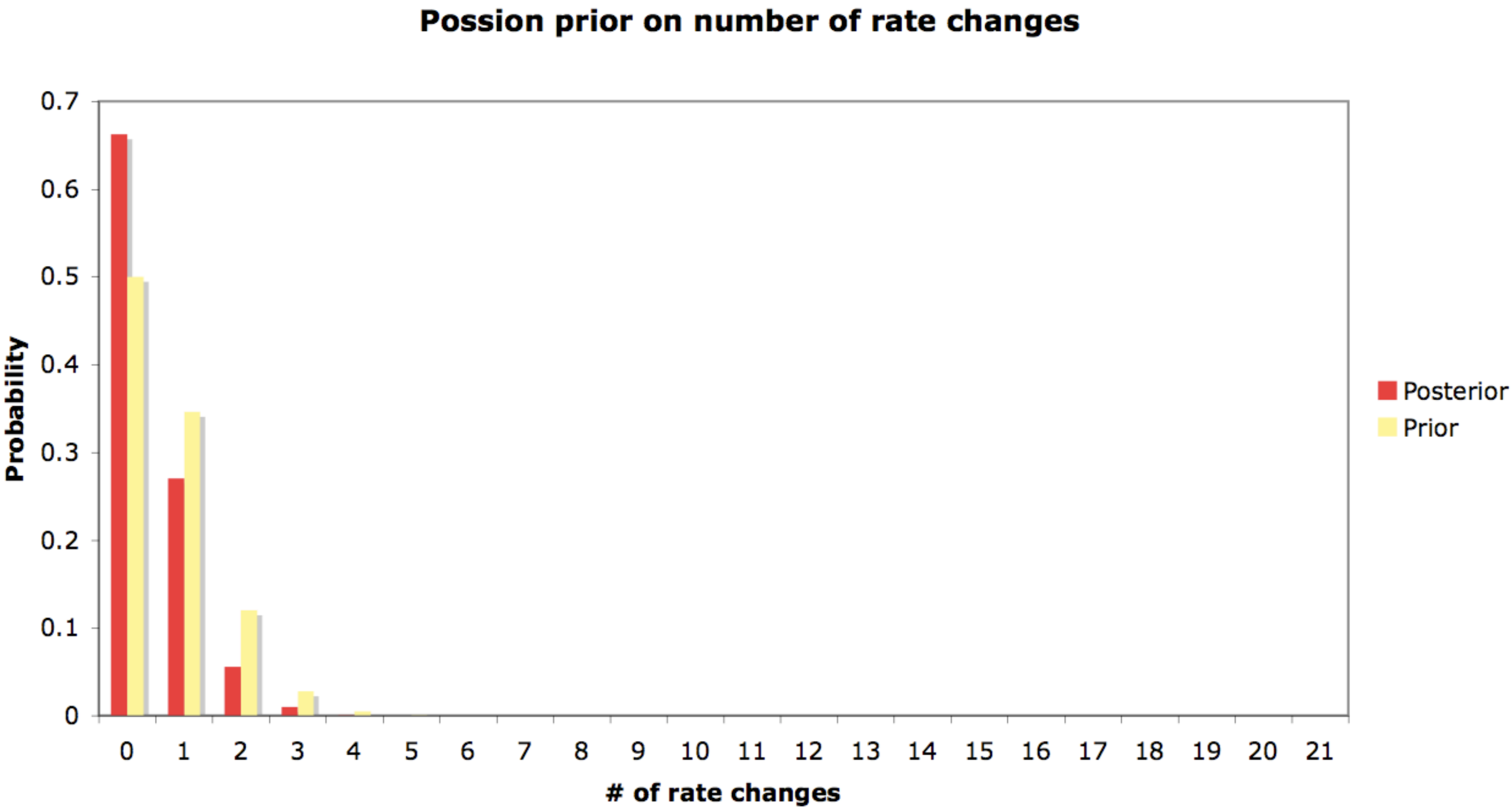
2^{2n-2} local molecular clock models

Random local molecular clocks

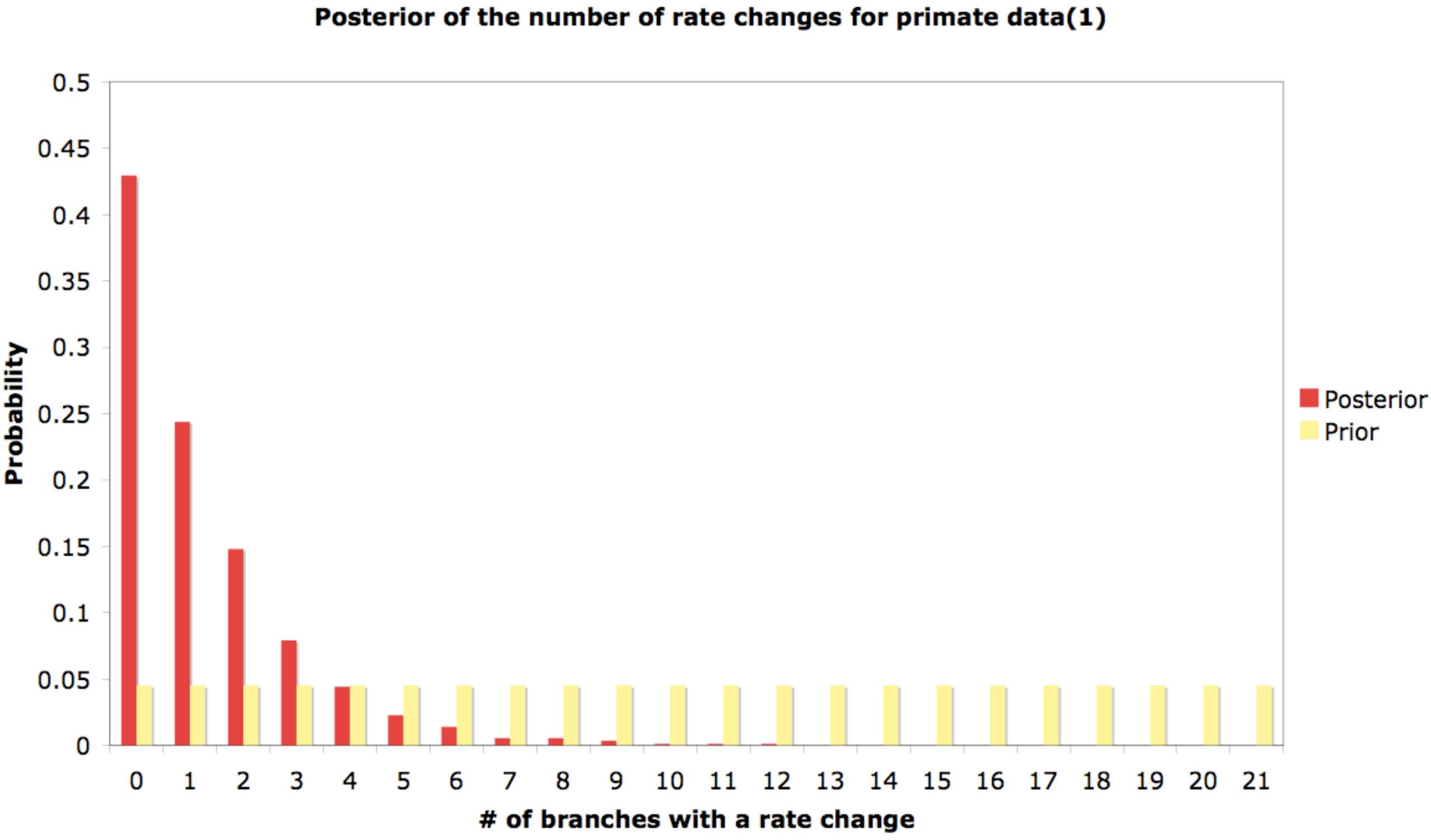


Red/Orange fast, Green/Blue slow

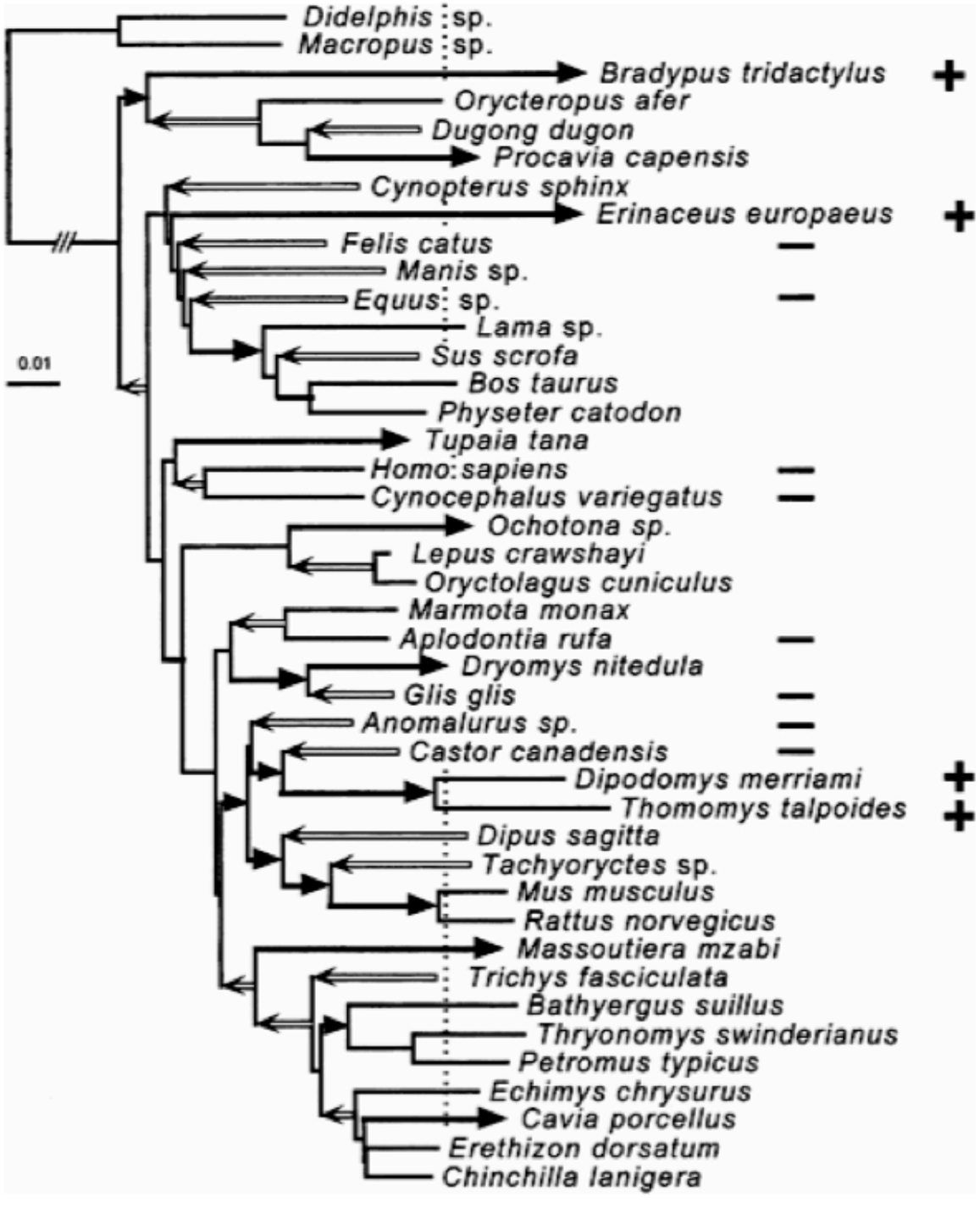
Primate data set Poisson prior



Primate data set - uniform prior



Rodents

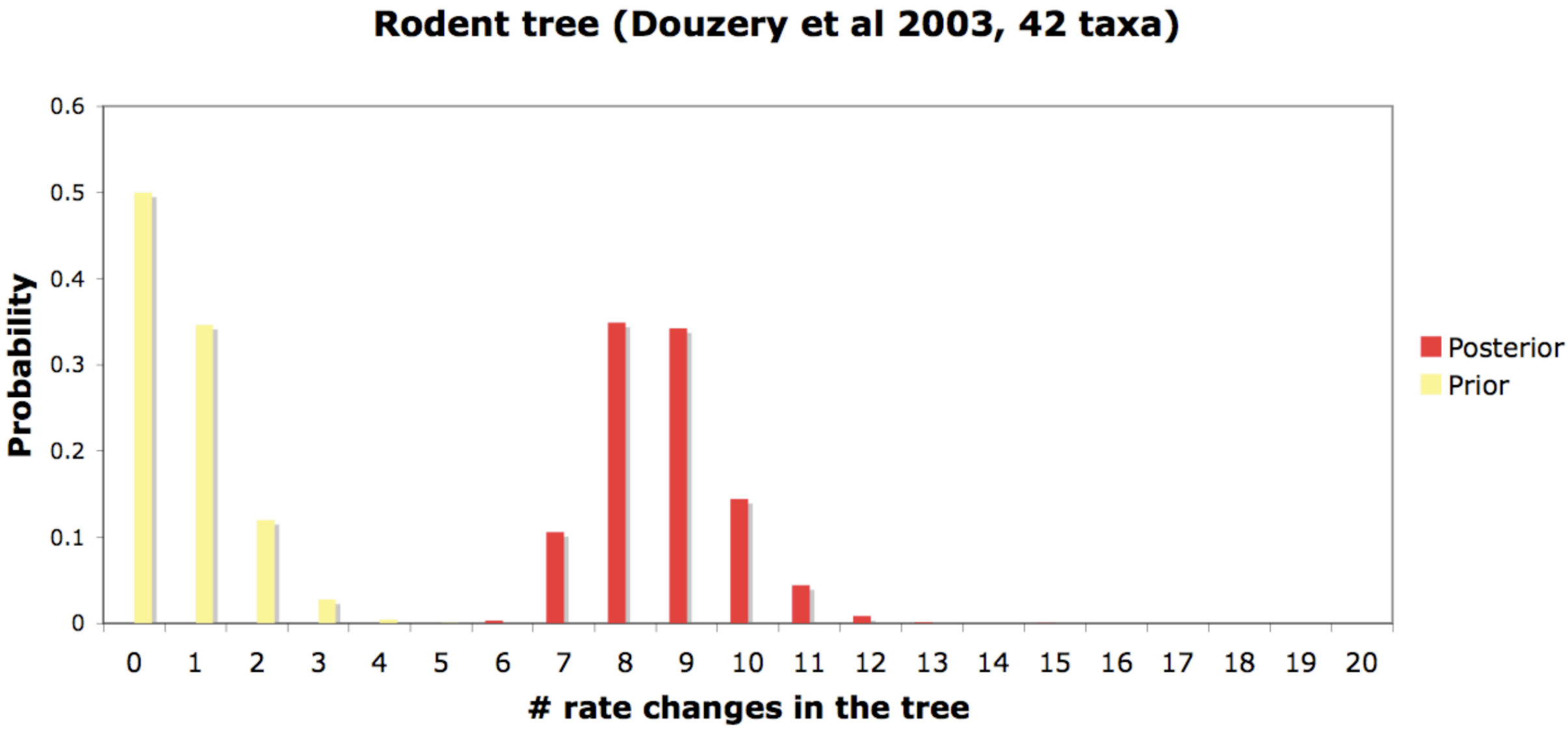


82 branches

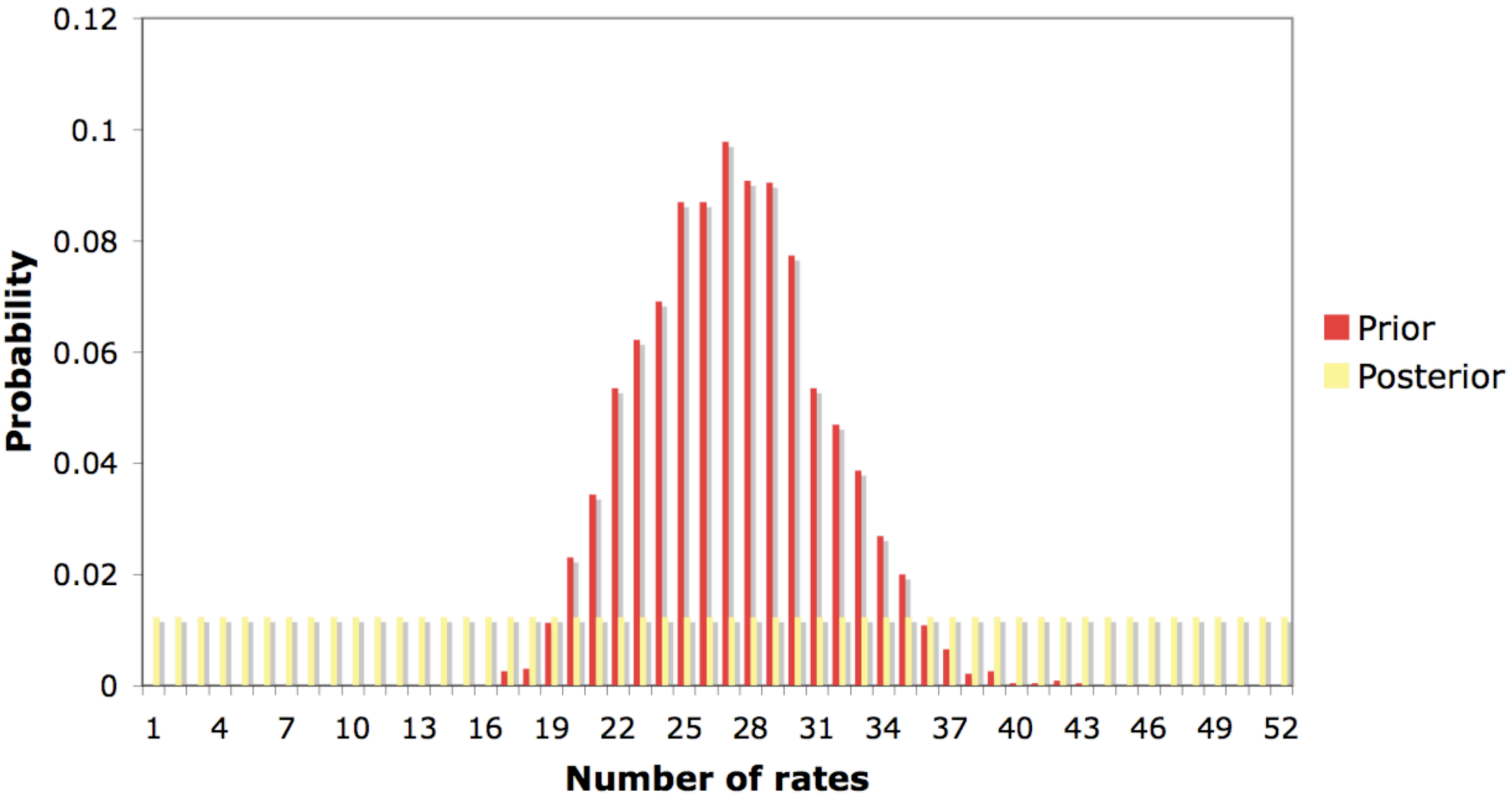
38 rate changes according to Douzery et al 2003

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).

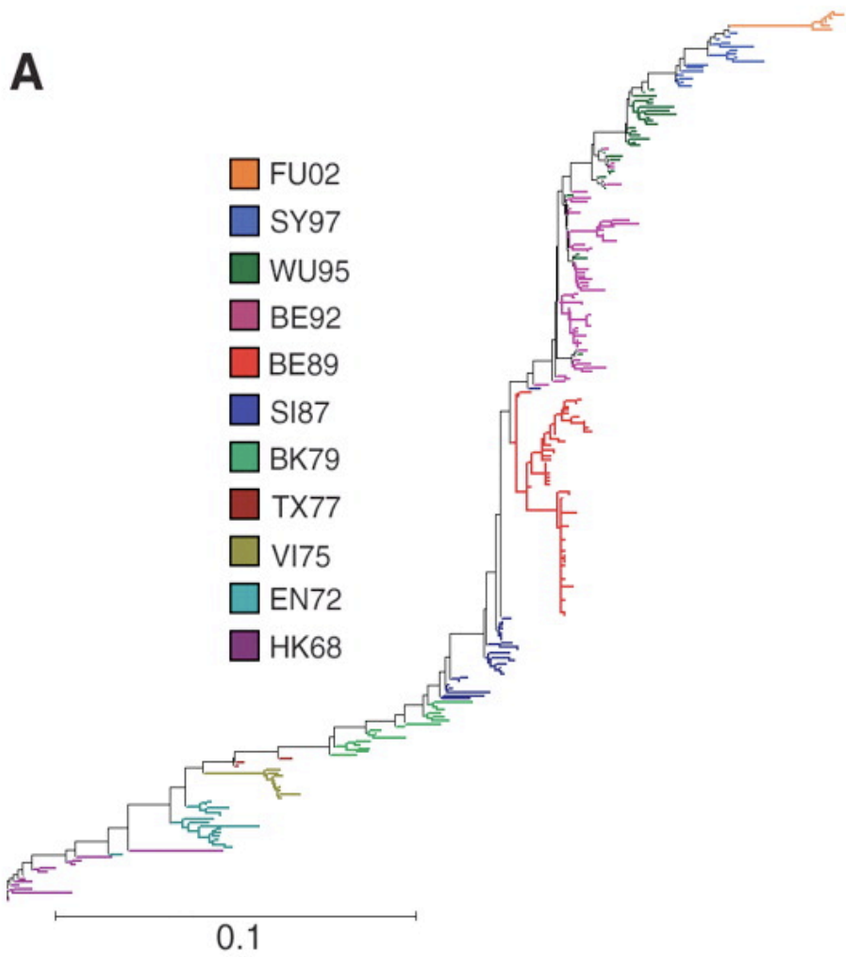
Rodent data set (Poisson prior on # changes)



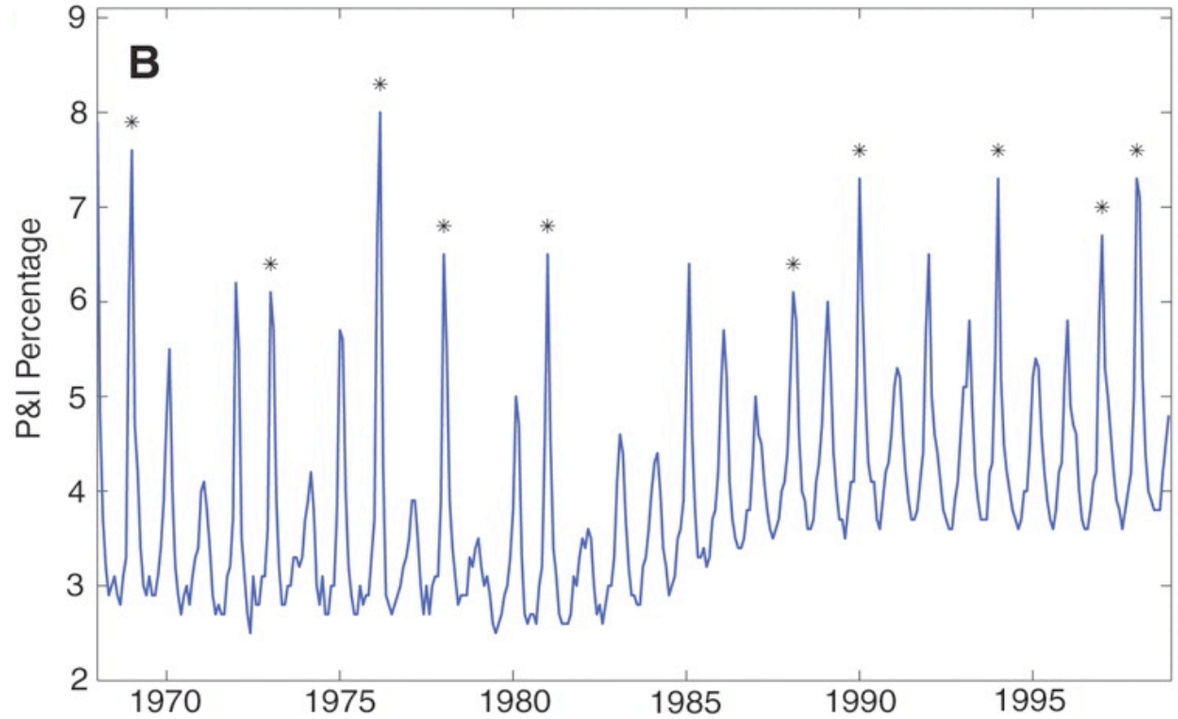
Rodent data set (uniform prior on # changes)



Phylodynamics



Statistical phylogenetics and population genetics models



Mathematical epidemiology, non-linear dynamical models

Goal: the integration of immunological, genomic and epidemiological data in a single coherent predictive model.

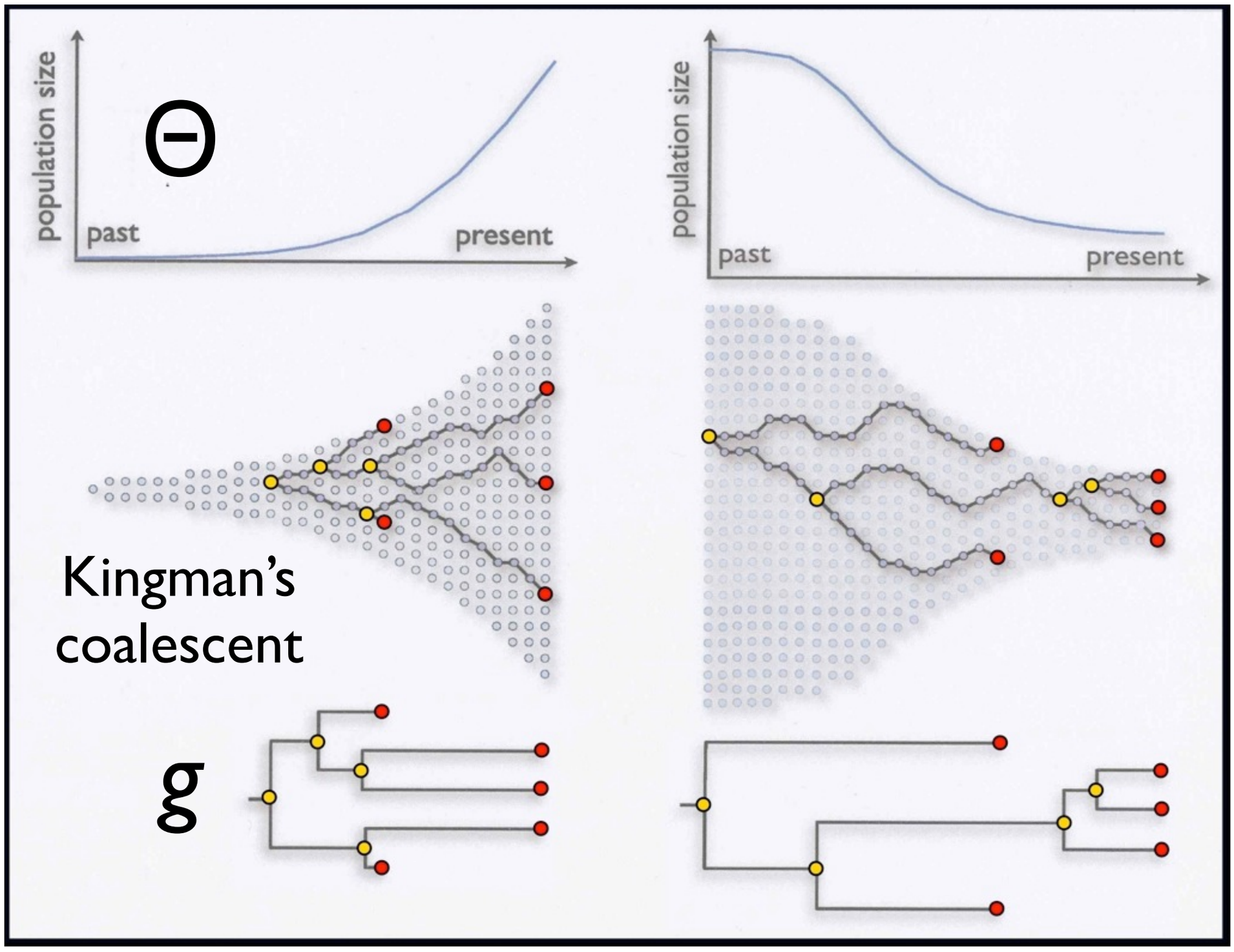
Coalescent tree priors

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)



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

The coalescent

Data: a small genetic sample from a large background population.

The coalescent

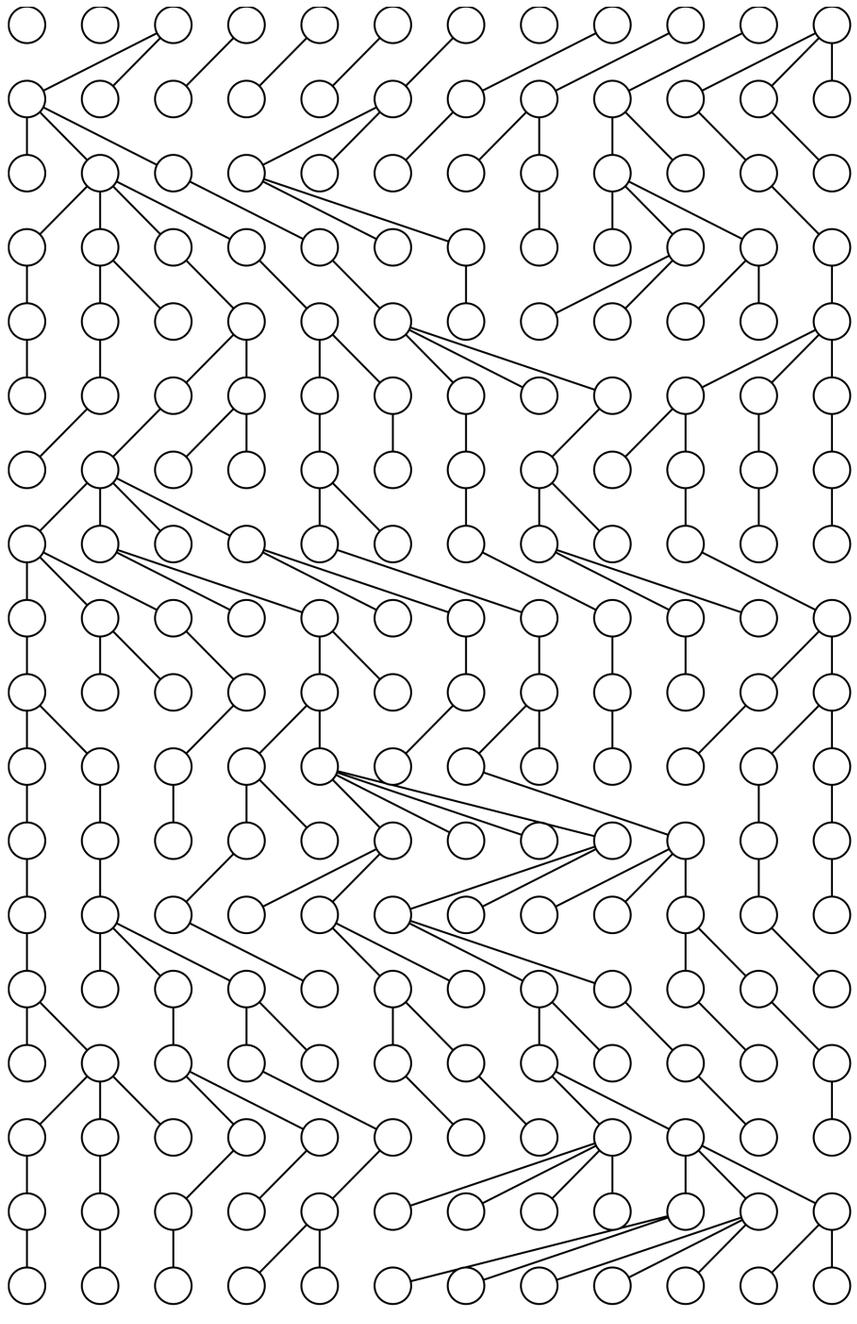
- is a model of the ancestral relationships of a sample of individuals taken from a larger population.
- describes a probability distribution on ancestral genealogies (trees) given a population history, N(t).
- Therefore the coalescent can convert information from ancestral genealogies into information about population history and vice versa.
- a model of ancestral genealogies, not sequences, and its simplest form assumes neutral evolution.
- can be thought of as a prior on the tree, in a Bayesian setting.

Theoretical population genetics

Most of theoretical population genetics is based on the idealized Wright-Fisher model of population which assumes

- Constant population size N
- Discrete generations
- Complete mixing

For the purposes of this presentation the population will be assumed to be haploid, as is the case for many pathogens.

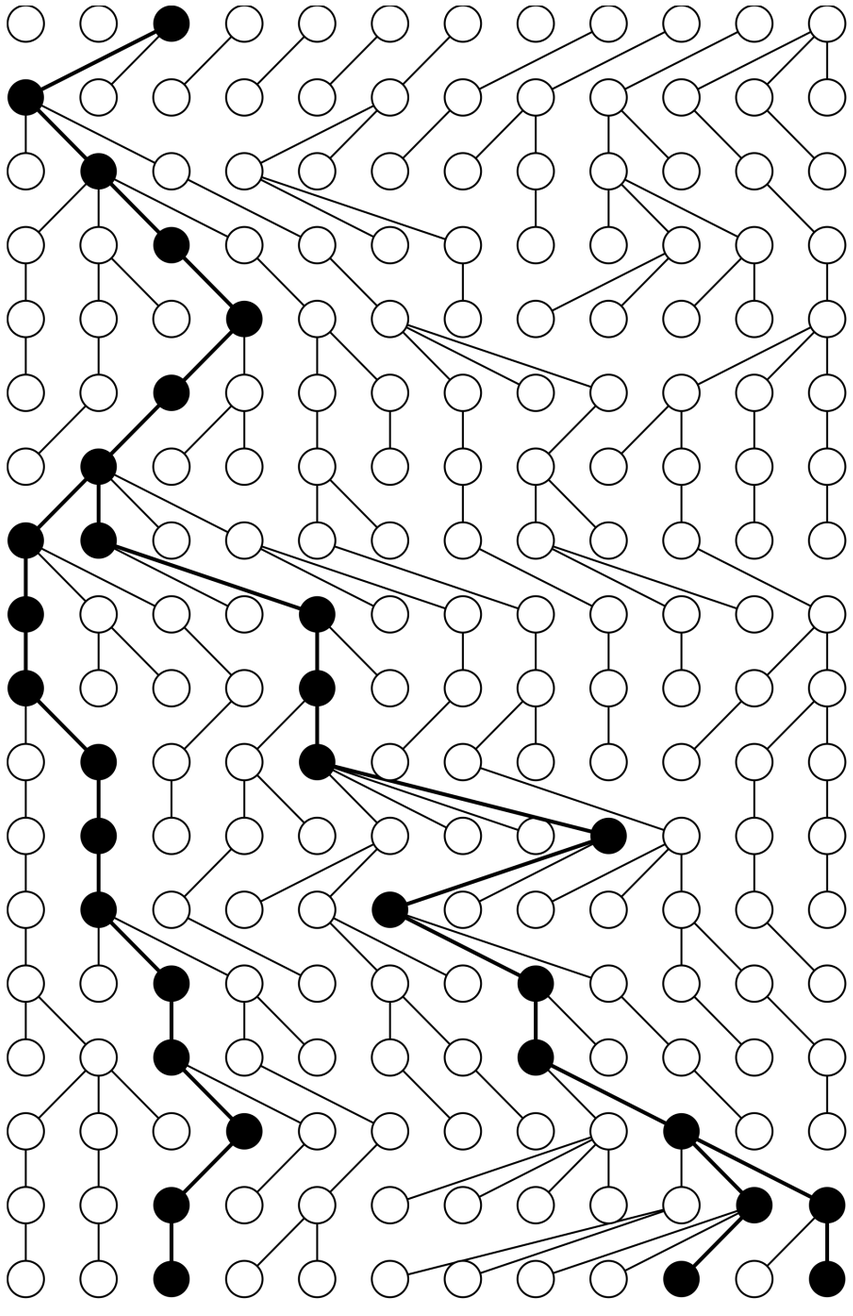


Kingman’s n-coalescent

Consider tracing the ancestry of a sample of k individuals from the present, back into the past.

This process eventually coalesces to a single common ancestor (concestor) of the sample of individuals.

Kingman’s n-coalescent describes the statistical properties of such an ancestry when k is small compared to the total population size N.



The coalescent of two lineages

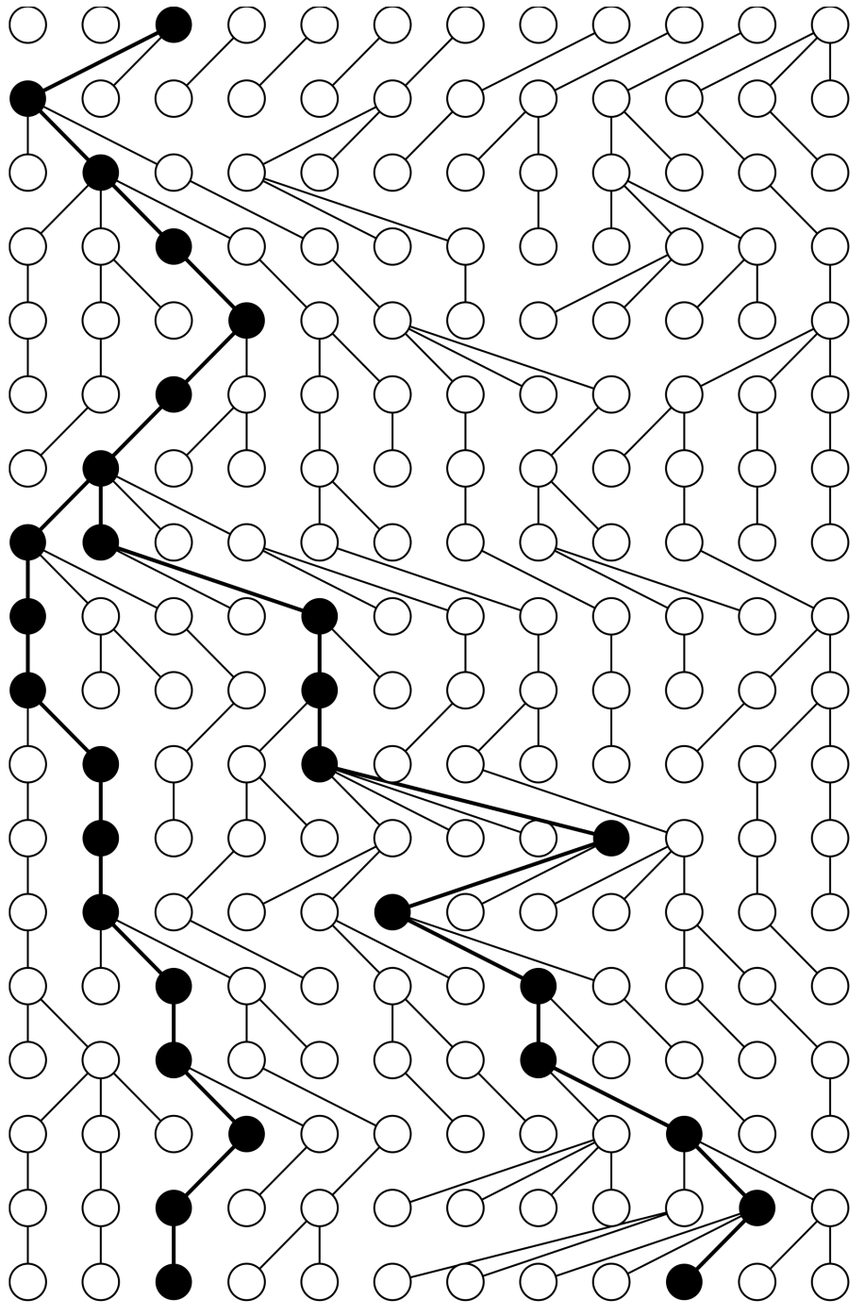
First, consider two random members from a population of fixed size N.

By perfect mixing, the probability they share a concestor in the previous generation is 1/*N.*

The probability the concestor is t generations back is

$$\Pr(t) = \frac{1}{N}(1 - \frac{1}{N})^{t-1}.$$

It follows that *g*=*t*−1, has a geometric distribution with a success rate of *λ*=1/*N*, and so has mean N and variance of $N^3/(N-1)$.



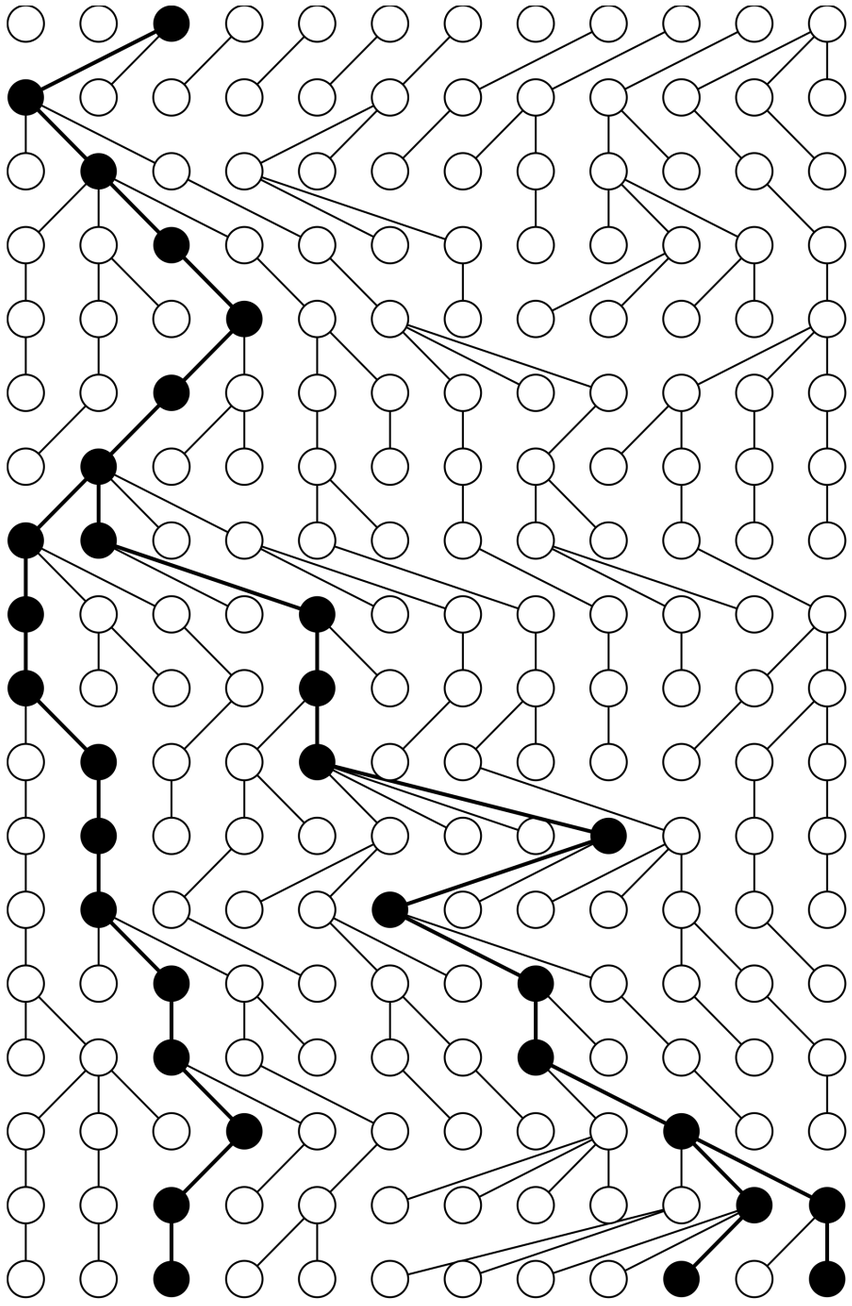
The coalescent of k lineages

With k lineages the time to the first coalescence is derived in the same way, only now there are $\binom{k}{2}$ possible pairs that may coalesce, resulting in a success rate of $\lambda = \binom{k}{2}/N$.

The mean time to first coalescence (t_k) is:

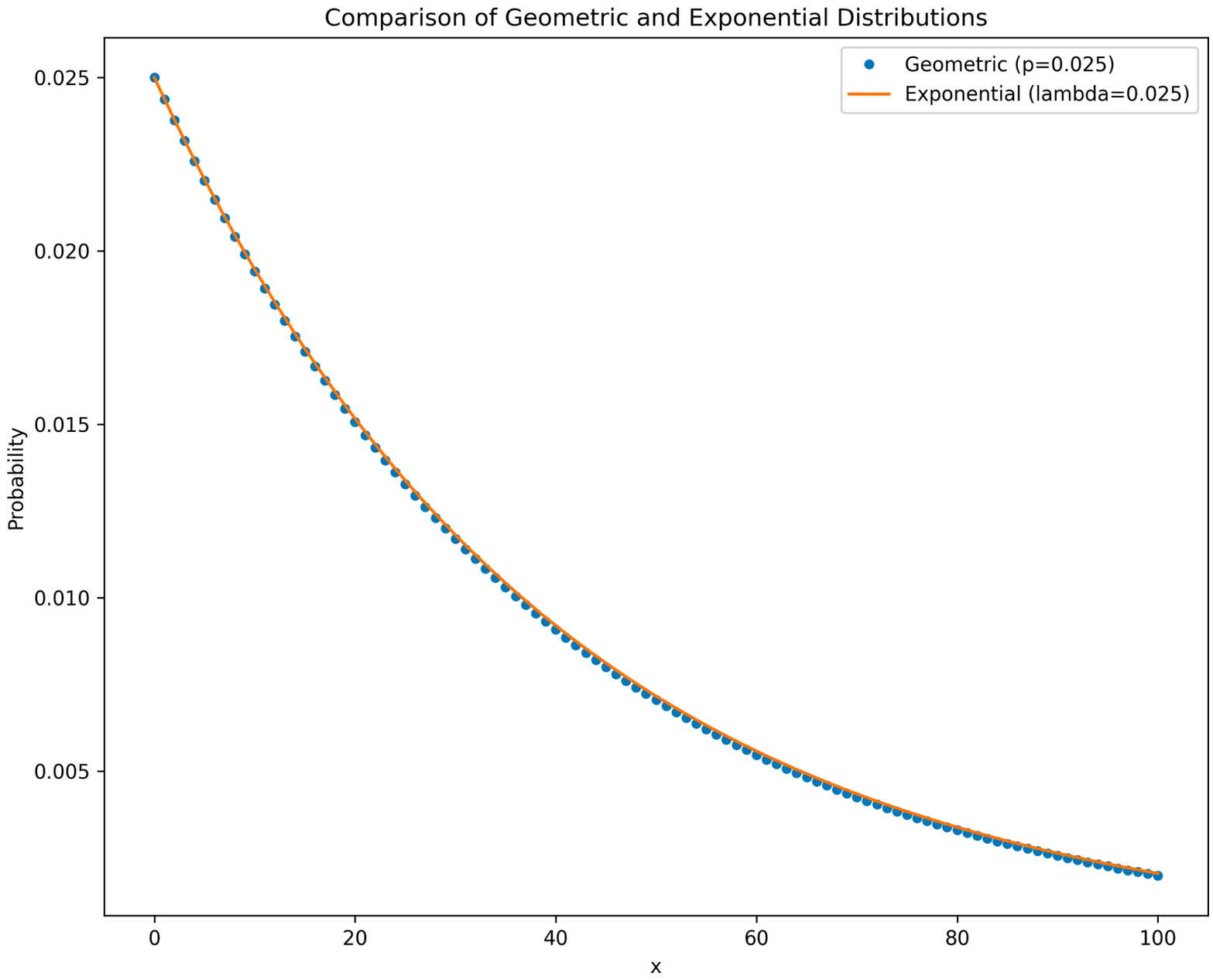
$$E[t_k] = \frac{N}{\binom{k}{2}}.$$

This implicitly assumes that N is much larger than $O(k^2)$, so that the probability of two coalescent events in the same generation is small.



The exponential distribution

The **geometric distribution converges to the exponential distribution** as the population size becomes larger and the rate of coalescence becomes smaller.



The coalescent is a *diffusion approximation*

Kingman (1982) showed that as N grows the coalescent process converges to a continuous-time Markov chain.

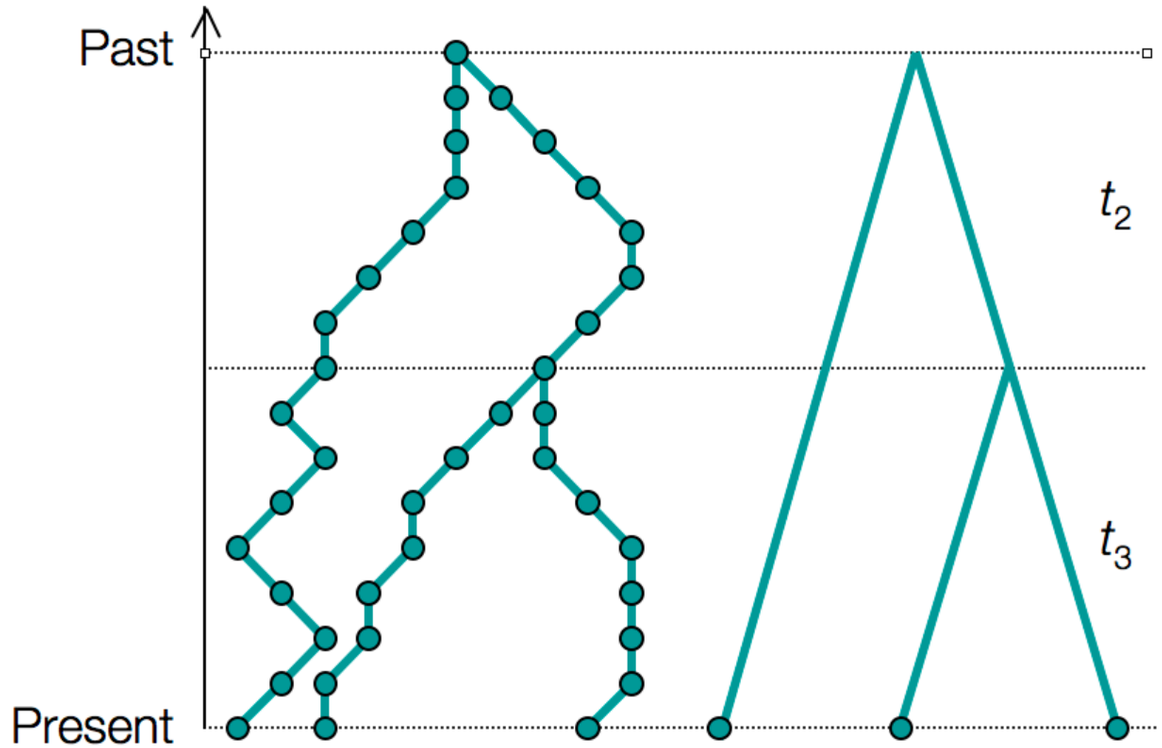
$\lambda = \binom{k}{2}/N$ is the rate of coalescence, i.e. the probability of coalescing a pair from k lineages on a short time interval Δt *is* $O(\lambda\Delta t)$.

Unsurprisingly the solution turns out to be the exponential distribution:

$$f(t_k) = \frac{\binom{k}{2}}{N} \exp(-\frac{\binom{k}{2} t_k}{N})$$

or more relevant to a tree hypothesis, for a *specific* pair of lineages:

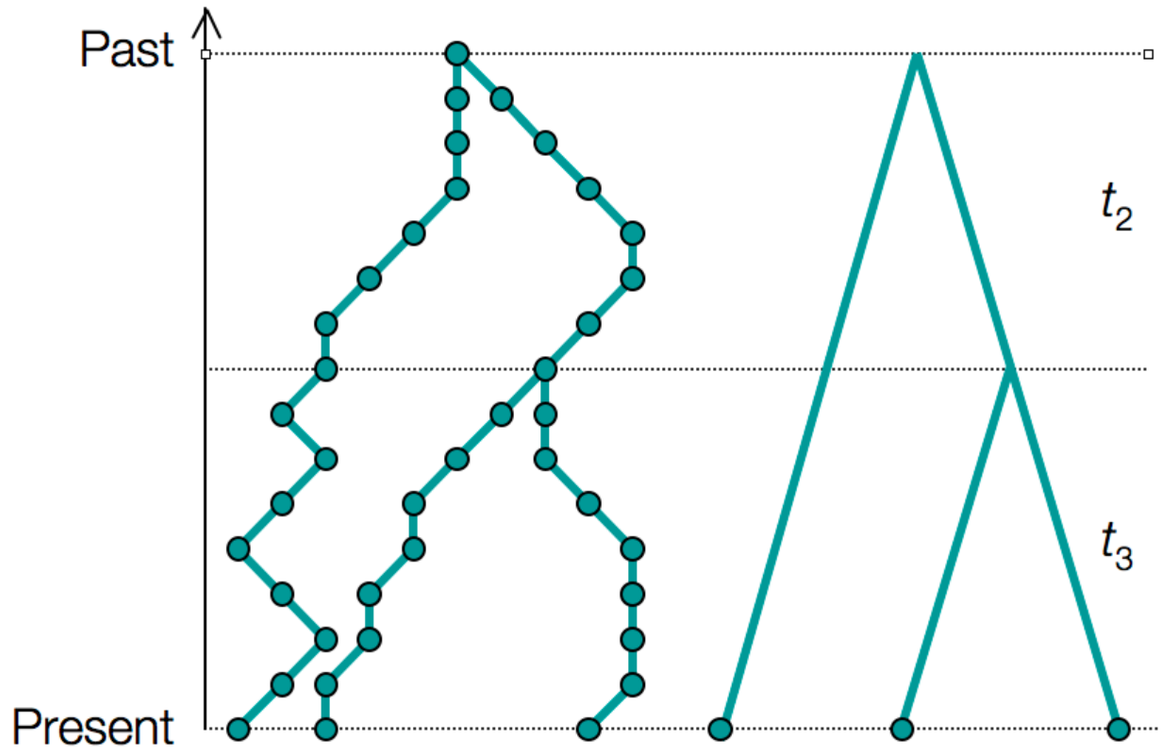
$$f(t_k) = \frac{1}{N} \exp(-\frac{\binom{k}{2} t_k}{N})$$



The coalescent density for a genealogy

For a genealogy g with coalescent times $\mathbf{t} = \{t_2, t_3, \ldots, t_n\}$ we can write the probability density of the genealogy:

$$P(g \mid N) = \frac{1}{N^{n-1}} \prod_{k=2}^{n} \exp\left(-\frac{\binom{k}{2} t_k}{N} \right)$$



The coalescent density with varying population size

The generalization of the coalescent for the case where the population size changes over time, N=N(t) is given by Griffiths and Tavare (1994).

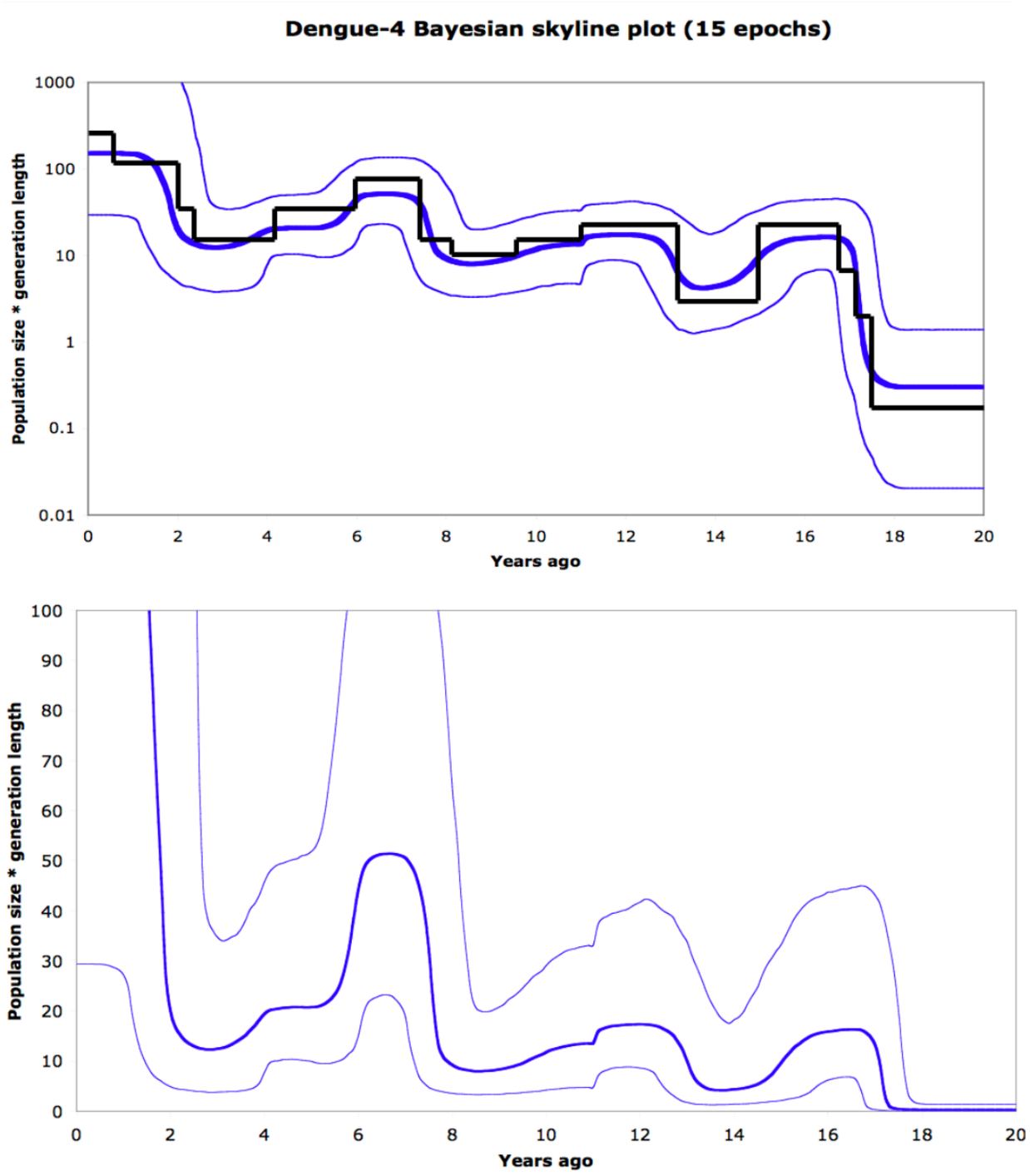
They showed that the coalescent density for the first coalescence event being at time t in the past given n lineages is:

$$f(t) = \frac{1}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(x)} dx \right)$$

So as long as the coalescent intensity function $\frac{1}{N(t)}$ is integrable with respect to t, then the coalescent density of a tree can be computed for population size function N(t)

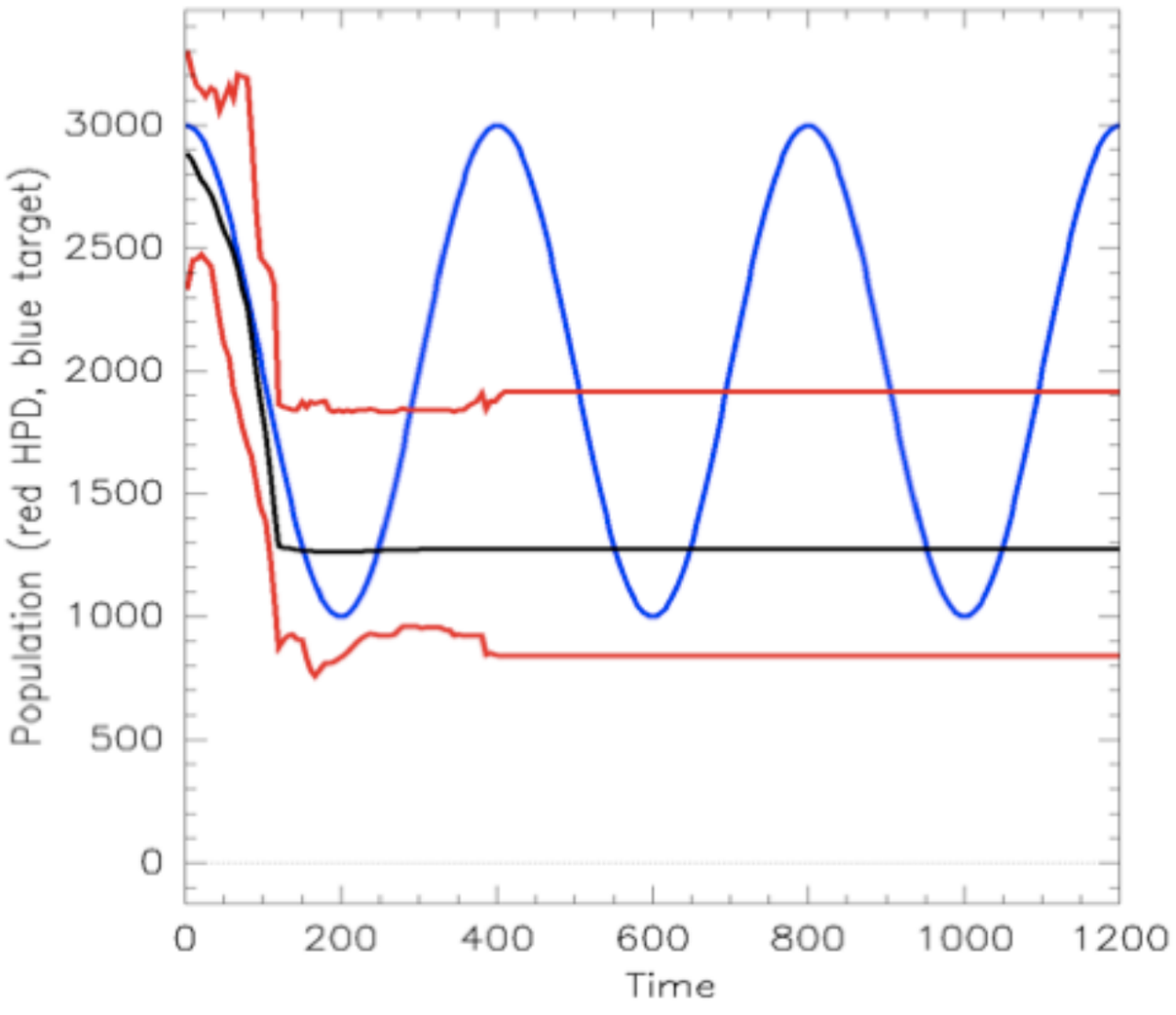
Bayesian skyline (Drummond et al, 2005)

The Bayesian skyline plot estimates a demographic function that has a certain fixed number of steps (in this example 15) and then integrates over all possible positions of the break points, and population sizes within each epoch.

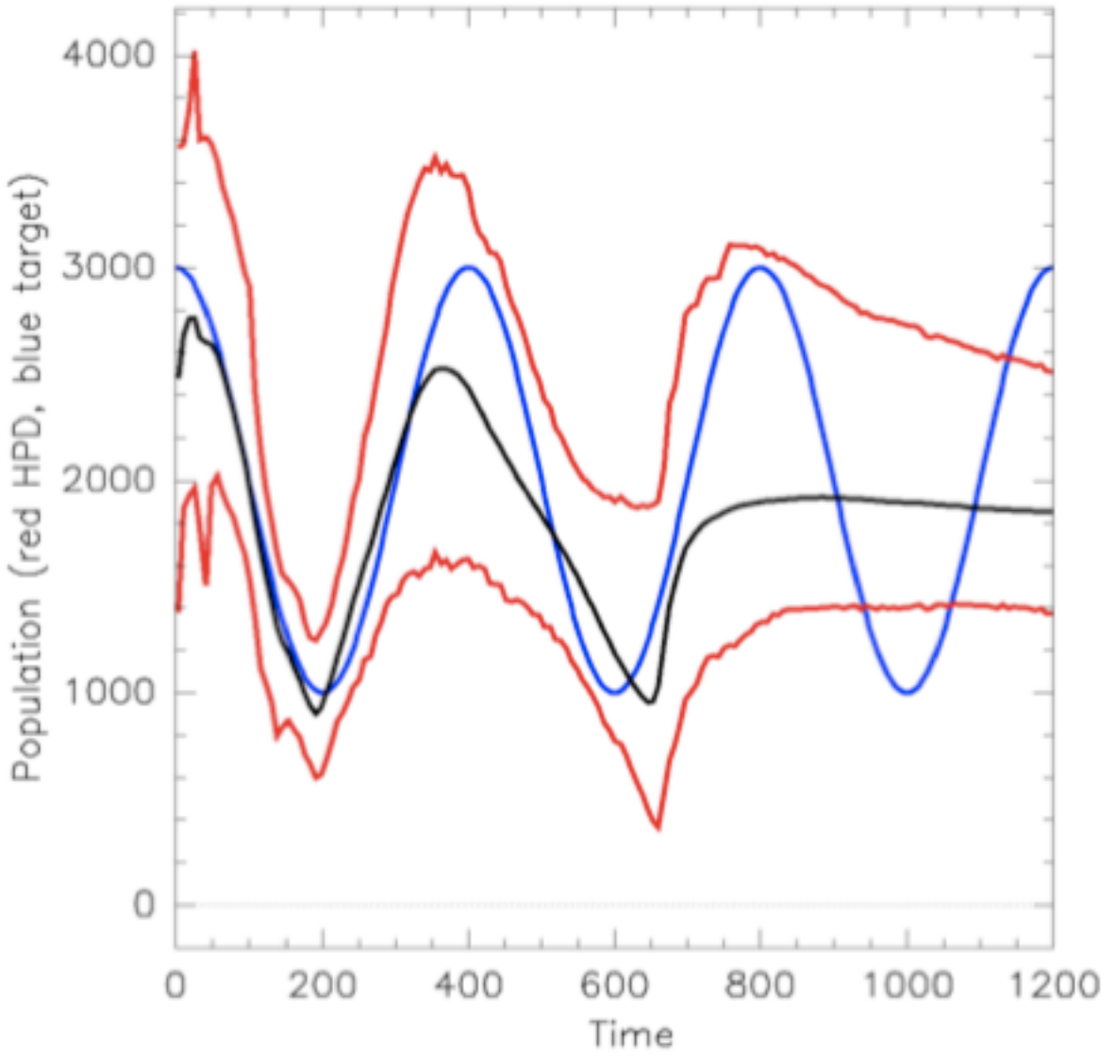


(Extended) Bayesian skyline plot

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



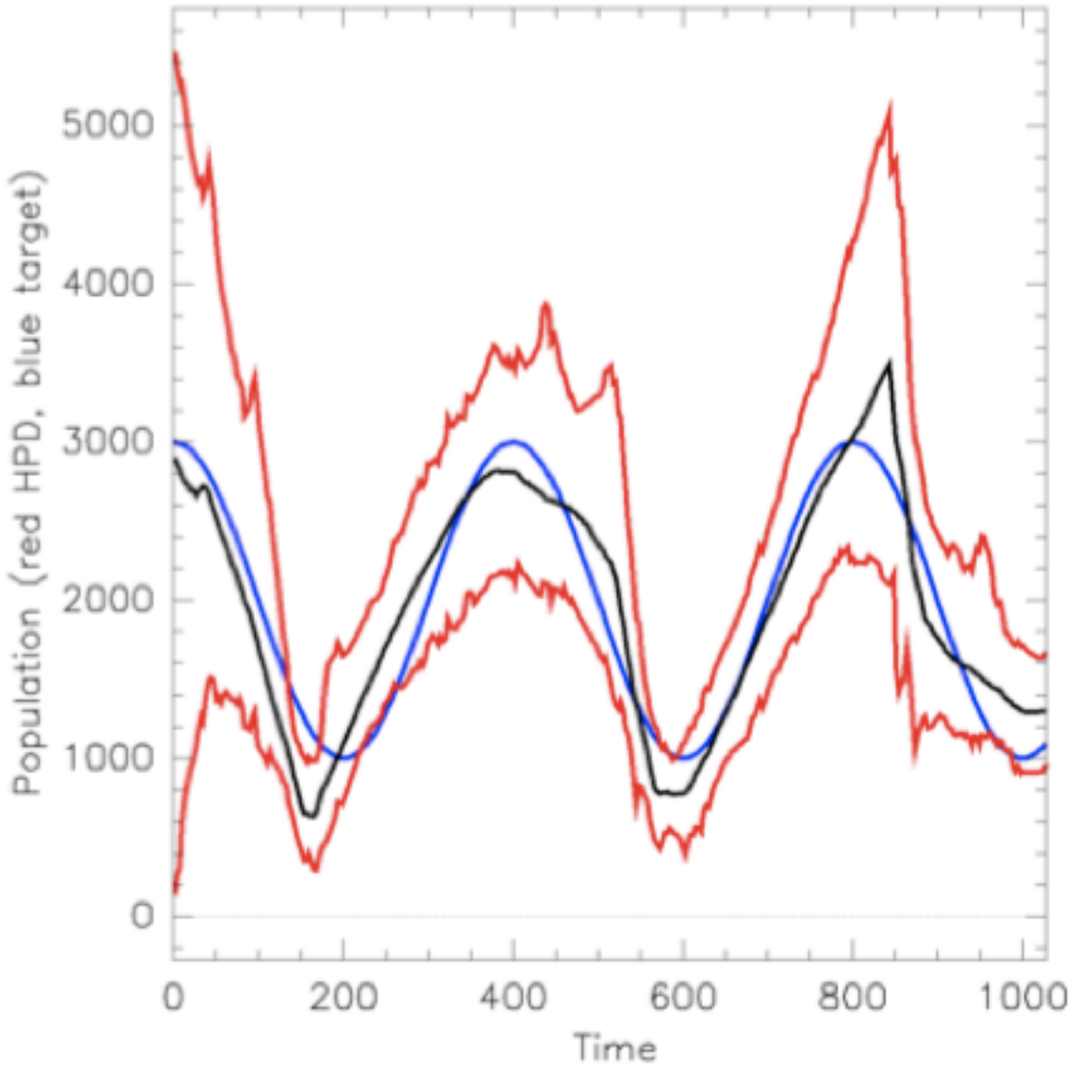
one gene sampled from 480 sampled individuals (480 gene sequences in total)



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

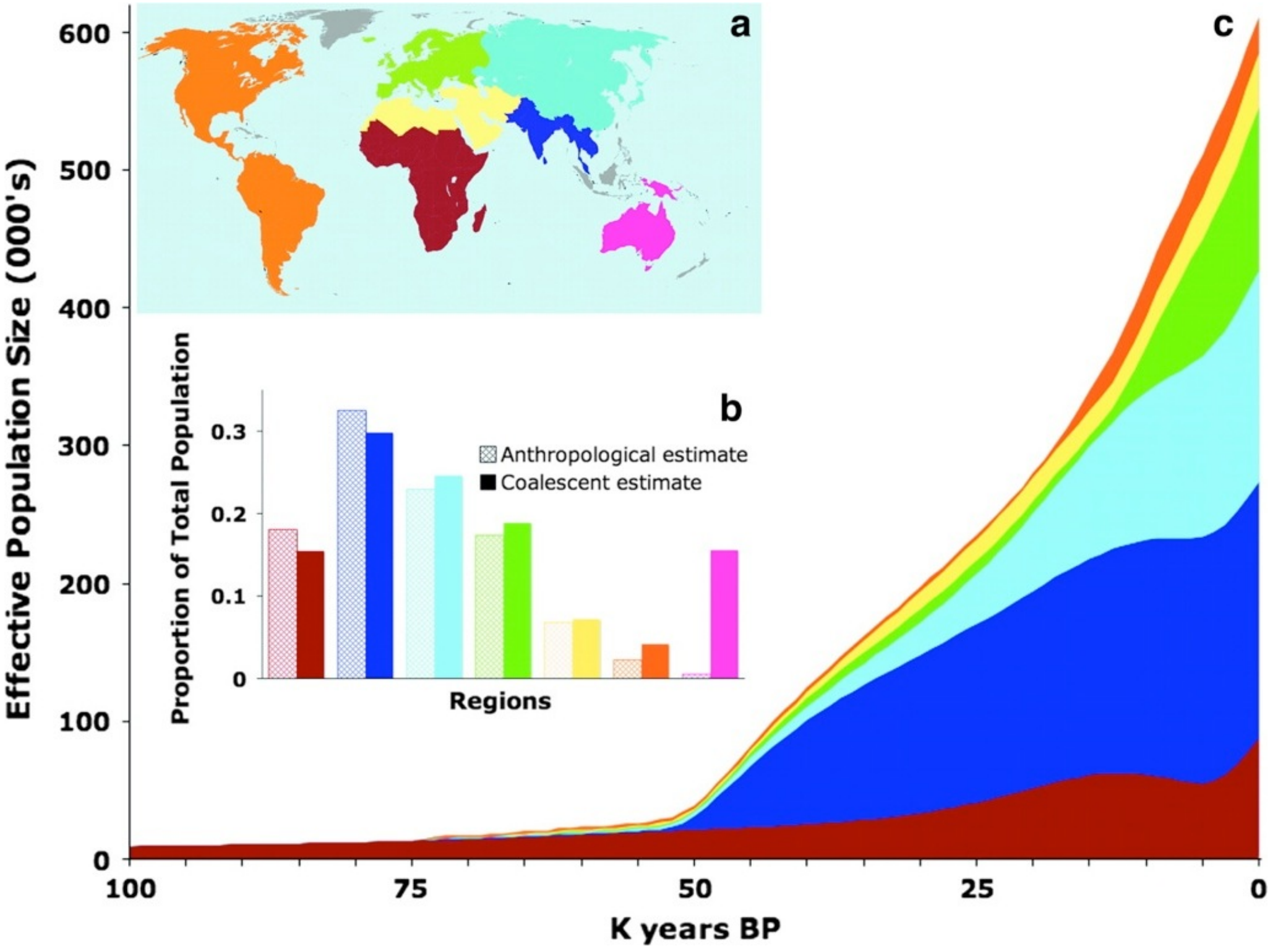
(Extended) Bayesian skyline plot

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



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

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

Bayesian skyline plots of influenza

Rambaut et al (2008)

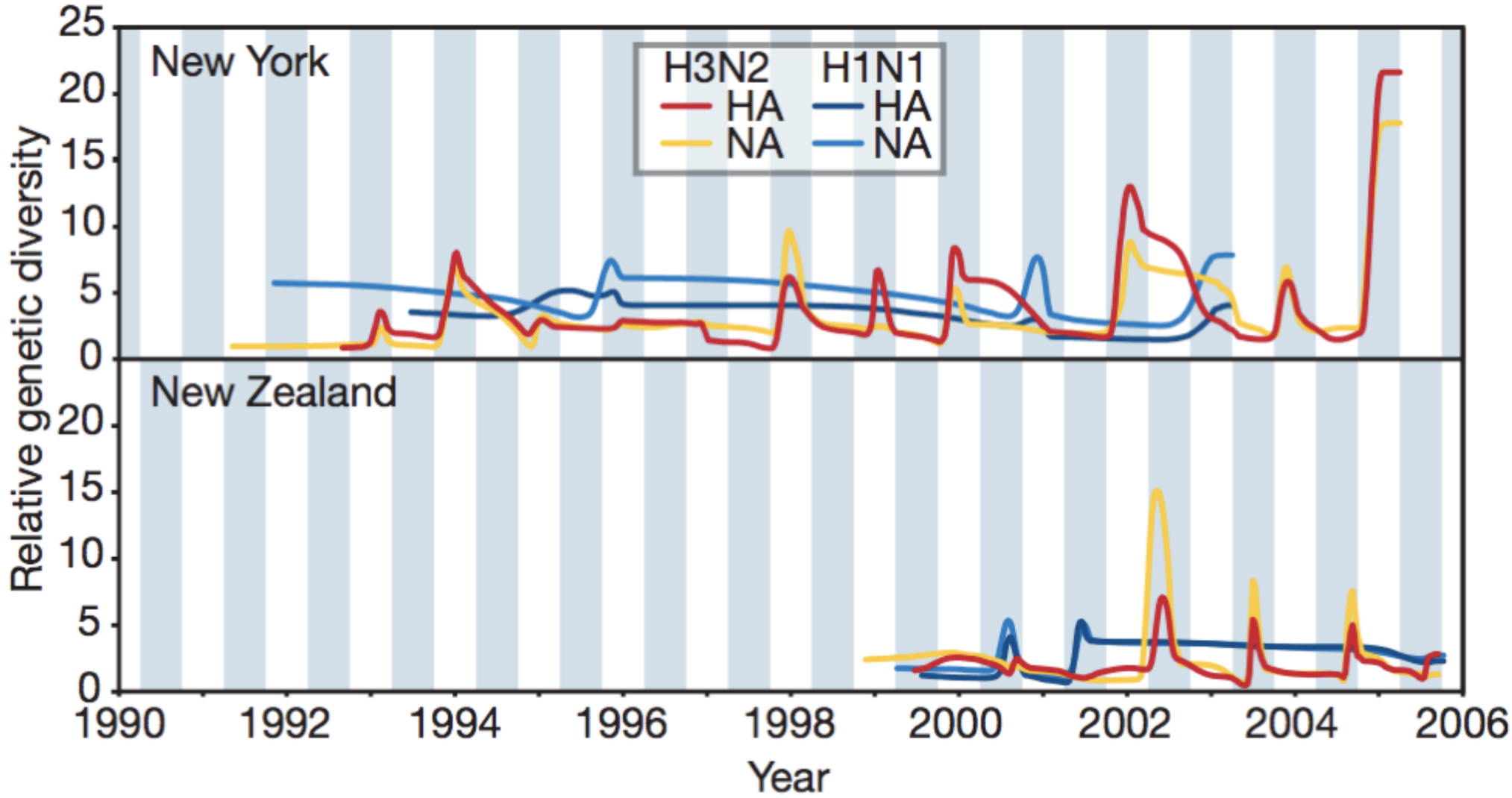


Figure 1 | Population dynamics of genetic diversity in influenza A virus. Bayesian skyline plots of the HA and NA segments for the A/H3N2 and A/H1N1 subtypes in New York state (top) and New Zealand (bottom). The horizontal shaded blocks represent the winter seasons. The *y*-axes represent a measure of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

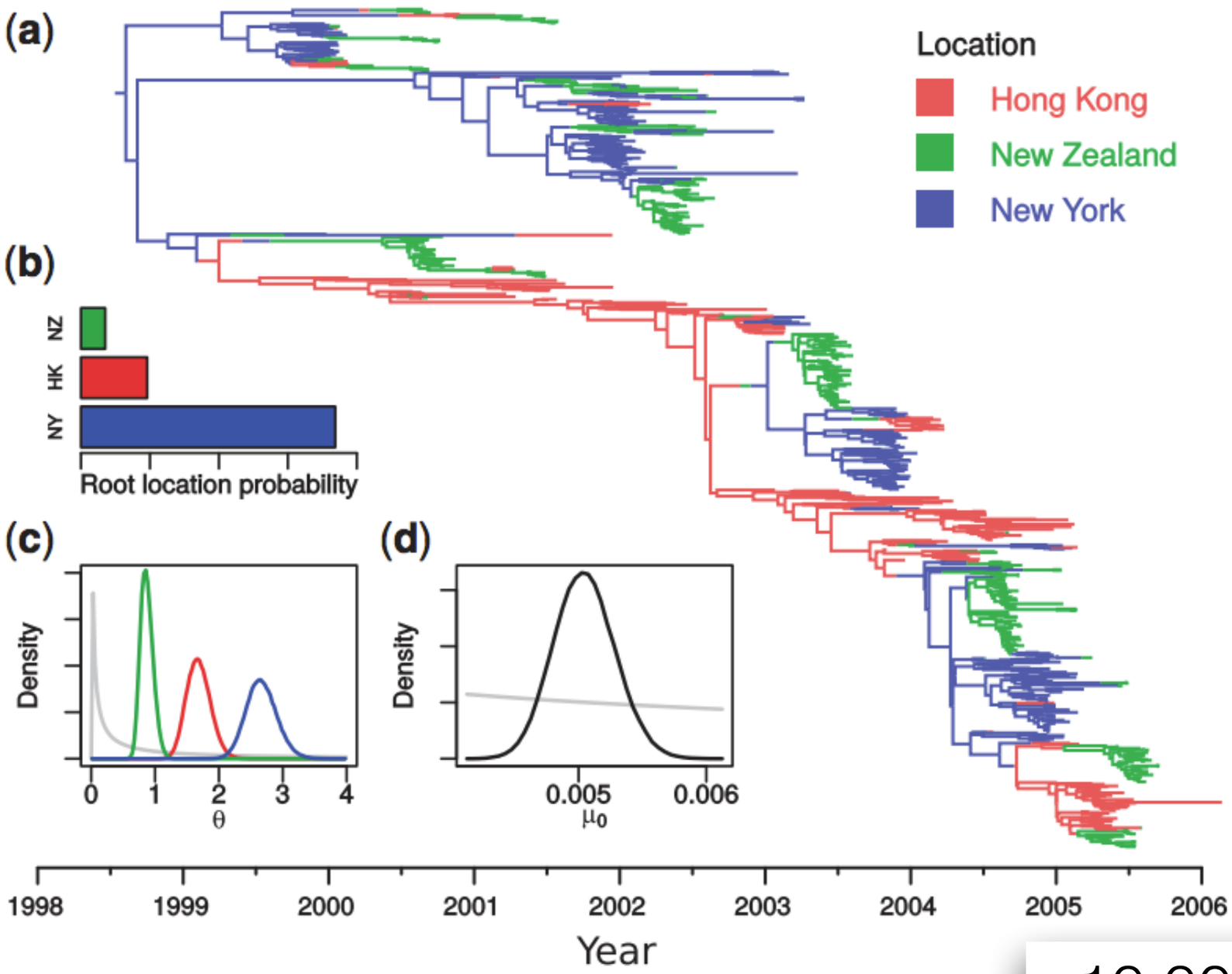
BIOINFORMATICS ORIGINAL PAPER 2014, pages 1–8
doi:10.1093/bioinformatics/btu201

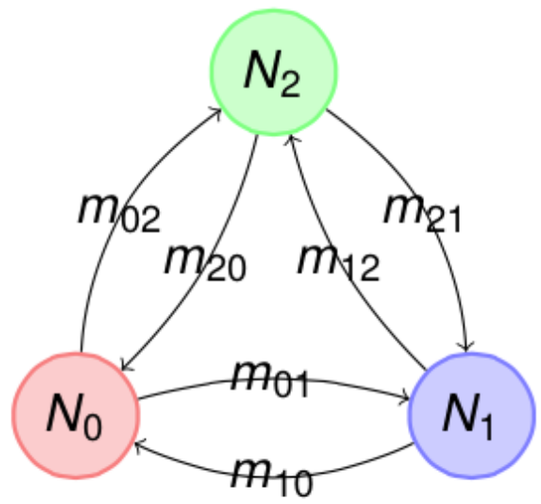
Phylogenetics Advance Access publication April 20, 2014

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

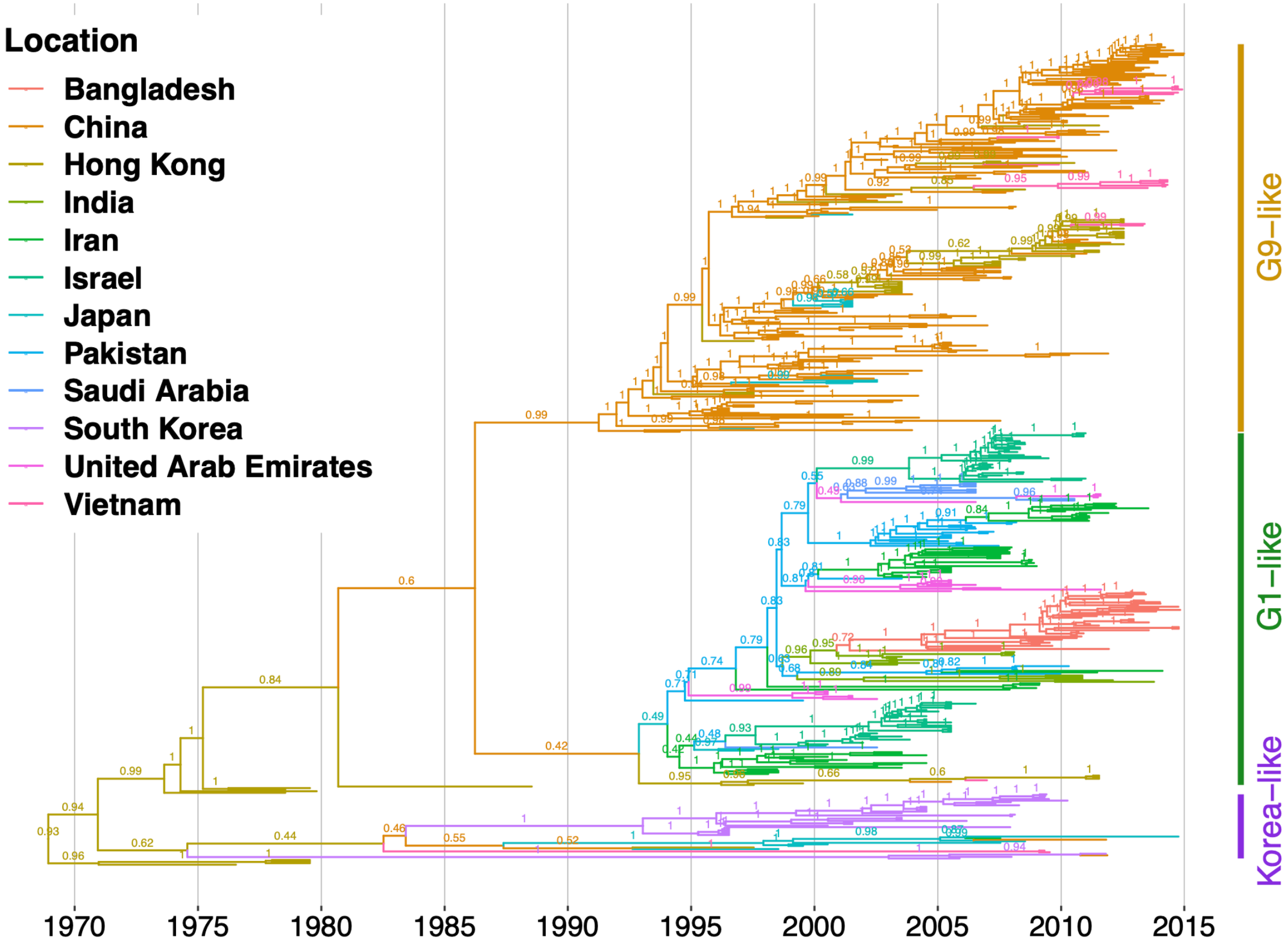




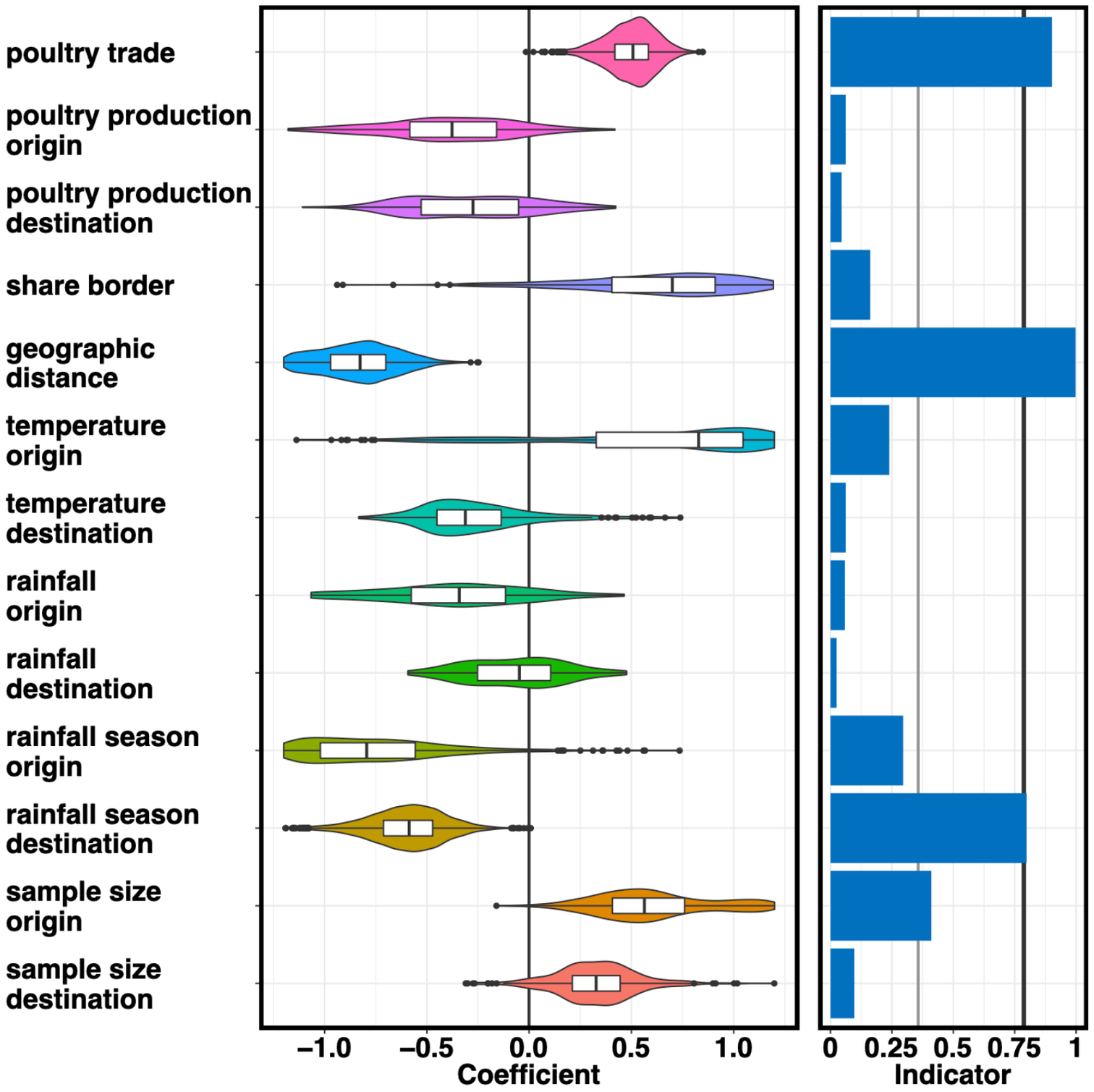
~10,000,000 nucleotides

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

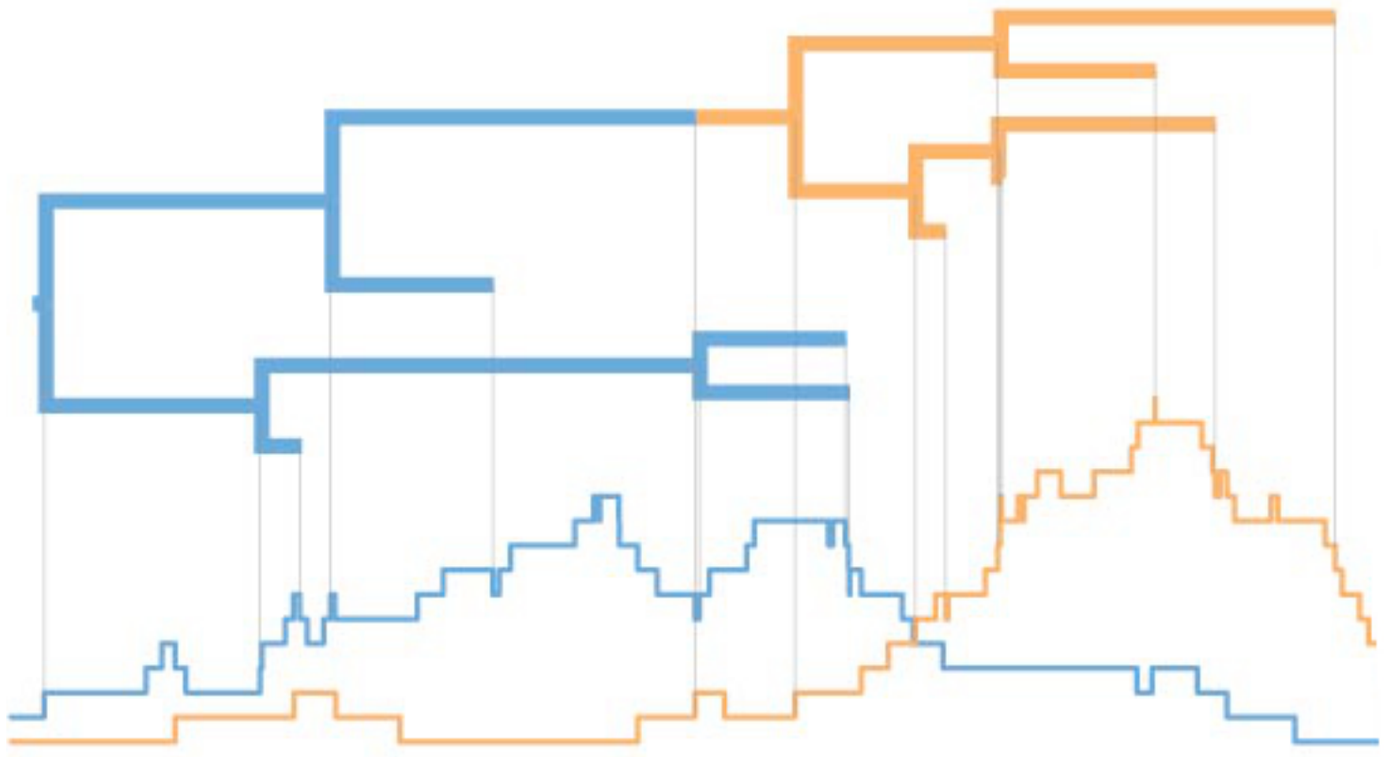


Birth-death-sampling models

- Cater for stochastic effects due to small population sizes and “founder effects”

- Mathematically more challenging than the coalescent.

- Provide a fundamental bridge between stochastic dynamical models and genealogical models.



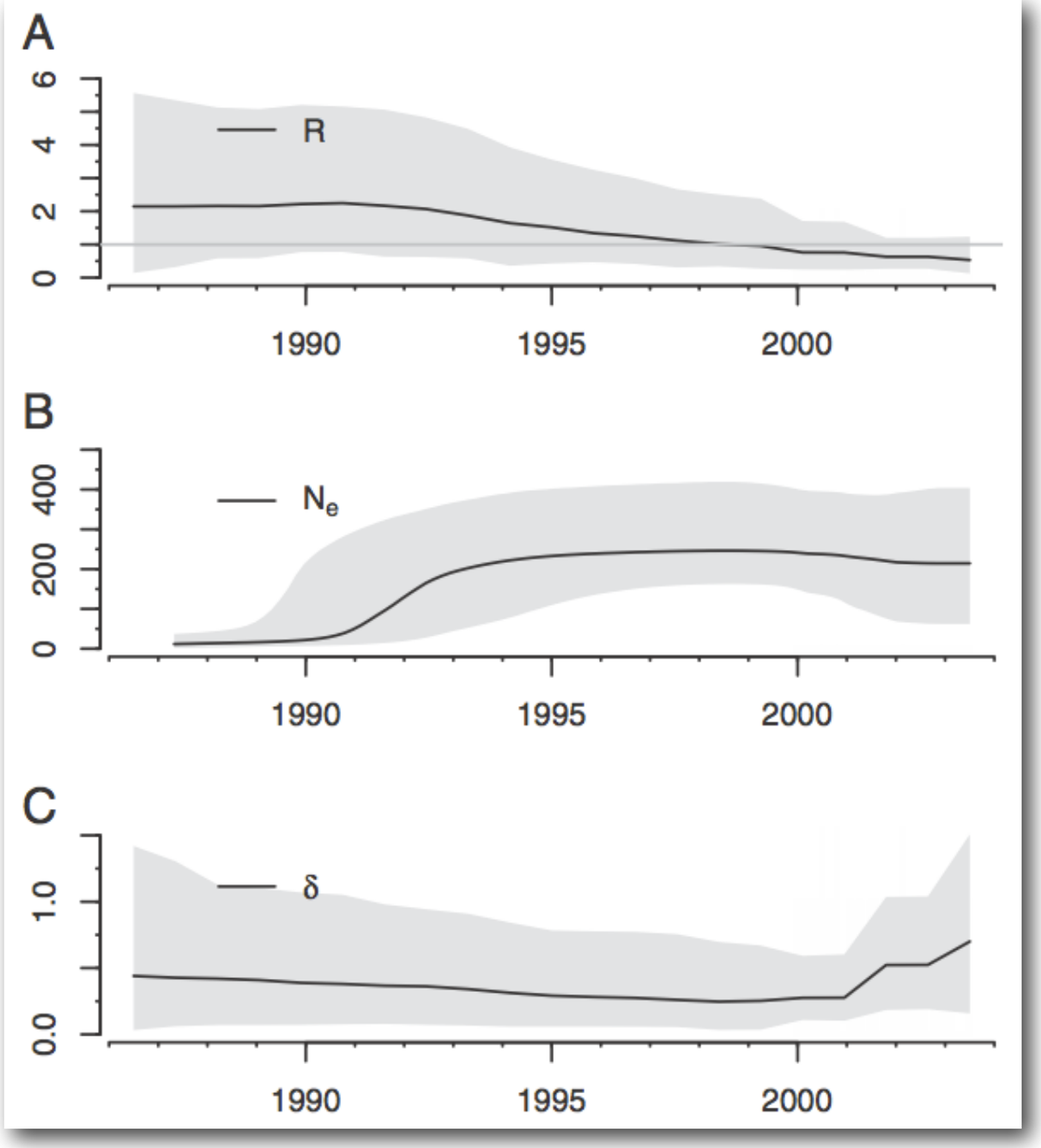
PNAS

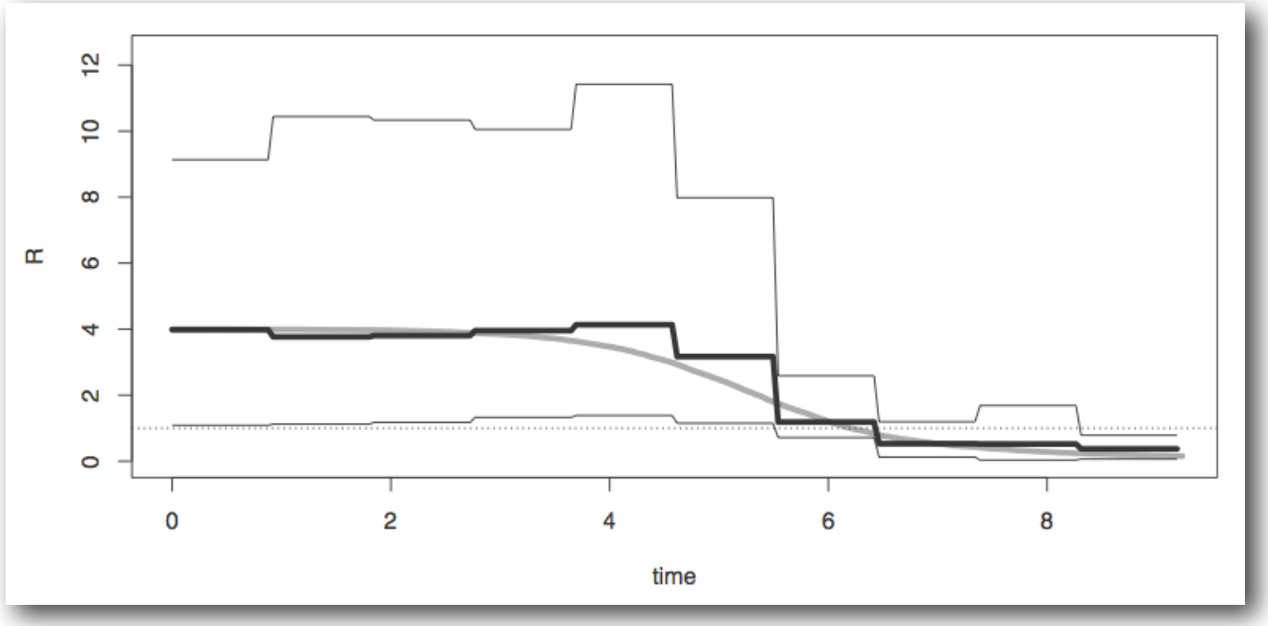
Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)

Tanja Stadler[a,1,2], Denise Kühnert[b,c,1], Sebastian Bonhoeffer[a], and Alexei J. Drummond[b,c]

[a]Department of Environmental Systems Science, Eidgenössische Technische Hochschule Zürich, 8092 Zürich, Switzerland; and [b]Department of Computer Science and [c]Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Auckland, New Zealand

Edited by Robert M. May, University of Oxford, Oxford, United Kingdom, and approved November 15, 2012 (received for review May 10, 2012)





Bayesian Inference of Species Trees from Multilocus Data

Joseph Heled*,[1] and Alexei J. Drummond[1,2,3]

[1]Department of Computer Science, University of Auckland, New Zealand
[2]Bioinformatics Institute, University of Auckland, New Zealand
[3]Allan Wilson Centre for Molecular Ecology and Evolution, University of Au

***Corresponding author:** E-mail: jheled@gmail.com.

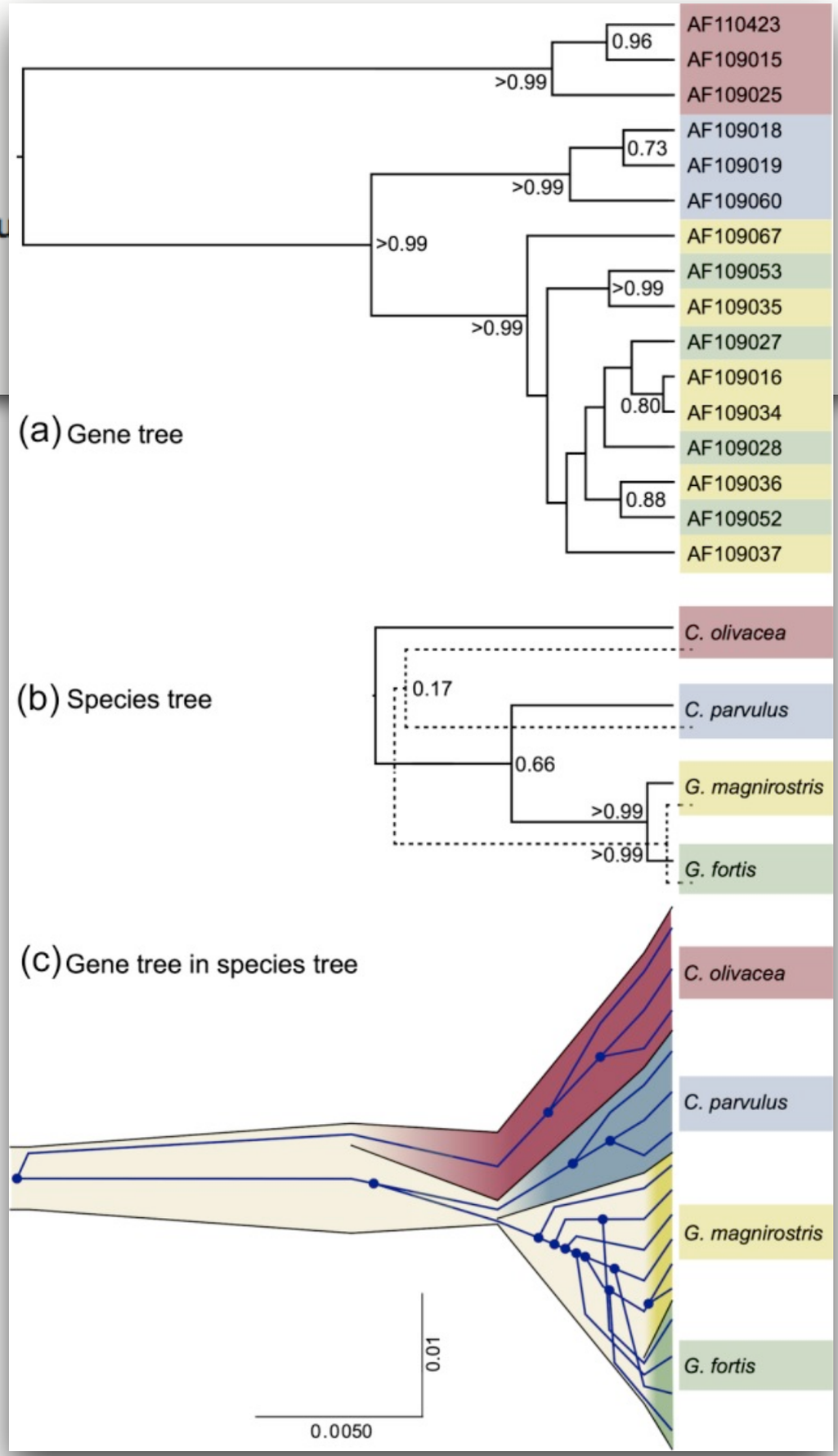
Associate editor: Dr. Jeffrey Throne

- **The multispecies coalescent** is a fundamental statistical model for explaining the relationship between a "gene tree" and a "species tree".

- Admits *incomplete lineage sorting* as the only cause of incongruence.

- Forms a theoretical basis for models that estimate species boundaries based on genomic data.

- Forms a theoretical basis for elaborations of phylogenomic models that include additional confounding factors: lateral gene transfer, admixture, hybridisation, gene flow.

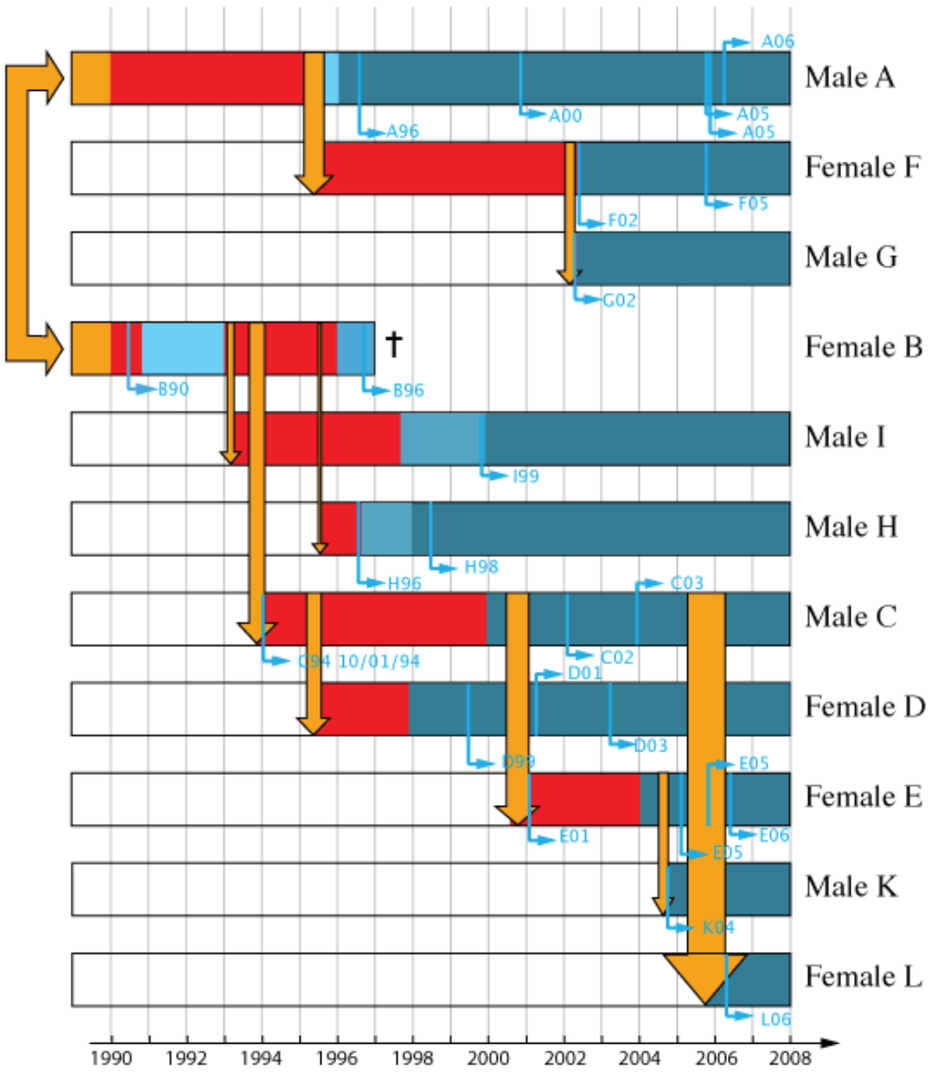


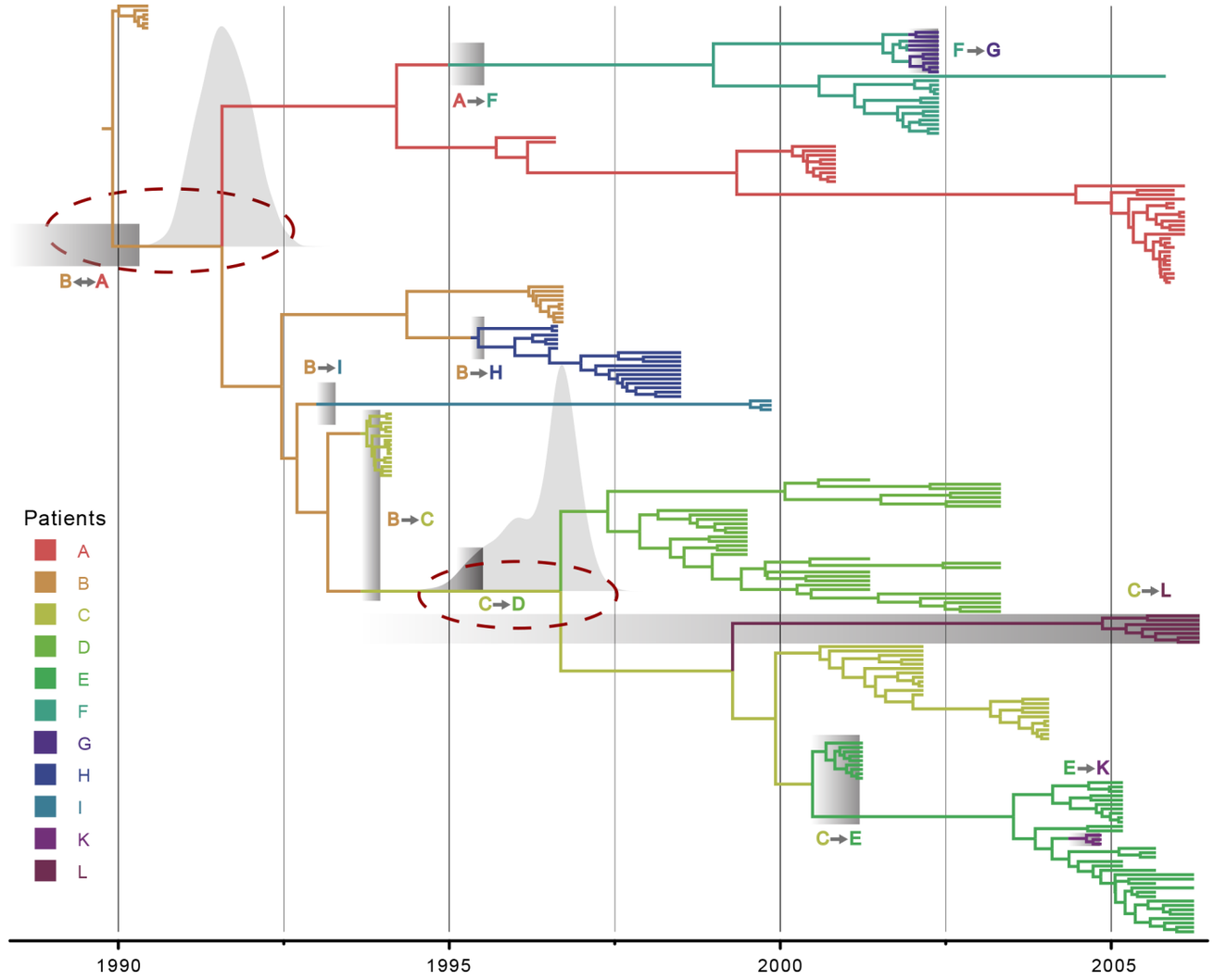
OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

The Genealogical Population Dynamics of HIV-1 in a Large Transmission Chain: Bridging within and among Host Evolutionary Rates

Bram Vrancken[1]*, Andrew Rambaut[2,3], Marc A. Suchard[4,5,6], Alexei Drummond[7], Guy Baele[1], Inge Derdelinckx[8], Eric Van Wijngaerden[8], Anne-Mieke Vandamme[1,9], Kristel Van Laethem[1], Philippe Lemey[1]





Integrative phylogenomics

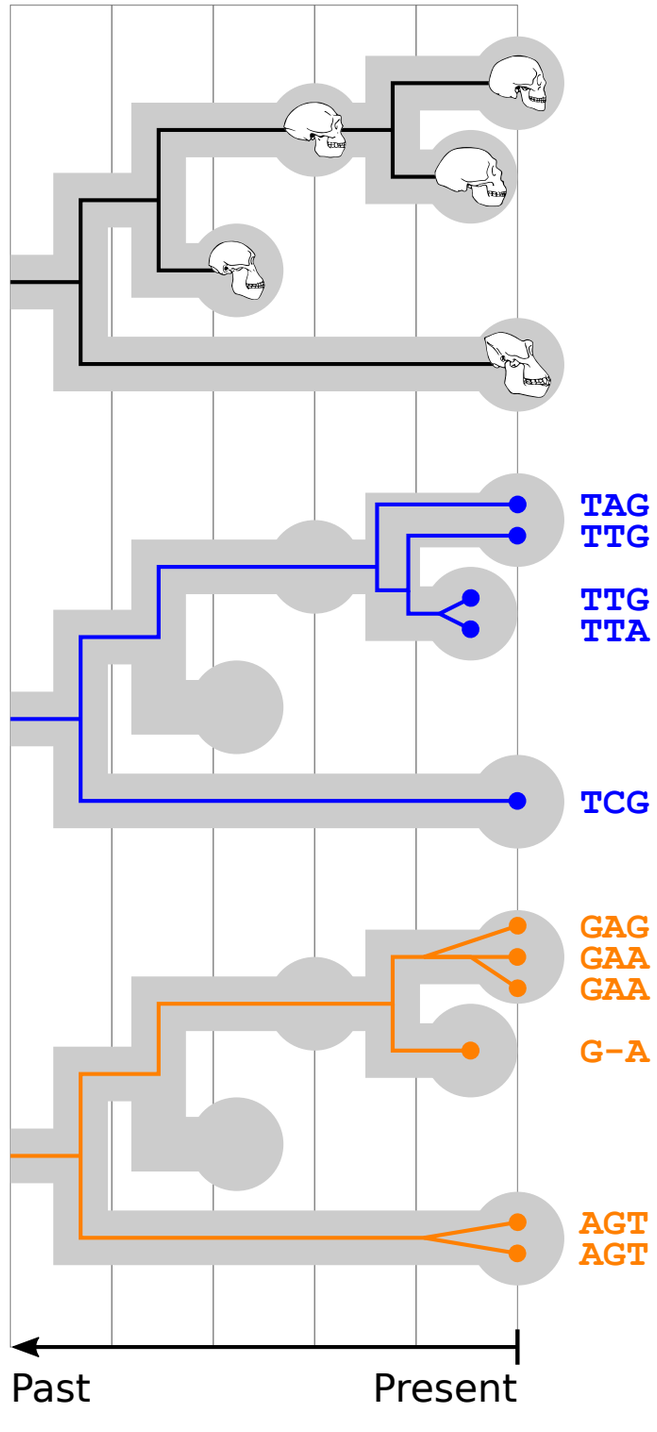
- The evidence for the evolutionary history of a clade of related species comes from a number of independent sources
  - **Genomic sequence data** from extant species
  - **Ancient DNA** from sub fossil species
  - **Phenotypic data** from extant and extinct species
  - **Fossil occurrence and age data**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

Bayesian phylogenetics

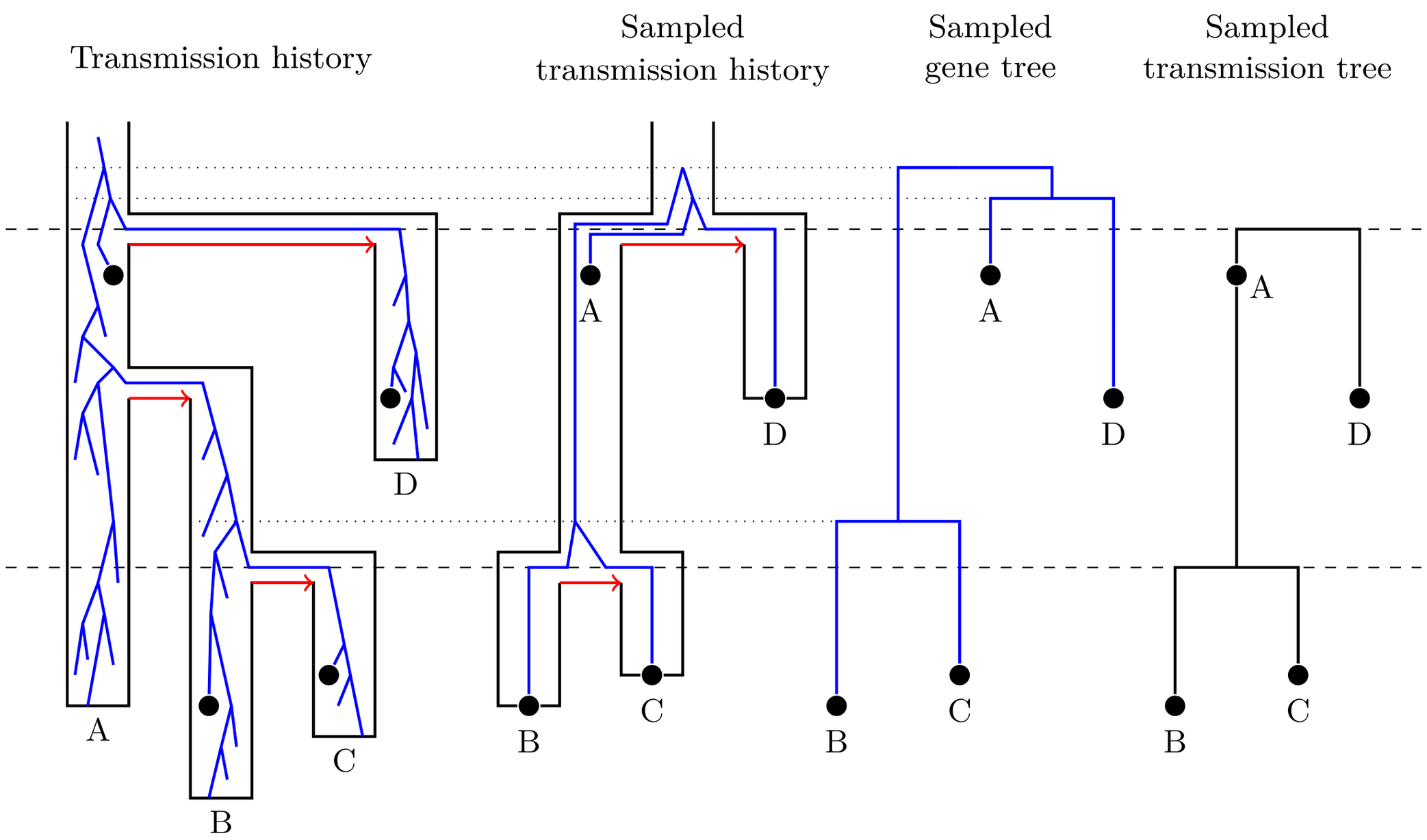
- **Bayesian phylogenetics** (Rannala & Yang,1996; Huelsenbeck & Ronquist, 2001)
- **Morphological substitution models** (Lewis, 2001)
- **Tip-dated phylogenetic models** (Drummond et al, 2002)
  - For ancient DNA and rapidly evolving viruses
- **Relaxed phylogenetics** (Drummond et al, 2006)
  - Reconciling 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

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

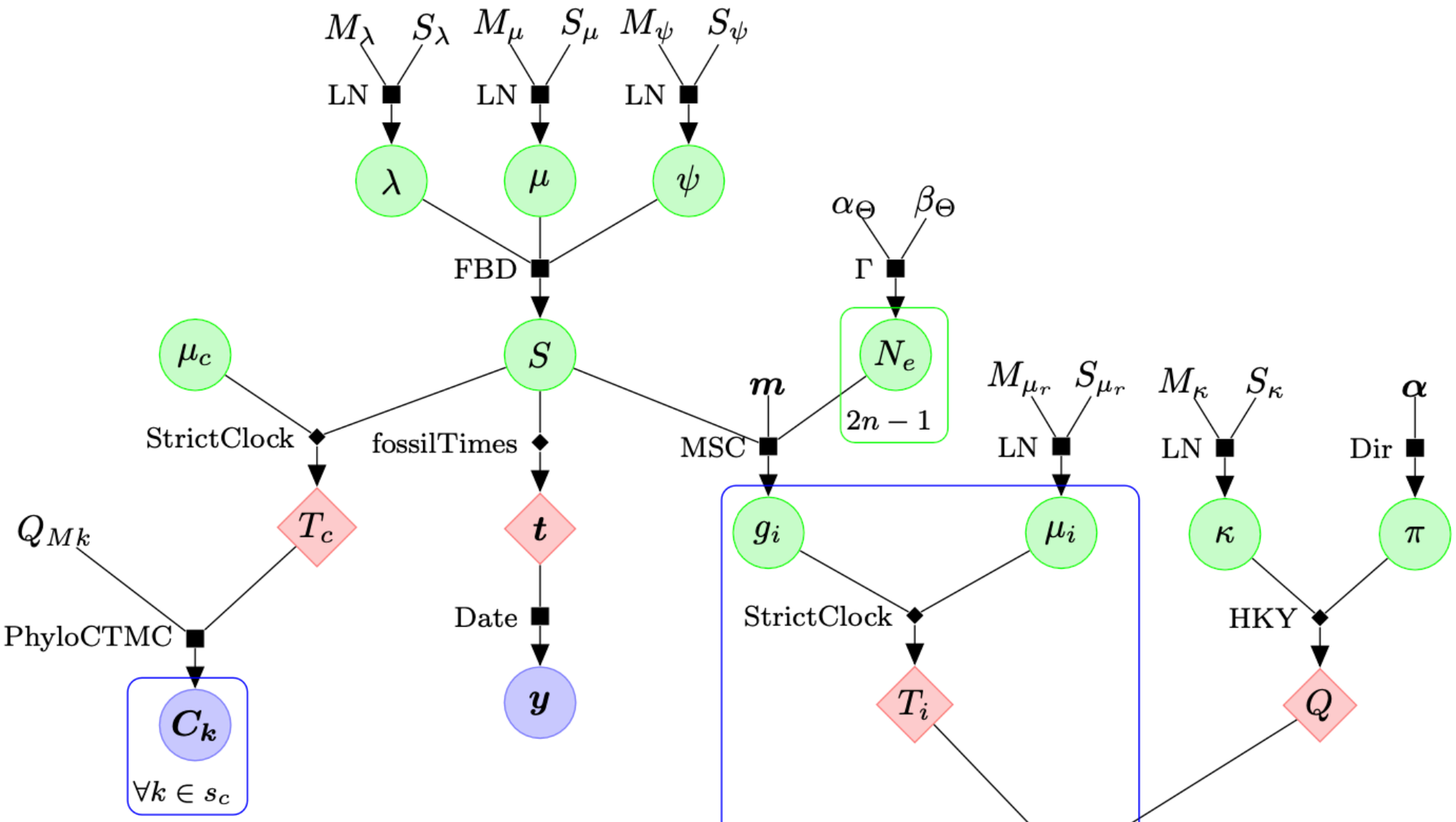


Another integrative model: pathogen evolution inside transmission trees



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



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

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.