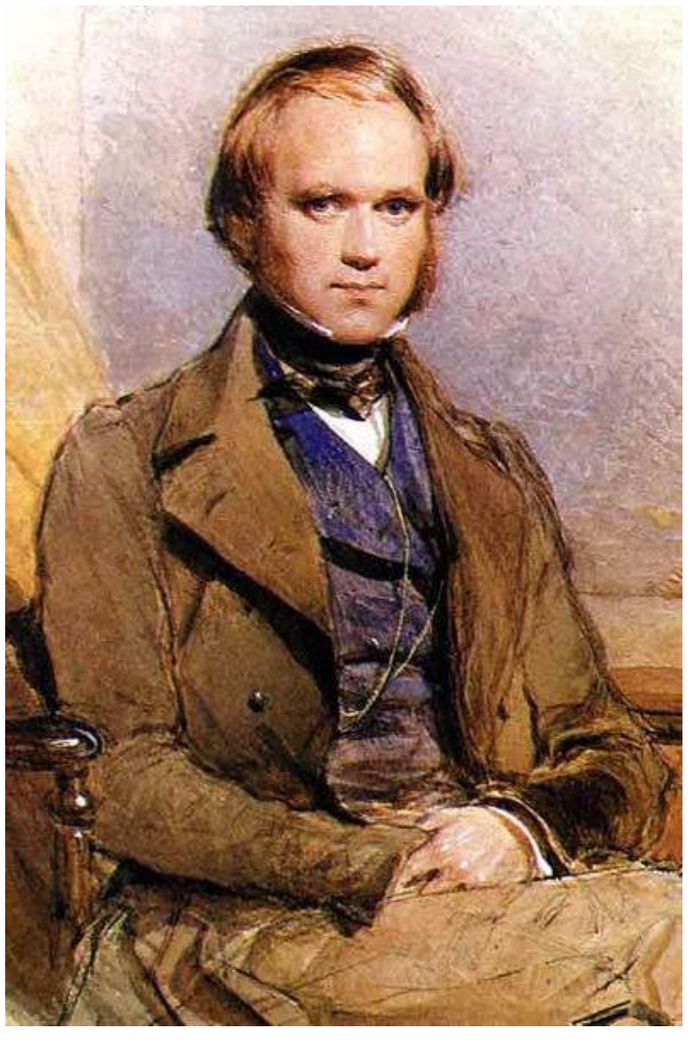
BEAST 2.5: An advanced platform for Bayesian evolutionary analysis

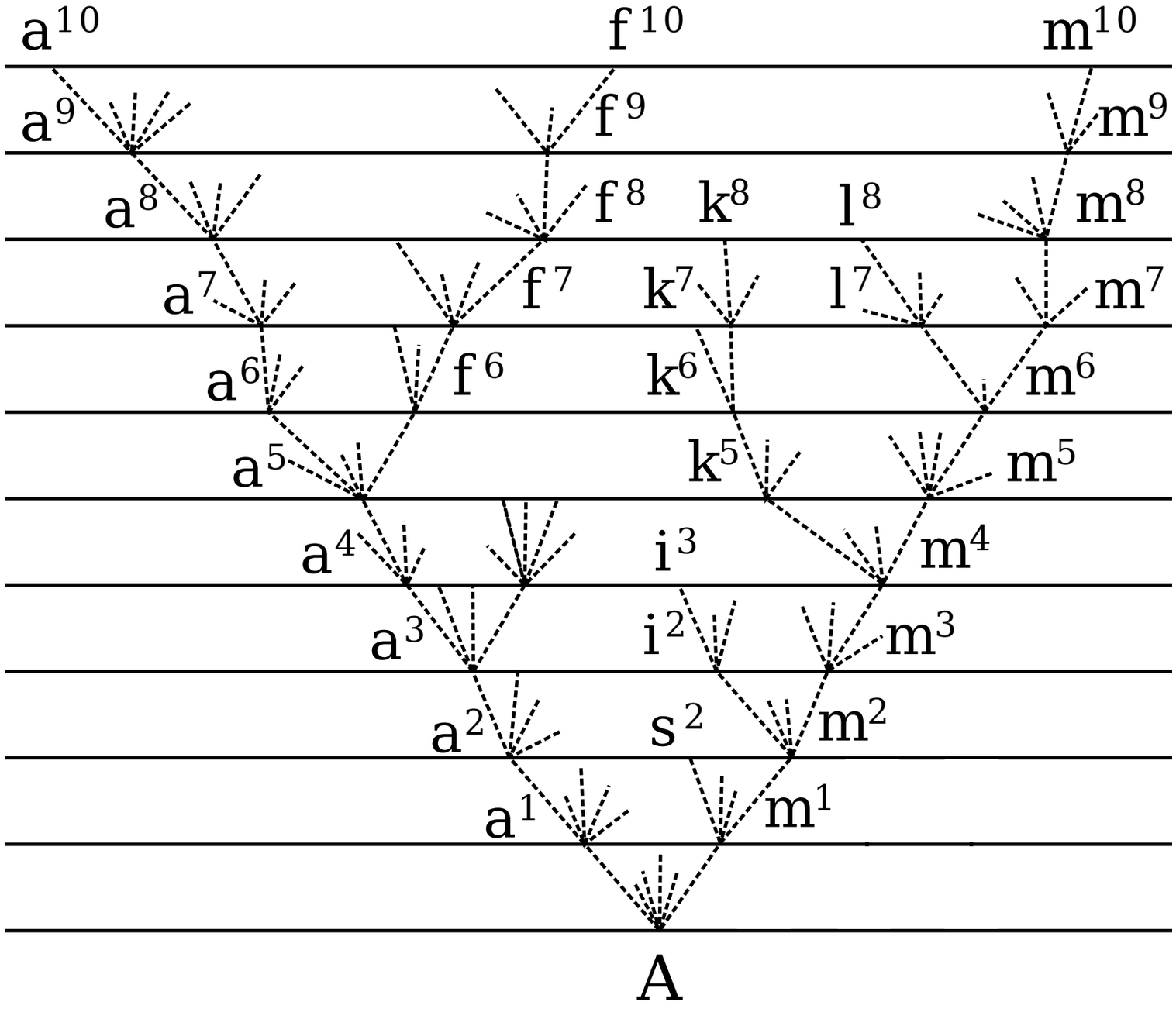
Professor Alexei Drummond

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

18th February 2019, Taming the BEAST Down Under, Sydney

Darwin’s Computer





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

Computational phylogenetics

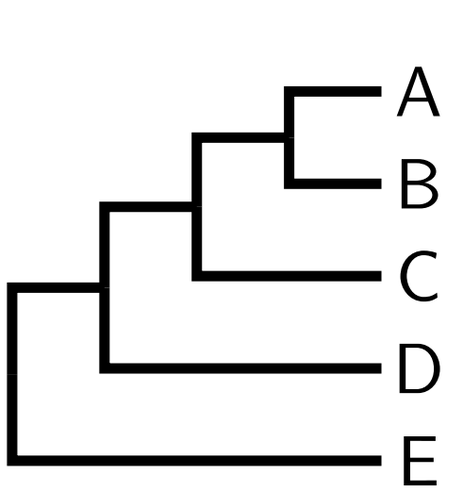
- **1960's-1970's - before the dark ages**
  - Maximum parsimony introduced
  - Least squares
  - Maximum likelihood for gene frequencies conceptualised but not practical (Cavalli-Sforza and Edwards, 1967)
- **1980's - idealogical warfare and the "Dark Ages" for systematics and molecular evolution**
  - Maximum Parsimony and cladistics peaked in the 80's
  - Maximum likelihood pruning algorithm introduced (Felsenstein 1981)
  - Neighbour-joining introduced (Nei 1987)

Computational phylogenetics (cont.)

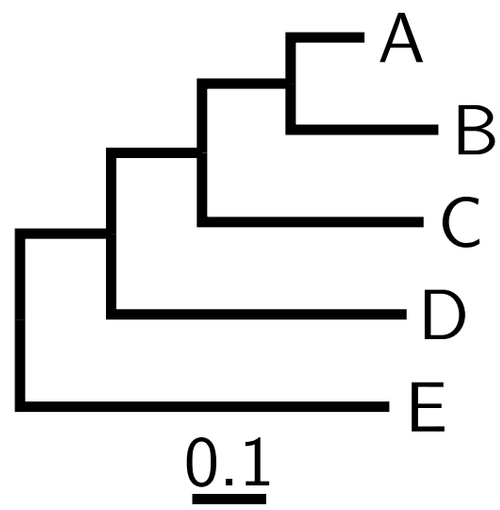
- **1990s - the statistical phylogenetics revolution**
  - maximum likelihood matures, parametric bootstrap, KH test et cetera.
  - Bayesian phylogenetics introduced (1996)
- **2000-2009 - Bayesian phylogenetics revolution**
  - MrBayes, BEAST 1.4, BayesPhylogeny, PhyloBayes et cetera
  - Relaxed phylogenetics
- **2010-now - Phylogenomics and modelling revolution**
  - Multispecies coalescent, fossilized birth-death models, Species Networks, Isolation with Migration models
  - Bayesian frameworks: BEAST 2.5, BEAST 1.10, RevBayes and more
- **Next - Integrative phylogenomics revolution?**

Phylogenies come in different classes

rooted trees

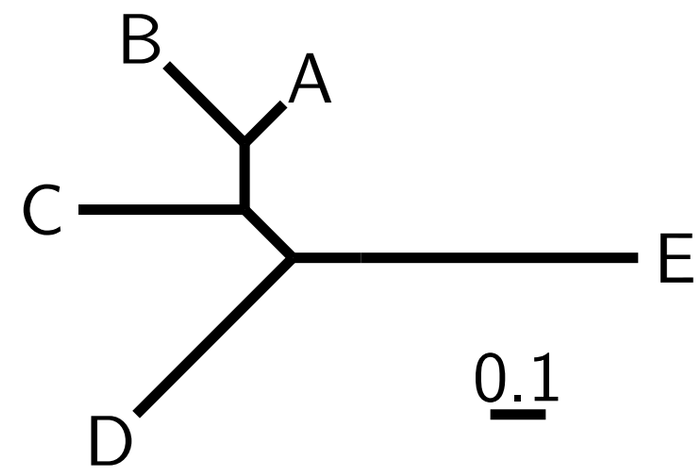


(a) cladogram



(b) phylogram

unrooted tree



(c) unrooted tree

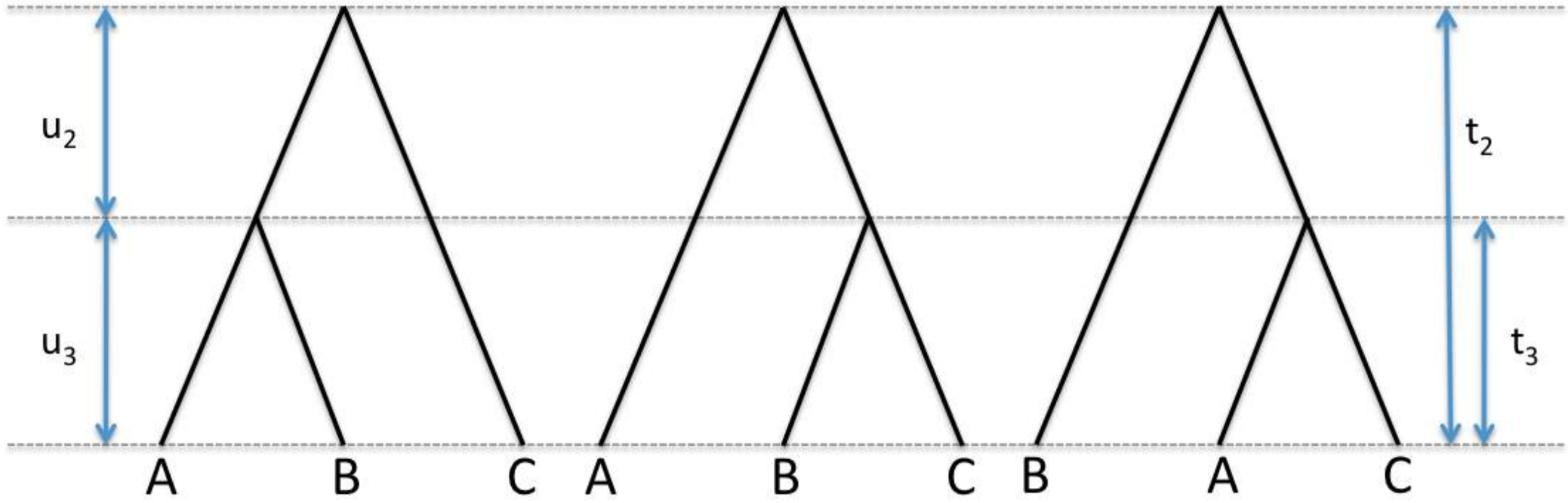
((((A, B), C), D), E);

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

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

The tip-labeled **time-tree**

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



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

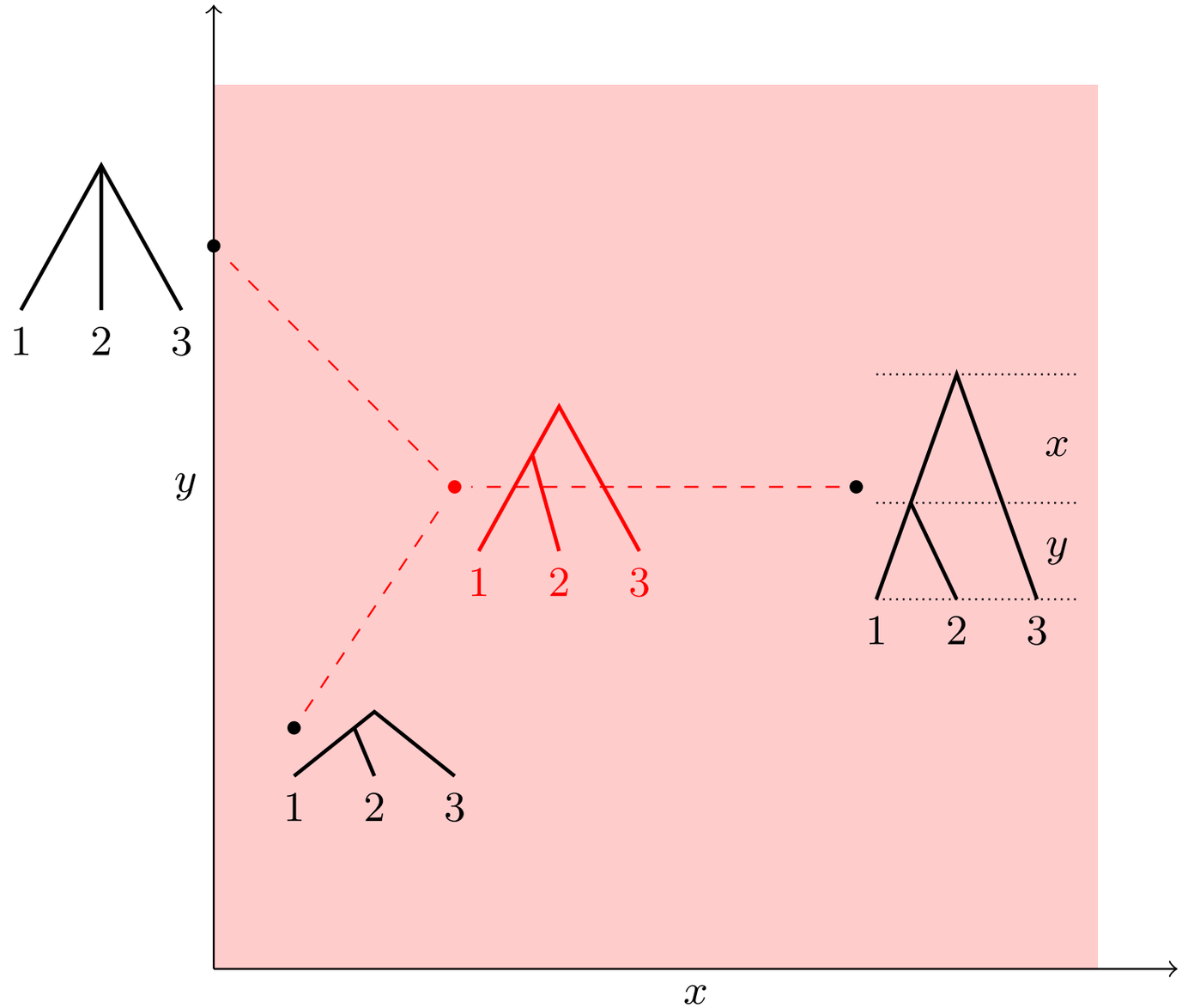
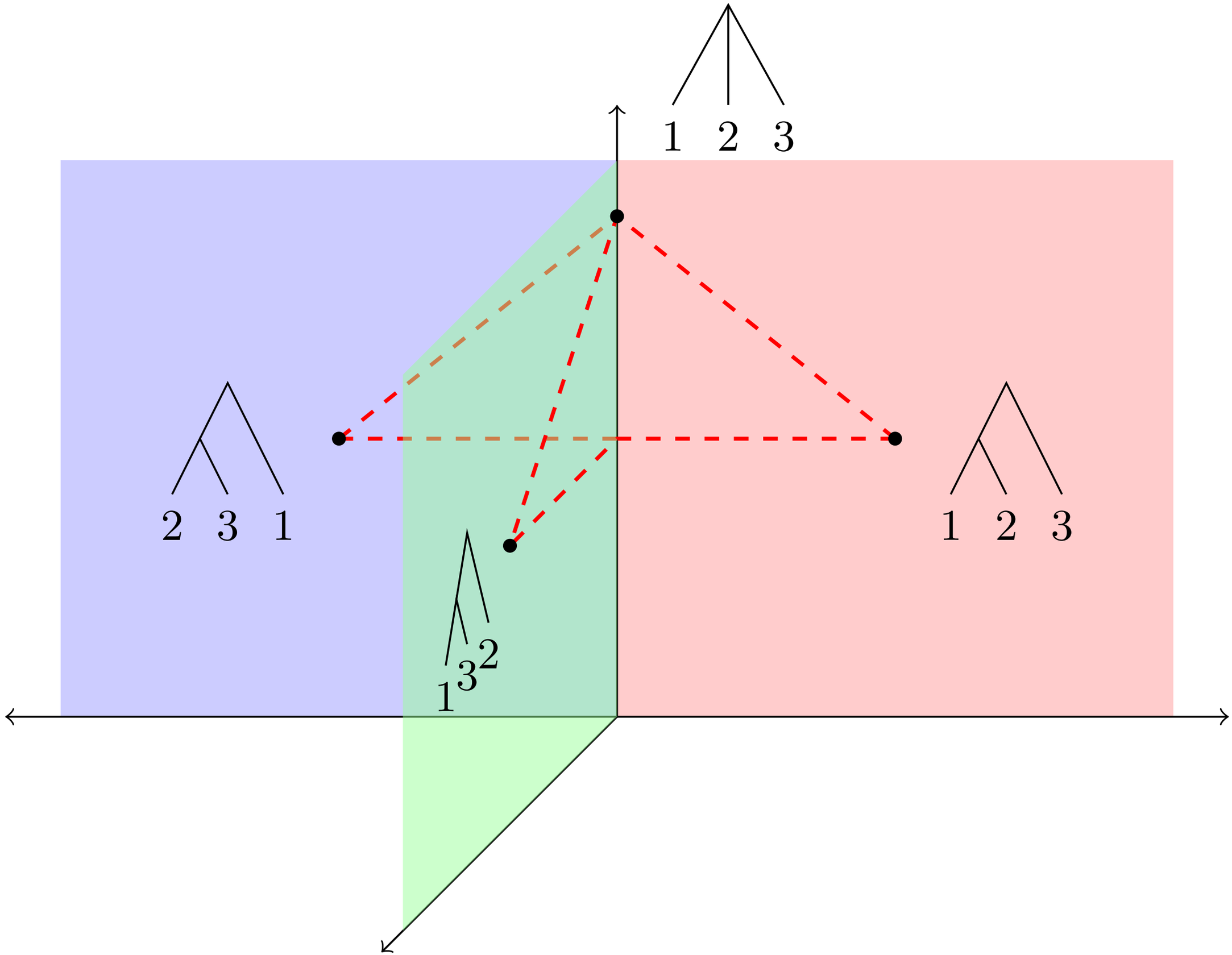


Figure: A Euclidean two-dimensional space representing the space of all possible time-trees for the topology ((1,2),3). There are two parameters, x and y, one for each of the two inter-coalescent intervals, the sum of which is the age of the root ($t_{root} = x + y$). Three trees are displayed, along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


1 2 3
2 3 1
1 2 3
1 3 2

Tip-labeled time-trees of size 4

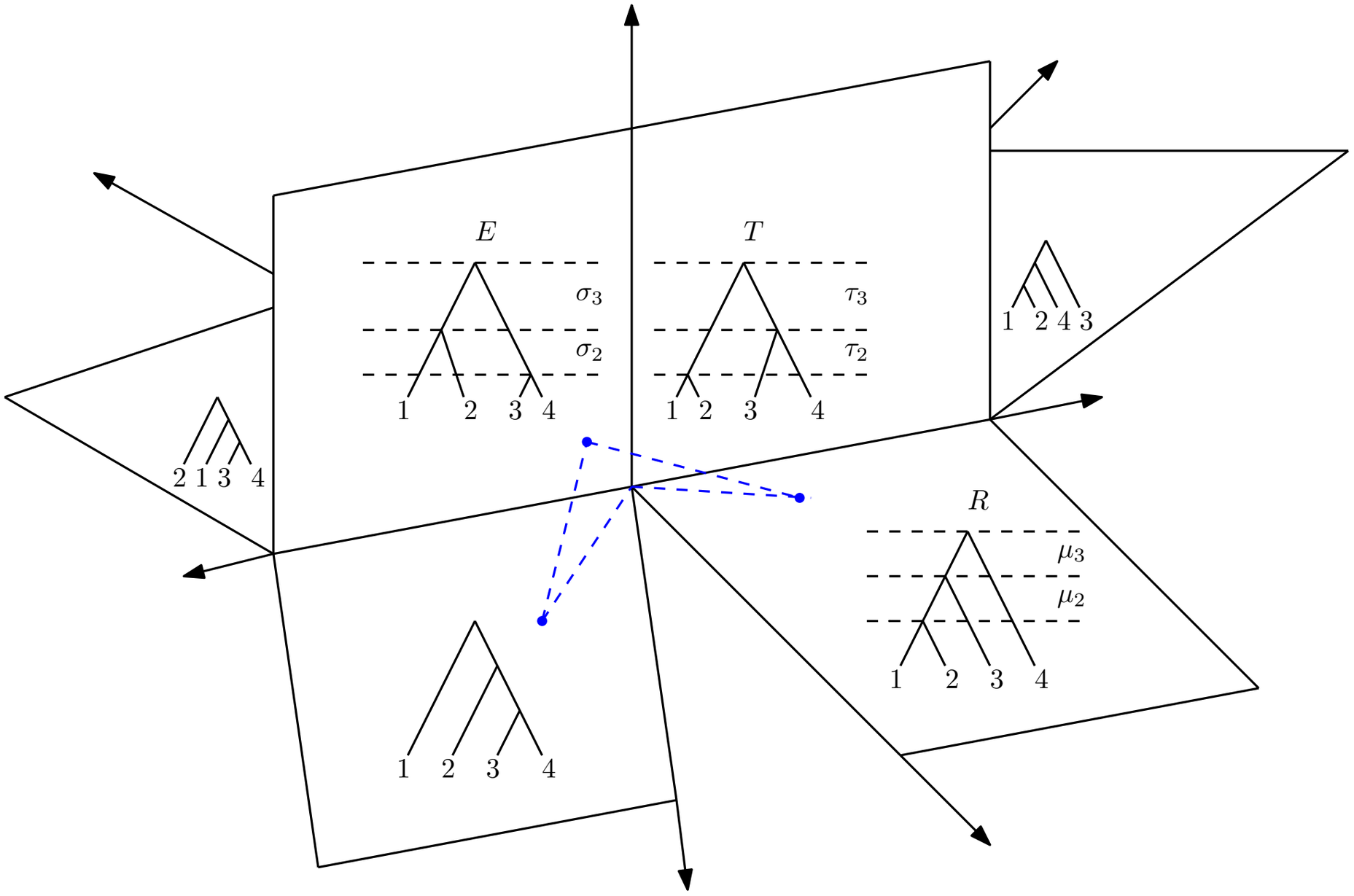
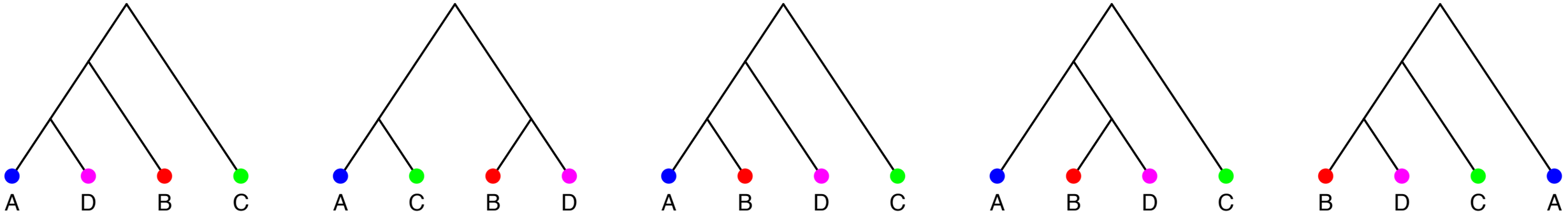


Figure: Three-dimensional projection of 4-dimensional τ-space T_4.

Figure from Gavryushkin and Drummond (2016)



15 possible (unranked) trees of 4 individuals/species

105 possible trees of 5 individuals/species

945 possible trees of 6 individuals/species

Question: How many possible trees are there relating seven taxa?

How many trees are there?

For n species there are

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

rooted, tip-labelled binary trees:

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

Counting different types of trees

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

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


Consensus
1. Tarsius_syrichta
2. Lemur_catta
3. Homo_sapiens
4. Pan
5. Gorilla
6. Pongo
7. Hylobates
8. Macaca_fuscata
9. M_mulatta
10. M_fascicularis
11. M_sylvanus
12. Saimiri_sciureus
~10,000 nucleotides

The phylogenetic likelihood

Felsenstein (1981)

Besides coding for function, DNA serves as a record of evolutionary history.

But, in order to reconstruct the phylogenetic tree we need to have a procedure to evaluate each tree in light of the sequence data.

One means of evaluating a tree would be to calculate the **probability of the data** under a statistical model of DNA evolution.

This is known as the **Likelihood** of the tree. One method of reconstructing the evolutionary history is then to find the tree that has the **Maximum Likelihood**.

Bayes theorem

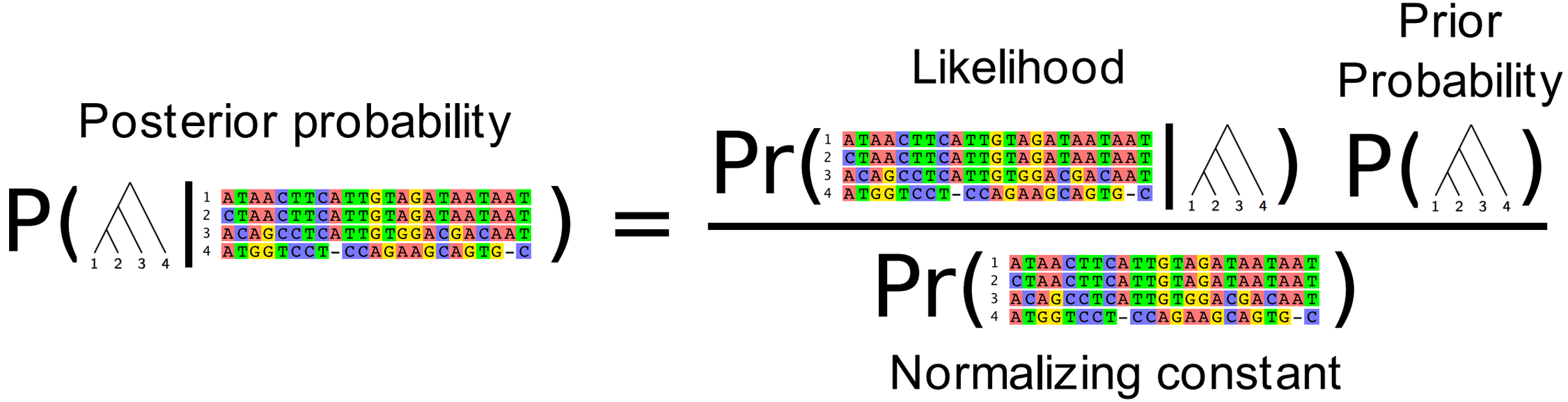
$$\underbrace{P(\theta \mid D)}_{\text{posterior}} = \frac{\overbrace{P(D \mid \theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(D)}_{\text{marginal likelihood}}}$$

Bayesian reconstruction of phylogenetic trees

Yang & Rannala (1997), Mau, Newton & Larget (1998), Wilson and Balding (1998)

In phylogenetics what we want is the **probability of each tree** given the aligned sequence data.

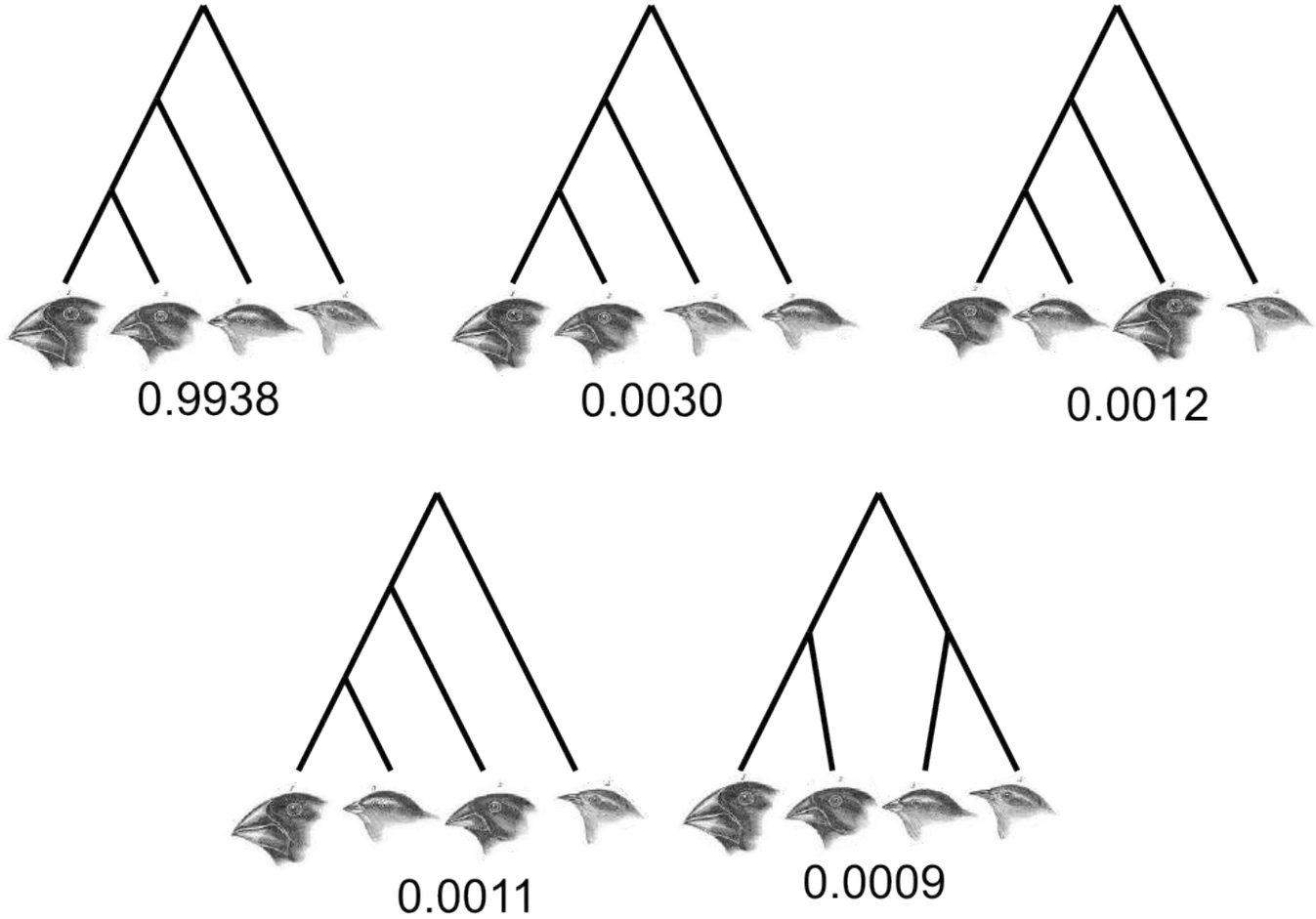
We can compute the probability of a tree using **Bayes Theorem**:



Using the **Markov chain Monte Carlo algorithm** we can produce a sample of trees from this posterior probability distribution **without knowing the marginal likelihood (normalizing constant)**.

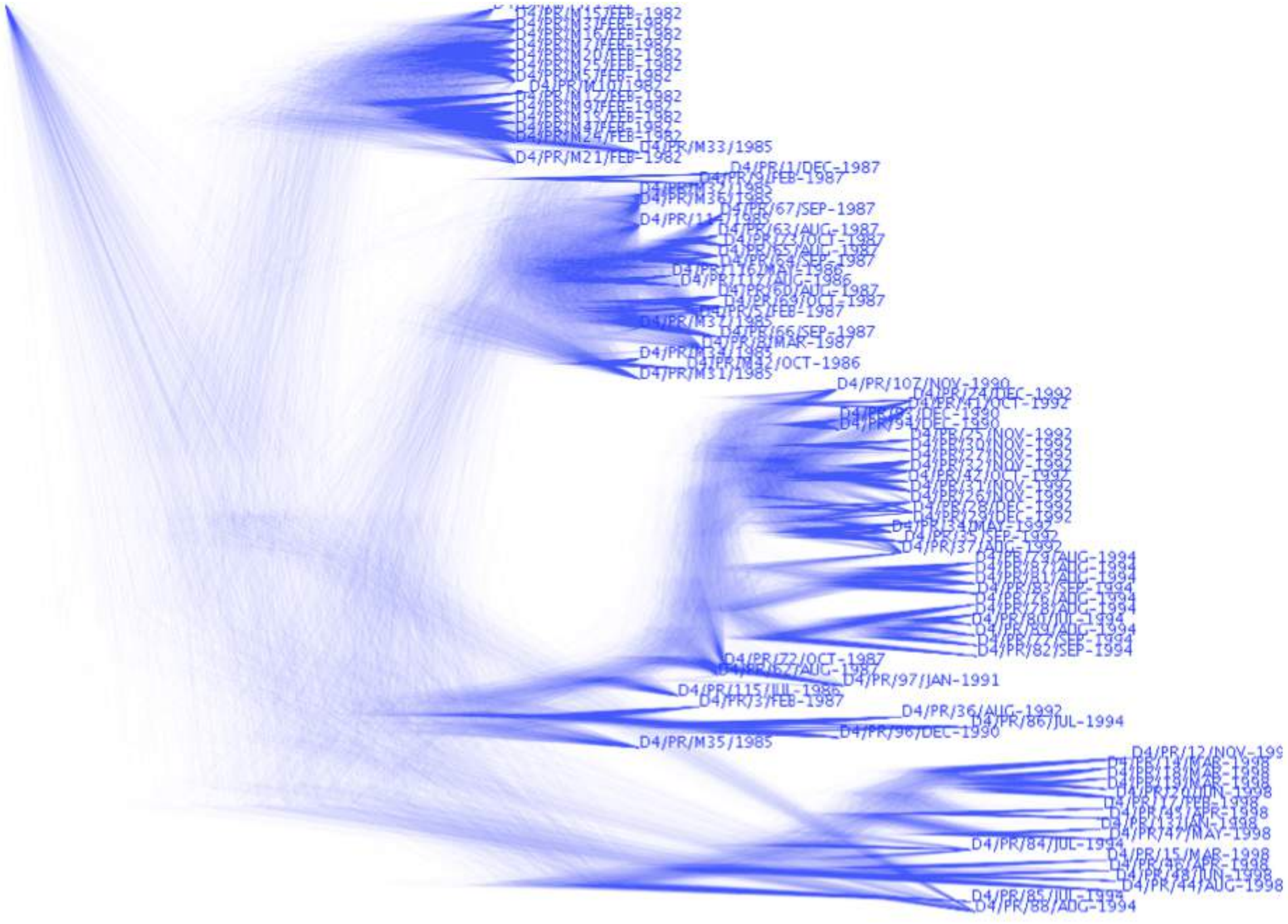
The posterior distribution on Darwin's Finches

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



This posterior probability distribution was computed using **Markov chain Monte Carlo** implemented in the BEAST software package.

The posterior distribution of larger trees



Elaborating the model

Basic model: (posterior proportional to likelihood x prior)

$$P(T \mid D) \propto \Pr(D \mid T)P(T)$$

Substitution model parameters:

Assuming independence

$$P(T, Q \mid D) \propto \Pr(D \mid T, Q)P(T)P(Q)$$

Substitution model and tree branching process parameters:

Assuming independence

$$P(T, Q, \theta \mid D) \propto \Pr(D \mid T, Q)P(T \mid \theta)P(\theta)P(Q)$$

The phylogenetic posterior

Standard application of Bayes theorem gives the posterior:

$$P(T, Q, \theta \mid D) = \frac{\Pr(D \mid T, Q, \theta) P(T, Q, \theta)}{\Pr(D)}$$

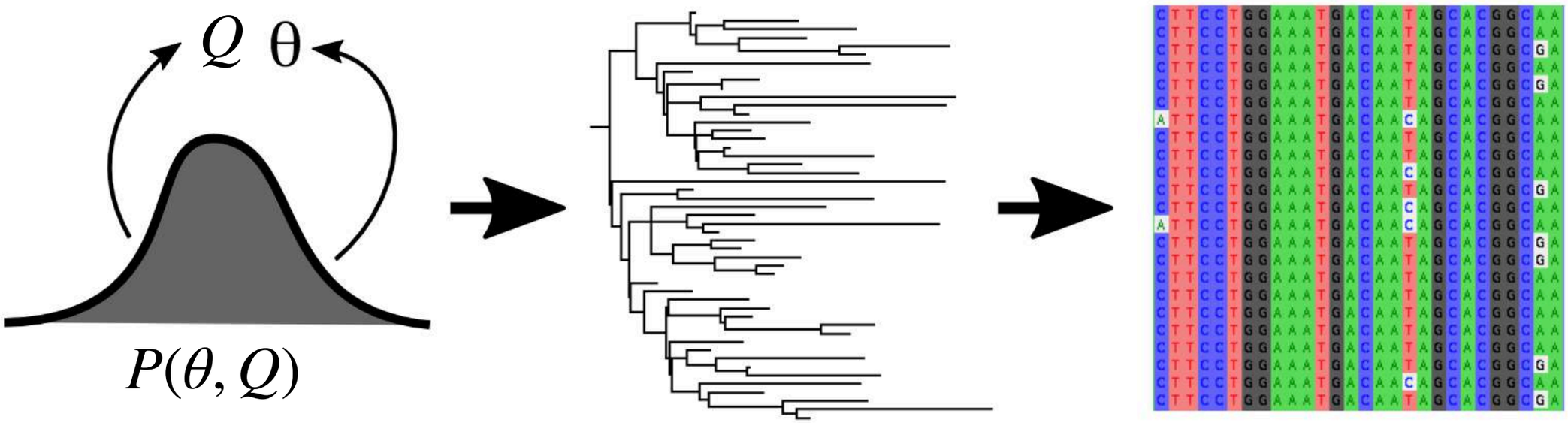
But you will normally see it written like this

$$P(T, Q, \theta \mid D) = \frac{1}{\Pr(D)} \Pr(D \mid T, Q) P(T \mid \theta) P(\theta) P(Q)$$

- the probability of the data doesn't depend on θ **except through the tree**.
- the prior probability of the tree depends on θ but not on Q.
- the prior probability of θ and the prior probability of Q are independent.

The neutrality assumption

Because of the way we've factorized the joint probability for the data and model parameters, we are implicitly assuming that our alignment could have been produced in the following fashion:

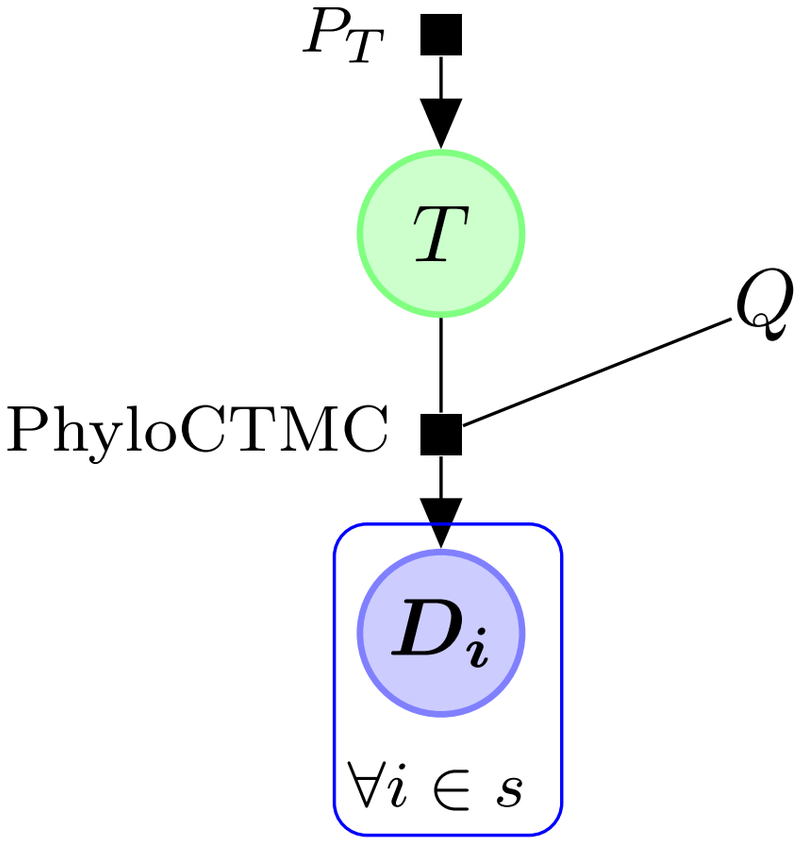


Sample parameters from prior

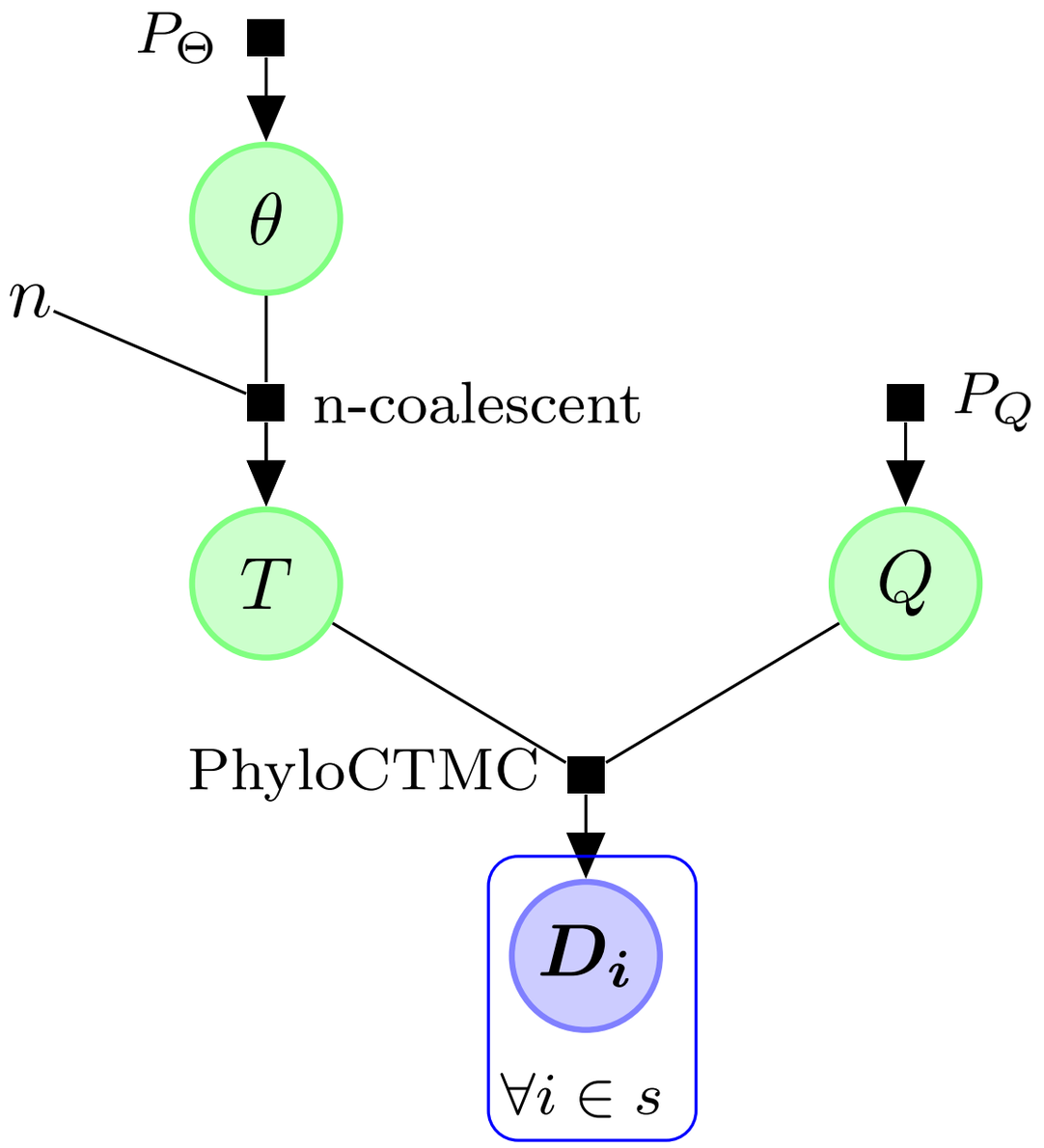
Sample tree from prior

Simulate sequence alignment

Graphical models

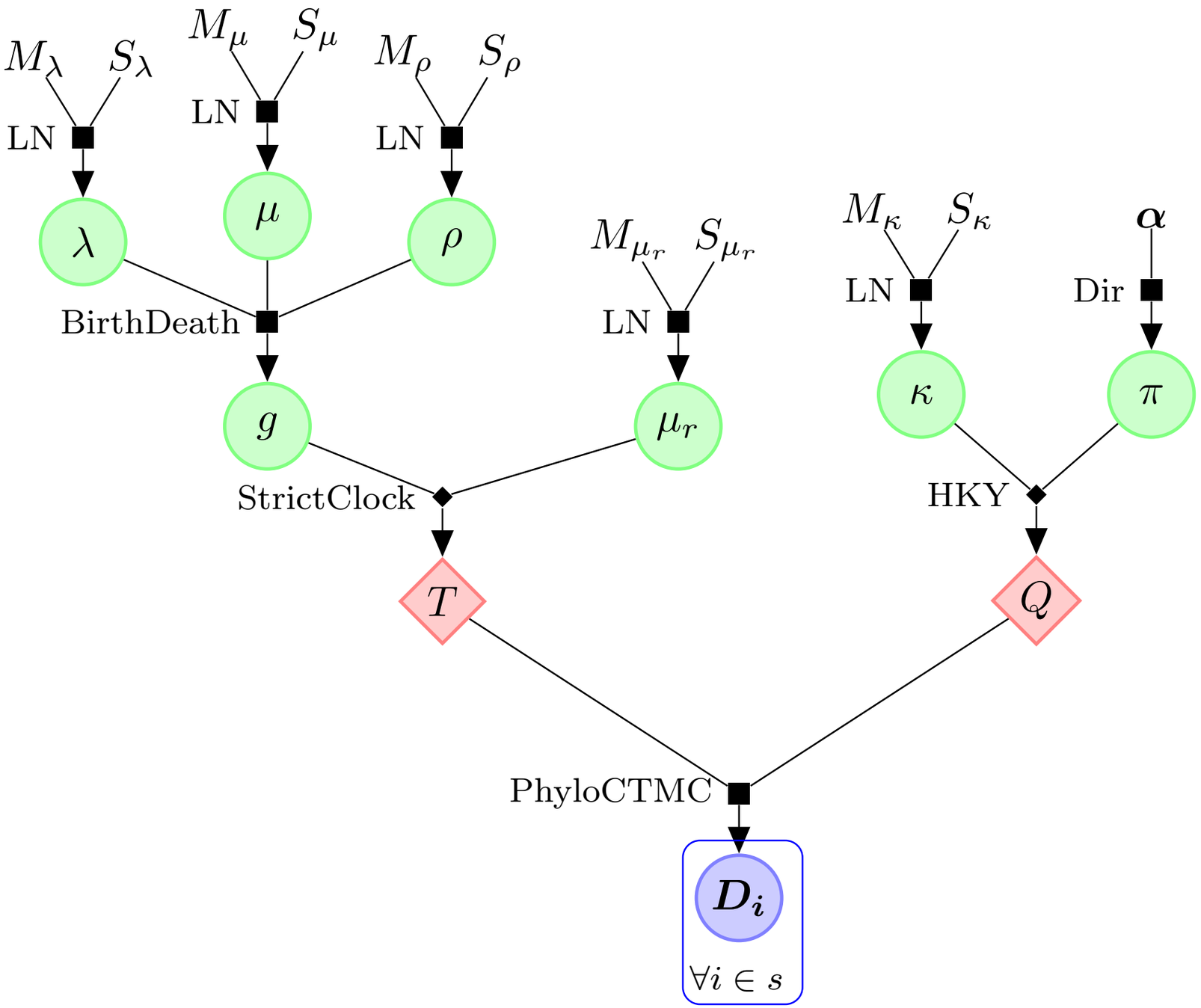


$$P(T|\boldsymbol{D}) \propto \Pr(\boldsymbol{D}|T)P(T)$$



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

Graphical models for phylogenomics



$$P(g, \mu_r, \lambda, \mu, \rho, Q|\boldsymbol{D}) \propto \Pr(\boldsymbol{D}|\mu_r g, Q) P(g|\lambda, \mu, \rho) P(\lambda) P(\mu) P(\rho) P(\mu_r) P(Q)$$

New Results

BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis

Remco Bouckaert, Timothy G Vaughan, Joelle Barido-Sottani, Sebastian Duchene, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kuhnert, Nicola de Maio, Michael Matschiner, Fabio K Mendes, Nicola Muller, Huw A Ogilvie, Louis du Plessis, Alex Popinga, Andrew Rambaut, David Rasmussen, Igor Siveroni, Marc Suchard, Chieh-Hsi Wu, Dong Xie, Chi Zhang, Tanja Stadler, Alexei J. Drummond

doi: https://doi.org/10.1101/474296

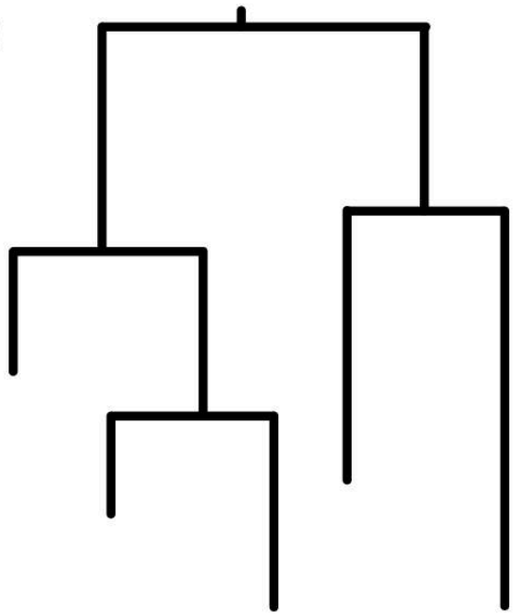
Authors hail from 18 institutions across 9 countries

BEAST 2.5

- Allows a user to construct a wide range of phylogenetic models to apply to sequence alignments and comparative data using the BEAUti user interface.
- The major components of the phylogenetic models are
  - the time-tree prior (coalescent or birth-death models)
  - the substitution model (nucleotide, codon, trait evolution)
  - the site model (how the substitution model varies among sites/loci)
  - the molecular clock model - strict, relaxed, random local clock
  - How such models are “partitioned” when there are multiple data partitions
- Major differences in models are handled by different top-level templates
- New sub-models and templates can be designed by 3rd party developers

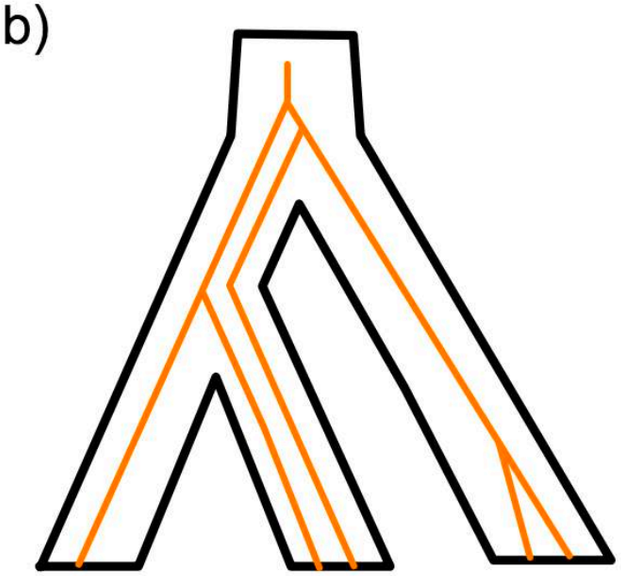
Some BEAST 2.5 tree priors

a)



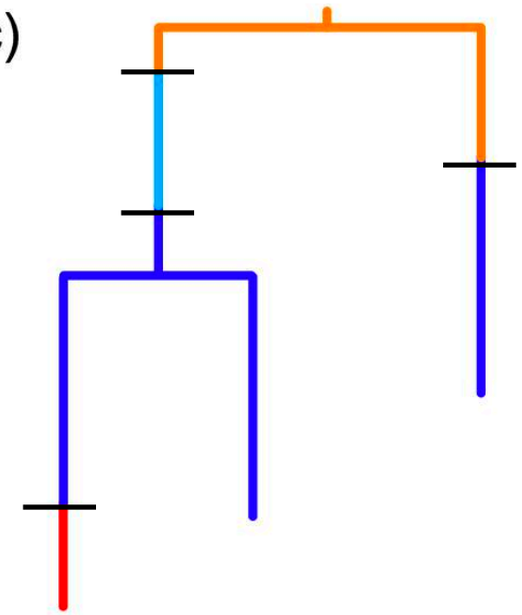
Tip-dated time tree
(leaf times conditioned on)

b)



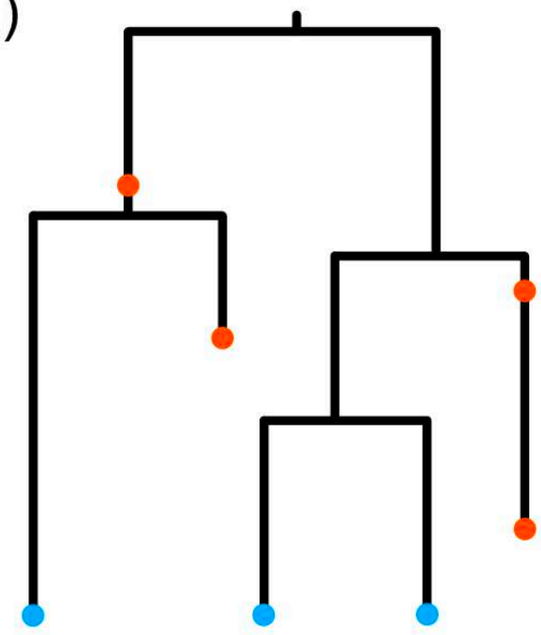
Multi-species coalescent

c)



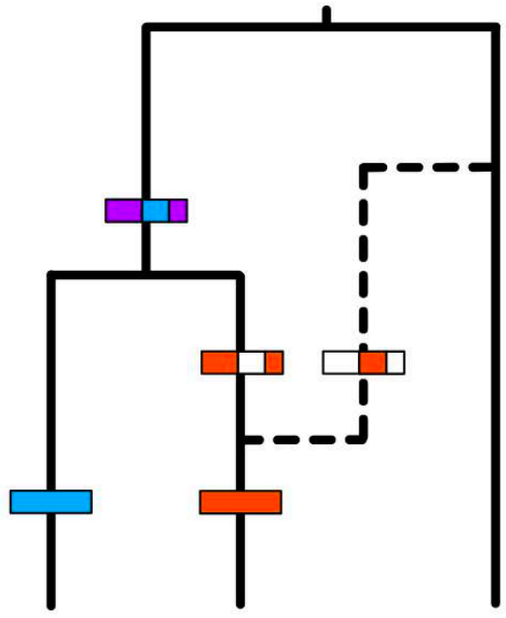
Multi-type time tree

d)



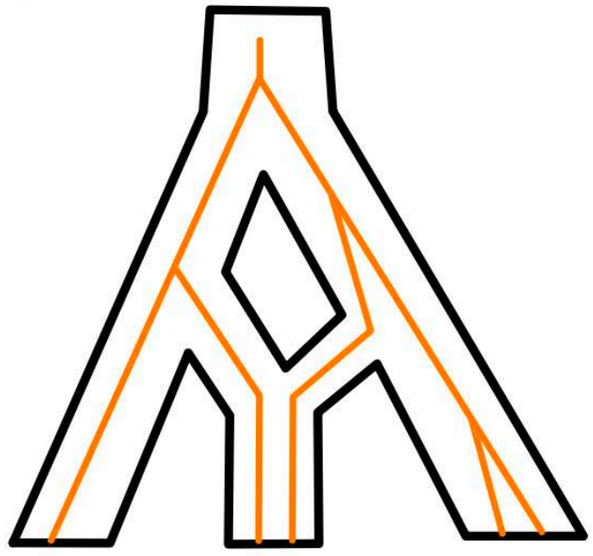
sampled ancestor
time tree

e)



ancestral gene conversion
graph

f)



Species network
with embedded gene tree

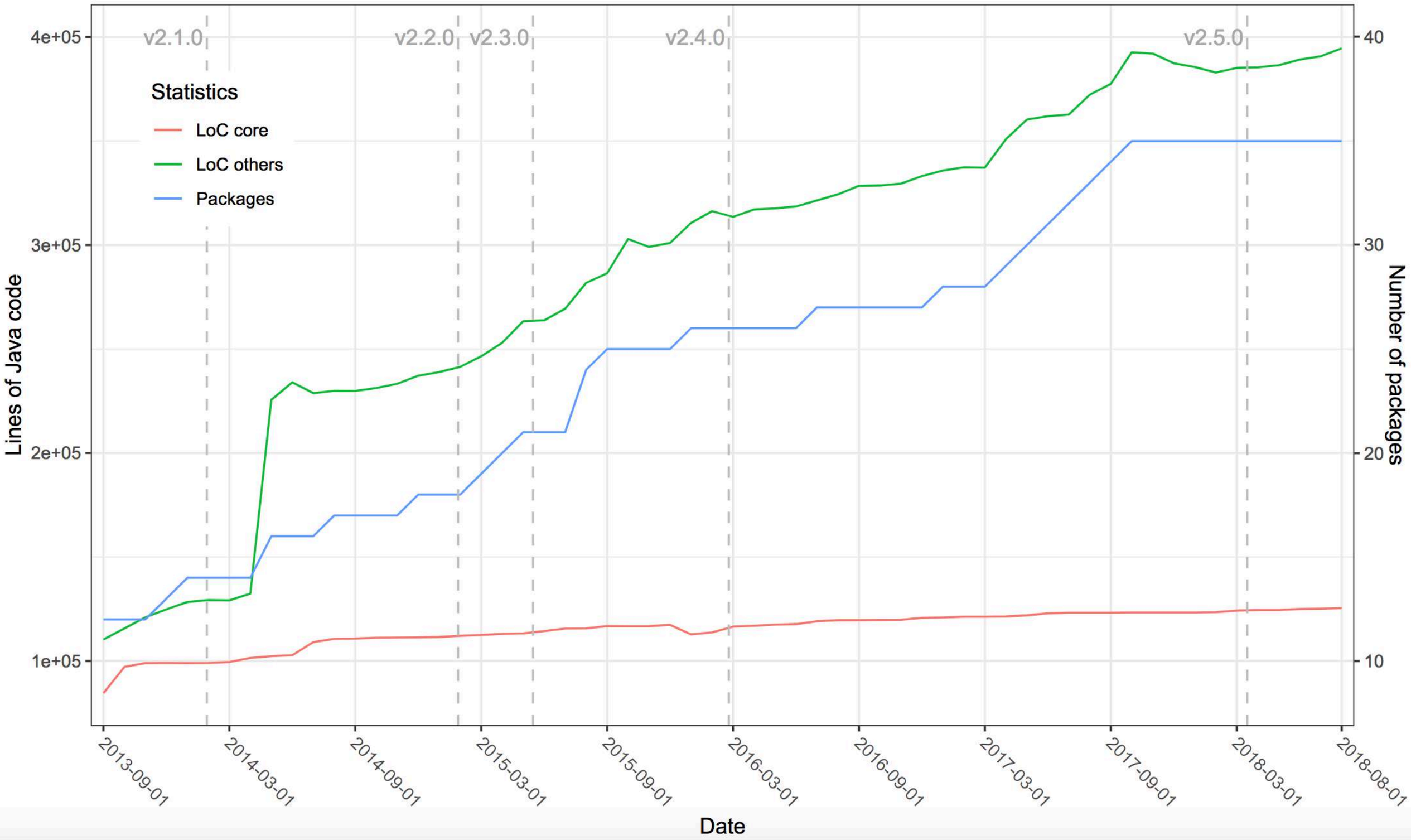
BEAST 2.5 packages

Table 1. BEAST 2 packages

Package	Subspecification	Special Feature	Reference
Substitution models :			
bModelTest	nucleotide subst.[1] model	model averaging, model comparison	[15]
SSM	nucleotide. subst. model	standard named nucleotide models	-
CodonSubstModels	codon subst. model	M0	[16, 17]
MM	morphological model	discrete	[18]
BEASTvntr	microsatellite model	variable number of tandem repeat data	[19, 20]
RBS	subst.[1] model	model averaging for contiguous site partitions	[6]
PoMo	nucleotide subst. model	mutation-selection	[21]
		& species tree	[13]
Site models :			
MGSM	site model	multi-gamma & relaxed gamma	[22]
substBMA	site model	Dirichlet mixture model for site partitions	[8]
Branch model :			
FLC	molecular clock model	strict and relaxed clocks within local clock model	[23]
Tree models :			
SA	unstructured population, non-par.[2]	sampled ancestor* / fossilized BD[3]	[10]
CA	unstructured population, non-par.	calibration density, sampling rate estimate	[24]
BDSKY	unstructured population, non-par.	BD serial skyline*, BD serial sampling	[9]
		BD incomplete sampling (no ψ)	[25]
phylodynamics	unstructured population, par.[2]	deterministic closed SIR, stochastic closed SIR	[26]
		birth-death SIR	[27]
EpiInf	unstructured population, par.	prevalence estimation, particle filtering	[28]
PhyDyn	structured population, par.	define epidemic model by ODEs[4]	[29]
MultiTypeTree	structured population	structured tree	[5]
BadTrIP	structured population	within-host, transmission inference	[14]
BDMM	structured population	multitype BD[3] model	[30]
BASTA	structured population	approx. structured coalescent	[31]
MASCOT	structured population	approx. structured coalescent and time variant GLM's	[32, 33]
SCOTTI	structured population	transmission inference	[34]
BREAK AWAY	geographical model	break-away model of phylogegraphy	[35]
GEO SPHERE	geographical model	whole world phylogeography	[36]
Network models :			
BACTER	network model	clonal frame ancestral recombination graph	[11, 37]
SpeciesNetwork	network model	species networks	[12]
Nested models :			
DENIM	multispecies coalescent	species tree estimation with gene flow	[38]
SNAPP	multispecies coalescent	from independent biallelic markers	[7]
STACEY	multispecies coalescent	species delimitation & species tree estimation	[39]
StarBEAST 2	multispecies coalescent	faster, species tree clocks, FBD-MSC, AIM	[40–43]
Model selection :			
MODEL SELECTION	model selection	path sampling, stepping stone	[44]
NS	model selection	nested sampling	[45]
Simulation tools :			
MASTER	simulation	stochastic population dynamics simulation	[46]

*

BEAST 2.5 development

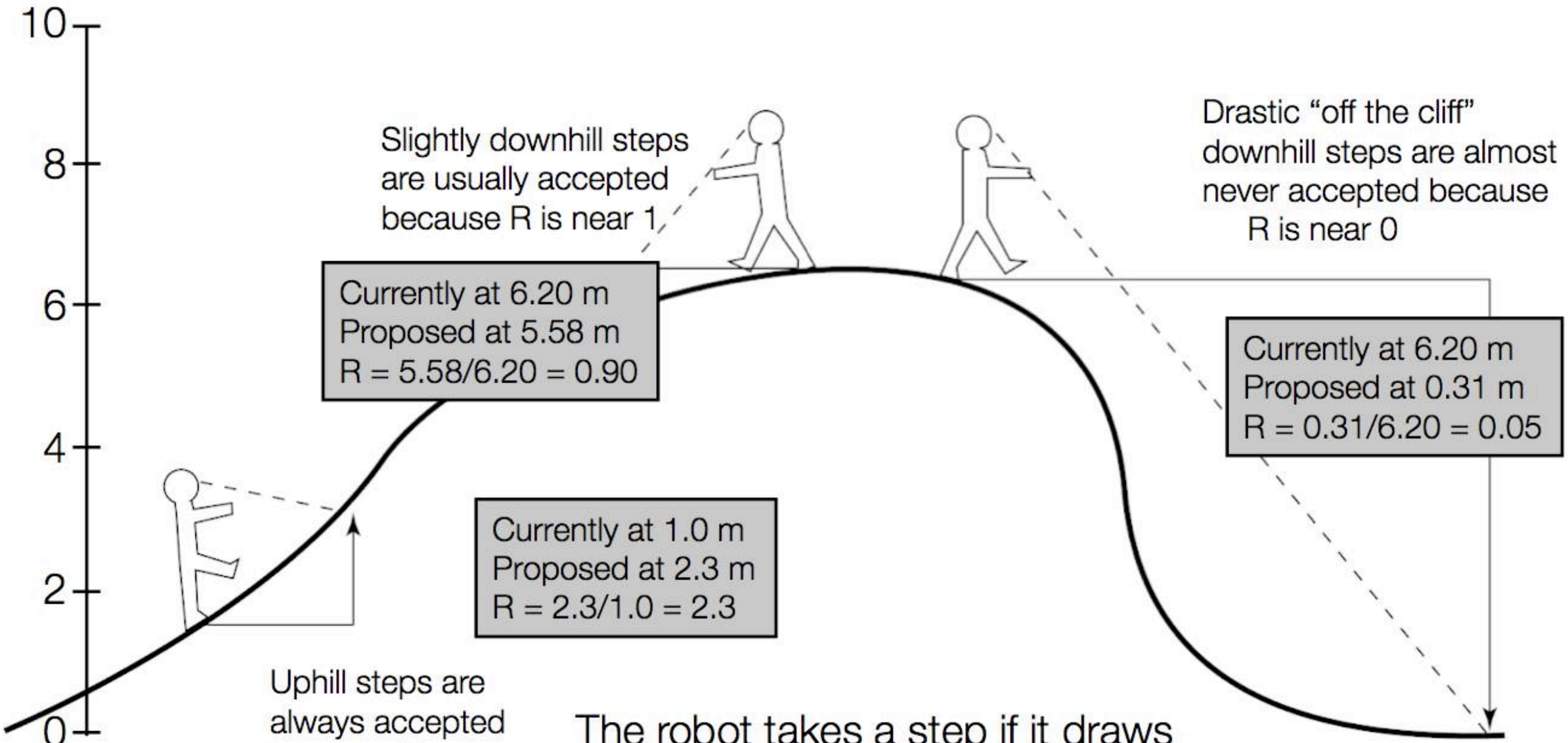


BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis

Remco Bouckaert[1*], Timothy G. Vaughan[2], Joelle Barido-Sottani[2], Sebastian Duchene[3], Mathieu Fourment[4], Alexandra Gavryushkina[5], Joseph Heled, Graham Jones, Denise Kühnert[6], Nicola De Maio[7], Michael Matschiner[8], Nicola Müller[2], Huw Ogilvie[9], Louis du Plessis[10], Alexandra Popinga[1], Andrew Rambaut[11], David Rasmussen[12], Igor Siveroni[13], Marc A. Suchard[14], Erik Volz[13], Chieh-Hsi Wu[15], Dong Xie[1], Chi Zhang[16], Tanja Stadler[2], Alexei J. Drummond[1*]

1 Centre of Computational Biology, University of Auckland, Auckland, New Zealand
2 ETH Zurich, Department of Biosystems Science and Engineering, 4058 Basel, Switzerland
3 University of Melbourne, Melbourne, Victoria, Australia
4 ithree institute, University of Technology Sydney, Sydney, Australia
5 University of Otago, Dunedin, New Zealand
6 Max Planck Institute, Jena, Germany
7 European Bioinformatic Institute, Cambridge, CB10 1SD, UK
8 University of Basel, Basel, Switzerland
9 Rice University, Houston, Texas, USA
10 Department of Zoology, University of Oxford, Oxford, UK
11 Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, Edinburgh, EH9 3FL UK
12 North Carolina State University, North Carolina, USA
13 Department of Infectious Disease Epidemiology, Imperial College London, Norfolk Place, W2 1PG, UK
14 Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA
15 Department of Statistics, University of Oxford, OX1 3LB, UK
16 Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing, China

Markov chain Monte Carlo (MCMC) robot

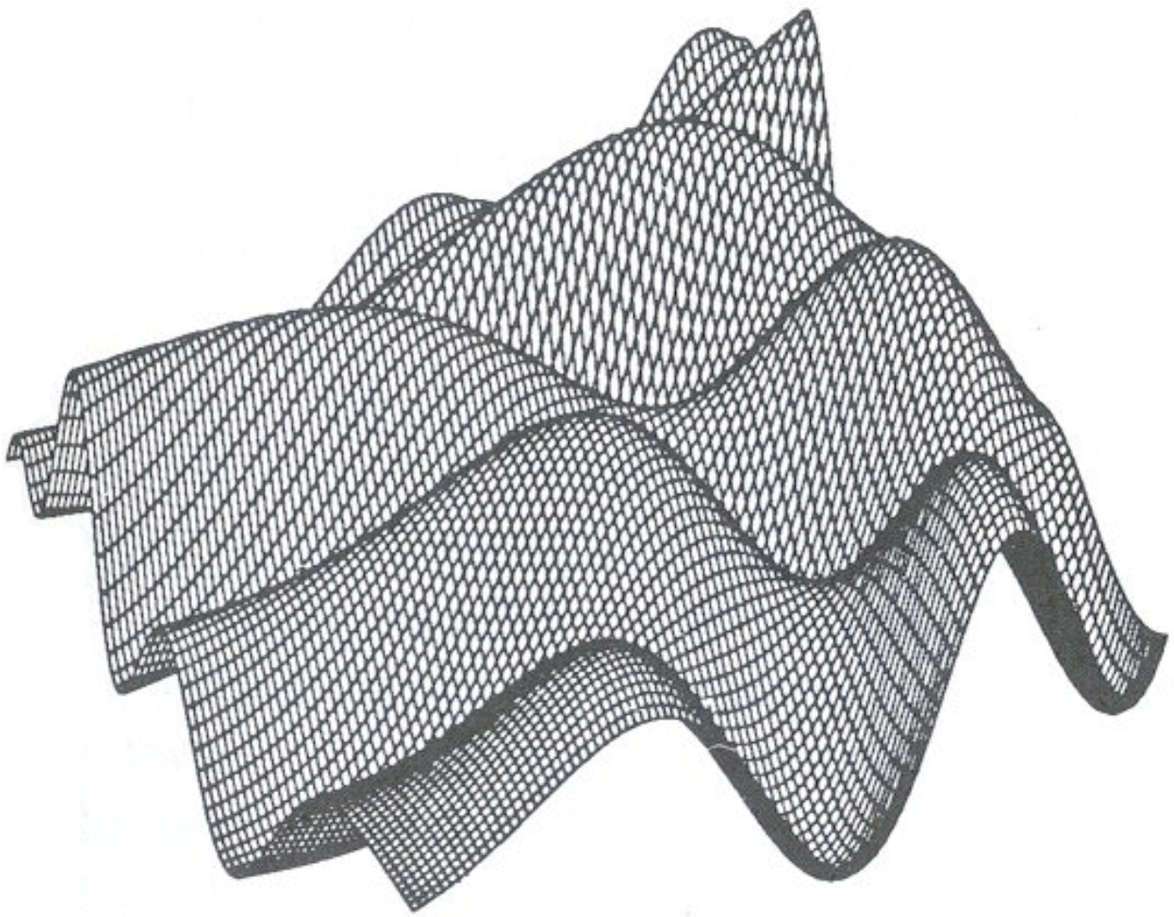


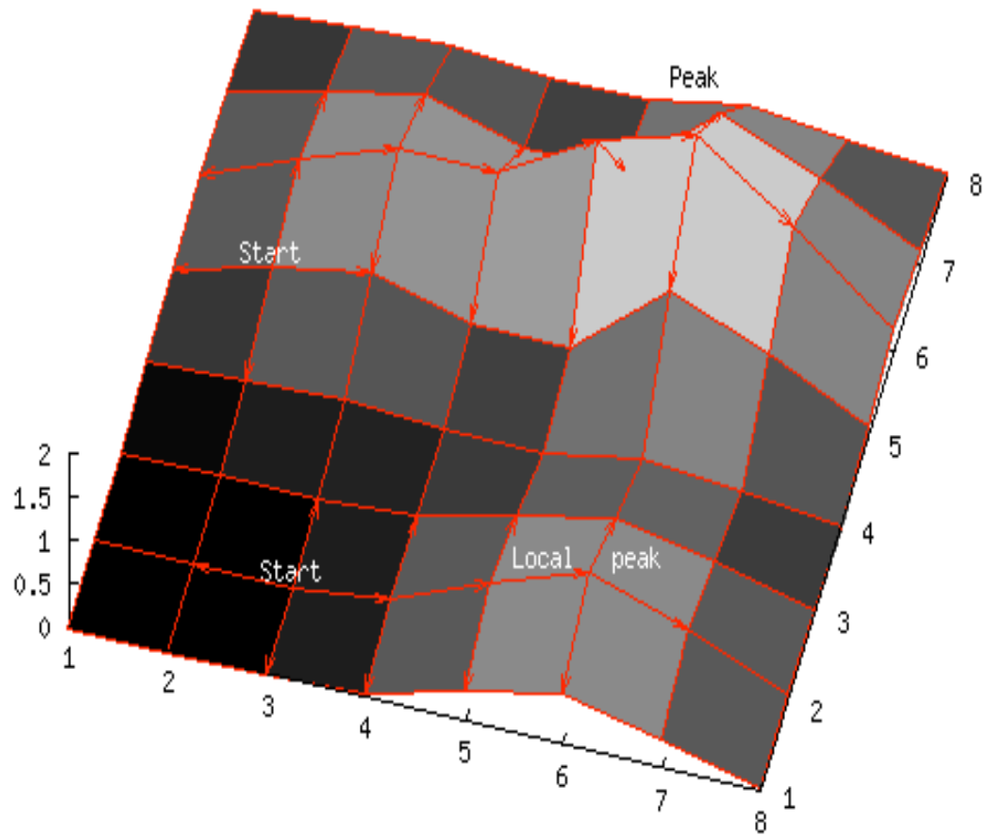
The robot takes a step if it draws a random number (uniform on 0.0 to 1.0), and that number is less than or equal to R

Slide courtesy of Paul O Lewis

MCMC animations

A probability distribution on tree space as a hilly landscape



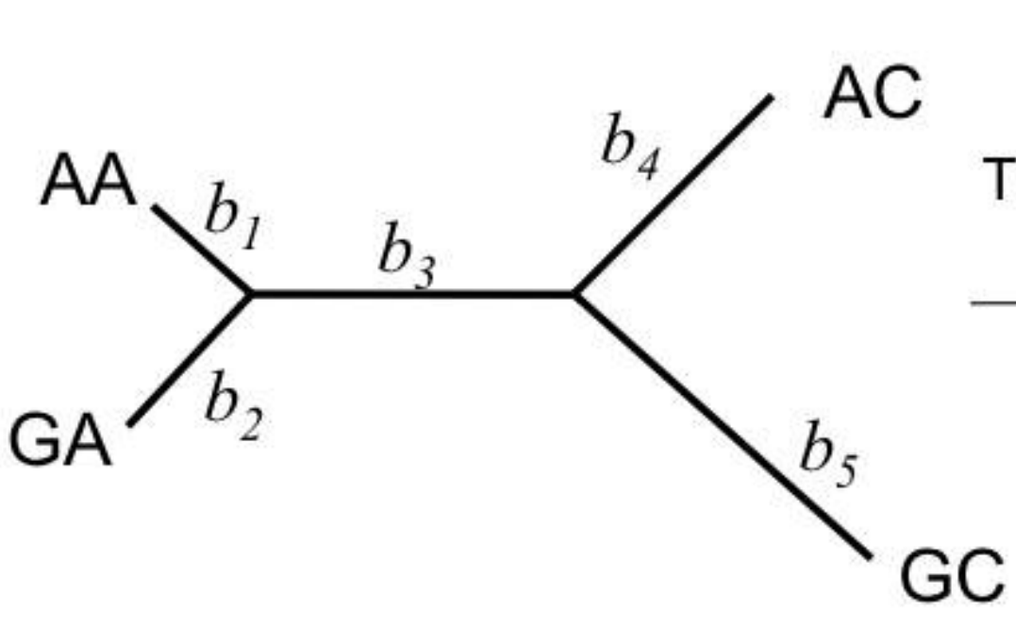


- This space can be **sampled** in a Bayesian analysis with MCMC
- The peak can be identified by a **search algorithm** in the context of maximum likelihoods

Clocks and calibrations

The molecular clock constraint

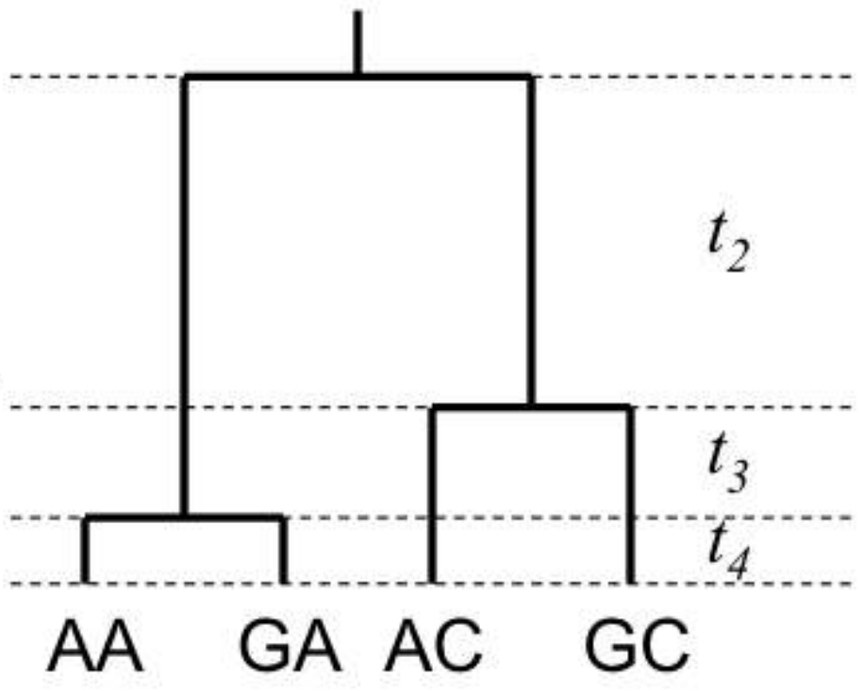
T



2n–3 branch lengths

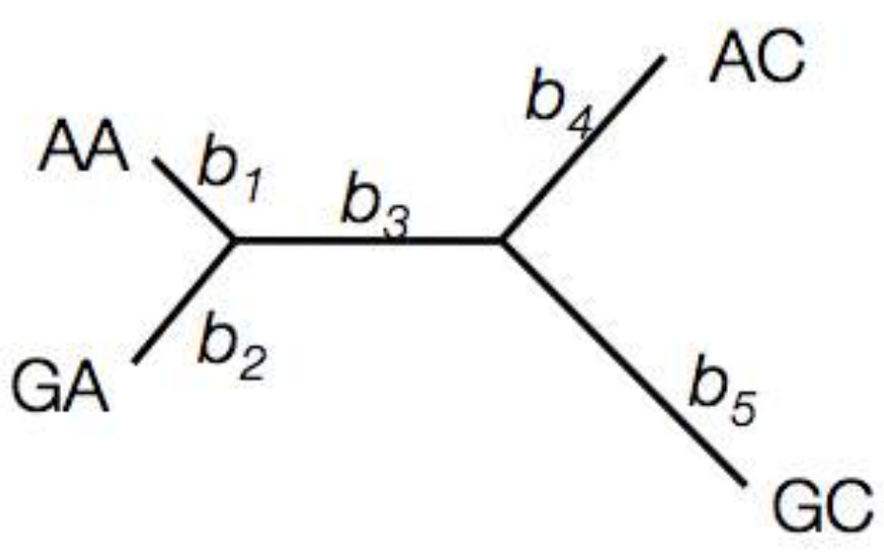
The "molecular clock" constraint

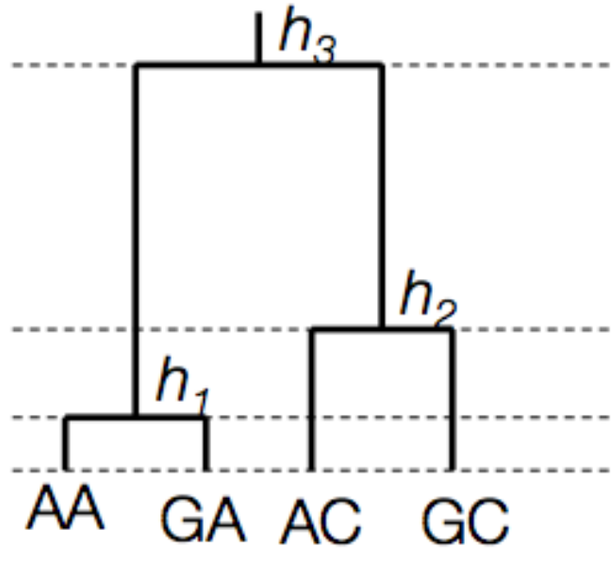
g



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:

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

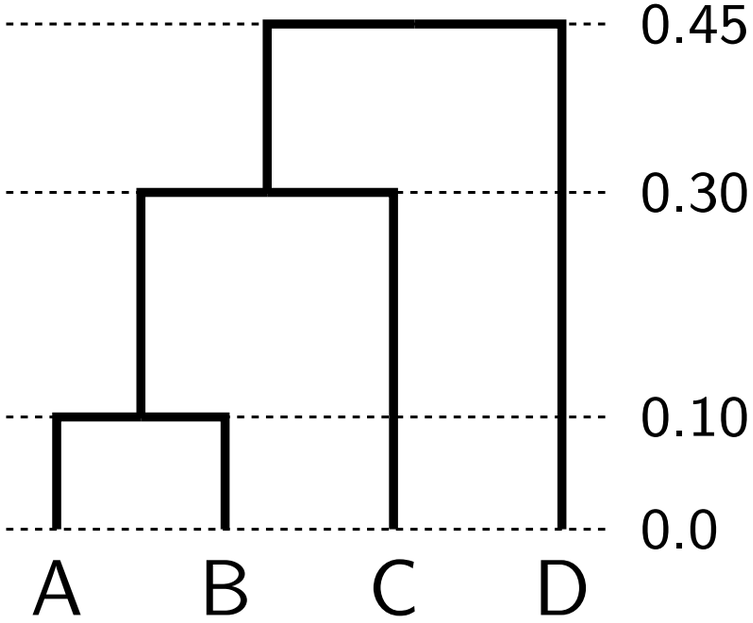
Estimate the global rate:

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

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

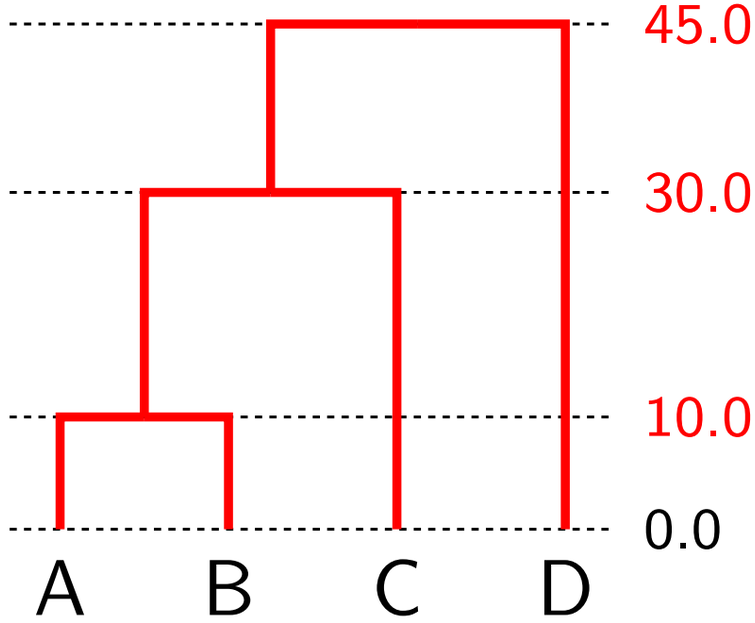
$$T = \mu \times g$$



"substitution tree"

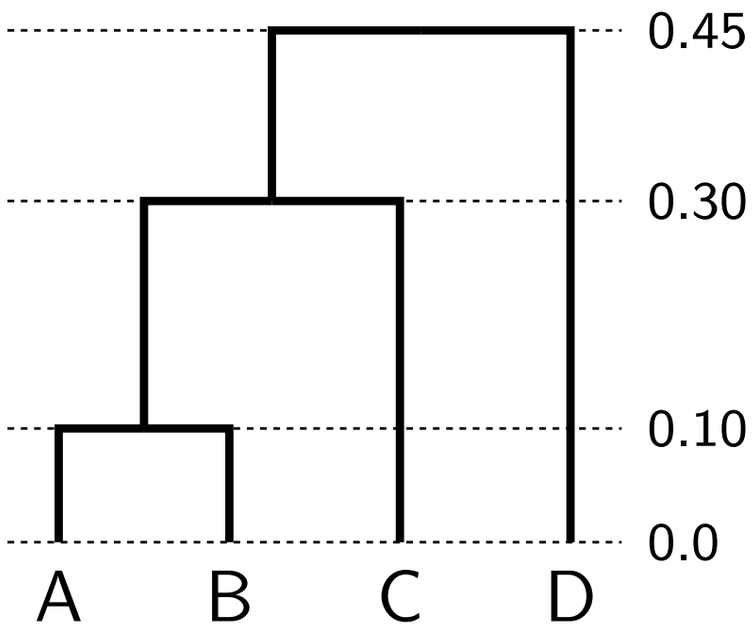
$=$ 0.01 $\times$

evolutionary rate
substitutions / site / unit time

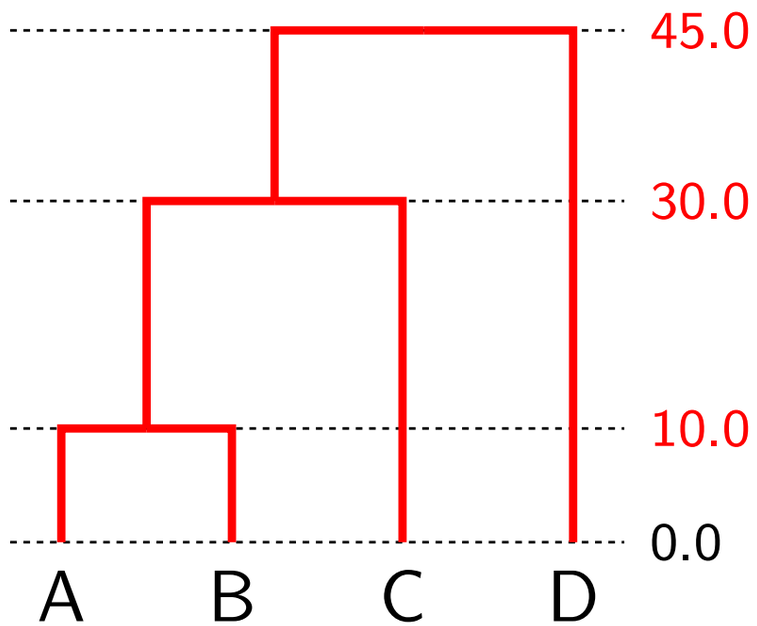


time tree

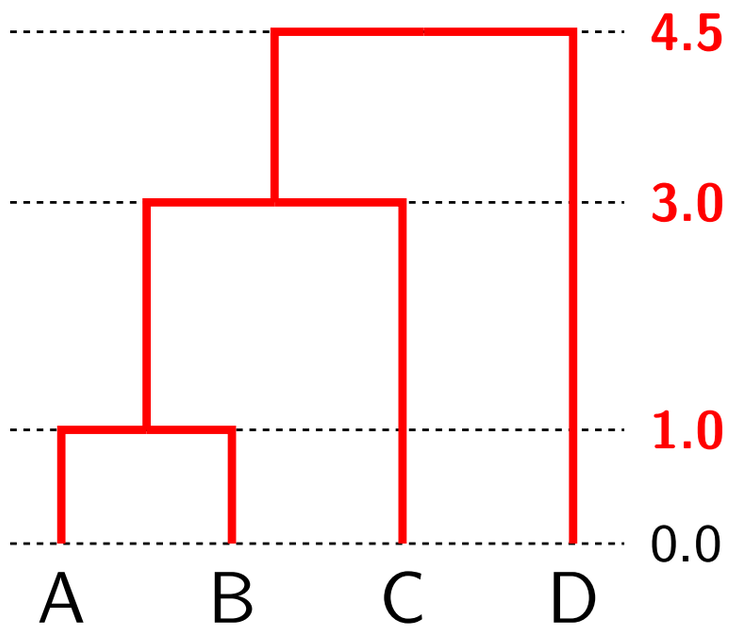
Non-identifiability of rates and times



$=$ 0.01 $\times$



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



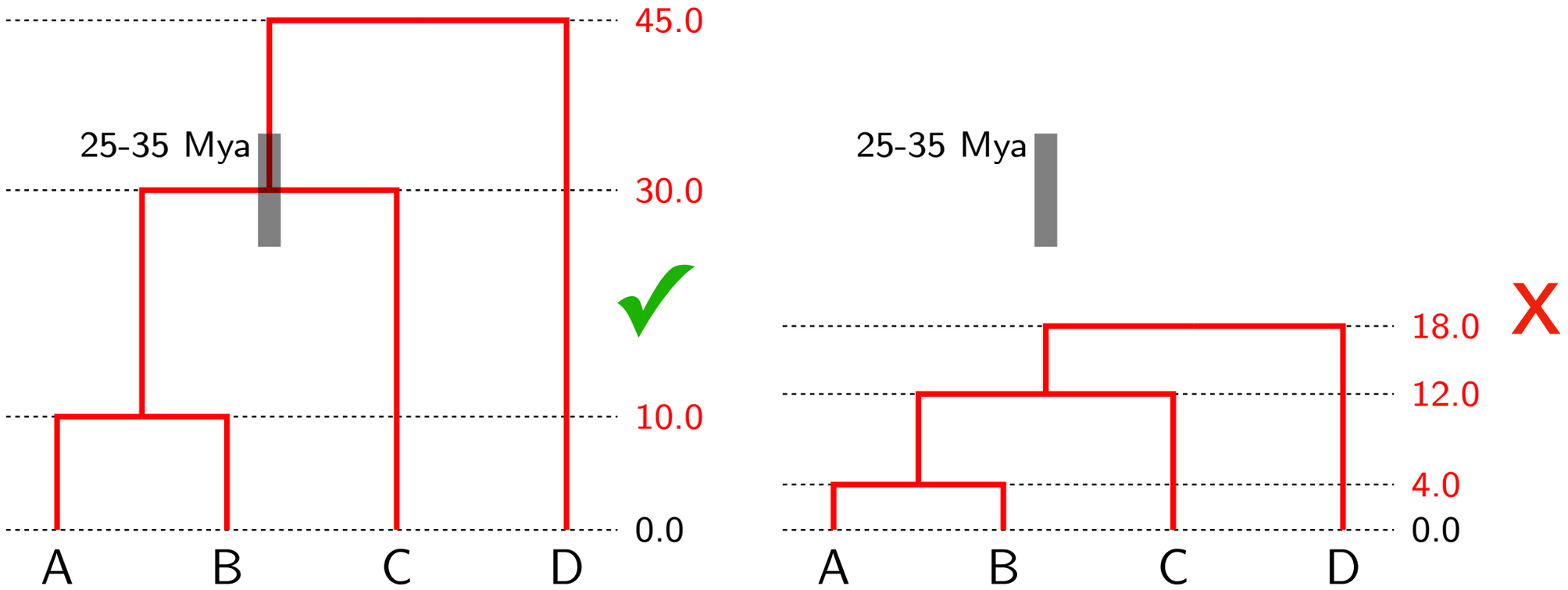
"substitution tree"

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**.

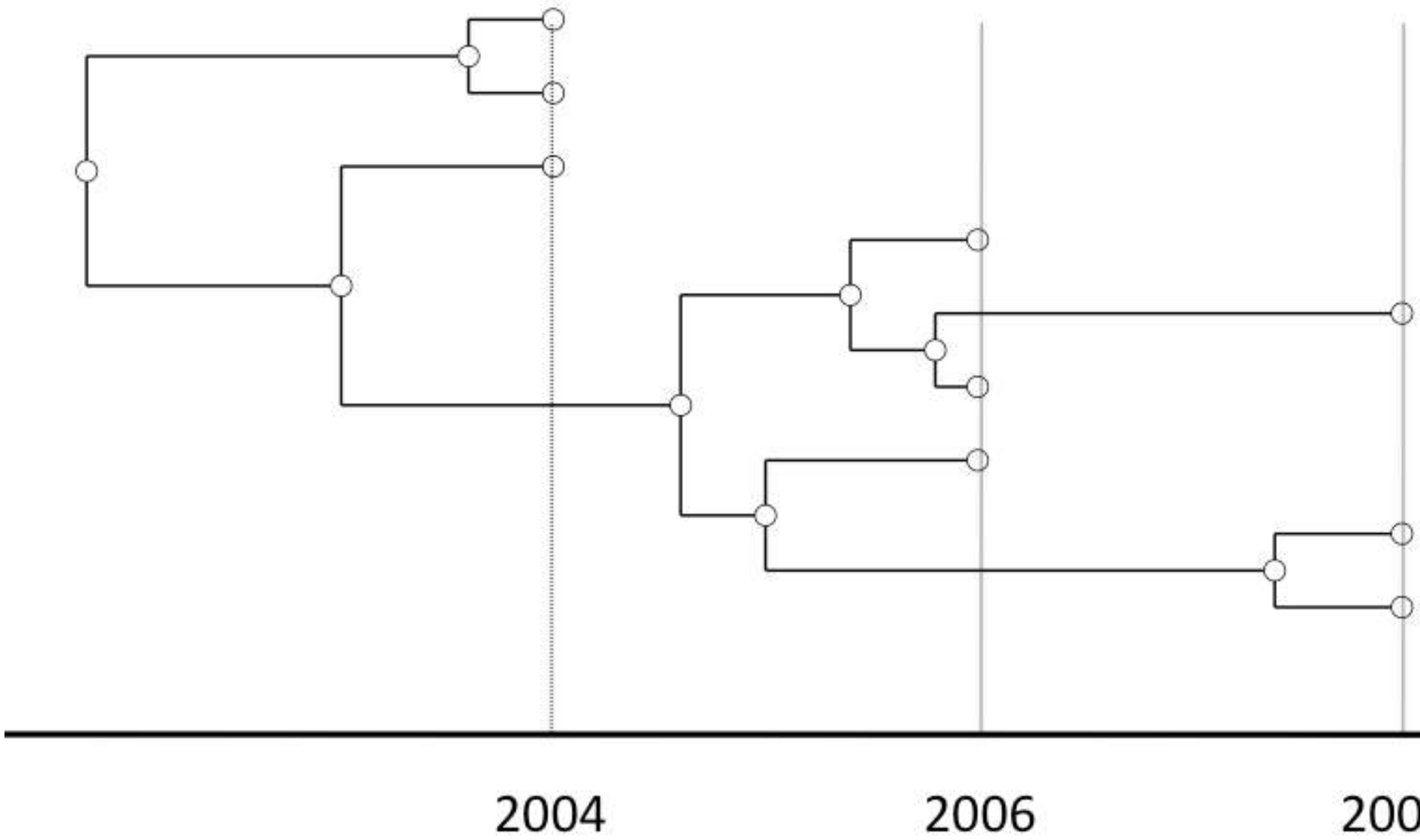


With a strict molecular clock, only the age (range) of a single node in the tree is needed in order to interpolate and extrapolate the ages of all other divergence times.

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

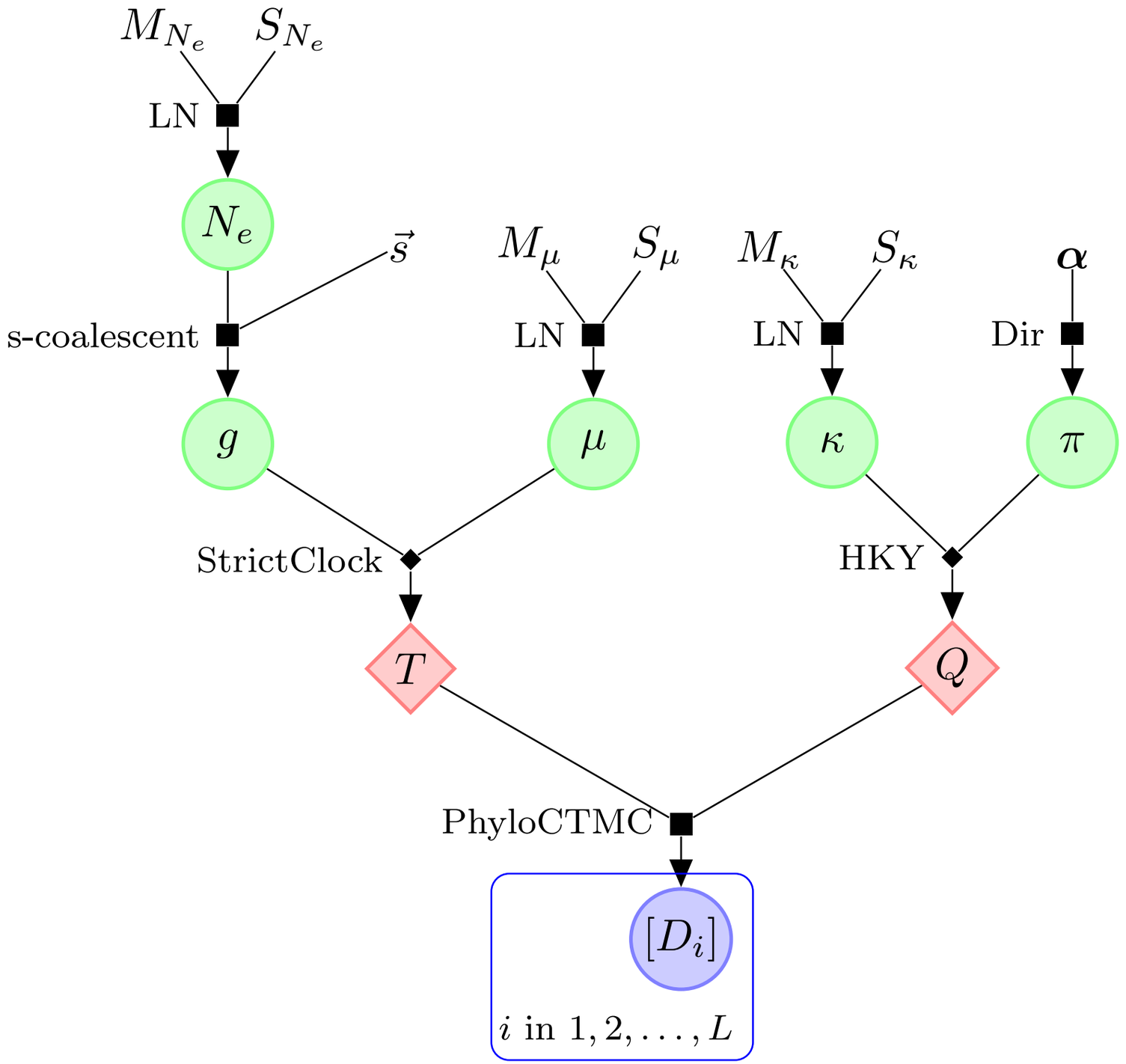
Drummond *et al* (2002)



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

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

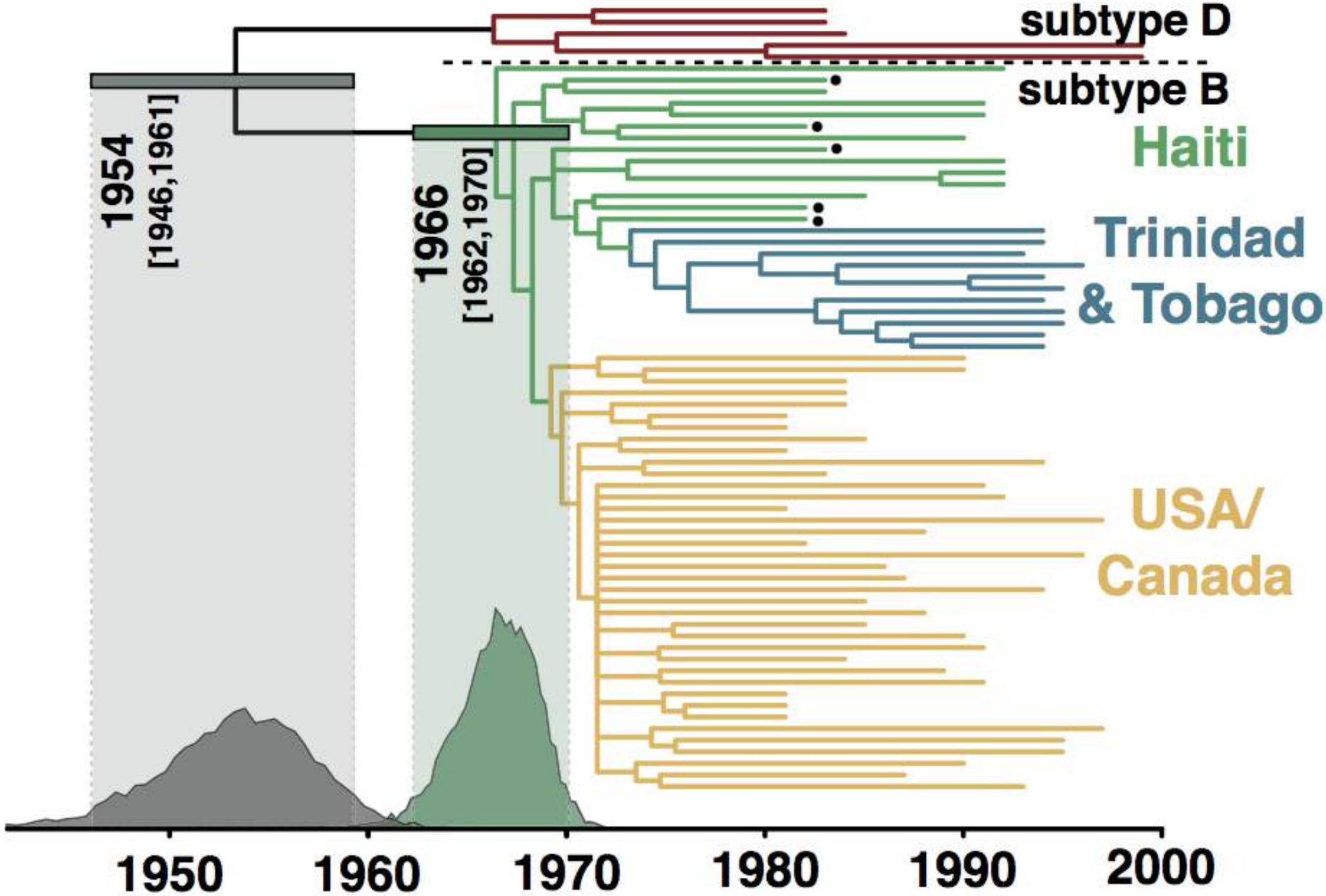
Bayesian s-coalescent model with uncertain times



$$P(g, \mu, N_e, Q|\boldsymbol{D}, \vec{s}) \propto \Pr(\boldsymbol{D}|g \cdot \mu, Q) P(g|N_e, \vec{s}) P(N_e) P(\mu) P(Q)$$

A calibrated phylogenetic inference

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



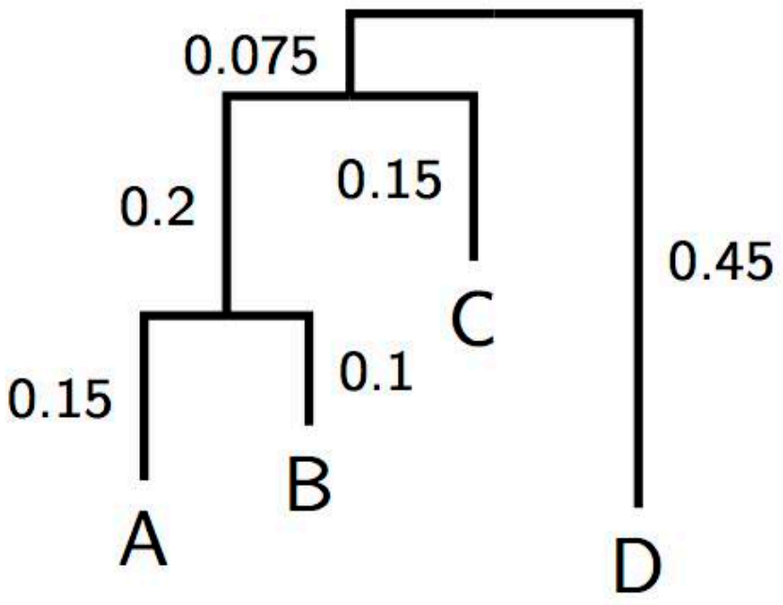
A phylogenetic reconstruction of samples of HIV-1 virus. Each tip represents a single infected individual from whom a blood sample has been taken.

Relaxed phylogenetics

Genetic distance = rate x time

Relaxed molecular clock

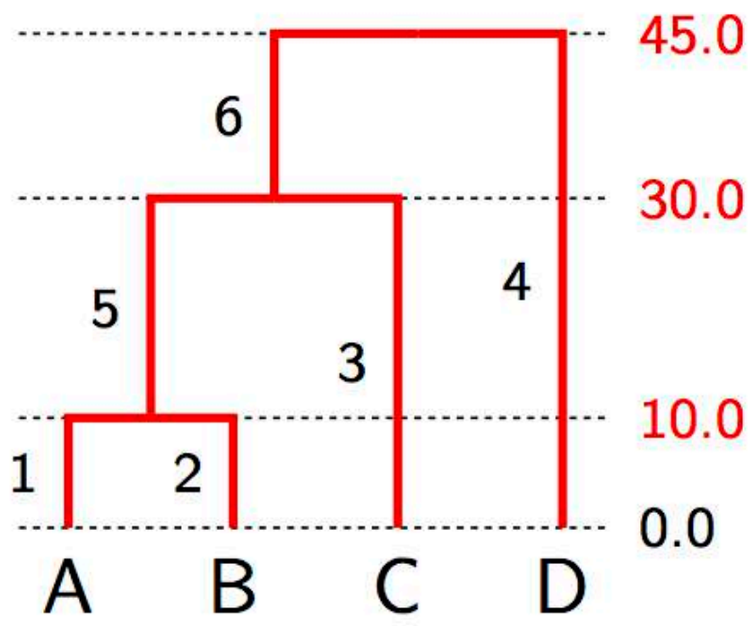
$$T = \vec{\mu} \star g$$



"substitution tree"

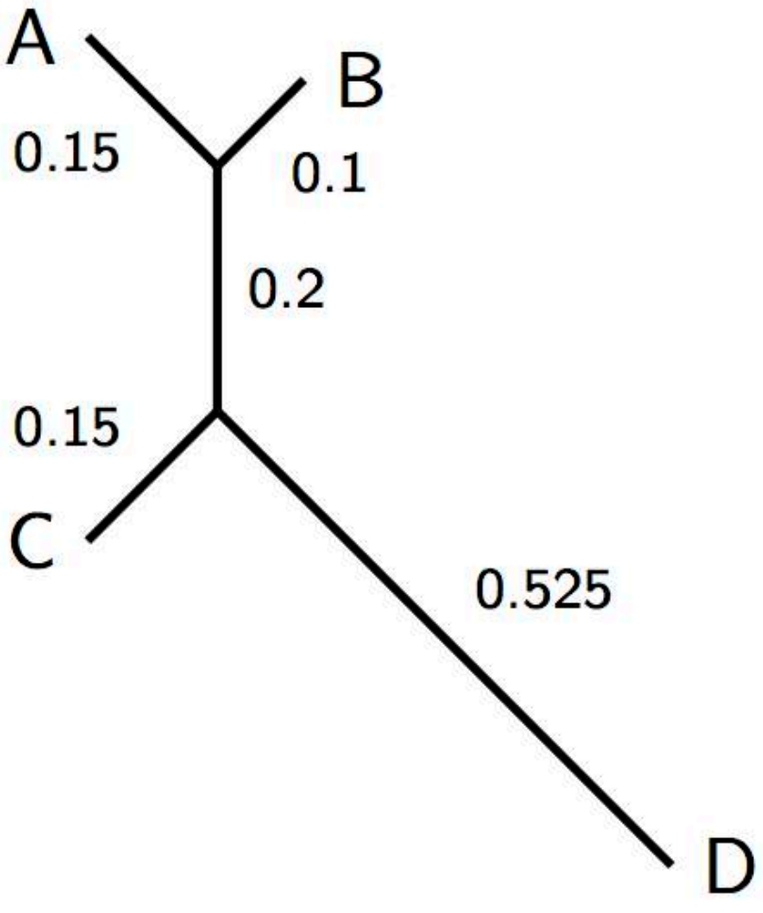
$$= \begin{pmatrix} 0.015 \\ 0.01 \\ 0.005 \\ 0.01 \\ 0.01 \\ 0.005 \end{pmatrix} \star$$

evolutionary rates
substitutions / site / unit time



time tree

Genetic distance = rate x time

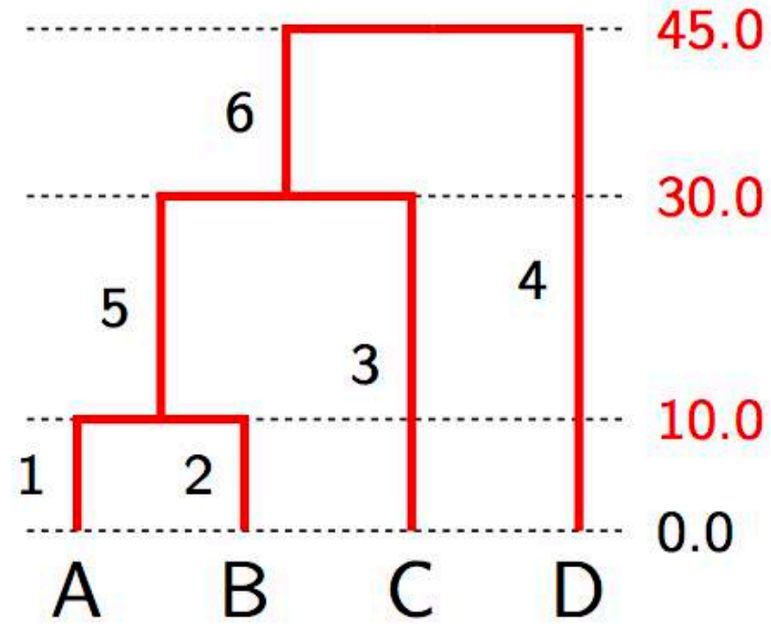


"substitution tree"

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

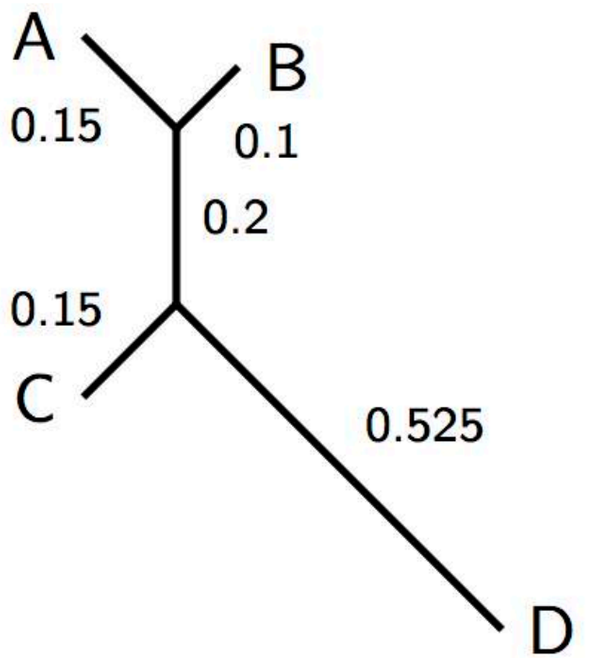
$$= \begin{pmatrix} 0.015 \\ 0.01 \\ 0.005 \\ 0.01 \\ 0.01 \\ 0.005 \end{pmatrix} \star$$

evolutionary rates
substitutions / site / unit time

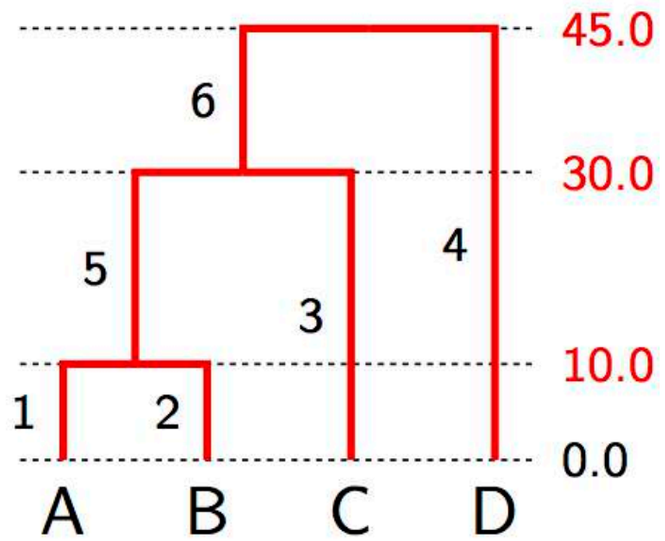


time tree

Non-identifiability of rates and times

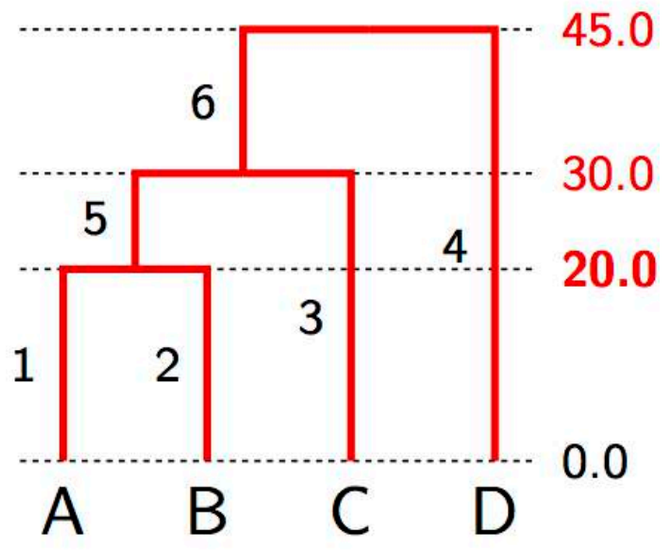


$$= \begin{pmatrix} 0.015 \\ 0.01 \\ 0.005 \\ 0.01 \\ 0.01 \\ 0.005 \end{pmatrix} \star$$



"substitution tree"

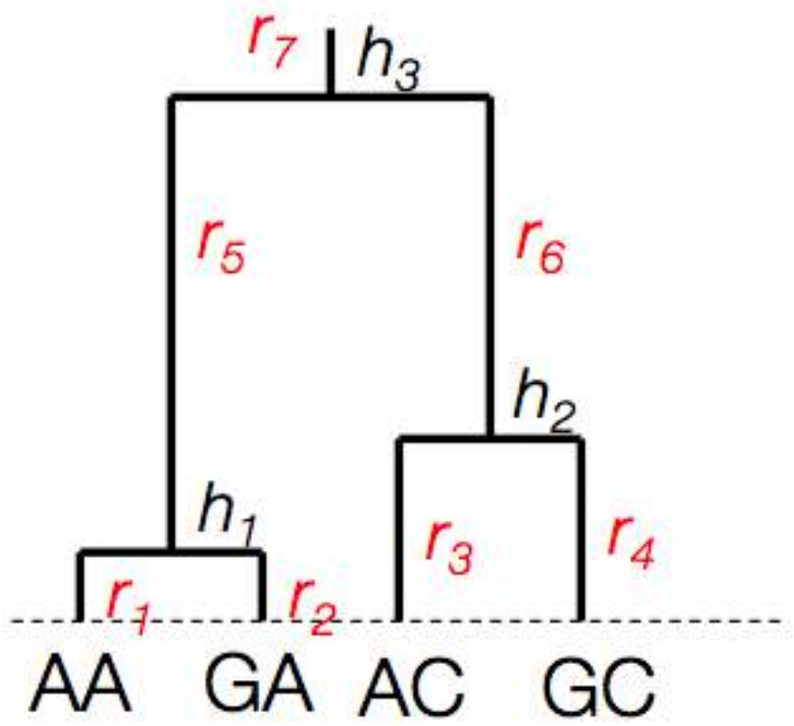
$$= \begin{pmatrix} \mathbf{0.0075} \\ \mathbf{0.005} \\ 0.005 \\ 0.01 \\ \mathbf{0.02} \\ 0.005 \end{pmatrix} \star$$



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

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

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

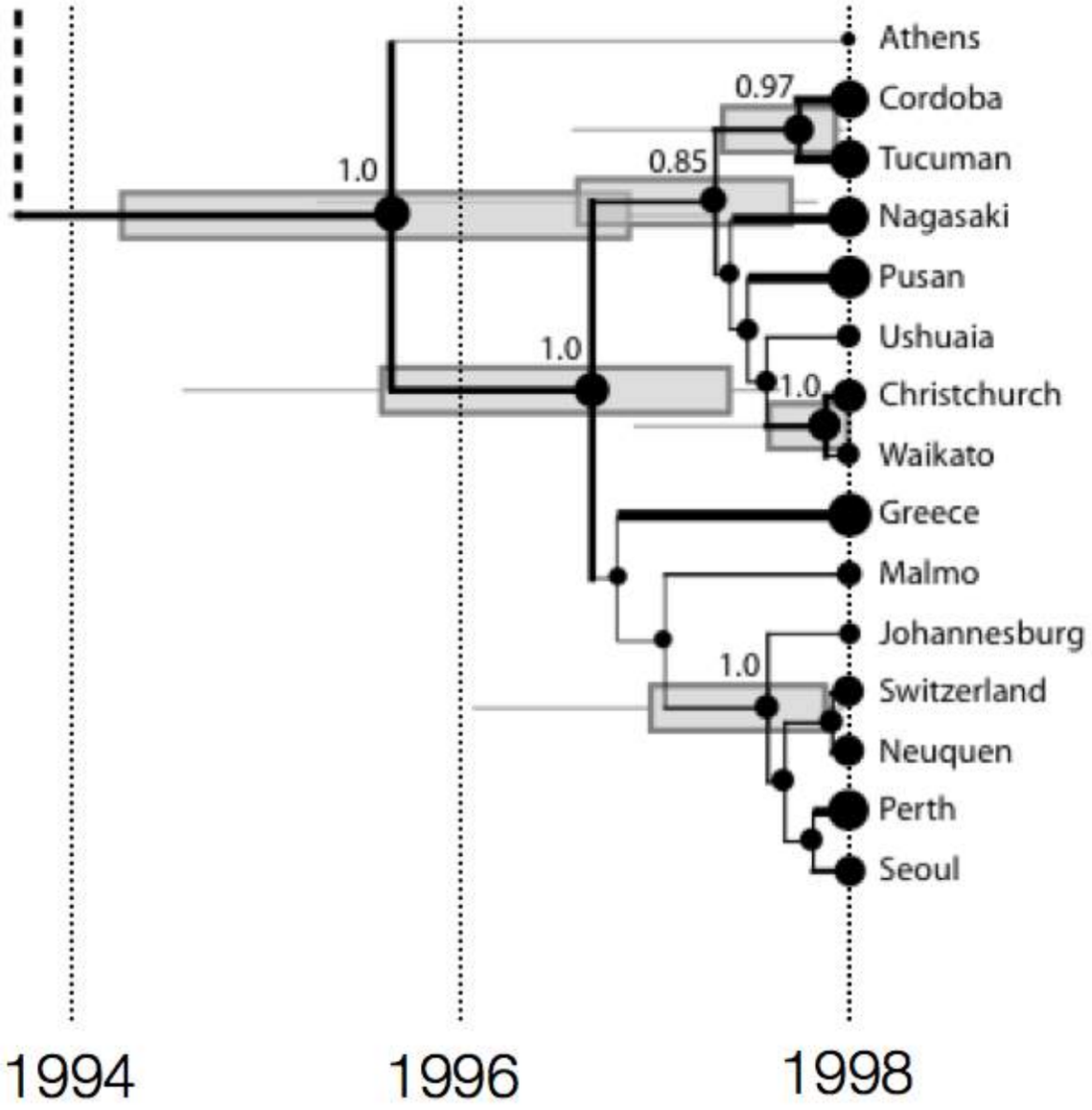
$$r \sim Gamma(\alpha, \beta)$$

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

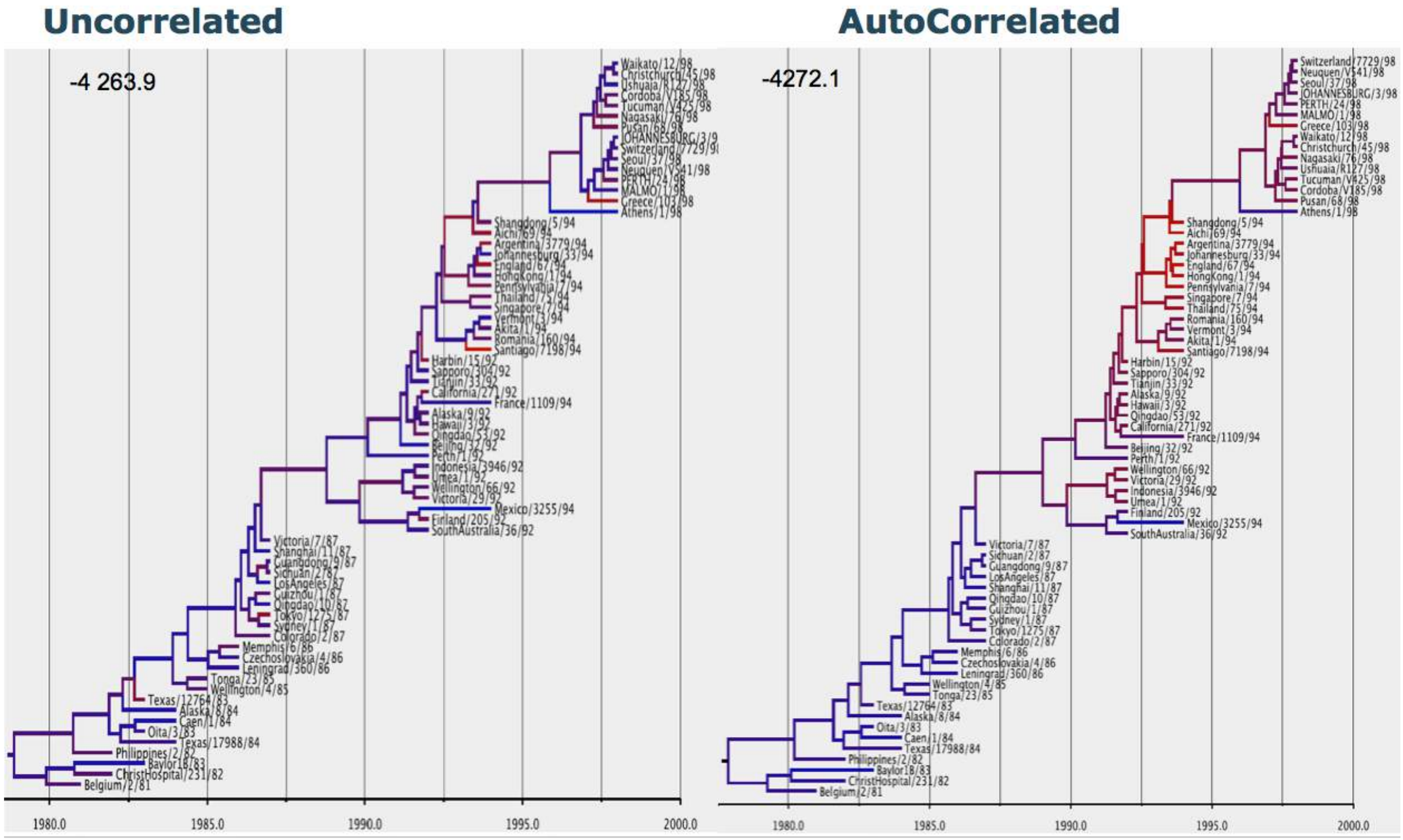
ML

Influenza A gene tree estimating using relaxed molecular clock

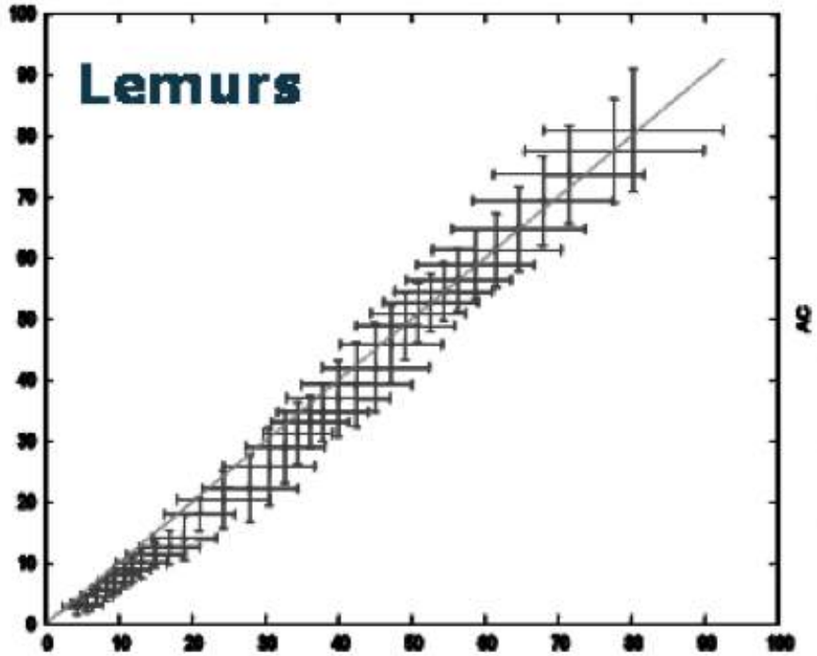


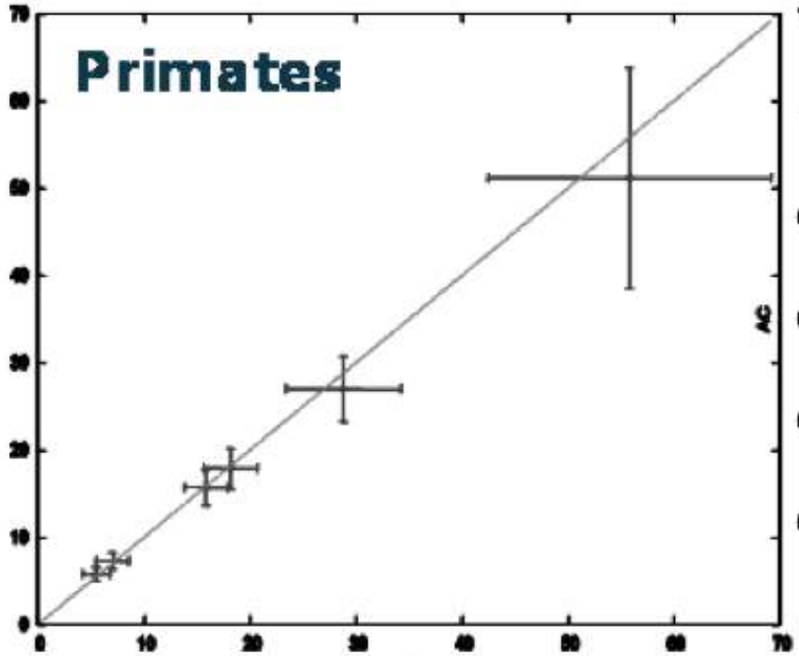
- Box-and-whisker plots show uncertainty in divergence times (only for splits with posterior probability > 0.5)
- Node size and branch thickness proportional to evolutionary rate.

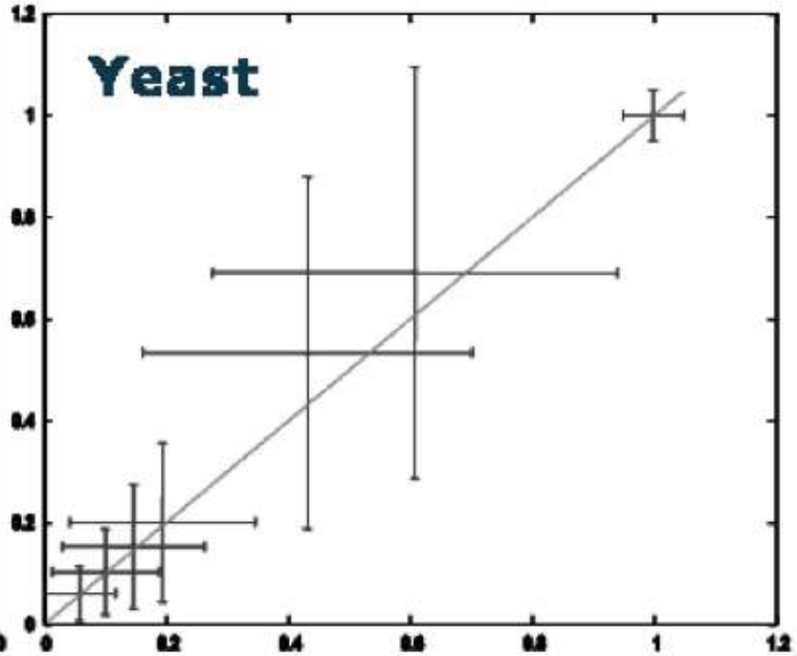
Influenza trees under different relaxed molecular clocks

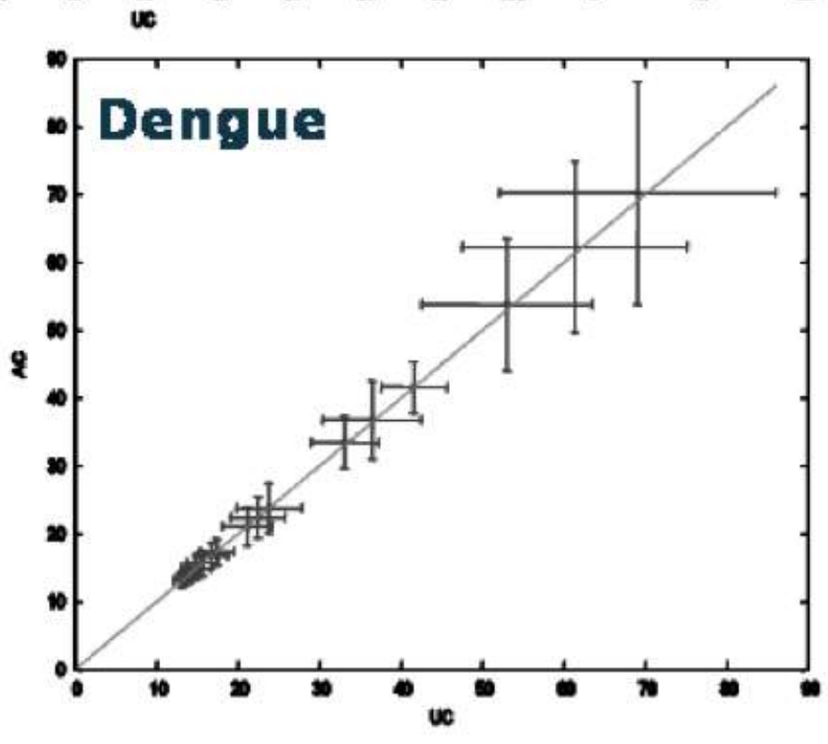


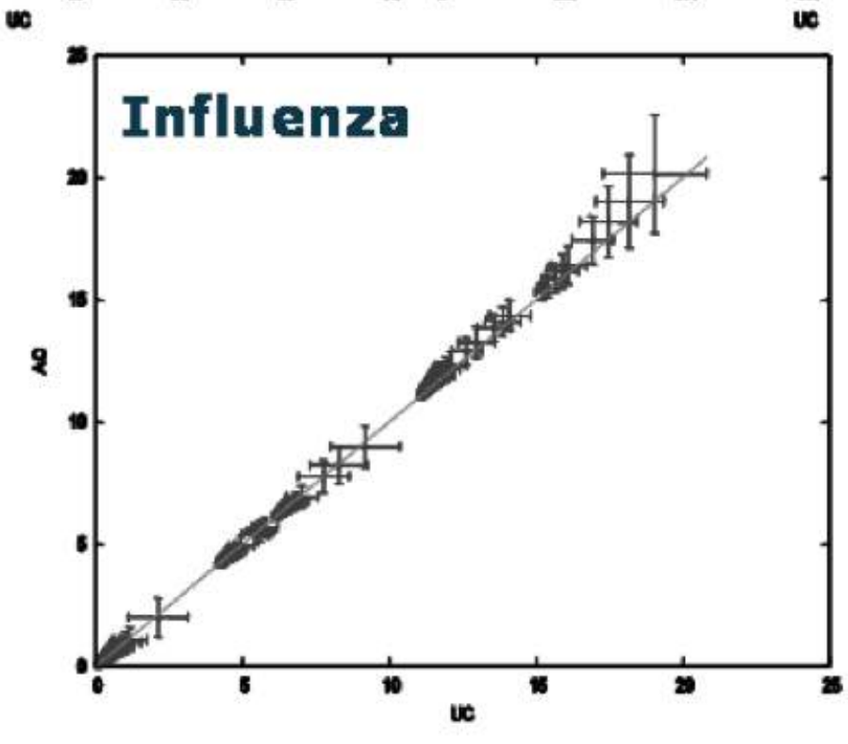
UC versus AC on five data sets



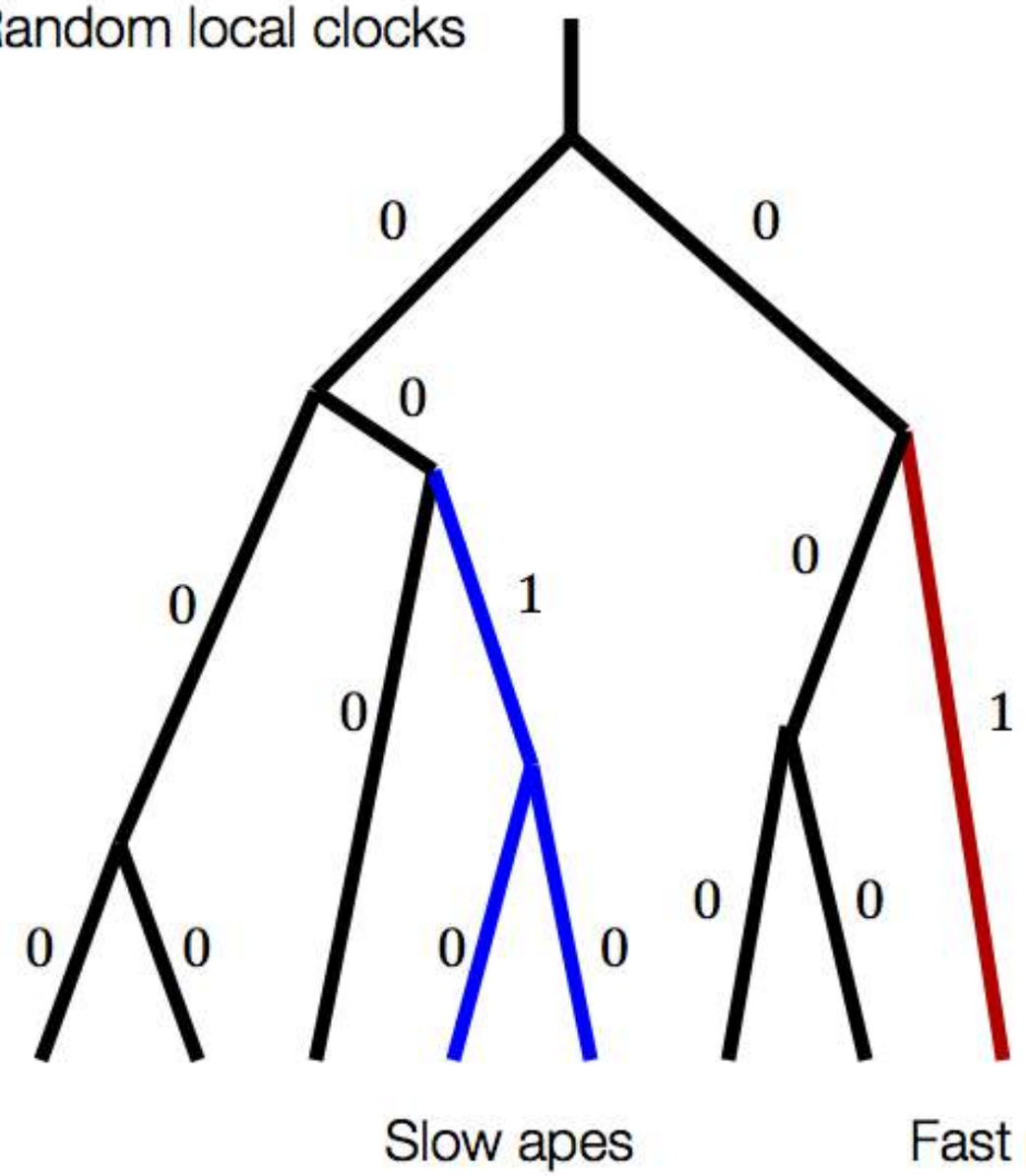








Random local molecular clocks



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

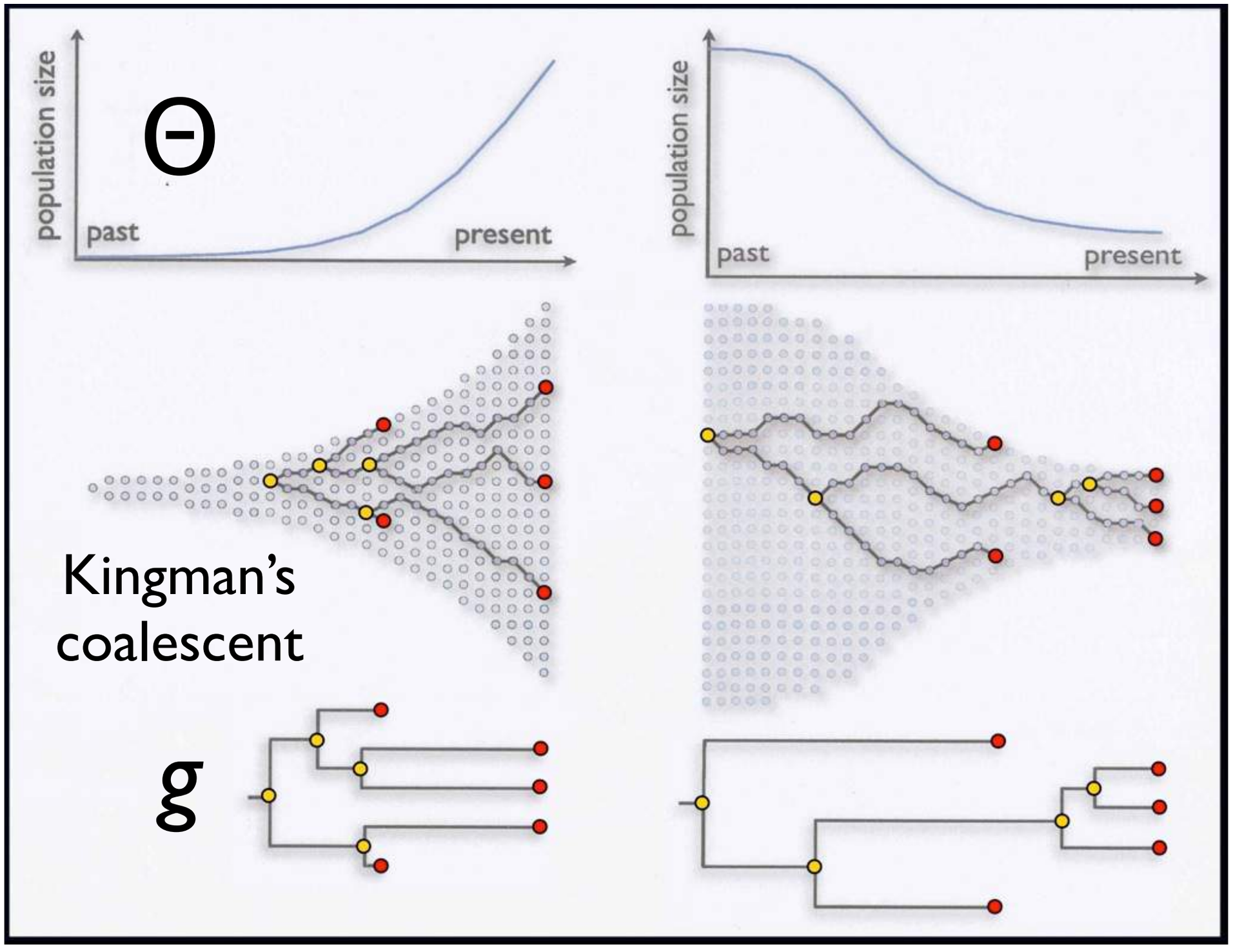
Tree priors

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)

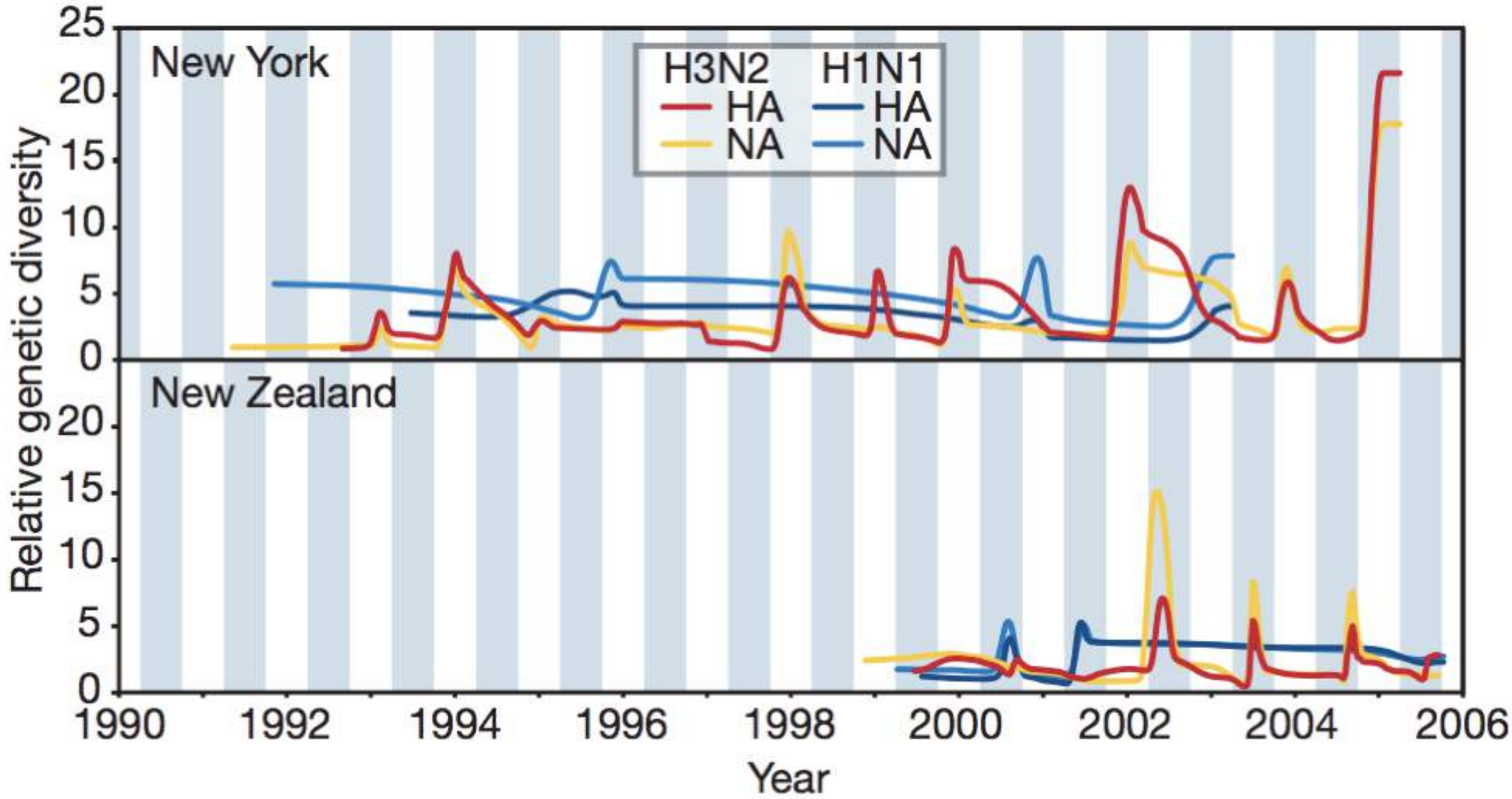
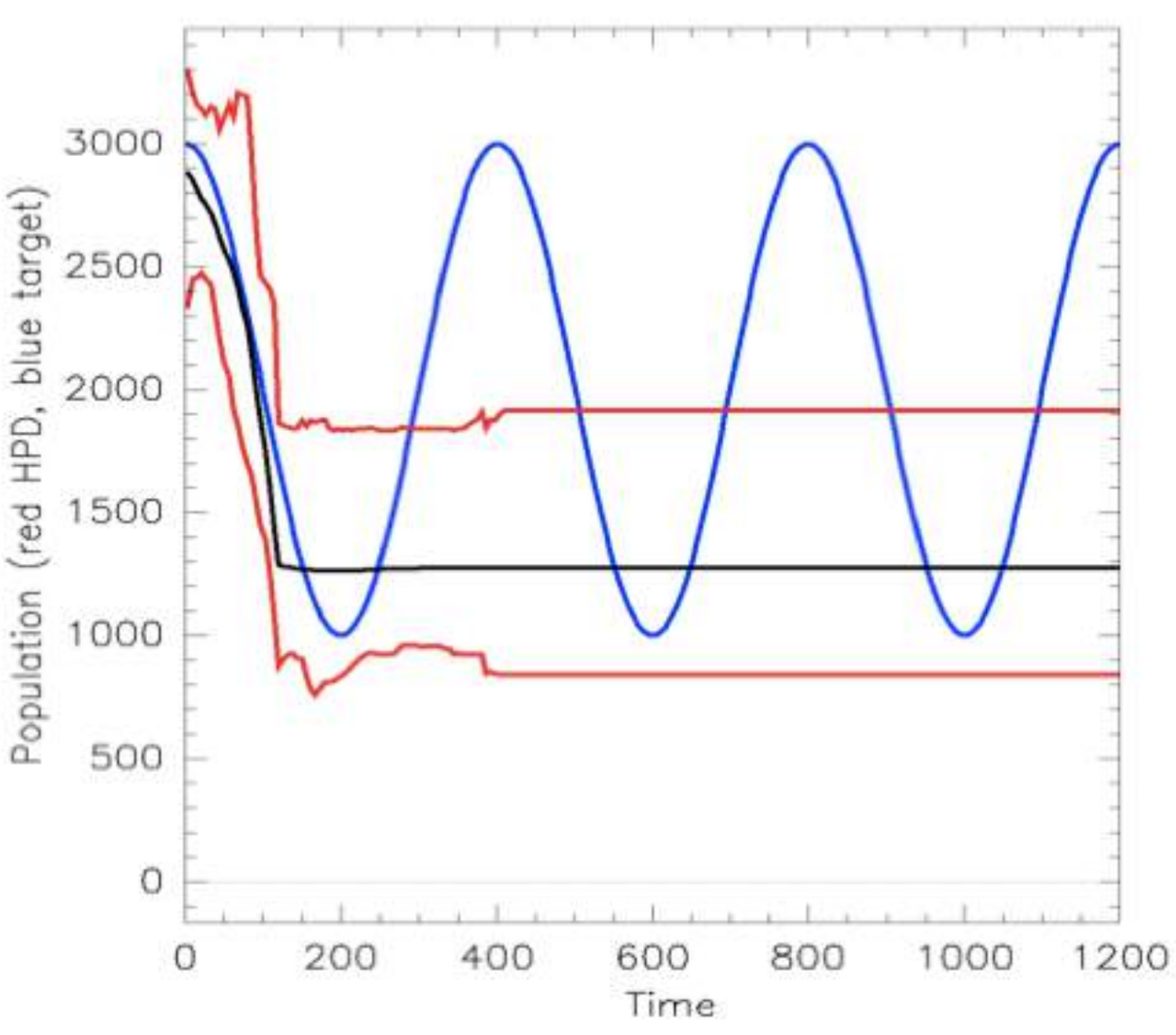


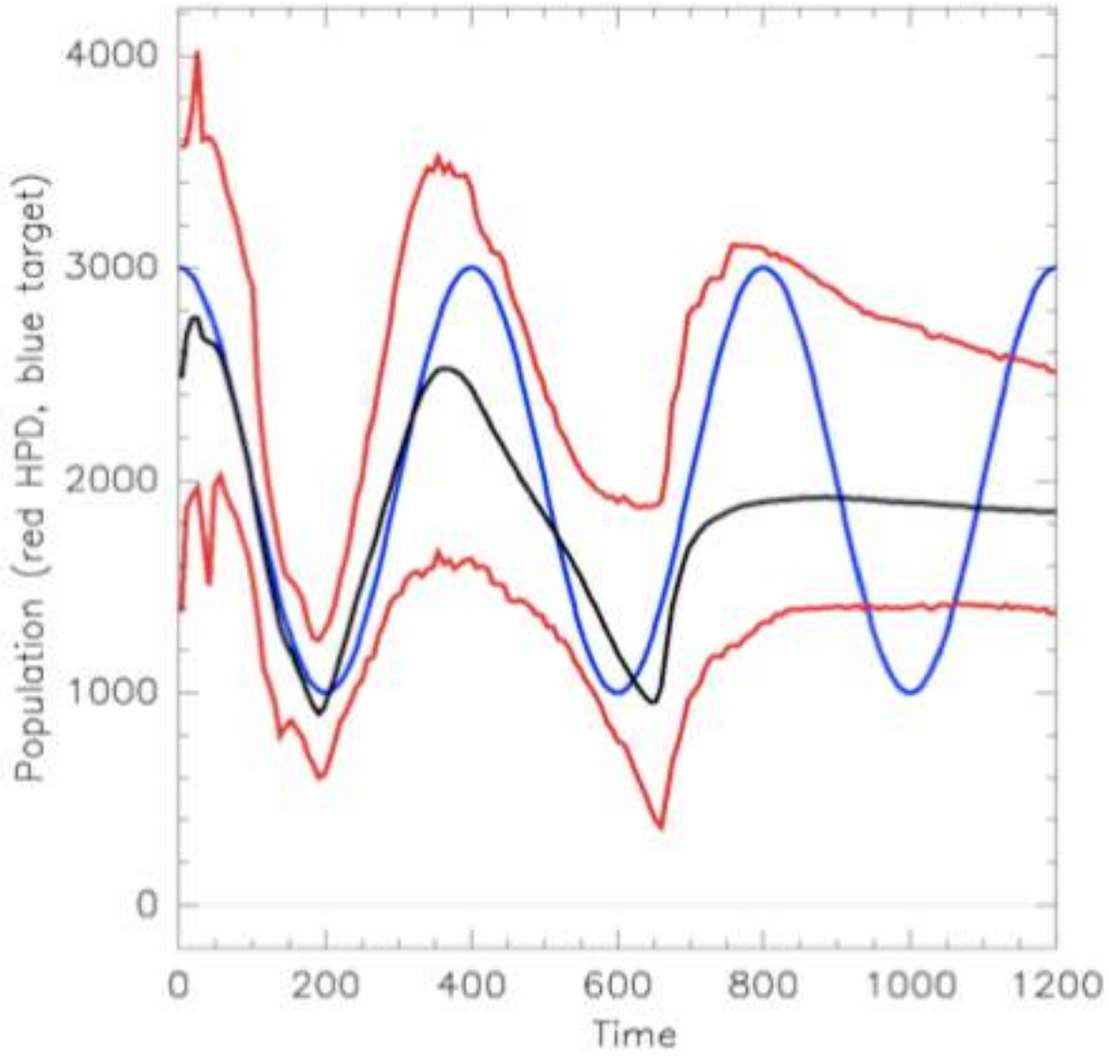
Figure 1 | Population dynamics of genetic diversity in influenza A virus. Bayesian skyline plots of the HA and NA segments for the A/H3N2 and A/H1N1 subtypes in New York state (top) and New Zealand (bottom). The horizontal shaded blocks represent the winter seasons. The *y*-axes represent a measure of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

(Extended) Bayesian skyline plot

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



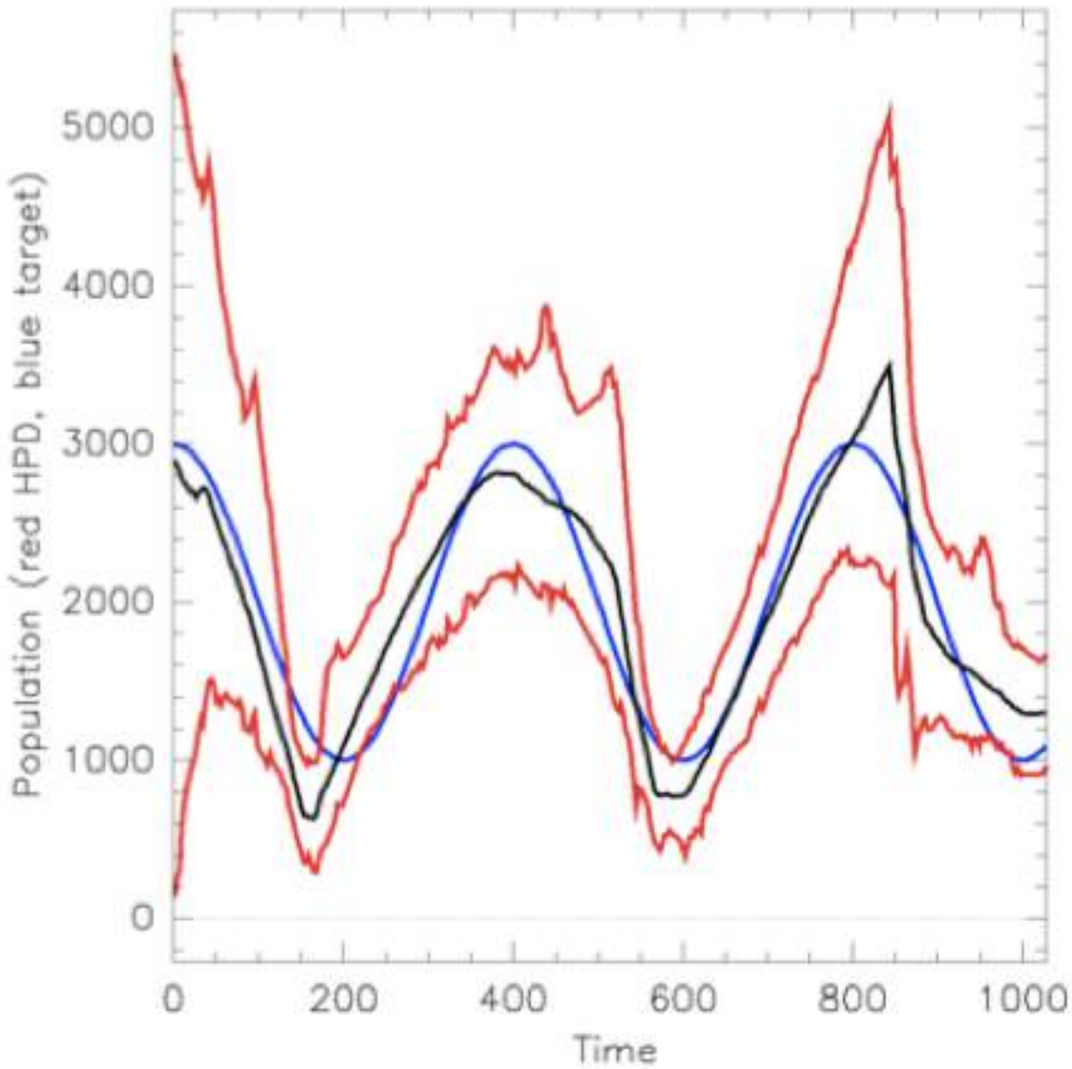
one gene sampled from 480 sampled individuals (480 gene sequences in total)



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

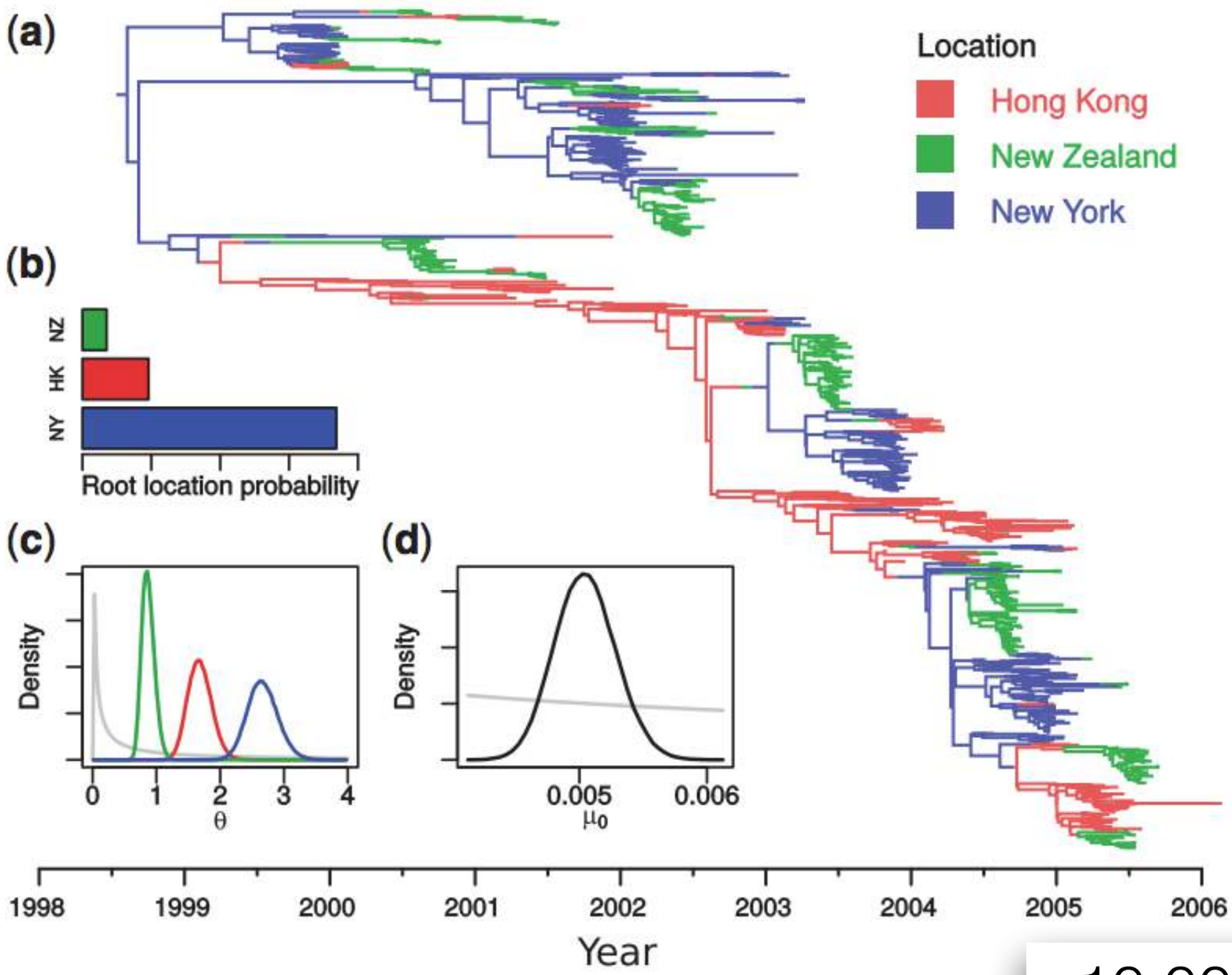
Phylogenetics

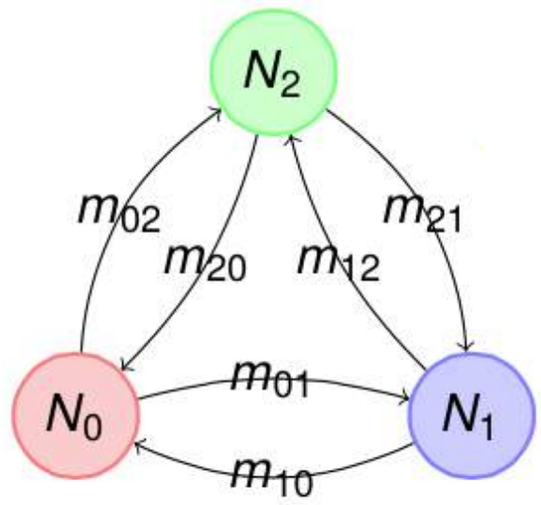
Advance Access publication April 20, 2014

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

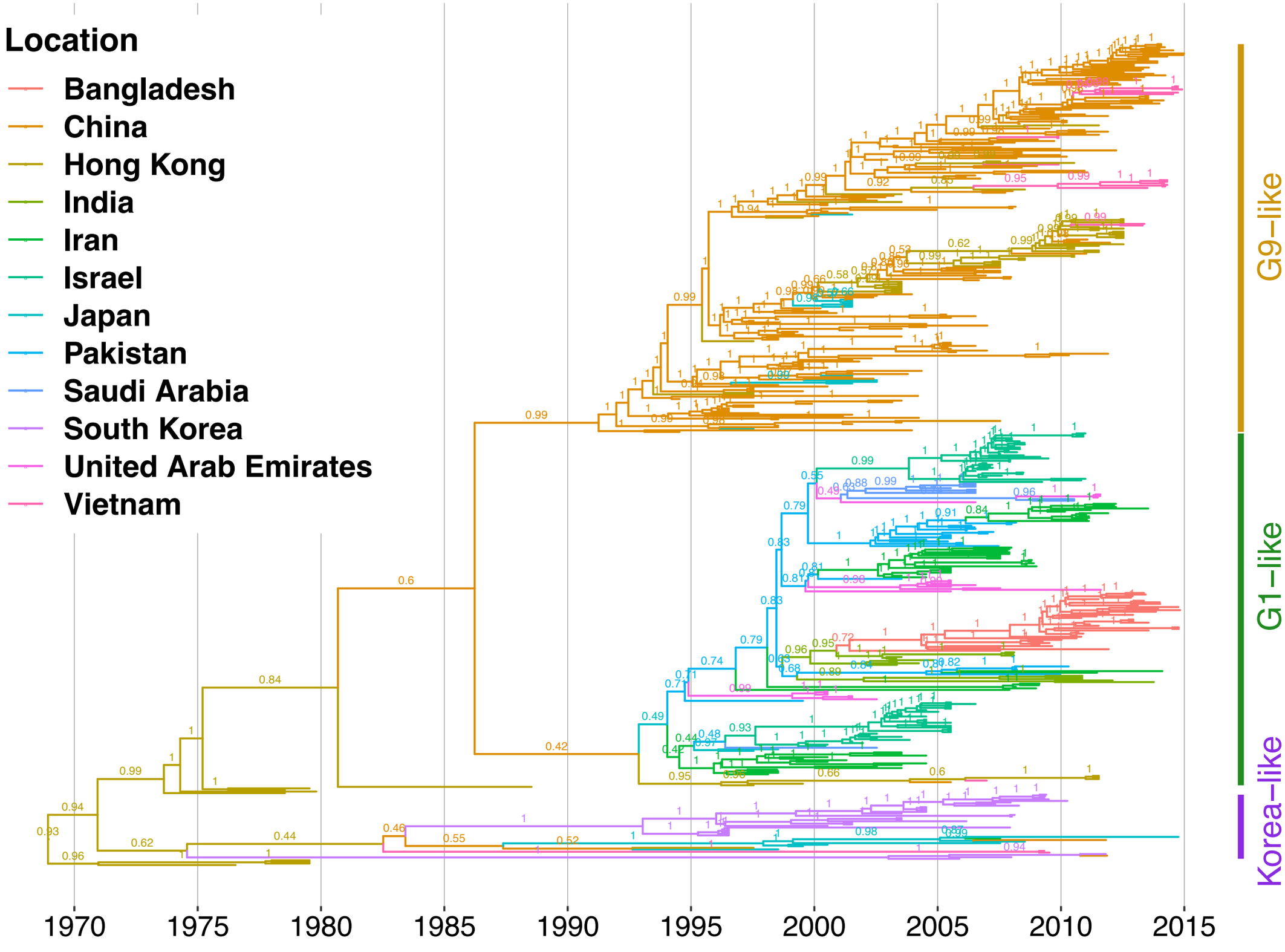




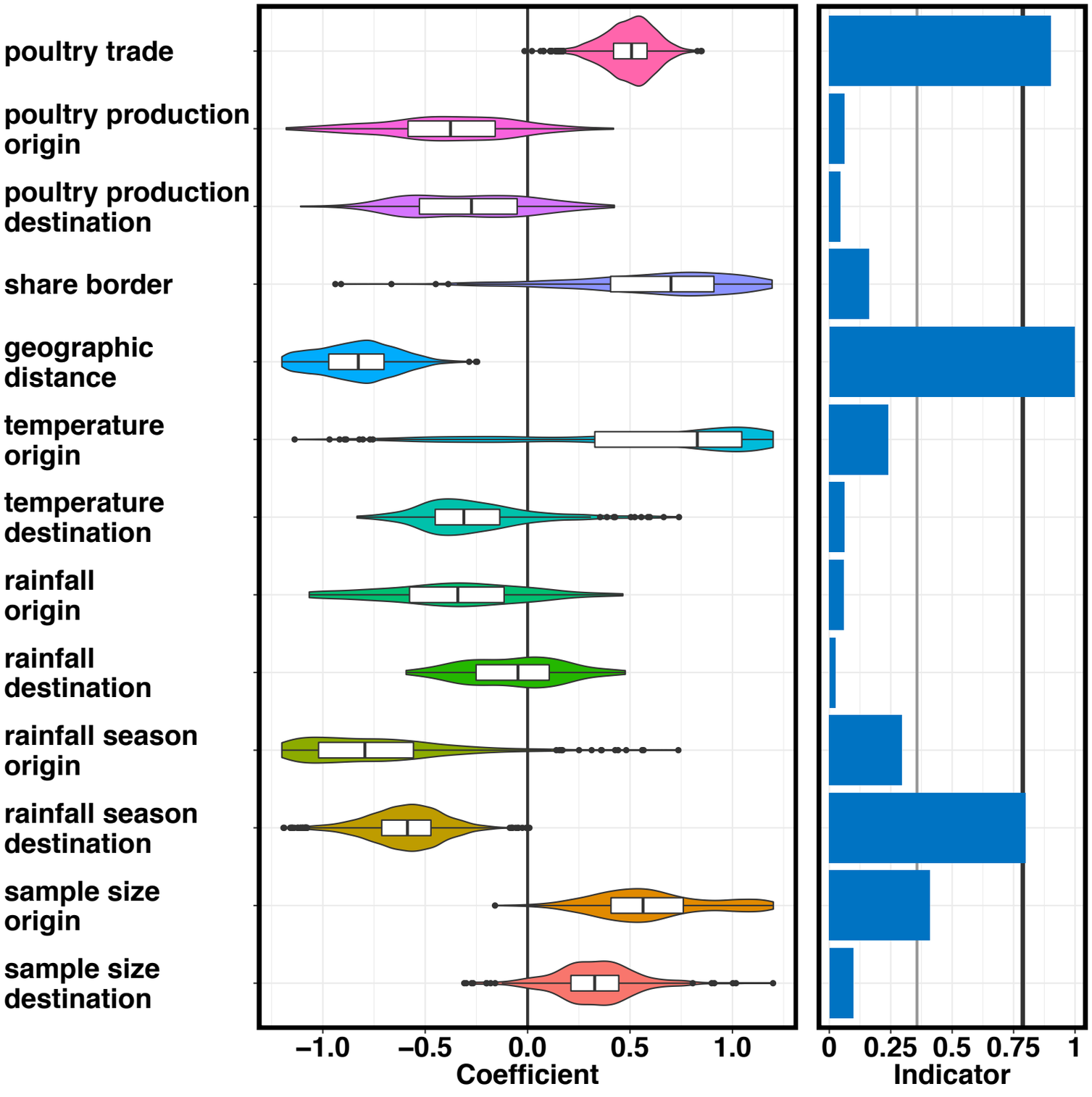
~10,000,000 nucleotides

Avian influenza H9N2 in asia

Yang, Mueller, Bouckaert, Xu, Drummond (in prep)



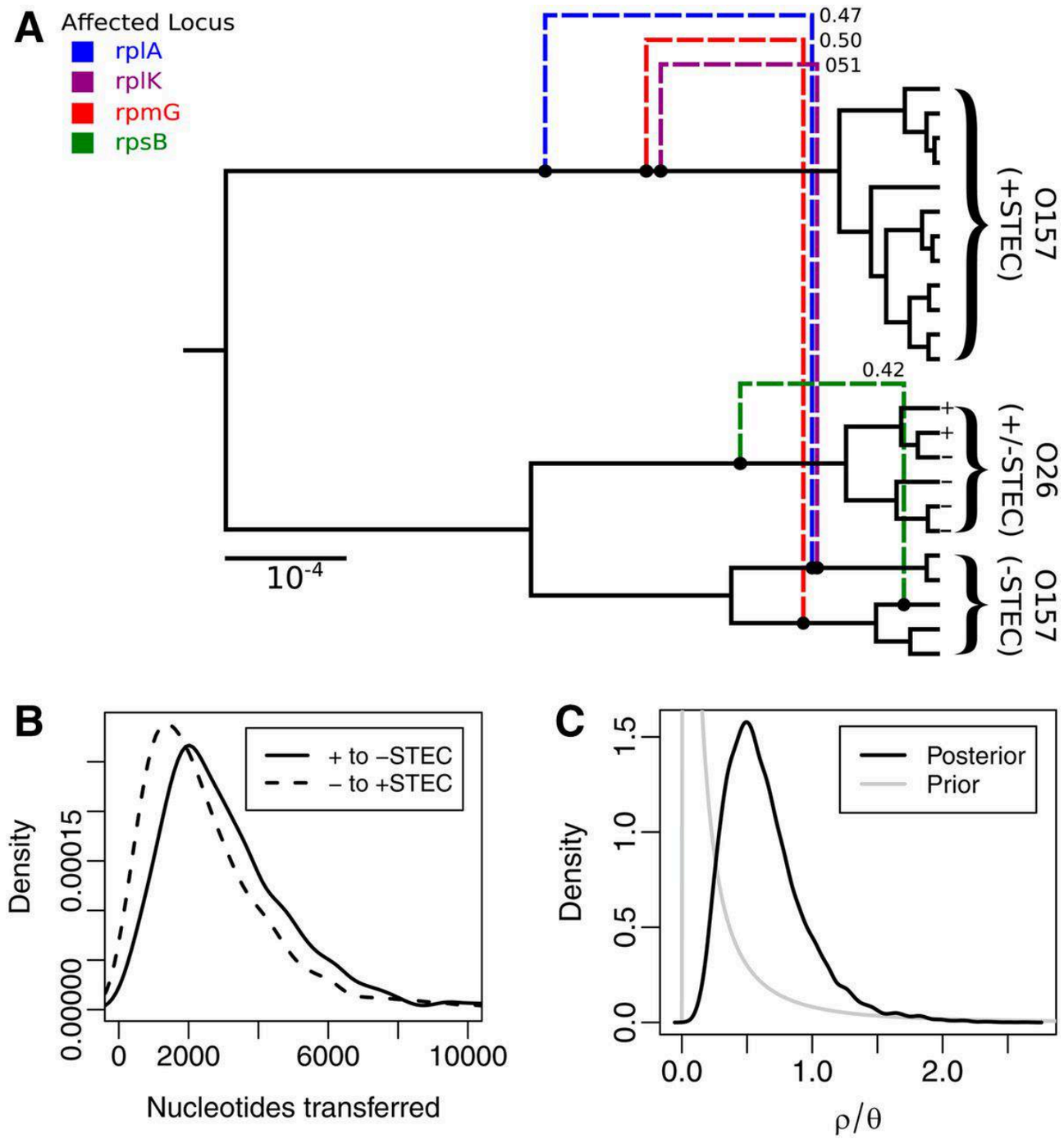
Phylogenetic GLM predictors of virus migration rate based on time-series data



Inferring Ancestral Recombination Graphs from Bacterial Genomic Data

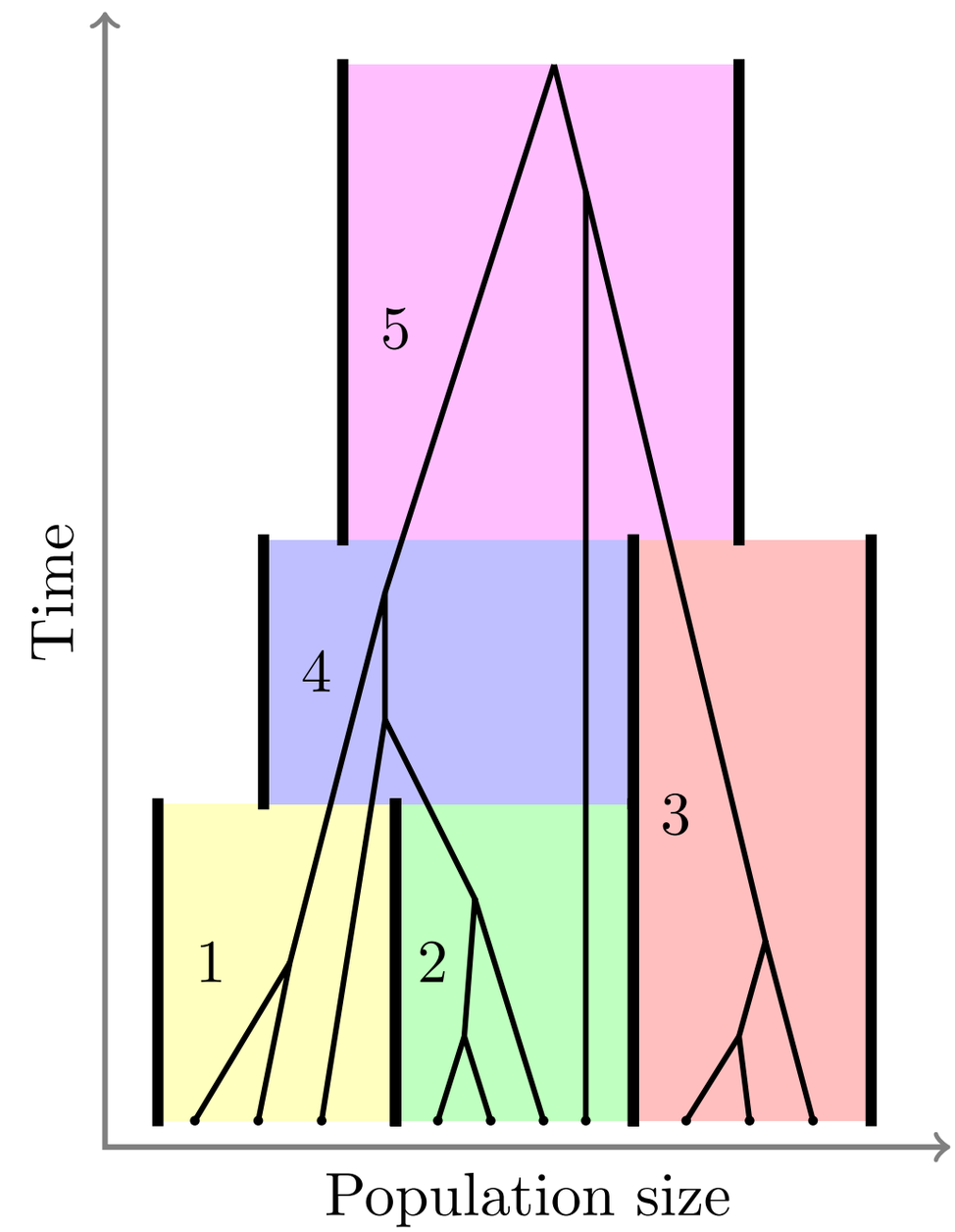
Timothy G. Vaughan, David Welch, Alexei J. Drummond, Patrick J. Biggs, Tessy George and Nigel P. French

GENETICS *February 1, 2017 vol. 205 no. 2 857-870; https://doi.org/10.1534/genetics.116.193425*



Trees inside trees

- Modelling **diversity and divergence**
- Central to understanding diverse areas including **speciation genomics** and **transmission** is building up a statistically rigorous model of the patterns of divergence and diversity spanning multiple species/hosts, multiple genes / gene families and multiple individuals per species/host.



$$p(g, S \mid D) \propto \Pr(D \mid g) P(g \mid S) P(S)$$

$$p(g, S \mid D) \propto \prod_i \Pr(D_i \mid g_i) P(g_i \mid S) P(S)$$

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.

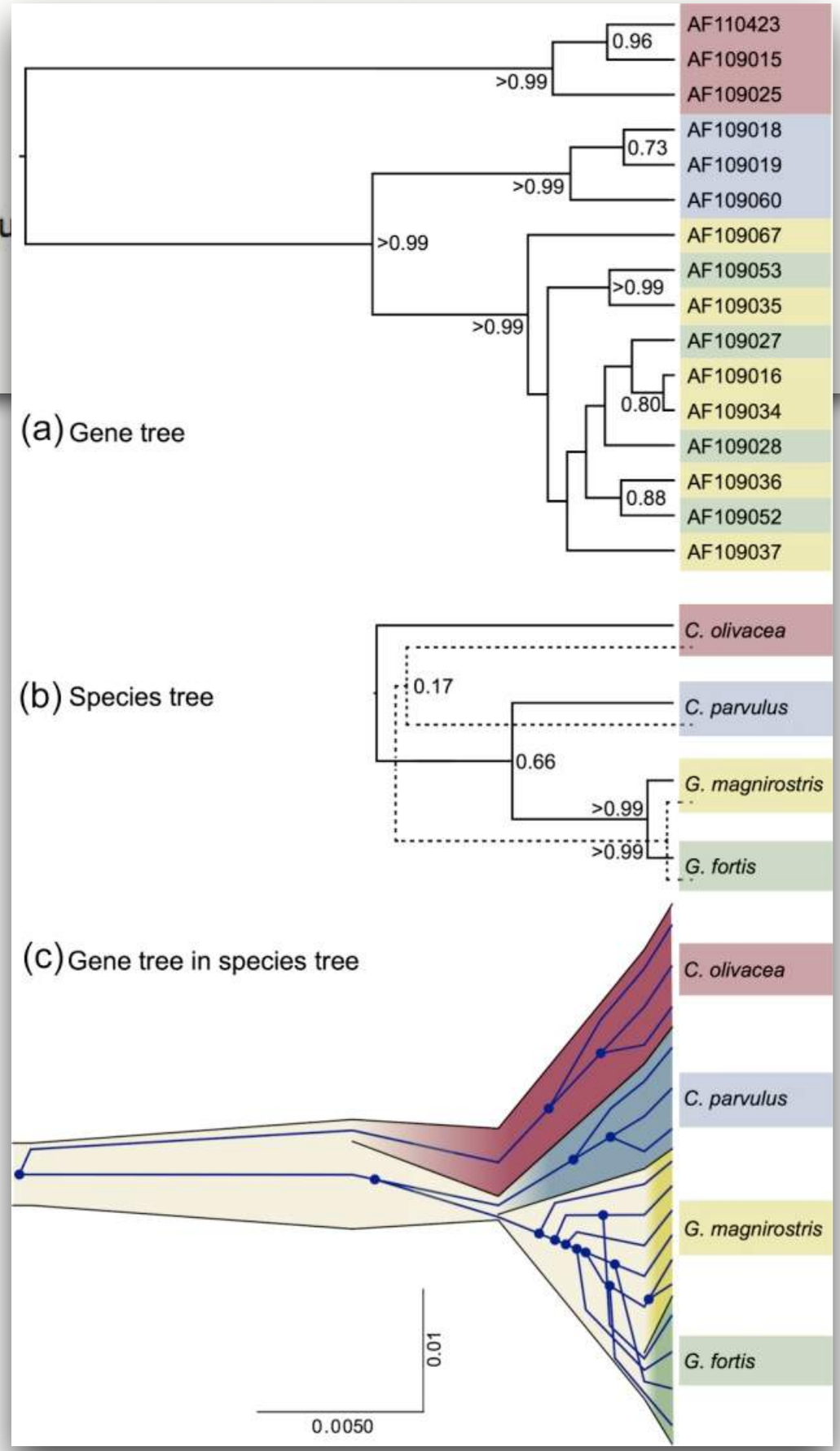
Associate editor: Dr. Jeffrey Throne

- **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.

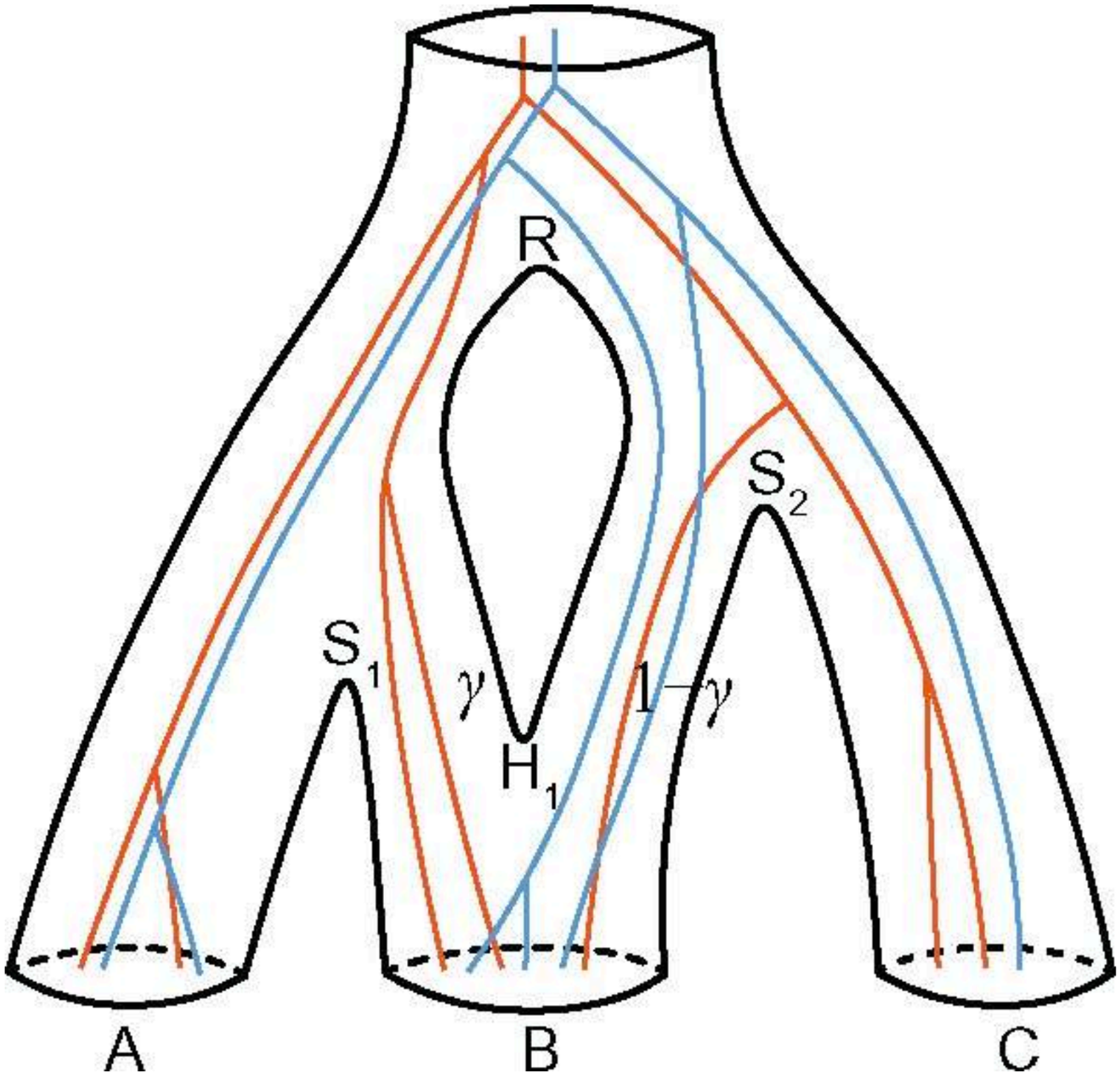
- Forms a theoretical basis for models that estimate species boundaries based on genomic data.

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

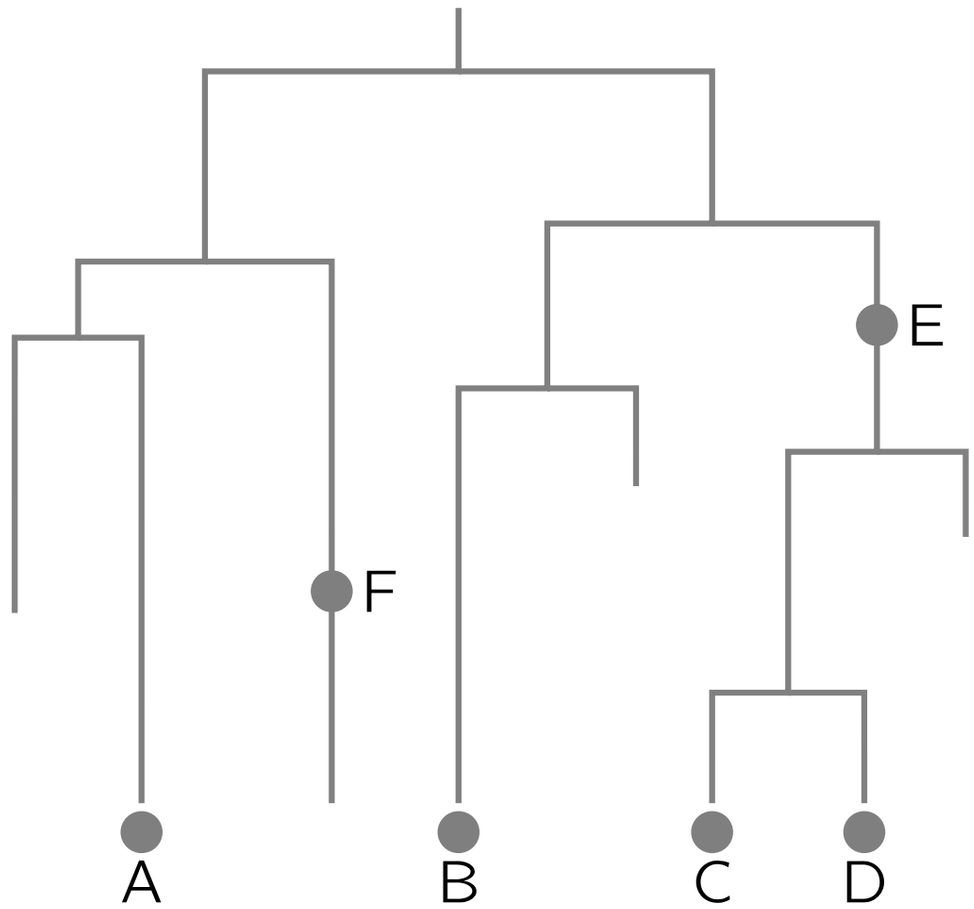


Species networks with embedded gene trees

Zhang, Ogilvie, Drummond, Stadler (2018)



Fossilized birth-death (FBD) model



Full tree

Stadler 2010, Gavryushkina *et al* 2014

The process starts at time $t_{or} > 0$ and ends at time zero (present time).

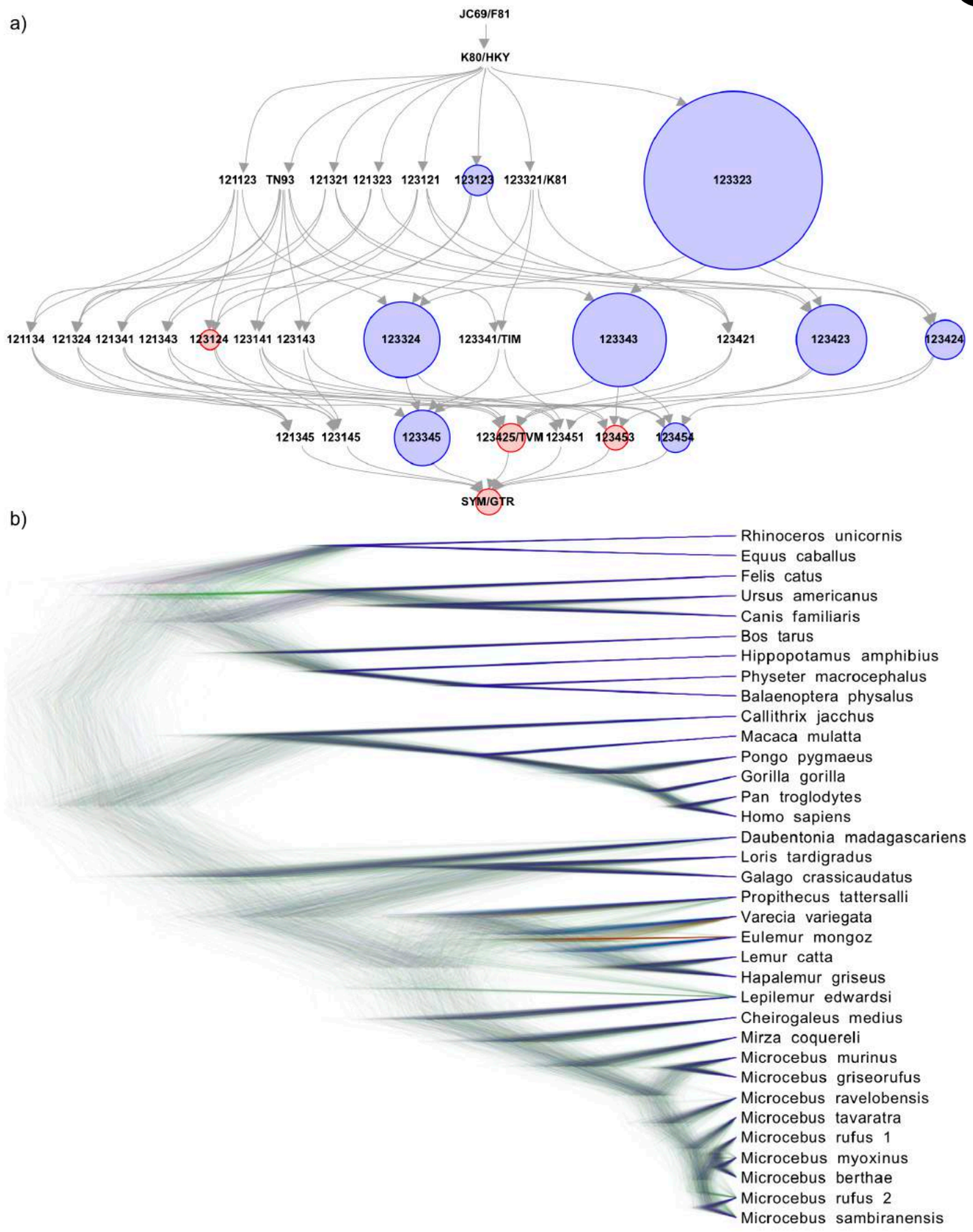
- birth rate λ
- death rate μ
- sampling rate ψ
- sampling at present probability ρ

Model parameters: $\eta = (t_{or}, \lambda, \mu, \psi, \rho)$.

All the parameters are identifiable.

Substitution and site models

“bModelTest” model averaging



Bouckaert, Drummond (2017)

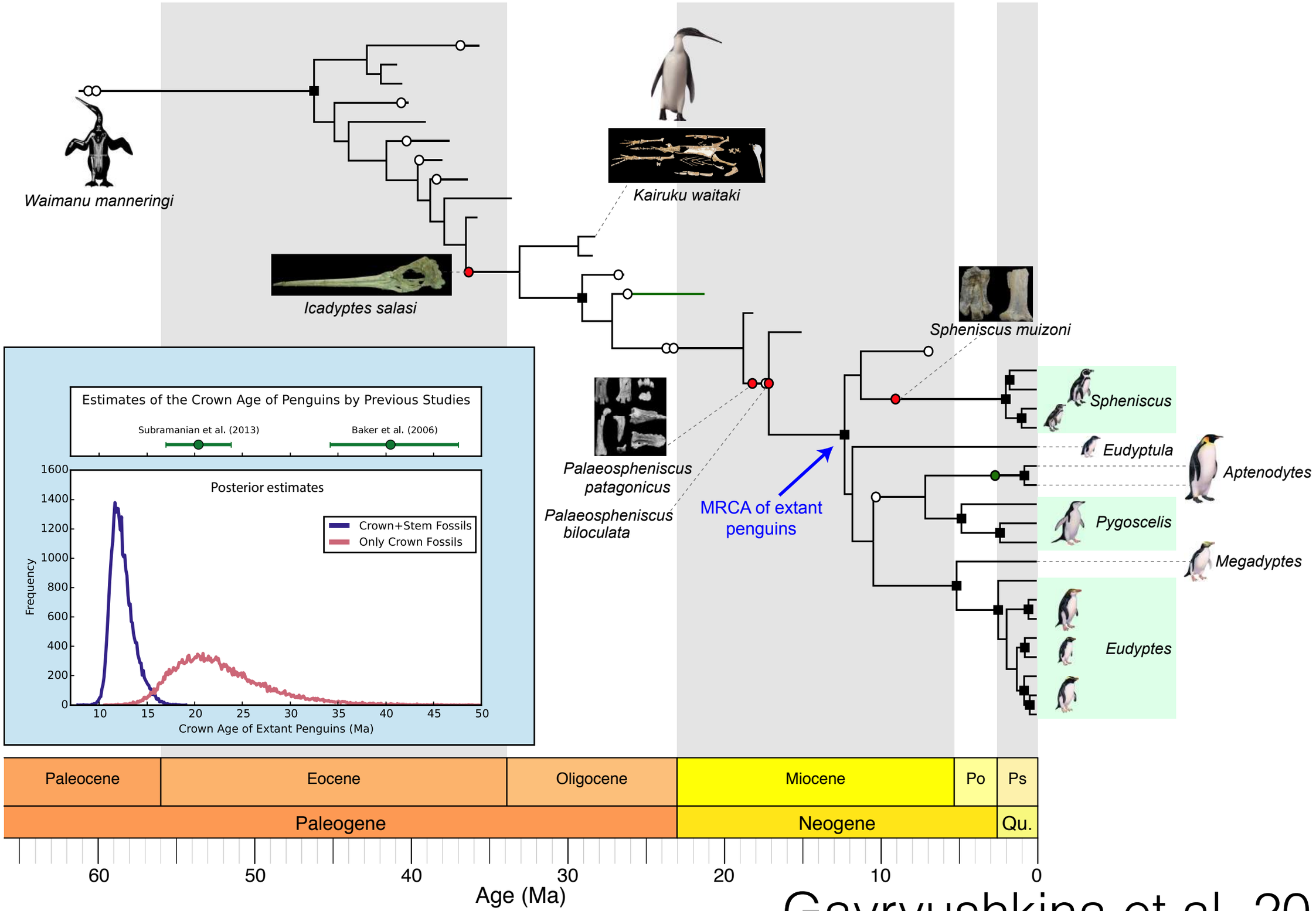
Integrative phylogenomics

- 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)
- **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

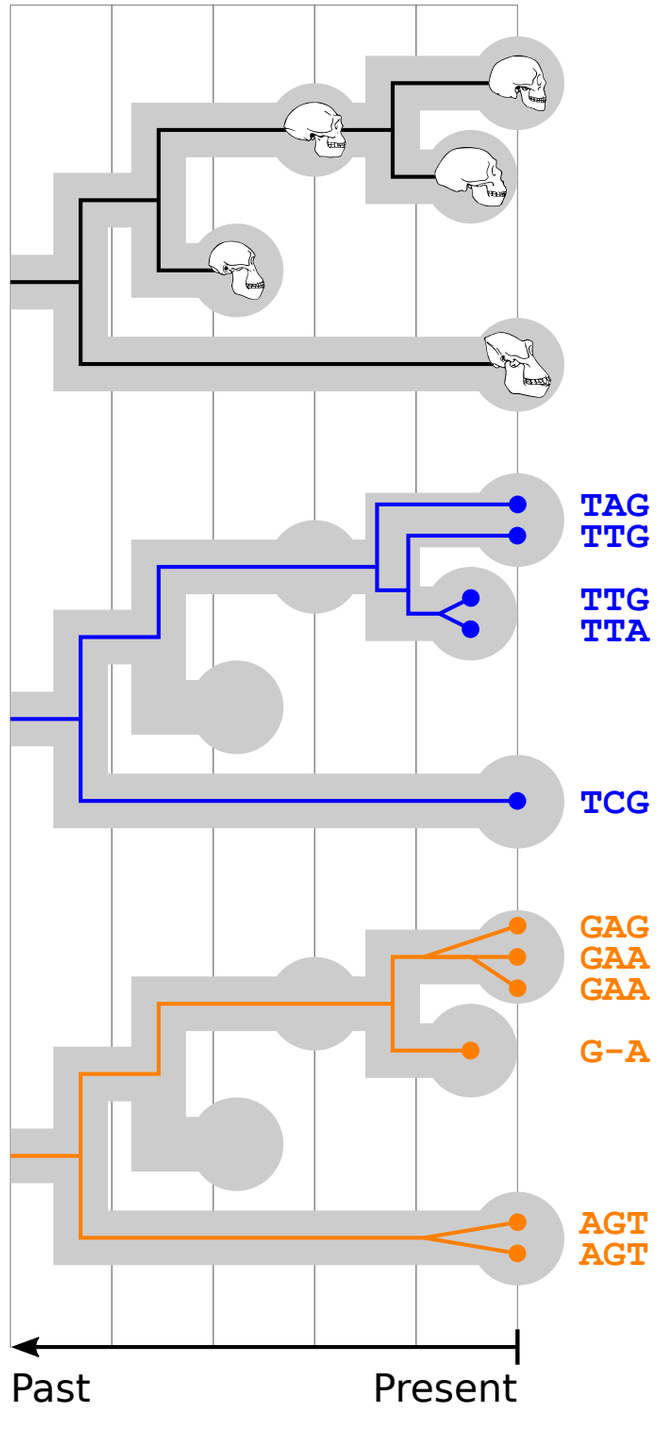
First step: Total-evidence of **DNA and fossils**



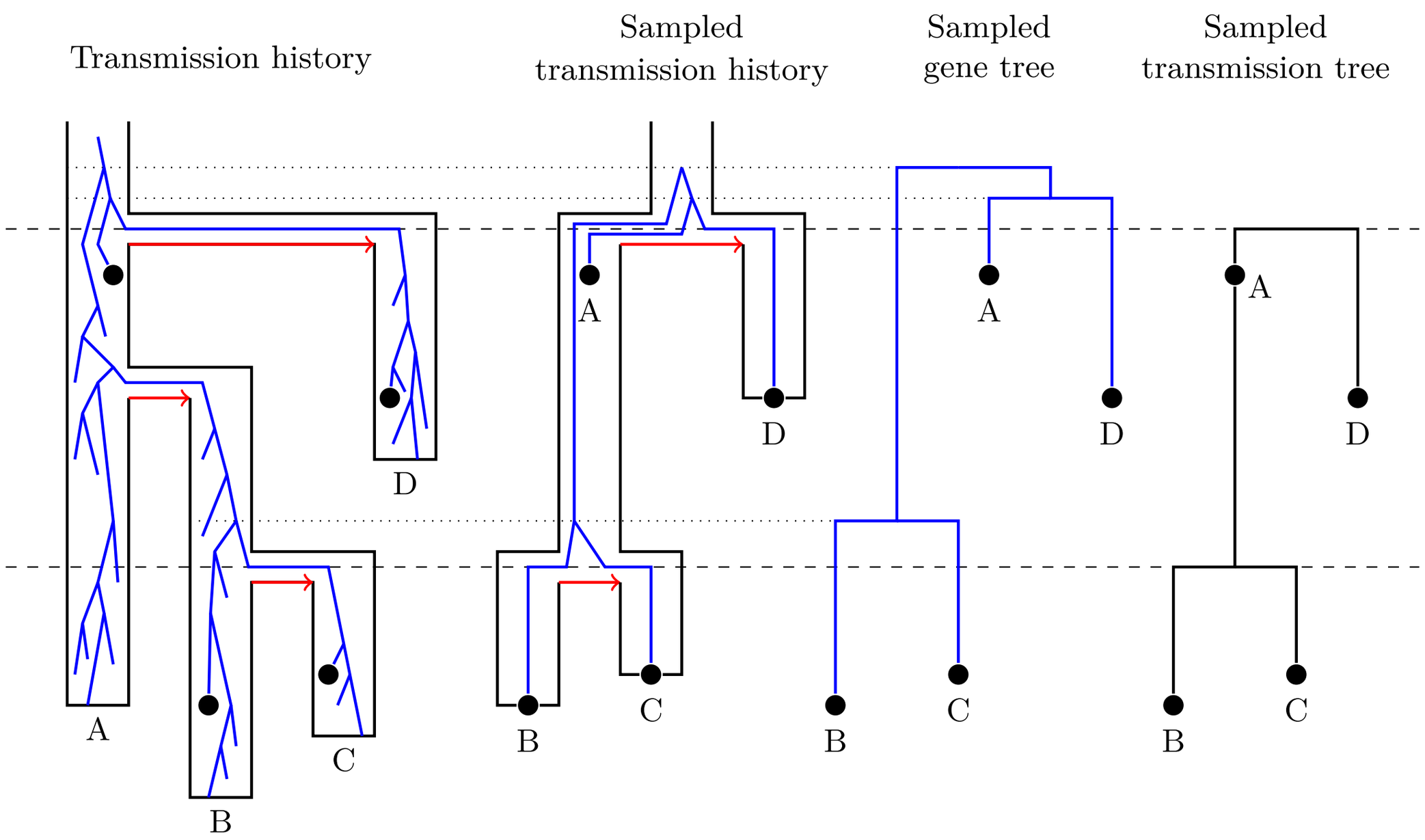
Gavryushkina et al, 2016

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

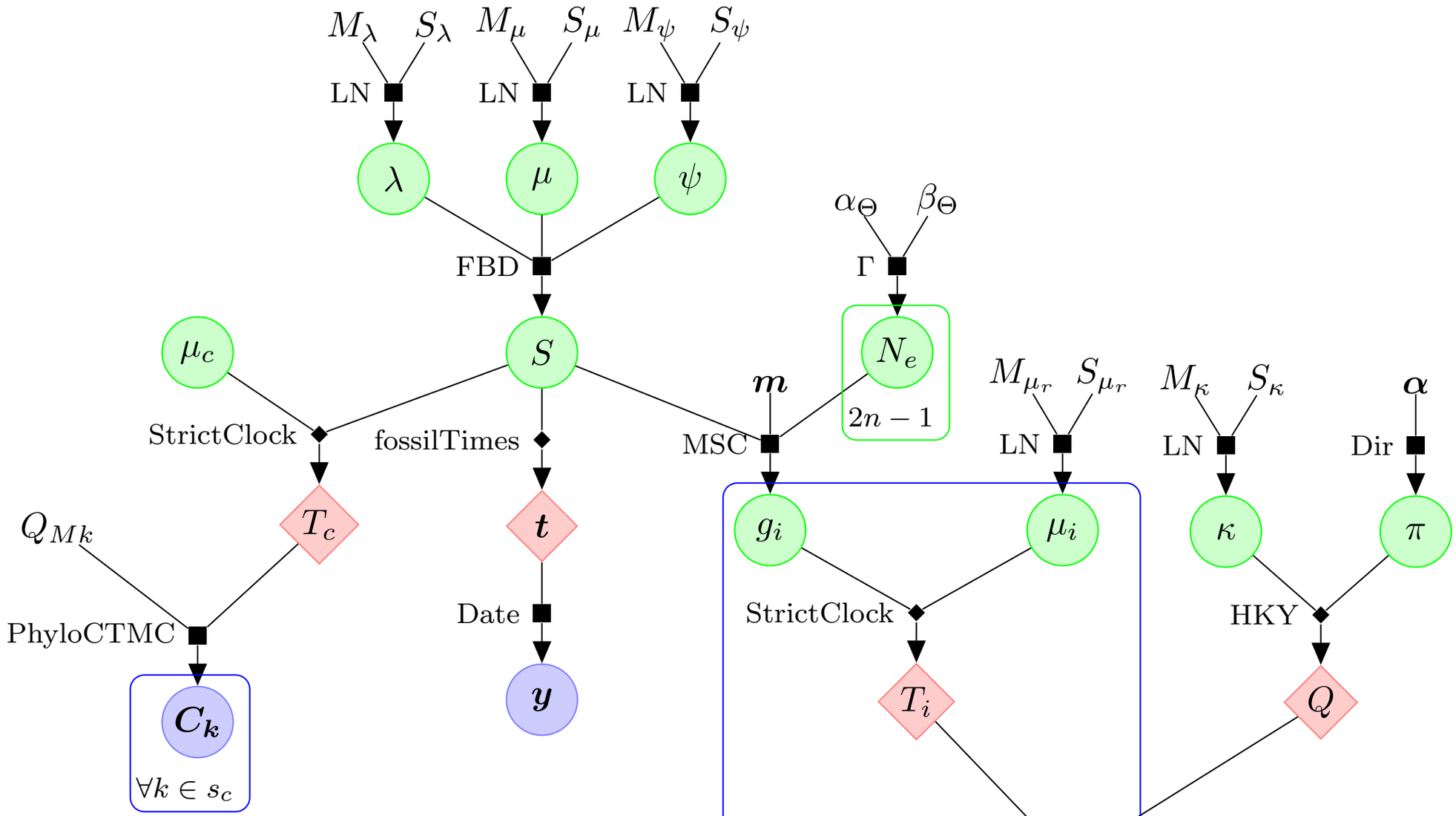


Another integrative model: pathogen evolution inside transmission trees



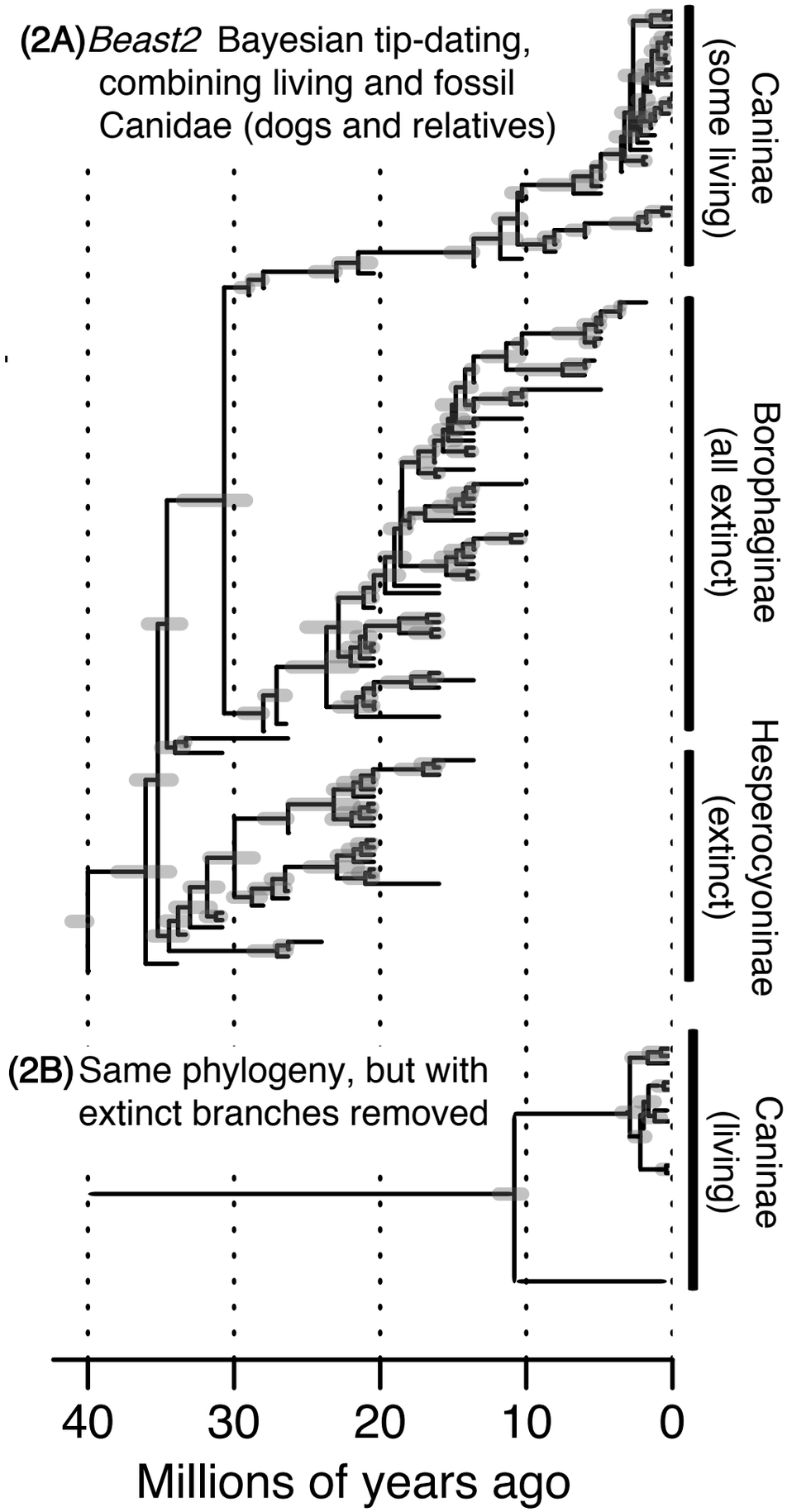
Goal: An integrated model of stochastic infection dynamics and viral genomic evolution. Integrating infection dynamics, genomic sequence evolution, (and phenotypic data?) into a single model.

Bayesian graphical model



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

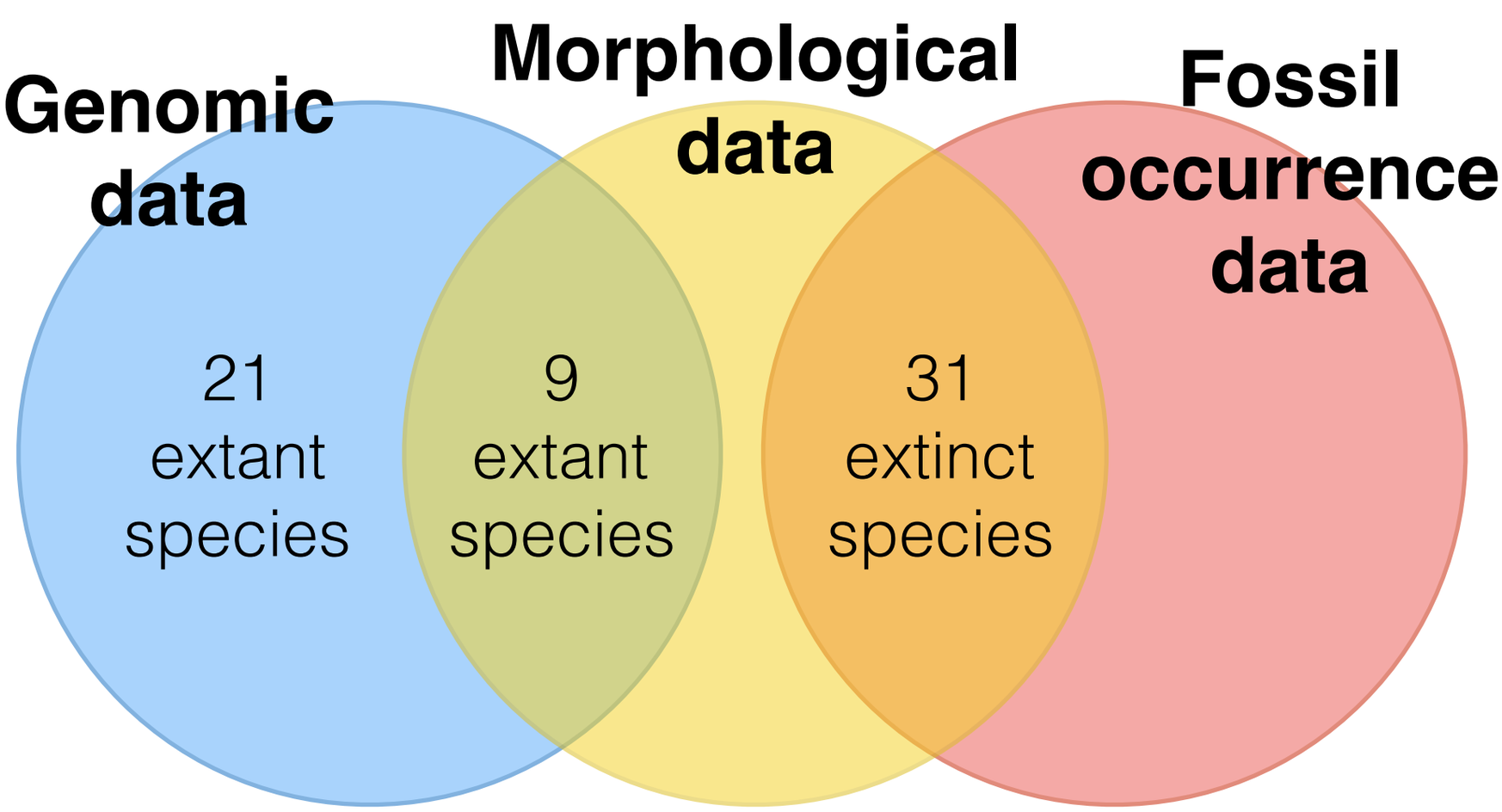
Caninae example



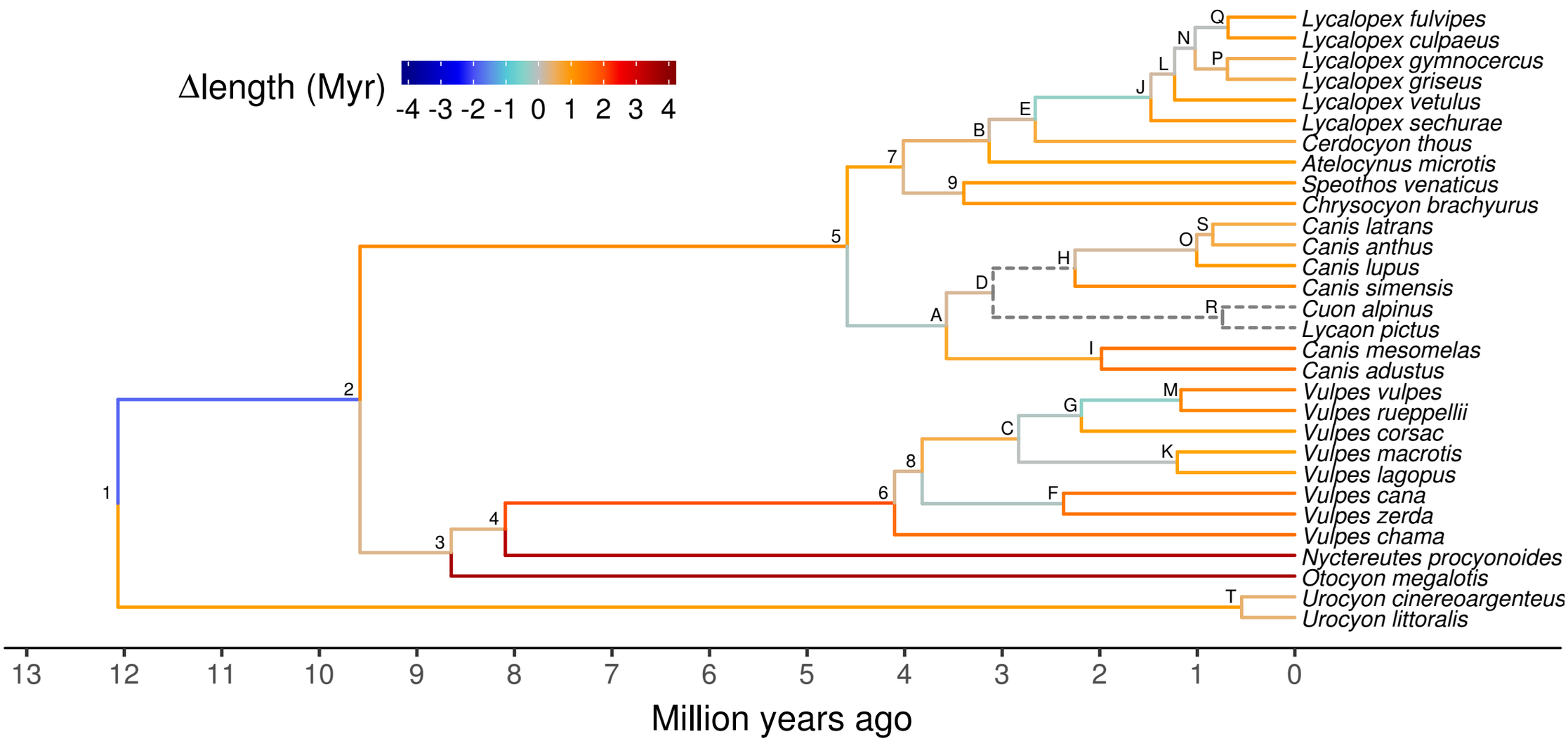
Caninae example

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

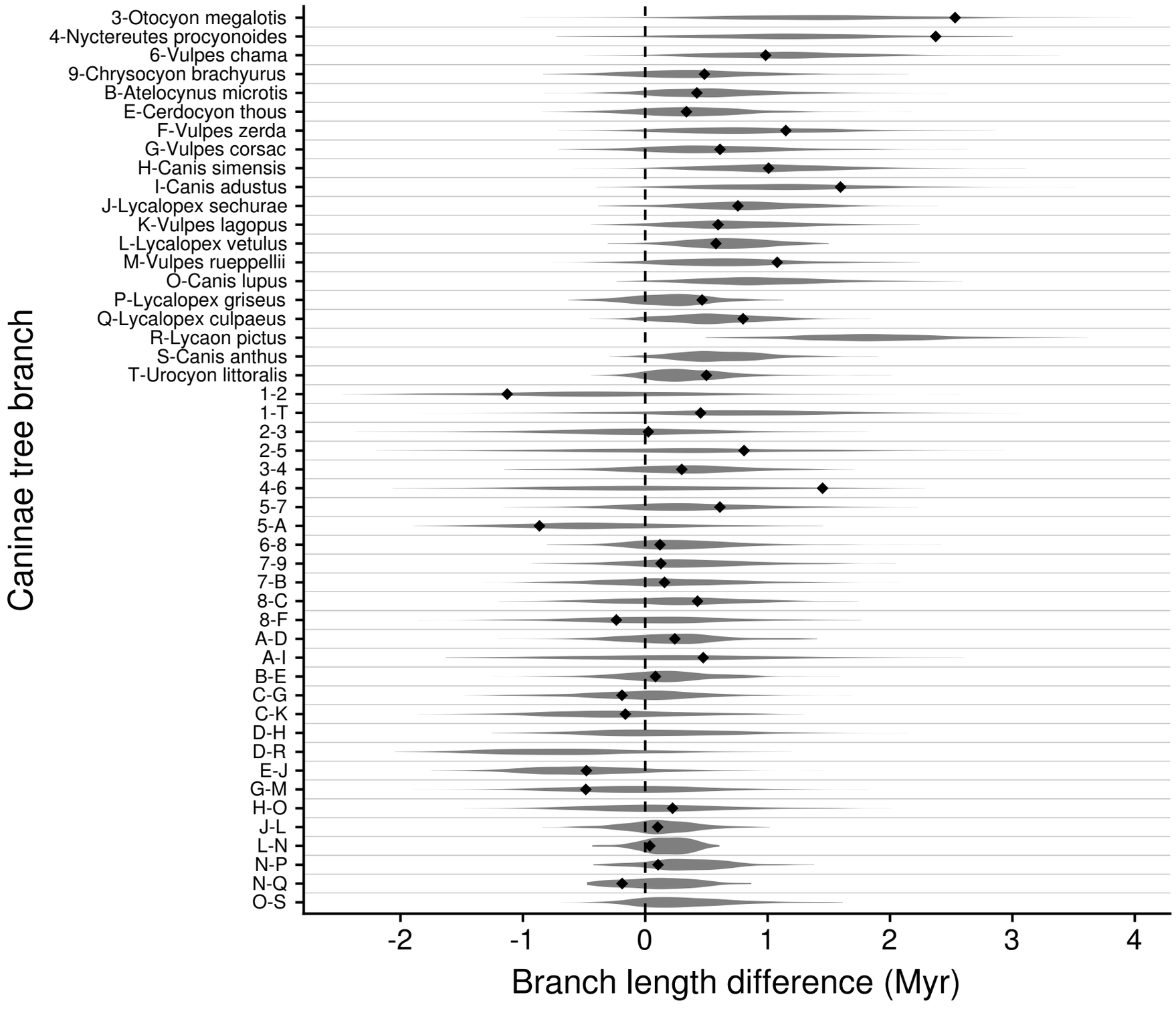
- 72 morphological characters from 40 species (2880 morphological character calls)



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



Divergence time estimation

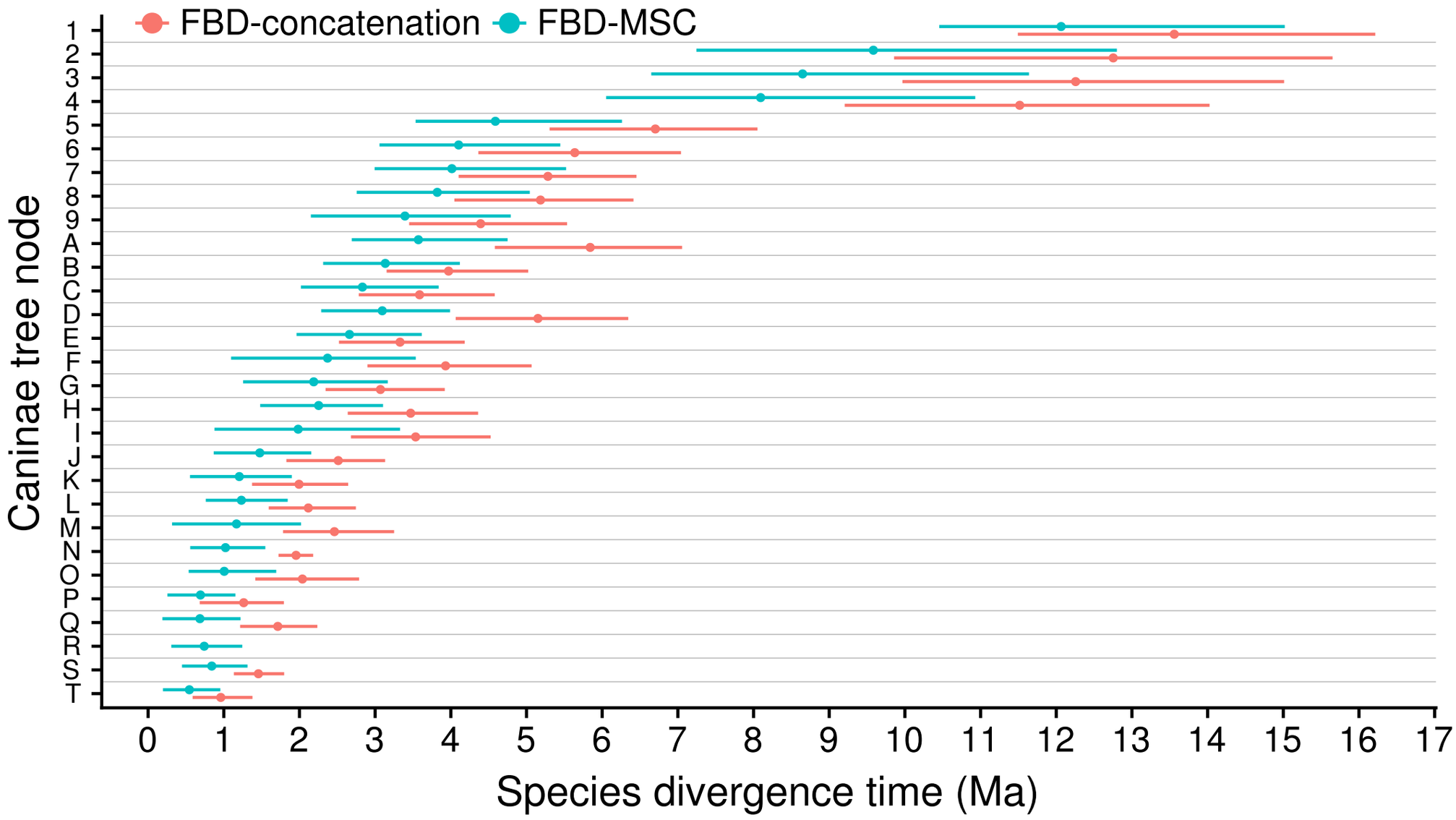


Figure 7: Speciation times estimated by fossilized birth-death with multispecies coalescent (FBD-MSC) and with concatenation (FBD-concatenation) models. Posterior mean FBD-MSC node ages (solid circles) and 95% highest posterior density (HPD) intervals (lines) are estimated from samples where that clade 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

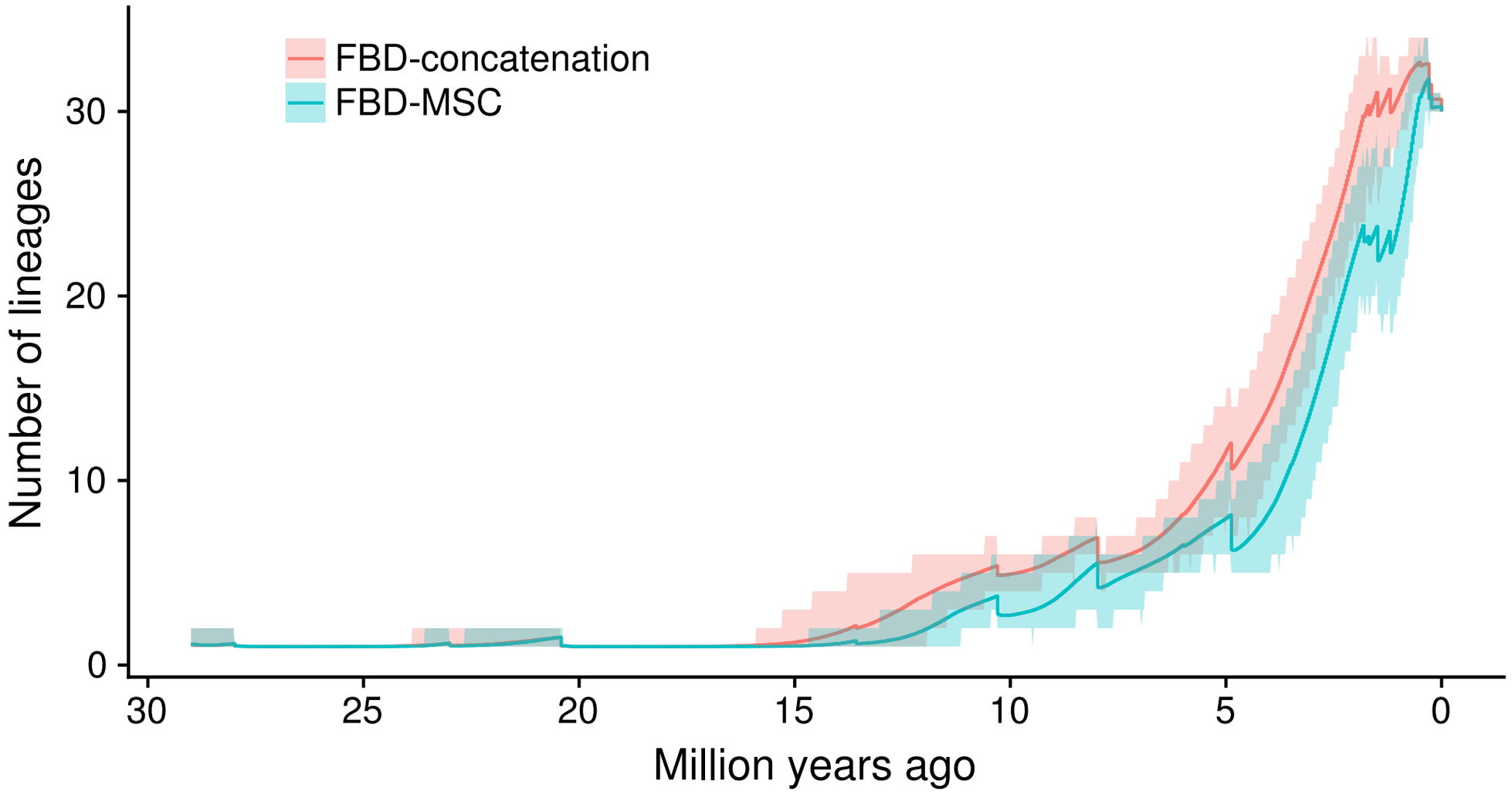


Figure 8: Lineages-through-time (LTT) plot of Caninae diversification. Posterior mean estimates (solid lines) of LTT are calculated for 1,001 evenly spaced time steps spanning 0 to 1, and include extant, fossil and ancestral taxa, and sampled ancestors (which are both fossil and ancestral). 95% highest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

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.

- **Computational sciences rely on continuous development and maintenance of large software packages**, and this is currently still a challenge for science funding models.