cEvo

virion

Understanding
virus-host interactions

Infectious disease phylodynamics

Louis du Plessis

ETH zürich

D BSSE

On the program for today

(1) **What is phylodynamics?**

(2) Bayesian inference recap

(3) BEAST2 introduction

Tutorial: Molecular clock dating (part i)

(4) Molecular clock models

Tutorial: Molecular clock dating (part ii)

(5) Setting priors

Tutorial: Phylodynamics (part i)

(6) Tree priors

Tutorial: Phylodynamics (part ii)

cEvo

virion

01110011 01110110 01101001 01110010 01110101

Understanding
virus-host interactions

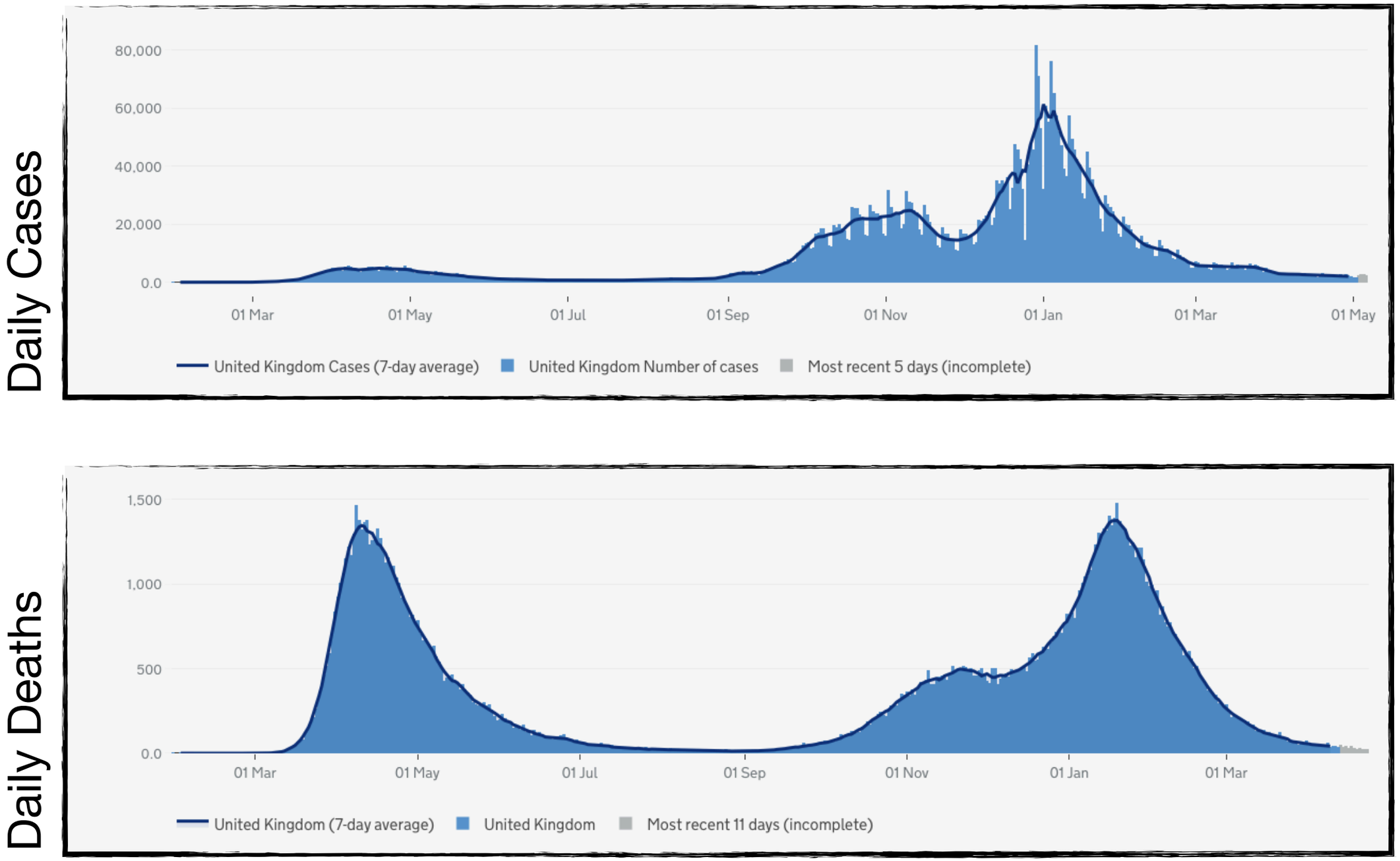
What is phylodynamics?

Louis du Plessis

ETH zürich

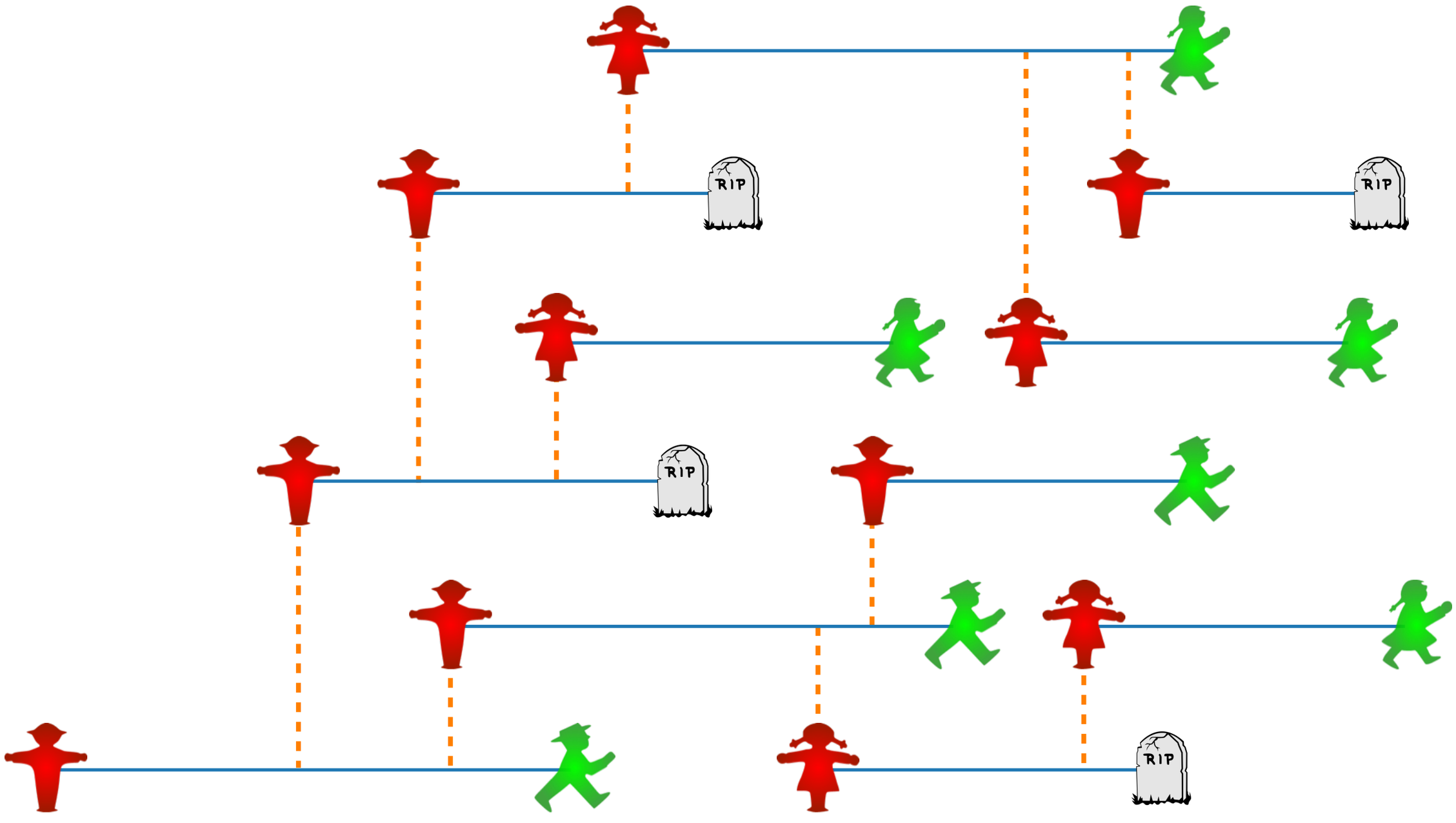
D BSSE

Epidemic time-series (United Kingdom)

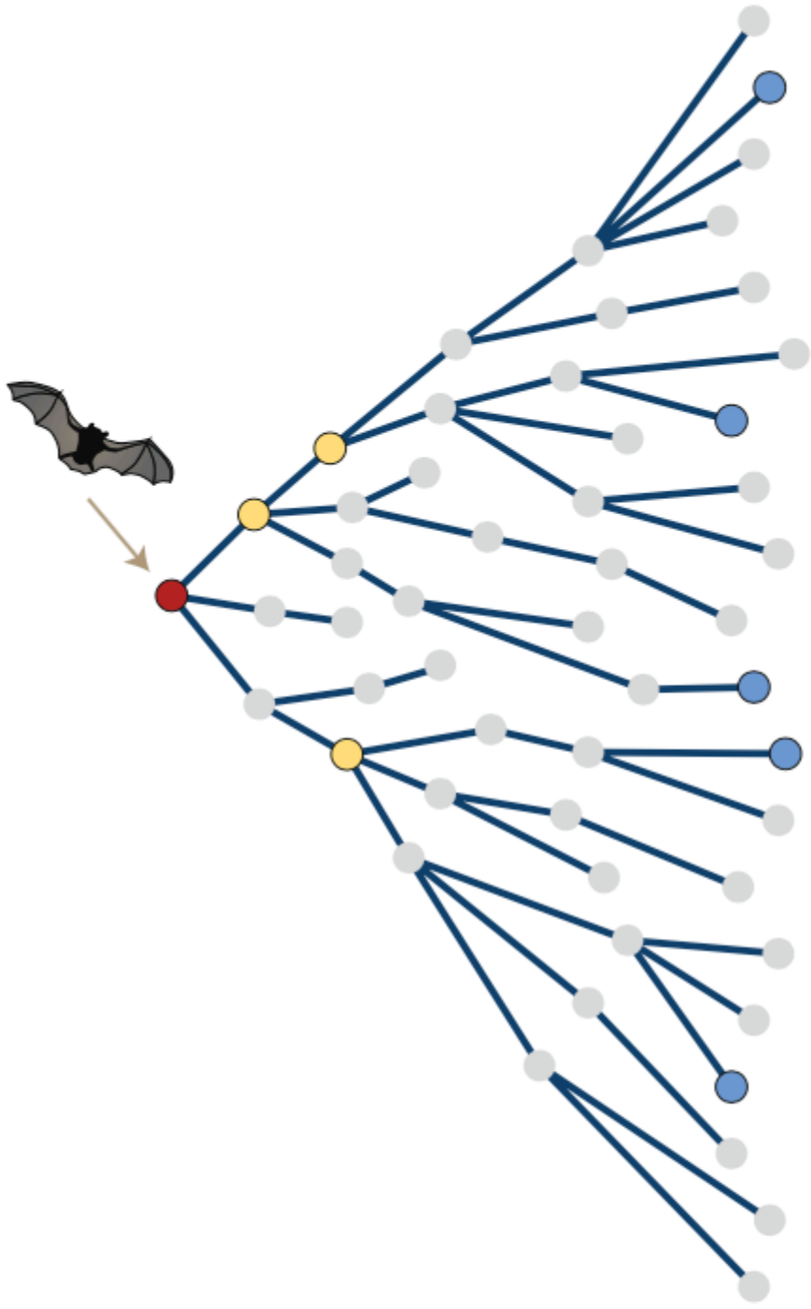


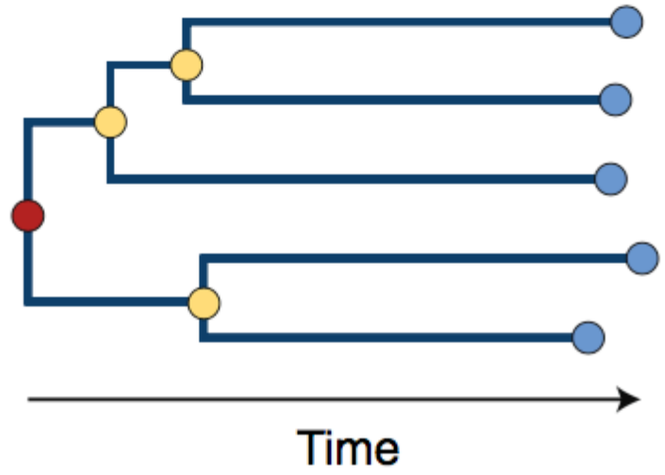
https://coronavirus.data.gov.uk

Definition: Transmission tree



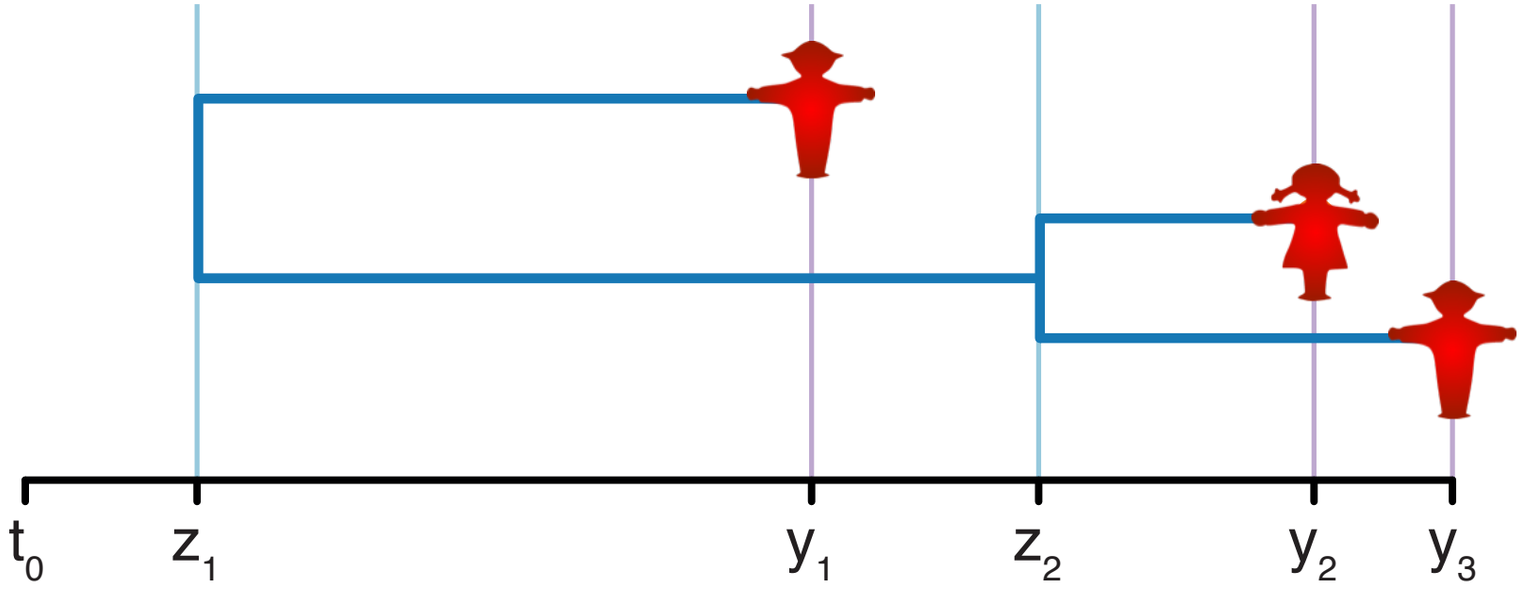
Definition: Partial transmission tree (phylogeny)





Grubaugh et al. (2019) *Nature Microbiology* 4:10-19

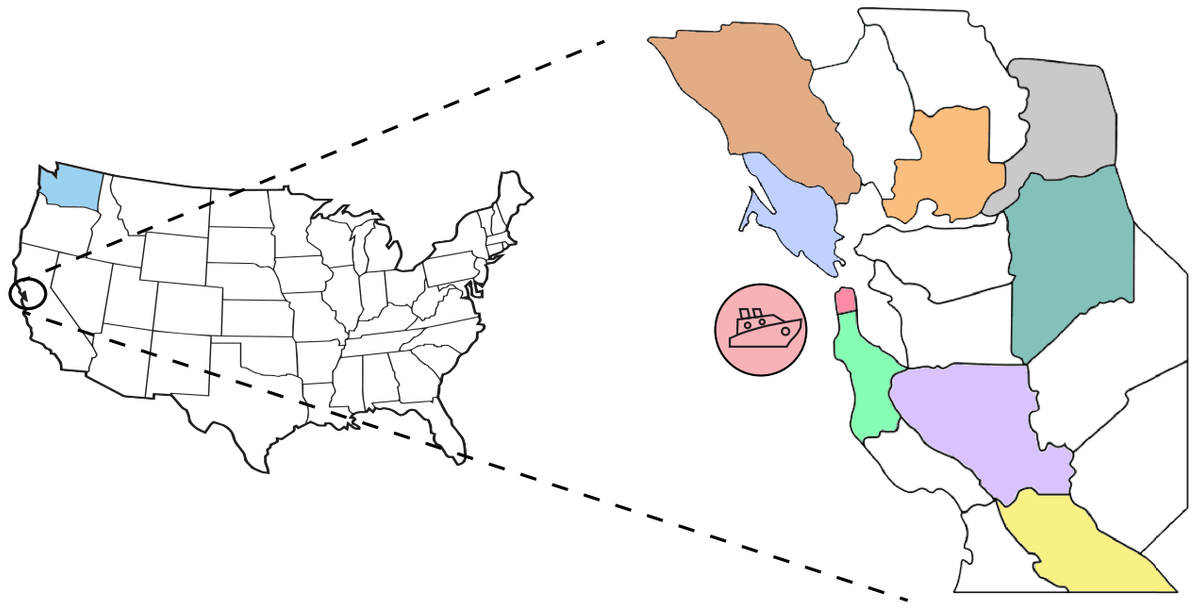

JUL
11
ACACACCC…
JUL
17
TCACACCT…
JUL
18
ACAGACTT…


t_0
z_1
y_1
z_2
y_2
y_3


GRAND PRINCESS
GRAND PRINCESS

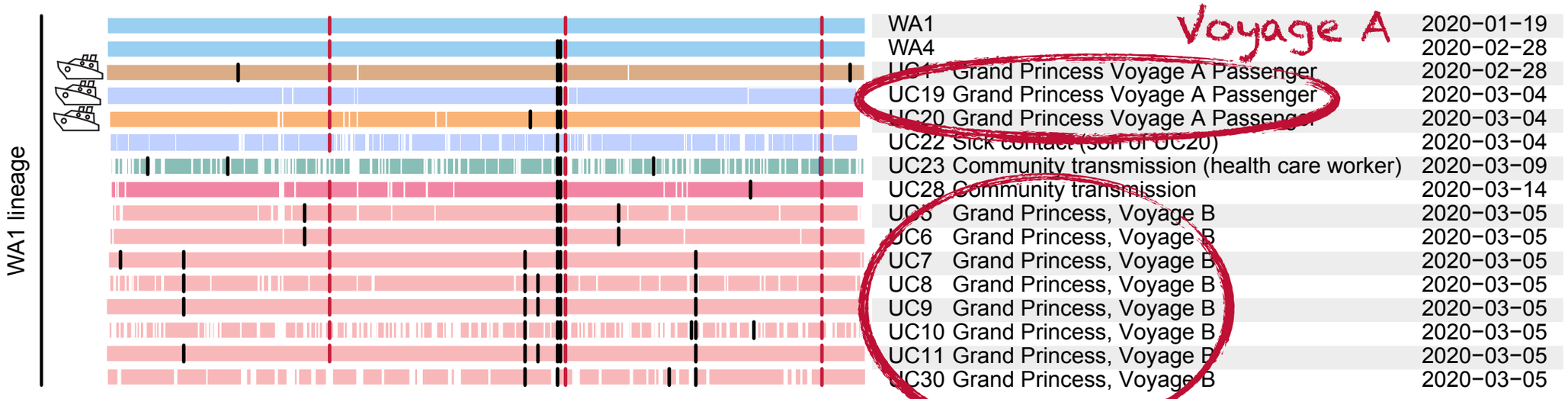
SARS-CoV-2 in Northern California

(Feb - Mar 2020)



Geographic Sampling Location

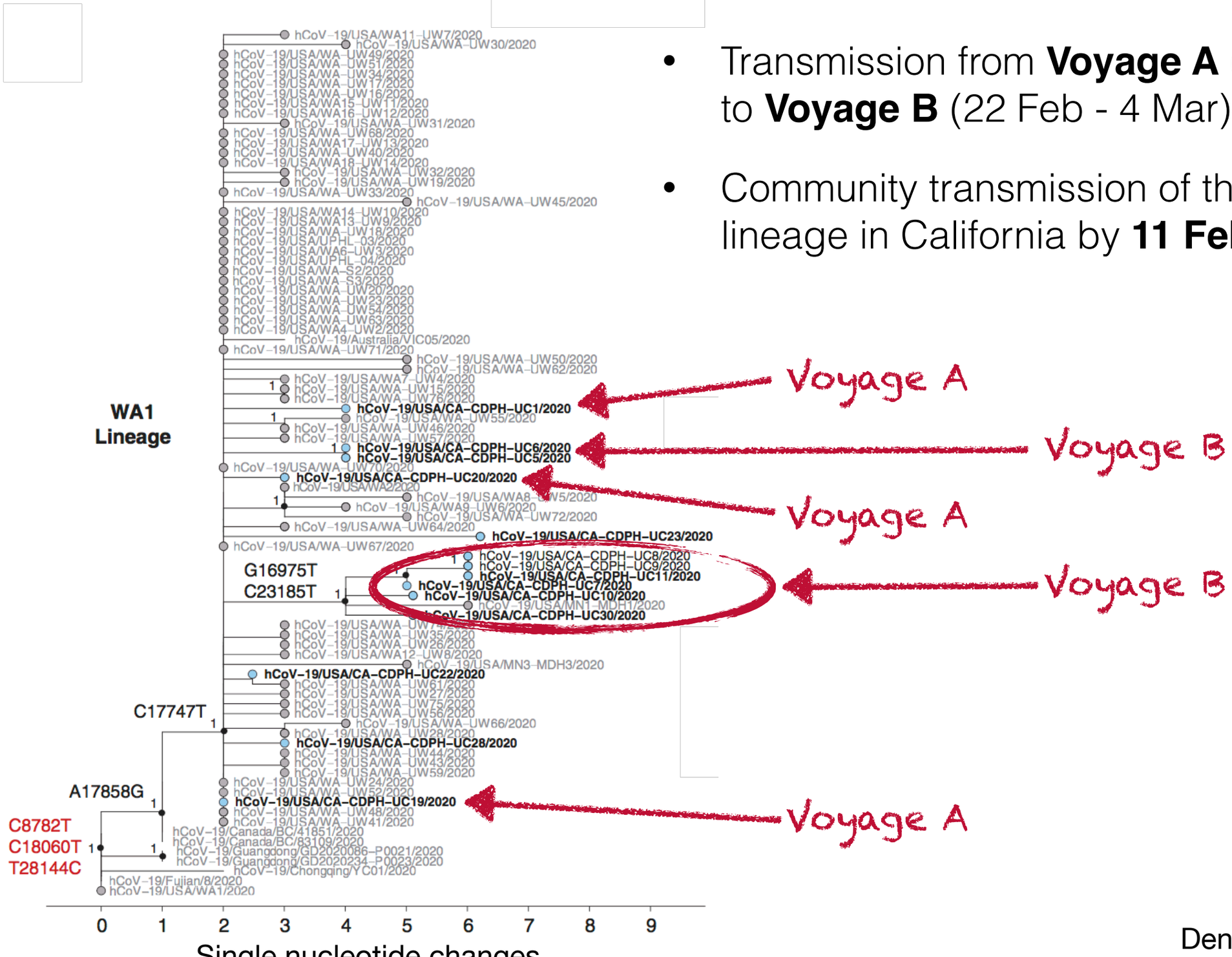
- Washington State
- Grand Princess Cruise Ship (Voyage B)
- Sacramento County
- San Benito County
- San Francisco County
- San Joaquin County
- San Mateo County
- Santa Clara County
- Solano County
- Sonoma County
- Marin County
- Grand Princess Voyage A Passenger



Deng et al. (2020) *Science*

SARS-CoV-2 in Northern California

(Feb - Mar 2020)

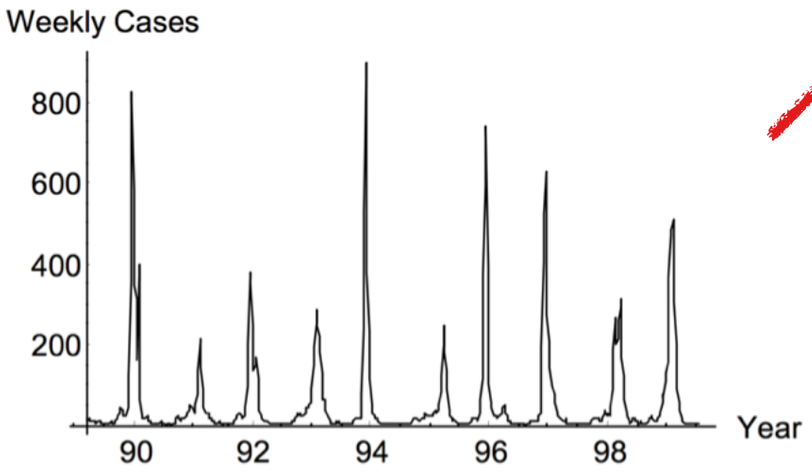


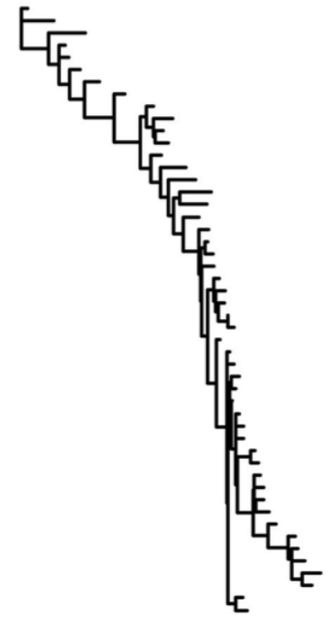
- Transmission from **Voyage A** (11 - 21 Feb) to **Voyage B** (22 Feb - 4 Mar)
- Community transmission of the WA1 lineage in California by **11 Feb**

Deng et al. (2020) *Science*

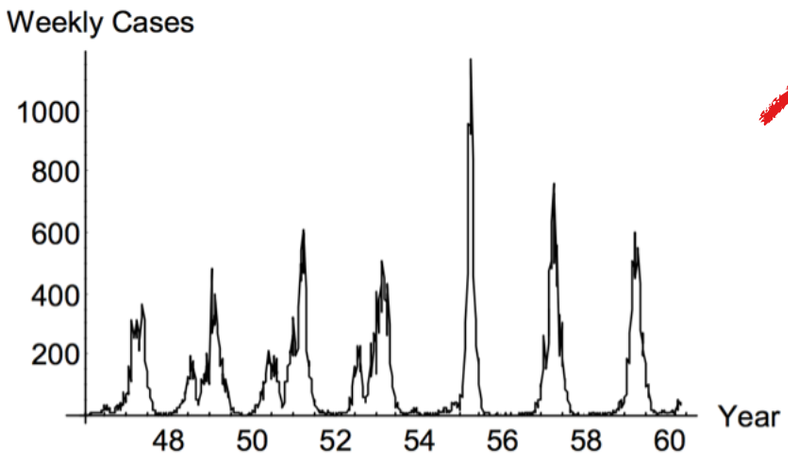
Case data alone is not enough!

Influenza A





Measles



Grenfell *et al.* **Science 2004**

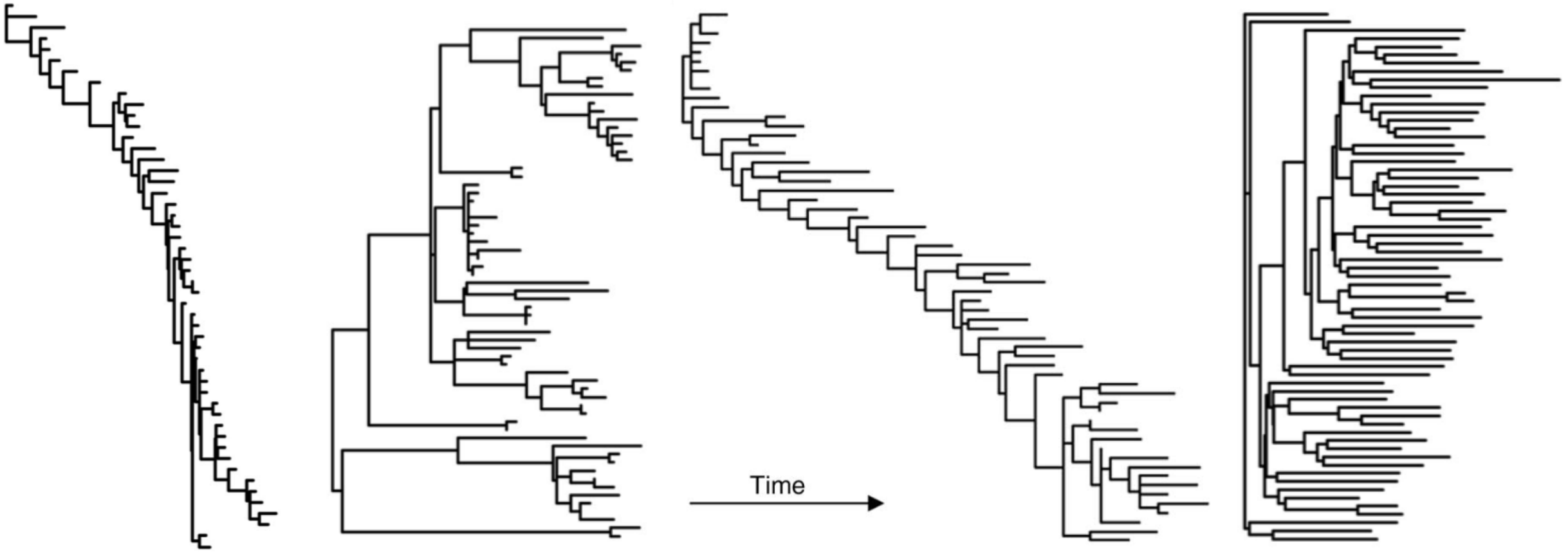
Genomes contain a signature of the epidemiological dynamics

Influenza A

Measles

HIV within-host

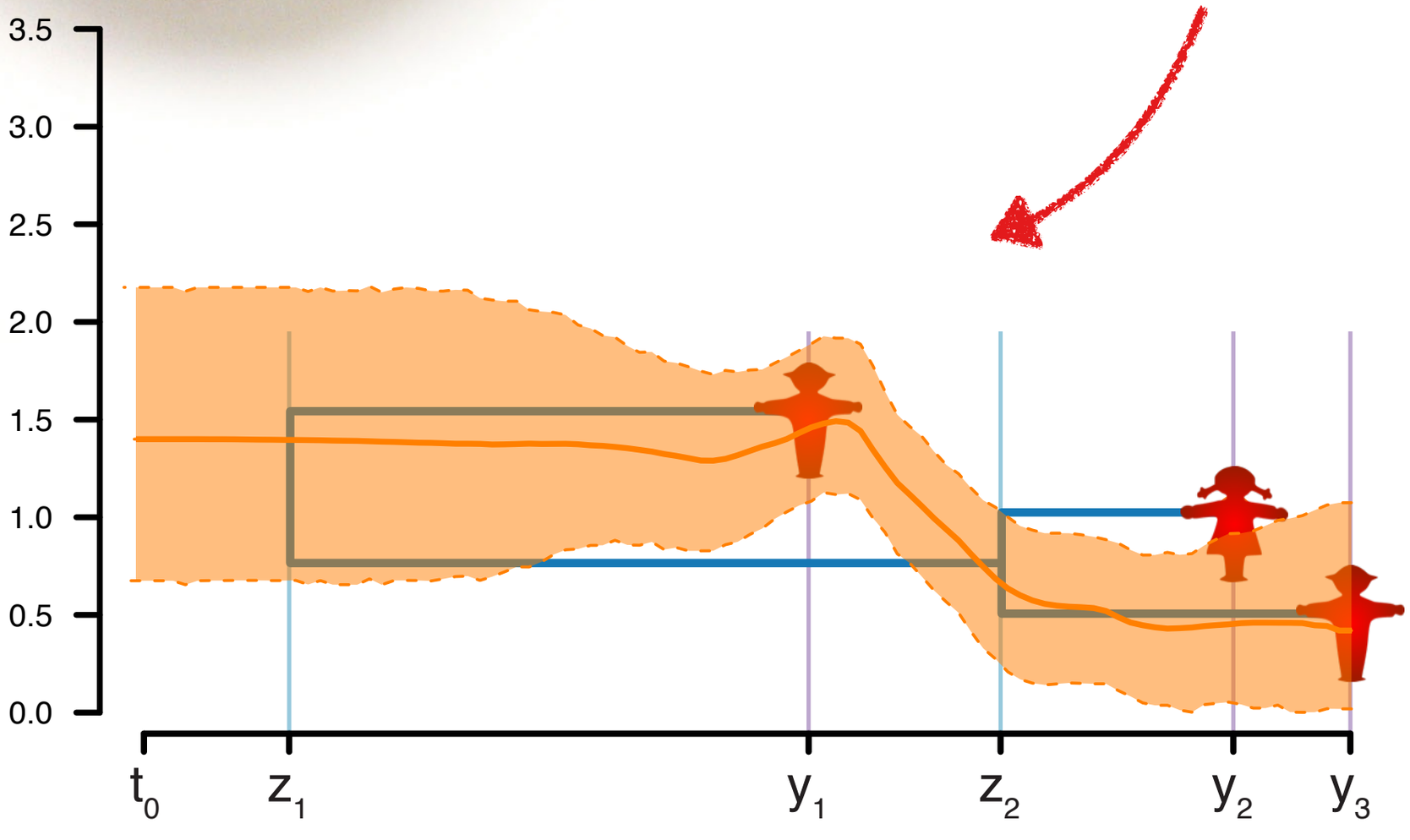
HCV



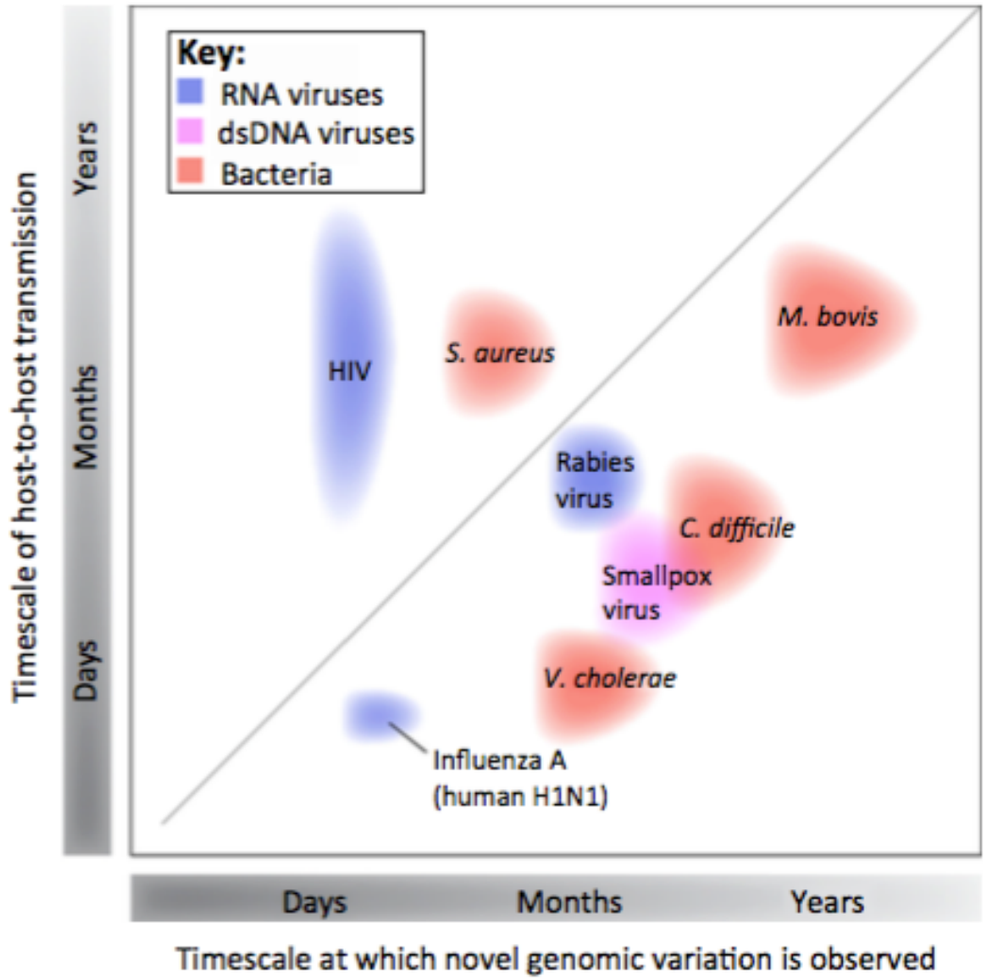
Phylodynamics aims to infer the **dynamics** responsible for the observed **phylogenies**

Grenfell *et al.* **Science 2004**


JUL
11
ACACACCC…
JUL
17
TCACACCT…
JUL
18
ACAGACTT…


3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
t_0
z_1
y_1
z_2
y_2
y_3

Measurably evolving populations



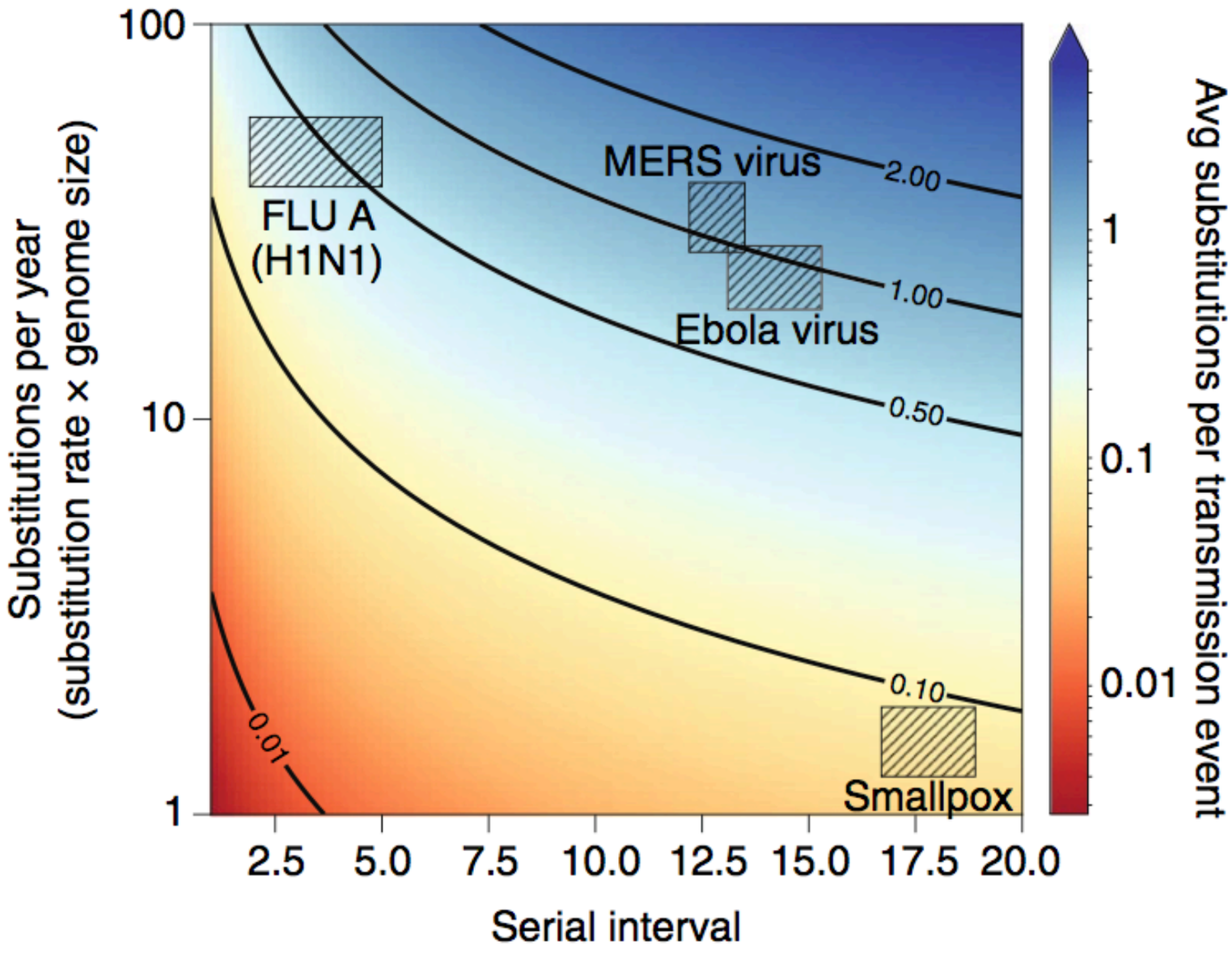
Novel genetic variation accumulates over the **sampling period**

- Large population size
- High mutation rate
- Short generation times
- This is the case for many pathogens (especially viruses)

Epidemiological and **evolutionary** dynamics occur on the same **timescale!**

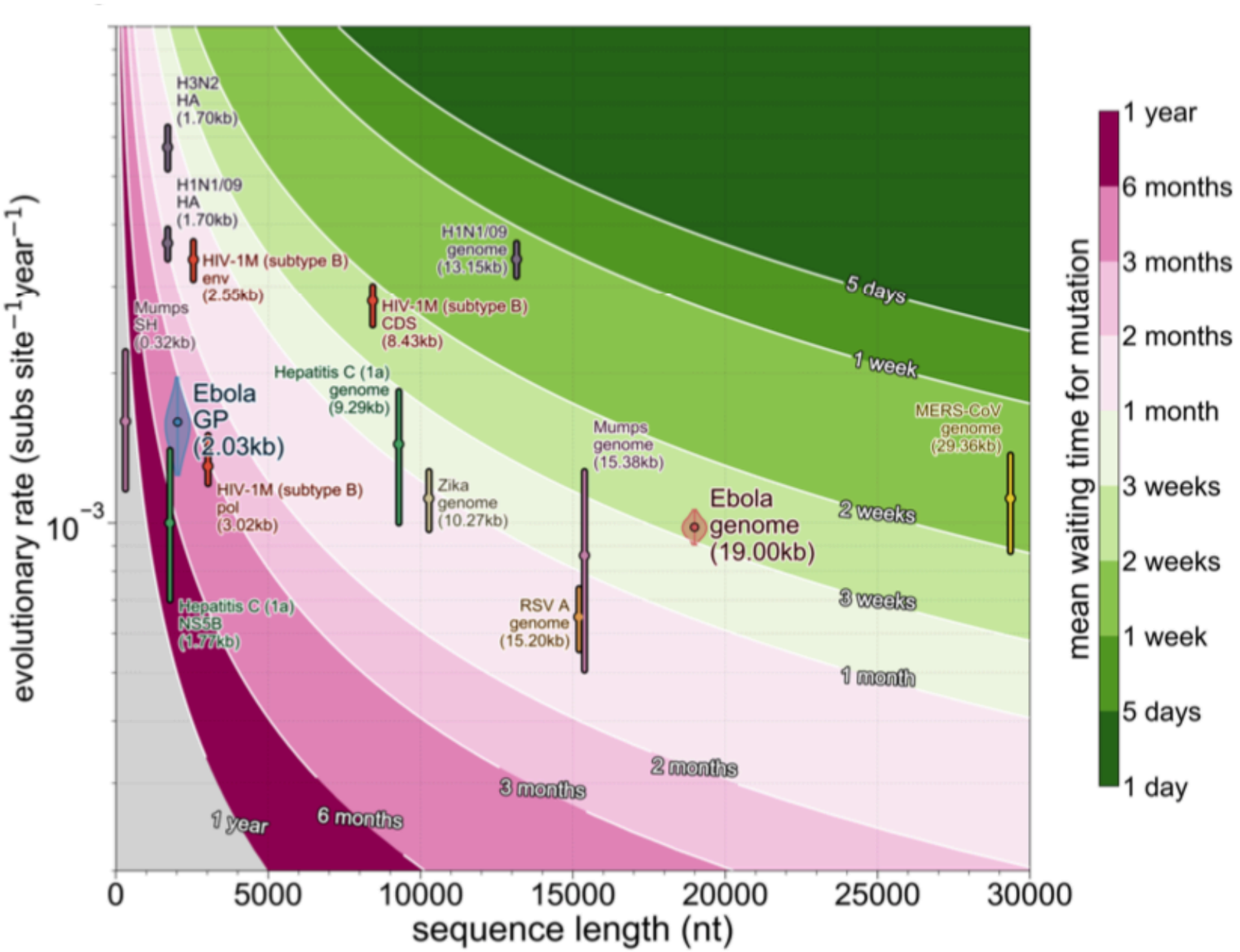
Biek *et al.*Trends in Ecology & Evolution **2015**

Measurably evolving populations and molecular clocks



Grubaugh et al. (2019) *Nature Microbiology* 4:10-19

Measurably evolving populations and molecular clocks

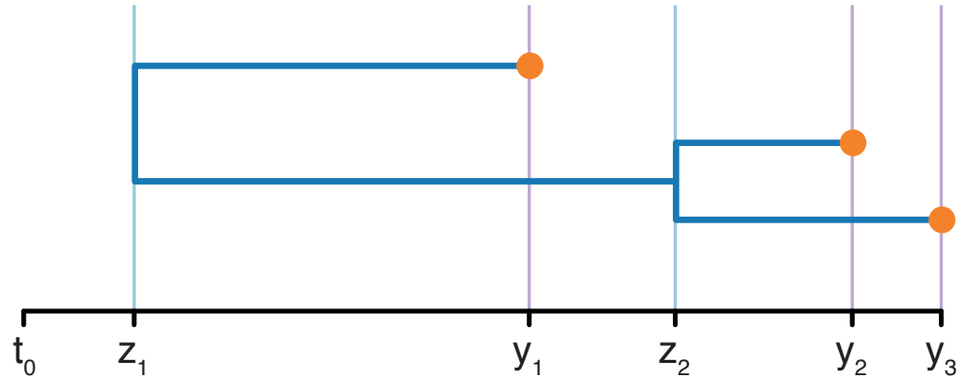


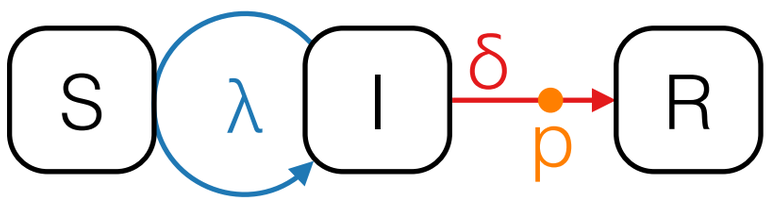
Dudas and Bedford (2019) *BMC Evol Biol 19*:232

What is phylodynamics?

Phylogenetics

- State of process
- Classification
- What are the ancestral relationships?





Phylodynamics

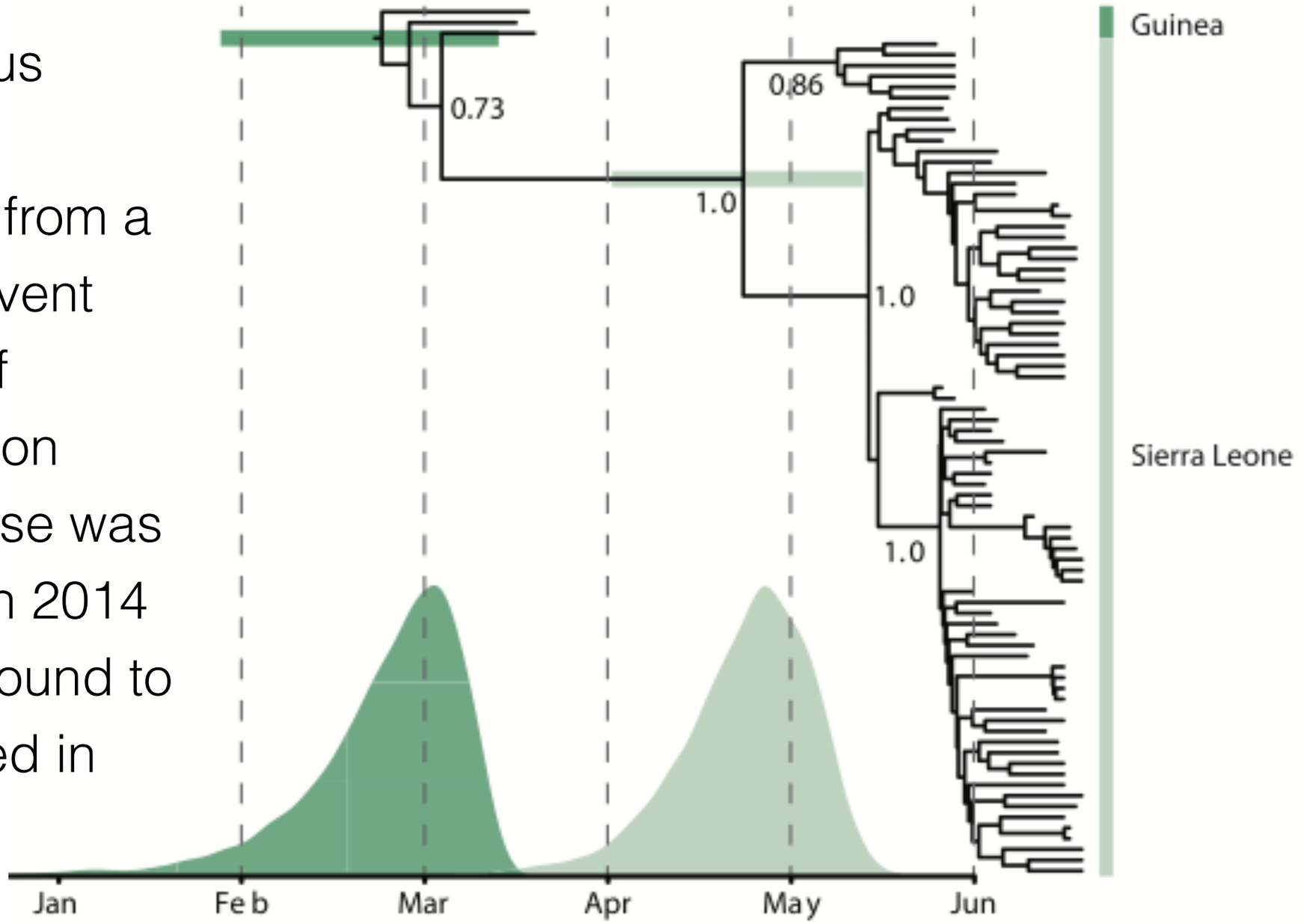
- Dynamics of process
- Epidemiological parameters
- How did we get here?

Phylodynamics questions

- How genetically diverse is a pathogen population?
- How do pathogen genomes change through time?
- How does pathogen genetic diversity vary through space?
- What processes or events determine these changes?
- When did an epidemic start?
- Where did it come from?
- How fast is it transmitting?
- In what direction is it spreading?
- Are hosts X,Y & Z epidemiologically linked?

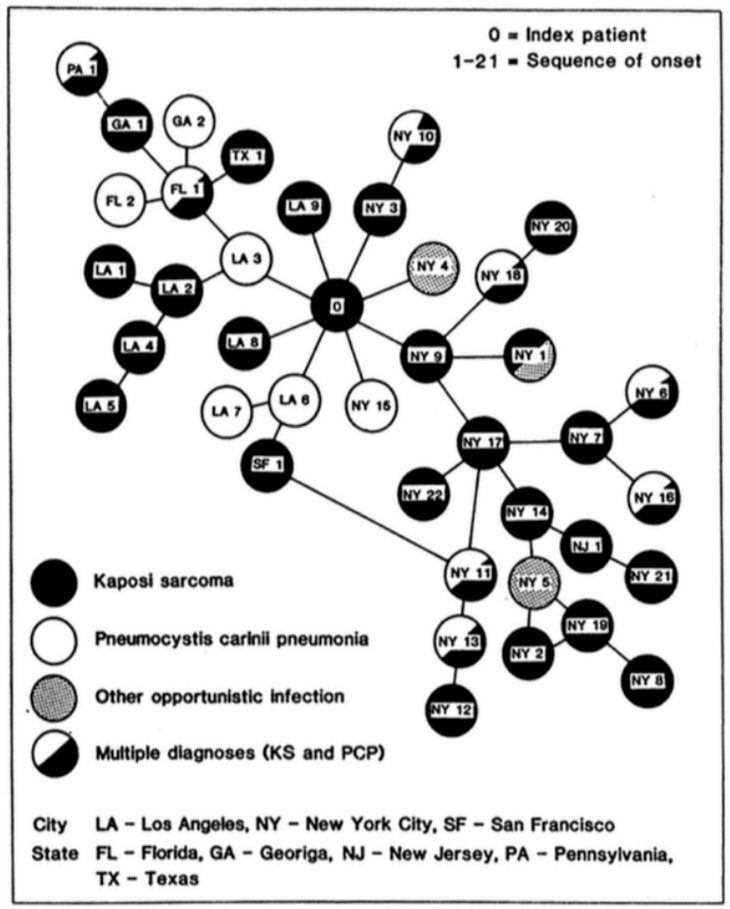
West African Ebola virus epidemic

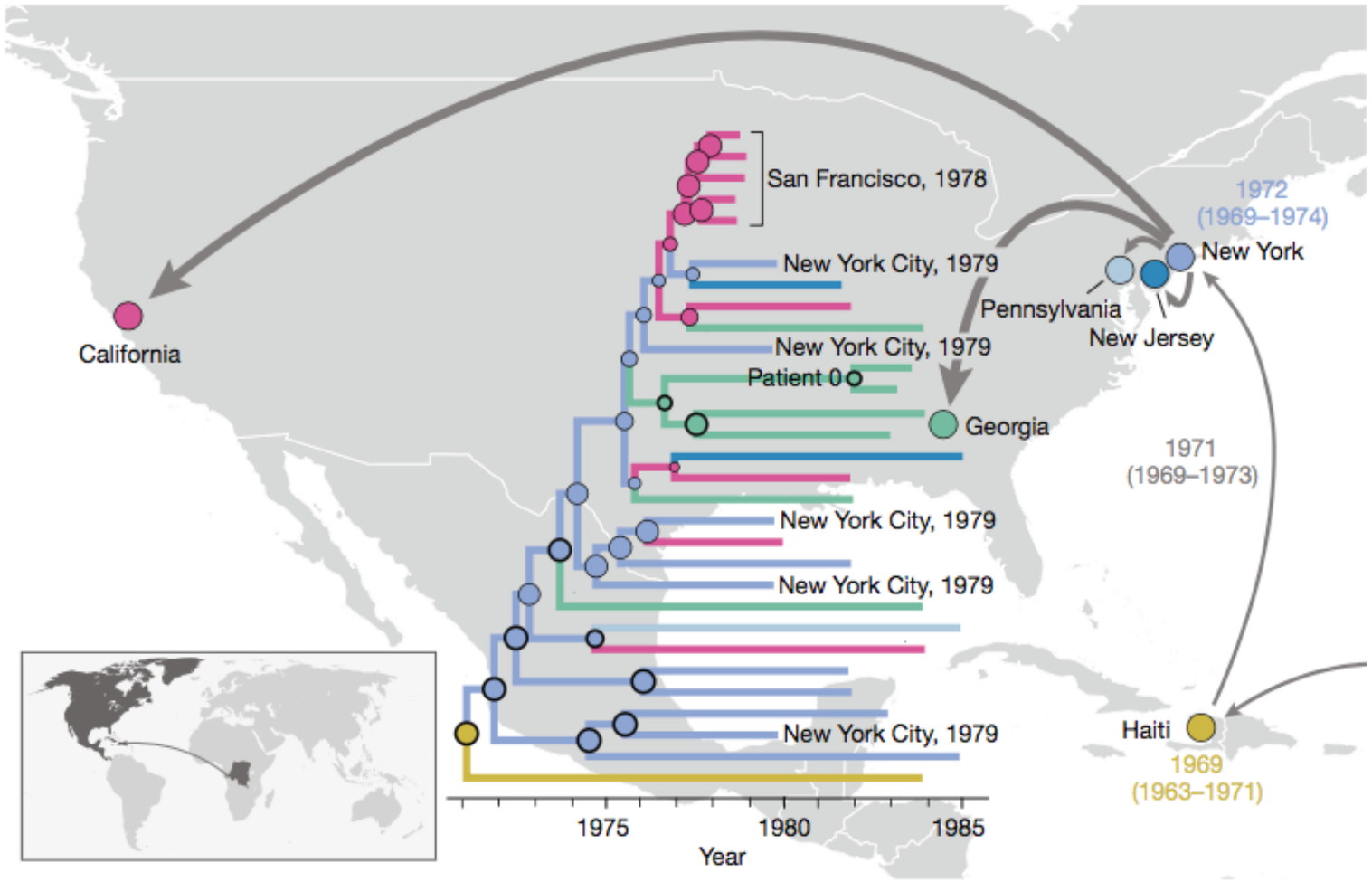
- Biggest Ebola virus epidemic
- Epidemic started from a single zoonosis event
- Several months of cryptic transmission before the first case was detected in March 2014
- Index case later found to have been infected in December 2013



Gire *et al.* Science **2014**

HIV-1 “Patient 0”



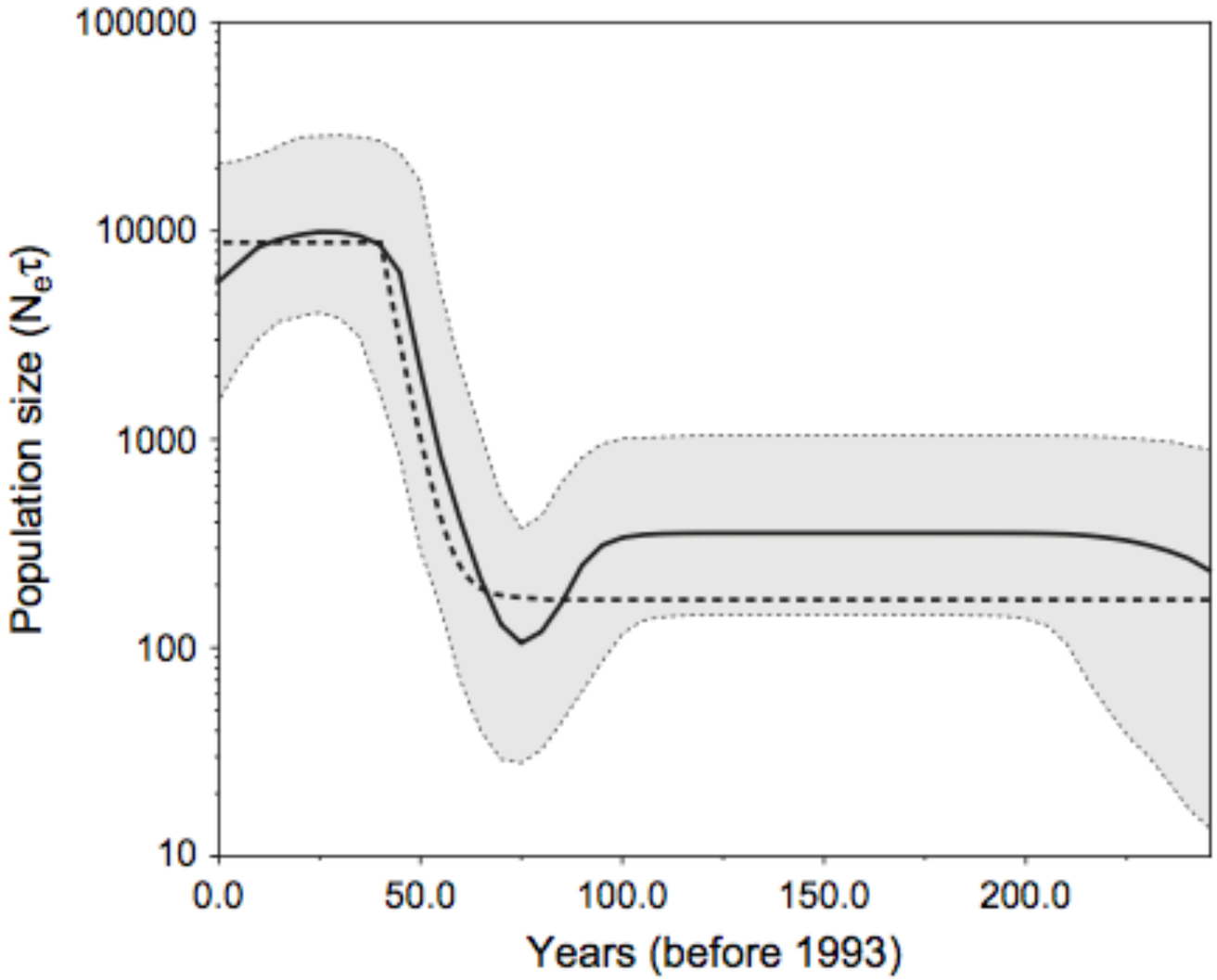


- So-called “Patient 0” could not be the index case for the US HIV-1 epidemic
- US epidemic emerged from a pre-existing Caribbean epidemic ~1970

Worobey *et al.* Nature **2016**

Hepatitus C virus in Egypt

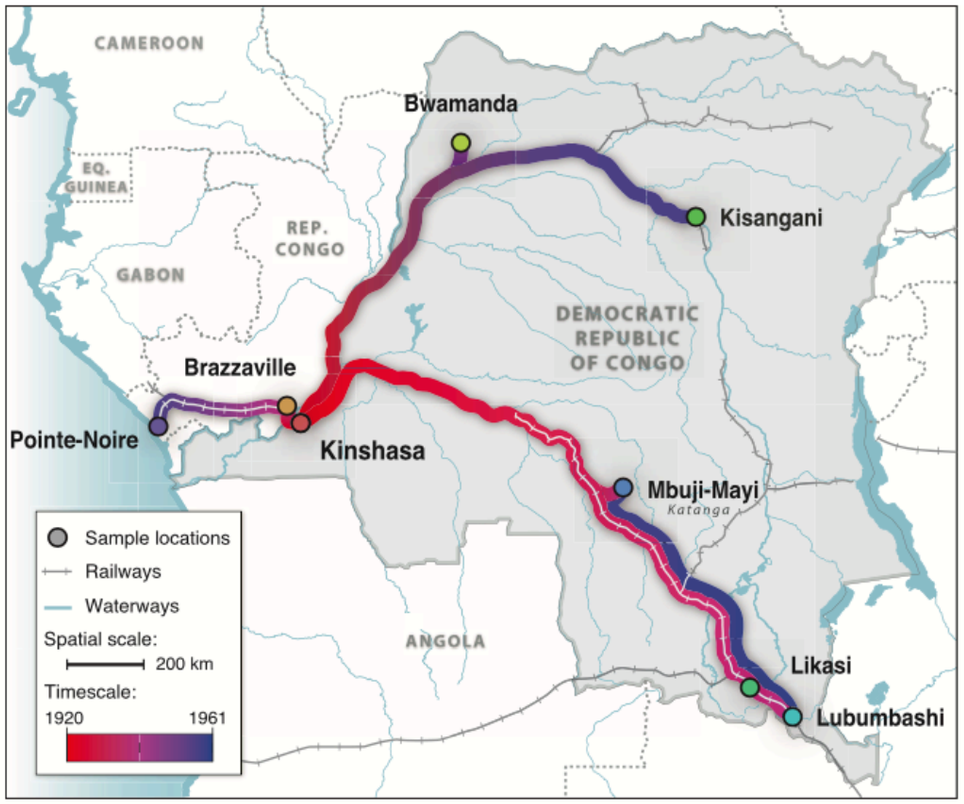
Combine with model of population dynamics

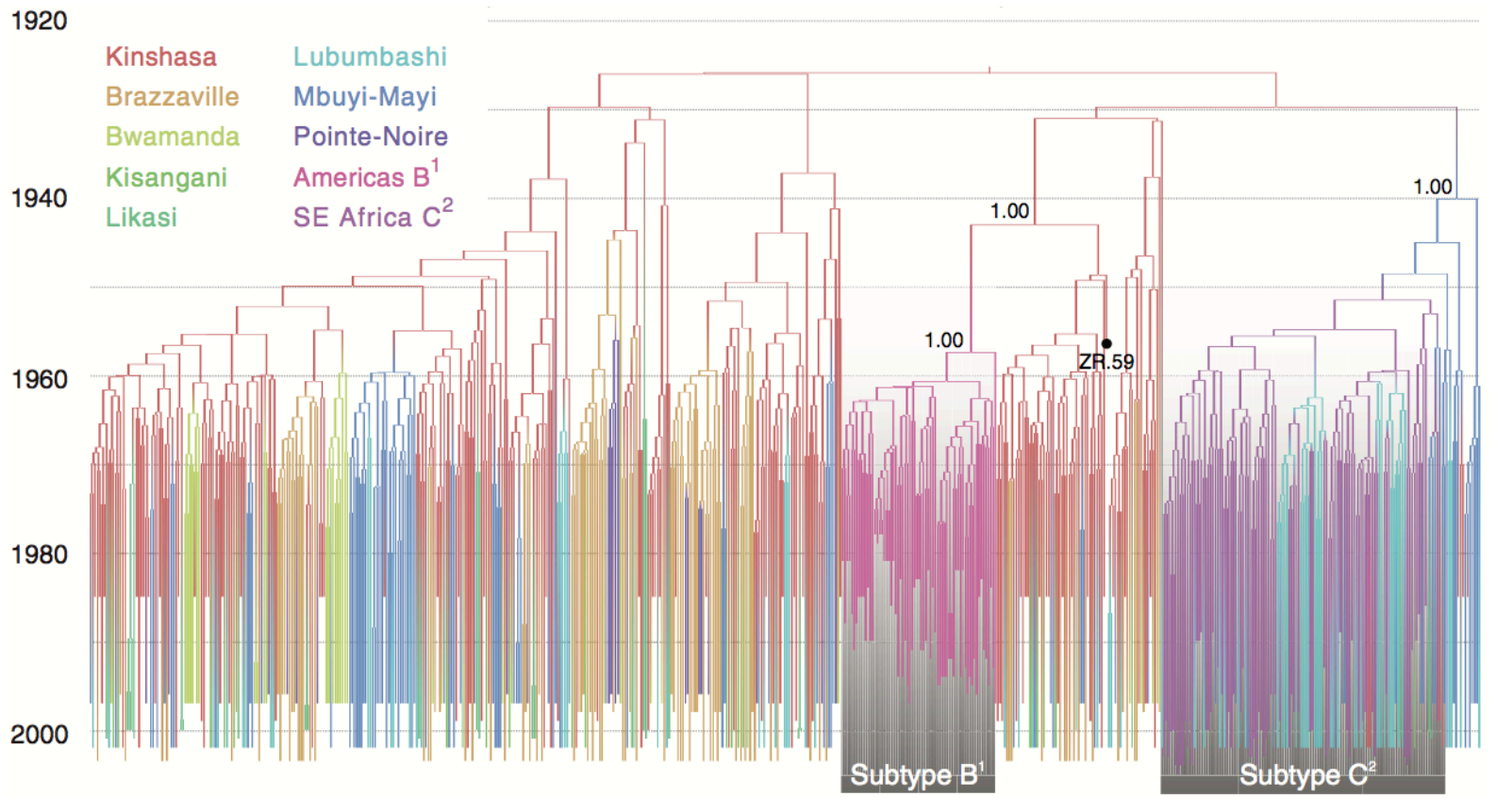


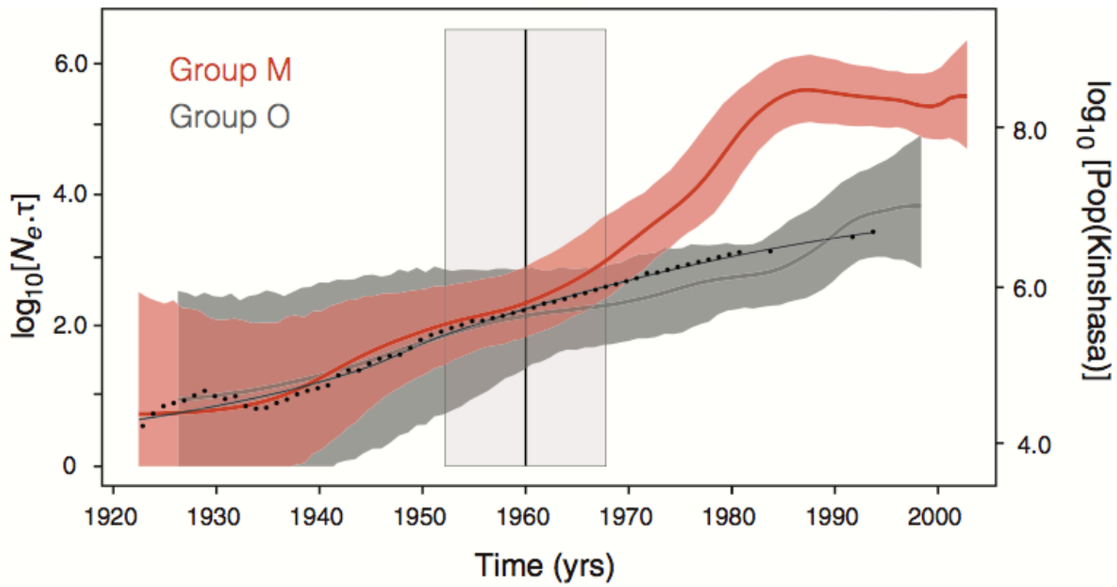
- Egypt has the highest prevalence (much higher than neighbouring countries)
- Exponential growth period during the middle of the 20th century
- Coincides with a period of possibly contaminated anti-schistosomiasis injections

Drummond *et al.* MBE **2005**

HIV-1 pandemic

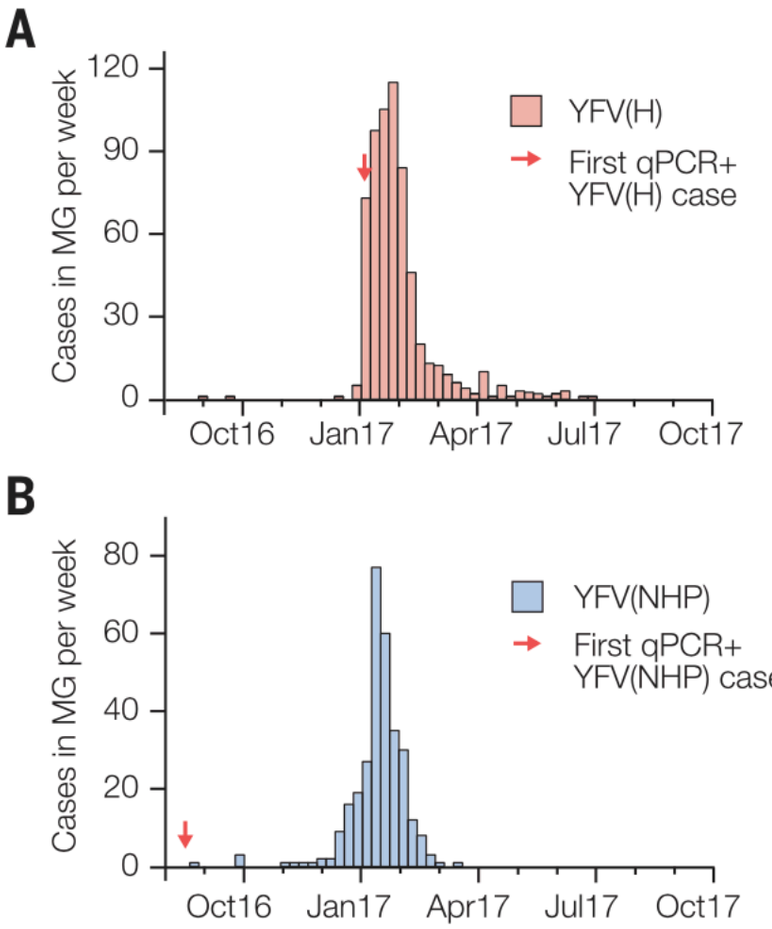


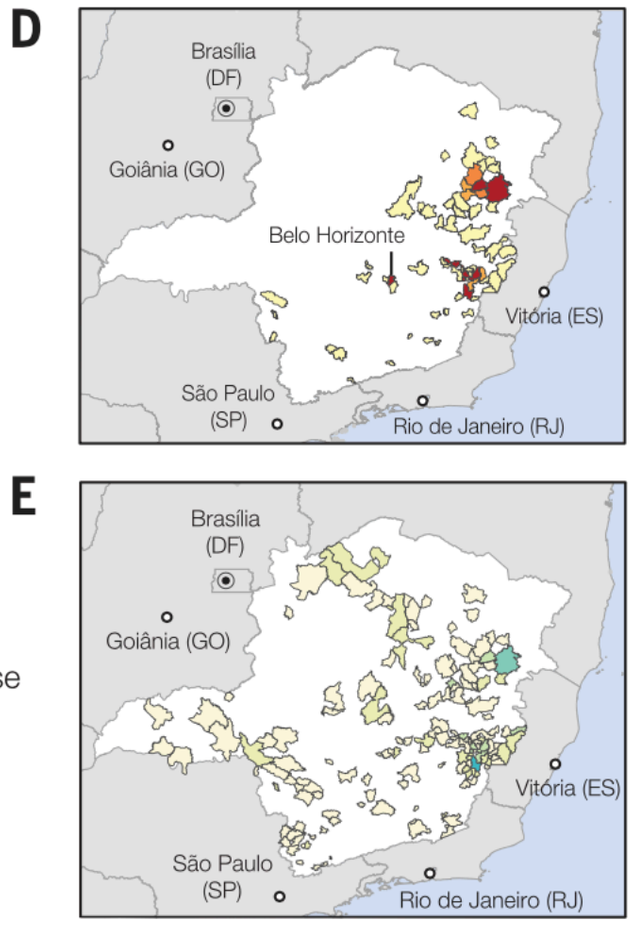


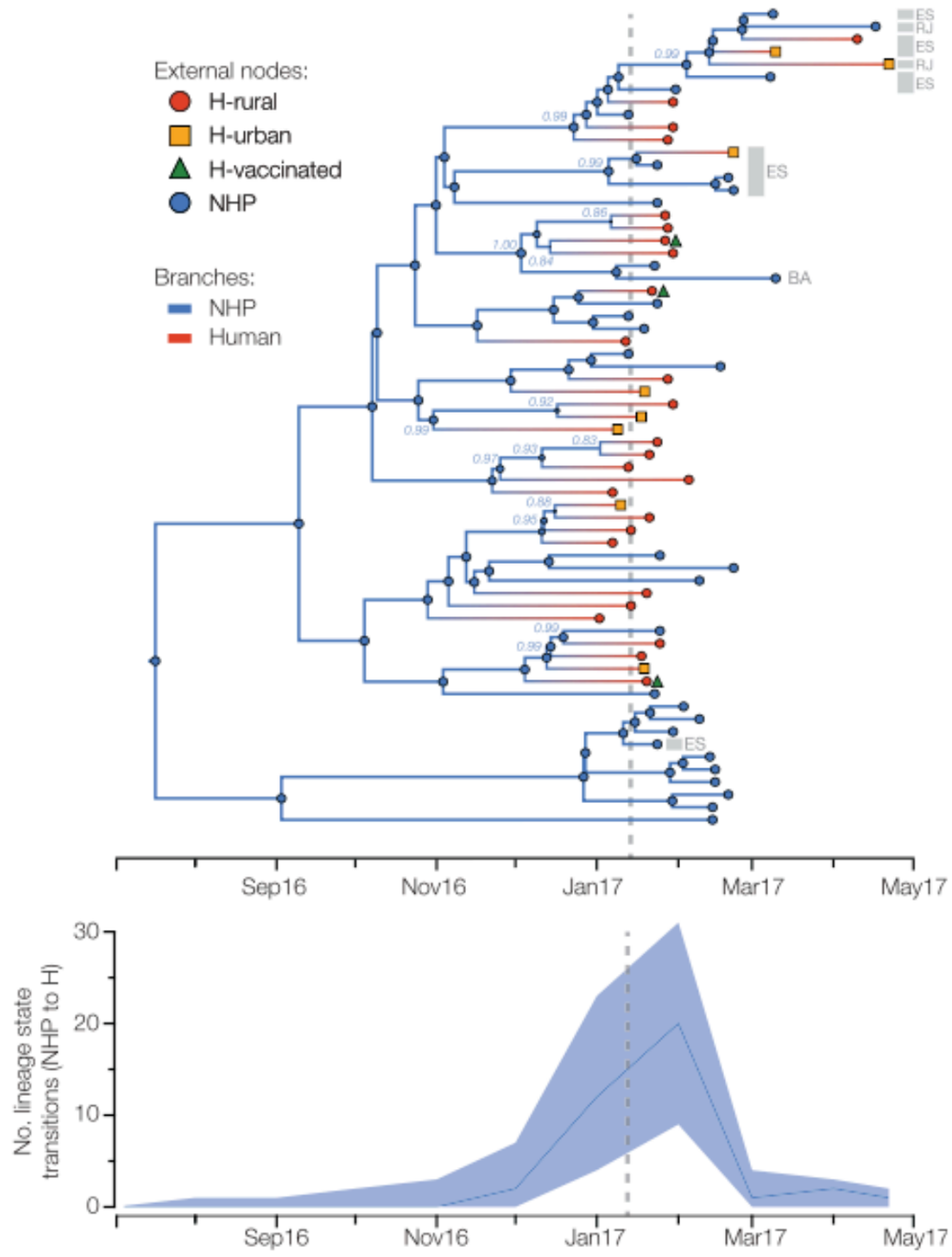


Faria *et al.* Science **2014**

Yellow fever in Brazil



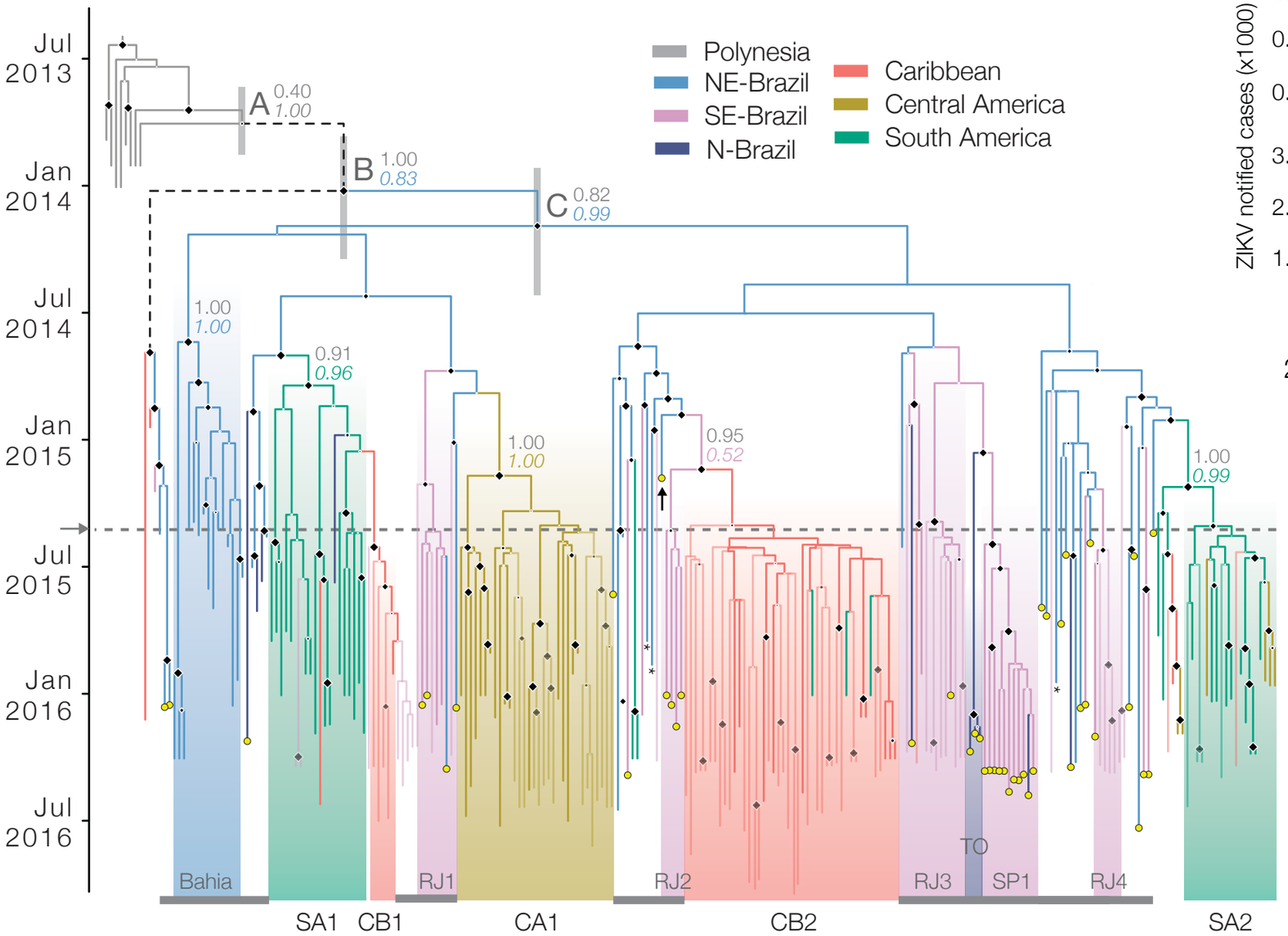


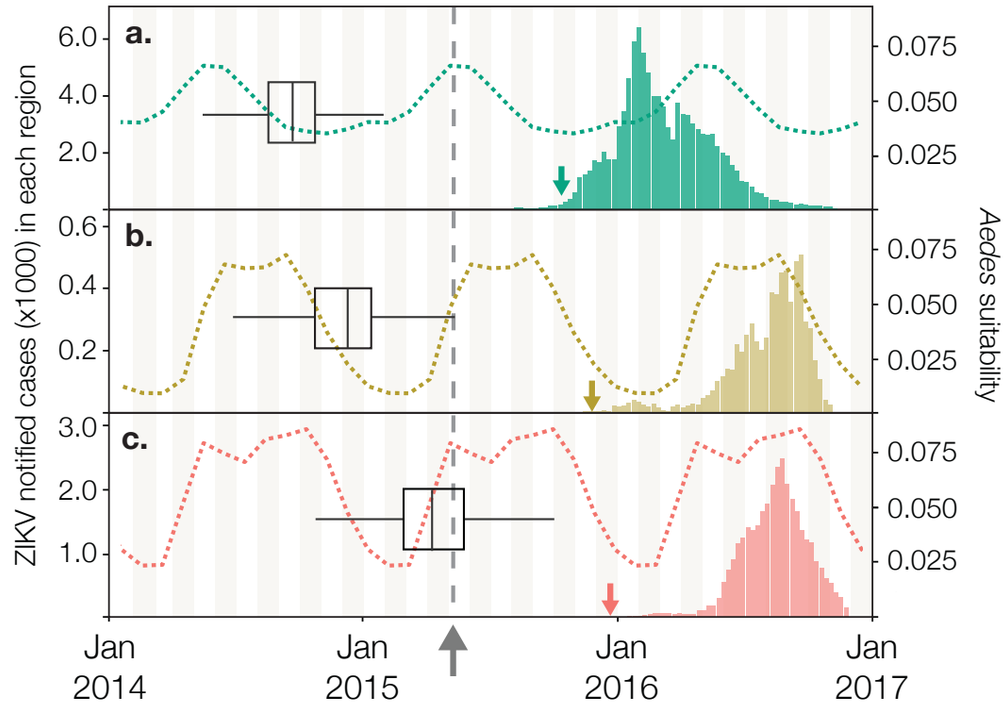


Faria *et al.* Science **2018**

Zika virus in the Americas

Combine with phylogeographic migration model

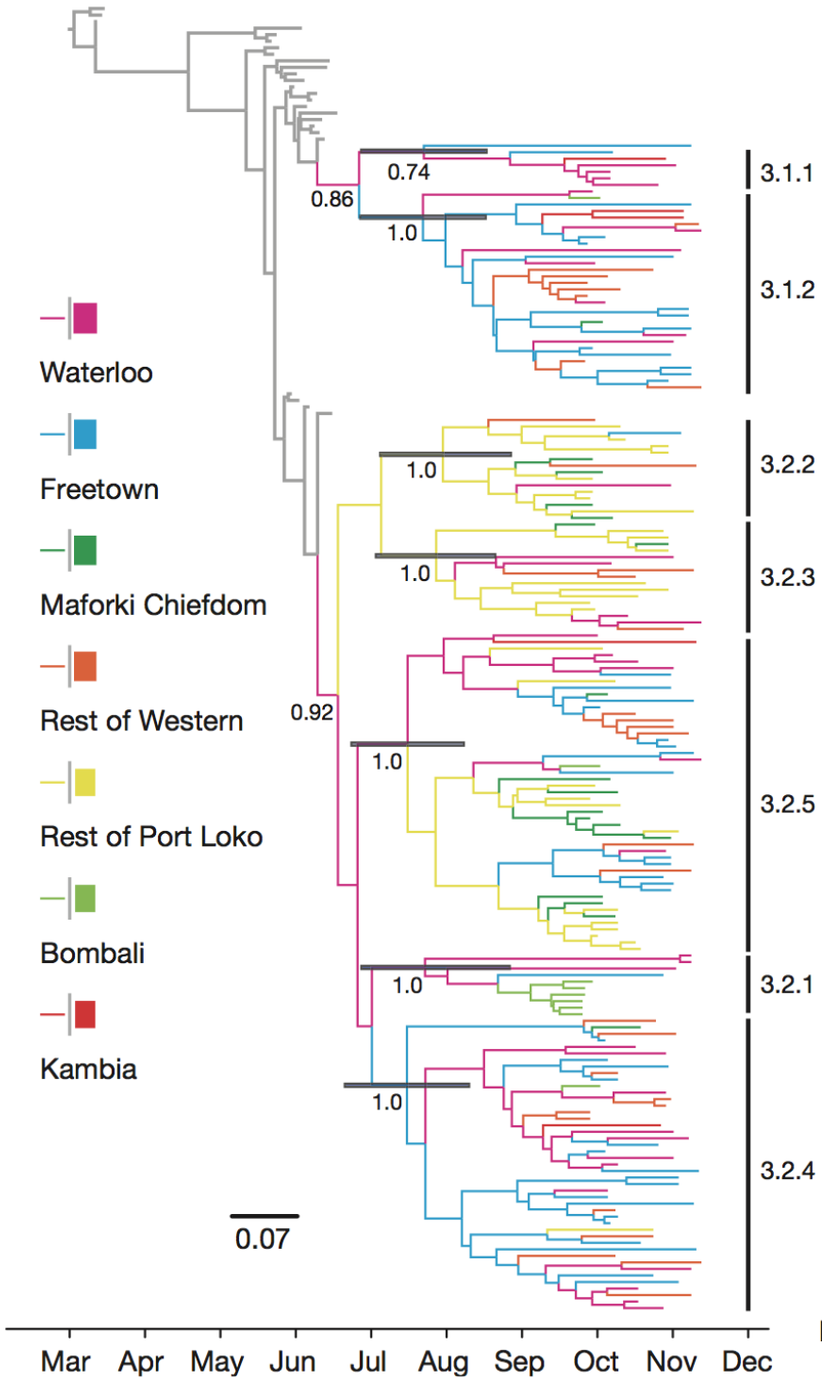


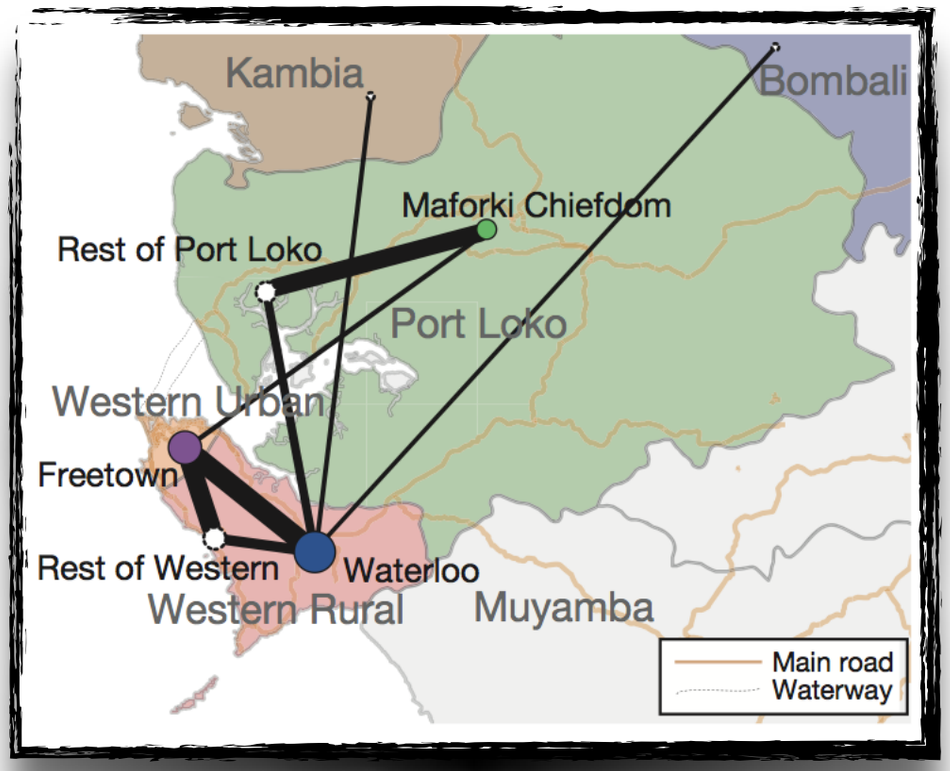


- Single introduction from French-Polynesia to Brazil
- ~1 year of cryptic transmission
- NE-Brazil is the source for the whole American epidemic

Faria *et al.* Nature **2017**

Phylogeography: Ebola in Sierra Leone

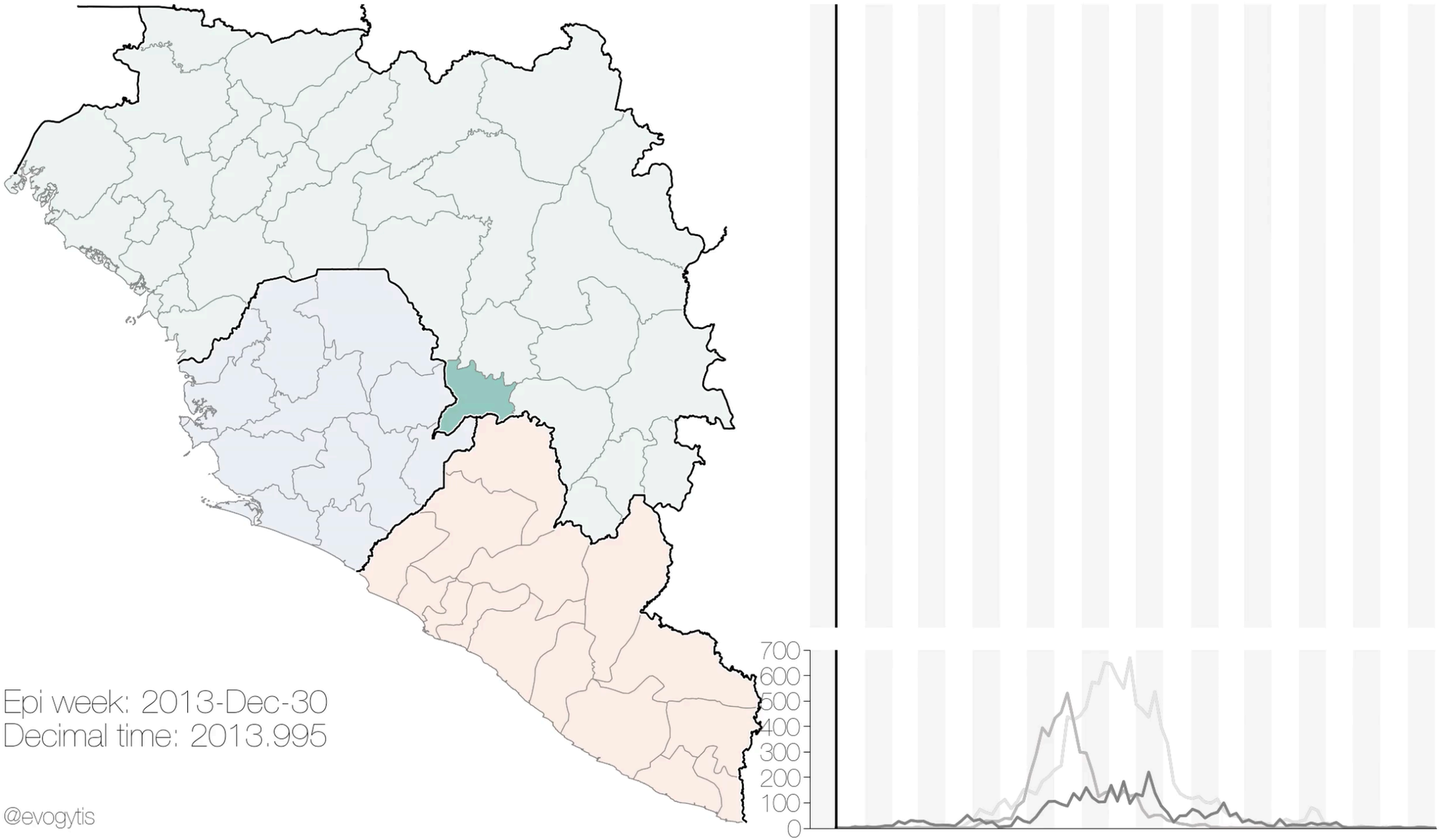




- Infer how the virus spread across the region
- Infer the most probable ancestral states
- Infer migration rates between different regions

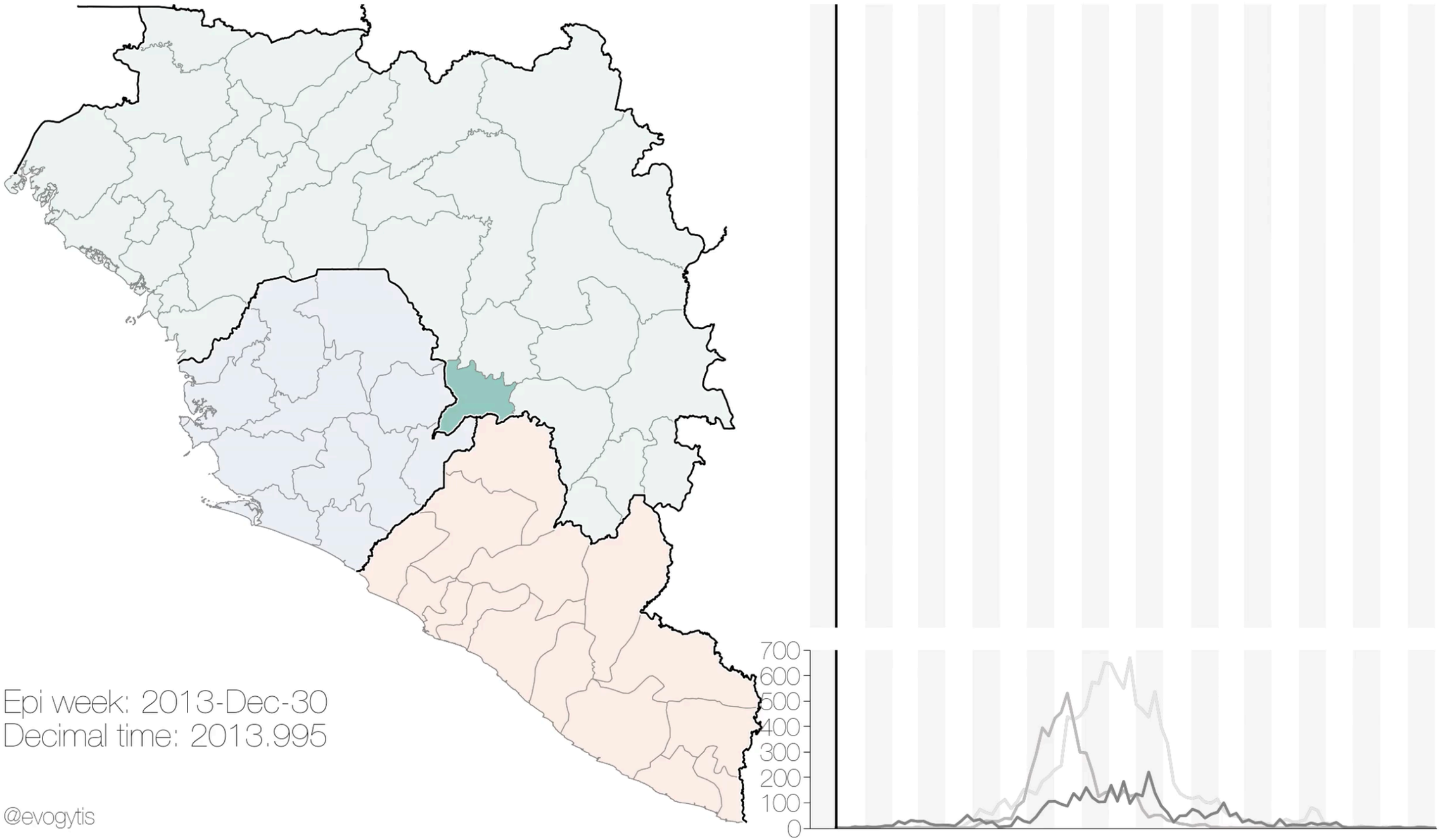
Tong *et al.* **Science 2015**

Phylogeography: Ebola in West Africa



Dudas *et al.* **Nature 2017**

Phylogeography: Ebola in West Africa



Dudas *et al.* **Nature 2017**

Seasonal Influenza (2010)

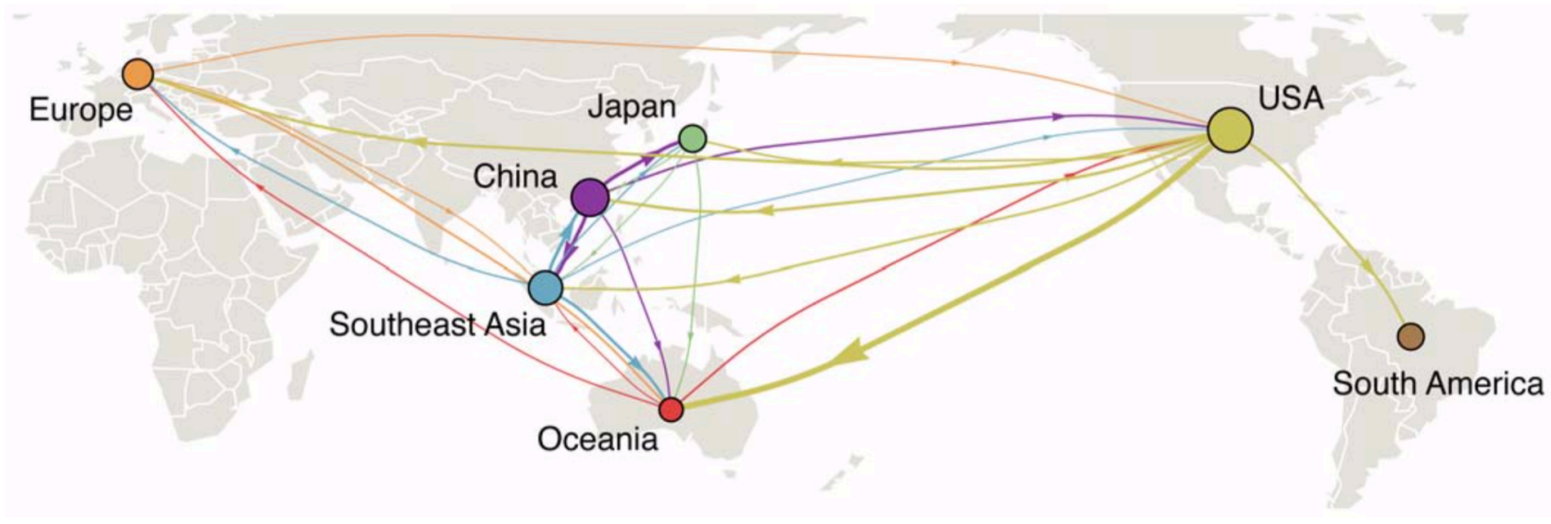
OPEN ACCESS Freely available online

PLoS PATHOGENS

Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2)

Trevor Bedford[1,2]*, Sarah Cobey[1,2], Peter Beerli[3], Mercedes Pascual[1,2]

1 Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, United States of America, **2** Howard Hughes Medical Institute, University of Michigan, Ann Arbor, Michigan, United States of America, **3** Department of Scientific Computing, Florida State University, Tallahassee, Florida, United States of America



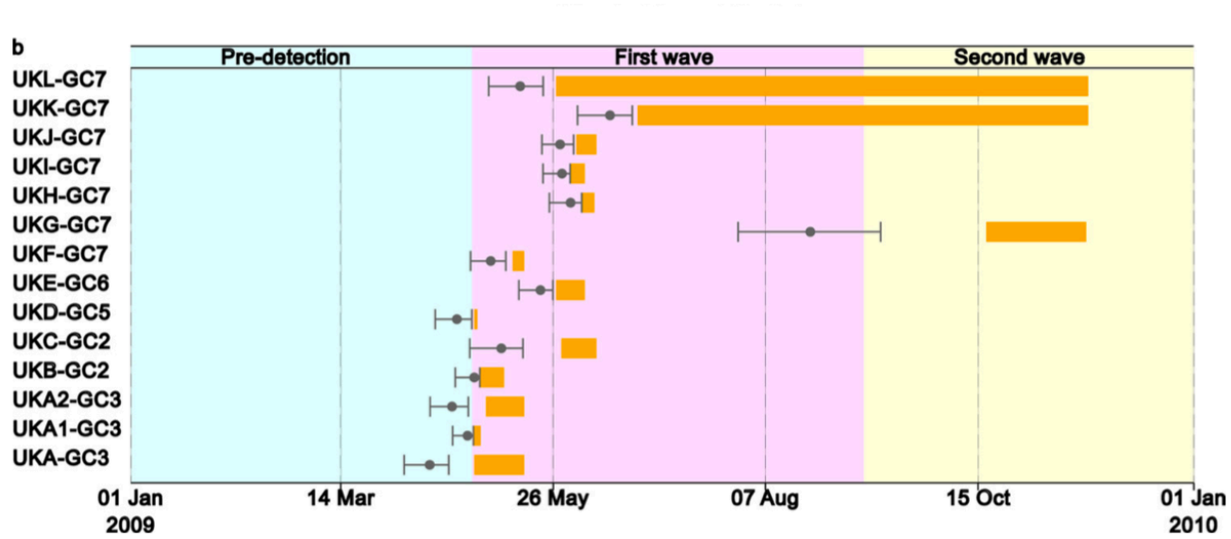
"Examining the influenza genealogy, it is apparent that regional outbreaks often result from **very few immigration events**, consistent with previous results. For example, the 2003 epidemic in Oceania appears almost completely monophyletic and can **trace its history to a single migration event** (or perhaps multiple migration events of identical strains) in early 2003."

Pandemic H1N1 (2012)

JVI
Journals.ASM.org

Evolutionary Dynamics of Local Pandemic H1N1/2009 Influenza Virus Lineages Revealed by Whole-Genome Analysis

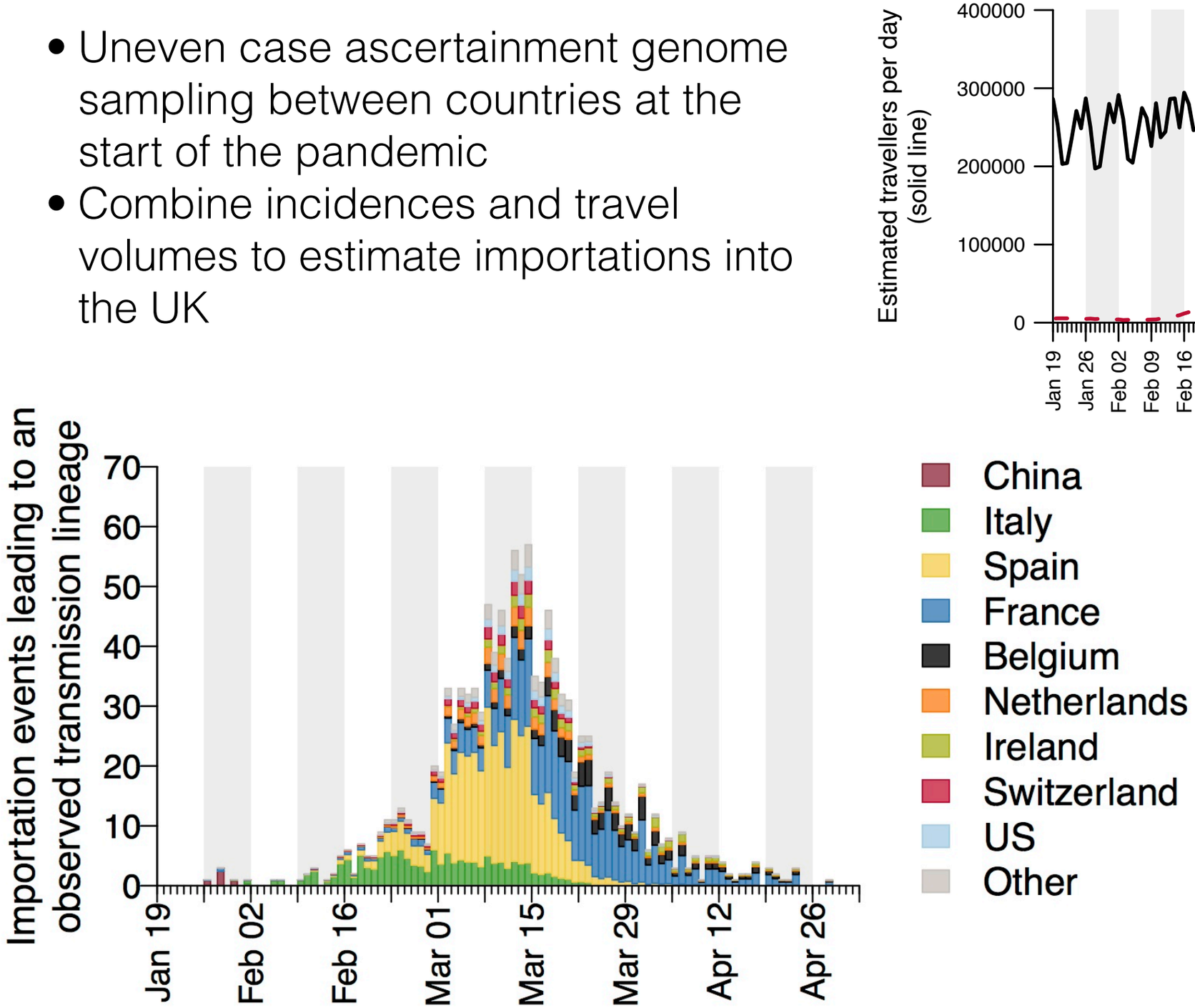
Gregory J. Baillie,[a] Monica Galiano,[b] Paul-Michael Agapow,[b] Richard Myers,[b] Rachael Chiam,[a] Astrid Gall,[a] Anne L. Palser,[a] Simon J. Watson,[a] Jessica Hedge,[c] Anthony Underwood,[b] Steven Platt,[b] Estelle McLean,[b] Richard G. Pebody,[b] Andrew Rambaut,[c,d] Jonathan Green,[b] Rod Daniels,[e] Oliver G. Pybus,[f] Paul Kellam,[a] and Maria Zambon[b]

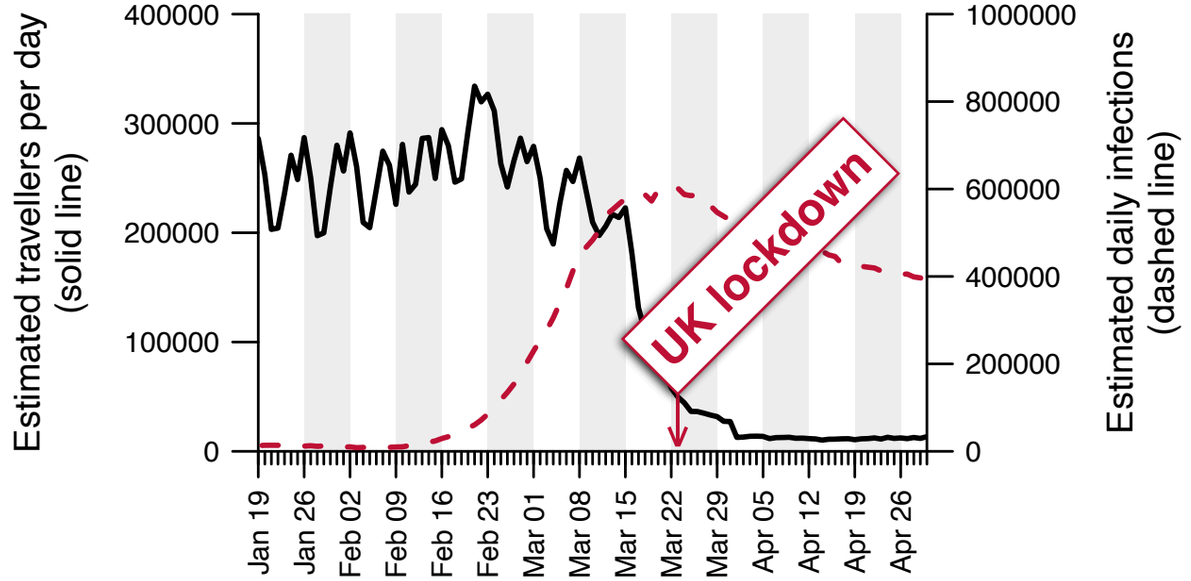


"This suggests that our sampling of the lineages that arrived in the United Kingdom at the start of the outbreak was **reasonably complete** and that the **13 clusters and 52 dispersed lineages** account for much of the United Kingdom pandemic H1N1/09 virus genetic diversity."

Introduction of SARS-CoV-2 into the UK

- Uneven case ascertainment genome sampling between countries at the start of the pandemic
- Combine incidences and travel volumes to estimate importations into the UK



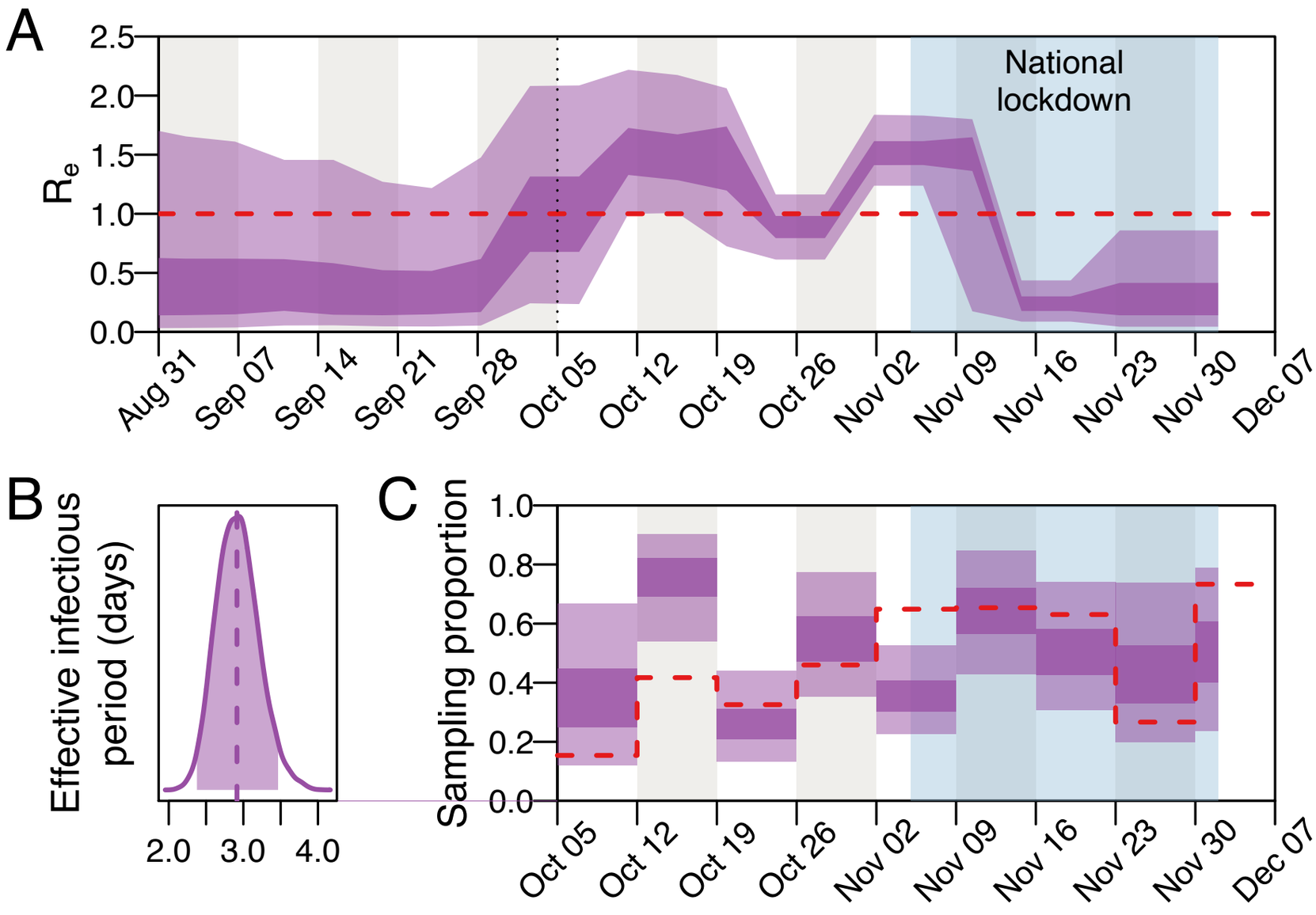


du Plessis et al. (2021) *Science*

R_e of SARS-CoV-2 on a university campus

(Michaelmas term 2020)

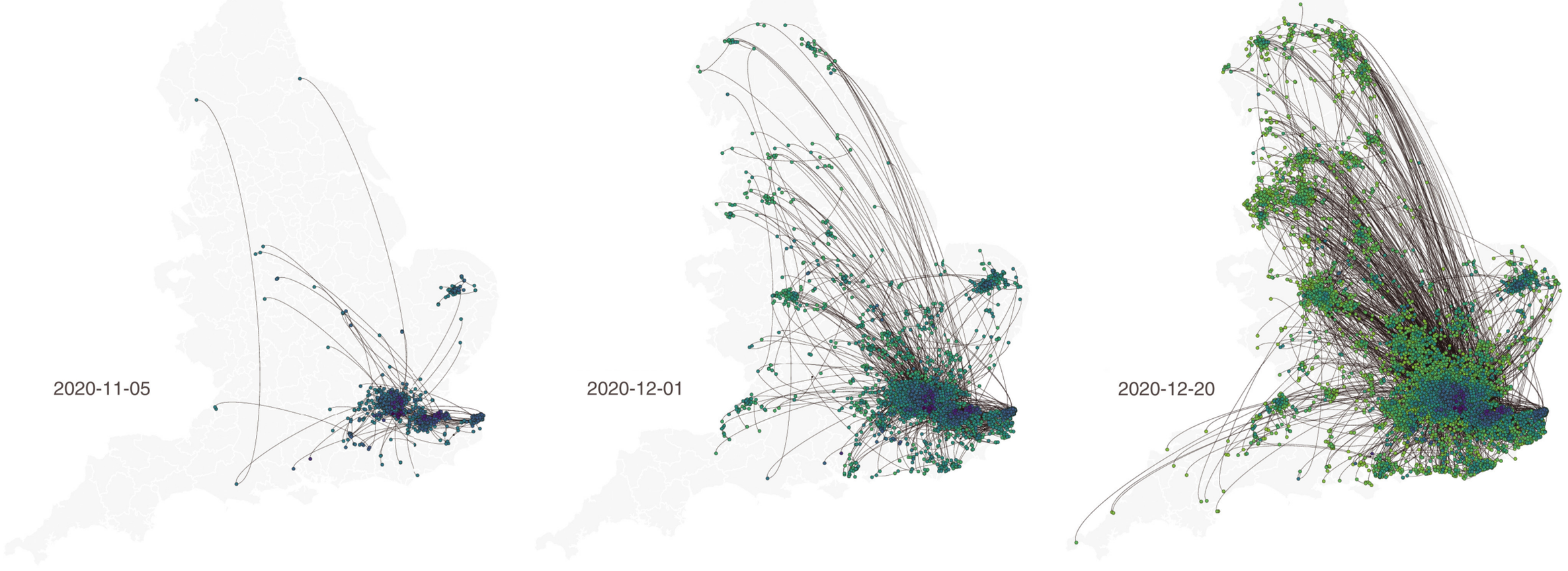
- Opt-in asymptomatic testing program for all Cambridge students
- Were campus-wide measures effective at stopping spread
- Did the national lockdown have an effect?



Aggarwal et al. (2021) *Nature Communications*

Spread of Alpha VOC in England

- A new variant (B.1.1.7) appeared in southeast England in September 2020 and spread rapidly
- Later designated as the Alpha variant of concern (VOC)
- Use a phylogeographic model to track spread



Kraemer et al. (2021) *Science*

cEvo

virojn

01110011 01110110 01101001 01110010 01110101

Understanding
virus-host interactions

Bayesian inference recap

Louis du Plessis

ETH zürich

D BSSE

On the program for today

(1) What is phylodynamics?

(2) Bayesian inference recap

(3) BEAST2 introduction

Tutorial: Molecular clock dating (part i)

(4) Molecular clock models

Tutorial: Molecular clock dating (part ii)

(5) Setting priors

Tutorial: Phylodynamics (part i)

(6) Tree priors

Tutorial: Phylodynamics (part ii)

data

- Samples drawn from a realisation of some **stochastic process**
- Typically an alignment of DNA or RNA sequences
- Can also be amino acids or codons

model

- **Model** is a mathematical description of the **process** that generated the data
- Each model has a number of **model parameters**
- Parameters are **random variables**

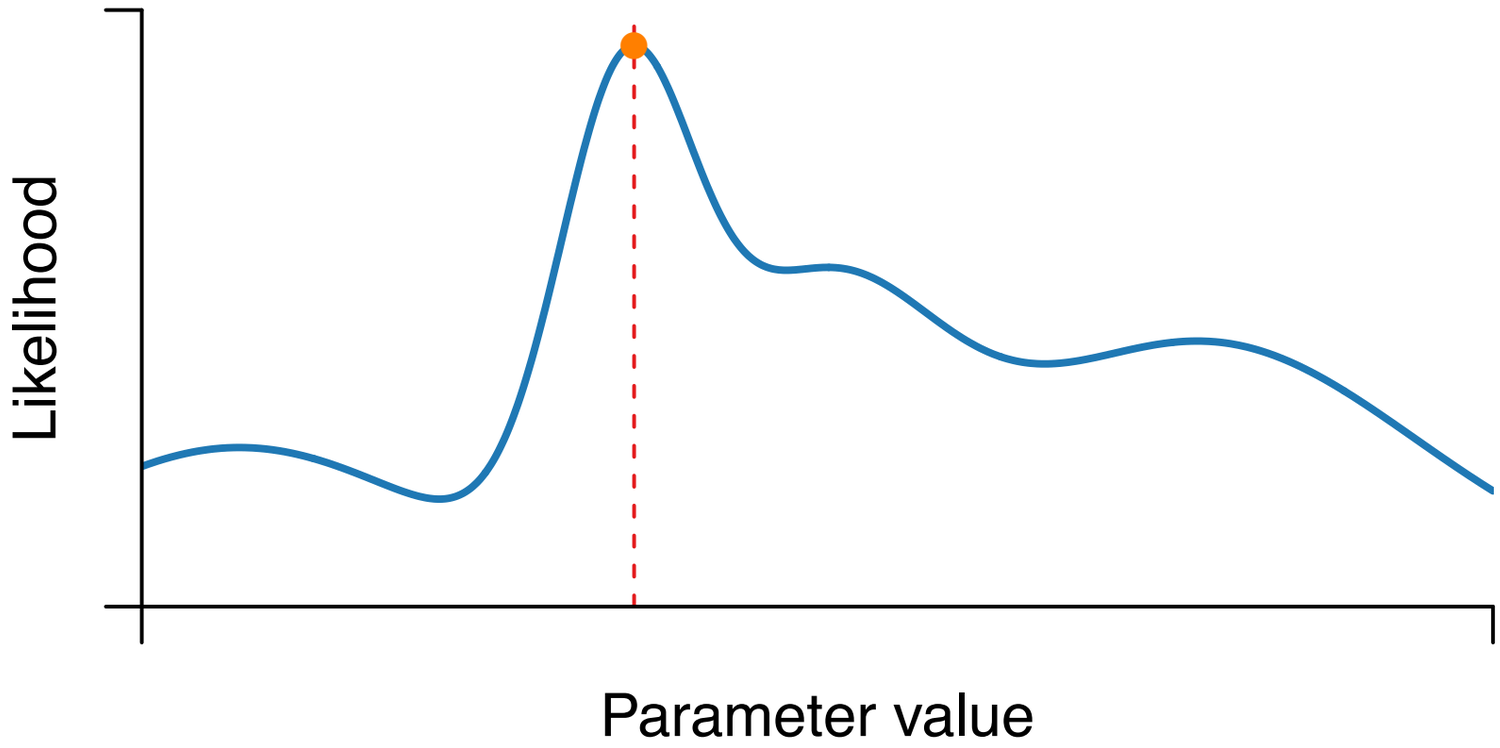
hypothesis

- **Hypothesis** is some statement about the model parameters
- Hypothesis assigns **values** to the model parameters
- Data used to decide if hypothesis is reasonable or not
- May only concern some model parameters. The rest are **nuisance parameters**

Maximum-likelihood inference

Likelihood → P(data | hypothesis, model)

- Likelihood is proportional to the probability of observing the data given a **hypothesis**
- Find the model parameters that maximise the likelihood



Inference frameworks

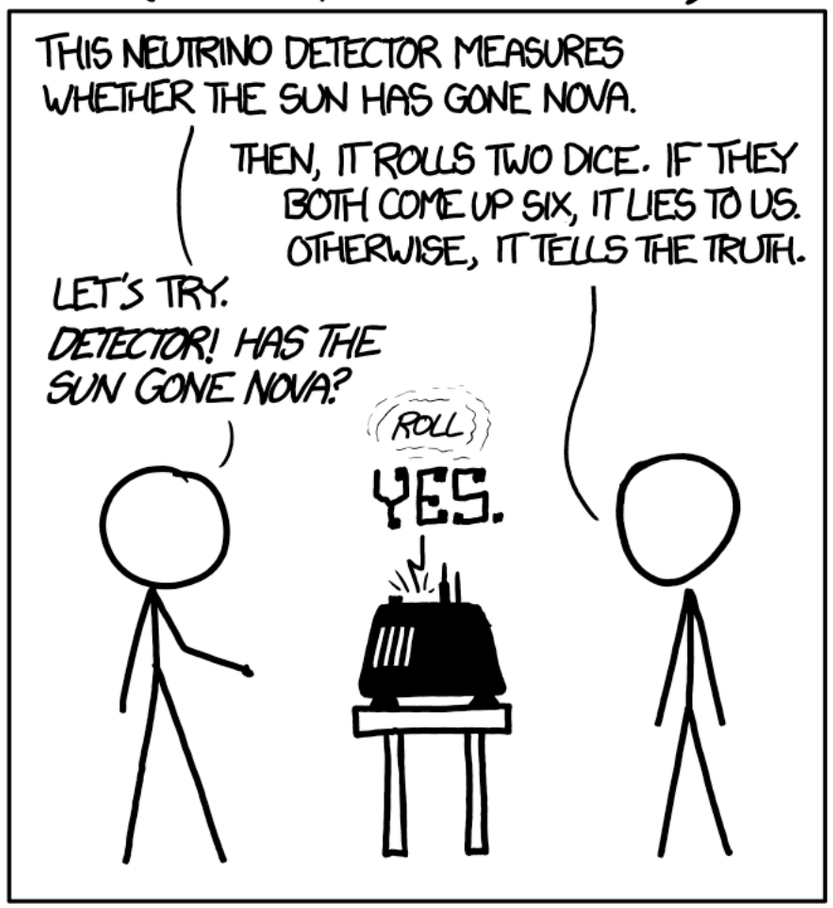
Frequentist

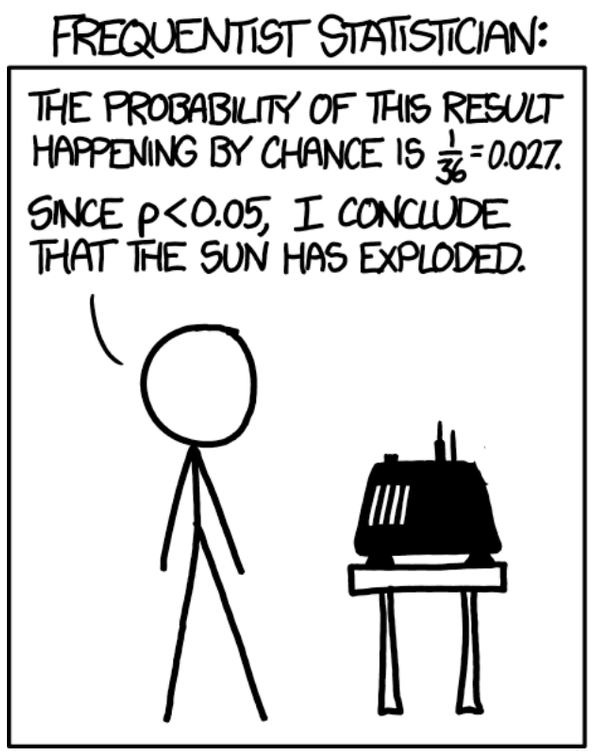
- Probabilities refer to the outcome of experiments
- Probability is the frequency of outcomes if an experiment is repeated many times
- Likelihood is the degree to which data supports a hypothesis

Bayesian

- Probability is a measure of the plausibility of propositions conditional on available information
- Data and model parameters are described by probabilities
- Probability reflects our belief in a hypothesis as well as representing the outcome of experiments
- Hypotheses have probabilities in the absence of data

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)





Bayesian inference

Prior → P(hypothesis I model)

- Have some degree of belief in our hypothesis
- From external sources, previous analyses etc.

Likelihood → P(data I hypothesis, model)

- Likelihood is proportional to the probability of observing the data given a hypothesis

Posterior → P(hypothesis I data, model)

- Combines information from the data **(likelihood)** and previous knowledge **(prior)**

Bayesian inference

Prior → P(hypothesis)

- Have some degree of belief in our hypothesis
- From external sources, previous analyses etc.

Likelihood → P(data | hypothesis)

- Likelihood is proportional to the probability of observing the data given a hypothesis

Posterior → P(hypothesis | data)

- Combines information from the data **(likelihood)** and previous knowledge **(prior)**

Bayesian inference estimates the posterior probability distribution instead of a likelihood curve

Bayesian inference

Likelihood

$$P(\textbf{hypothesis} \mid \textbf{data}) = \frac{P(\textbf{data} \mid \textbf{hypothesis})P(\textbf{hypothesis})}{P(\textbf{data})}$$

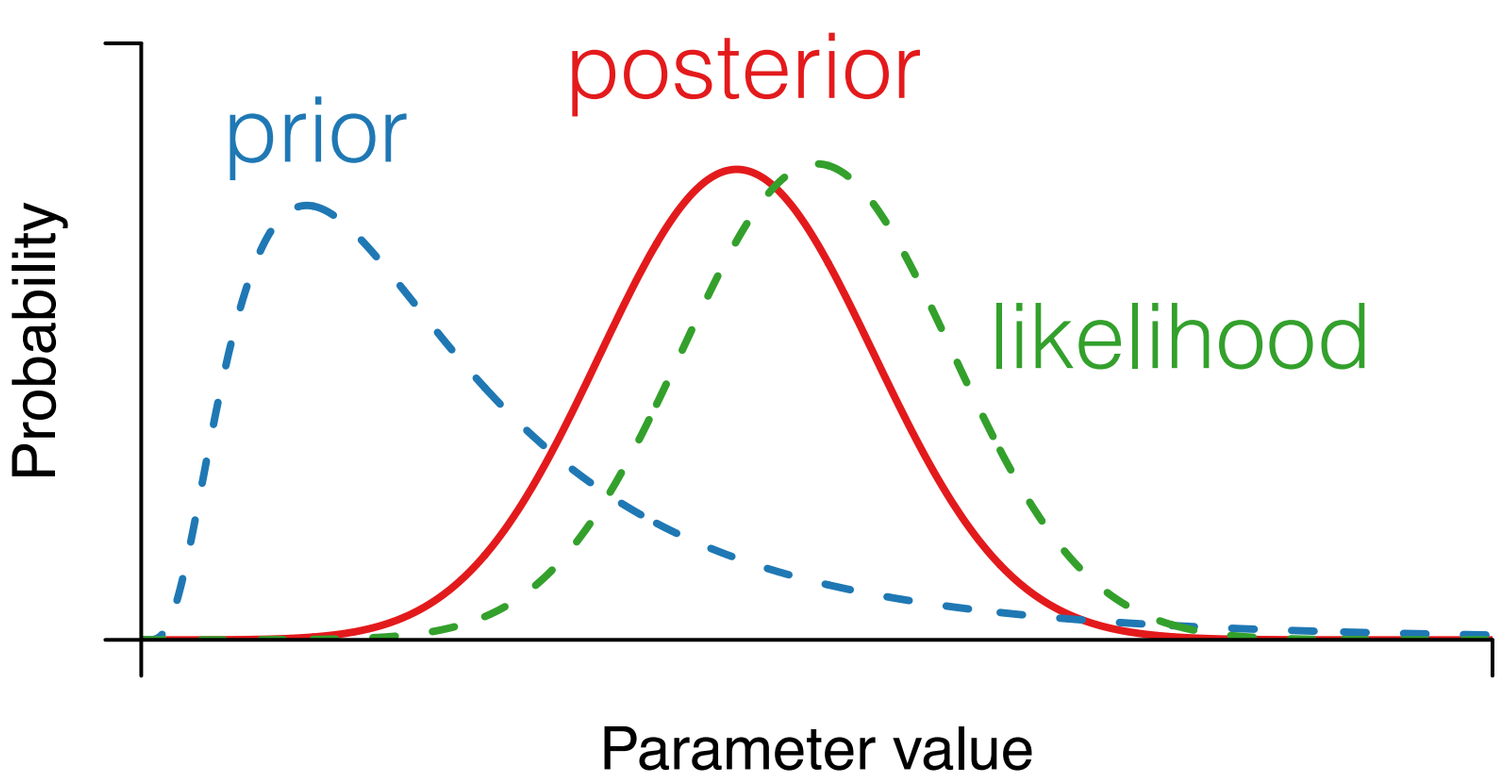
Posterior

Model evidence

Prior

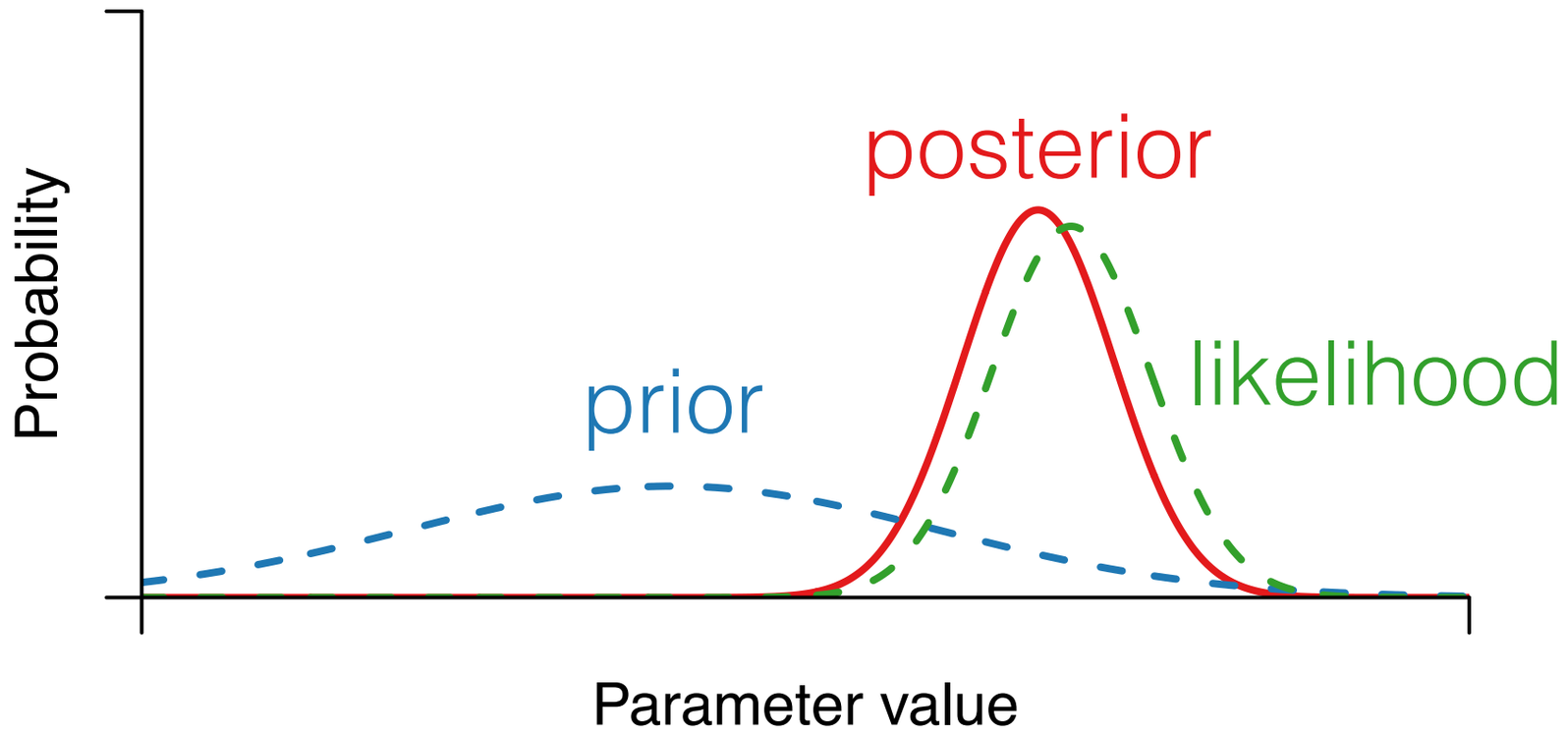
Bayesian inference

Prior distribution is combined (updated) with the likelihood to yield the **posterior distribution**



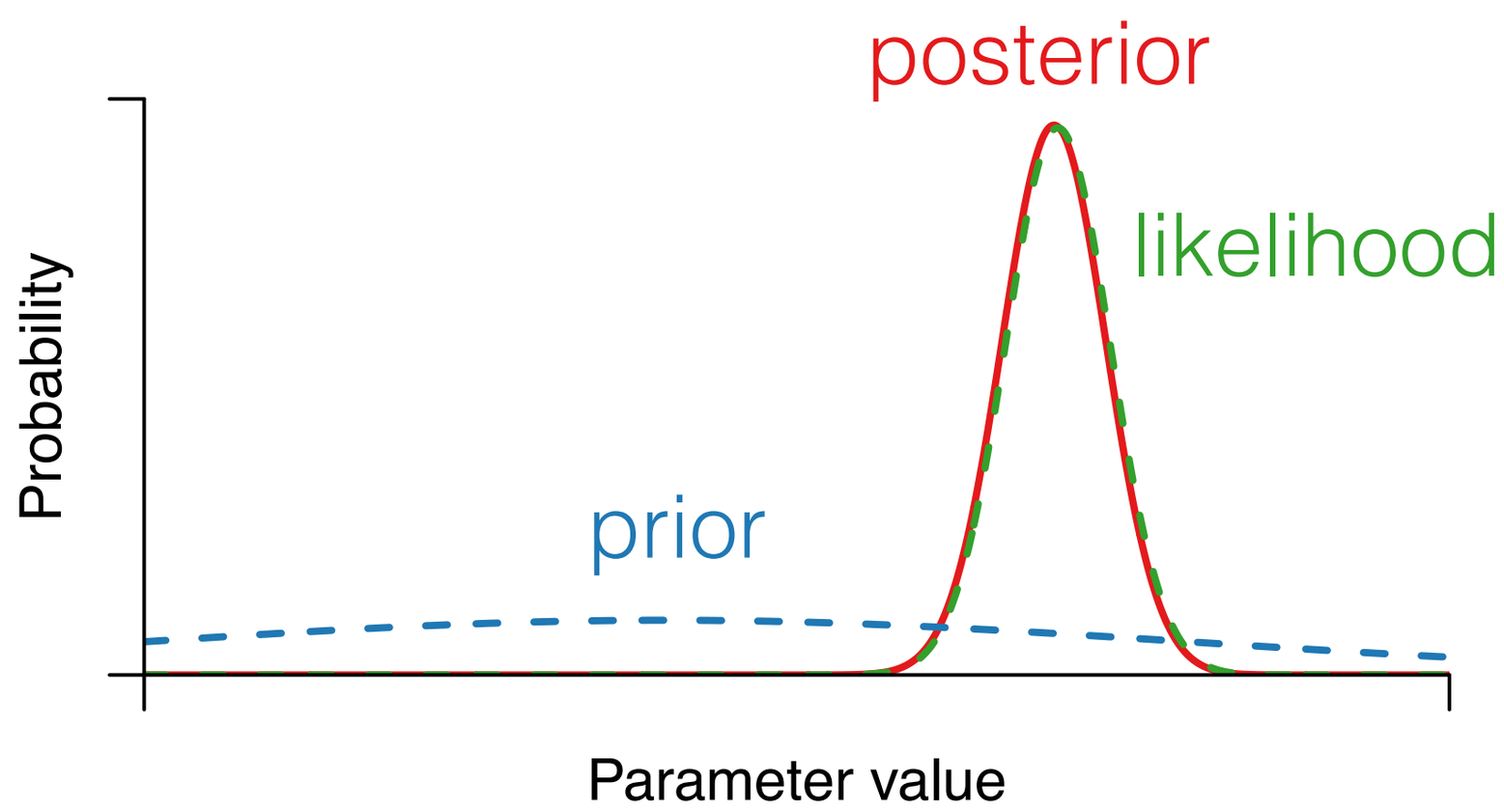
Bayesian inference

The more informative the data, the less effect the prior has



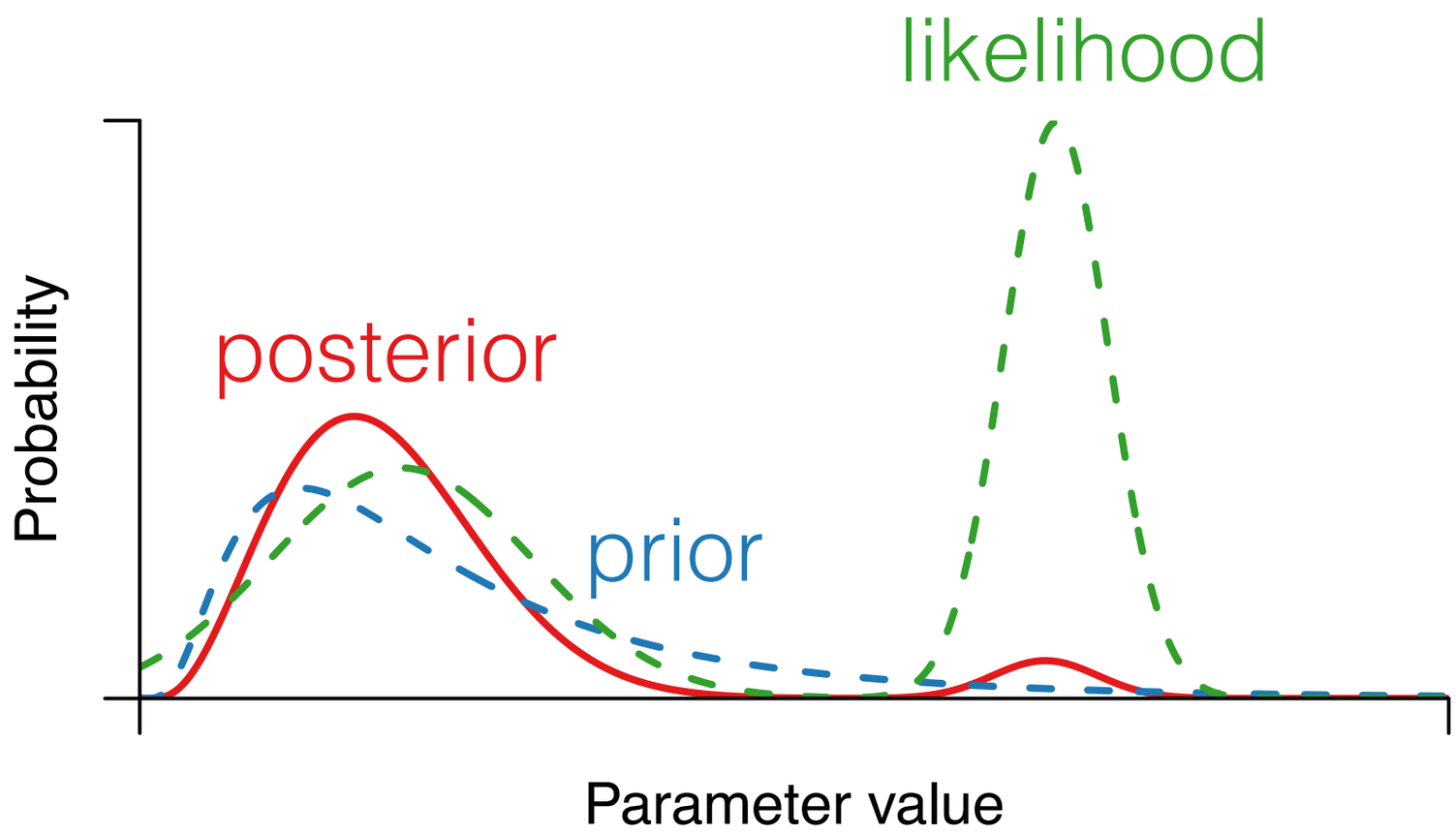
Bayesian inference

If the prior is very uninformative and the likelihood very informative the posterior will be almost exactly the likelihood



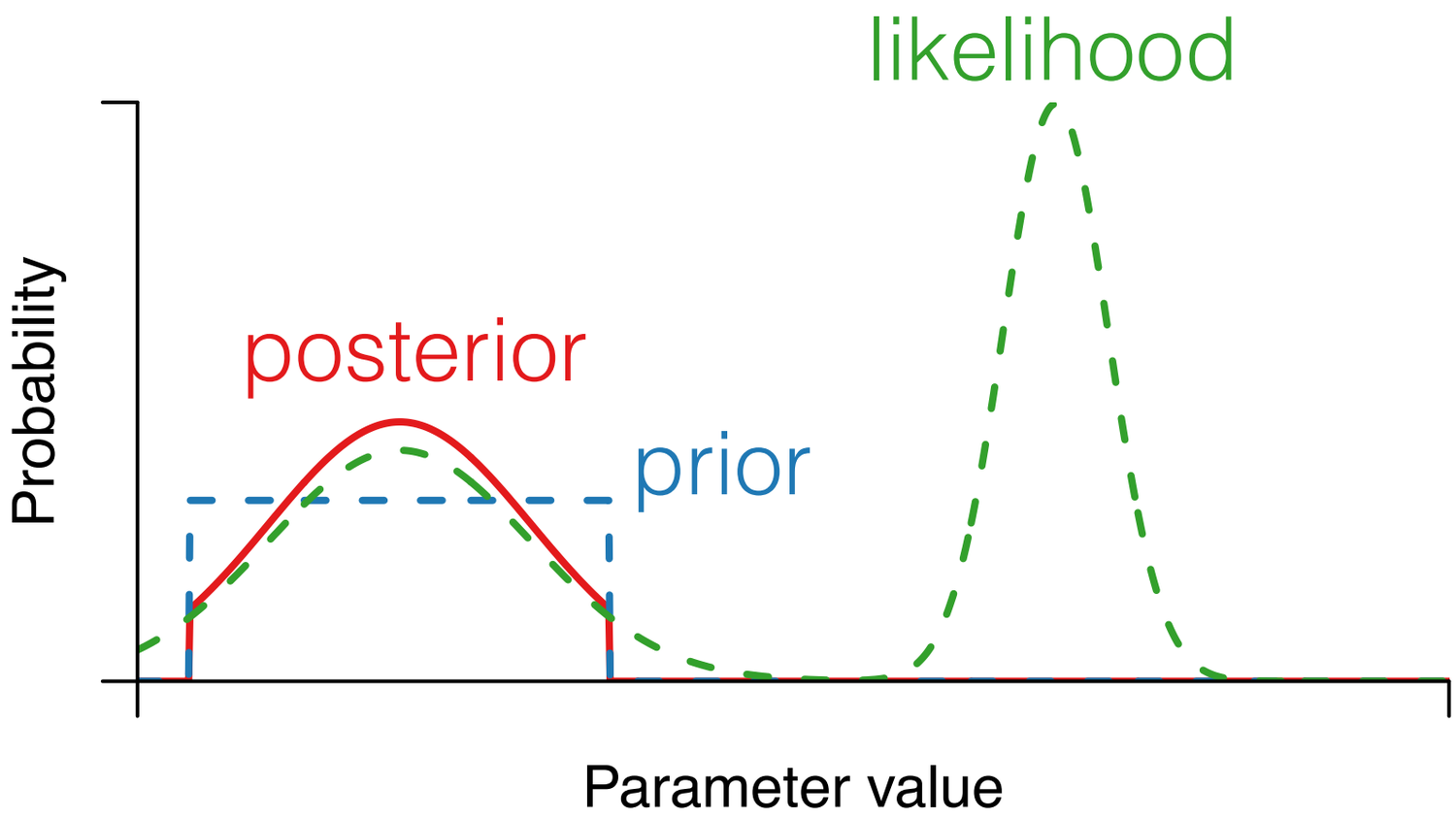
Bayesian inference

If the likelihood is multimodal the posterior is usually also multimodal even when the prior is informative

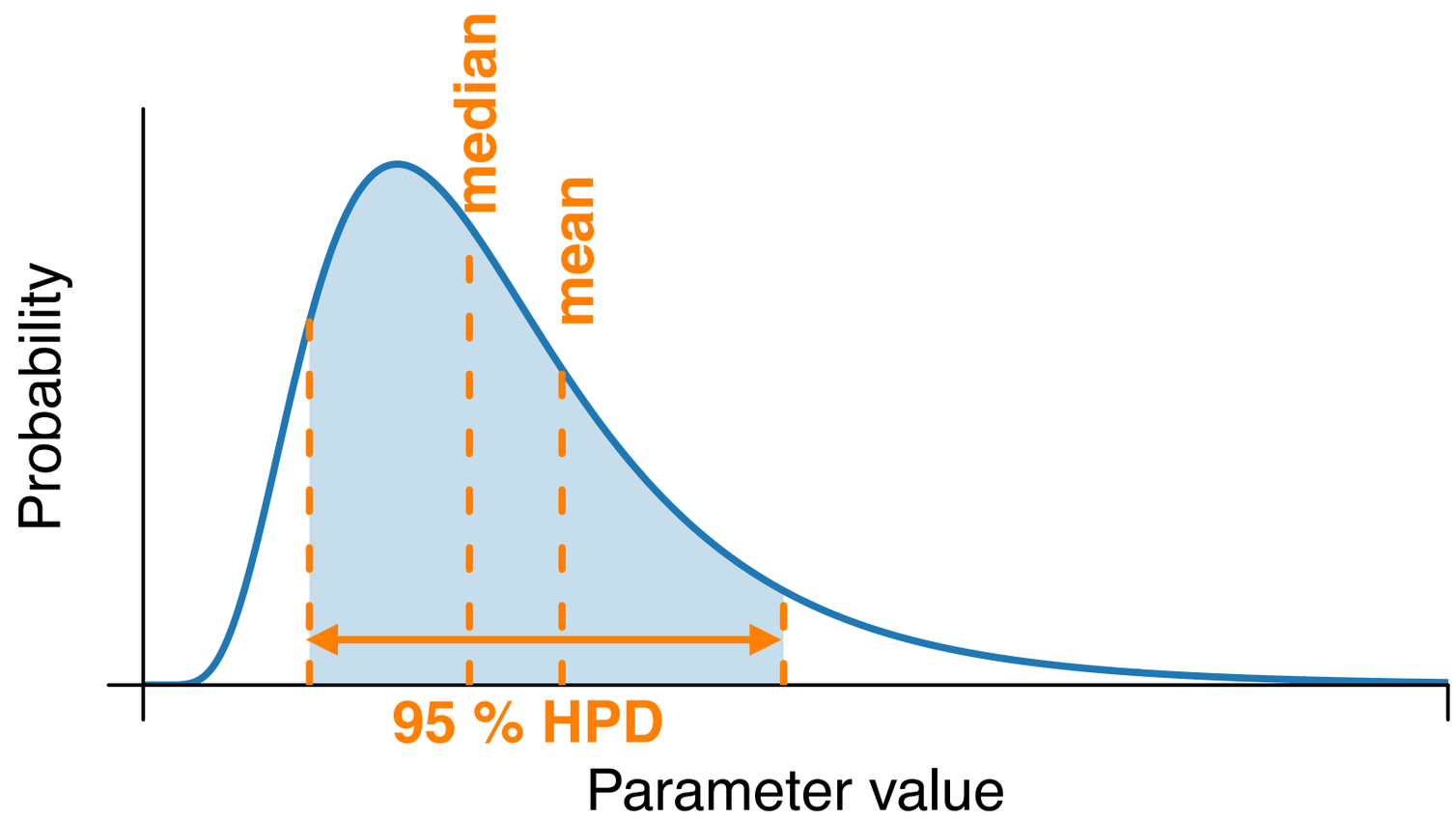


Bayesian inference

Can use the prior to restrict the posterior to specific ranges, e.g. biologically realistic values



Summarising posteriors



95% HPD interval is the smallest interval that contains 95% of the posterior probability

Bias and precision

Biased and **imprecise**

Biased and **precise**

Unbiased and **imprecise**

Unbiased and **precise**

https://manoa.hawaii.edu/exploringourfluidearth/physical/world-ocean/map-distortion/practices-science-precision-vs-accuracy

$$P(\textbf{hypothesis} \mid \textbf{data}) = \frac{P(\textbf{data} \mid \textbf{hypothesis})P(\textbf{hypothesis})}{P(\textbf{data})}$$

Prior → P(hypothesis)

- Original probability for the model parameters/components
- Belief in our hypothesis
- All parameters have priors, whether you specify them or not!

Likelihood → P(data I hypothesis)

- Probability of data given parameters (defined by model)

Posterior → P(hypothesis I data)

- Updated probability for the model parameters in light of the data

Model evidence → P(data) = P(data I model)

- Probability for data given model (any combination of parameters)
- Not generally used for parameter inference
- Used for Bayesian model selection

cEvo

VIROIN

Understanding
virus-host interactions

BEAST2 introduction

Louis du Plessis

ETH zürich

D BSSE

On the program for today

(1) What is phylodynamics?

(2) Bayesian inference recap

(3) BEAST2 introduction

Tutorial: Molecular clock dating (part i)

(4) Molecular clock models

Tutorial: Molecular clock dating (part ii)

(5) Setting priors

Tutorial: Phylodynamics (part i)

(6) Tree priors

Tutorial: Phylodynamics (part ii)

beast **noun**

\bēst\

Definition of *beast*

1. any nonhuman animal, especially a large, four-footed mammal
2. a contemptible person
3. something formidably difficult to control or deal with
4. **the beast**, the Antichrist. Rev. 13:18

beast2 noun

\bēst-tōō\

Definition of *beast2*

1. **B**ayesian **E**volutionary **A**nalysis by **S**ampling **T**rees **2**
2. a cross-platform package for performing Bayesian inference using MCMC

beast2 noun

\bēst-tōō\

Definition of *beast2*

1. **B**ayesian **E**volutionary **A**nalysis by **S**ampling **T**rees **2**
2. a cross-platform package for performing Bayesian inference using MCMC *with emphasis on phylogenetic analysis of molecular sequences*

beast2 noun

\bēst-tōō\

Definition of *beast2*

1. **B**ayesian **E**volutionary **A**nalysis by **S**ampling **T**rees **2**
2. a *modular, extensible,* cross-platform package for performing Bayesian inference using MCMC *with emphasis on phylogenetic analysis of molecular sequences*

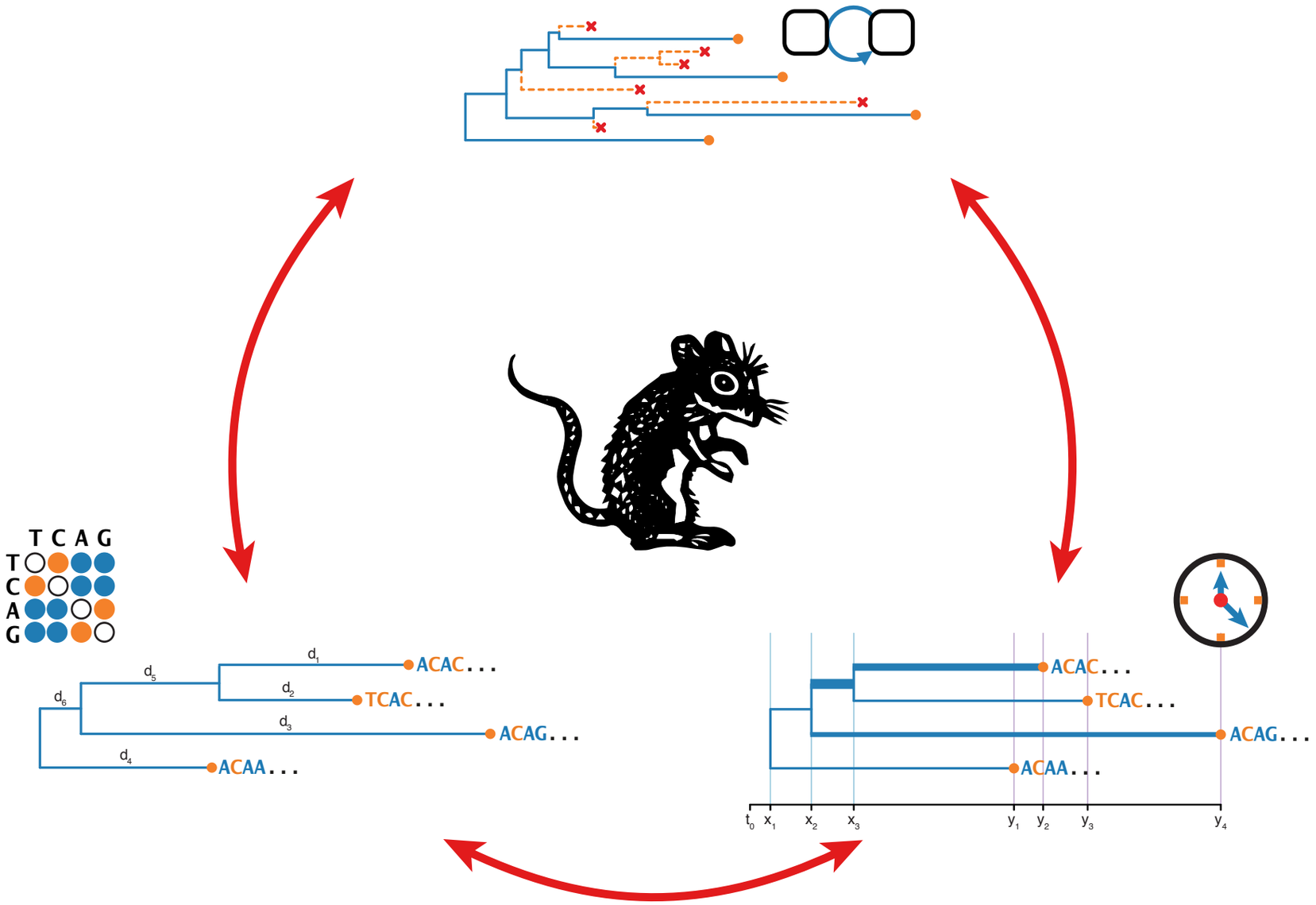
beast2 noun

\bēst-tōō\

Definition of *beast2*

1. **B**ayesian **E**volutionary **A**nalysis by **S**ampling **T**rees **2**
2. a *modular, extensible,* cross-platform package for performing Bayesian inference using MCMC *with emphasis on phylogenetic analysis of molecular sequences*
3. something formidably difficult to control or deal with

What goes into a **BEAST** analysis?



ACAC...
TCAC...
ACAG...
genetic sequences

genealogy

demographic model

site model

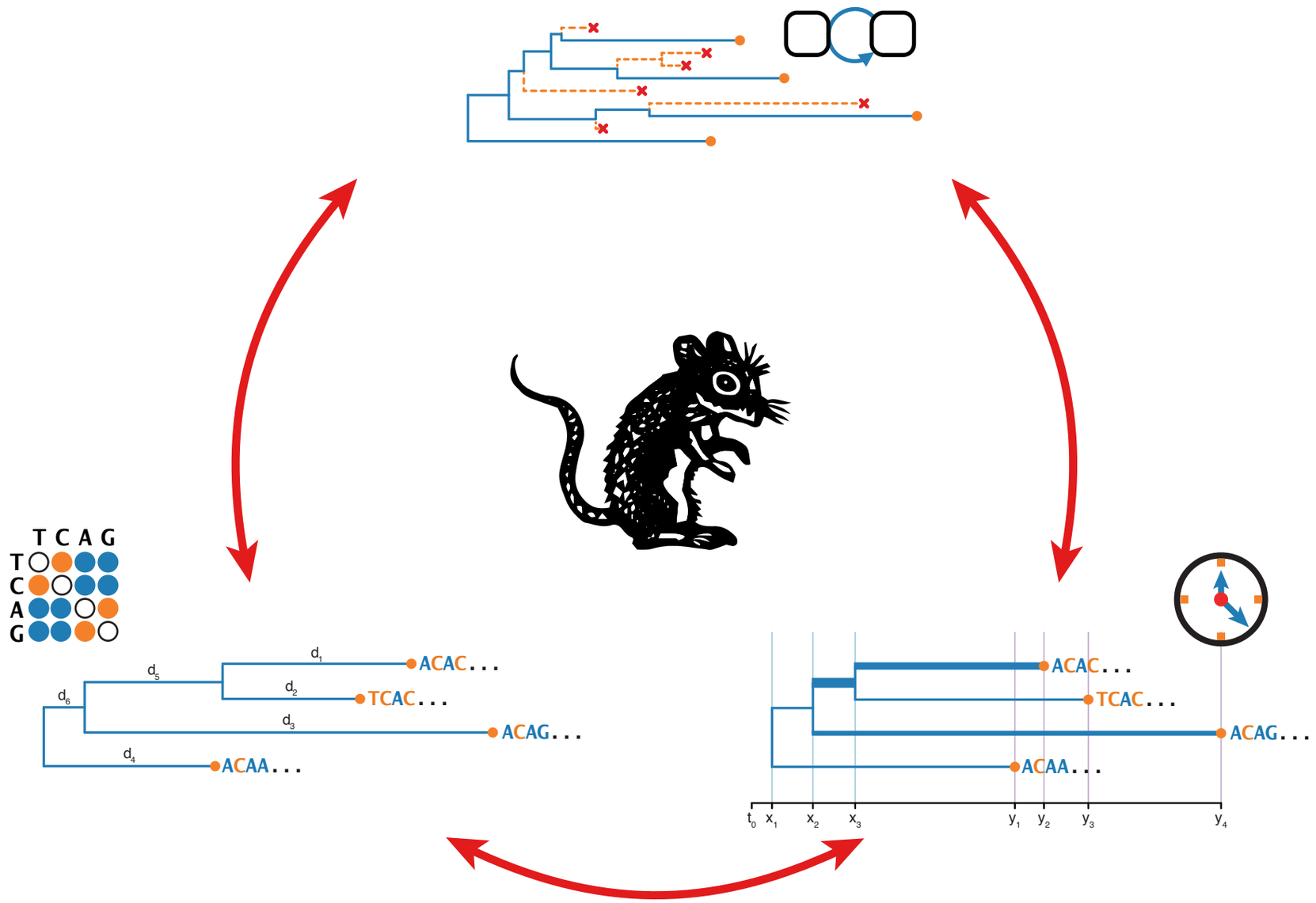
molecular clock model

ACAC...
TCAC...
ACAG... The data

- Samples drawn from a realisation of some stochastic process
- Assume that the data are correct

- Typically one or more alignments of genetic sequencing data (DNA, RNA, amino acids, codons)
- Sampled at one or many time points
- May also contain sampling location or phenotypic trait data

What goes into a **BEAST** model?



ACAC...
TCAC...
ACAG...
genetic sequences

genealogy

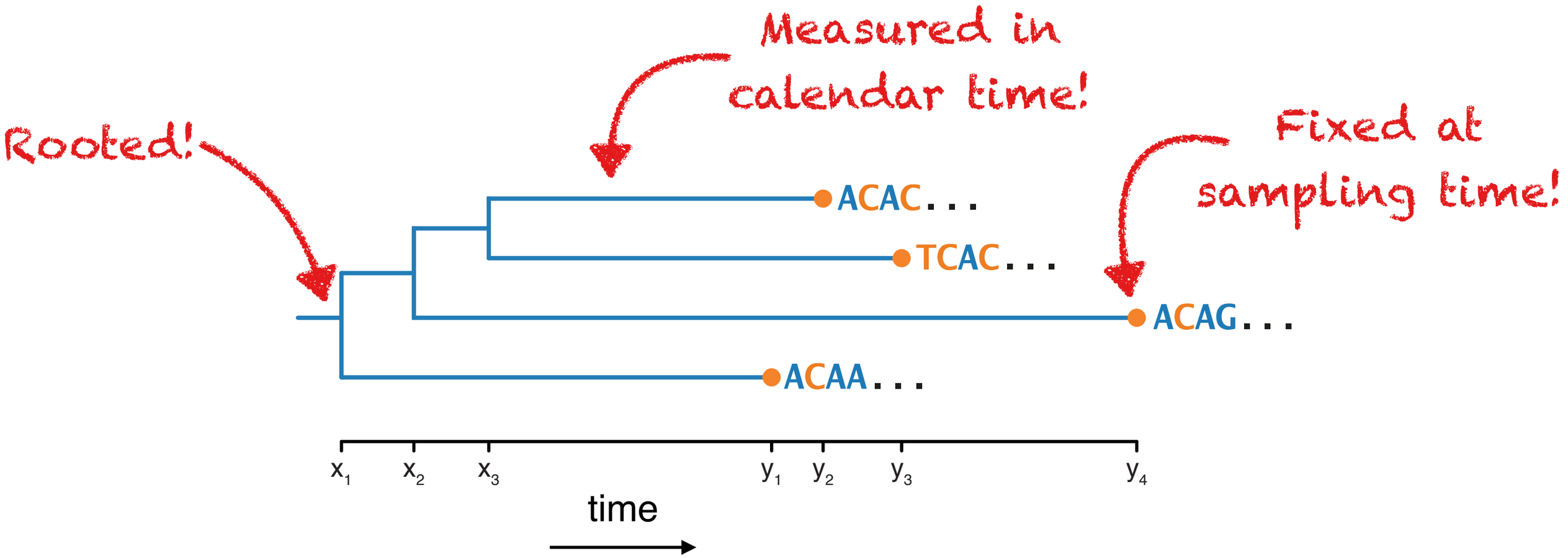
demographic model

site model

molecular clock model

The genealogy (tree)

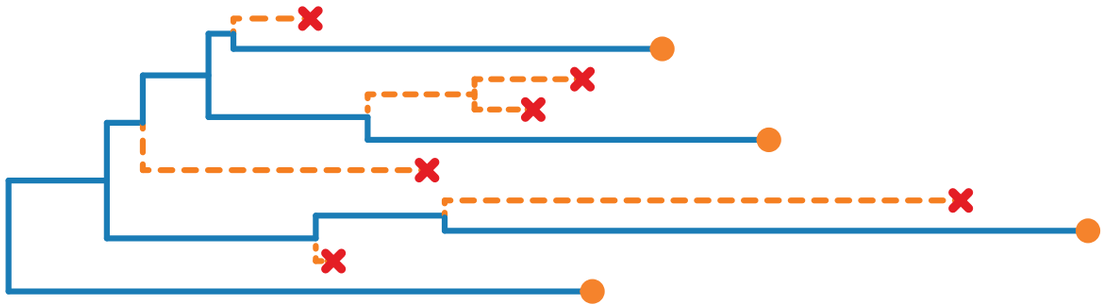
The fundamental genealogical structure
in **BEAST2** is the **rooted time-tree**

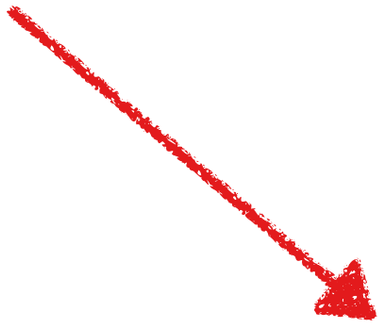


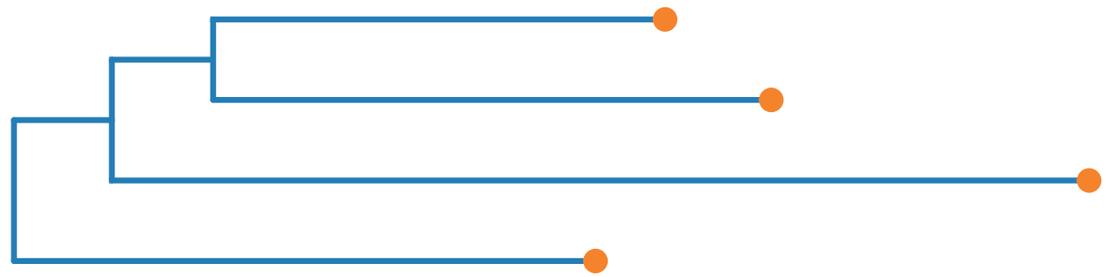
- This tree is a "sampled" or "reconstructed" tree
- Displays ancestral relationships between **sampled sequences** (individuals/taxa)

The genealogy (tree)

full tree





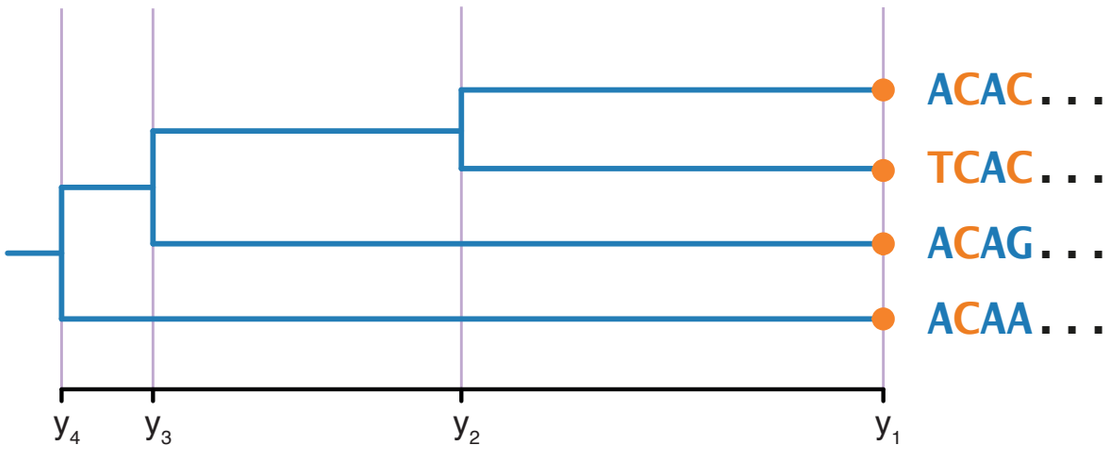


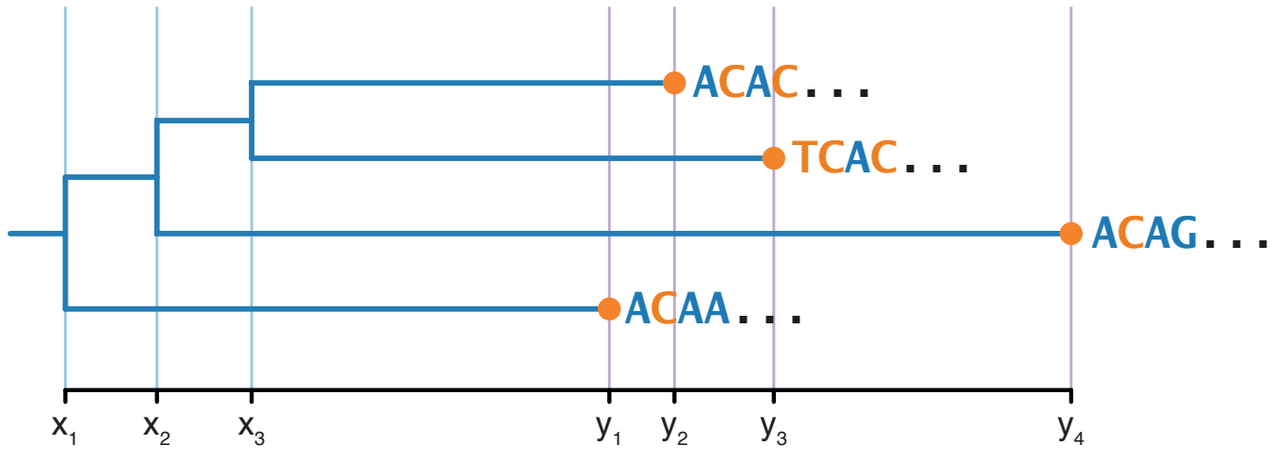
sampled/reconstructed tree

The genealogy (tree)

Sequences sampled
at one time point

Sequences sampled
at many time points

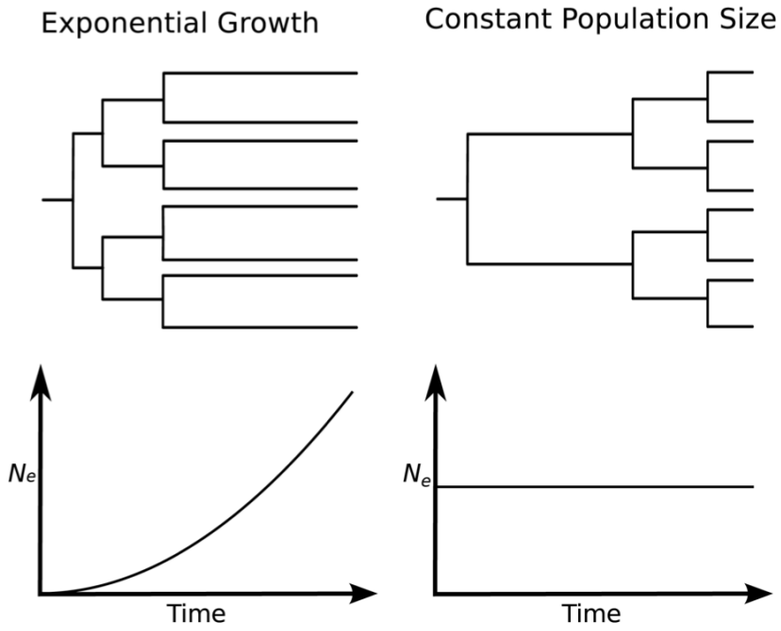


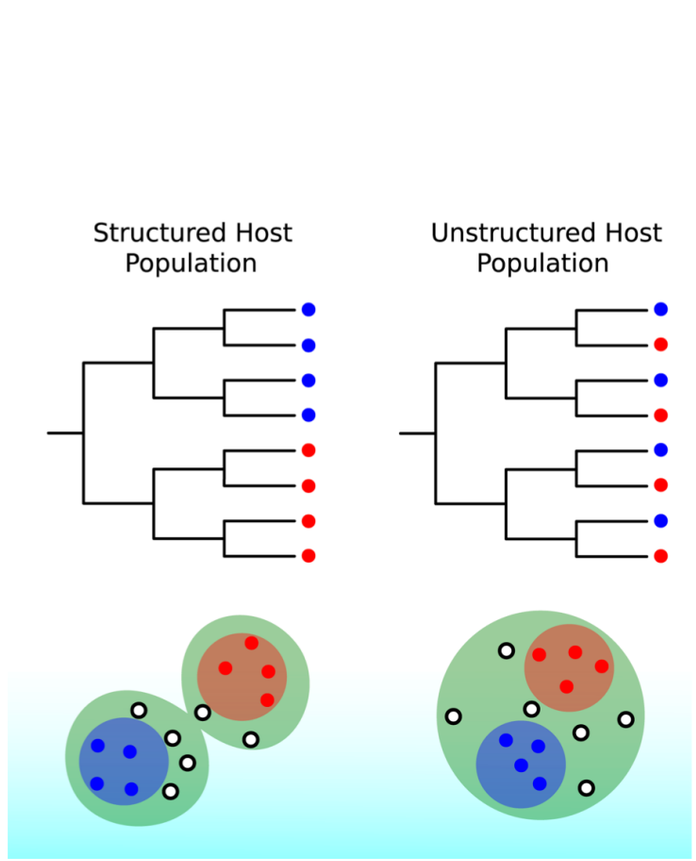


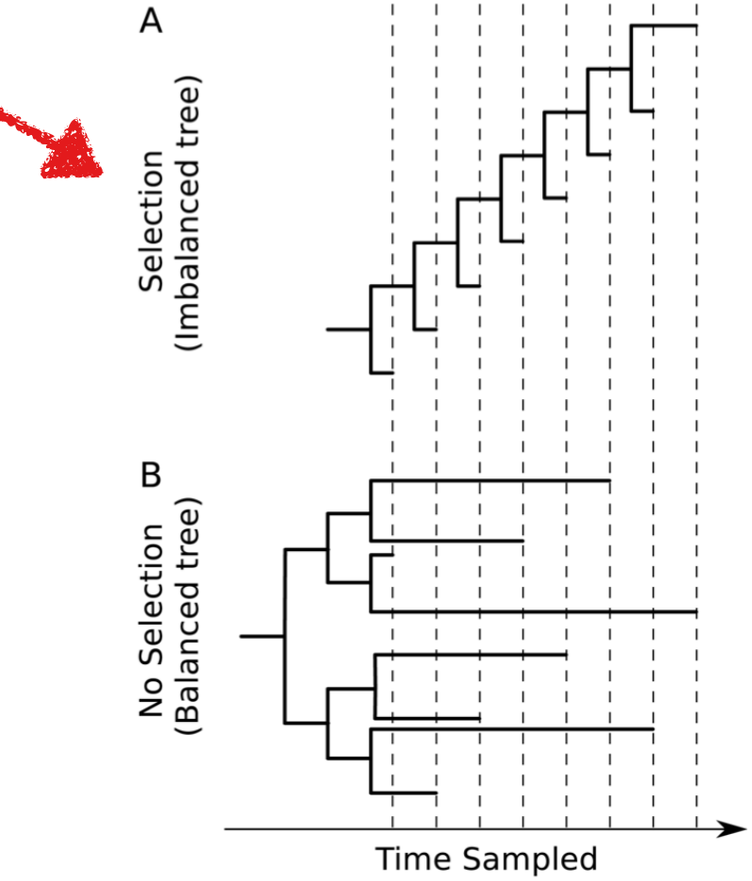
homochronous tree

heterochronous tree

Different population dynamics generate trees that look different



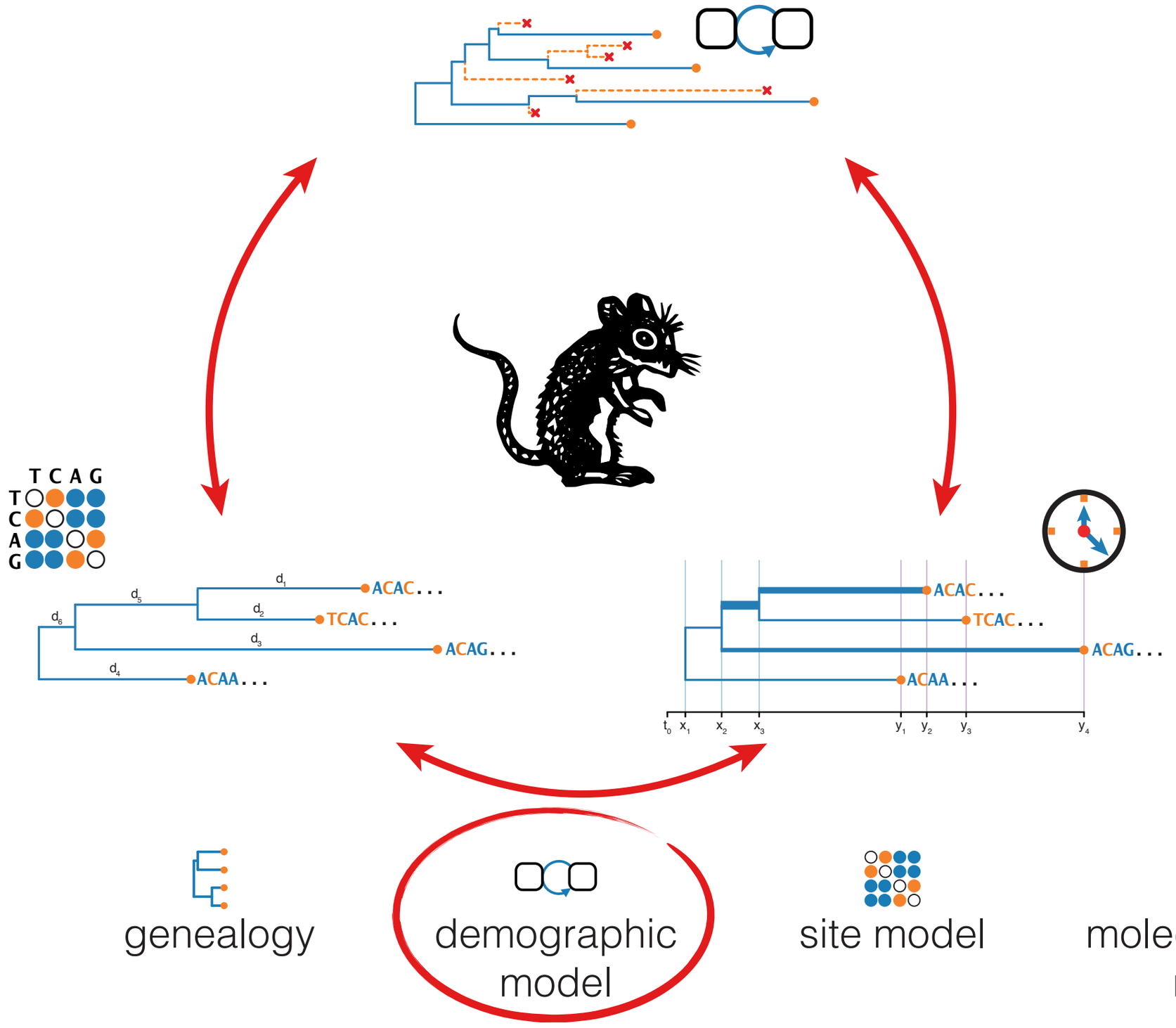




Volz *et al.* **PLoS Comp Biol 2013**

Grenfell *et al.* **Science 2004**

What goes into a **BEAST** model?



ACAC...
TCAC...
ACAG...
genetic sequences

genealogy

demographic model

site model

molecular clock model

Demographic model

- Describes the population/speciation dynamics
- How does the population demographics / species diversity change over time?

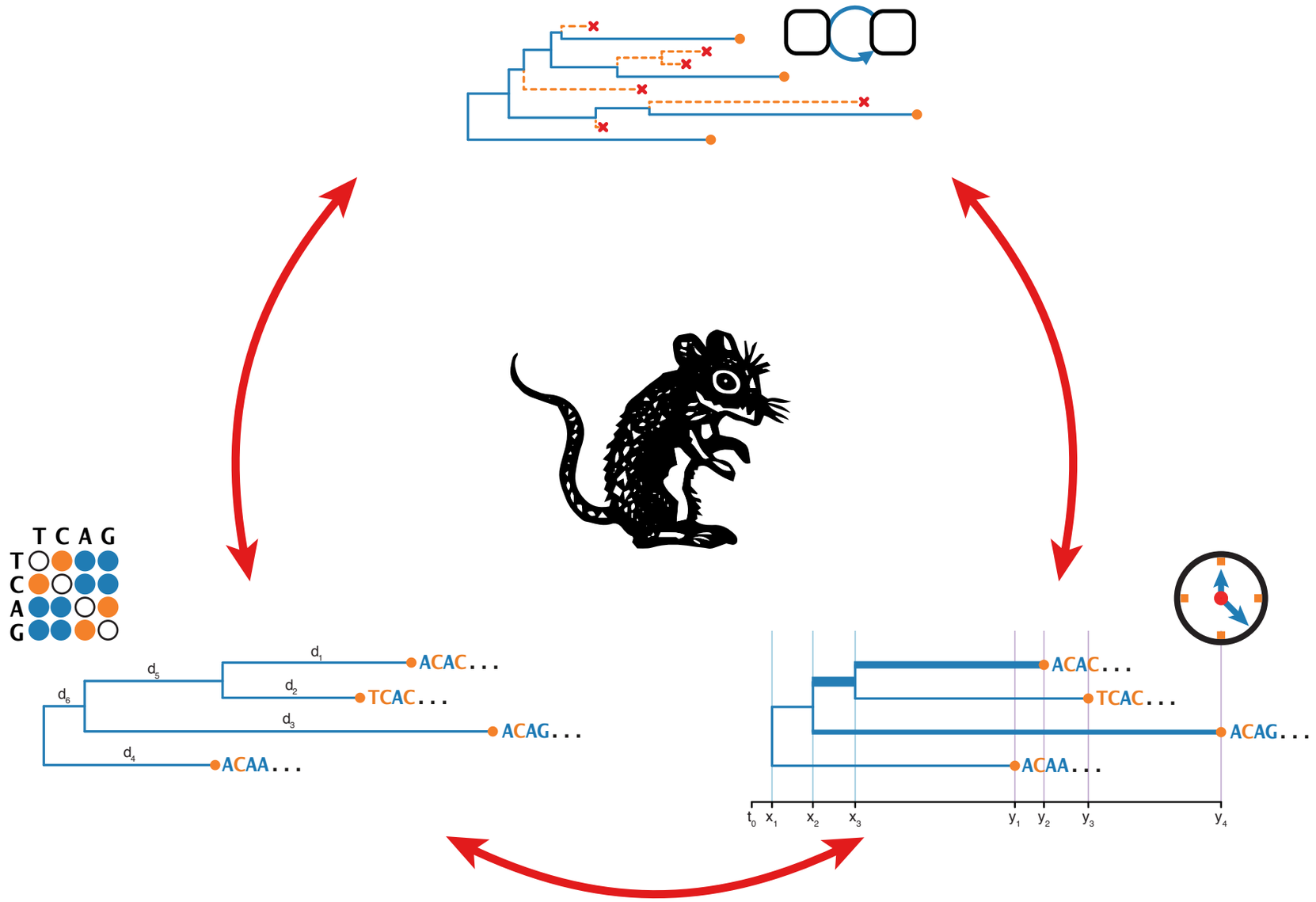
Tree
Realisation of a stochastic process

$P(\text{tree} \mid \text{demographic model})$

Demographic model
Describes the population dynamics (growth of the tree)

- How likely is the genealogy given a demographic model?
- Sometimes called a **tree prior**
- Usually a **coalescent** or **birth-death** model

What goes into a **BEAST** model?



ACAC...
TCAC...
ACAG...
genetic sequences

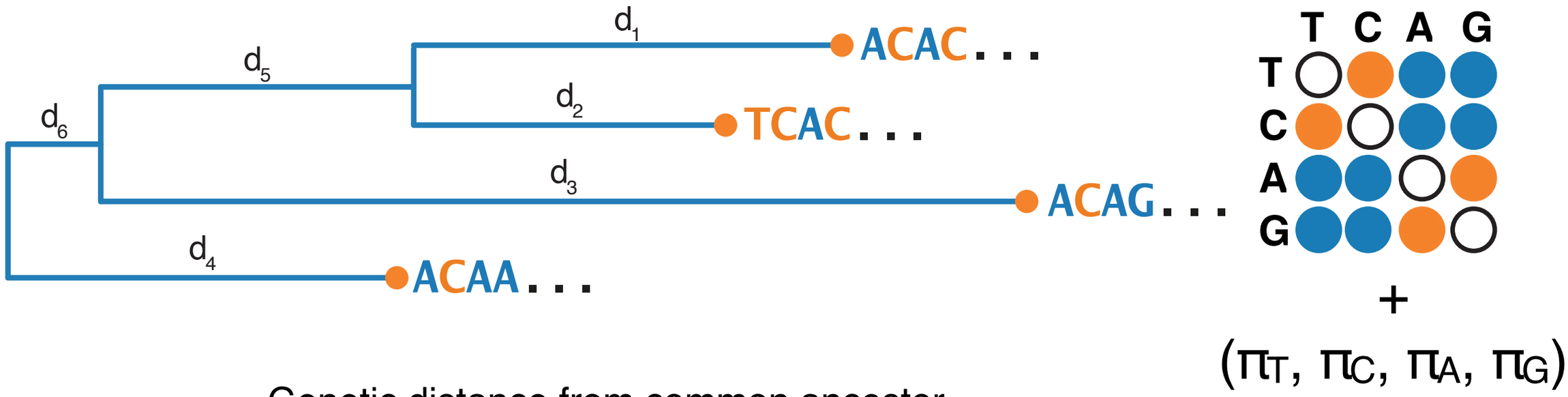
genealogy

demographic model

site model

molecular clock model

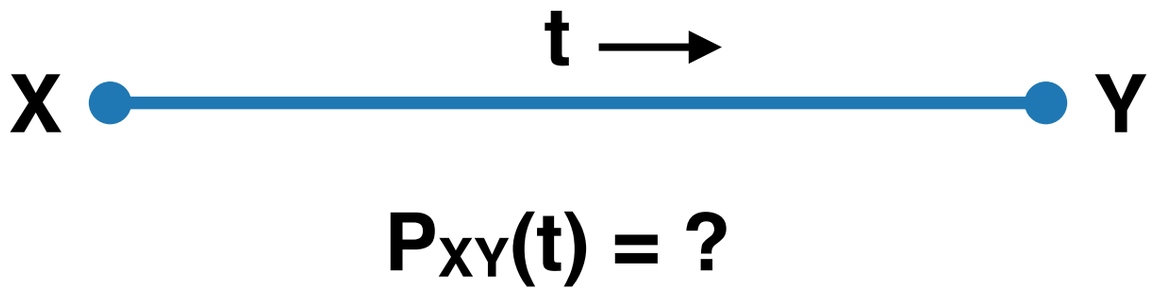
Site model



Genetic distance from common ancestor

- We observe sequences at the tips, not at internal nodes
- **Substitution model** describes rates of substitution between available characters relative to genetic distance (expected substitutions/site), as well as equilibrium frequencies of characters
- **Site model** describes how the substitution model varies from site-to-site
- **Site model links sequences to the genealogy**
  → using Felsenstein's pruning algorithm we can calculate the likelihood: $P($ ACAC... / TCAC... / ACAG... $|$  $)$

Substitution model



- What is the probability of observing **Y** at the end of the branch?
- Multiple substitutions at the same site means not all substitutions are observed
- Need to account for **all** possible trajectories from **X** to **Y**

Substitutions as a Markov process

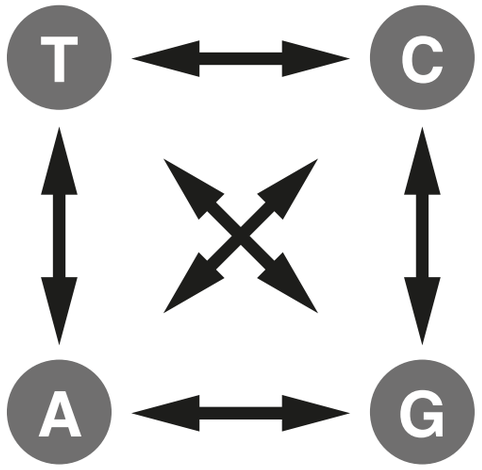
Assume:

- Every site is evolving independently
- Substitutions at each site is governed by a (usually reversible) Markov process

$$\mathbf{Q} = \begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} -(a+b+c) & a & b & c \\ d & -(d+e+f) & e & f \\ g & h & -(g+h+i) & i \\ j & k & l & -(j+k+l) \end{pmatrix} \end{array}$$

Process is governed by a rate matrix (Q) which gives the relative rates of substitutions between nucleotides

Jukes-Cantor model (JC69)



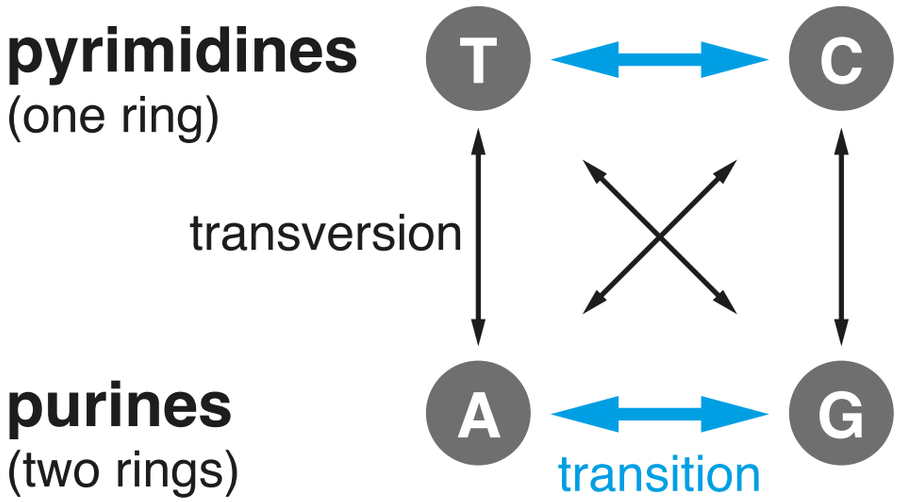
$$\begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} \cdot & \lambda & \lambda & \lambda \\ \lambda & \cdot & \lambda & \lambda \\ \lambda & \lambda & \cdot & \lambda \\ \lambda & \lambda & \lambda & \cdot \end{pmatrix} \end{array}$$

$$\pi_T = \pi_C = \pi_A = \pi_G$$

- Simplest model
- All rates and frequencies are equal!

Kimura 2-parameter model (K80)

(courtesy of Carsten Magnus)

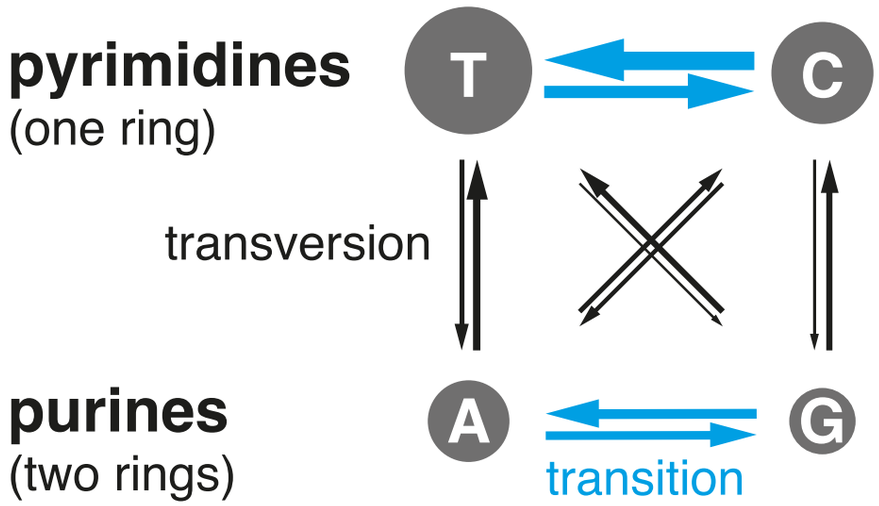


$$\begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} \cdot & \alpha & \beta & \beta \\ \alpha & \cdot & \beta & \beta \\ \beta & \beta & \cdot & \alpha \\ \beta & \beta & \alpha & \cdot \end{pmatrix} \end{array}$$

$$\pi_T = \pi_C = \pi_A = \pi_G$$

- Accounts for transition/transversion bias
- Still symmetric ($r_{ij} = r_{ji}$)
- Equilibrium frequencies still equal
  After a long period of evolution p(T) = p(C) = p(A) = p(G) = 0.25

HKY-model (HKY85)



$$(\pi_T, \pi_C, \pi_A, \pi_G)$$

$$\begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} \cdot & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & \cdot & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & \cdot & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & \cdot \end{pmatrix} \end{array}$$

$$= \begin{pmatrix} \cdot & \alpha & \beta & \beta \\ \alpha & \cdot & \beta & \beta \\ \beta & \beta & \cdot & \alpha \\ \beta & \beta & \alpha & \cdot \end{pmatrix} \cdot \begin{pmatrix} \pi_T & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_A & 0 \\ 0 & 0 & 0 & \pi_G \end{pmatrix}$$

- Accounts for transition/transversion bias
- Accounts for unequal equilibrium frequencies
- Not symmetric anymore ($r_{ij} \neq r_{ji}$)
- Still time-reversible ($\pi_i q_{ij} = \pi_j q_{ji}$)

General time-reversible model (GTR/REV) (courtesy of Carsten Magnus)

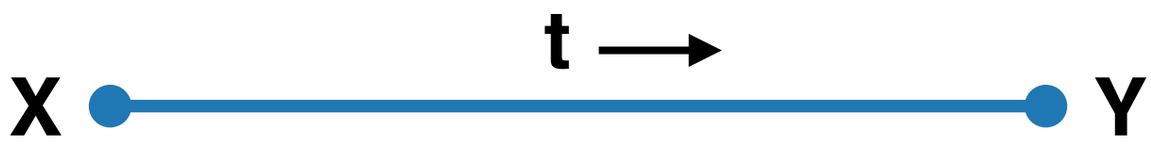
$$
\begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array}
\begin{array}{c}
\begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\
\begin{pmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \cdot \end{pmatrix}
\end{array}
= \begin{pmatrix} \cdot & a & b & c \\ a & \cdot & d & e \\ b & d & \cdot & f \\ c & e & f & \cdot \end{pmatrix} \cdot \begin{pmatrix} \pi_T & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_A & 0 \\ 0 & 0 & 0 & \pi_G \end{pmatrix}
$$

- Most general time-reversible model
- More flexible models are possible, but mathematically inconvenient

Transition probability matrix

$$\mathbf{P}(t) = e^{\mathbf{Q}t} \qquad \mathbf{P}(t) = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} p_{tt}(t) & p_{tc}(t) & p_{ta}(t) & p_{tg}(t) \\ p_{ct}(t) & p_{cc}(t) & p_{ca}(t) & p_{cg}(t) \\ p_{at}(t) & p_{ac}(t) & p_{aa}(t) & p_{ag}(t) \\ p_{gt}(t) & p_{gc}(t) & p_{ga}(t) & p_{gg}(t) \end{pmatrix} \end{array}$$

- Transition probabilities take into account every possible evolutionary trajectory (Chapman-Kolmogorov theorem)

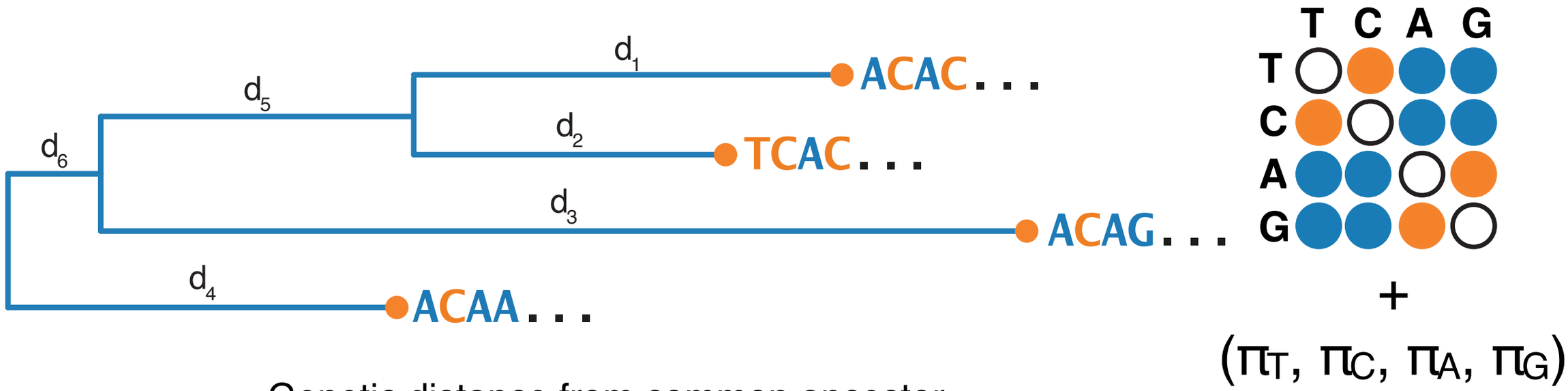


$$P_{XY}(t) = ?$$

- **Q** only gives the **relative substitution rates**

⇒ distance is measured in expected number of substitutions per site

Site model



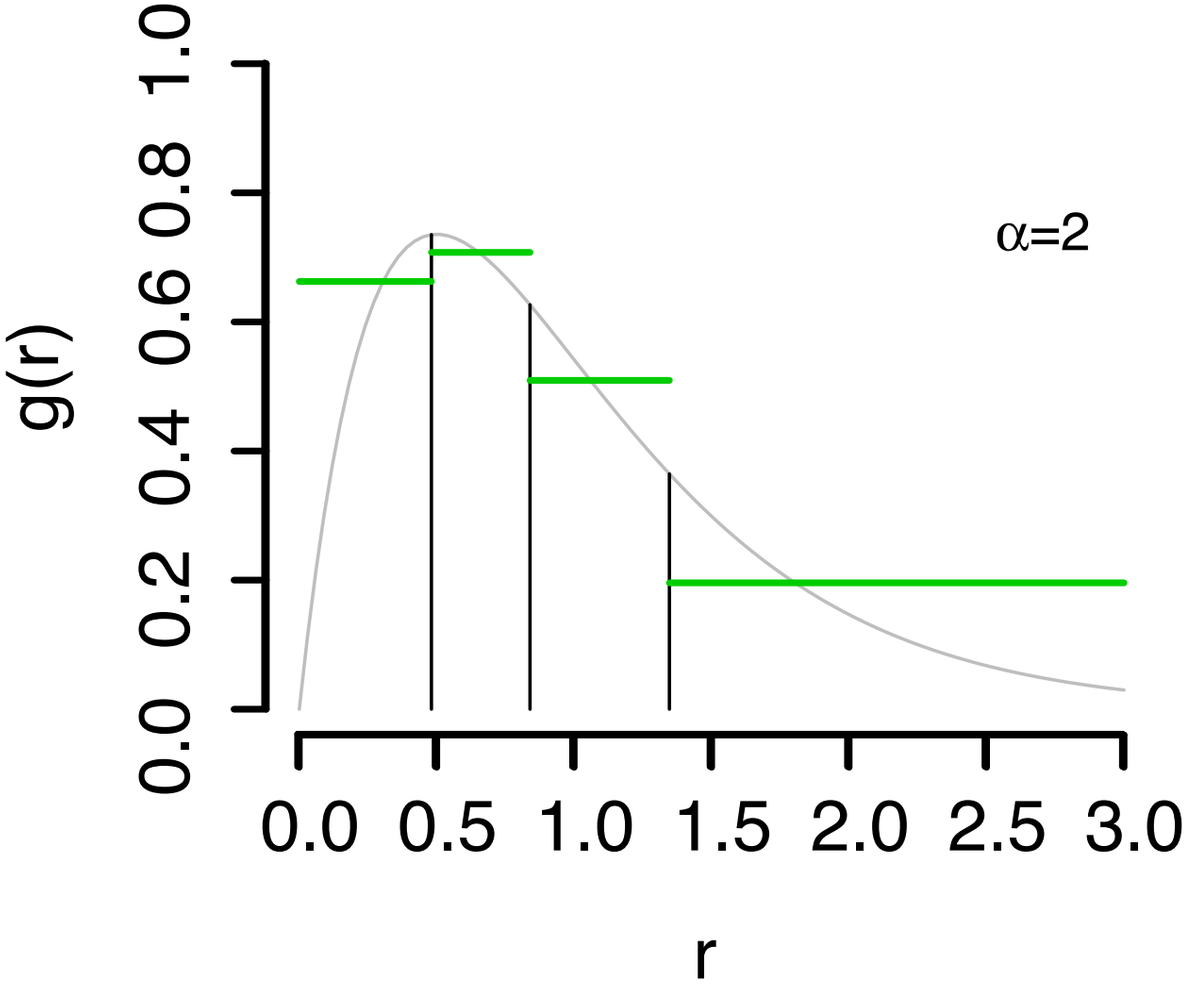
Genetic distance from common ancestor

- Describes how the substitution model varies from site-to-site
- Assume every site is evolving independently
- Account for rate heterogeneity between sites:
  - Proportion of invariant sites
  - Gamma rate heterogeneity
  - Multi-locus models

Gamma rate heterogeneity

(courtesy of Carsten Magnus)

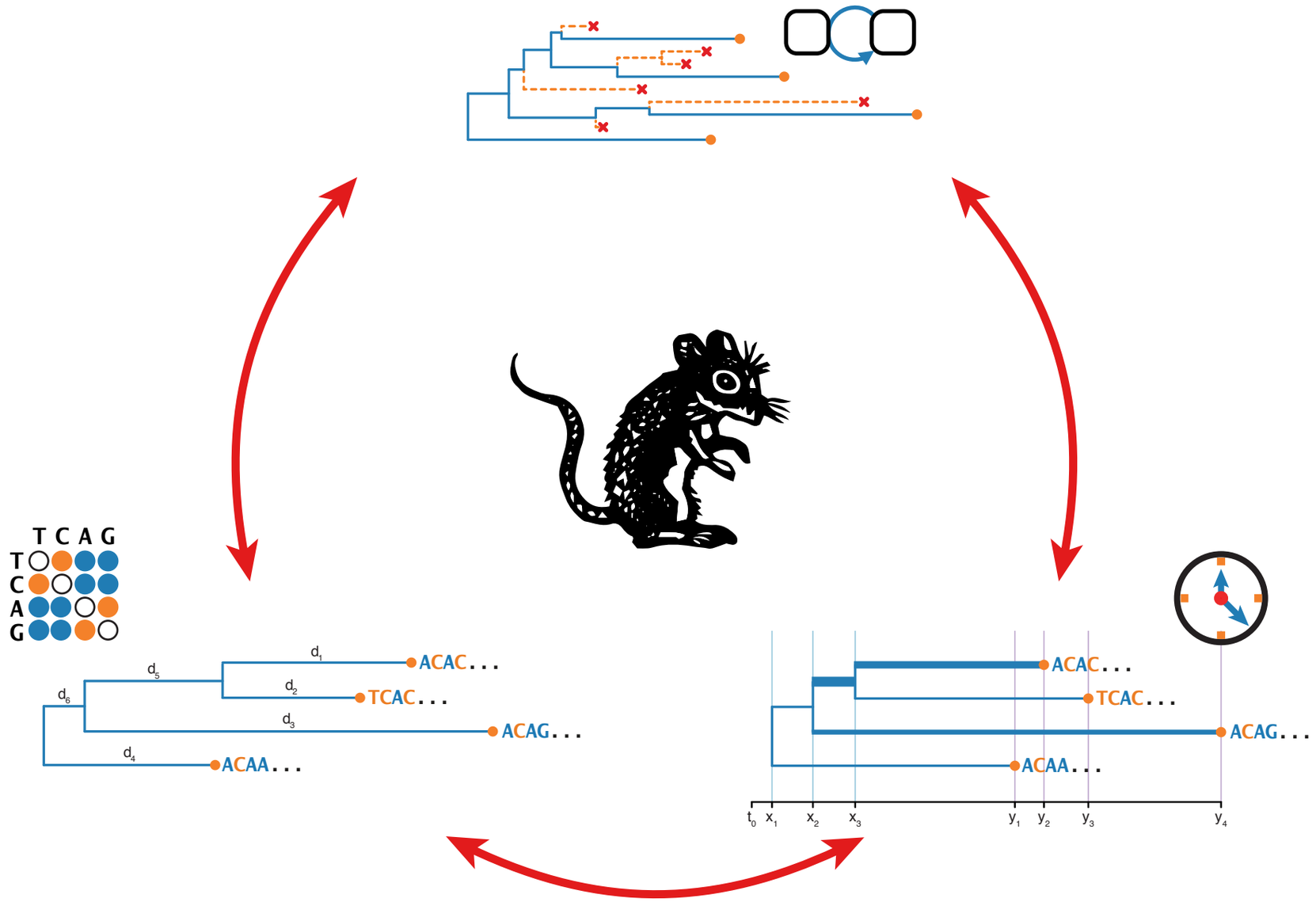
- Not all sites evolve at the same rate
- Assume rate heterogeneity among sites is Γ-distributed
- Discretise Γ-distribution to ***n*** discrete rate categories for computational reasons



Multi-locus models

- Γ-distributed rate variation is not always flexible enough to model differences between different loci
- Use a separate substitution model for each locus
- Can also use separate models for different codon positions
- Loci could be different genes, different codon positions, etc.

What goes into a **BEAST** model?



ACAC...
TCAC...
ACAG...
genetic sequences

genealogy

demographic model

site model

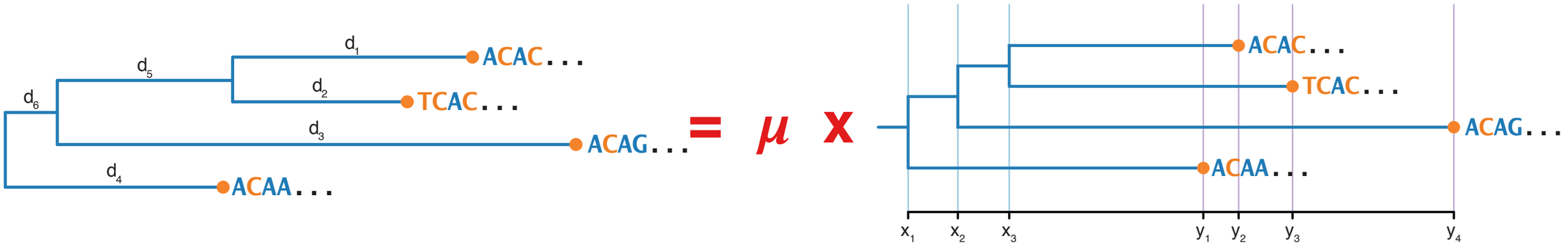
molecular clock model

Molecular clock model

genetic distance tree (subst/site)

clock rate (subst/site/year)

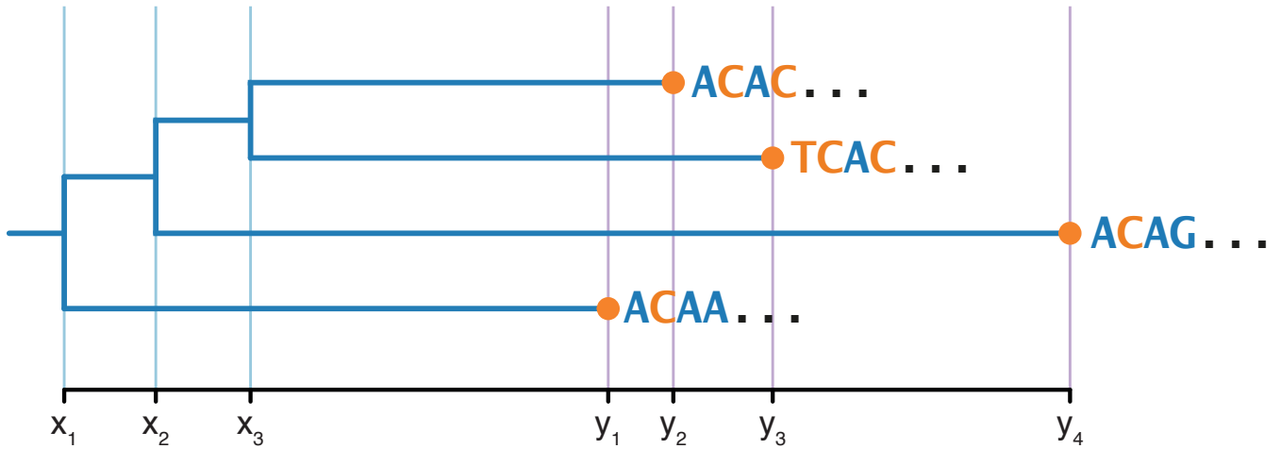
time tree (years)



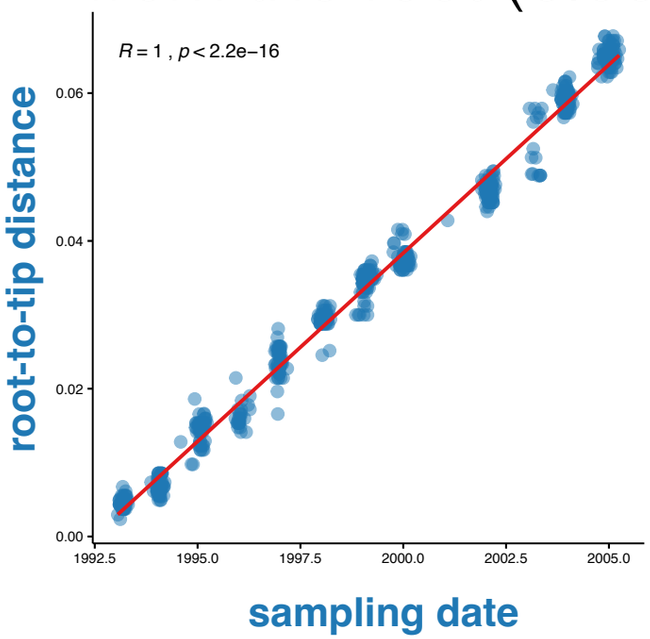
- Determines how quickly sequences are evolving along the tree
- **Genetic distance = Rate x Time**

Molecular clock model

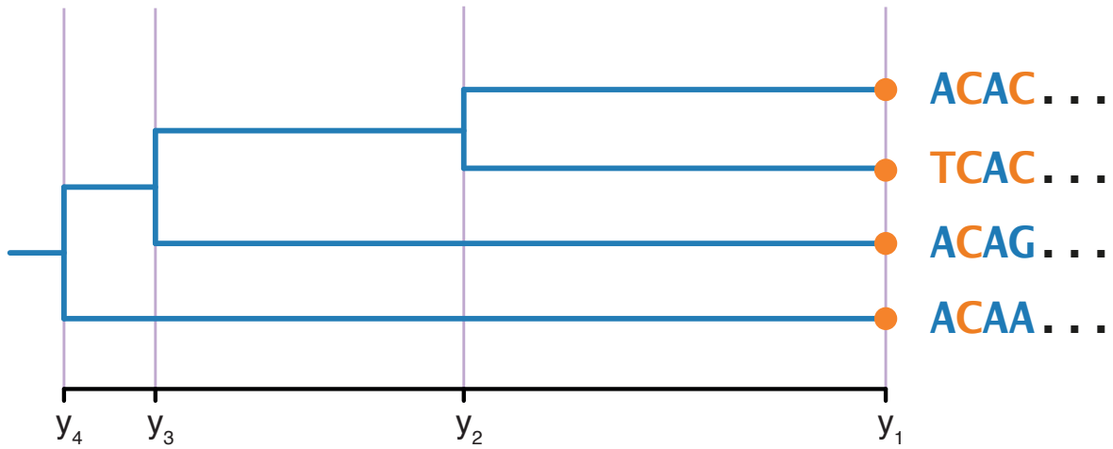
heterochronous tree



- Correlation between sampling time and genetic distance from the root (clock signal)

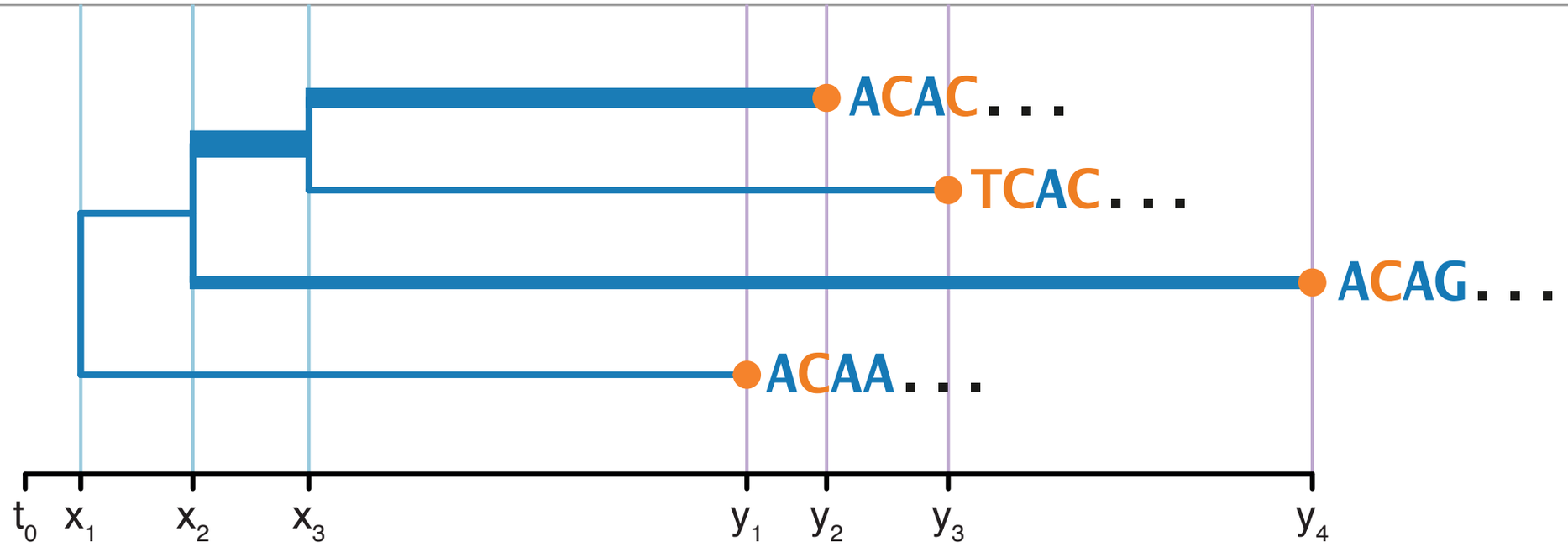


homochronous tree



- Need external information to calibrate the clock
- Fix clock rate or calibrate some internal nodes with external information

Relaxed and local clocks



Some branches may have different clock rates!

Putting it all together

genetic sequences


genealogy

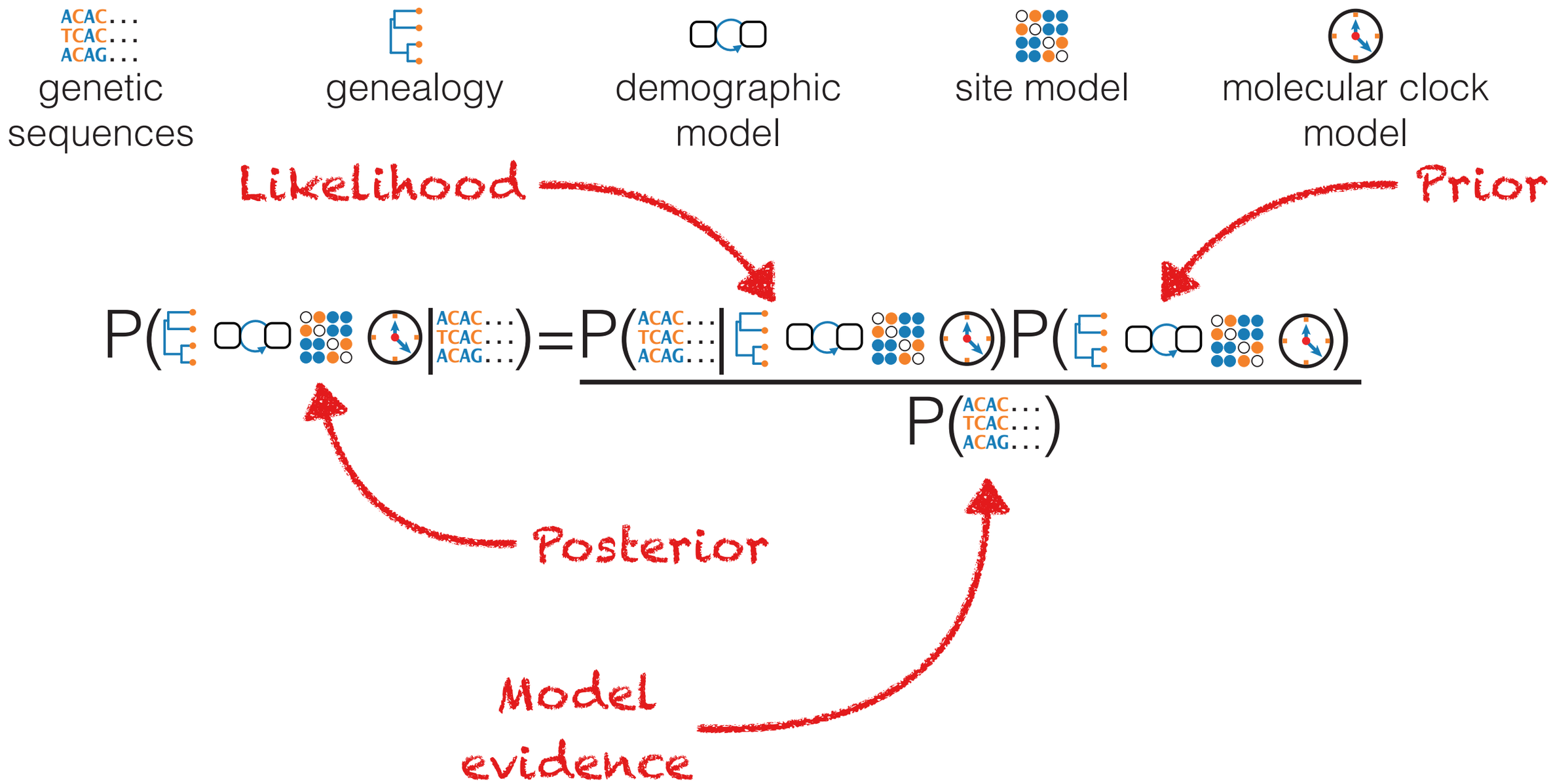

demographic model


site model

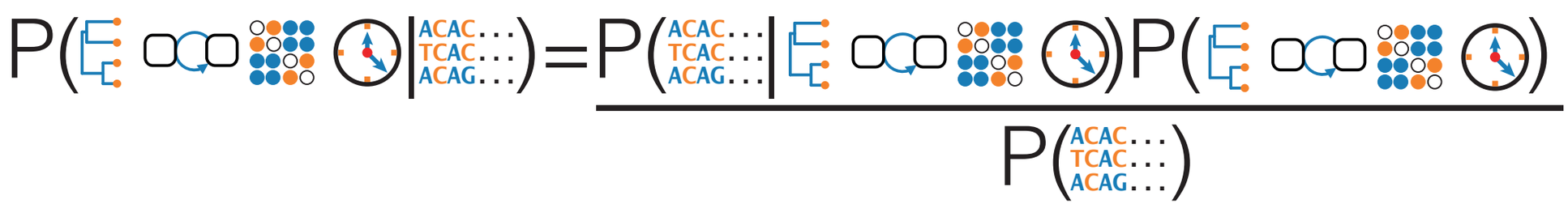

molecular clock model

$$P(\textbf{hypothesis} \mid \textbf{data}) = P(\textbf{data} \mid \textbf{hypothesis})\frac{P(\textbf{hypothesis})}{P(\textbf{data})}$$

Putting it all together

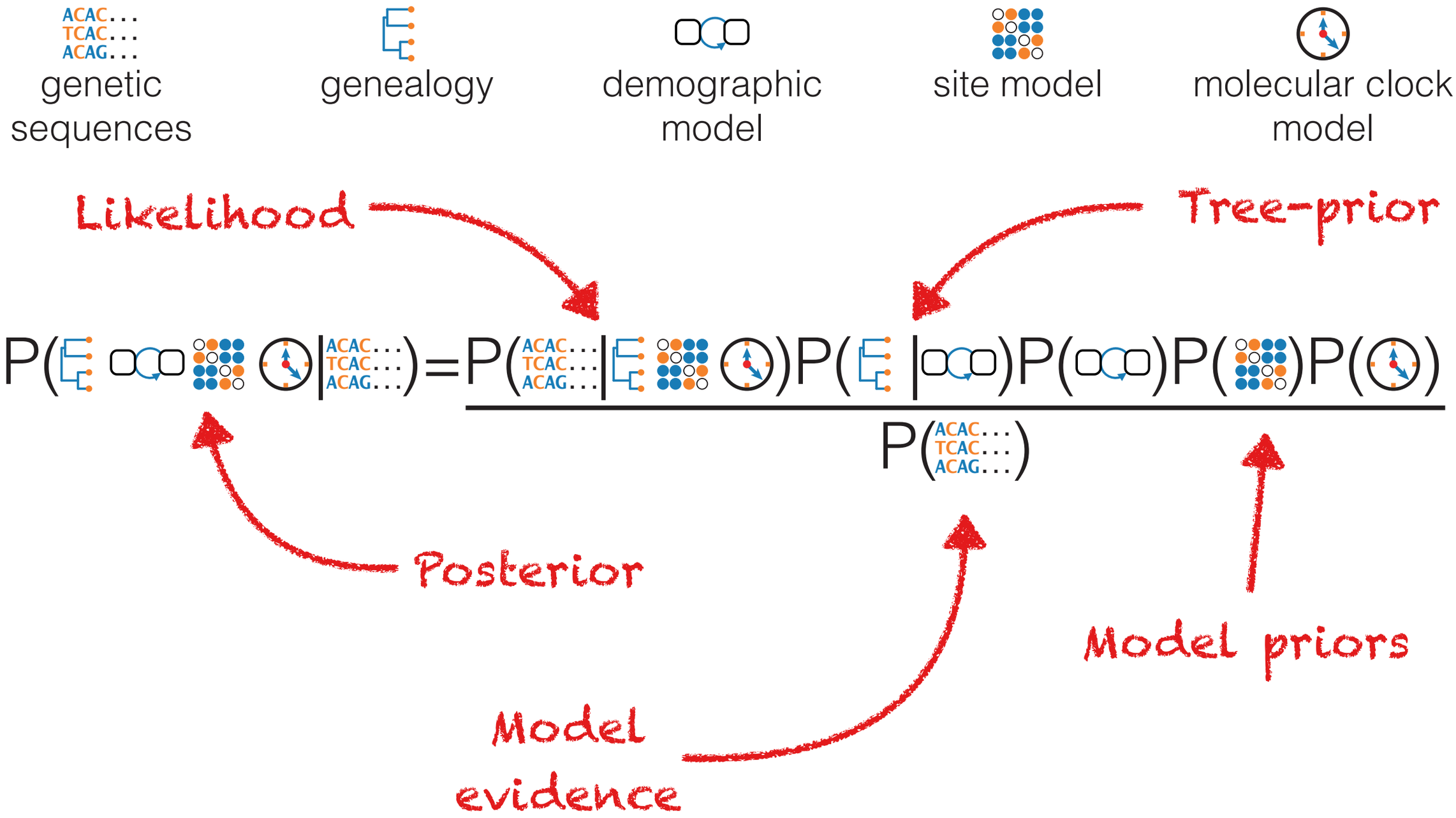


Putting it all together

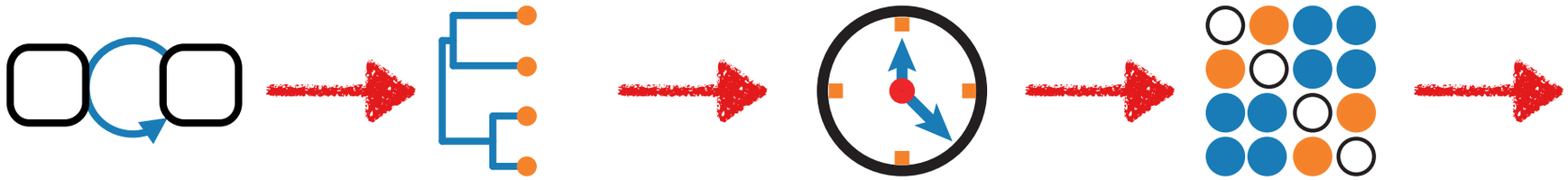


Assume independence

Posterior distribution in BEAST2



Neutrality assumption



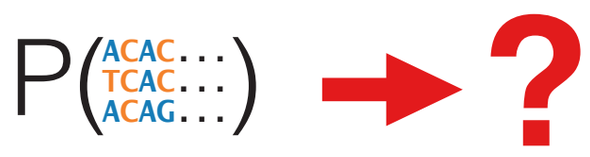
Assumed tree-generating process independent of sequence data

$$P(\text{tree}, \text{model}, \text{clock} \mid \text{ACAC}\ldots, \text{TCAC}\ldots, \text{ACAG}\ldots) = \frac{P(\text{ACAC}\ldots, \text{TCAC}\ldots, \text{ACAG}\ldots \mid \text{tree}, \text{model}, \text{clock})\, P(\text{tree} \mid \text{model})\, P(\text{model})\, P(\text{clock})}{P(\text{ACAC}\ldots, \text{TCAC}\ldots, \text{ACAG}\ldots)}$$

How can we find the posterior?

- We want to calculate the posterior distribution

- But we cannot easily calculate the marginal likelihood (model evidence)



→ use **MCMC**! (Markov-chain Monte Carlo)

- MCMC is a stochastic algorithm that performs a random walk on the posterior, preferentially sampling high-density areas

MCMC

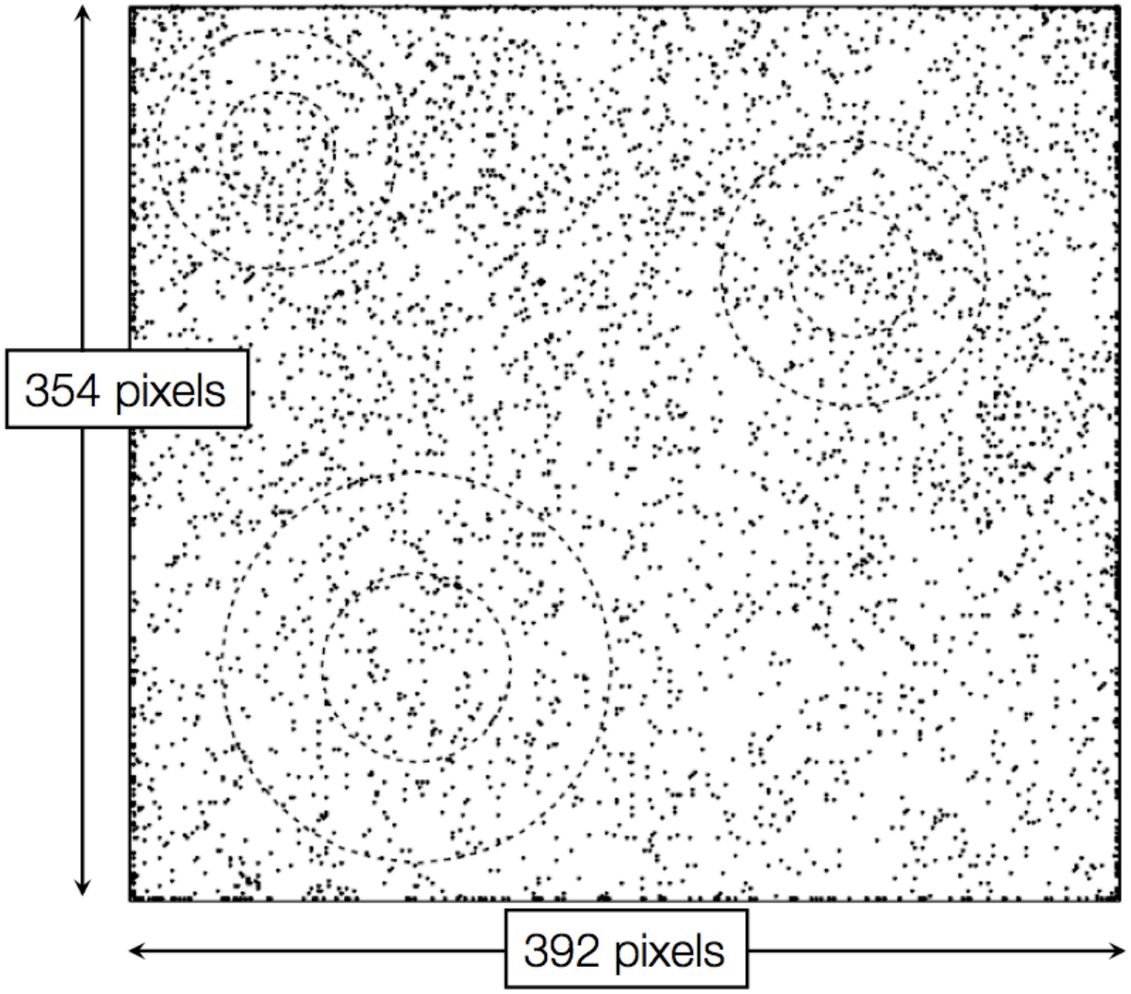
(Markov-chain Monte Carlo)

- MCMC draws samples from the posterior
  → output is a list of values that can approximate the posterior

- Only need to compare which posterior density is higher
  → So we only need the ratio of posteriors
  (marginal likelihoods cancel out!)

$$\frac{P(\theta_1 \mid D)}{P(\theta_2 \mid D)} = \frac{\frac{P(D \mid \theta_1)P(\theta_1)}{\cancel{P(D)}}}{\frac{P(D \mid \theta_2)P(\theta_2)}{\cancel{P(D)}}}$$

Pure random walk (courtesy of Paul Lewis)



Random walk

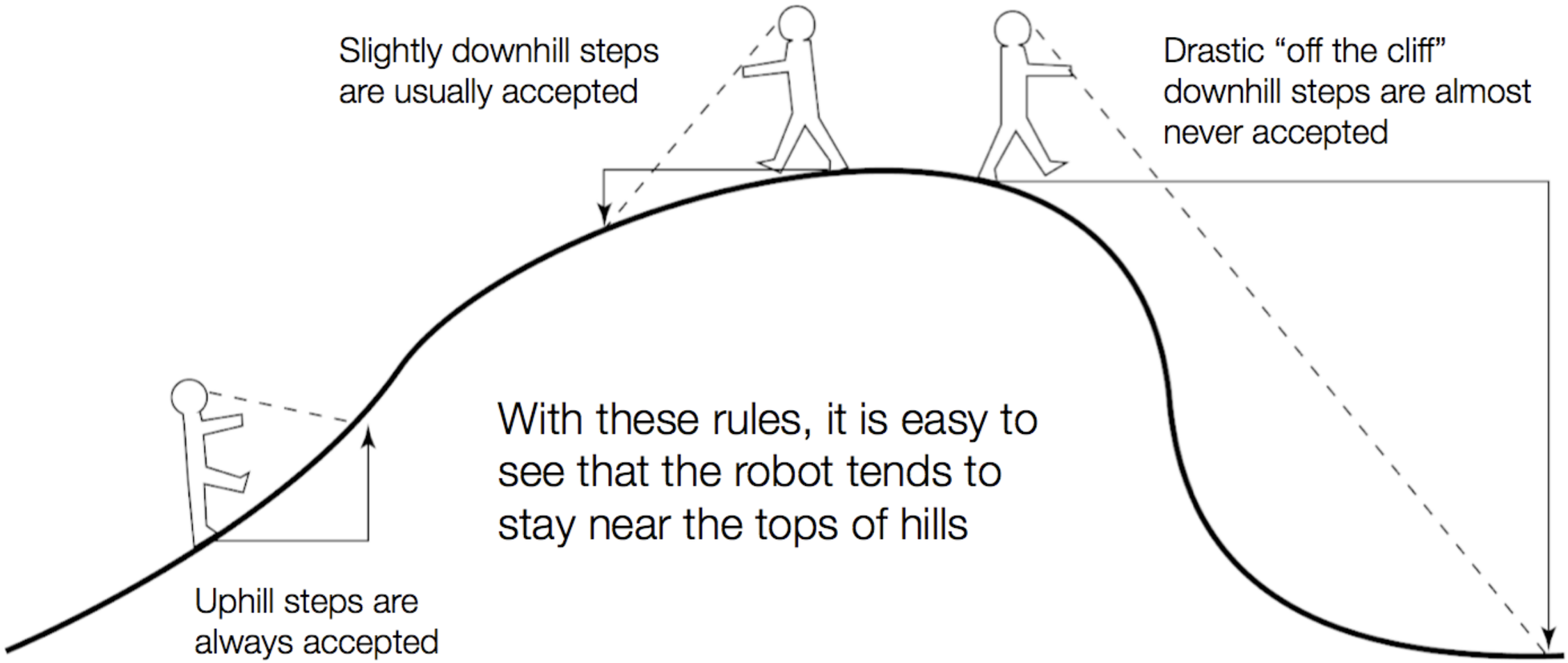
- Random direction
- Gamma distributed step size
- Reflection at edges

Target distribution

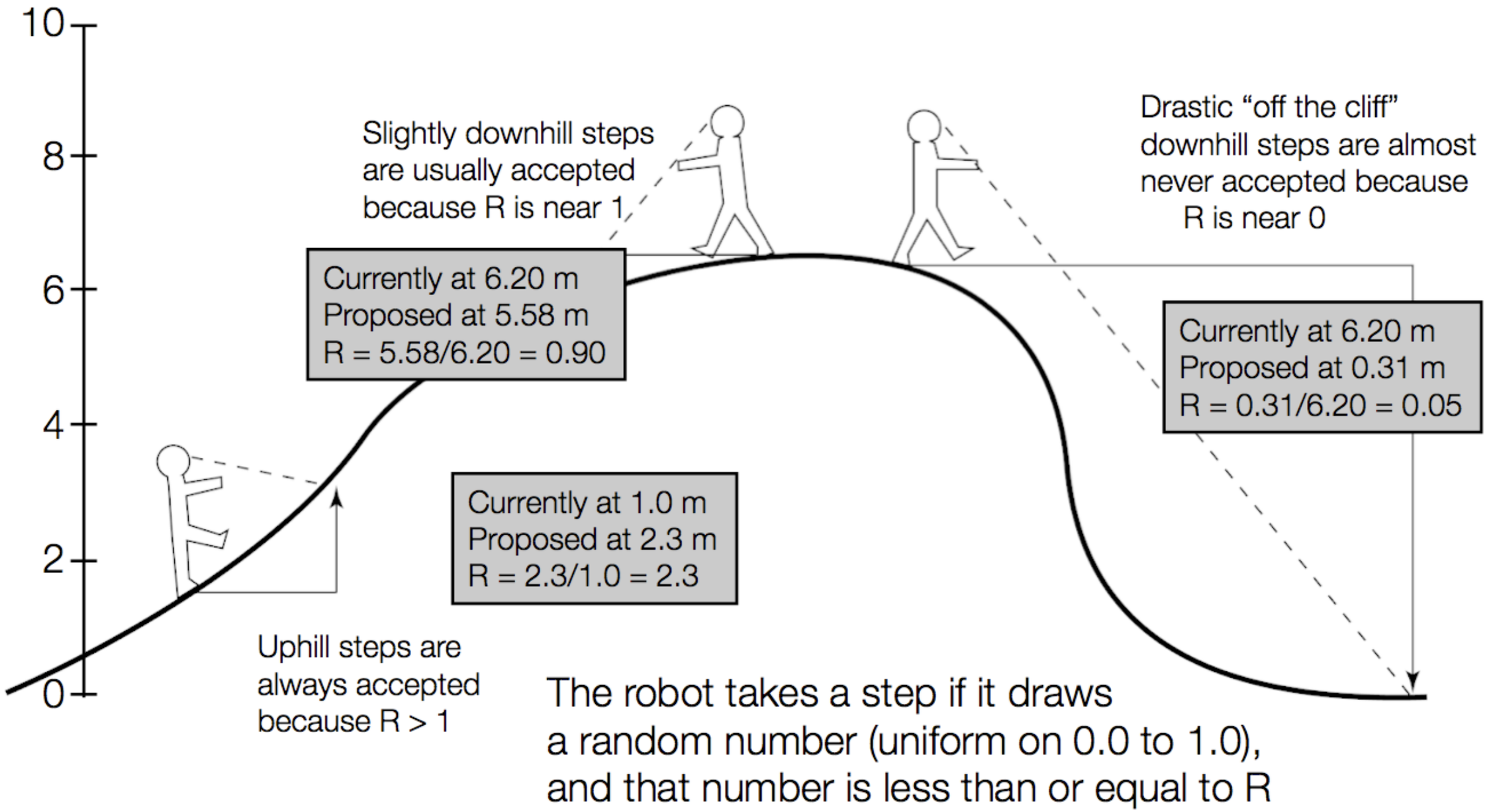
- Equal mixture of 3 bivariate normal hills
- Inner contours: 50%
- Outer contours: 95%

5000 steps by the random walk - not informative at all!

MCMC robot (courtesy of Paul Lewis)

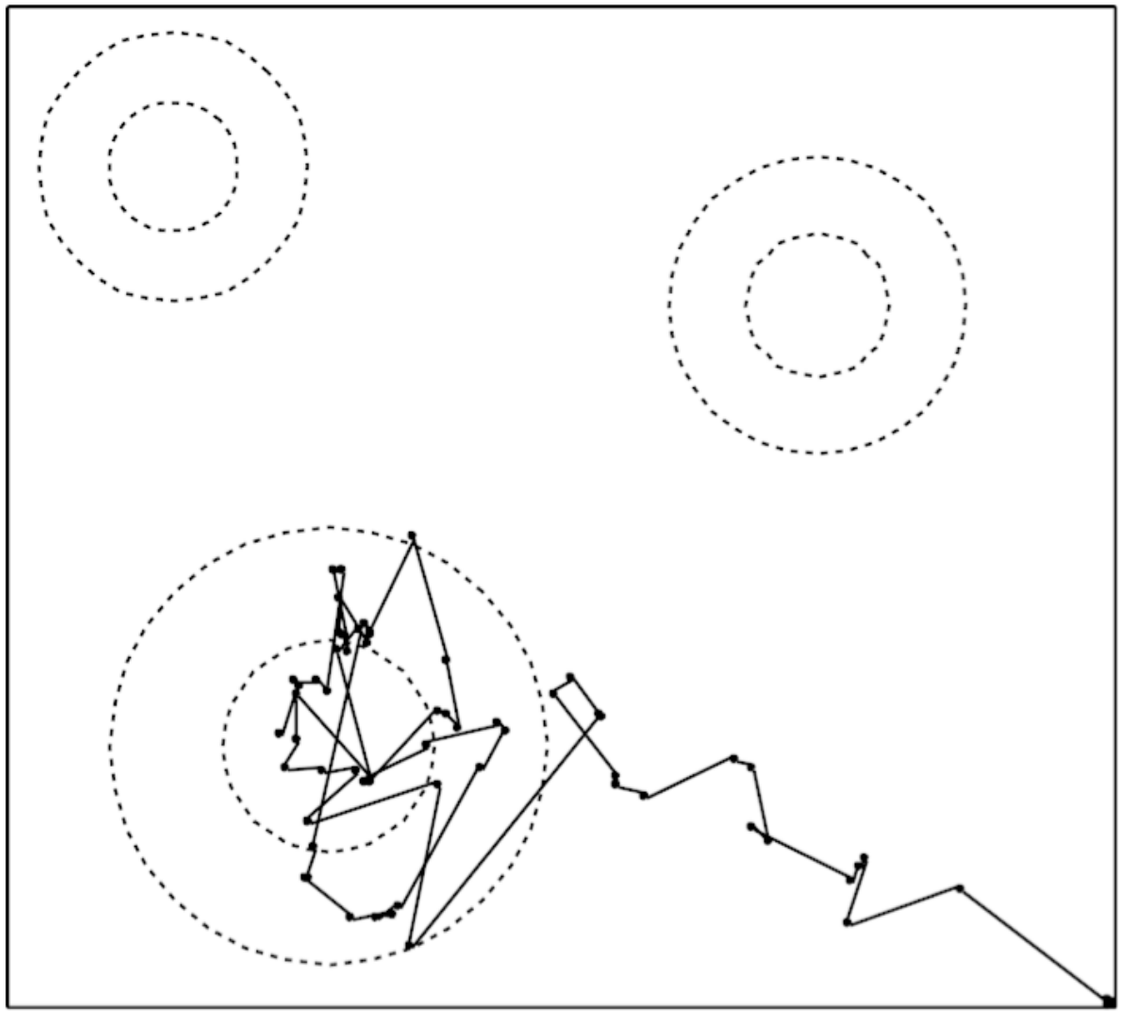


MCMC robot (courtesy of Paul Lewis)



(R is the ratio between the posterior densities)

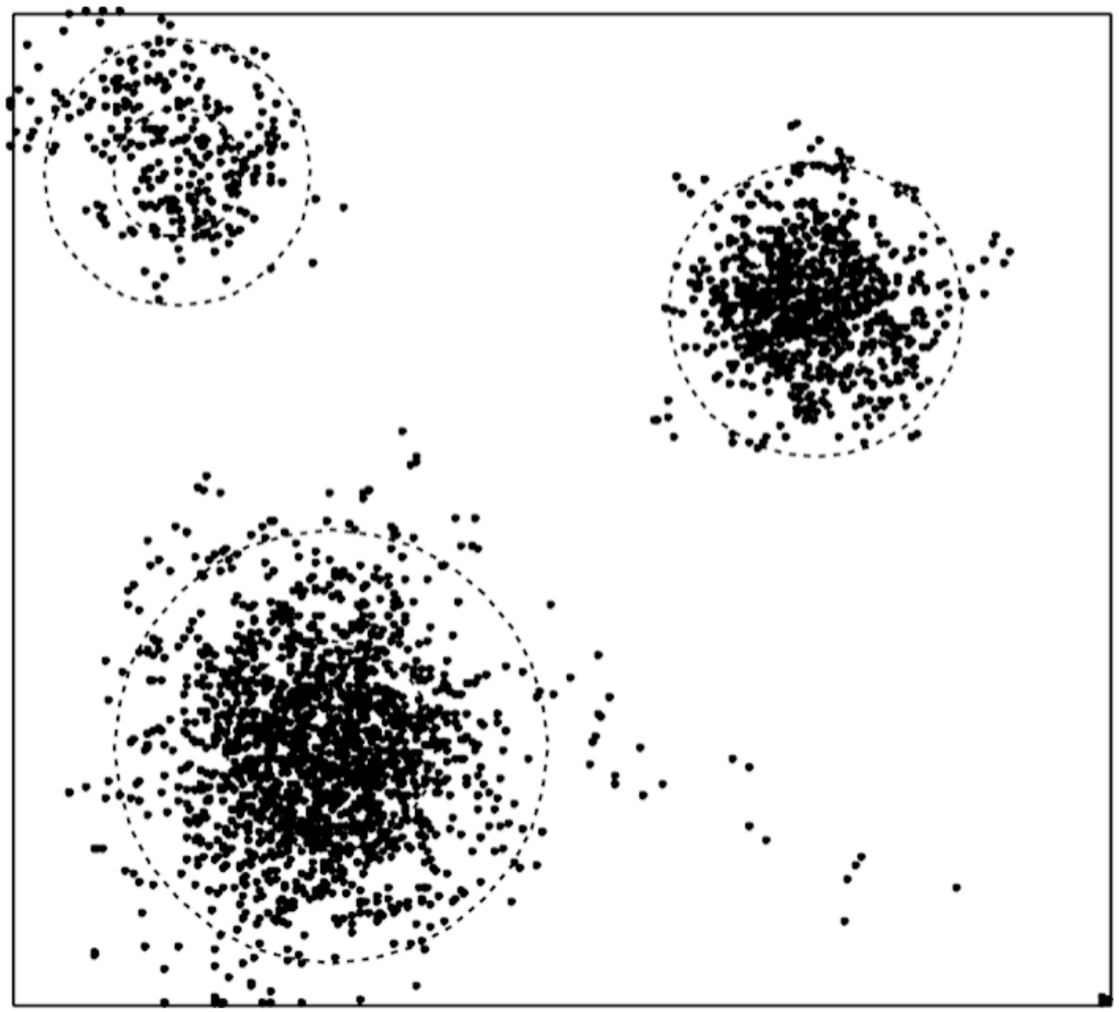
Burn in (courtesy of Paul Lewis)



- Using MCMC rules the robot quickly finds one of the 3 hills
- First few steps are not representative of the distribution

First 100 steps by the robot

MCMC approximation (courtesy of Paul Lewis)



5000 steps by the robot

How good is the approximation?

- 51.2% of points inside 50% contours
- 93.6% of points inside 95% contours

The more steps, the better the accuracy

Operators

Target distribution

- This is the **posterior** in BEAST2: P()
- MCMC is a directed random walk through the state space and samples the target distribution
- How to pick the next state to sample?

Proposal distribution

- Used to decide where to step to next
- The choice only affects the **efficiency** of the algorithm
- In BEAST1 and BEAST2 operators are used to propose the next step
- A parameter (or multiple parameters) are selected and perturbed to propose a step

Operators

Target

- This
- MC
  spa
- Hov

- Operators are a part of the MCMC **algorithm**, not the **model!**
- Tuning operators can help to improve efficiency, but should not change the results

Proposal distribution

- Used to decide where to step to next
- The choice only affects the **efficiency** of the algorithm
- In BEAST1 and BEAST2 operators are used to propose the next step
- A parameter (or multiple parameters) are selected and perturbed to propose a step

How many trees are there?

Genomes	Possible trees
4	15
5	105
6	945
7	10395
8	135135
9	2027025
10	34459425
20	8.2×10^{21}
48	3.21×10^{70}

≈ The number of particles in the universe

MCMC in practice

Before

- Decide on the length of the chain
  (total number of steps to take)
- Decide on the sampling frequency
  (how often to record samples so
  that they are uncorrelated)

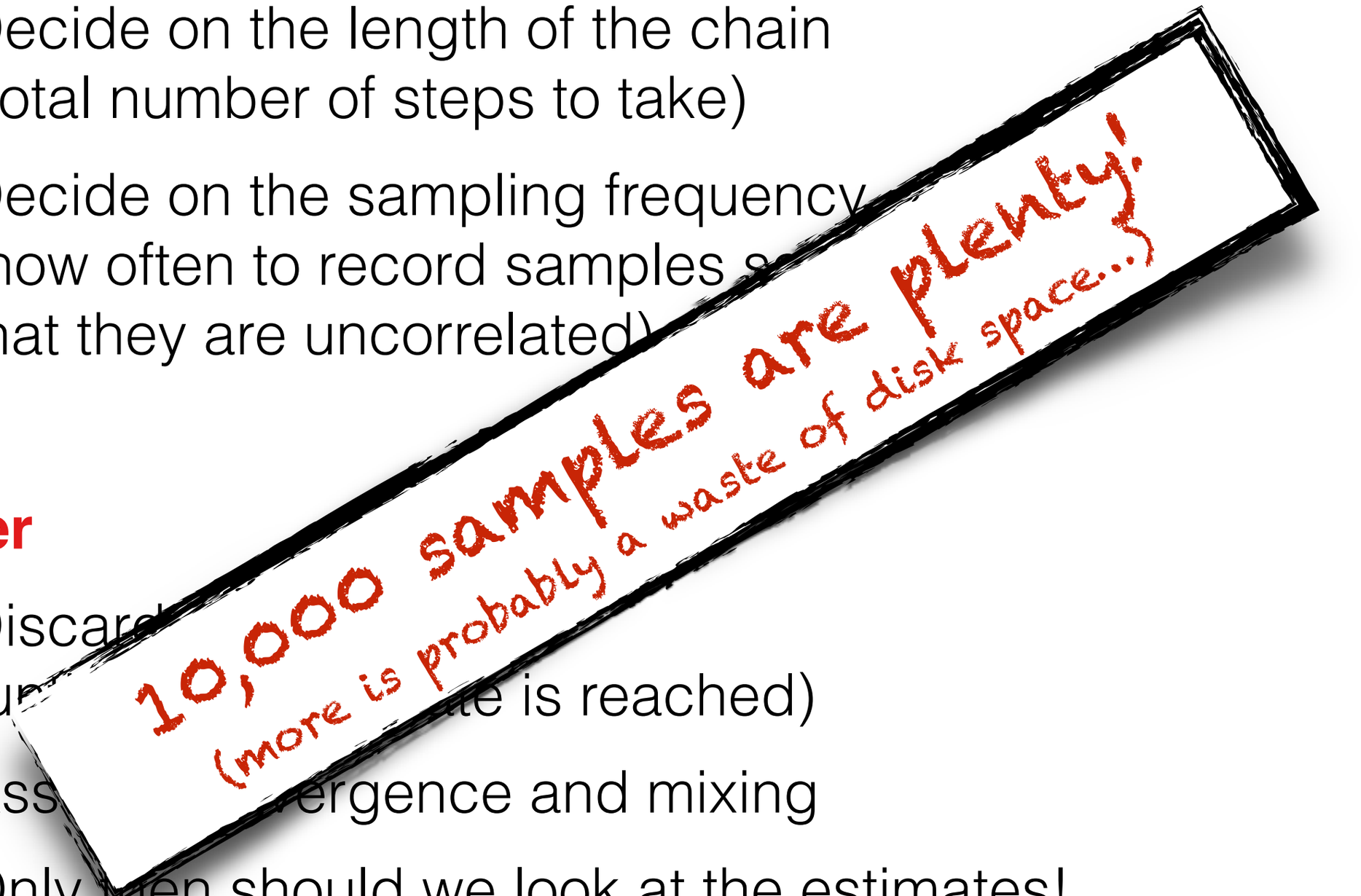
After

- Discard burn-in
  (until stationary state is reached)
- Assess convergence and mixing
- Only then should we look at the estimates!

MCMC in practice

Before

- Decide on the length of the chain (total number of steps to take)
- Decide on the sampling frequency (how often to record samples s… that they are uncorrelated)

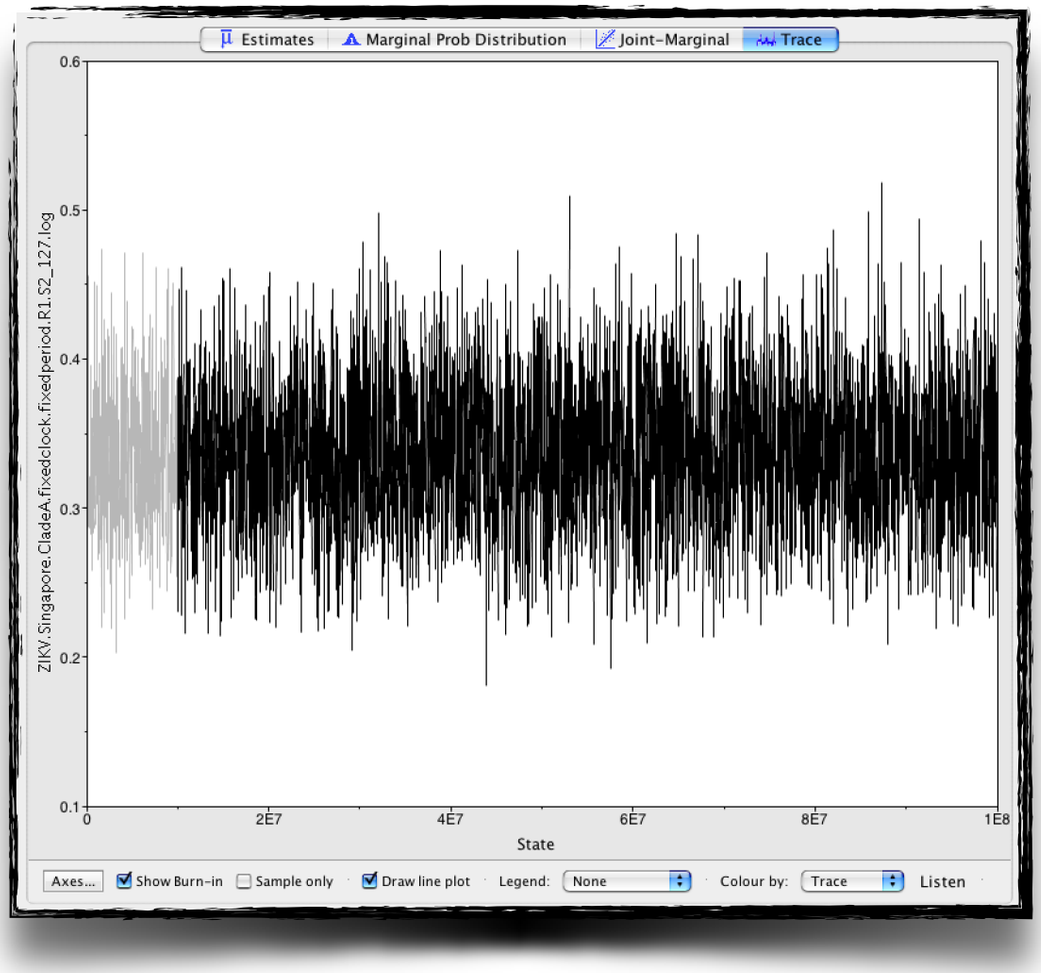


After

- Discard … (u… …te is reached)
- Ass… …ergence and mixing
- Only then should we look at the estimates!

What we hope will happen

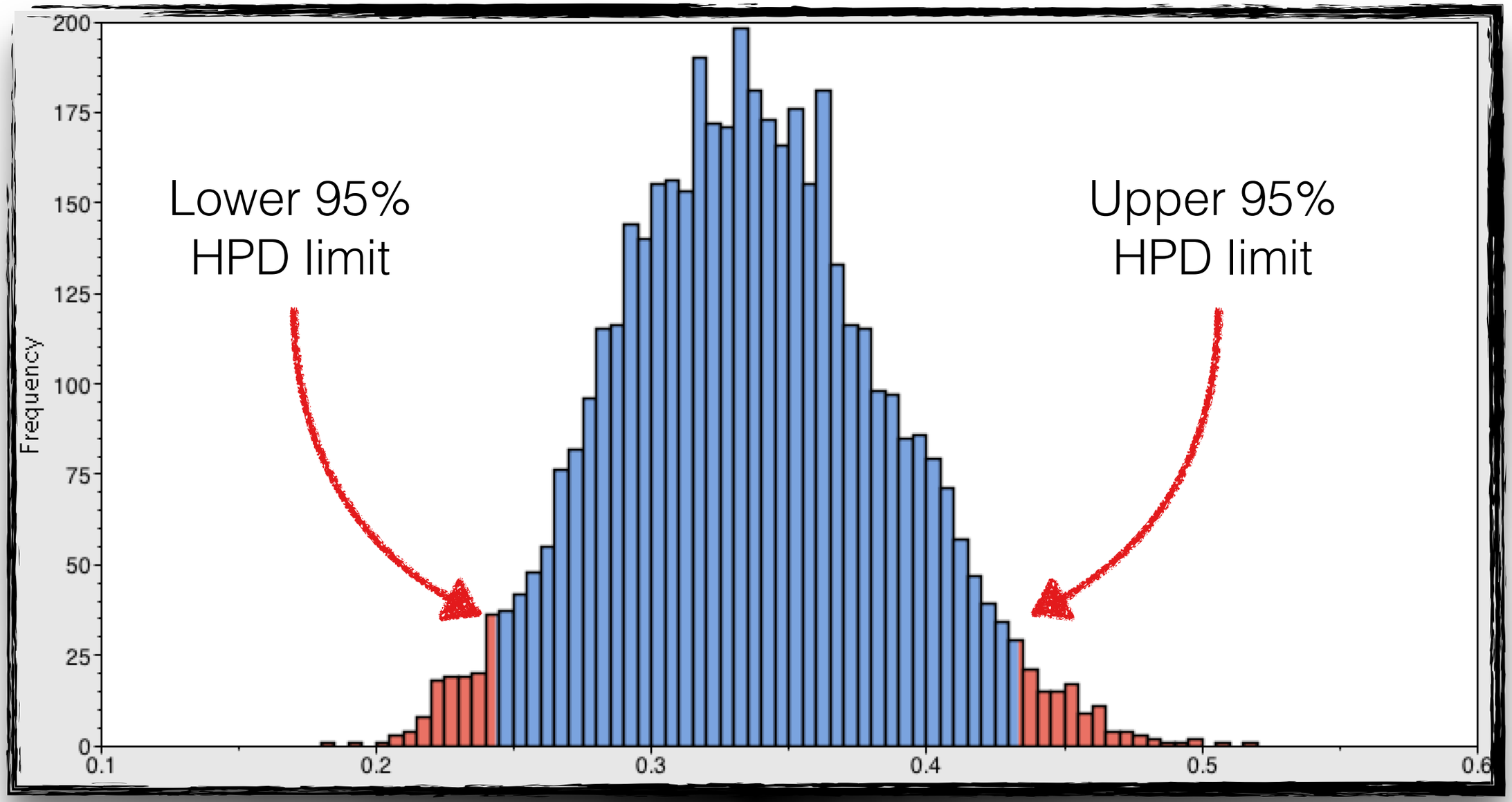
- The MCMC algorithm samples efficiently from high density areas of the posterior distribution
- We end up with a **good** approximation of the posterior distribution in **finite** time
- Appearance of white noise
- Everything is awesome!



Mixing well! 😄

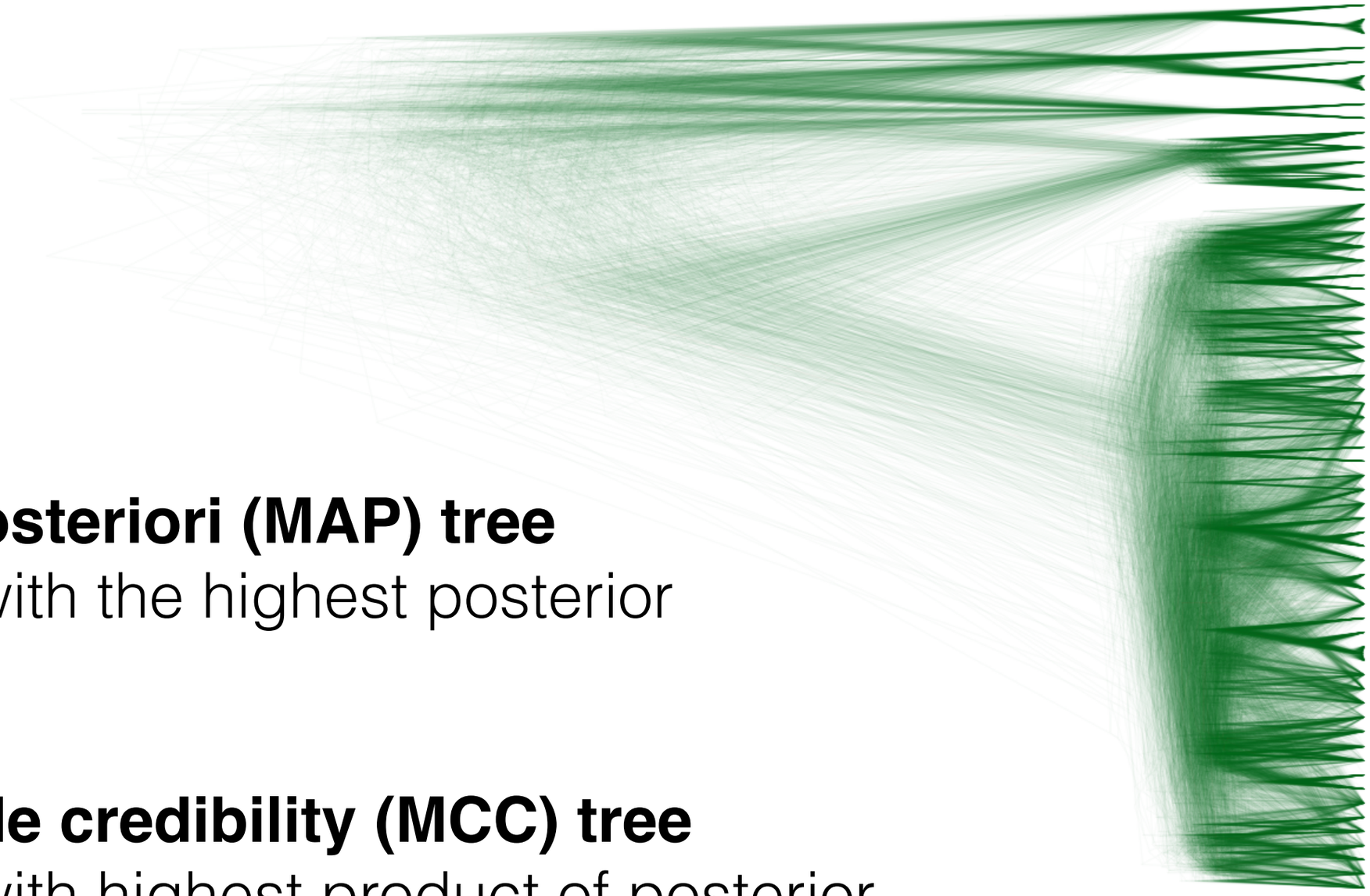
HPD intervals

(Highest Posterior Density)



Smallest region that contains 95% of the posterior probability

What about trees?

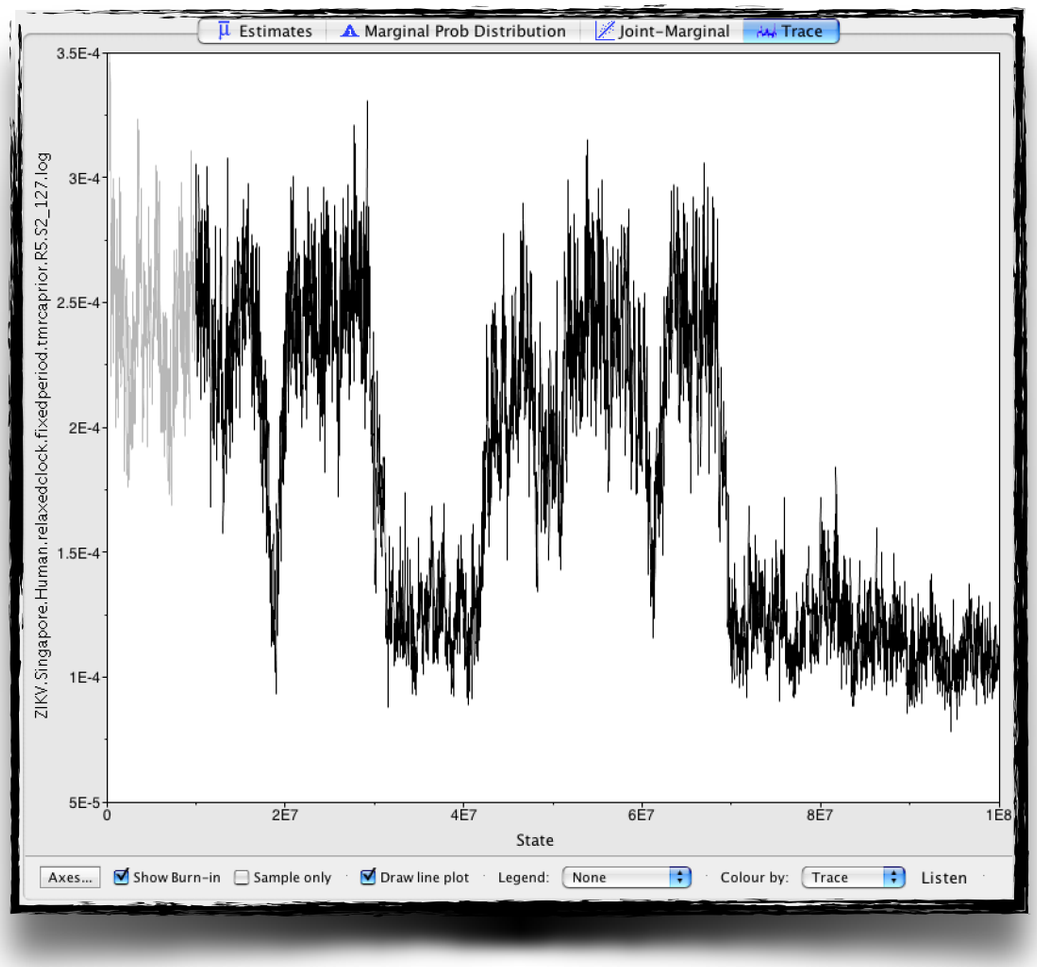


Maximum a posteriori (MAP) tree
Sampled tree with the highest posterior probability

Maximum clade credibility (MCC) tree
Sampled tree with highest product of posterior node probabilities

Questions to ask…

- Is the chain **mixing** well?
- Are samples uniformly drawn from all over the stationary distribution?
- Do we have a “sticky chain?”



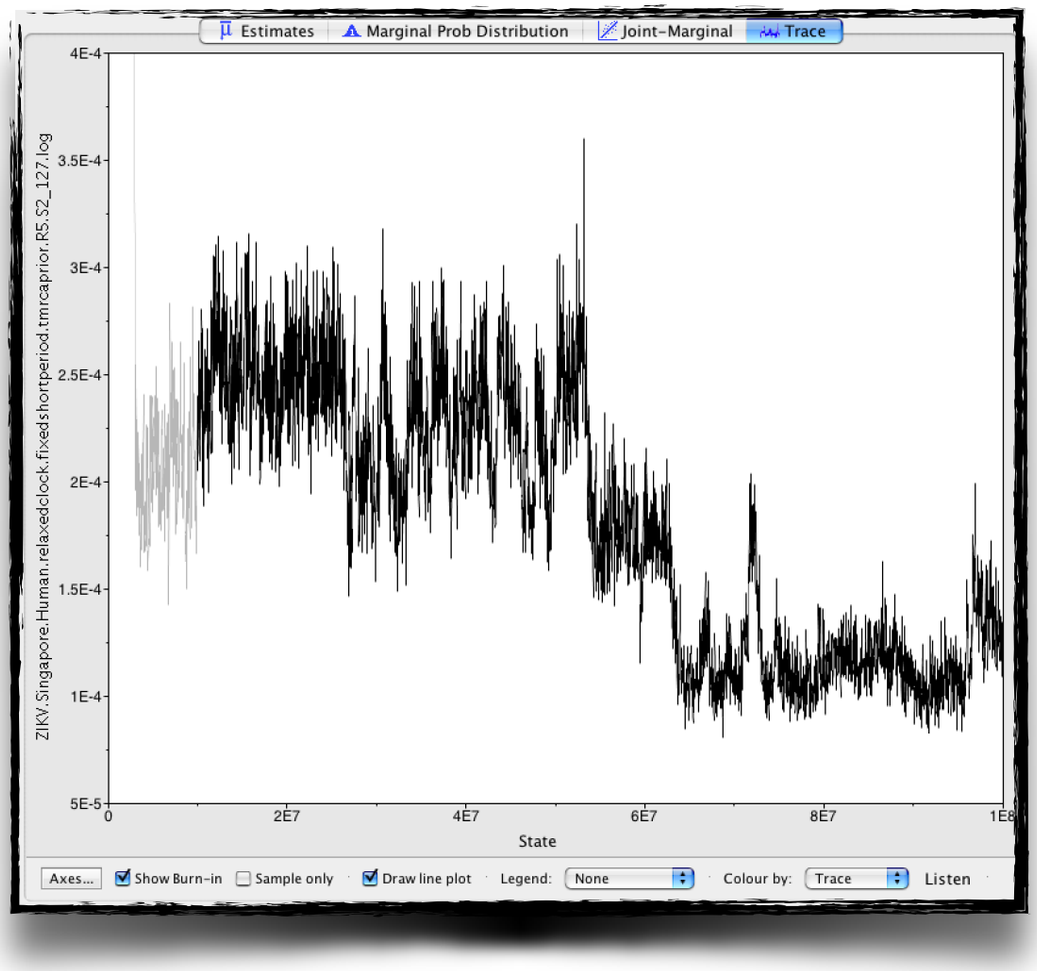
Not mixing! 😧

Solutions

- MCMC gets stuck in some states for long times
- Tune operators to make better proposals

Questions to ask…

- Has the chain **converged** to the stationary distribution?
- Did we pass the burn-in?

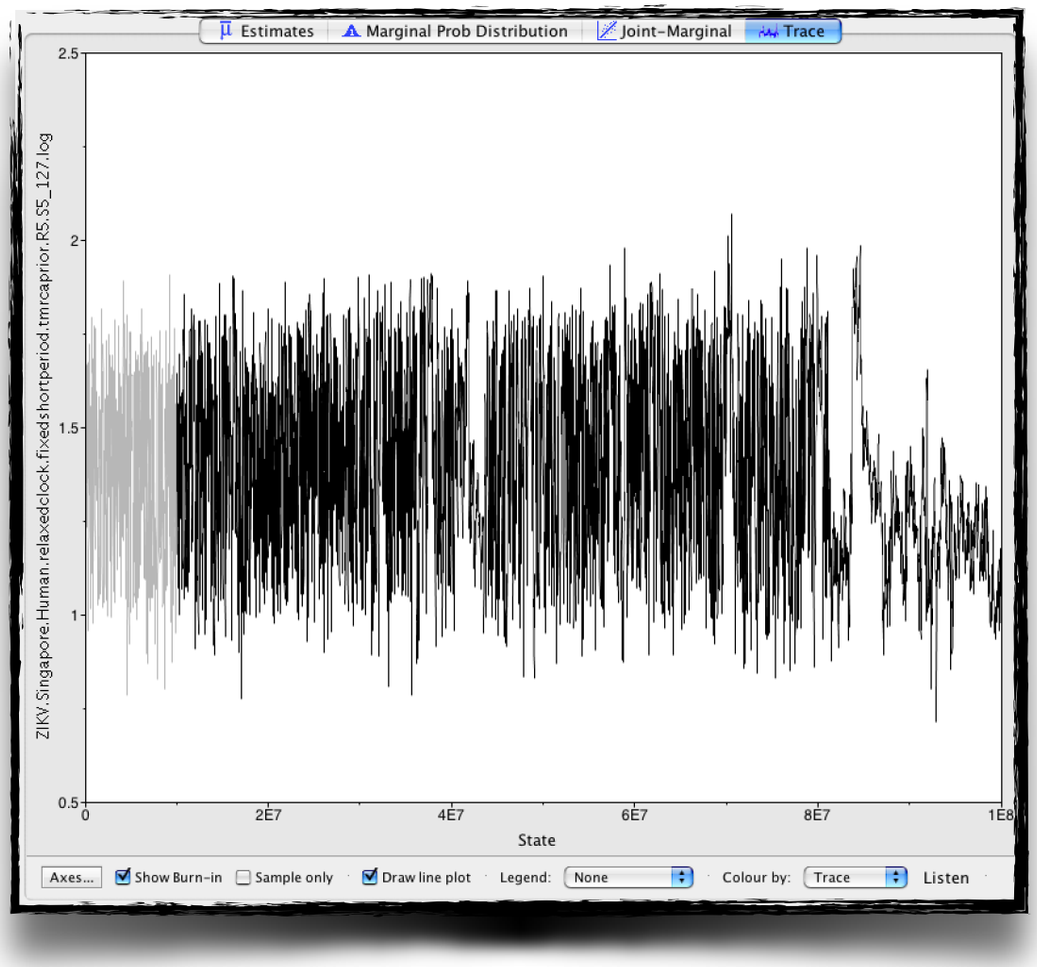


Not converged! 😓

Solution: Run for longer

Questions to ask…

- Are we there yet?
- How do we know if the chain is long enough?



Still not converged! 😵

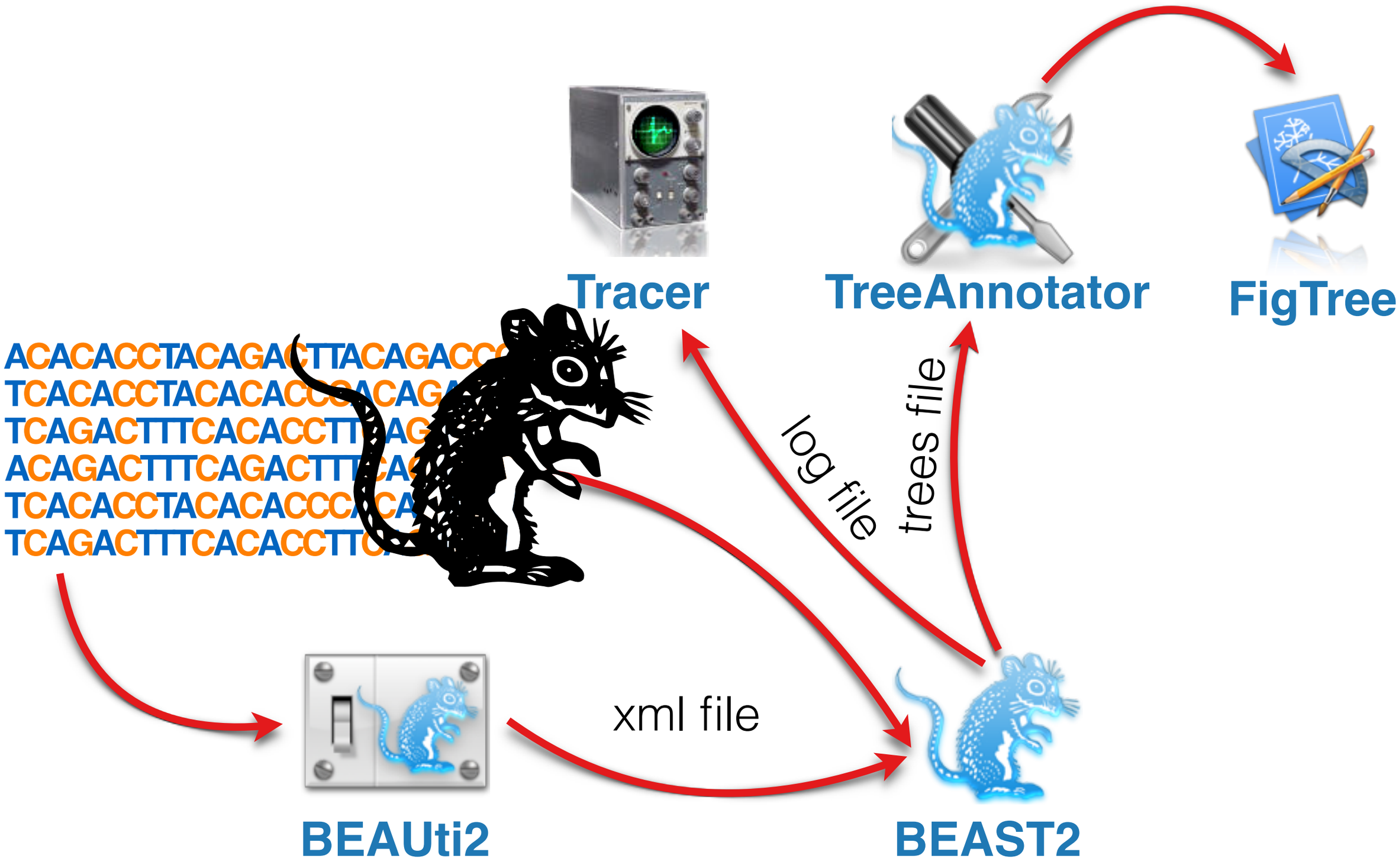
Solution

- Run multiple chains
- Combine chains
- Check that all chains give the same result

FANTASTIC BEASTS

AND WHERE TO FIND THEM

BEAST2 workflow



BEAUti2

(http://beast2.org)

Graphical tool for setting up a BEAST2 analysis

Input:

- Genetic sequence data
- Optional:
  - Sampling times
  - Sampling locations
  - Traits
  - etc.

Output:

- Compact XML description of data, model and prior distributions that can be run in BEAST2

BEAUti 2: Standard /Users/louis/Documents/Taming_the_BEAST/Tutorials-Git/Introduction-to-BEAST2/xml/Primates.xml

Partitions | Tip Dates | Site Model | Clock Model | Priors | MCMC

Link Site Models | Unlink Site Models | Link Clock Models | Unlink Clock Models | Link Trees | Unlink Trees

Name	File	Taxa	Sites	Data Type	Site Model	Clock Model	Tree	...
noncoding	primate-mtDNA	12	205	nucleotide	noncoding	clock	tree	
1stpos	primate-mtDNA	12	231	nucleotide	1stpos	clock	tree	
2ndpos	primate-mtDNA	12	231	nucleotide	2ndpos	clock	tree	
3rdpos	primate-mtDNA	12	231	nucleotide	3rdpos	clock	tree	

+ | - | r | Split

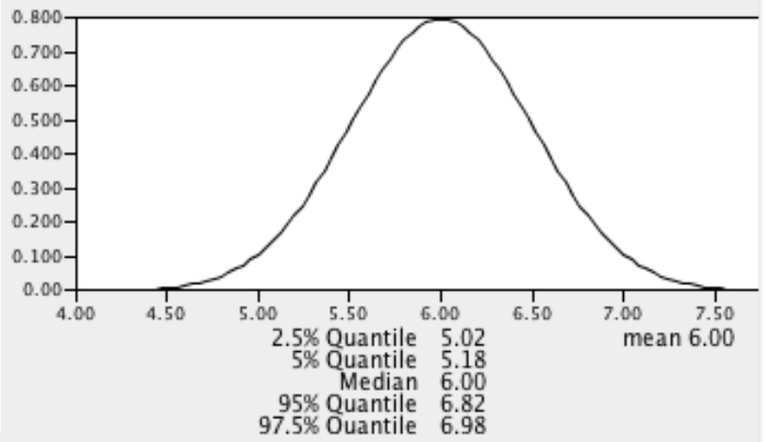
BEAUti 2: Standard /Users/louis/Documents/Taming_the_BEAST/Tutorials-Git/Introduction-to-BEAST2/xml/Primates.xml

Partitions | Tip Dates | Site Model | Clock Model | Priors | MCMC

Tree.t:tree	Calibrated Yule Model		
birthRateY.t:tree	Gamma	initial = [1.0] [−∞,∞]	Calibrated Yule speciation process birth rate for t:3rdpos
clockRate.c:clock	Uniform	initial = [1.0] [−∞,∞]	substitution rate of partition c:3rdpos
gammaShape.s:1stpos	Exponential	initial = [1.0] [−∞,∞]	Prior on gamma shape for partition s:1stpos
gammaShape.s:2ndpos	Exponential	initial = [1.0] [−∞,∞]	Prior on gamma shape for partition s:2ndpos
gammaShape.s:3rdpos	Exponential	initial = [1.0] [−∞,∞]	Prior on gamma shape for partition s:3rdpos
gammaShape.s:noncoding	Exponential	initial = [1.0] [−∞,∞]	Prior on gamma shape for partition s:noncoding
kappa.s:1stpos	Log Normal	initial = [2.0] [0.0,∞]	HKY transition-transversion parameter of partition s:1stpos
kappa.s:2ndpos	Log Normal	initial = [2.0] [0.0,∞]	HKY transition-transversion parameter of partition s:2ndpo
kappa.s:3rdpos	Log Normal	initial = [2.0] [0.0,∞]	HKY transition-transversion parameter of partition s:3rdpos
kappa.s:noncoding	Log Normal	initial = [2.0] [0.0,∞]	HKY transition-transversion parameter of partition s:nonco

human-chimp.prior — Normal — monophyletic

Mean	6.0	estimate
Sigma	0.5	estimate
Offset	0.0	



Tipsonly

Use Originate

Primates_long.xml — UNREGISTERED

Primates_long.xml

```
39
40  <run id="mcmc" spec="MCMC" chainLength="2500000">
41      <state id="state" storeEvery="5000">
42          <tree id="Tree.t:tree" name="stateNode">
43              <taxonset id="TaxonSet.noncoding" spec="TaxonSet">
44                  <alignment id="noncoding" spec="FilteredAlignment" filter="1,458-659,897-898">
45                      <data idref="primate-mtDNA"/>
46                  </alignment>
47              </taxonset>
48          </tree>
49          <parameter id="mutationRate.s:noncoding" name="stateNode">1.0</parameter>
50          <parameter id="gammaShape.s:noncoding" name="stateNode">1.0</parameter>
51          <parameter id="kappa.s:noncoding" lower="0.0" name="stateNode">2.0</parameter>
52          <parameter id="kappa.s:1stpos" lower="0.0" name="stateNode">2.0</parameter>
53          <parameter id="gammaShape.s:1stpos" name="stateNode">1.0</parameter>
54          <parameter id="mutationRate.s:1stpos" name="stateNode">1.0</parameter>
55          <parameter id="kappa.s:2ndpos" lower="0.0" name="stateNode">2.0</parameter>
56          <parameter id="gammaShape.s:2ndpos" name="stateNode">1.0</parameter>
57          <parameter id="mutationRate.s:2ndpos" name="stateNode">1.0</parameter>
58          <parameter id="kappa.s:3rdpos" lower="0.0" name="stateNode">2.0</parameter>
59          <parameter id="gammaShape.s:3rdpos" name="stateNode">1.0</parameter>
60          <parameter id="mutationRate.s:3rdpos" name="stateNode">1.0</parameter>
61          <parameter id="birthRateY.t:tree" name="stateNode">1.0</parameter>
62          <parameter id="clockRate.c:clock" name="stateNode">1.0</parameter>
63      </state>
64
65      <init id="RandomTree.t:tree" spec="beast.evolution.tree.RandomTree" estimate="false" initial="@Tree.t:tree" taxa="
        @noncoding">
66          <populationModel id="ConstantPopulation0.t:tree" spec="ConstantPopulation">
67              <parameter id="randomPopSize.t:tree" name="popSize">1.0</parameter>
68          </populationModel>
69      </init>
70
71      <distribution id="posterior" spec="util.CompoundDistribution">
72          <distribution id="prior" spec="util.CompoundDistribution">
73              <distribution id="CalibratedYuleModel.t:tree" spec="beast.evolution.speciation.CalibratedYuleModel"
                birthRate="@birthRateY.t:tree" tree="@Tree.t:tree"/>
74              <prior id="CalibratedYuleBirthRatePrior.t:tree" name="distribution" x="@birthRateY.t:tree">
75                  <Gamma id="Gamma.0" name="distr">
76                      <parameter id="RealParameter.0" estimate="false" name="alpha">0.001</parameter>
77                      <parameter id="RealParameter.01" estimate="false" name="beta">1000.0</parameter>
78                  </Gamma>
79              </prior>
80              <prior id="ClockPrior.c:clock" name="distribution" x="@clockRate.c:clock">
81                  <Uniform id="Uniform.0" name="distr" upper="Infinity"/>
82              </prior>
```

Line 1, Column 3 | 0 misspelled words | Spaces: 4 | XML

BEAST2

(http://beast2.org)

- **B**ayesian **e**volutionary **a**nalysis by **s**ampling **t**rees
- Performs MCMC analyses of sequences under selected sequence evolution and tree model
- Similar to BEAST 1.10 but completely separate and generally incompatible
- BEAST2 and BEAST1 have a common origin, have much of the same functionality but have diverged over time
- BEAST2 has a modular design that makes it easy to extend
- GUI interface but can also be run from command line (e.g. on a cluster)

Input:

- xml model description file

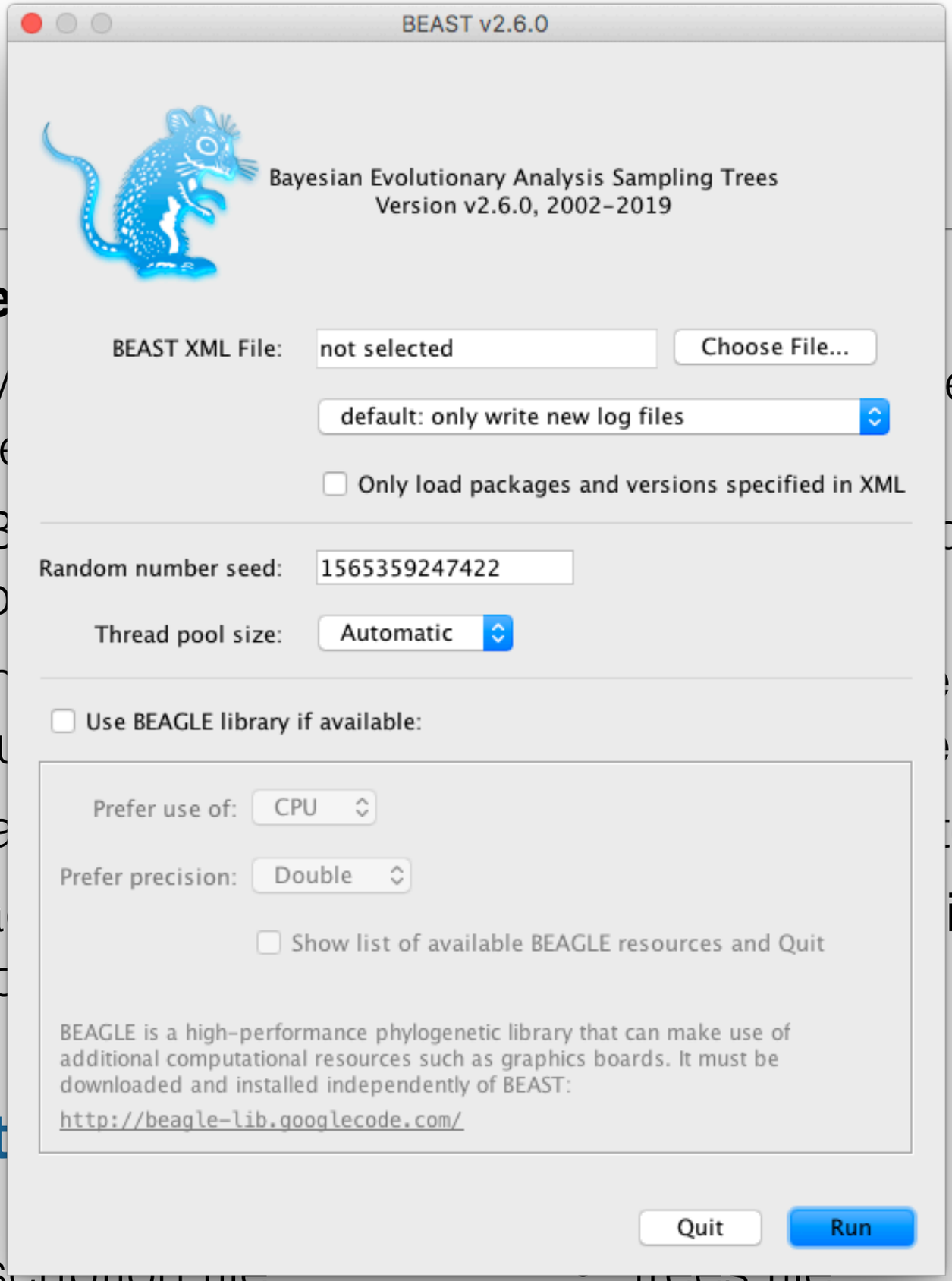
Outputs:

- log file
- trees file
- state file

BEAST2

(http://beast2.org)

- **B**ayesian **e**
- Performs M ected sequence e
- Similar to B d generally incompatib
- BEAST2 ar much of the same fu
- BEAST2 ha to extend
- GUI interfa ine (e.g. on a c



Input

- xml description file
- trees file
- state file

BEAST2 packages

- BEAST2 is organised into a central "core" together with a large number of separate packages
- Packages can be developed by anybody — **including you!**
- Can be directly integrated into BEAST2 and updated frequently without waiting for a full BEAST2 release
- Packages add new models or completely new functionality
  - phylogeography
  - bacterial ARG inference
  - morphological models
  - multispecies coalescent
  - model selection and averaging
  - stochastic simulations
  - linguistic analyses
  - …
- Install new packages through the package manager in BEAUti

BEAST 2 Package Manager

List of available packages for BEAST v2.6.*

Name	Installed	Latest	Dependencies	Link	Detail
BEAST	2.6.0	2.6.1			BEAST core
Babel		0.2.1	BEASTLabs		BABEL = BEAST analysis backing effective linguistics
bacter	2.2.3	2.2.3			Bacterial ARG inference.
BADTRIP		1.0.0			Infer transmission time for non-haplotype data and epi data
BASTA		3.0.1			Bayesian structured coalescent approximation
bdmm	0.3.5	0.3.5	MultiTypeTree, MASTER		pre-release of multitype birth-death model (aka birth-death-migration model)
BDSKY	1.4.5	1.4.5			birth death skyline - handles serially sampled tips, piecewise constant rate changes through time and sampled ancestors
BEAST_CLASSIC		1.5.0	BEASTLabs		BEAST classes ported from BEAST 1 in wrappers
BEASTLabs	1.9.0	1.9.0			BEAST utilities, such as Script, multi monophyletic constraints
BEASTvntr		0.1.3			Variable Number of Tandem Repeat data, such as microsattelites
Beasy		0.0.2	BEASTLabs		Makes it easier to construct models: Automatic methods text generator, Beasy XML generator, and more
bModelTest	1.2.1	1.2.1	BEASTLabs		Bayesian model test for nucleotide subst models, gamma rate heterogeneity and invariant sites
BREAK_AWAY		1.0.0	BEASTLabs, GEO_SPHERE		break-away model of phylogegraphy
CA		2.0.0			Bayesian estimation of clade ages based on probabilities of fossil sampling
CoalRe		0.0.4			Infer viral reassortment networks
CodonSubstModels		1.1.3			Codon substitution models
CoupledMCMC		0.1.7	BEASTLabs		Coupled MCMC (parallel Tempering or MC3)
DENIM		1.0.0			Divergence Estimation Notwithstanding ILS and Migration
EpiInf		7.1.5	SA		BD/SIR/SIS epidemic trajectory inference.
FLC		1.1.0			Flexible local clock model
GEO_SPHERE		1.3.0	BEASTLabs		Whole world phylogeography
Mascot		1.2.2			Marginal approximation of the structured coalescent
MASTER	6.1.1	6.1.1			Stochastic population dynamics simulation
MGSM		0.3.0			Multi-gamma and relaxed gamma site models
MM		1.1.1			Enables models of morphological chatacter evolution
MODEL_SELECTION		1.5.1	BEASTLabs		Select models through path sampling/stepping stone analysis
MSBD		1.1.0			Multi-state birth-death prior with state-specific birth and death rates
MultiTypeTree	7.0.1	7.0.1			Structured coalescent inference
NS		1.1.0	MODEL_SELECTION, BEASTLabs		Nested sampling for model selection and posterior inference
PhyDyn		1.3.4			PhyDyn: Epidemiological modelling with BEAST
phylodynamics		1.3.0	BDSKY		BDSIR and Stochastic Coalescent
PoMo		1.0.1			PoMo, a substitution model that separates mutation and drift processes
SA	2.0.2	2.0.2	BEASTLabs		Sampled ancestor trees
SCOTTI		2.0.1			Structured COalescent Transmission Tree Inference
SNAPP		1.5.0			SNP and AFLP Phylogenies
SpeciesNetwork		0.12.2			Multispecies network coalescent (MSNC) inference of introgression and hybridization
SSM		1.1.0			Standard Nucleotide Substitution Models
STACEY		1.2.5			Species delimitation and species tree estimation
StarBEAST2		0.15.5	SA, MM		Multispecies coalescent inference using multi-locus and fossil data
substBMA		1.2.3			Substitution Bayesian Model Averaging
TMA		1.0.0	MASTER, BEASTLabs, phylodynamics, BDSKY, TreeStat2		Tree model adequacy: test whether the tree prior used is adequate for your data
TreeStat2		0.0.2			Utility for calculating tree statistics from tree log file

Latest | Install/Upgrade | Uninstall | Package repositories | Close | ?

...

- Install new packages through the package manager in BEAUti

Tracer

(http://beast.community)

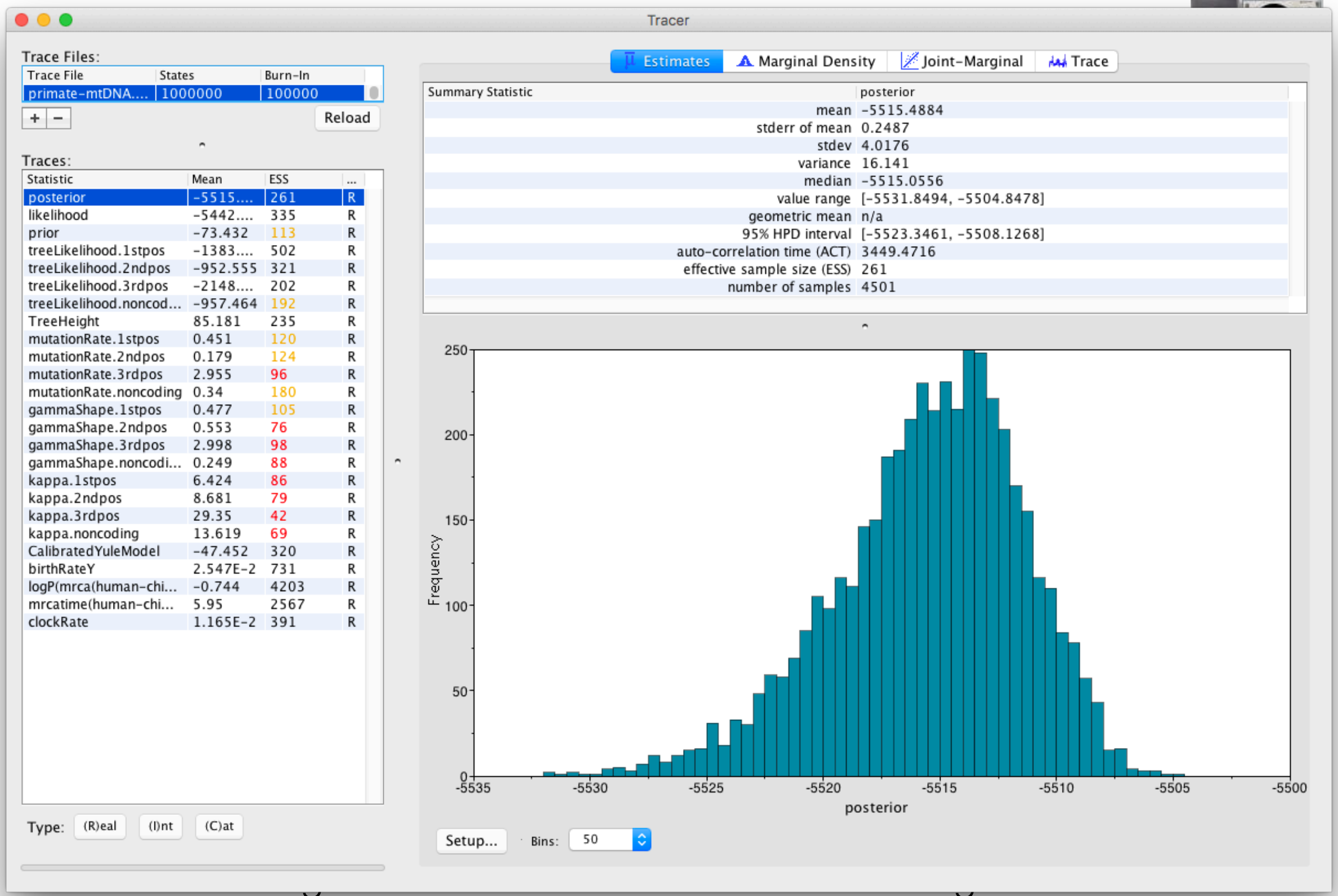
- Analyse (parameter) log files from BEAST2 runs
- Check mixing, ESS, ACT, parameter correlations
- Provides overview of posterior parameter estimates
- Easily compare several analyses
- Demographic reconstruction for some models (e.g. Bayesian Skyline Plot)
- Tracer is **primarily** a diagnostic tool — usually perform final analyses in a statistical package like R!

Input:

- log file

Output:

- Gain insight
- Demographic reconstructions



- Demographic reconstructions

Tracer

Trace Files:

Trace File	States	Burn-In
primate-mtDNA....	1000000	100000

+ − Reload

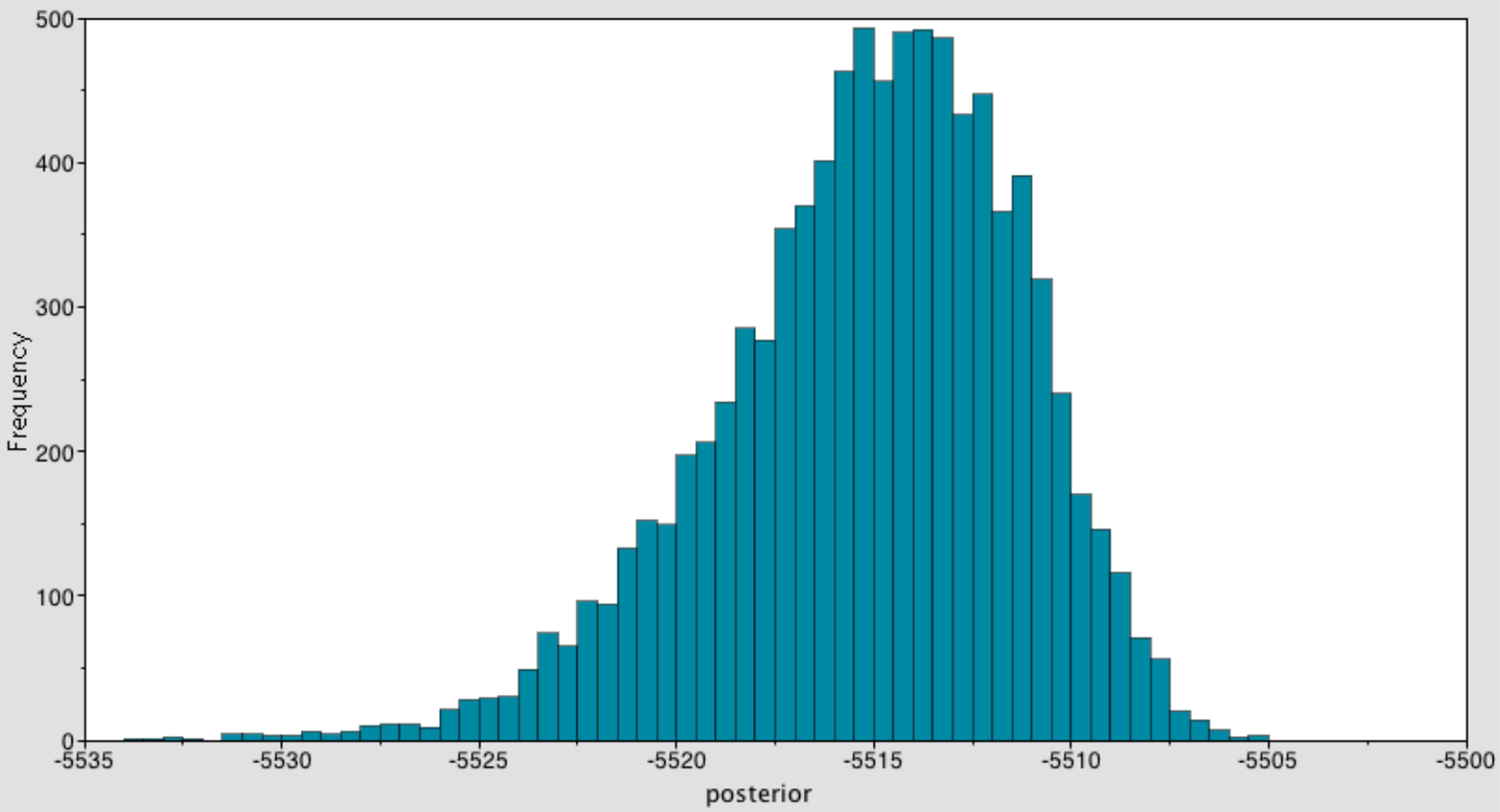
Traces:

Statistic	Mean	ESS	...
posterior	-5515....	2387	R
likelihood	-5441....	2349	R
prior	-73.169	1379	R
treeLikelihood.1stpos	-1383....	3189	R
treeLikelihood.2ndpos	-952.37	2885	R
treeLikelihood.3rdpos	-2148....	1687	R
treeLikelihood.noncod...	-957.267	1731	R
TreeHeight	83.827	1409	R
mutationRate.1stpos	0.45	852	R
mutationRate.2ndpos	0.182	714	R
mutationRate.3rdpos	2.949	646	R
mutationRate.noncodi...	0.346	1344	R
gammaShape.1stpos	0.496	889	R
gammaShape.2ndpos	0.575	911	R
gammaShape.3rdpos	3.022	726	R
gammaShape.noncodi...	0.244	1006	R
kappa.1stpos	6.235	719	R
kappa.2ndpos	8.5	1359	R
kappa.3rdpos	28.777	365	R
kappa.noncoding	13.478	875	R
CalibratedYuleModel	-47.285	1755	R
birthRateY	2.561E-2	3805	R
logP(mrca(human-chi...	-0.731	9001	R
mrcatime(human-chi...	5.949	8655	R
clockRate	1.161E-2	1836	R

Type: (R)eal (I)nt (C)at

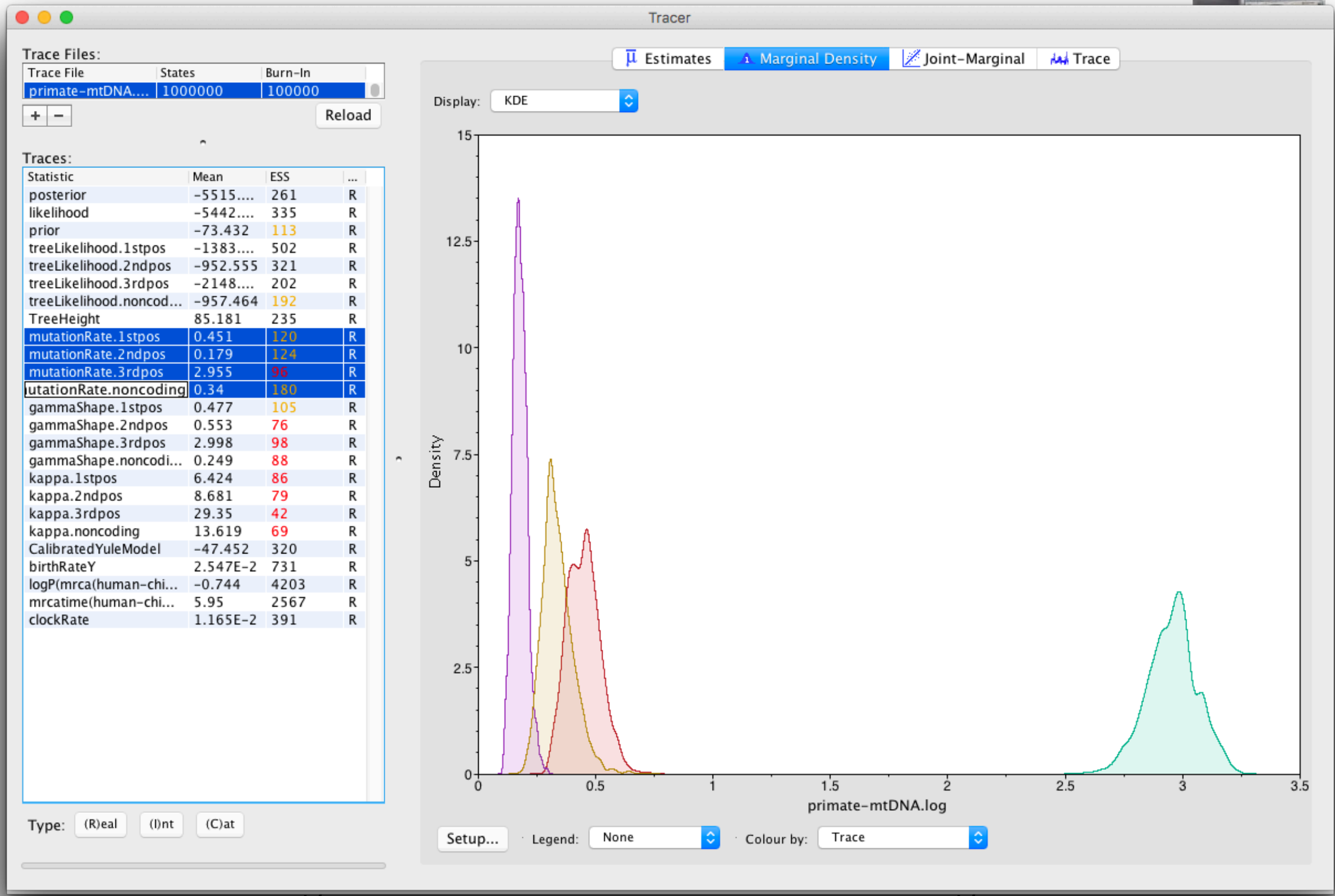
Estimates | Marginal Density | Joint-Marginal | Trace

Summary Statistic	posterior
mean	-5515.1348
stderr of mean	0.0786
stdev	3.8428
variance	14.7672
median	-5514.7368
value range	[-5533.9705, -5505.0922]
geometric mean	n/a
95% HPD interval	[-5523.187, -5508.5802]
auto-correlation time (ACT)	3770.478
effective sample size (ESS)	2387.2
number of samples	9001

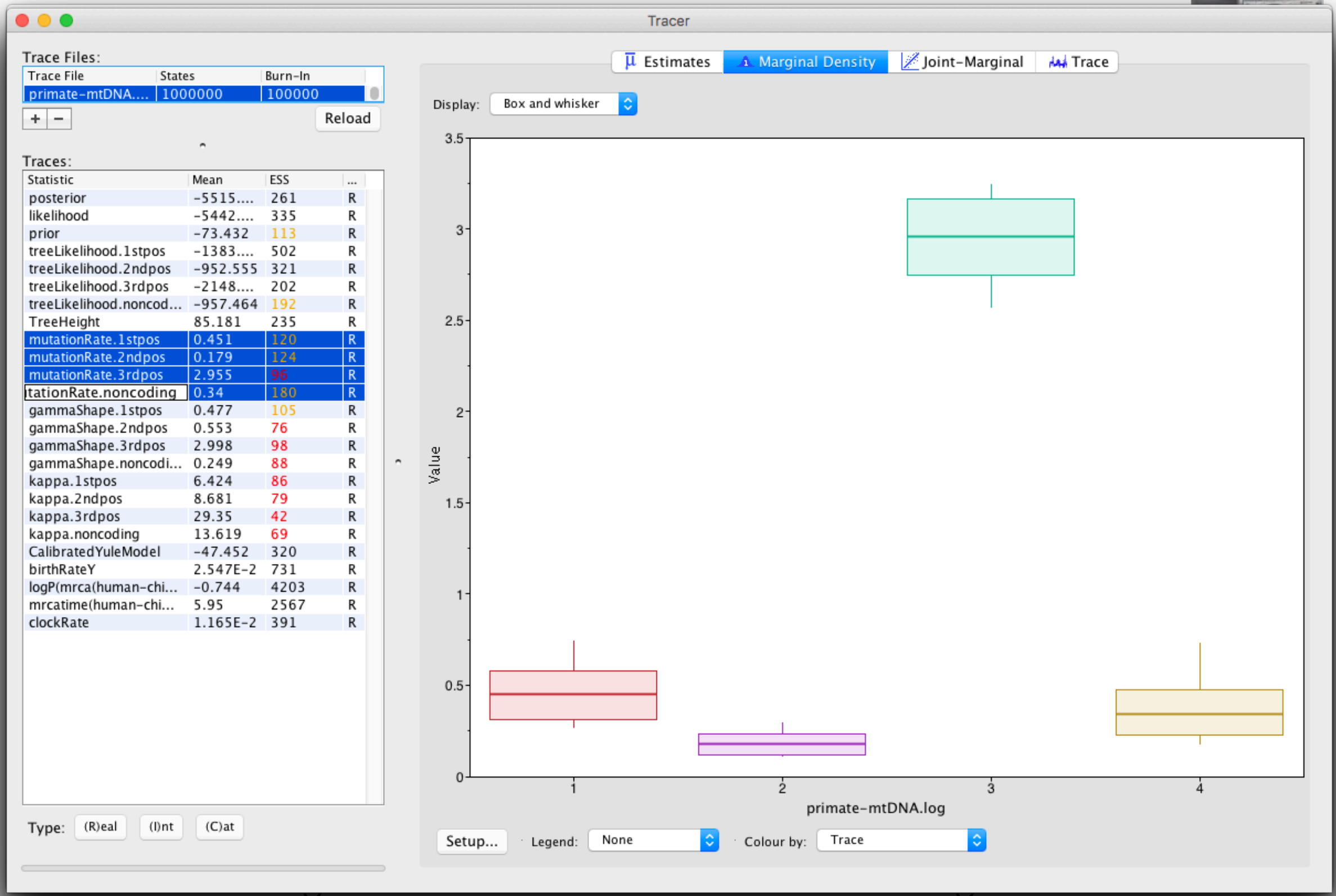


Setup... Bins: 50

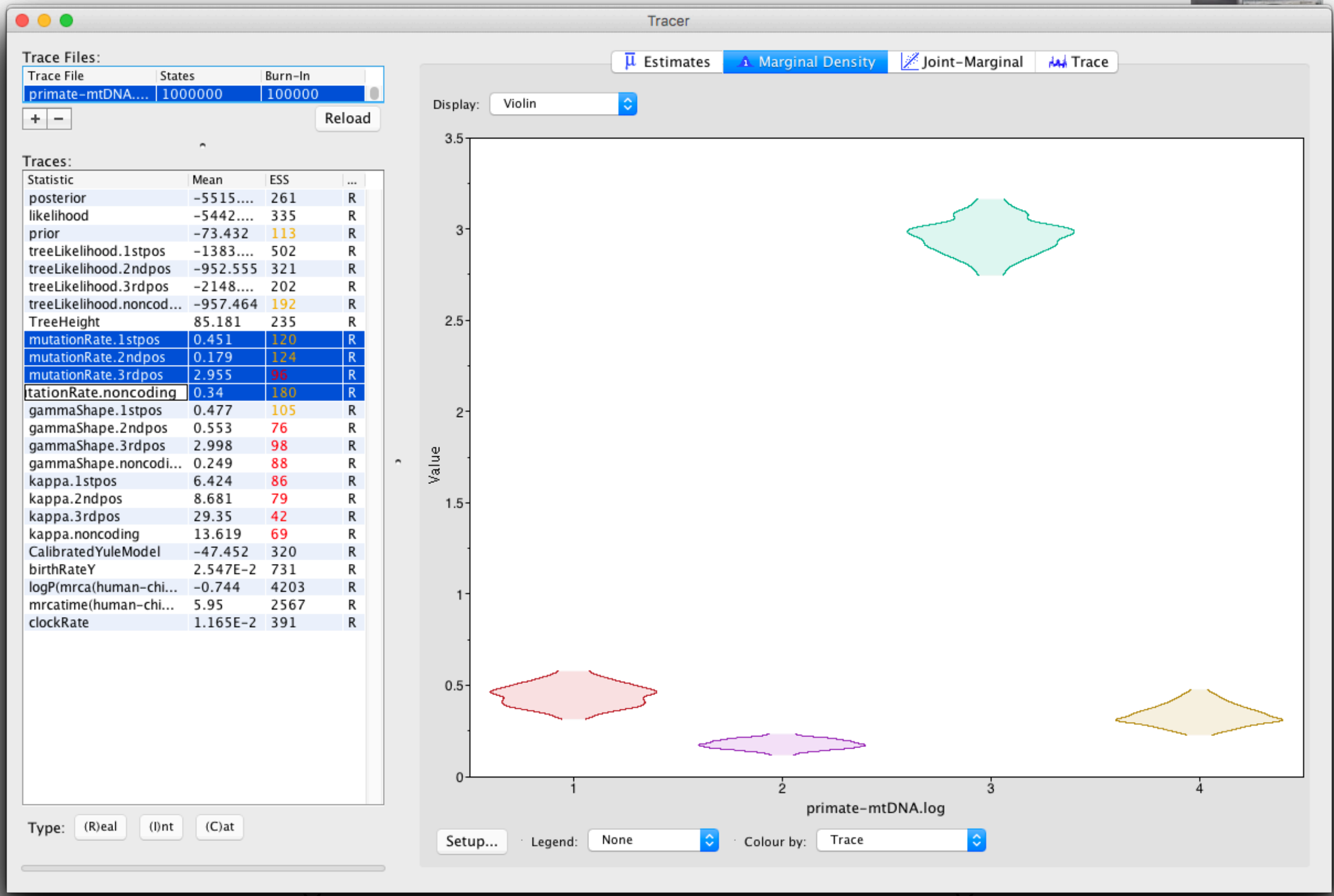
- Demographic reconstructions



- Demographic reconstructions



- Demographic reconstructions



- Demographic reconstructions

Tracer

Trace Files:

Trace File	States	Burn-In
primate-mtDNA....	1000000	100000

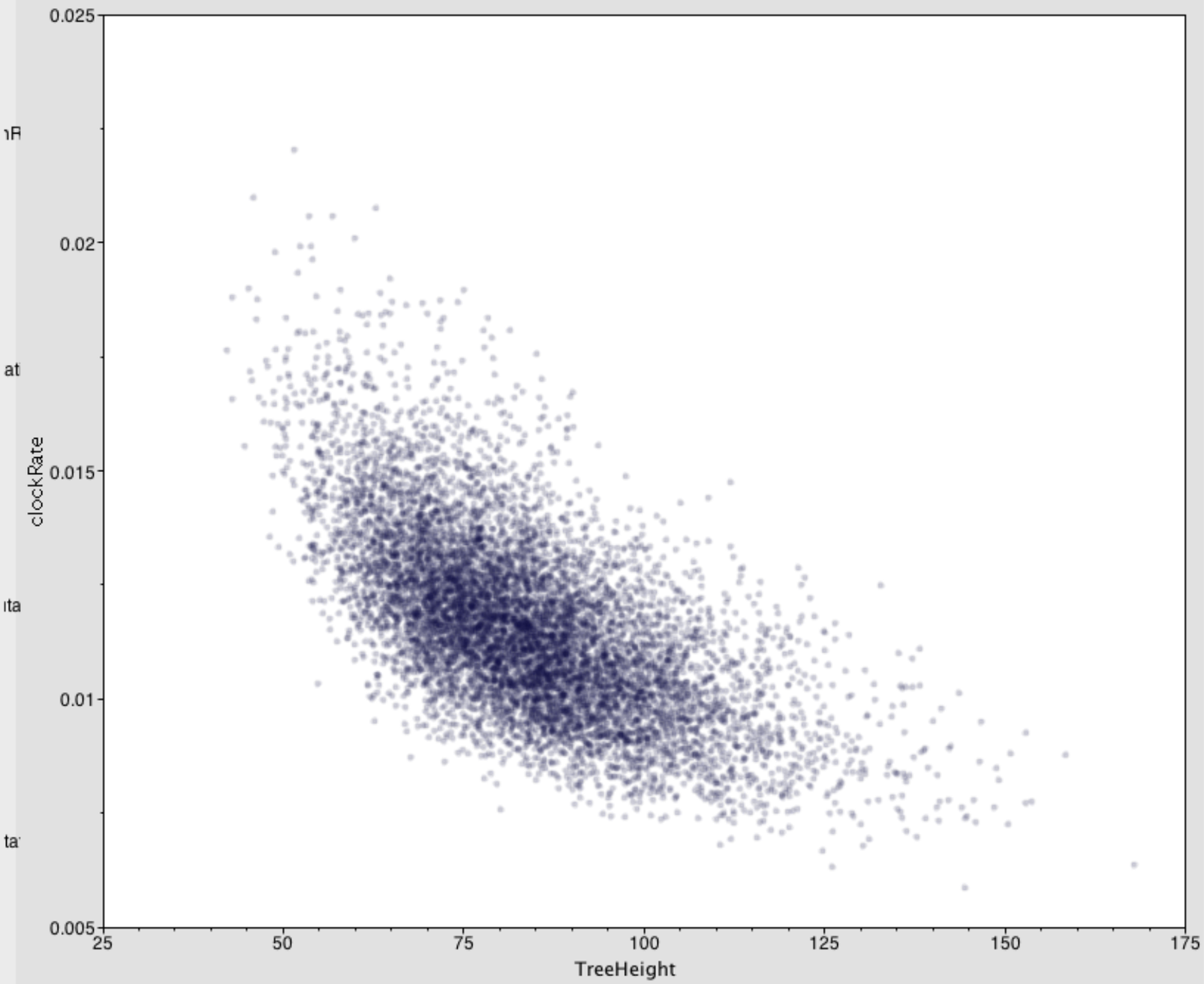
+ − Reload

Traces:

Statistic	Mean	ESS	...
posterior	-5515....	2387	R
likelihood	-5441....	2349	R
prior	-73.169	1379	R
treeLikelihood.1stpos	-1383....	3189	R
treeLikelihood.2ndpos	-952.37	2885	R
treeLikelihood.3rdpos	-2148....	1687	R
treeLikelihood.noncod...	-957.267	1731	R
TreeHeight	83.827	1409	R
mutationRate.1stpos	0.45	852	R
mutationRate.2ndpos	0.182	714	R
mutationRate.3rdpos	2.949	646	R
mutationRate.noncodi...	0.346	1344	R
gammaShape.1stpos	0.496	889	R
gammaShape.2ndpos	0.575	911	R
gammaShape.3rdpos	3.022	726	R
gammaShape.noncodi...	0.244	1006	R
kappa.1stpos	6.235	719	R
kappa.2ndpos	8.5	1359	R
kappa.3rdpos	28.777	365	R
kappa.noncoding	13.478	875	R
CalibratedYuleModel	-47.285	1755	R
birthRateY	2.561E-2	3805	R
logP(mrca(human-chi...	-0.731	9001	R
mrcatime(human-chi...	5.949	8655	R
clockRate	1.161E-2	1836	R

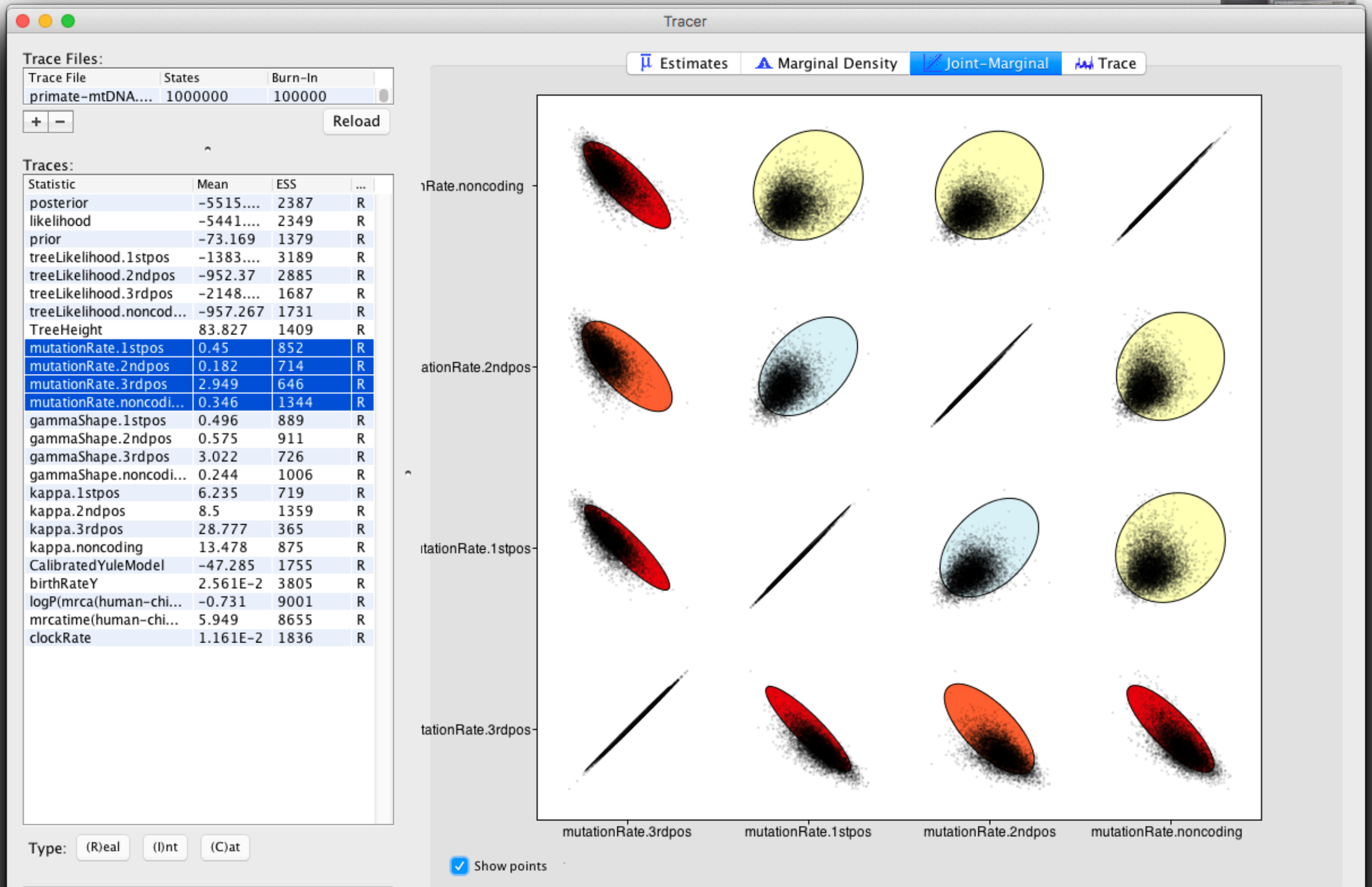
Type: (R)eal (I)nt (C)at

Estimates | Marginal Density | Joint-Marginal | Trace

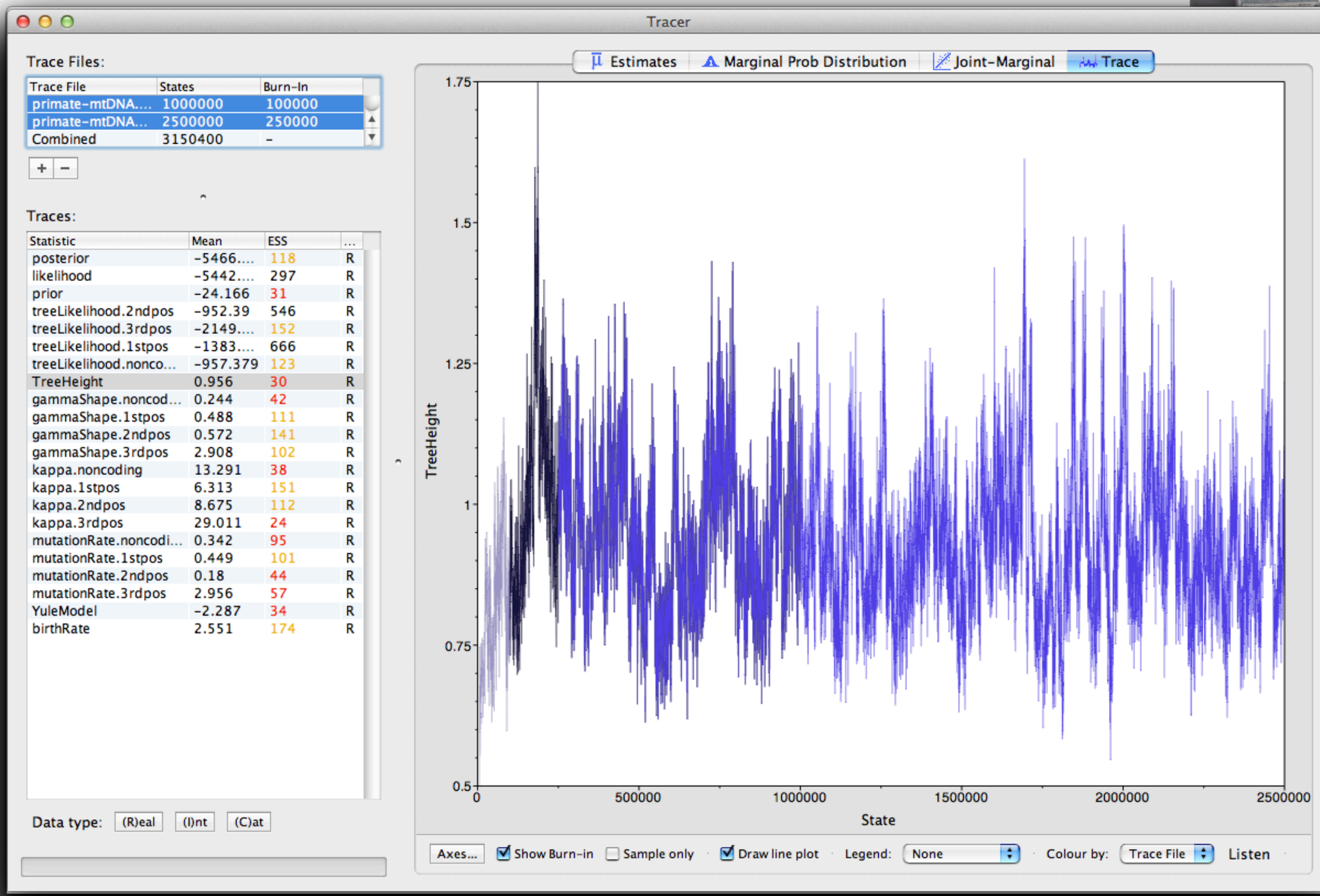


Setup... Sample only Draw as points Use translucency

- Demographic reconstructions

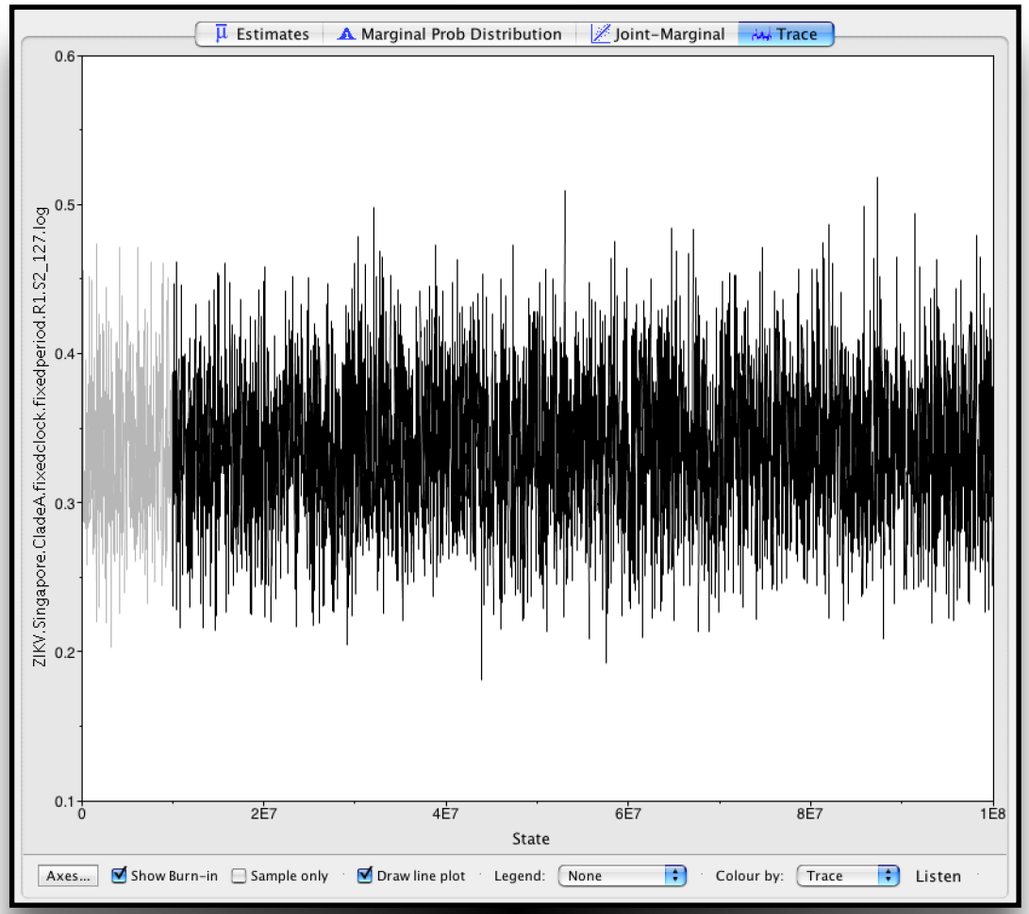


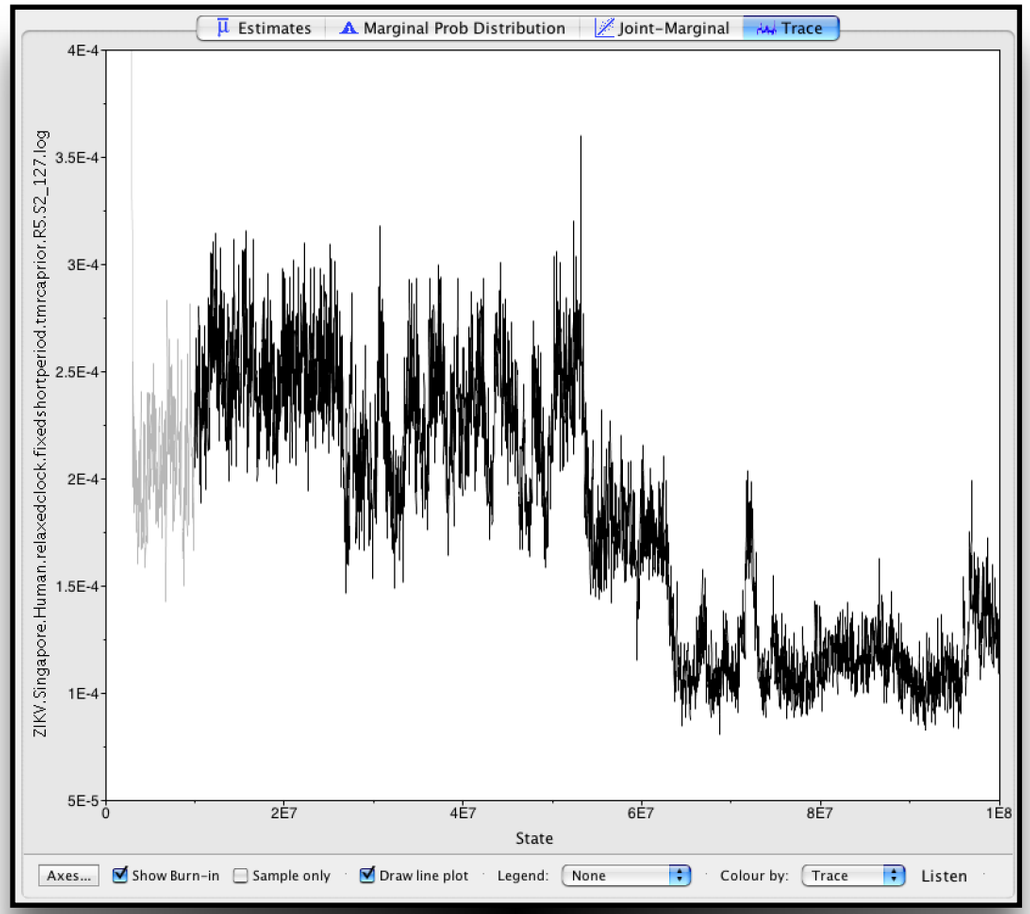
- Demographic reconstructions



- Demographic reconstructions

Look at the chains first!





Mixing well! 😄

Not mixing! 😭

- Demographic reconstructions

TreeAnnotator

(Included with BEAST2)

- Analyse trees file from BEAST2 runs
- Produces single summary tree (MCC) with node annotations (including clade posterior probabilities)
- Positions internal nodes according to average taxon set MRCA times in trees file
- Note that the MCC tree is just a heuristic summary: may produce negative branch lengths when topological uncertainty is large!

Input:

- Tree log file (many trees)

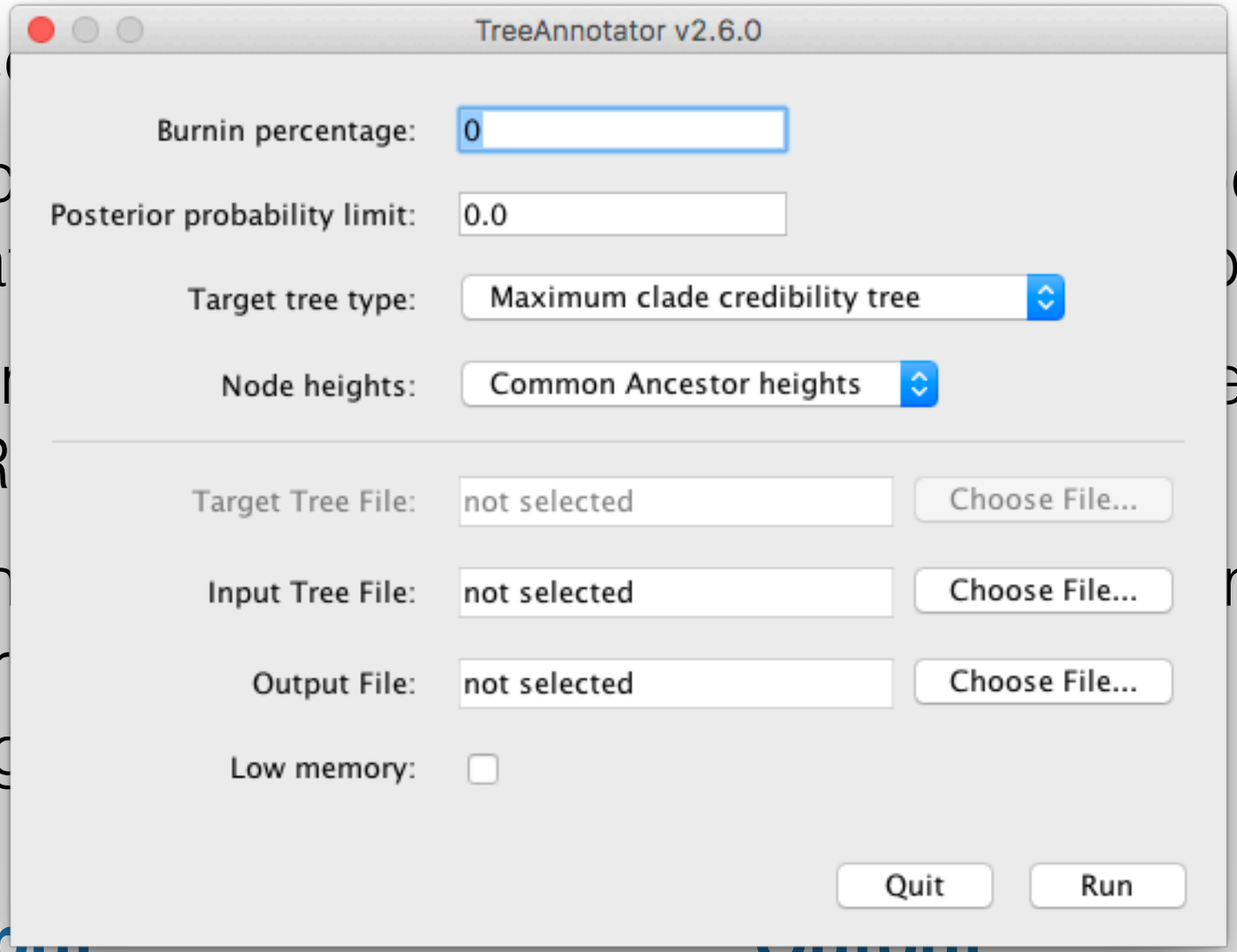
Output:

- MCC tree (one annotated summary tree)

TreeAnnotator

(Included with BEAST2)

- Analys
- Produc … de annota… …ilities)
- Positio… …e taxon set MR…
- Note th… …mary: may pr… topolog…



Input:

- Tree log file (many trees)

Output:

- MCC tree (one annotated summary tree)

FigTree

(http://tree.bio.ed.ac.uk/software/figtree/)

- Visualise trees from BEAST2 runs
- Annotate branches and nodes with probabilities and labels

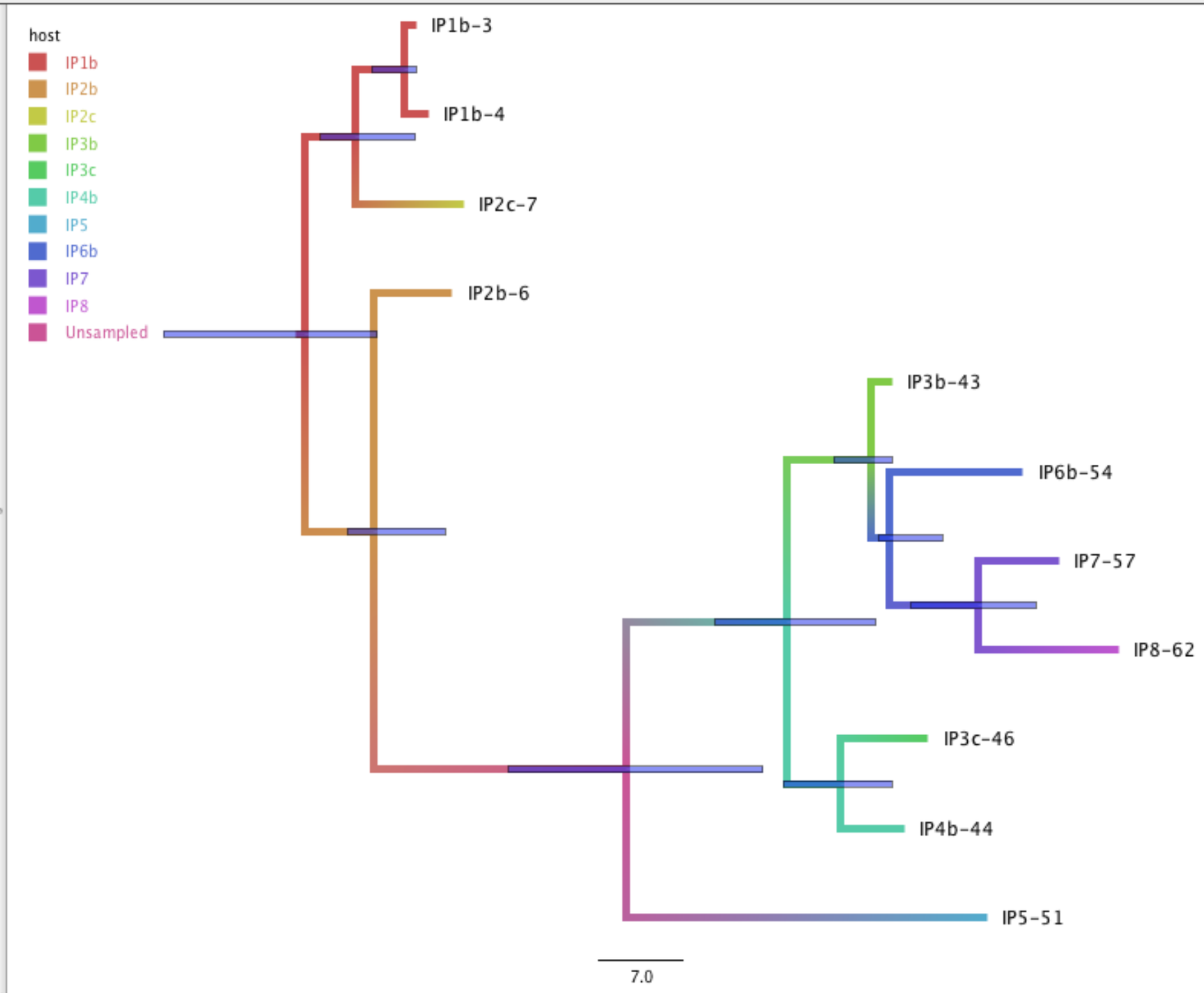
FigTree v1.4.3 - FMDV.MCC.tree

Cartoon Collapse Reroot Rotate Annotate Colour Hilight Find | Node Clade Taxa (Selection Mode) | Prev/Next

Layout

Zoom:
Expansion:
Fish Eye:
Root Length:
Curvature:
Align Tip Labels

Current Tree: 1 / 1
Appearance
Trees
Time Scale
Tip Labels
Tip Shapes
Node Labels
Node Shapes
Node Bars
Branch Labels
Scale Bar
Scale Axis
Legend



Tree: 11 tips [height = 68.4, length = 0E0]

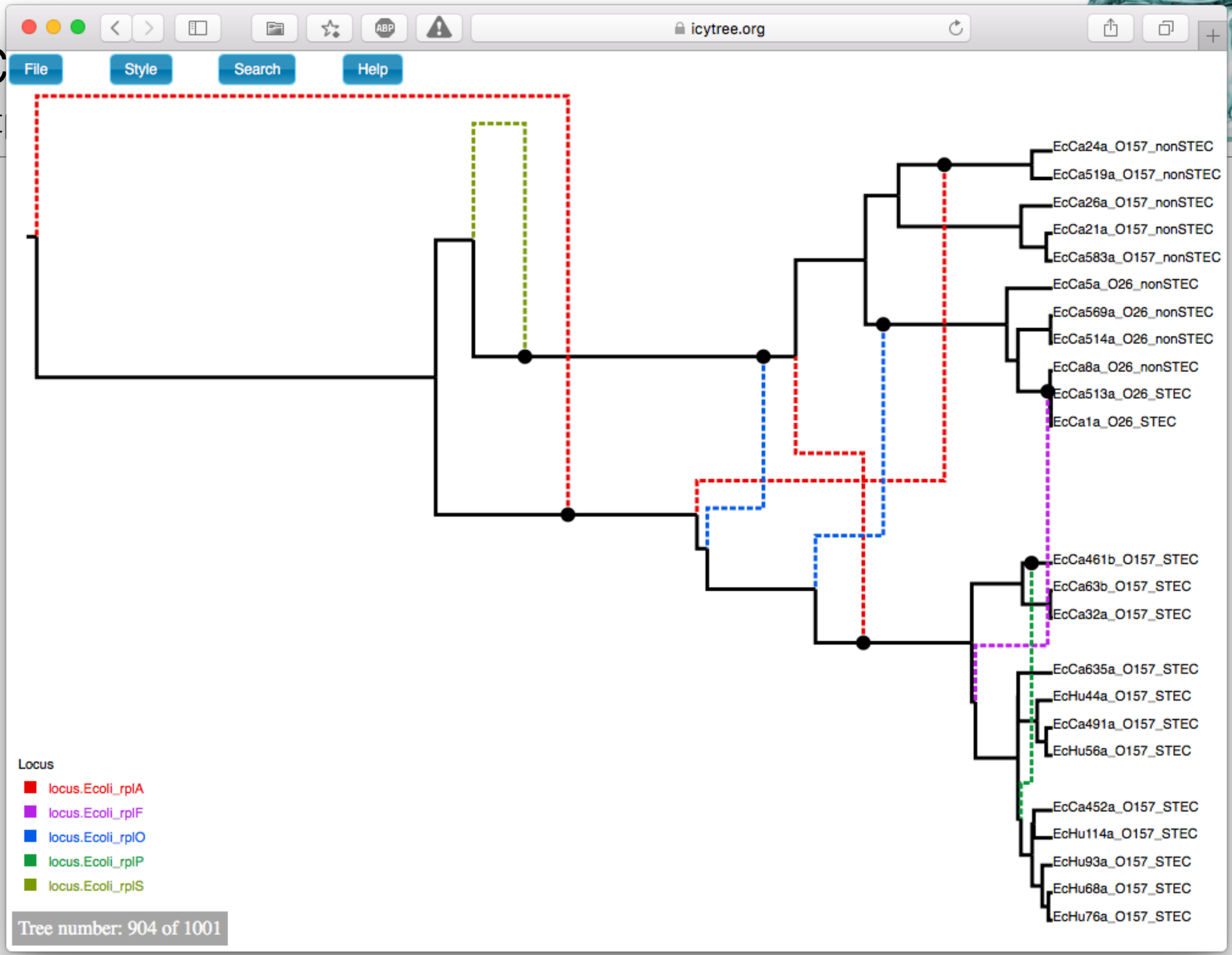
IcyTree

(https://icytree.org)

- Similar to FigTree, but places an emphasis on quick visualisation rather than publication quality output
- Annotate branches and nodes with probabilities and labels
- Better suited for structured models and ancestral recombination graphs (ARGs)
- Faster than FigTree for analysing many trees
- Web app (no installation required)

Ic

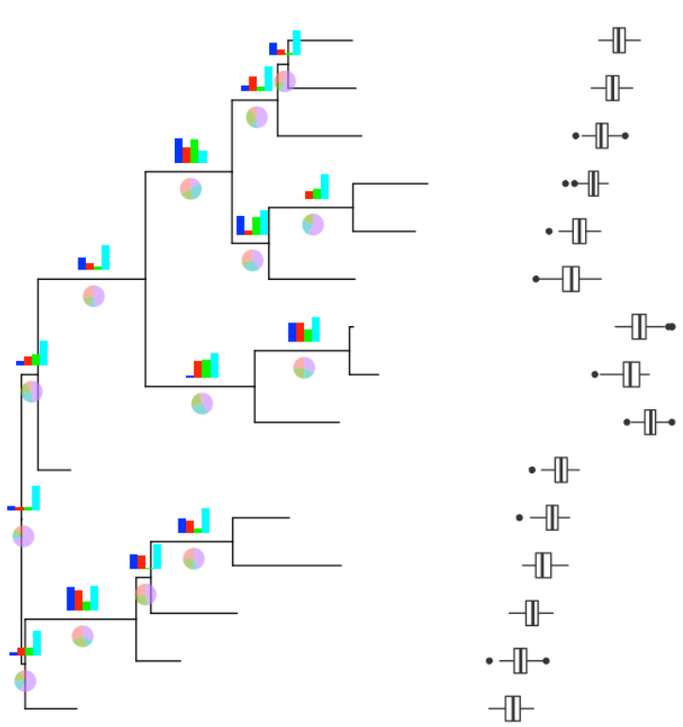
(htt

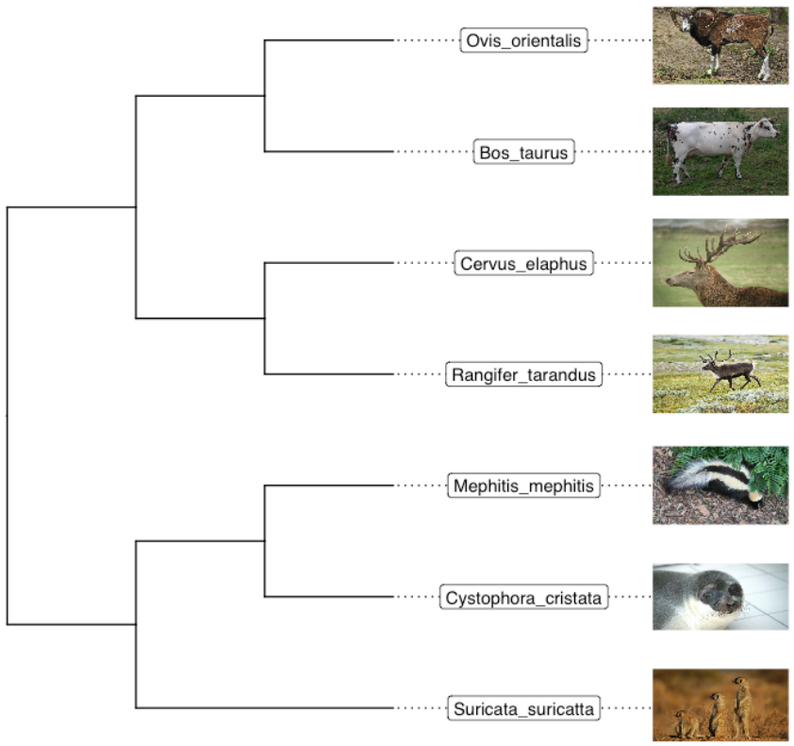

icytree.org
File
Style
Search
Help
EcCa24a_O157_nonSTEC
EcCa519a_O157_nonSTEC
EcCa26a_O157_nonSTEC
EcCa21a_O157_nonSTEC
EcCa583a_O157_nonSTEC
EcCa5a_O26_nonSTEC
EcCa569a_O26_nonSTEC
EcCa514a_O26_nonSTEC
EcCa8a_O26_nonSTEC
EcCa513a_O26_STEC
EcCa1a_O26_STEC
EcCa461b_O157_STEC
EcCa63b_O157_STEC
EcCa32a_O157_STEC
EcCa635a_O157_STEC
EcHu44a_O157_STEC
EcCa491a_O157_STEC
EcHu56a_O157_STEC
EcCa452a_O157_STEC
EcHu114a_O157_STEC
EcHu93a_O157_STEC
EcHu68a_O157_STEC
EcHu76a_O157_STEC
Locus
locus.Ecoli_rplA
locus.Ecoli_rplF
locus.Ecoli_rplO
locus.Ecoli_rplP
locus.Ecoli_rplS
Tree number: 904 of 1001

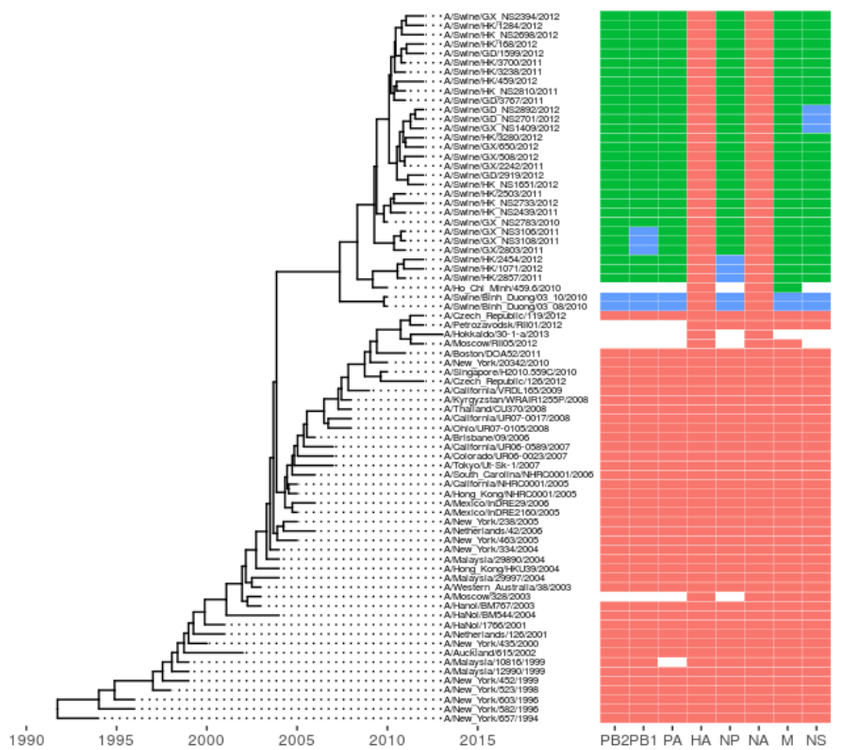
ggtree

(https://guangchuangyu.github.io/software/ggtree/)

- R-package to visualise trees using ggplot grammar
- Works with BEAST2 tree files (and many other packages)
- Can easily annotate trees with other analyses in R







On the program for today

(1) What is phylodynamics?

(2) Bayesian inference recap

(3) BEAST2 introduction

Tutorial: Molecular clock dating (part i)

(4) Molecular clock models

Tutorial: Molecular clock dating (part ii)

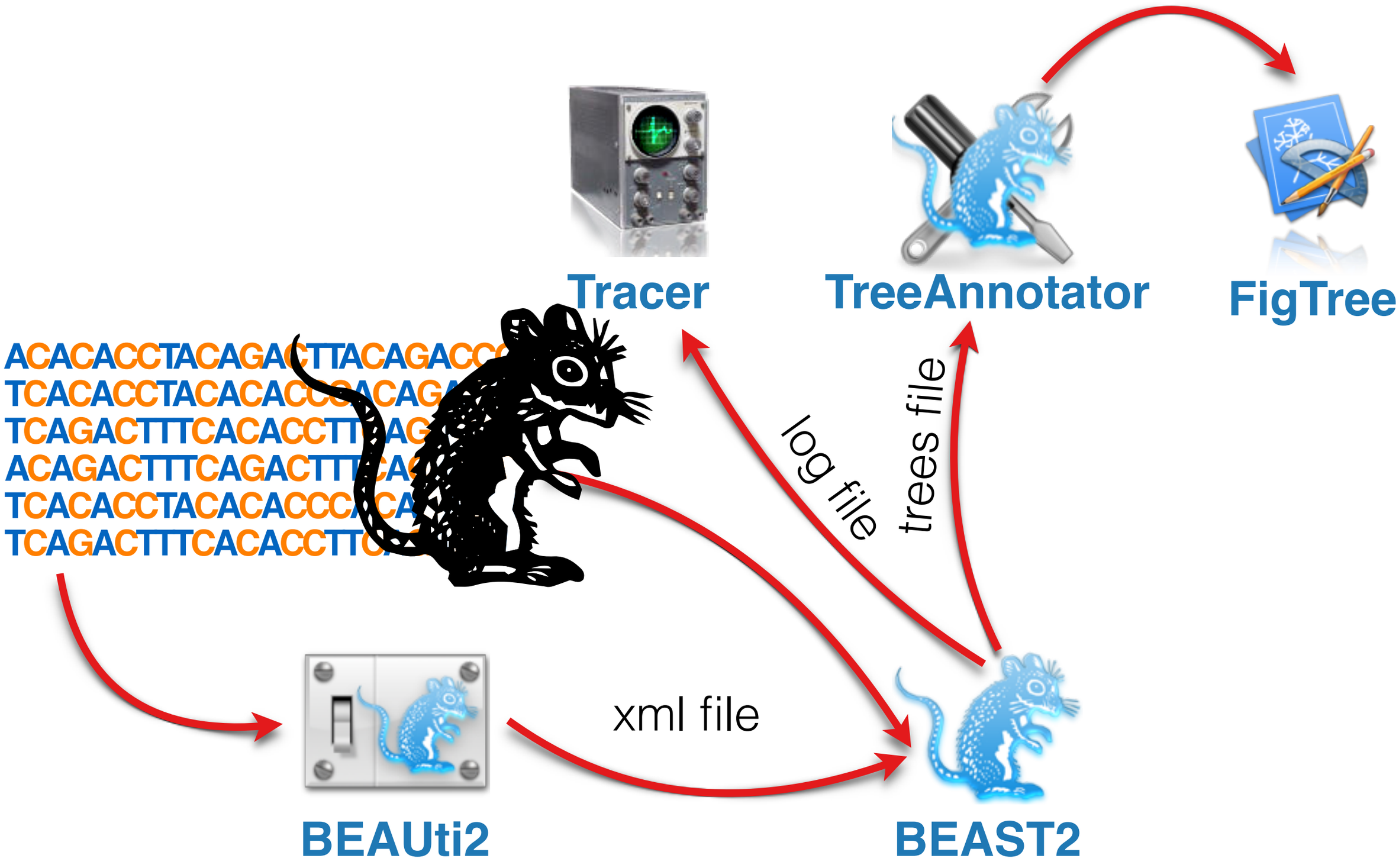
(5) Setting priors

Tutorial: Phylodynamics (part i)

(6) Tree priors

Tutorial: Phylodynamics (part ii)

BEAST2 workflow



Tools of the trade

BEAST2

Software implementing MCMC for model parameter and tree inference

BEAUti2

Part of BEAST2 package for setting up the input file (.xml)

Tracer

Analysis of BEAST output files (.log)

TreeAnnotator

Analysis of BEAST output files (.trees)

FigTree, IcyTree, ggtree

Visualisation of trees (.trees)

BEAST best practice

(This is just a guideline and each analysis is unique)

Before you begin

1) Know your data
   (check alignment, remove outliers, check clock signal etc.)
2) Plan your analysis carefully

Before you run the analysis

3) Ask someone else to look at your XML file
4) Sample from the prior (run without data)

Actually running the analysis

5) Run analysis with multiple chains

After the analysis

6) Combine chains
7) Assess convergence and mixing
8) Ask someone else to look at your log files

BEAST troubleshooting priors

(Common reasons for lack of joy)

My analysis won't start! 😫

P(X = "parameter out of bounds") = 0.9
→ Check initial values fall within prior bounds!

P(X = "numerical underflow") = 0.05
→ Increase number of initialisation attempts
→ Use a different random number seed
→ **Use a better starting tree**

P(X = "package not installed") = 0.02
→ Install the package…

P(X = "something weird") = 0.01
→ Contact developers

BEAST troubleshooting priors

(Common reasons for lack of joy)

My analysis won't start! 😫

P(X = "output files already exist" | new user) = 0.9

→ Check "overwrite existing log and tree files"

P(X = "XML syntax error" | first time editing XML) = 0.9

→ Create template in BEAUti

→ Look at examples from the developers

→ **Look at the source code**

cEvo

virion

Understanding
virus-host interactions

Molecular clock models

Louis du Plessis

ETH zürich

D BSSE

On the program for today

(1) What is phylodynamics?

(2) Bayesian inference recap

(3) BEAST2 introduction

Tutorial: Molecular clock dating (part i)

(4) Molecular clock models

Tutorial: Molecular clock dating (part ii)

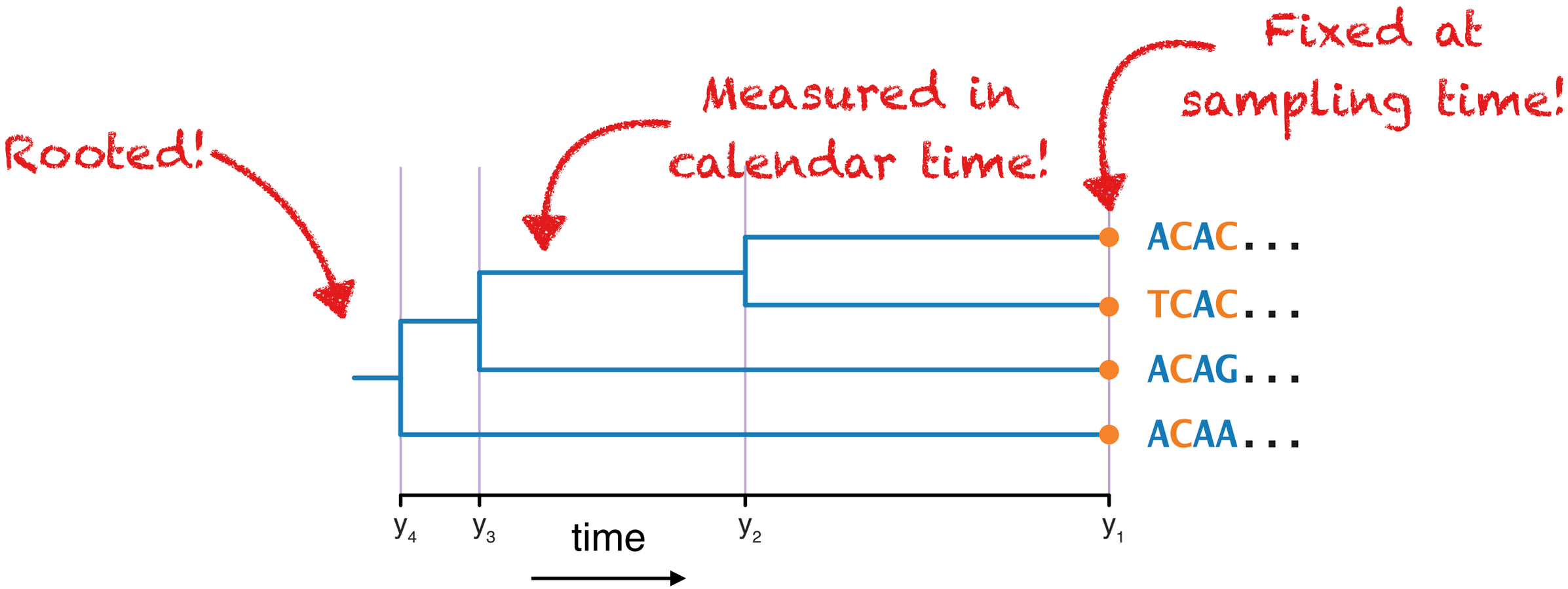
(5) Setting priors

Tutorial: Phylodynamics (part i)

(6) Tree priors

Tutorial: Phylodynamics (part ii)

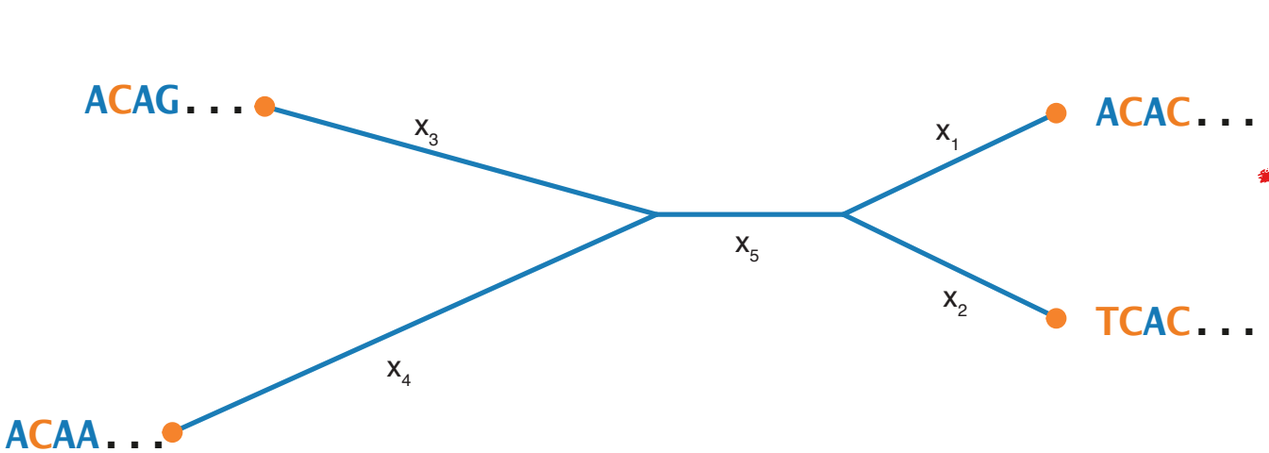
What is a time tree?



- **Rooted** phylogeny (direction of evolution is known)
- Branch lengths are measured in **calendar time** (units of years, months, days)
- Leaves are fixed at **sampling times**

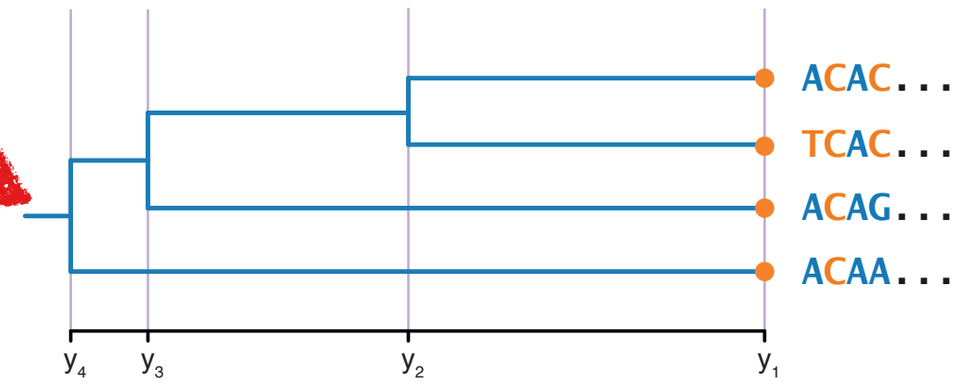
Molecular clock constraint

genetic distance tree
(subst/site)



2n-3 branch lengths

time tree
(years)

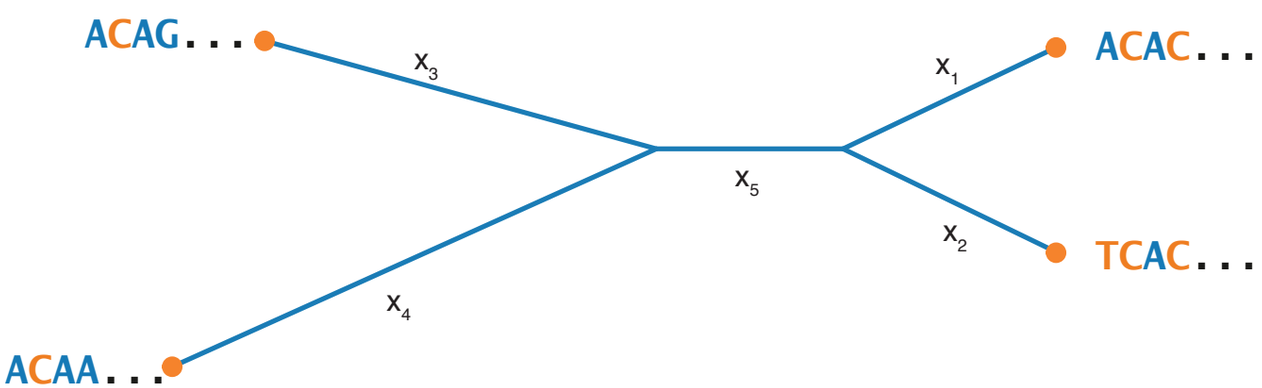


n-1 waiting times

Molecular clock constraint

- Assume the rate of evolution is constant across all branches
- **Linear** accumulation of substitutions over time

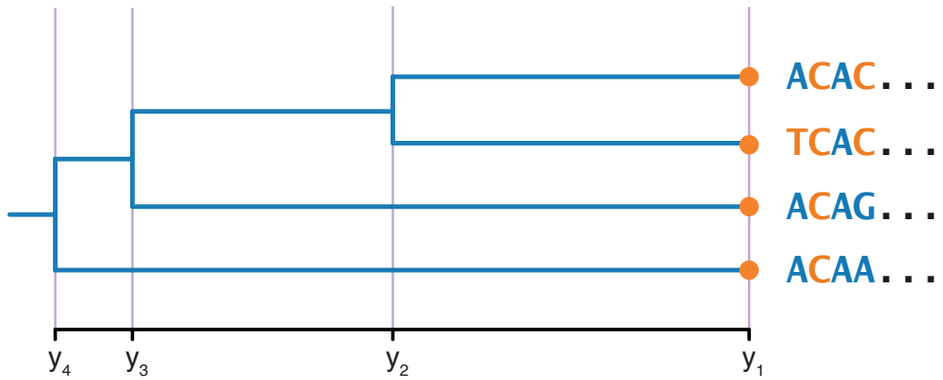
genetic distance tree



- 2n-3 **branch lengths**
- Branch lengths are **independent and identically distributed** among branches
- Root location or direction of evolution is not known

- Topology implies nothing about individual branch lengths

time tree

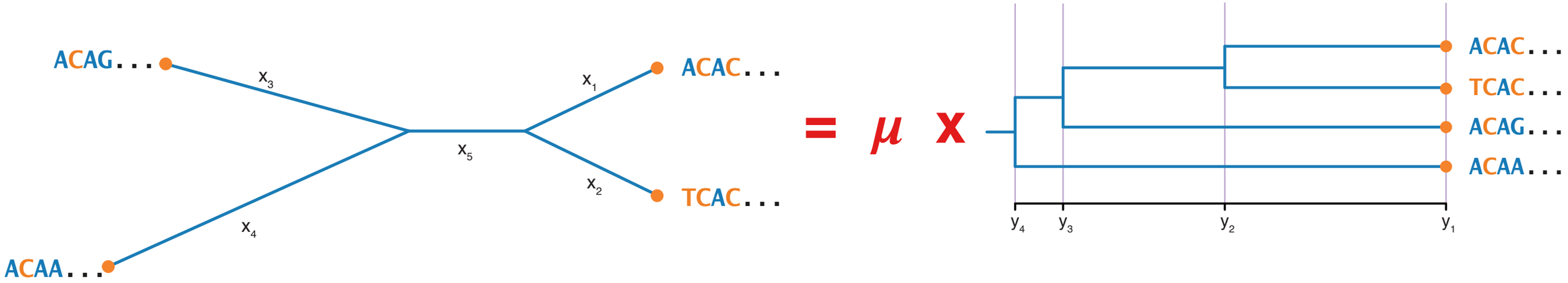


- n-1 **waiting times**
- Rate of evolution is the **same** on all branches
  (strict clock assumption)
- Root of the tree is equidistant from all tips
  (for contemporaneous samples)
- Topology constrains some branch lengths (2 branches in a cherry must be of equal length)

Clock rate: μ

- Rate at which substitutions appear
- Usually measured in substitutions/site/year
- **Genetic distance = Rate x Time**

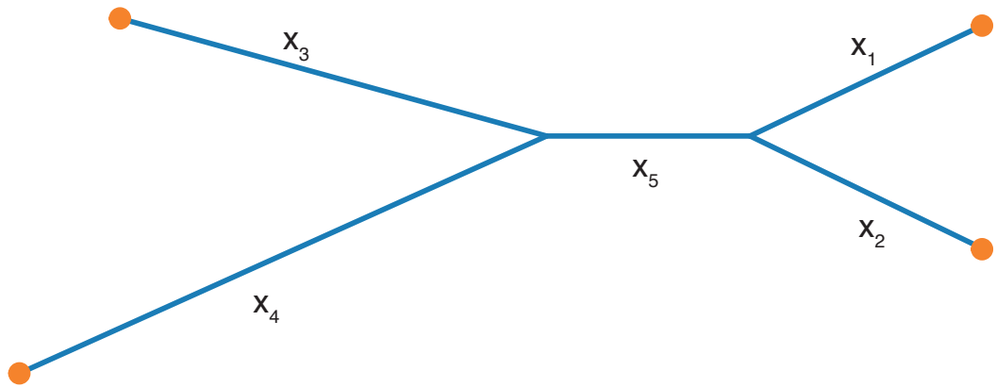
genetic distance tree (subst/site) | **clock rate** (subst/site/year) | **time tree** (years)



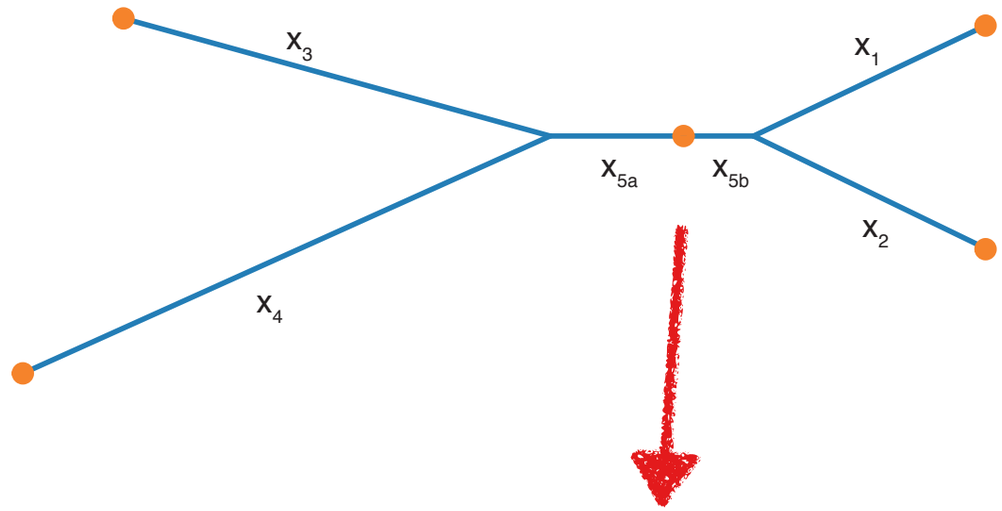
Clock rate converts branch lengths of time tree into genetic distances

Rooting the tree

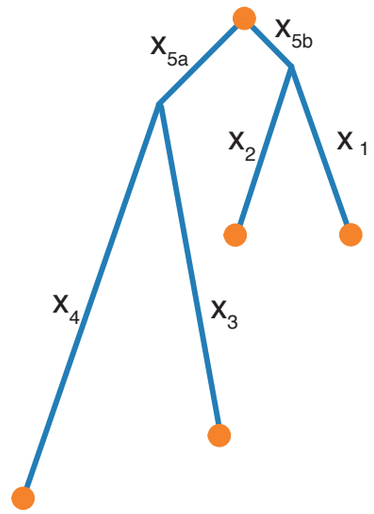
1) Estimate genetic distance tree



2) Pick root location and split branch

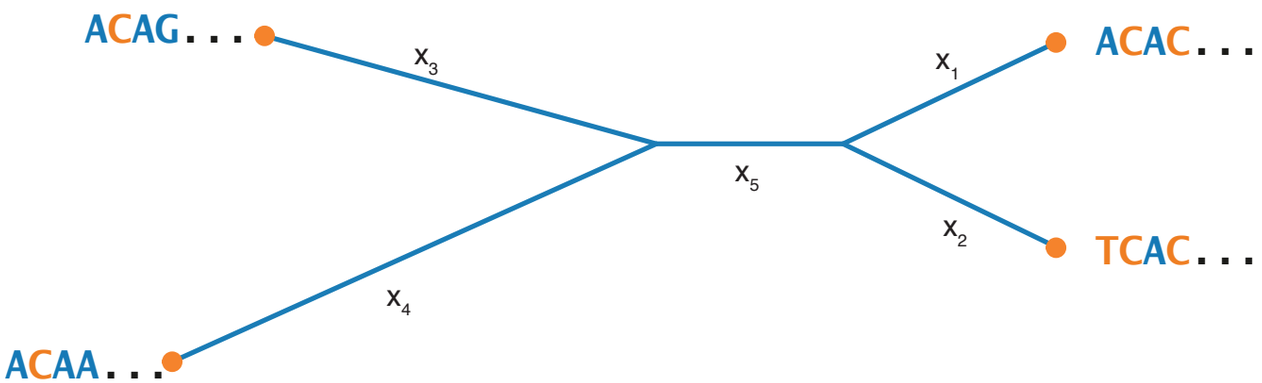


3) PULL!

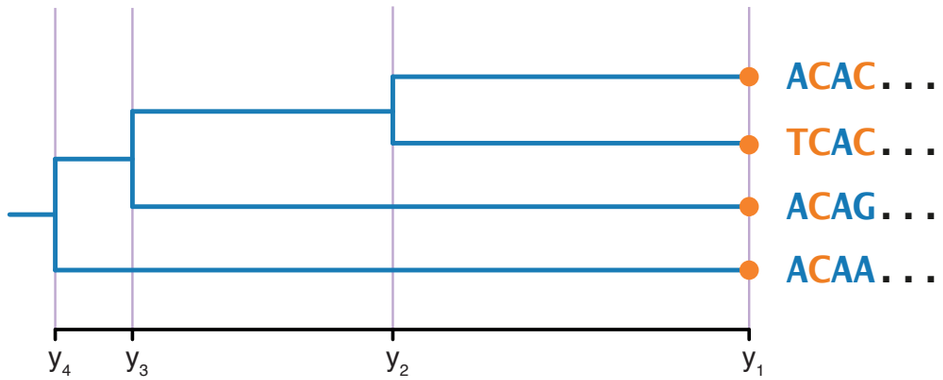


- **The likelihood stays the same!**
  - Substitution model is **reversible**
  - $x_5 = x_{5a} + x_{5b}$
- **Pulley principle:** Root location has no effect on the likelihood

genetic distance tree
(subst/site)



time tree
(years)



if

$$x_1 = \mu(y_2 - y_1)$$

$$x_2 = \mu(y_2 - y_1)$$

$$x_3 = \mu(y_3 - y_1)$$

$$x_4 = \mu[(y_4 - y_1) + (y_4 - y_3)]$$

$$x_5 = \mu(y_3 - y_2)$$

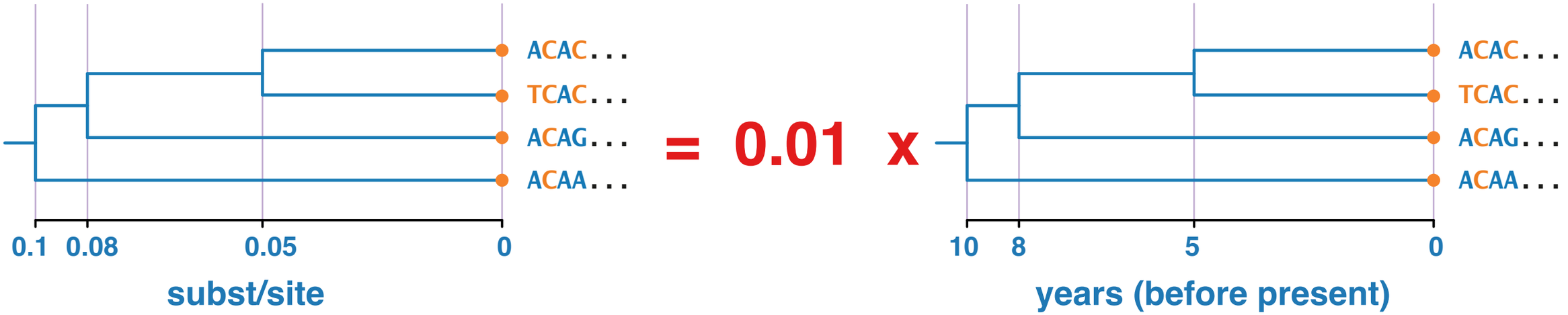
(Pulley principle)

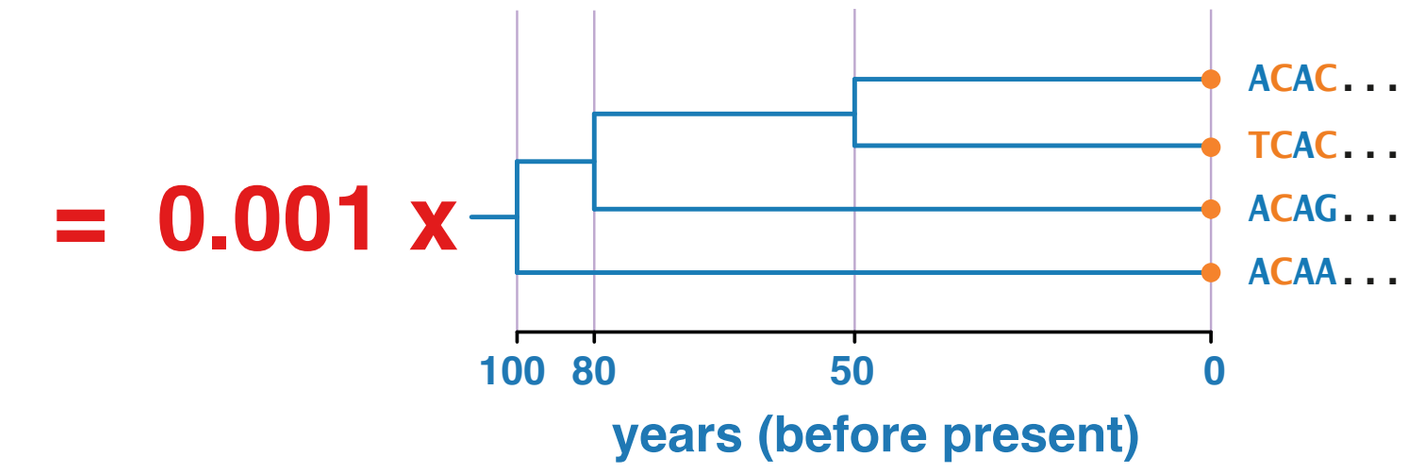
Non-identifiability of rates and times

Genetic distance tree
(subst/site)

Clock rate
(subst/site/year)

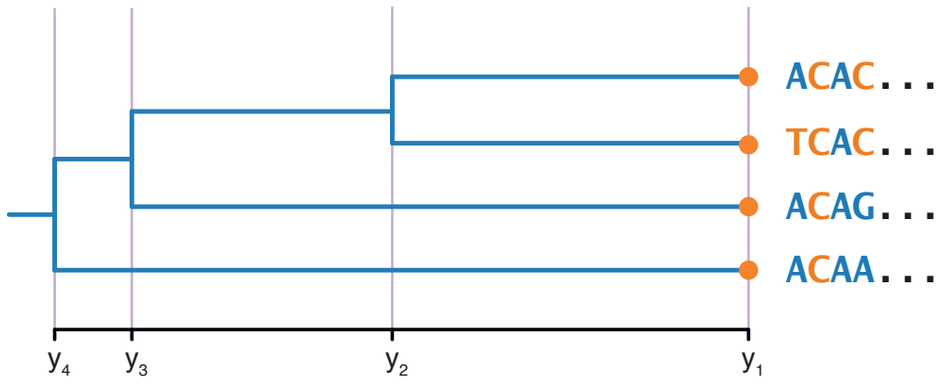
Time tree
(year)





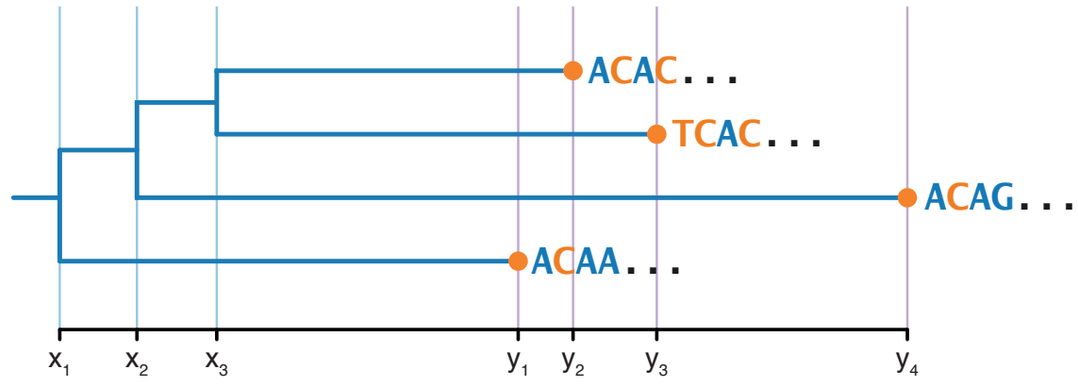
Need extra information to calibrate the clock!

Homochronous data



- All sequences sampled at the same time (contemporaneous)
- Root of the tree is equidistant from all tips
- Non-identifiability of clock rate and time

Heterochronous data

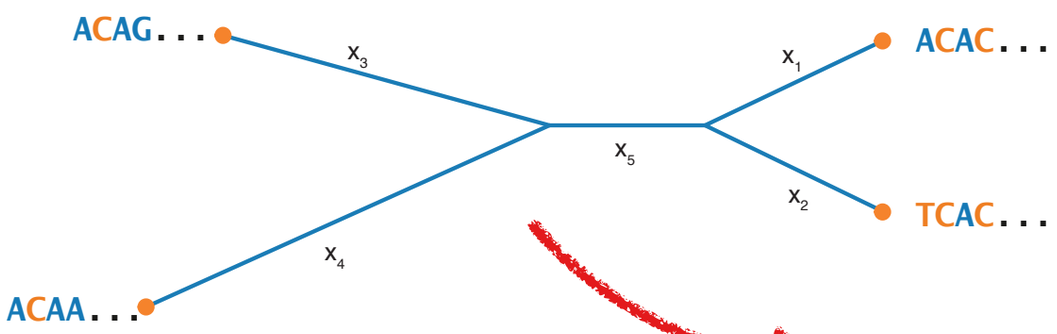


- Sampling period is a significant part of the tree height
- Time to the root is different for each sequence
- **We can use the sampling times to calibrate the clock!**

Calibrating the clock

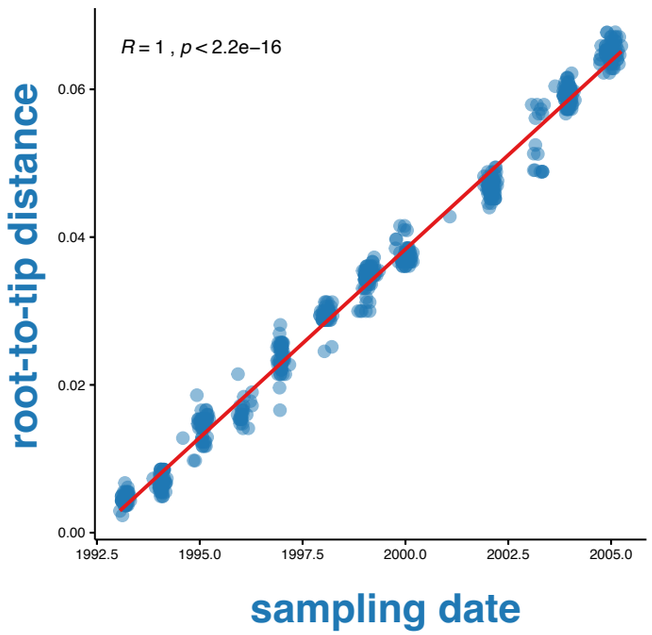
(heterochronous data)

1) Estimate genetic distance tree



2) Pick a root

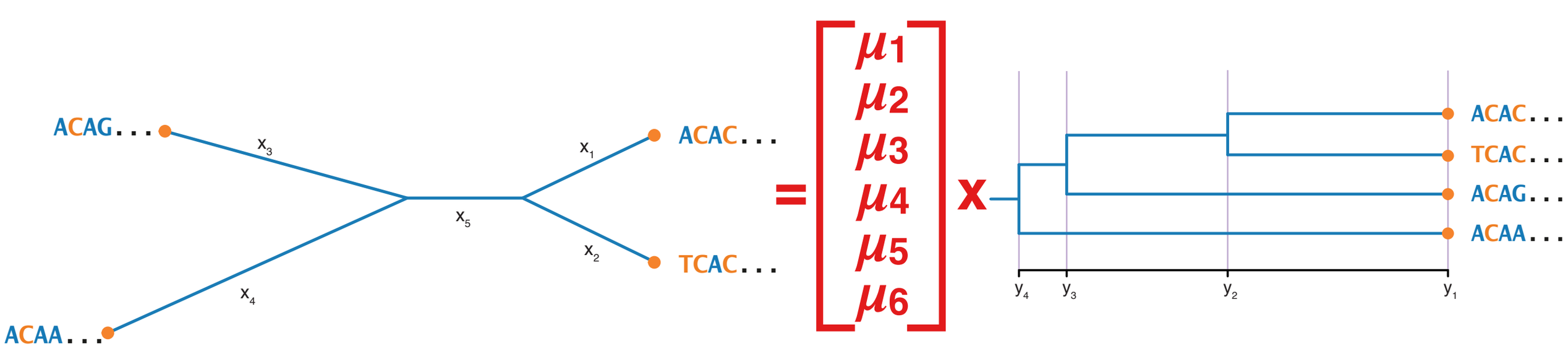
3) Plot root-to-tip distance vs sampling date



- Use linear regression (least squares) to fit a line through the points ($\alpha + \beta t$)
- Clock rate is the slope (β)
- Tree height (tMRCA) is the x-axis intercept ($-\alpha/\beta$)
- Repeat steps **2** and **3** to maximise correlation, R^2 etc. or minimise the residuals (only if the true root is not known)

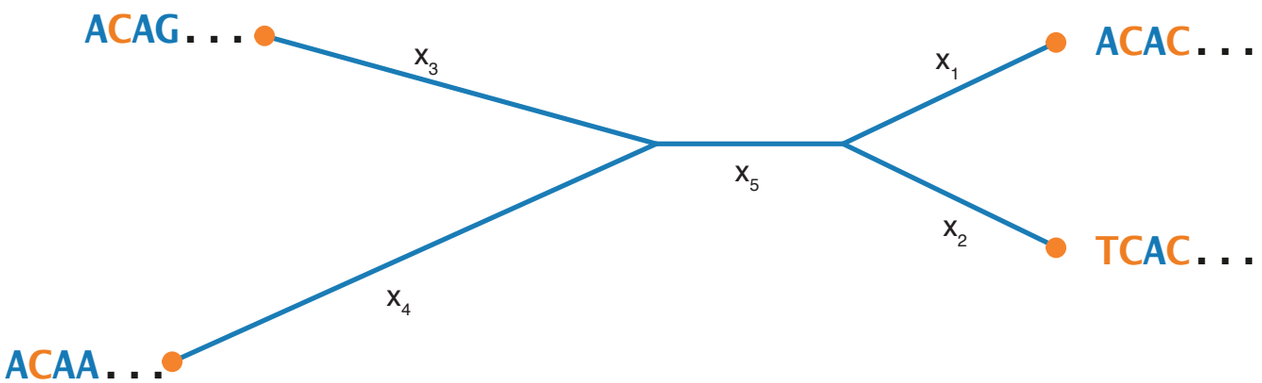
Relaxed clock: $\bar{\mu}$

- Different branches can have different clock rates
- **Genetic distance = Rate x Time**

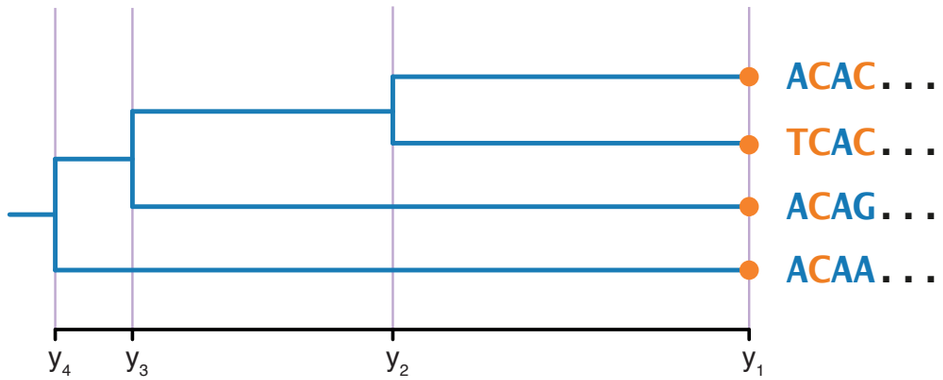


Clock rate converts branch lengths of time tree into genetic distances

genetic distance tree
(subst/site)



time tree
(years)



if

$$x_1 = \bar{\mu}_1(y_2 - y_1)$$
$$x_2 = \bar{\mu}_2(y_2 - y_1)$$
$$x_3 = \bar{\mu}_3(y_3 - y_1)$$
$$x_4 = \bar{\mu}_4(y_4 - y_1) + \bar{\mu}_6(y_4 - y_3)$$
$$x_5 = \bar{\mu}_5(y_3 - y_2)$$

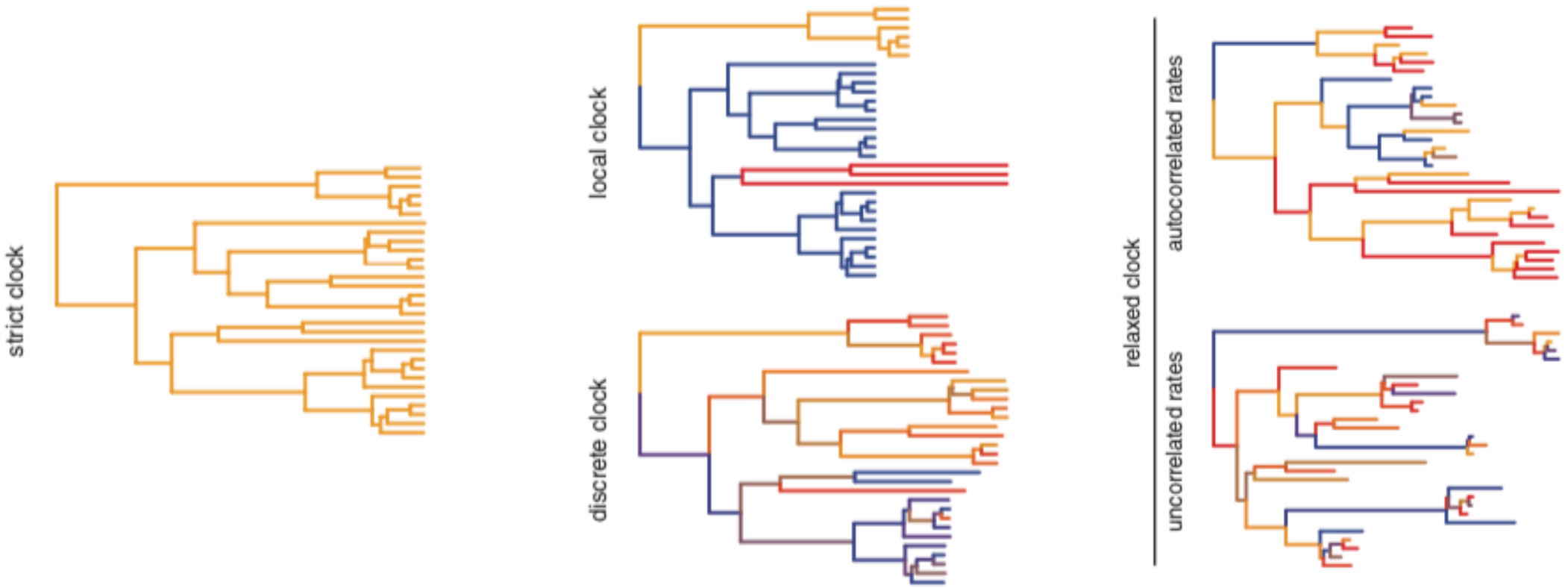
(Pulley principle)

Causes of rate variation

Differences in

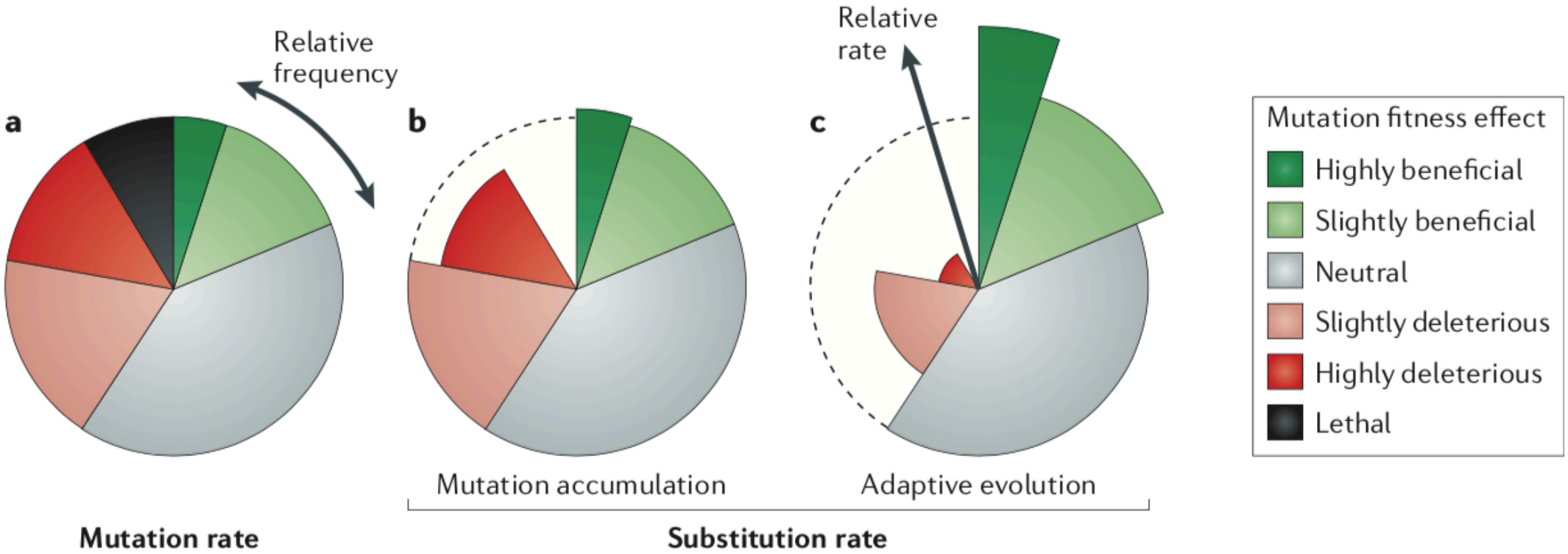
- Generation time
- Metabolic rate
- Error-rate and repair mechanisms
- Population size
- Selective pressure

} **Lineage effects**



Bromham *et al*. ***Biol Rev*** **2018**

What are we actually estimating?



Barrick and Lenski. ***Nat Rev Genetics* 2013**

On the program for today

(1) What is phylodynamics?

(2) Bayesian inference recap

(3) BEAST2 introduction

Tutorial: Molecular clock dating (part i)

(4) Molecular clock models

Tutorial: Molecular clock dating (part ii)

(5) Setting priors

Tutorial: Phylodynamics (part i)

(6) Tree priors

Tutorial: Phylodynamics (part ii)

cEvo

virion

01110011 0110110 0110100 01110010 01110101

Understanding
virus-host interactions

Setting priors

Louis du Plessis

ETH zürich

D BSSE

On the program for today

(1) What is phylodynamics?

(2) Bayesian inference recap

(3) BEAST2 introduction

Tutorial: Molecular clock dating (part i)

(4) Molecular clock models

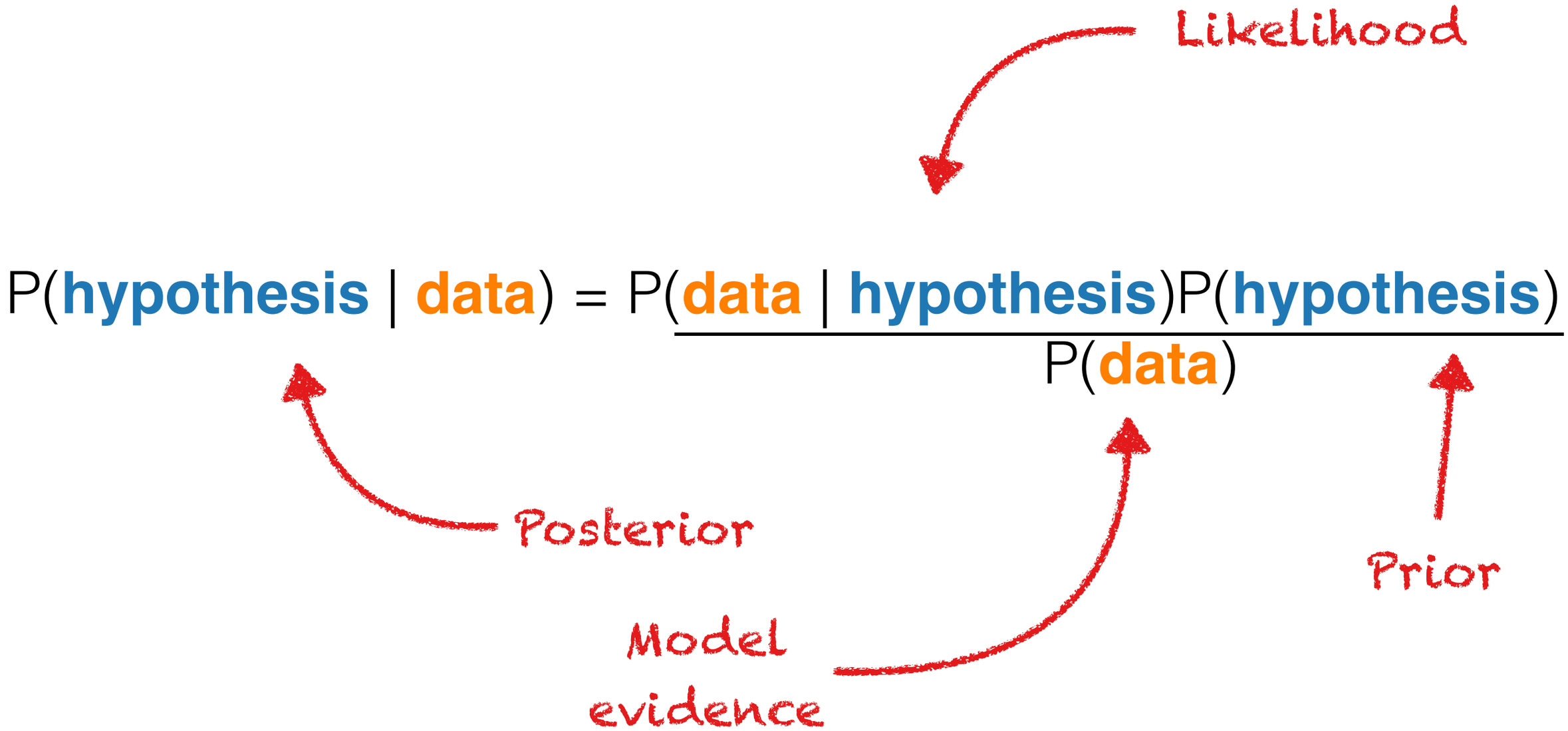
Tutorial: Molecular clock dating (part ii)

(5) Setting priors

Tutorial: Phylodynamics (part i)

(6) Tree priors

Tutorial: Phylodynamics (part ii)



Prior → P(hypothesis)

- Original probability for the model parameters/components (before data collected or analysed)
- Belief in our hypothesis
- All parameters have priors, whether you specify them or not!

What is a **prior** probability?

The probability of whatever we are interested in, in the absence of possibly relevant data

- Usually define a **prior distributions** for model **parameters**
  e.g. clock rate, tree, …
- Often use **parametric distributions**
  e.g. Uniform, Normal, Gamma, Beta, Lognormal, Laplace, …
- Sometimes a **prior** on a model component
  e.g. substitution model (HKY, GTR, JC, …)
- Priors can have priors which can have priors *ad infinitum* (**hyperpriors**)
- Parameter bounds are part of the prior
  e.g. normal distribution with lower bound 0.

Improper **priors**

Suppose we want to pick a prior for the molecular clock rate, μ, and we want to use an uninformative prior

- What about a uniform prior (constant)? $f(\mu) = c$

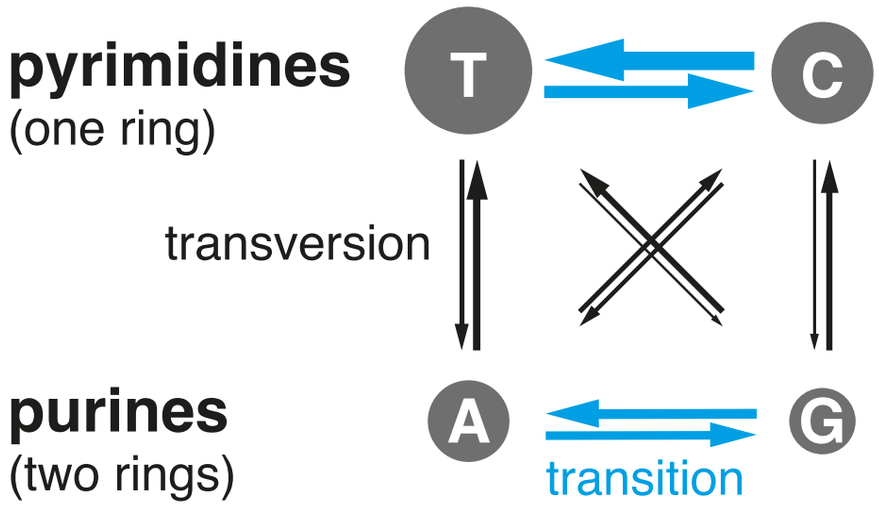
  Is this uninformative?

  - But this is defined from 0 to infinity!
  - Places almost all probability on very large (and biologically unrealistic) values!
  - There is no normalising constant such that it integrates to 1!
- f(μ) = 1/μ is a better choice for rates (uniform in log space)
  - But it's also impossible to find a normalising constant!
- Probably better to use something like a lognormal distribution
  - Always >0, long tail, integrates to 1

How to pick a **prior**?

- Should be chosen based on your beliefs about model parameters (from independent evidence/experiments)
- Should be chosen with a particular analysis in mind (no priors are universal!)
- Think carefully about plausible ranges of parameters
- Be conservative if you are unsure about the parameter (use diffuse priors)
- Conjugate priors? (no — those are just for mathematical convenience)
- Reference priors? (yes — if you are an objective Bayesian)
- Do not use improper priors if you can help it (priors that integrate to infinity, e.g. $1/x$)

Example: HKY-model (HKY85)



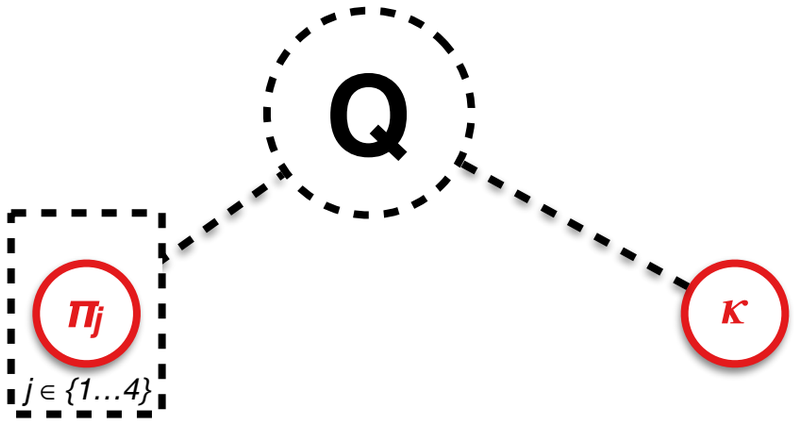
$$(\pi_T, \pi_C, \pi_A, \pi_G)$$

$$\begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} \cdot & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & \cdot & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & \cdot & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & \cdot \end{pmatrix} \end{array}$$

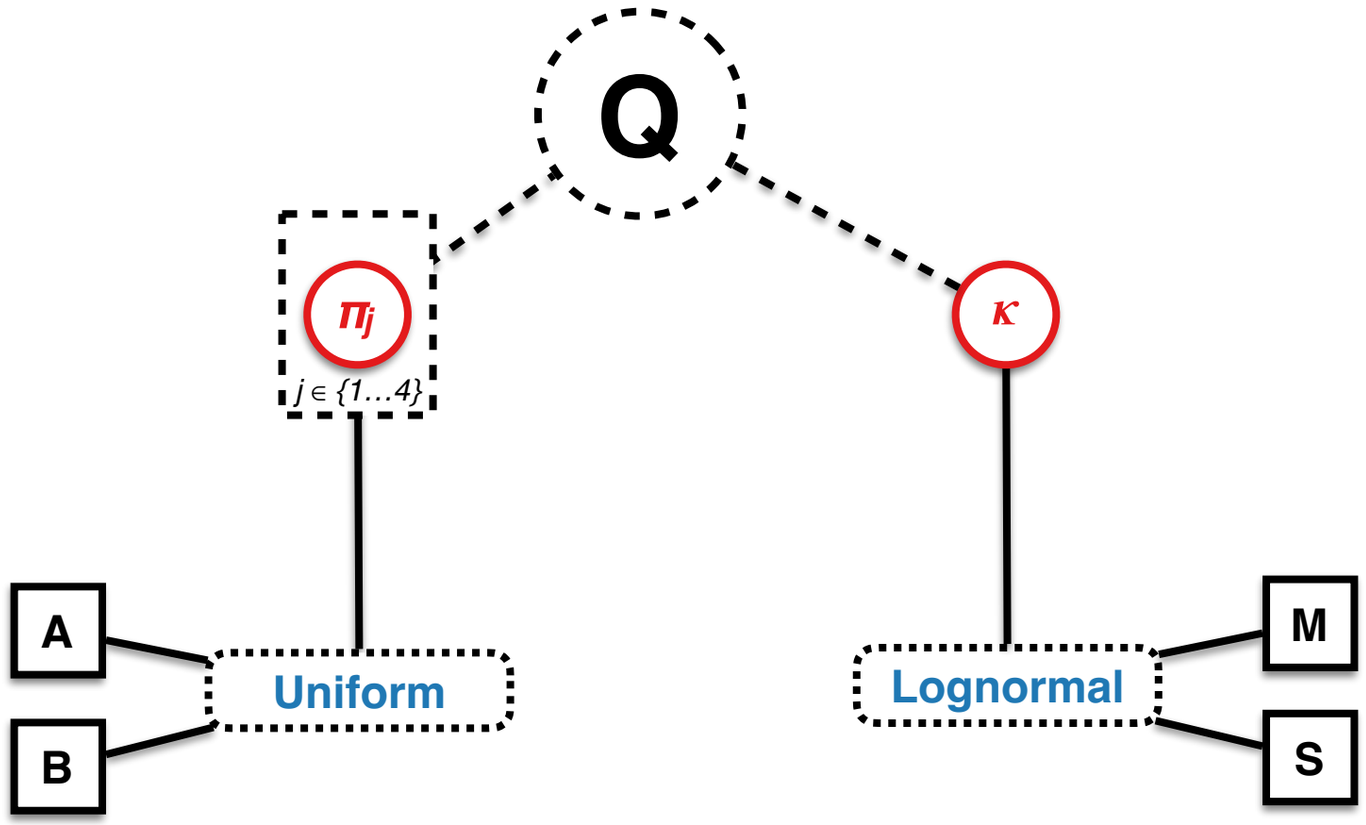
$$= \begin{pmatrix} \cdot & \alpha & \beta & \beta \\ \alpha & \cdot & \beta & \beta \\ \beta & \beta & \cdot & \alpha \\ \beta & \beta & \alpha & \cdot \end{pmatrix} \cdot \begin{pmatrix} \pi_T & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_A & 0 \\ 0 & 0 & 0 & \pi_G \end{pmatrix}$$

- **Q** matrix gives relative rates of substitution between nucleotides
- 5 parameters:
  - $\kappa = \alpha/\beta$
  - $\pi_T, \pi_C, \pi_A, \pi_G$ — assume these all have the same prior probability

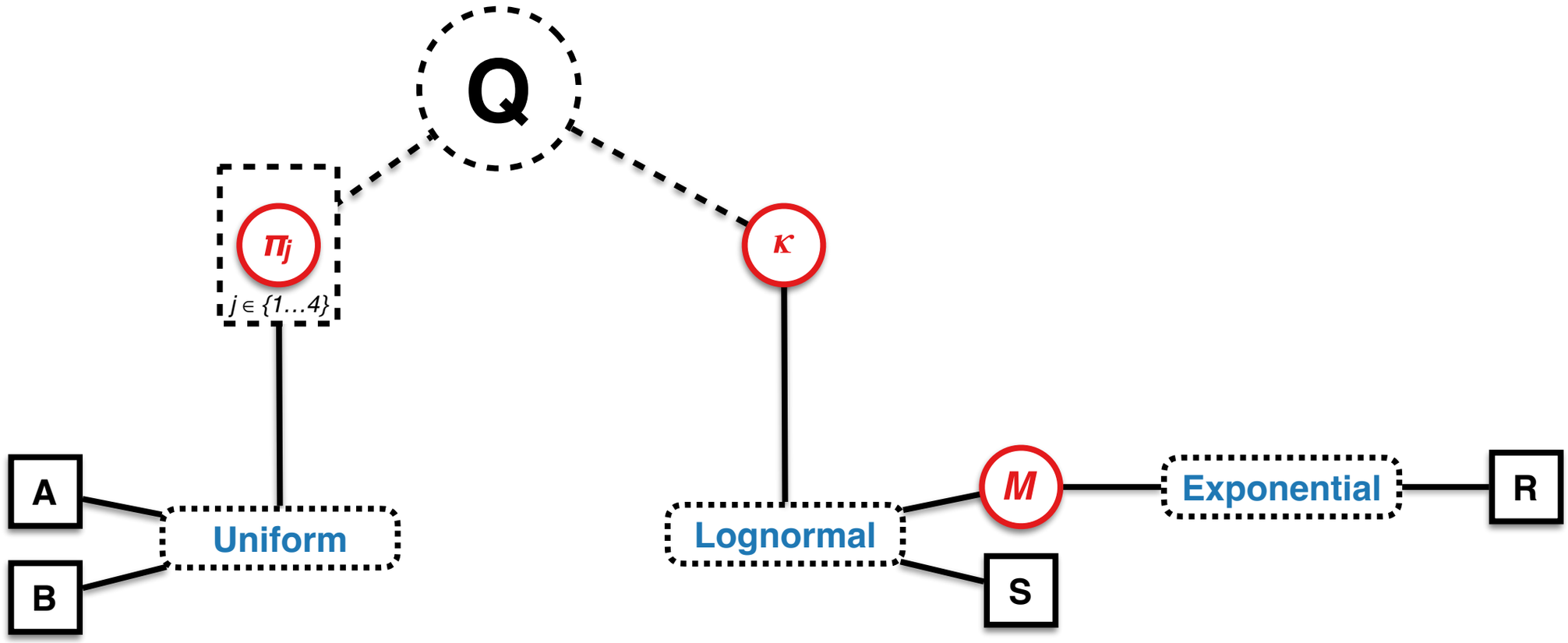
Example: HKY-model (HKY85)



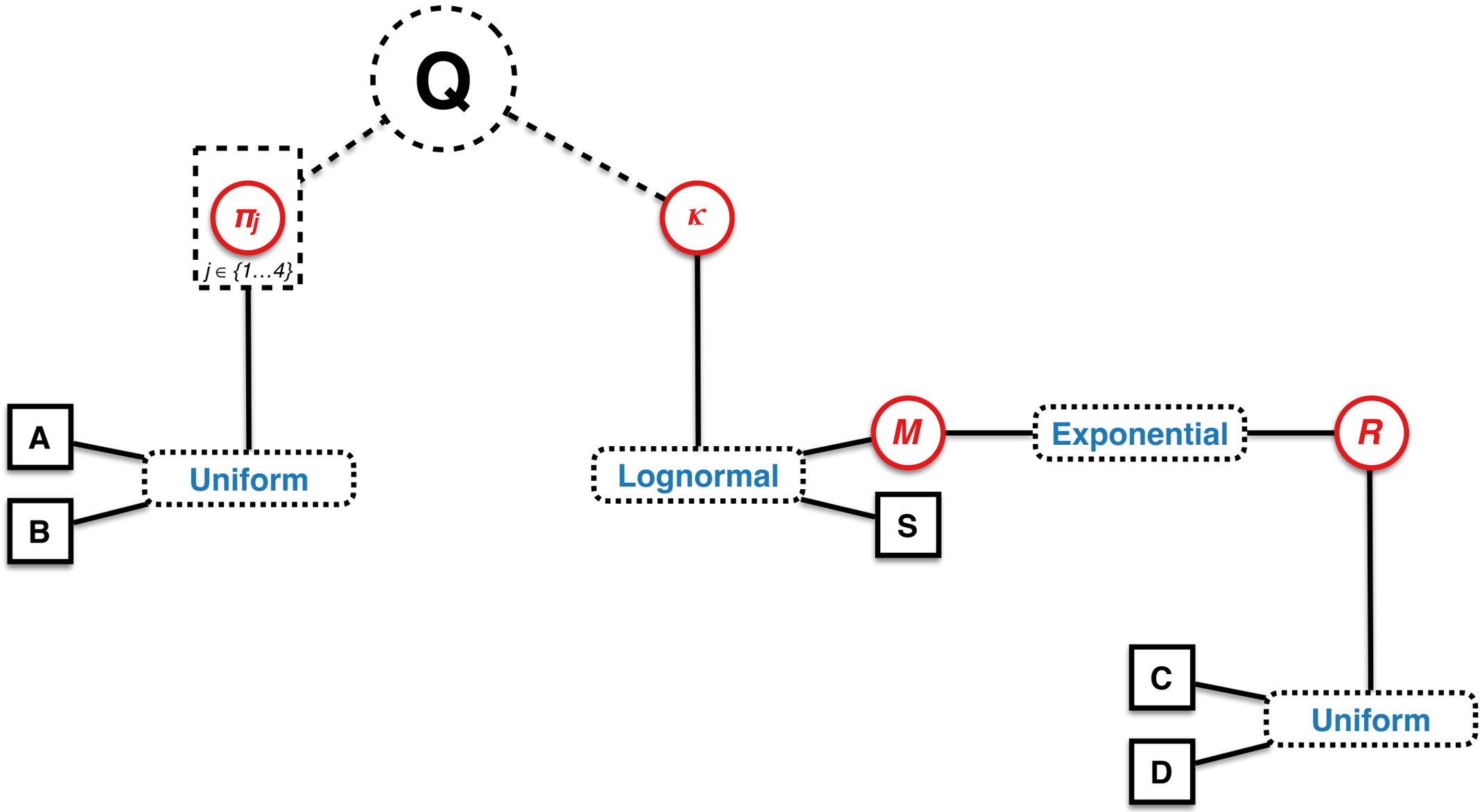
Example: HKY-model (HKY85)



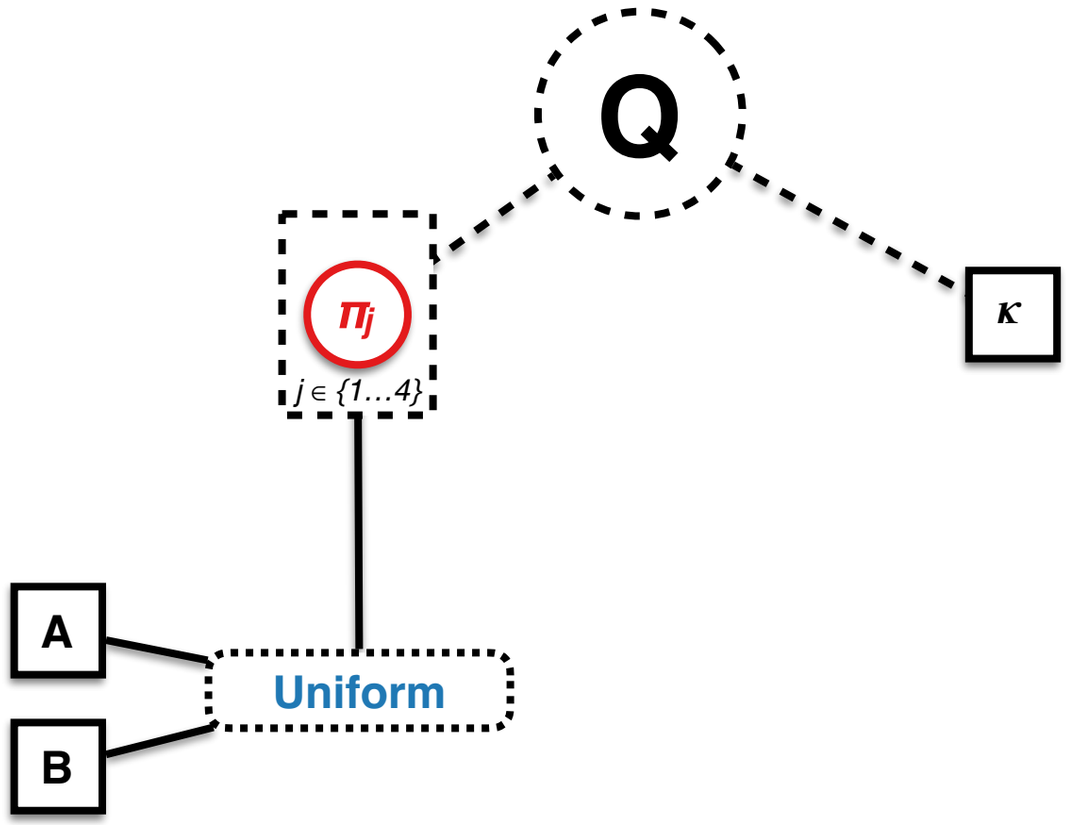
Example: HKY-model (HKY85)



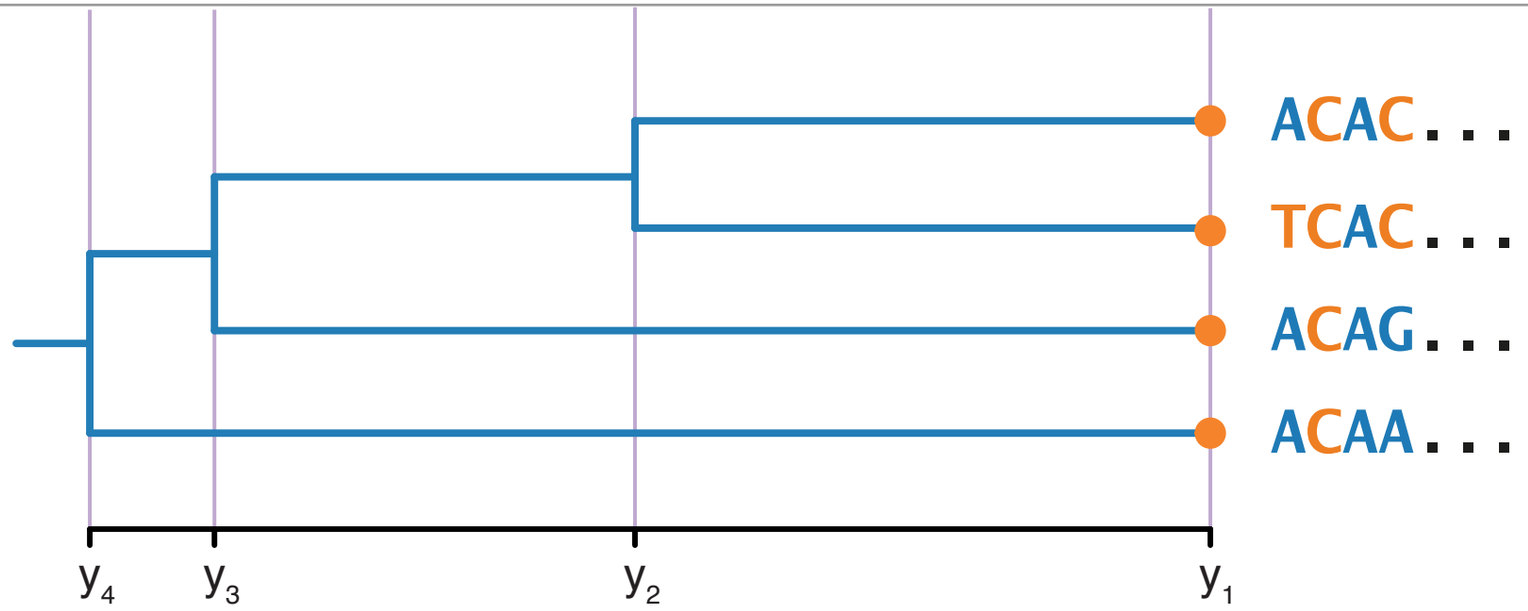
Example: HKY-model (HKY85)



Example: HKY-model (HKY85)



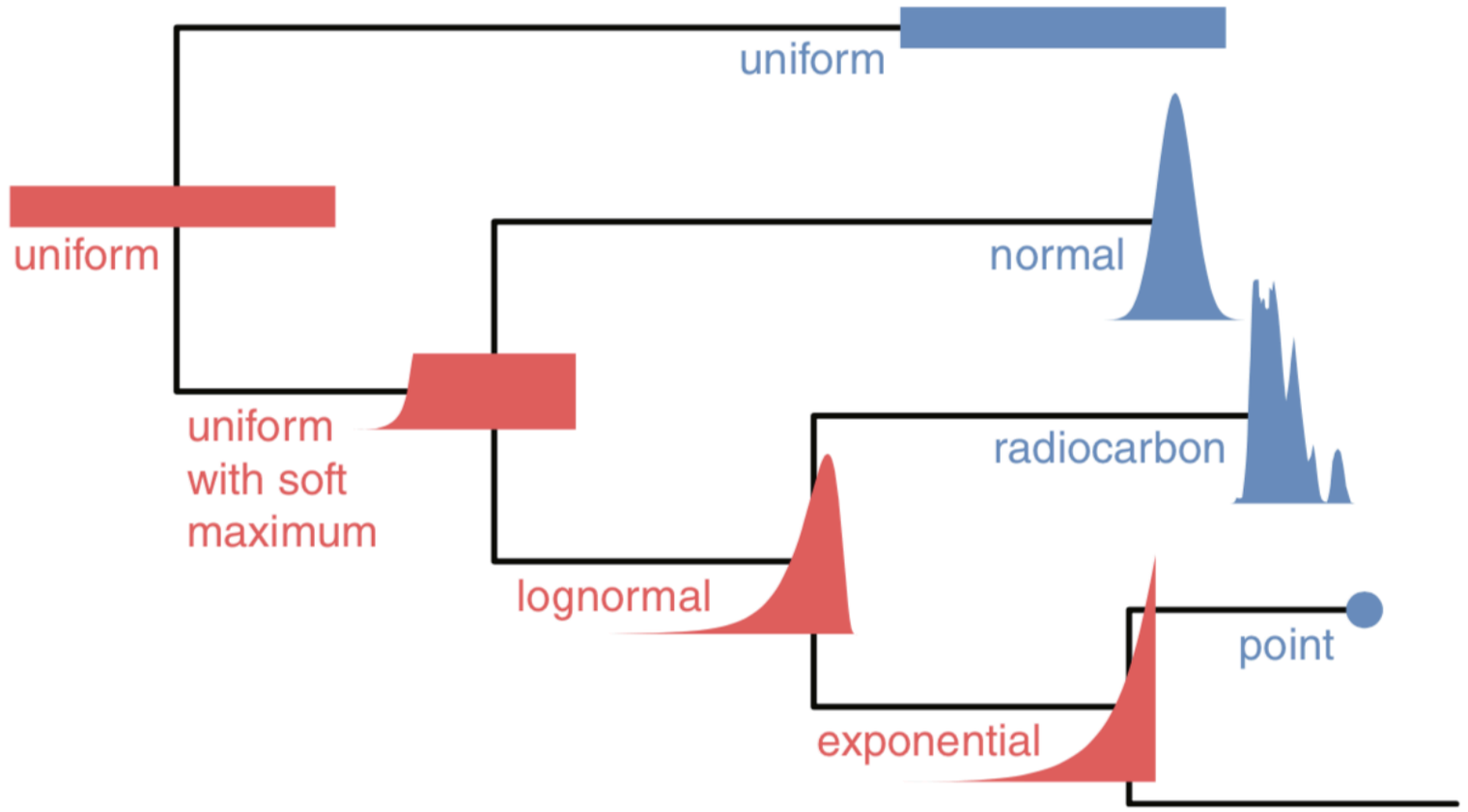
Calibration nodes



Homochronous trees need external **(prior)** information to date:

- Fix the clock rate
- Use a **calibration node** (or nodes)

Calibration nodes

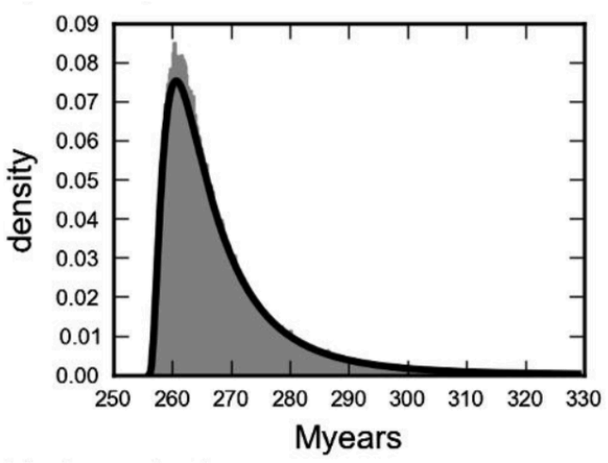


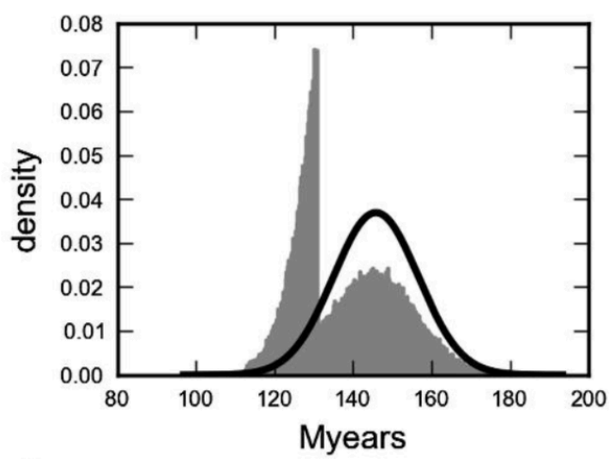
Bromham *et al.* **2018**

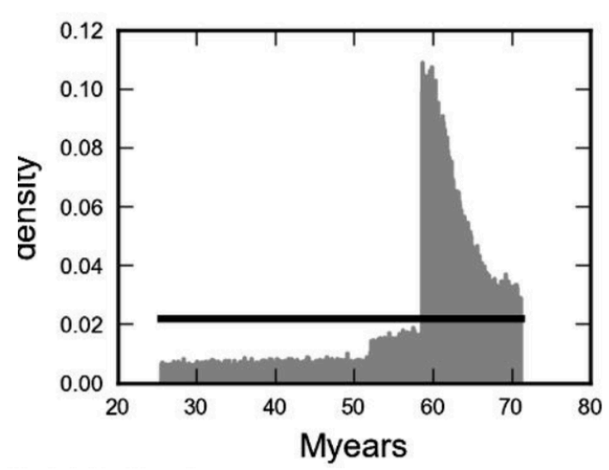
Calibration nodes

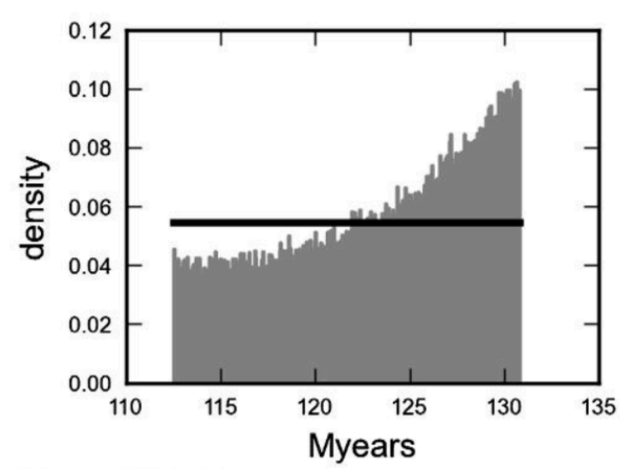
More calibration nodes mean more information to calibrate the clock

but is more always better?





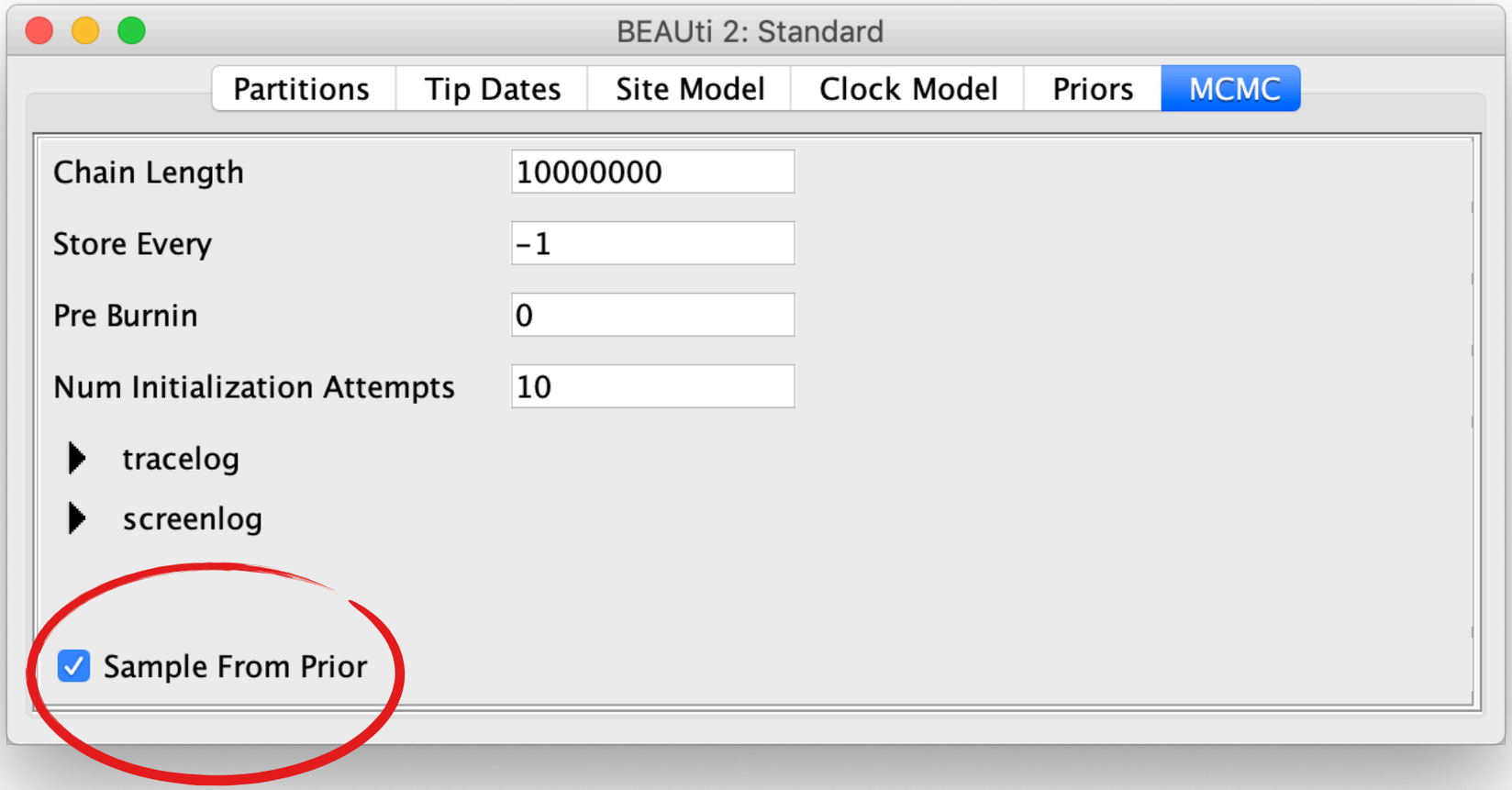




Induced prior may be different to the prior you set!

Heled and Drummond **2012**

Setting priors best practice



- Sample or simulate from prior
- For key parameters plot prior and posterior together
- Mess with priors and see how sensitive/robust posterior is to prior perturbations

On the program for today

(1) What is phylodynamics?

(2) Bayesian inference recap

(3) BEAST2 introduction

Tutorial: Molecular clock dating (part i)

(4) Molecular clock models

Tutorial: Molecular clock dating (part ii)

(5) Setting priors

Tutorial: Phylodynamics (part i)

(6) Tree priors

Tutorial: Phylodynamics (part ii)

Understanding
virus-host interactions

Tree priors

Louis du Plessis

ETH zürich

D BSSE

On the program for today

(1) What is phylodynamics?

(2) Bayesian inference recap

(3) BEAST2 introduction

Tutorial: Molecular clock dating (part i)

(4) Molecular clock models

Tutorial: Molecular clock dating (part ii)

(5) Setting priors

Tutorial: Phylodynamics (part i)

(6) Tree priors

Tutorial: Phylodynamics (part ii)

Demographic model

- Describes the population/speciation dynamics
- How does the population demographics / species diversity change over time?

Tree
Realisation of a stochastic process

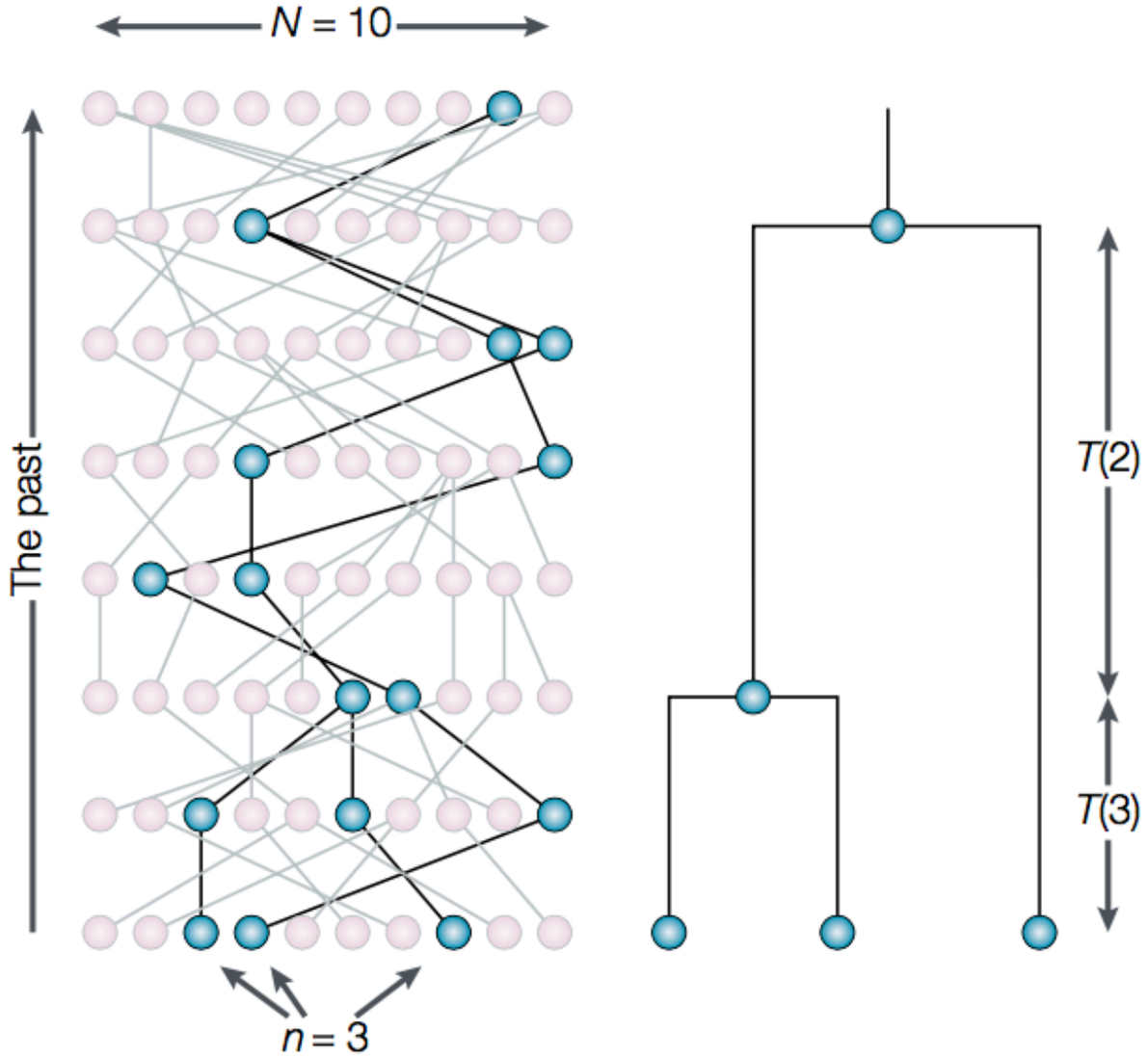
$P(\text{tree} \mid \text{demographic model})$

Demographic model
Describes the population dynamics (growth of the tree)

- How likely is the genealogy given a demographic model?
- Sometimes called a **tree prior**
- Usually a **coalescent** or **birth-death** model

Coalescent models

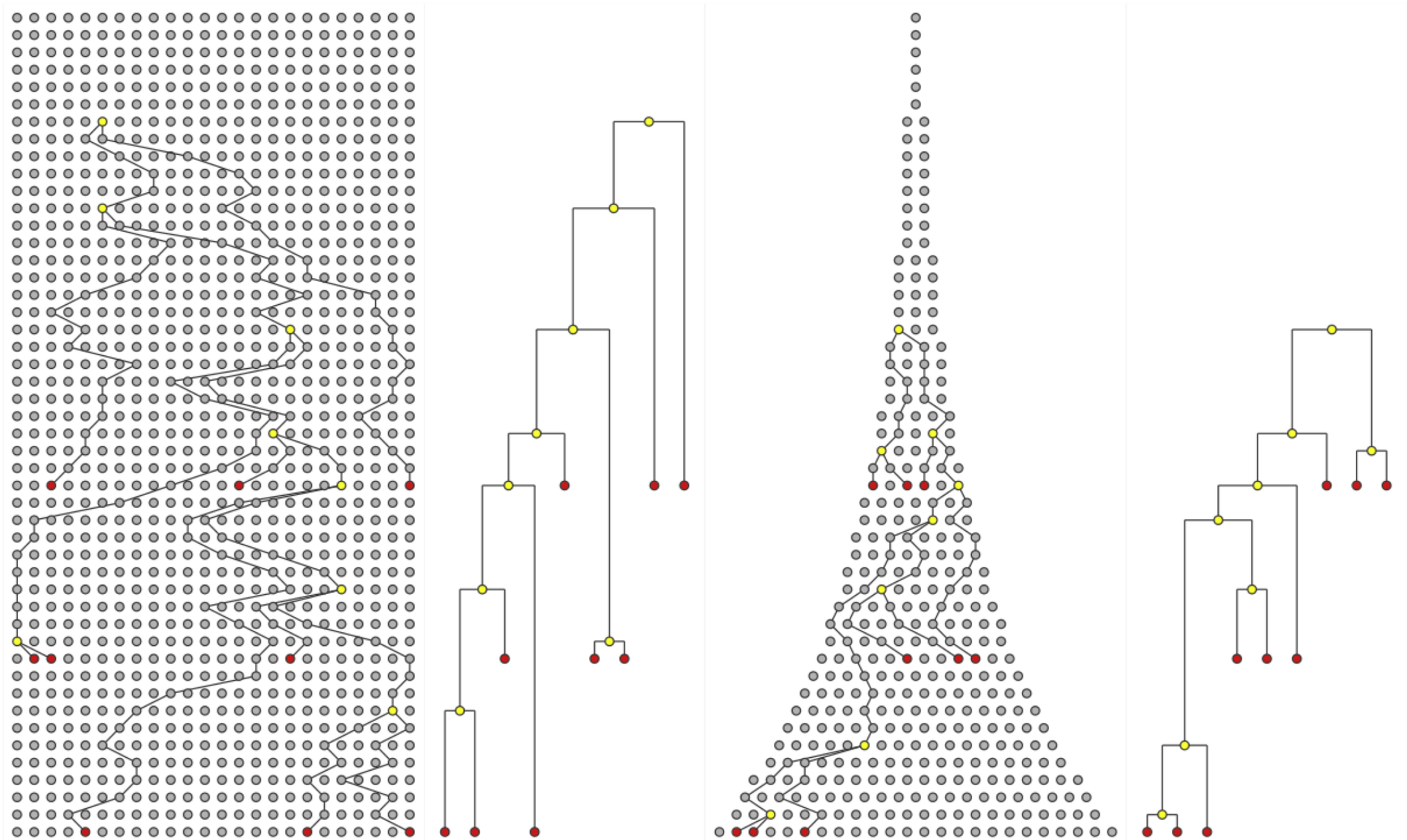
Full population **Sampled tree**



- Approximation to Wright-Fisher population dynamics (with large **N**)
- Discrete, non-overlapping generations, complete mixing
- Trace ancestry of **n** samples in a population of size **N**
- Given **N** it is easy to calculate the probability for **2** nodes to coalesce in time **t**
- Extend to the probability of **2** of **n** nodes coalescing in time **t**
- Calculate the probability of observing a given **tree** for a particular **N**
  → estimate **N** ($\mathbf{N_e}$ in practice)

Rosenberg and Nordborg **Nature Reviews Genetics 2002**

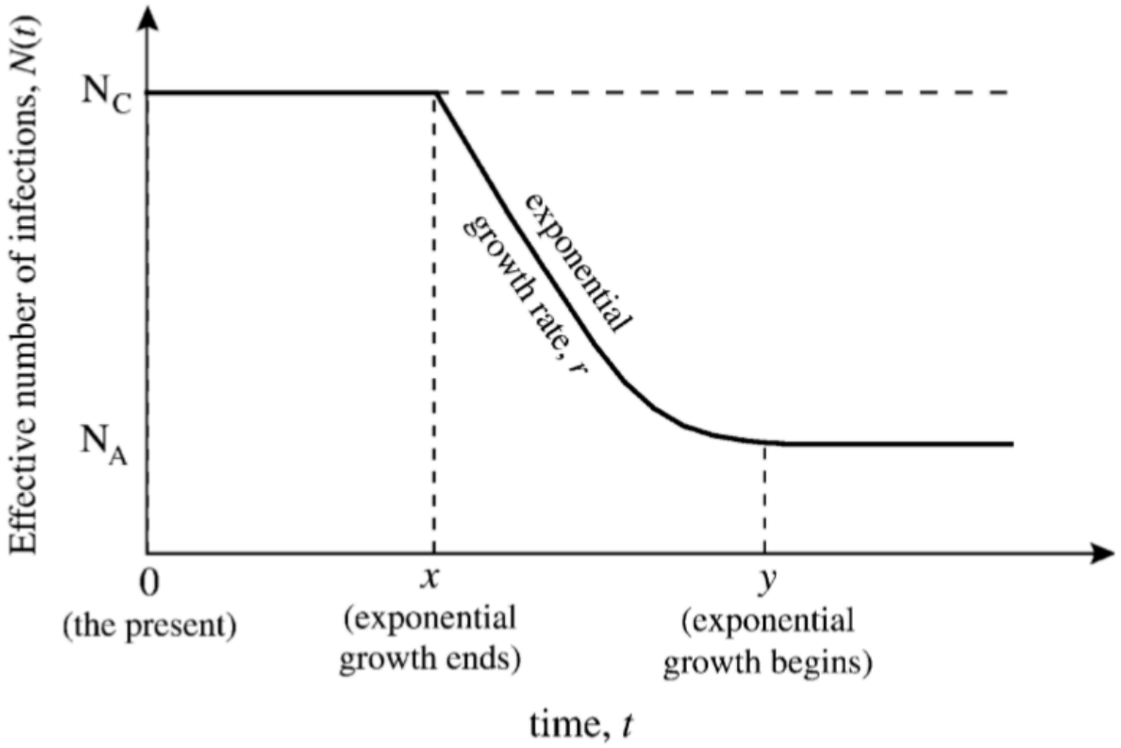
Changes in N_e over time



Kühnert *et al.* **Infection, Genetics and Evolution 2011**

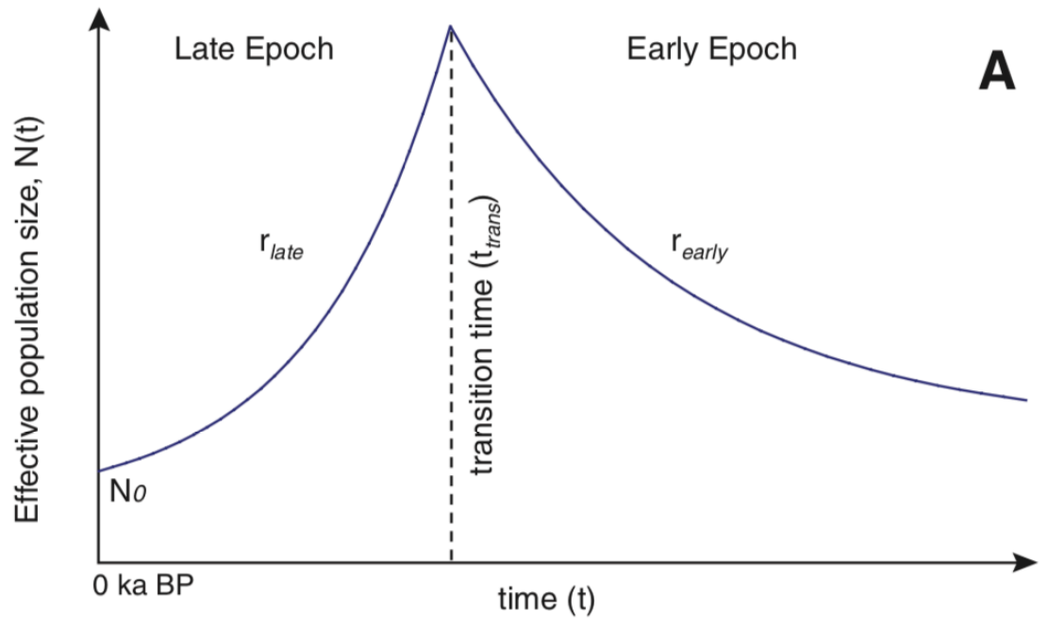
Parametric models for $N_e(t)$

Egyptian HCV



Constant-Exponential-Constant dynamics

Steppe Bison

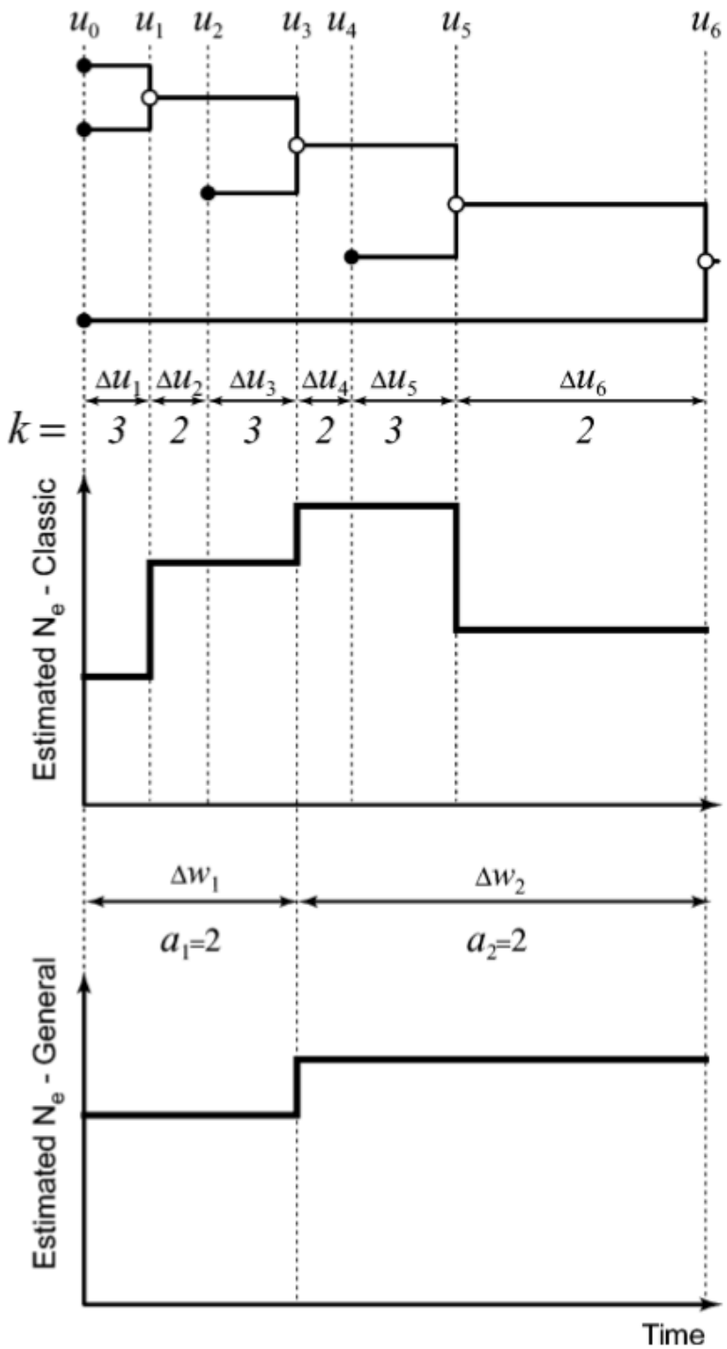


Boom-Bust dynamics

Pybus *et al.* **MBE 2003**

Shapiro *et al.* **Science 2004**

Skyline plots



Classic skyline plot

- Piecewise constant N_e
- Change-points at coalescence events
- Noisy estimate

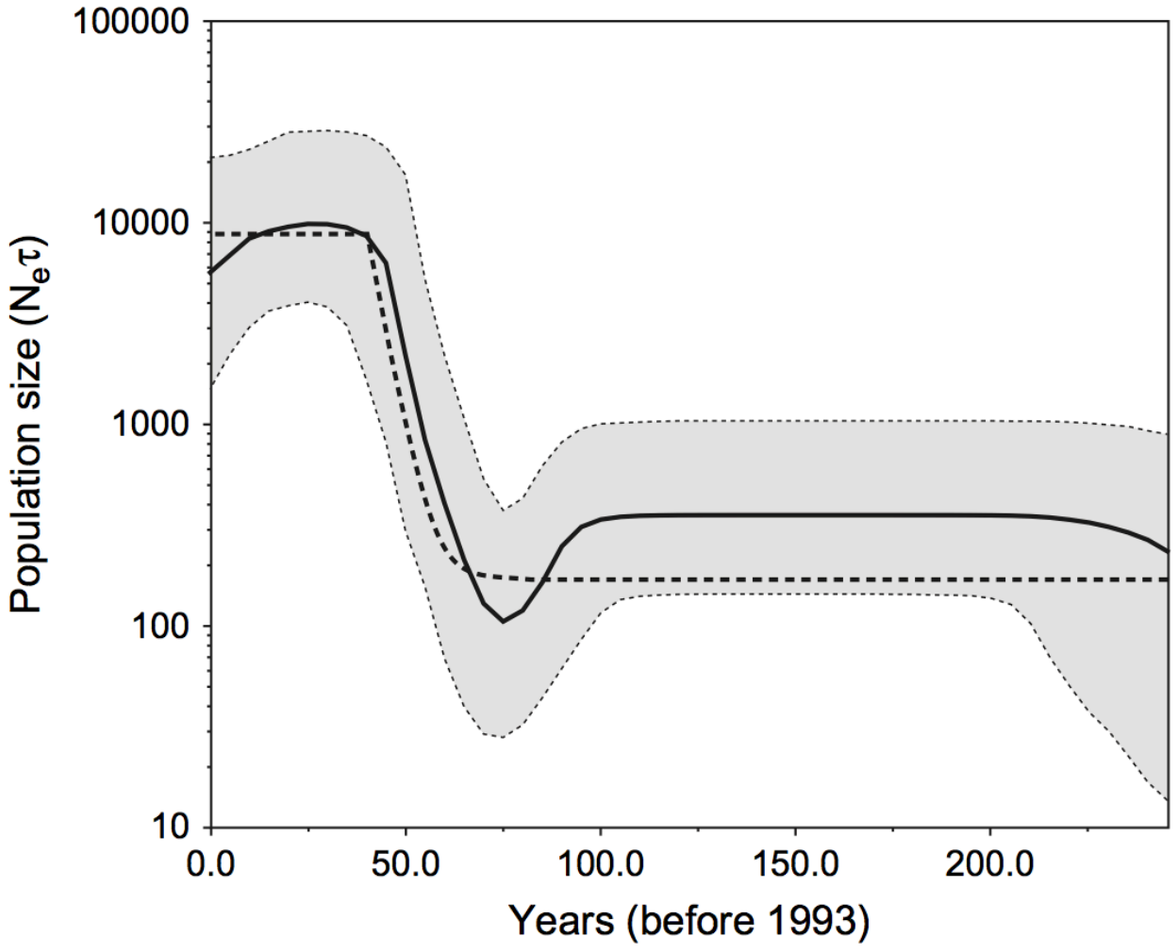
Generalised skyline plot

- Group neighbouring segments
- Smoother estimate

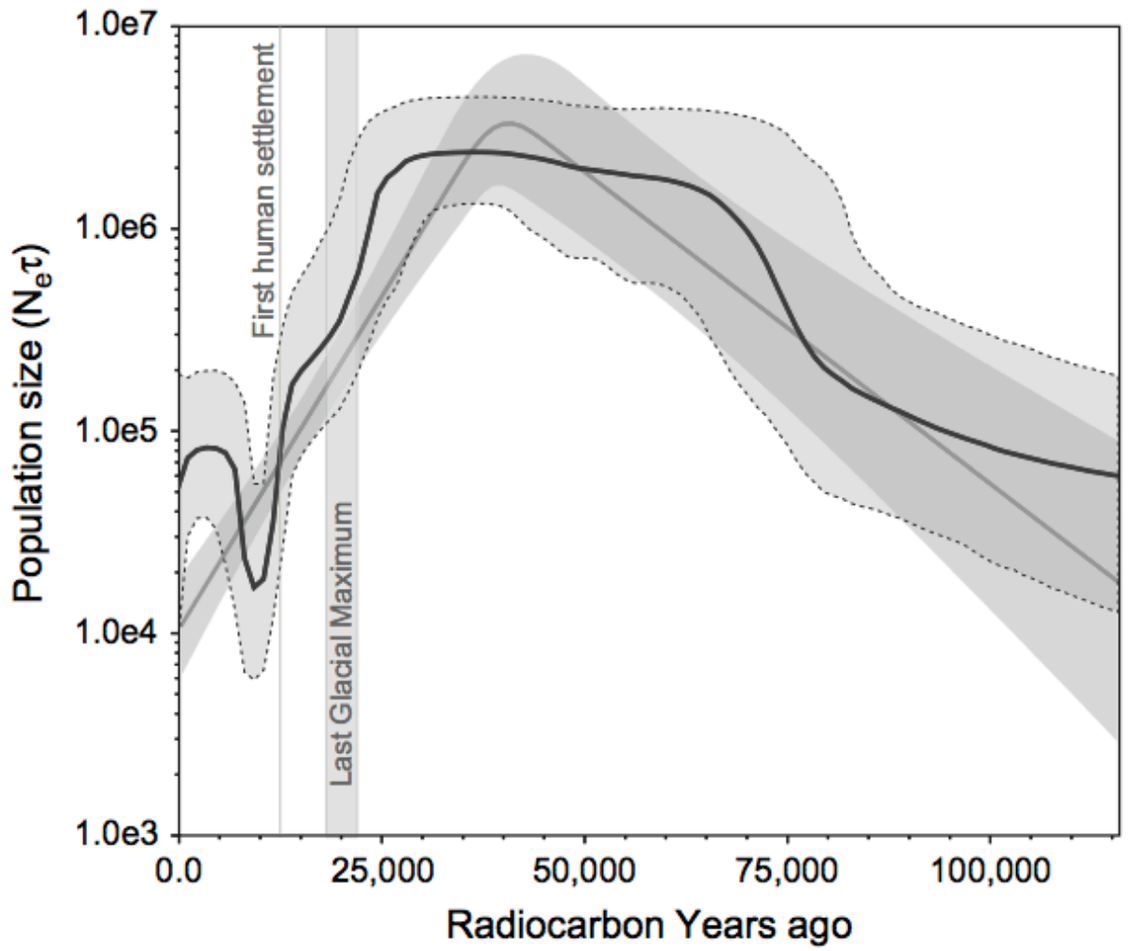
Drummond *et al.* **MBE 2005**

Bayesian Skyline plot

Egyptian HCV



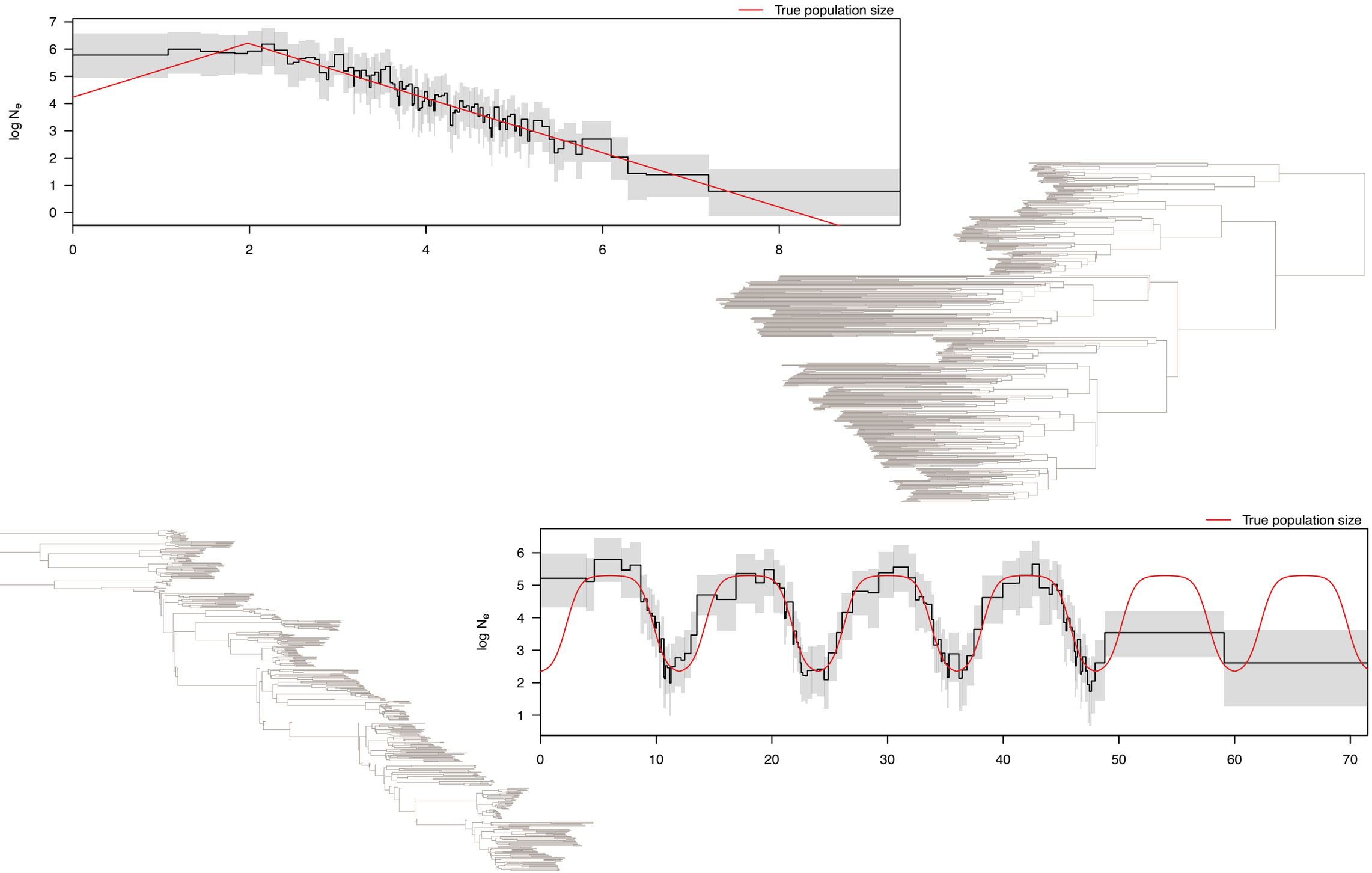
Steppe Bison



- Recovers the same dynamics as the parametric model
- More flexible but also more uncertain

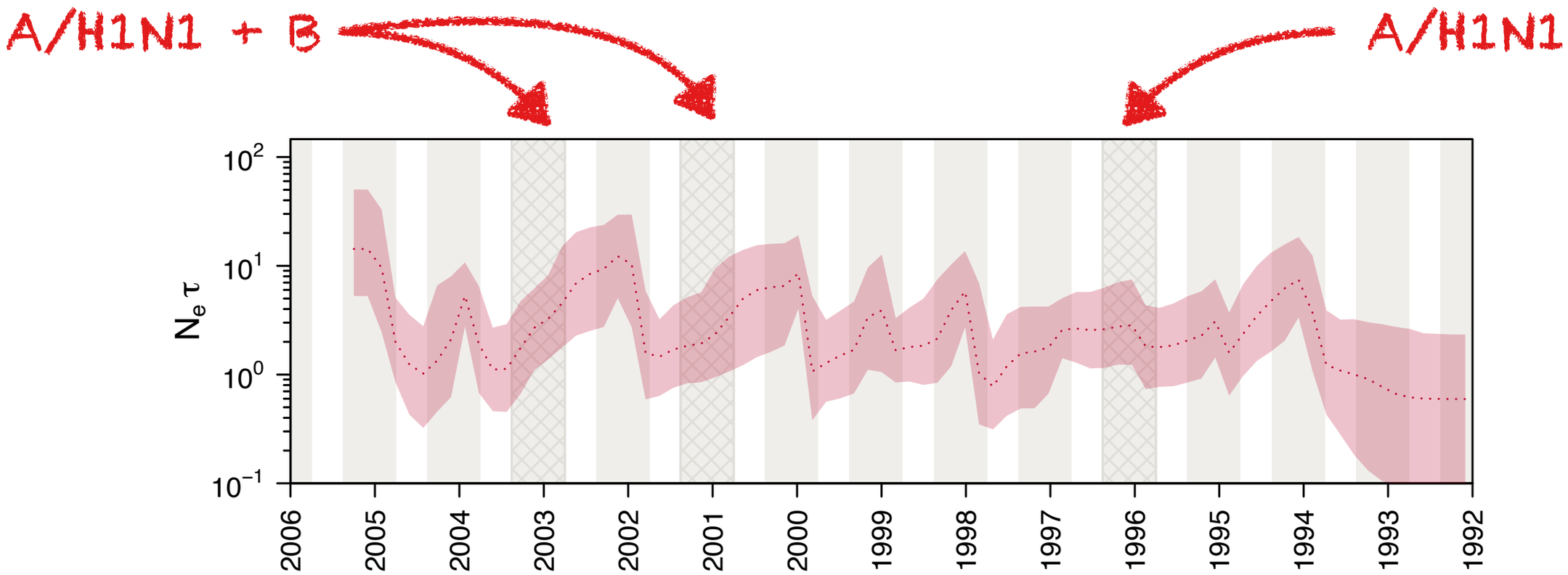
Drummond *et al.* **MBE 2005**

How well does it work?



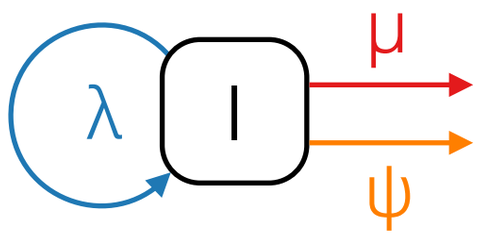
Bayesian Skyline plot

Influenza A H3N2 in New York



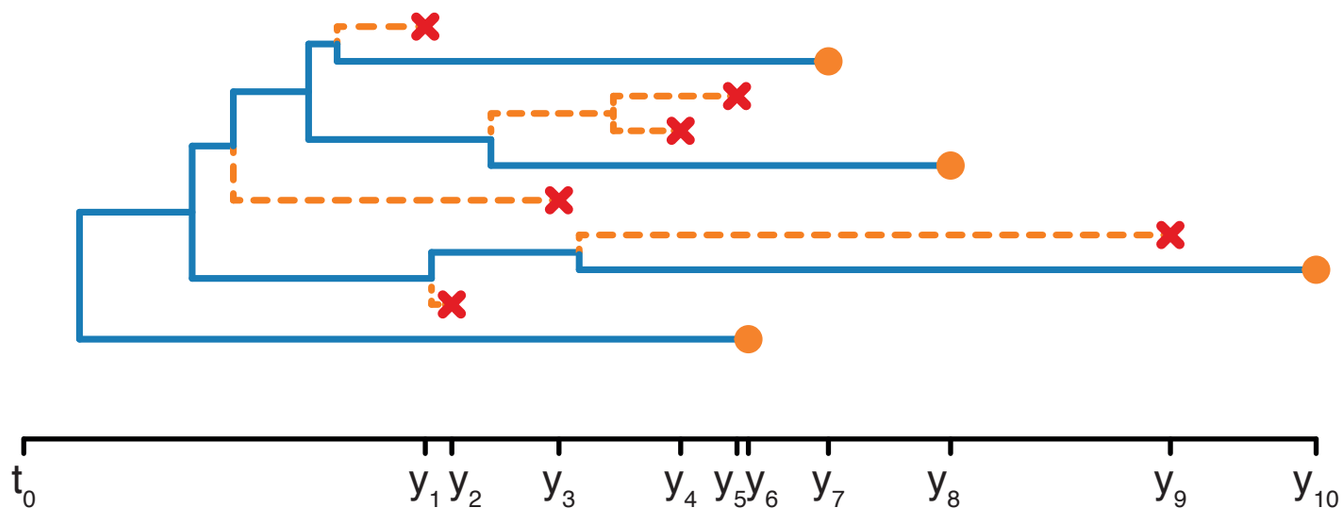
Parag *et al.* **MBE 2020**

Birth-death-sampling models



- λ — birth/transmission rate (lineages added to **full** tree)
- μ — death/recovery rate (lineages removed from the **full** tree)
- ψ — sampling rate (samples added to **sampled** tree)

- Forward-in-time branching process
- Events happen at different rates
  - infection/recovery
  - speciation/extinction
  - sampling/fossilization
  - …



- Calculate the probability of a series of events at specific times to generate a (sampled) tree

Birth-death-sampling models

λ

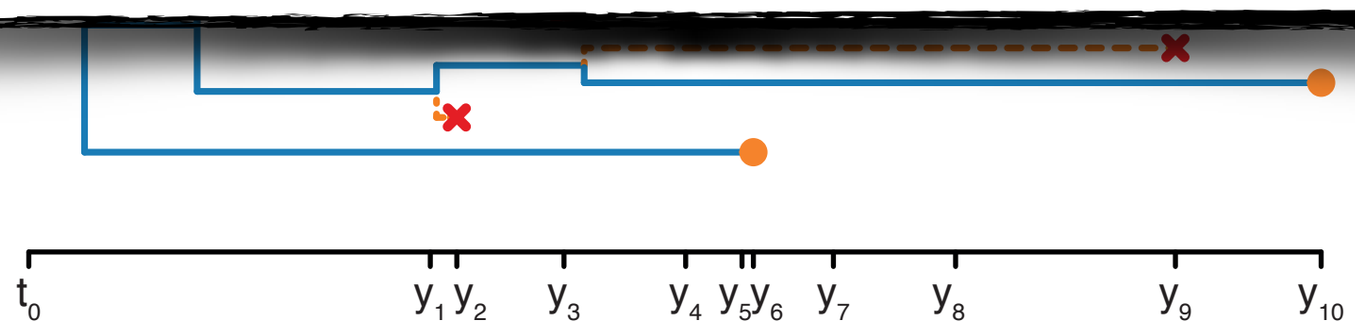
ee)

full tree)

Infectious disease parameterisation

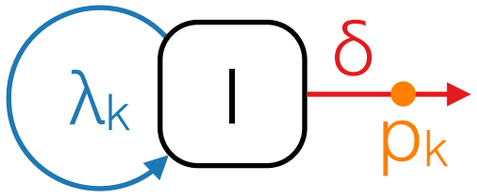
- $R_e = \lambda/(\mu+\psi)$ — Effective reproductive number (is it spreading or not?)
- $\delta = \mu + \psi$ — Becoming uninfectious rate ($1/\delta$ = infectious period)
- $p = \psi/(\mu + \psi)$ — Sampling proportion (Proportion of sampled removals)

More natural priors!

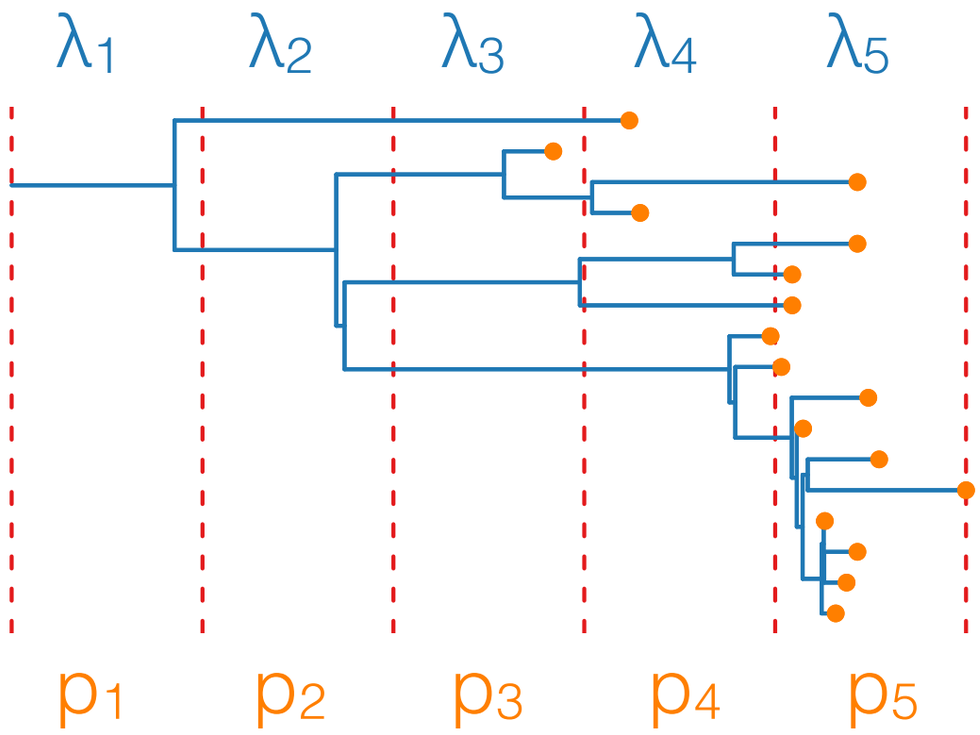


- Calculate the probability of a series of events at specific times to generate a (sampled) tree

Birth-death skyline

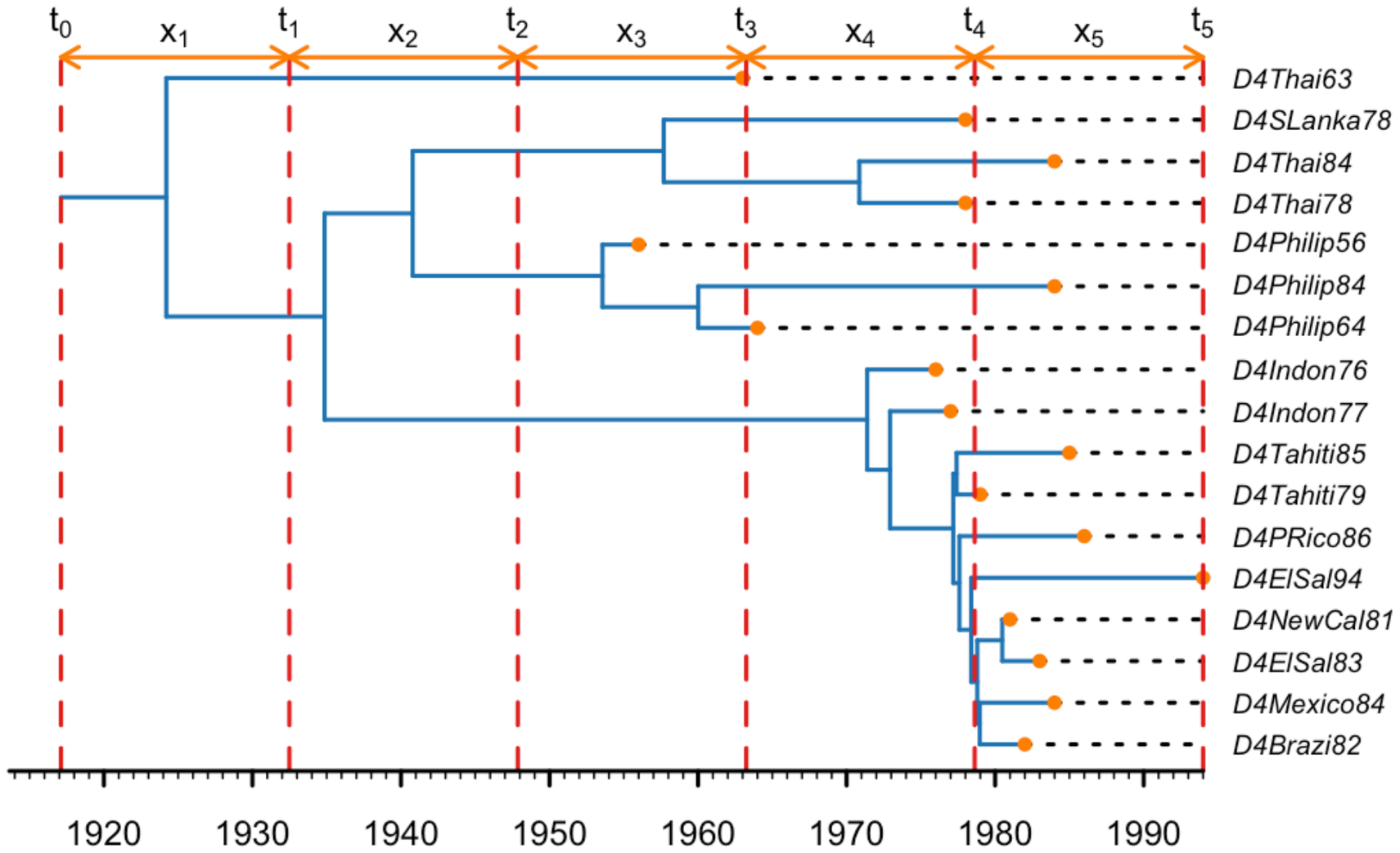


- λ — infection rate
- δ — becoming-noninfectious rate
- p — sampling proportion
- t_0 — Origin of the process



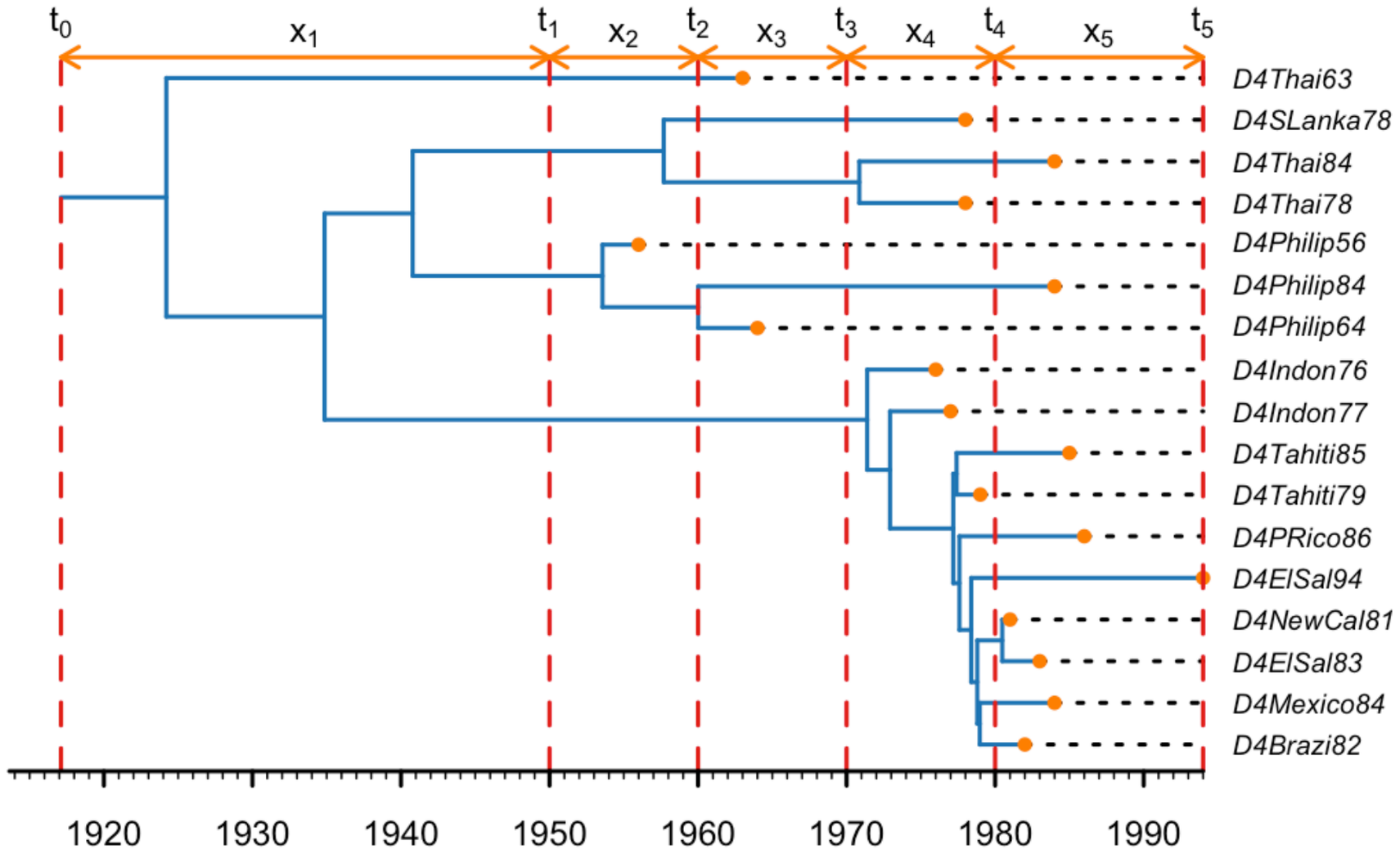
- Model parameters are rates that describe the growth of the tree
- Effective reproduction number = λ/δ
  (is it spreading or not?)
- Allow parameters to change through time
- Shifts in rates can be anywhere

Birth-Death Skyline (BDSKY)



- Time-changing parameter can be **any** or **all** of the model rates (birth, death, sampling)

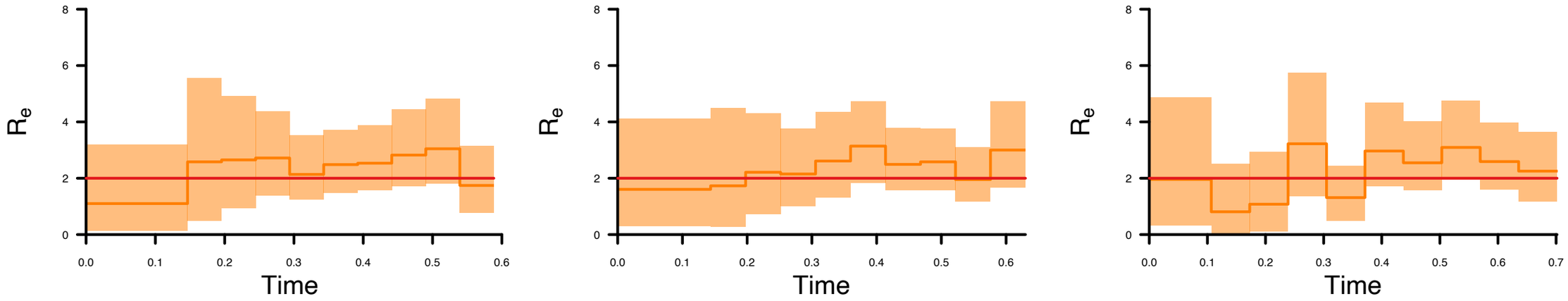
Birth-Death Skyline (BDSKY)



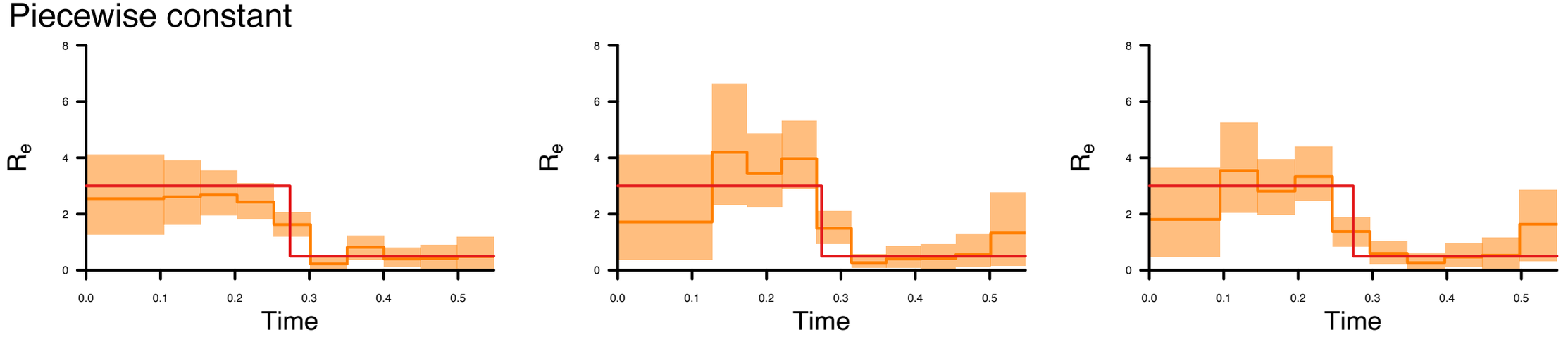
- Change-points can be anywhere between origin and present
- More difficult to set up XML if not equally-spaced

How well does it work?

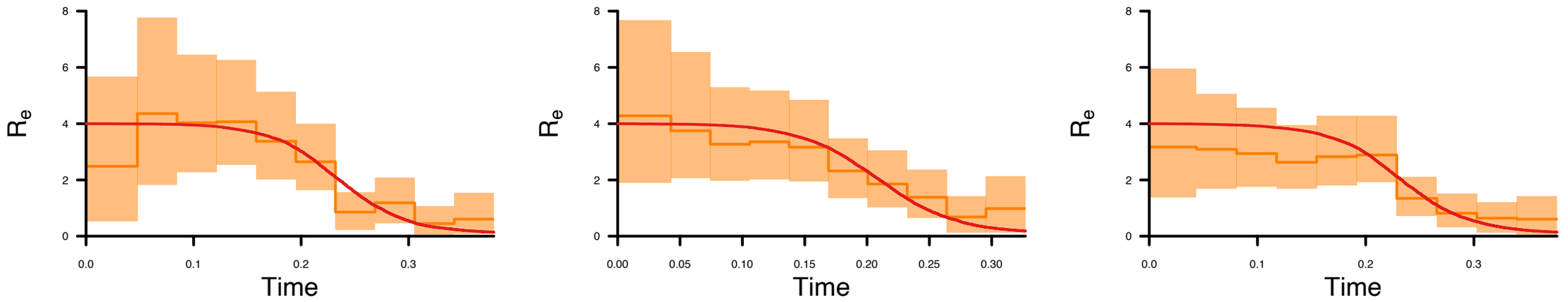
Constant



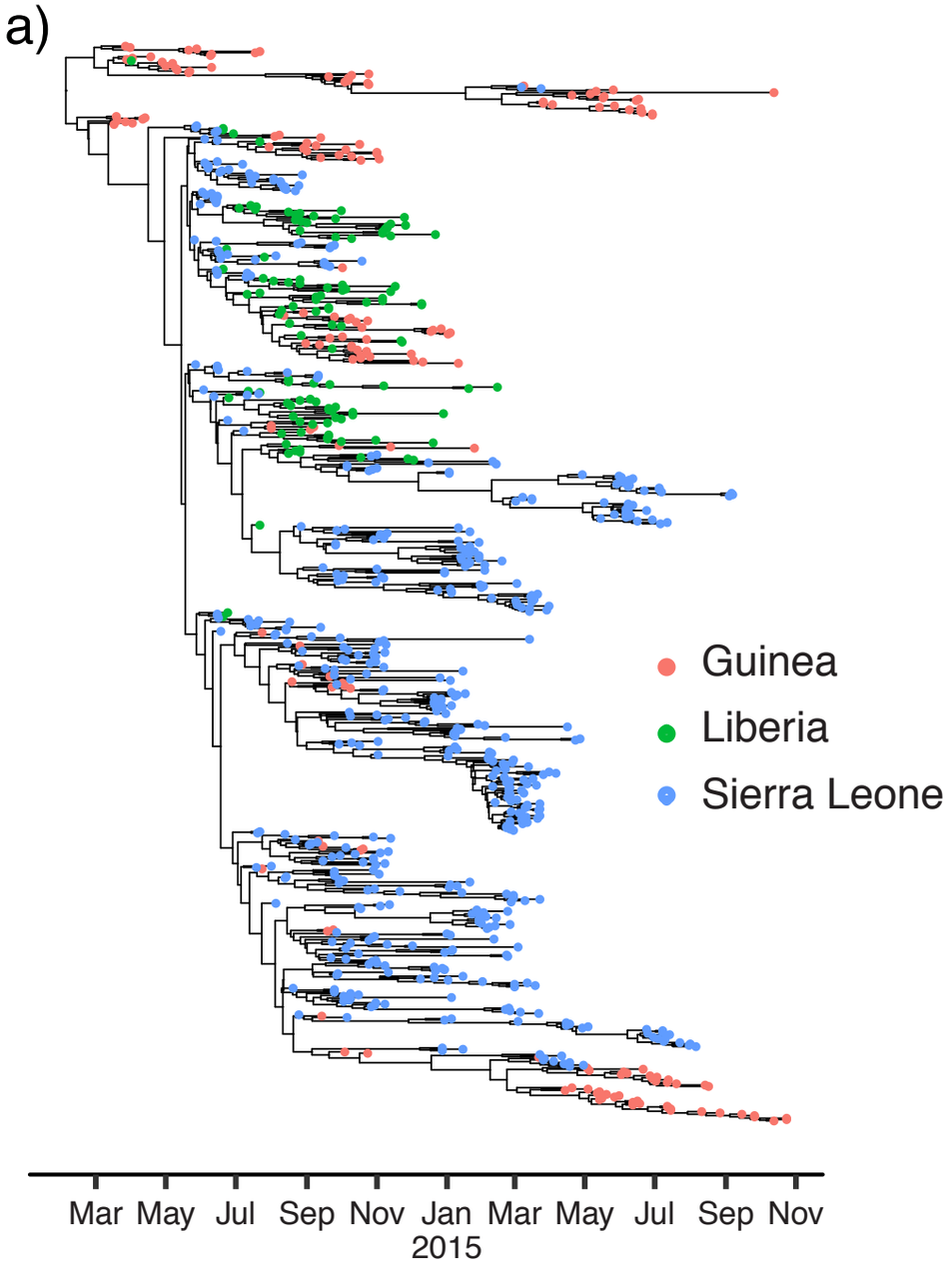
Piecewise constant

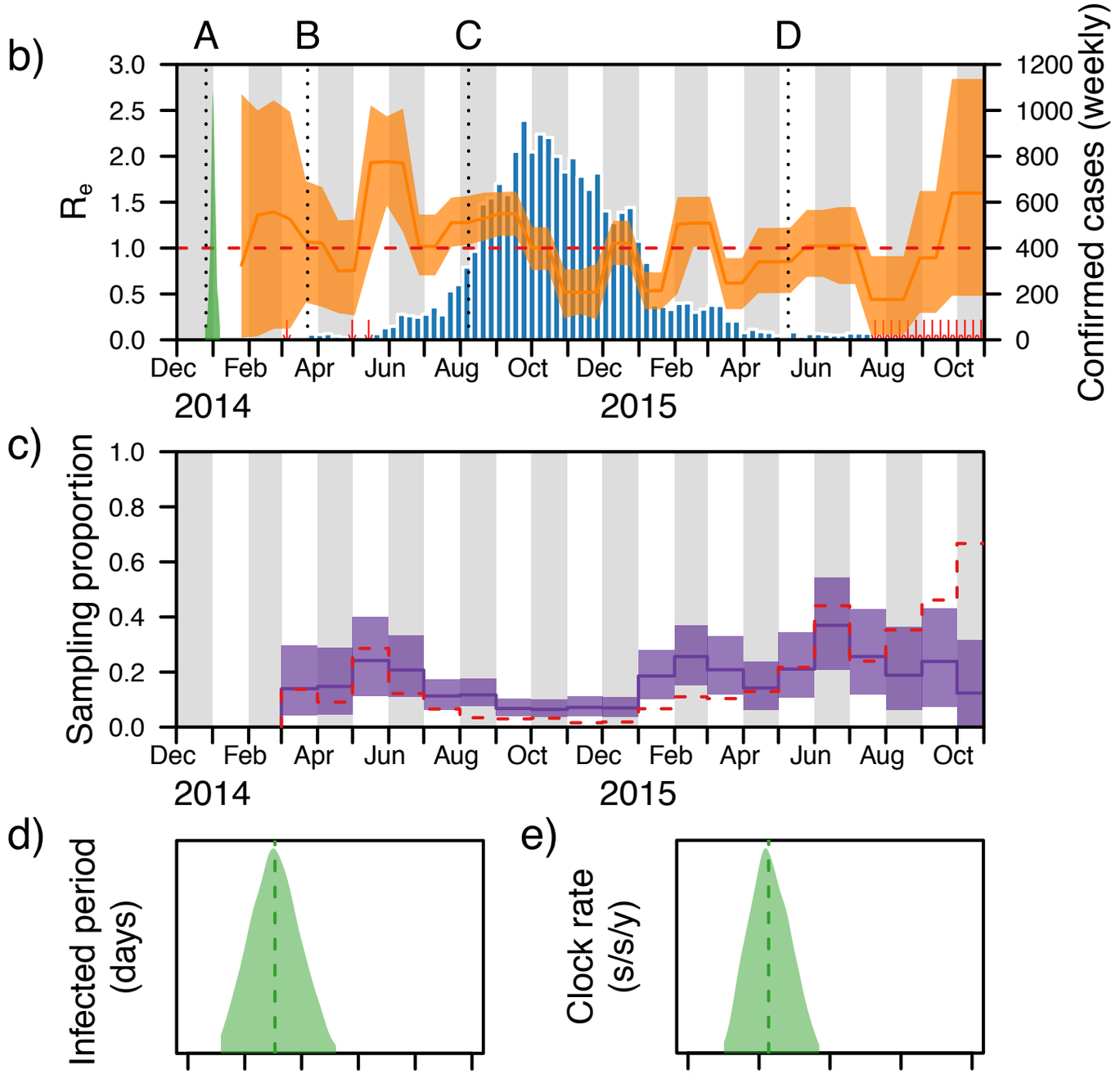


SEIR model



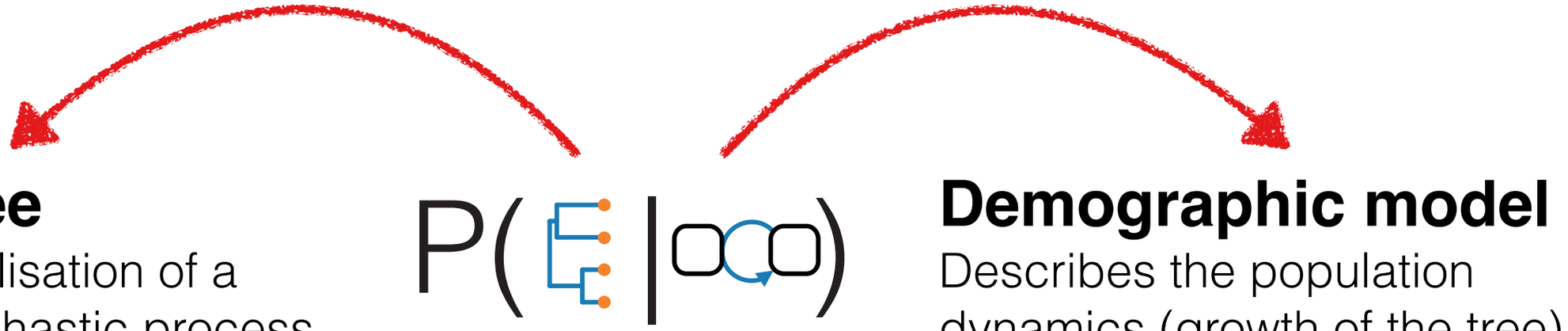
EBOV in West Africa





Bouckaert *et al.* **PLOS Comp Biol 2019**

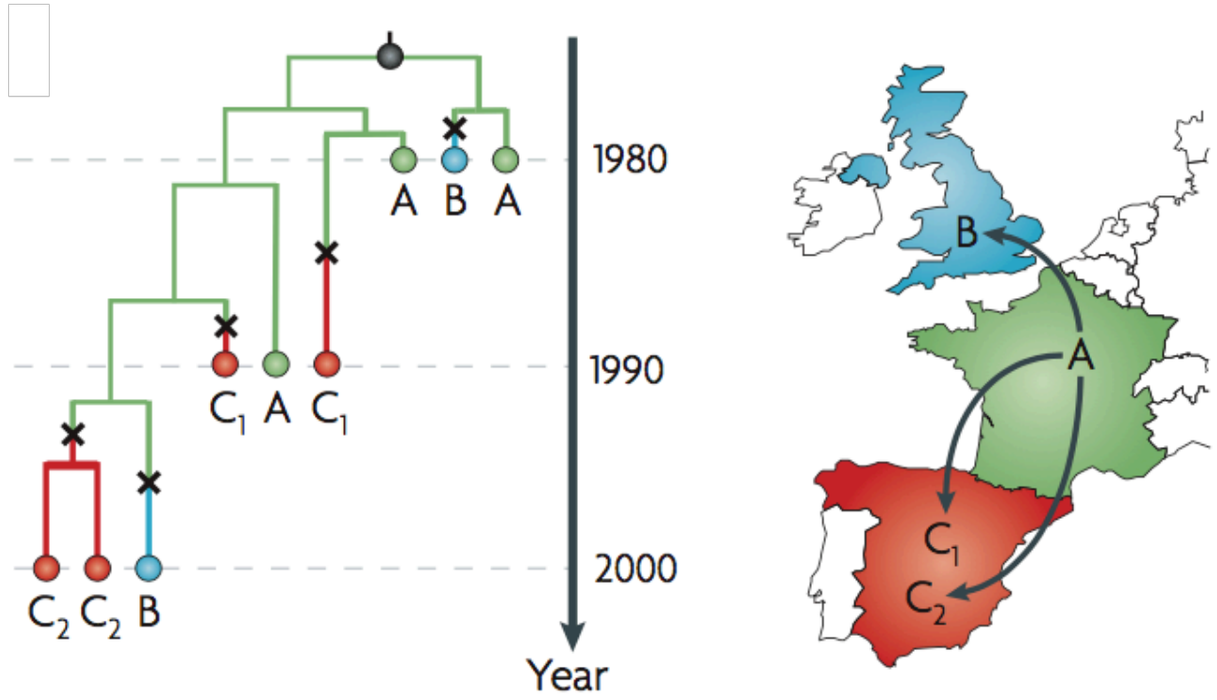
Demographic model



- **Coalescent:**
  Given **n** sampling times and an estimate for **$N_e(t)$**,
  out of all the ways we can connect the samples,
  what is the probability of the current tree?

- **Birth-death:**
  Given an estimate for the **origin** time, **birth**, **death** and **sampling** rates, if we simulate a tree forward-in-time from the origin to the time of the most recent sample, out of all the trees with **n** samples, what is the probability of the current tree?

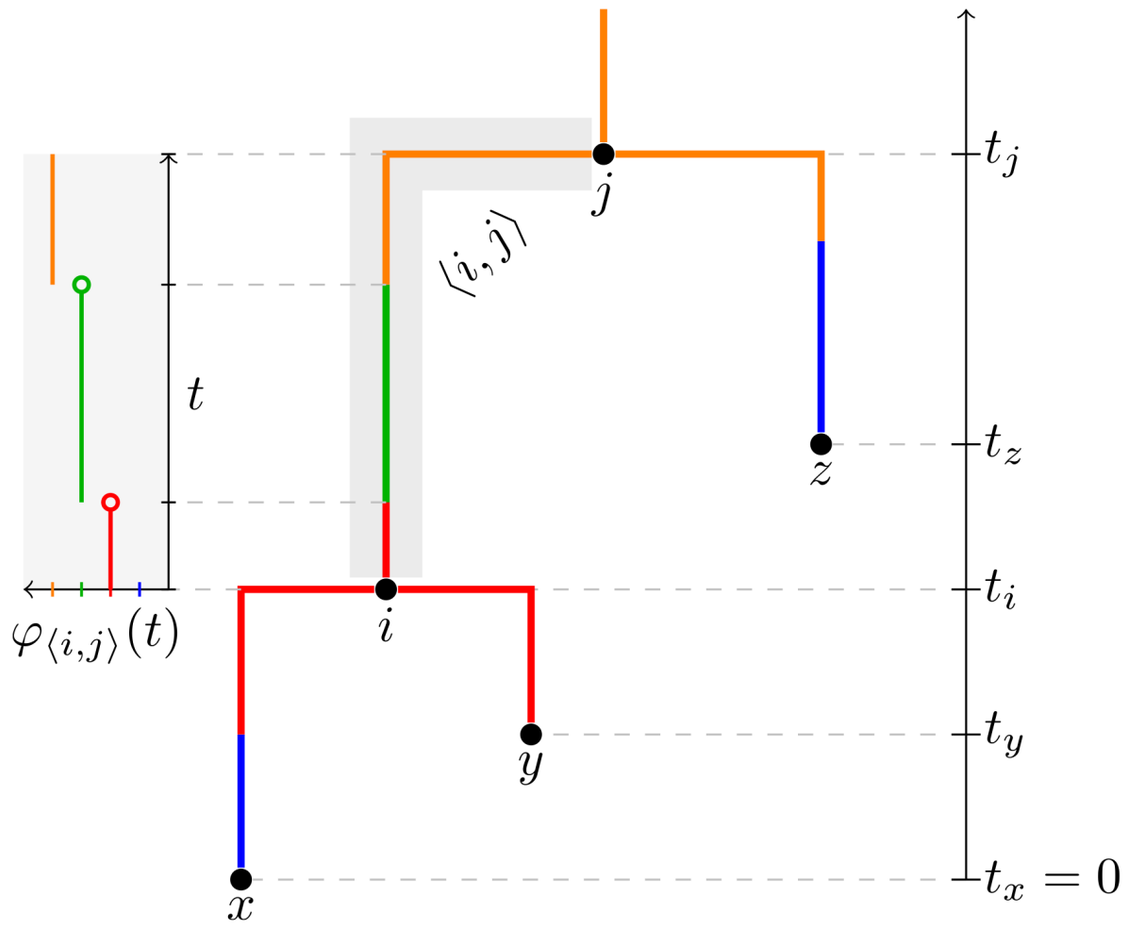
Phylogeography



- "Mugration" model assumes a migration model independent of the tree prior (treats migrations like substitutions - extra alignment with one site for location)
- True structured models (structured coalescent, multi type birth-death model) model it as part of the tree-prior

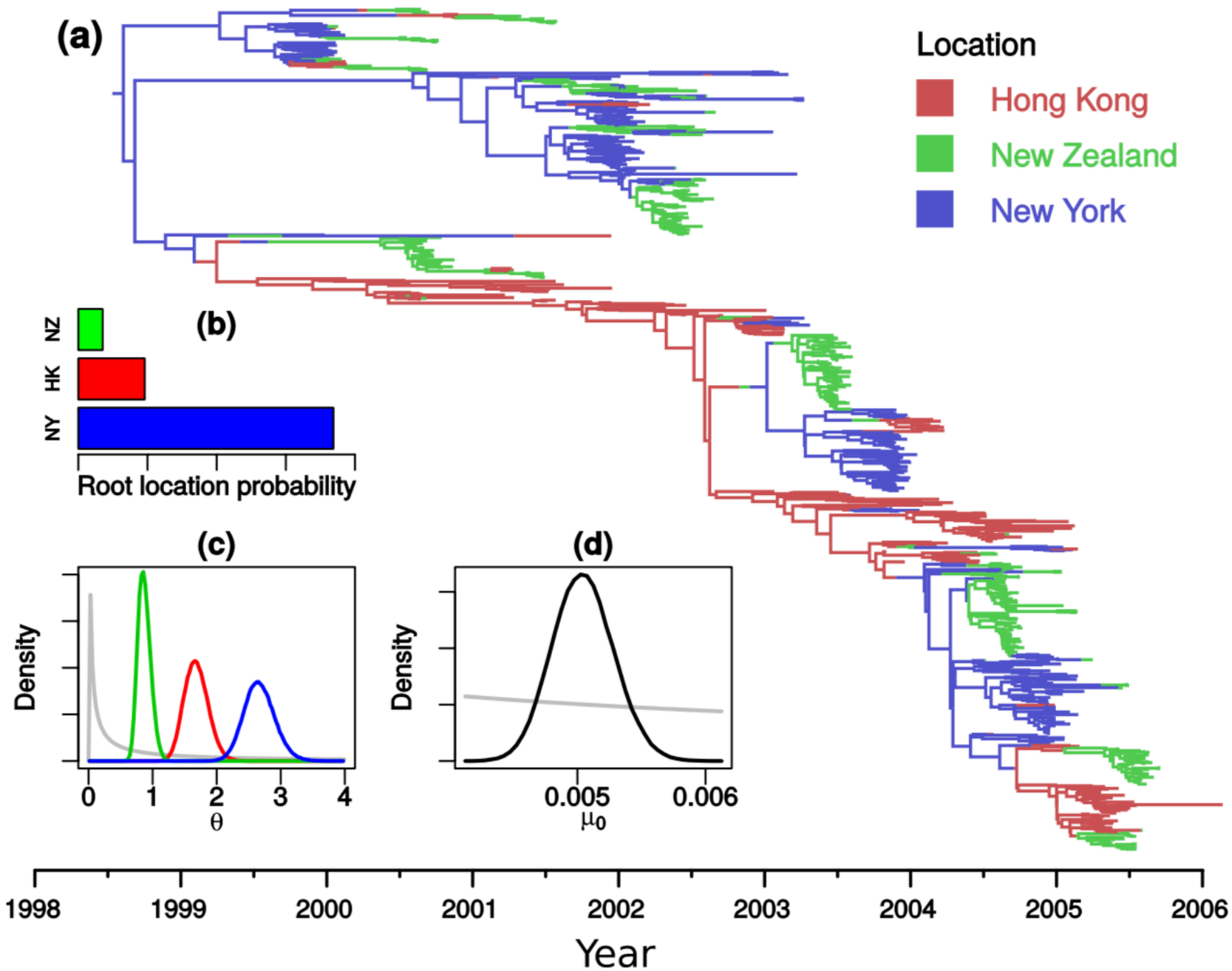
Pybus and Rambaut (2009) *Nat Rev Genet* 10:540-550

Structured models



- Explicitly model different types of subpopulations or locations
- Different types may be associated with different epidemiological dynamics
- Can also model migration between types

Seasonal Influenza H3N2 phylogeography



Vaughan *et al.* **Bioinformatics 2015**

Thank you for listening!

Slides are my own, but some slides were inspired by (or copied from) slides by **Alexei Drummond**, **David Rasmussen, Tanja Stadler, Simon Ho** and **Oliver Pybus**