Introducing Bayesian Phylogenetics

Remco R. Bouckaert
r.bouckaert@auckland.ac.nz

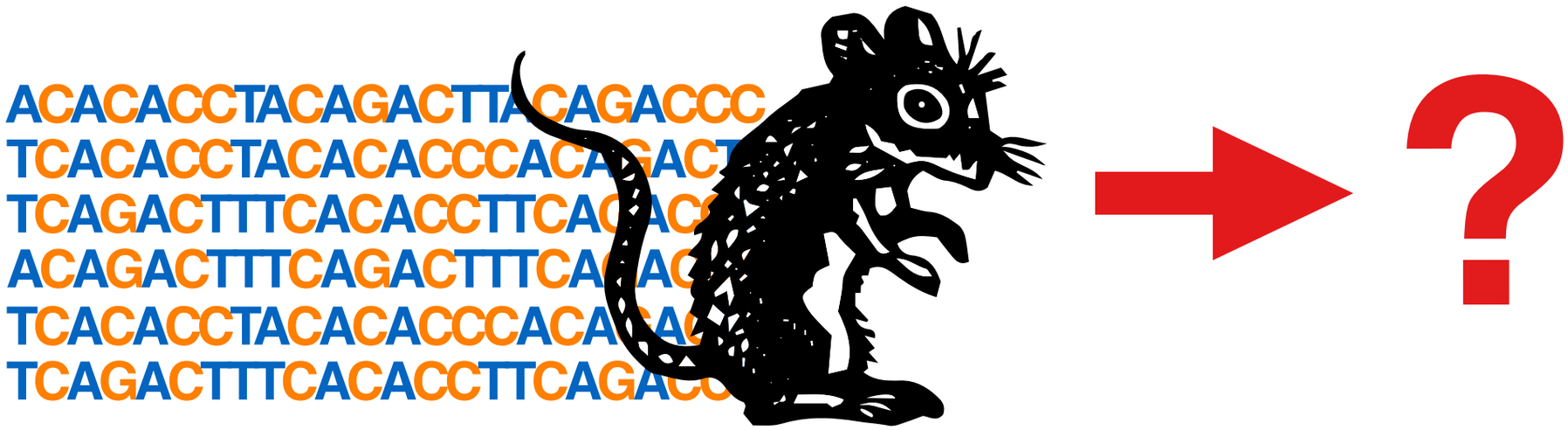
Centre of Computational Evolution, University of Auckland

20 November 2023

Workshop funded by:

Australian National University | CENTRE FOR BIODIVERSITY ANALYSIS

We all have one thing in common…



All of us use genomic sequencing data
to answer questions in **BEAST**

Bayesian **phylogenetic** and **phylodynamic** inference

Likelihood

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

Posterior

Prior

Marginal likelihood of the data

Bayesian inference

(Data and model parameters are both described by probabilities)

Prior → P(model)

- Have some degree of belief in our hypothesis
- All model parameters have priors, whether you specify them or not

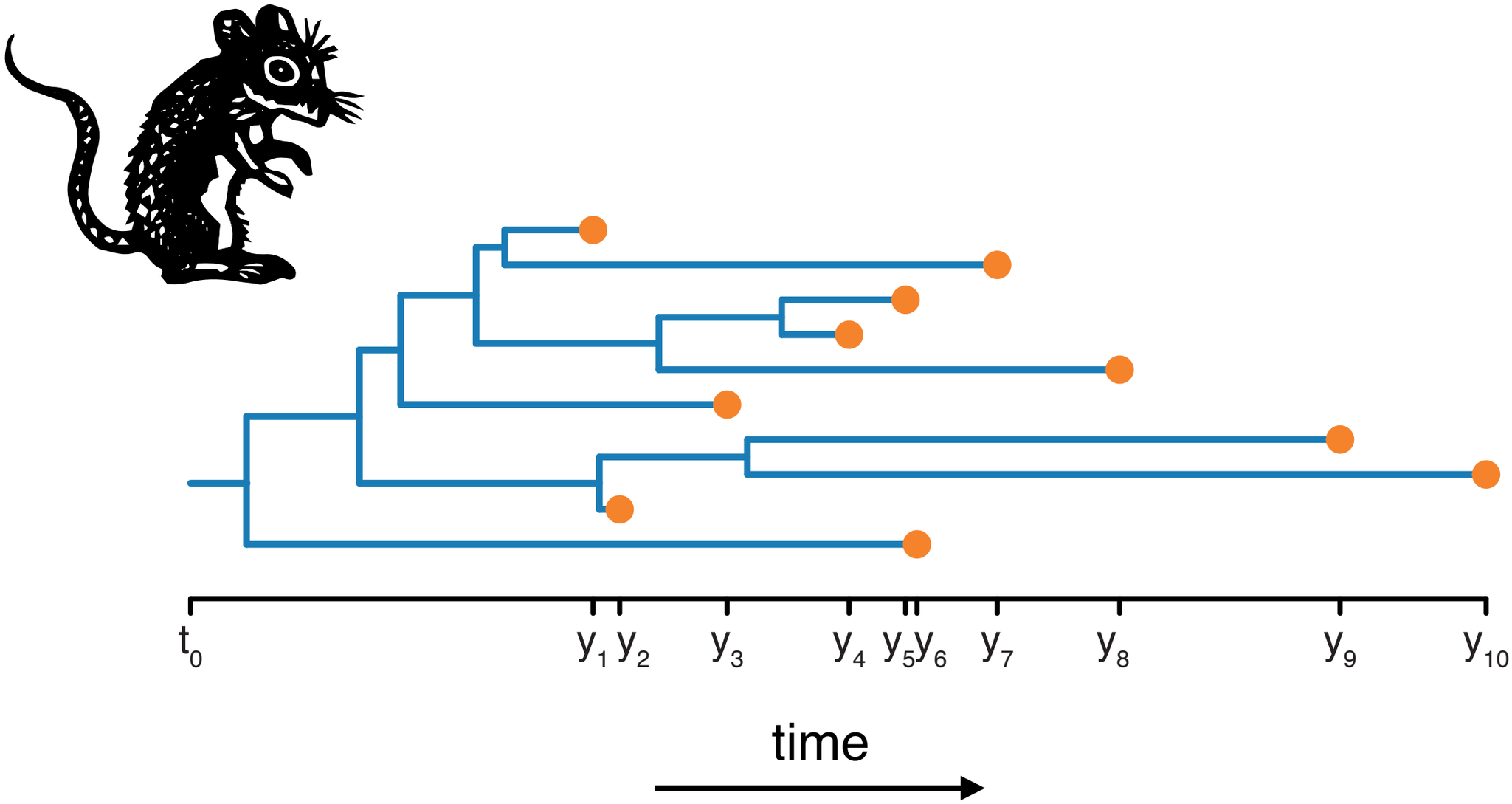
Likelihood → P(data | model)

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

Posterior → P(model | data)

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

Rooted time-trees



Fundamental data structure in **BEAST**
is a **rooted time-tree**

How many trees are there?

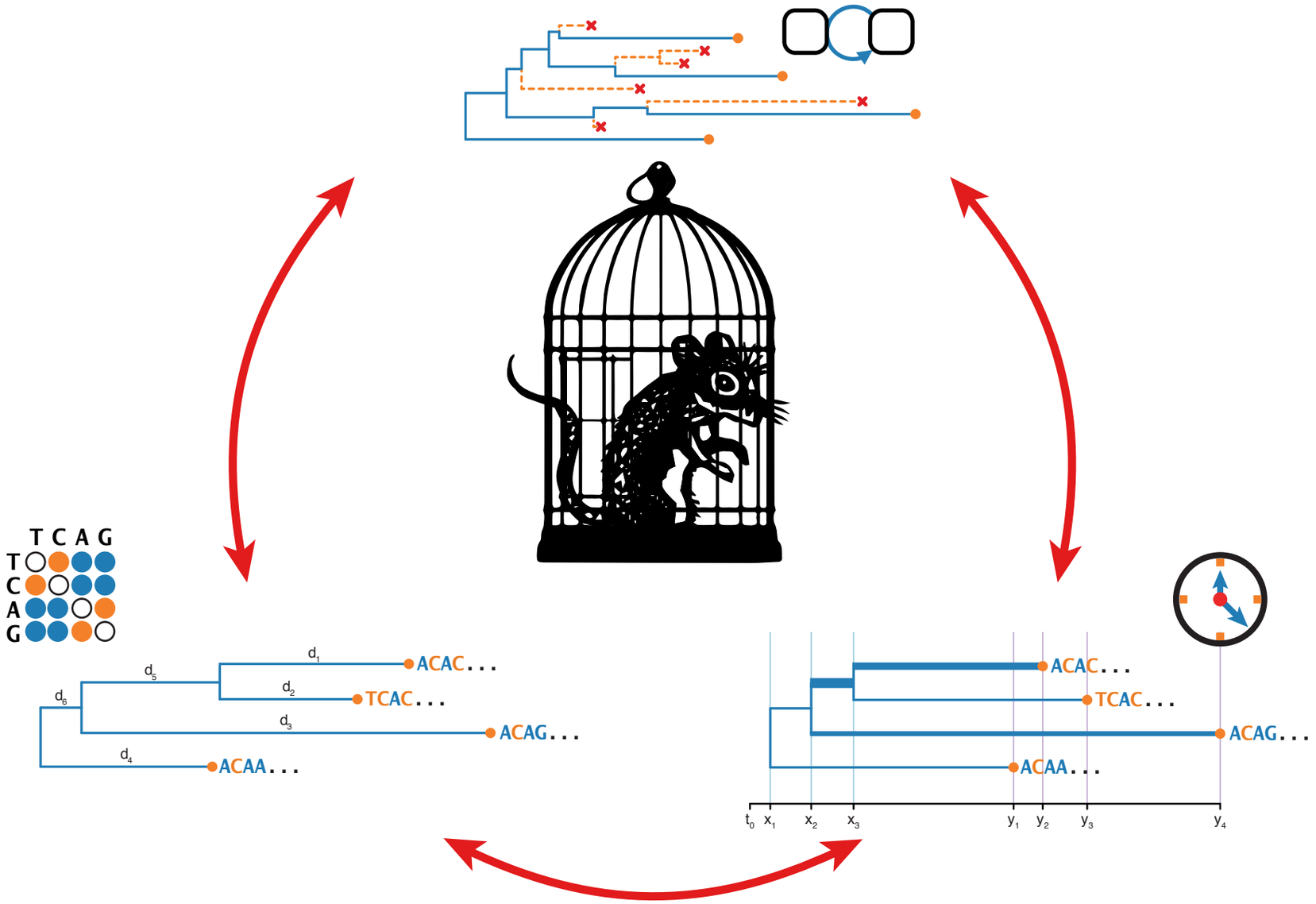
For n species there are

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

rooted, tip-labelled binary trees:

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

What goes into a **BEAST** model?



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

genealogy

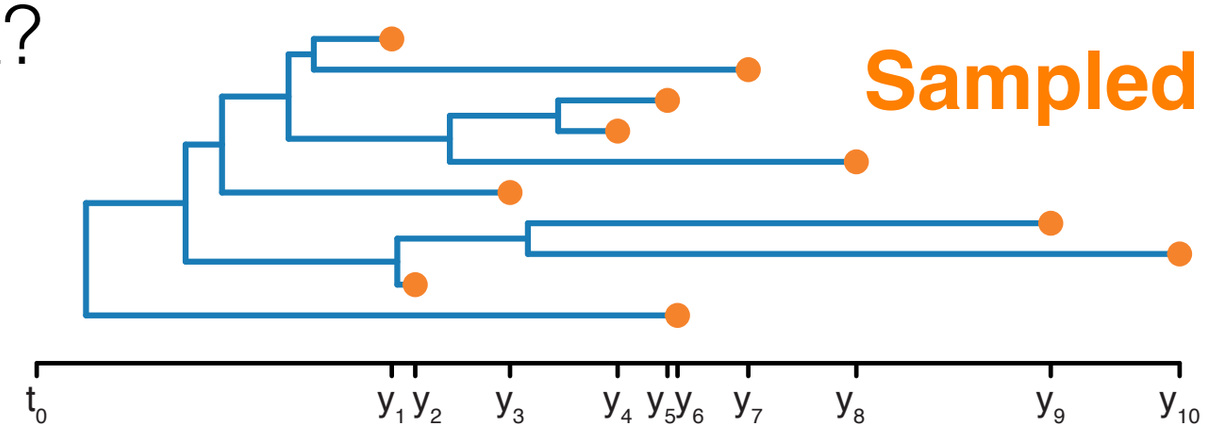
demographic model

site model

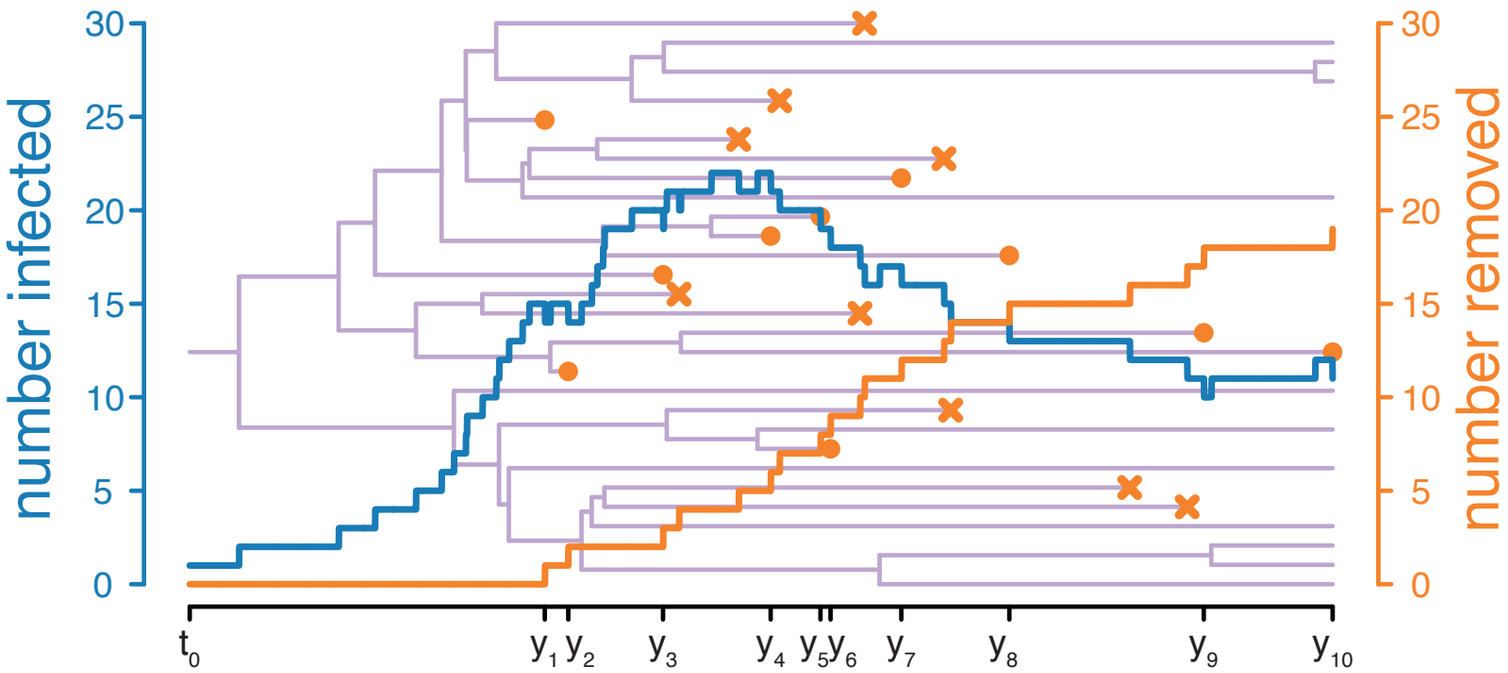
molecular clock model

The genealogy (tree)

- What are the ancestral relationships between the sequences in our dataset?

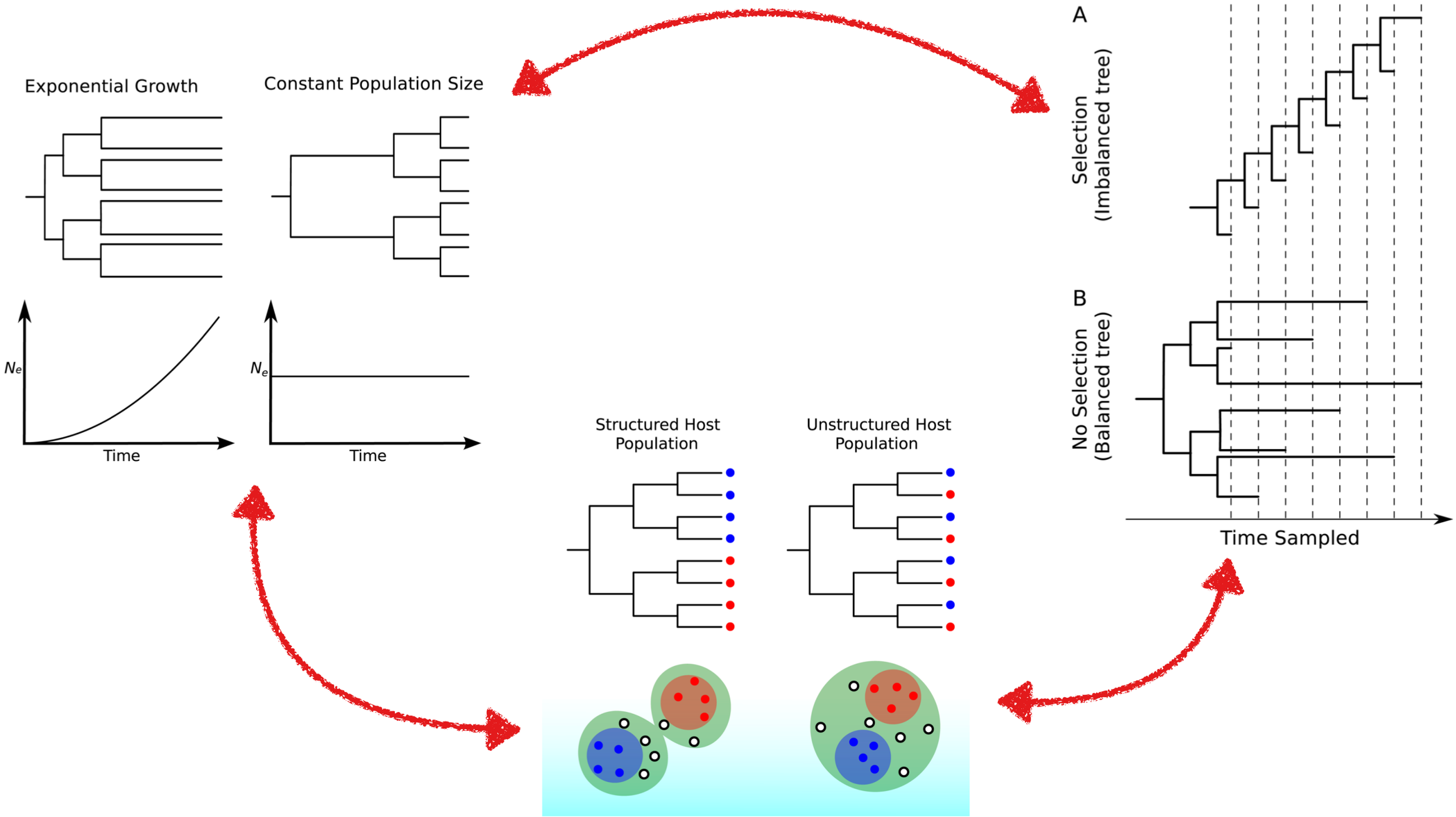


Full tree

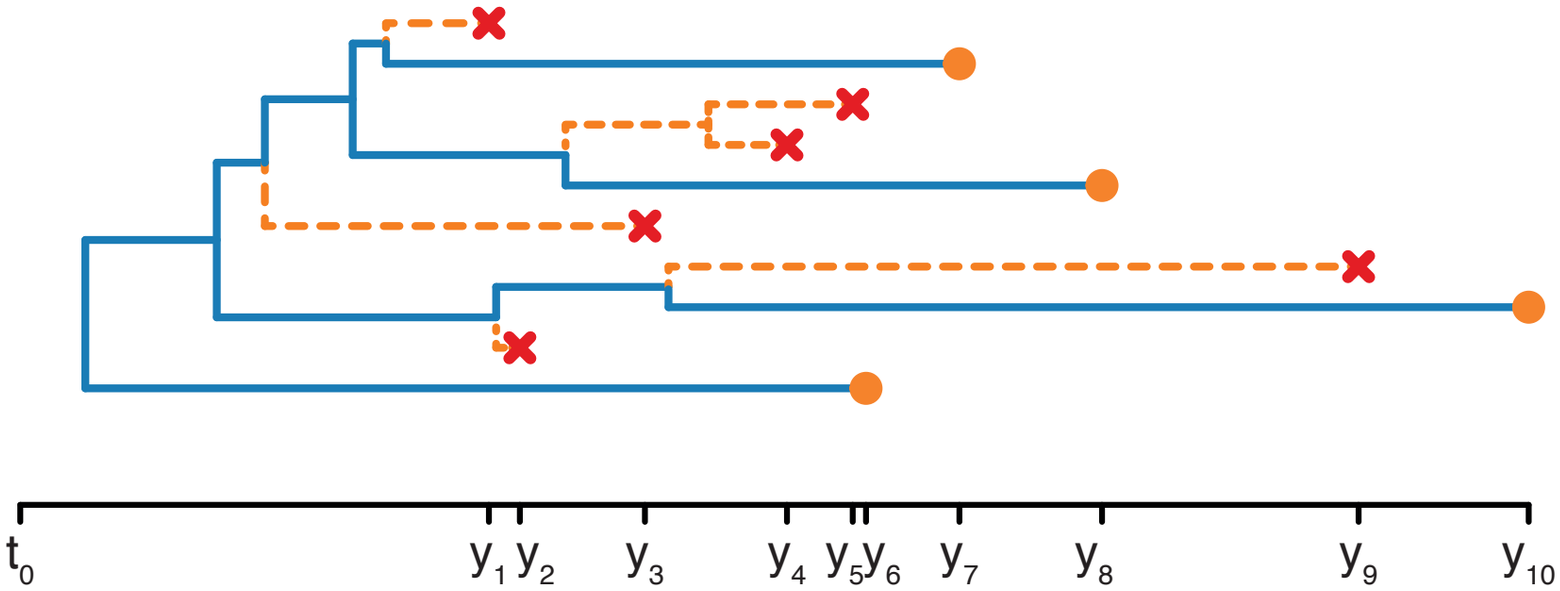


- Only the relationships between the **sampled** sequences!

Different population dynamics generate different trees



Demographic model

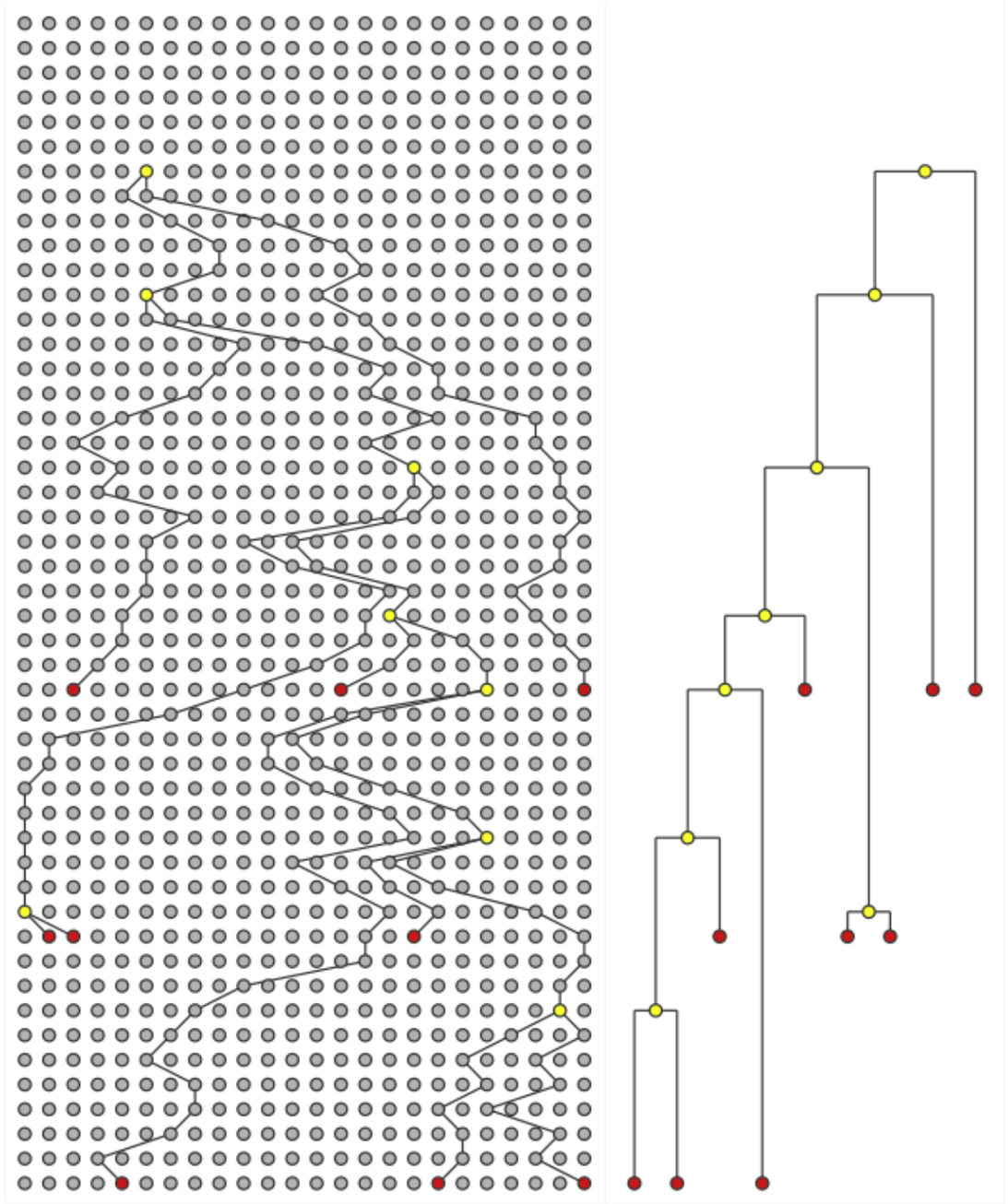


- Describes the population dynamics
- How does the population grow over time?

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

- How likely is the genealogy given a demographic model?
- Usually a birth-death or a coalescent model

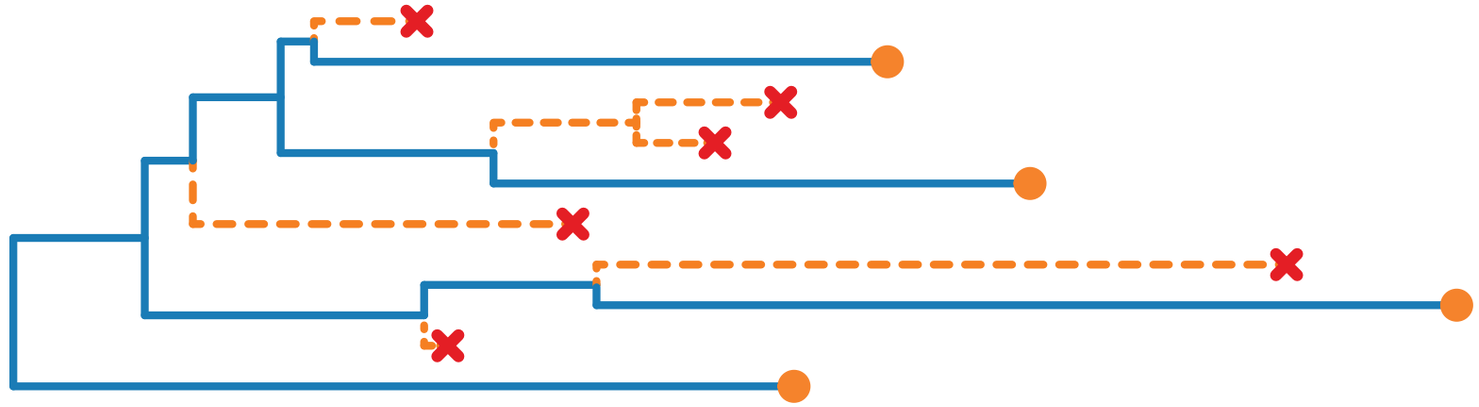
Coalescent model

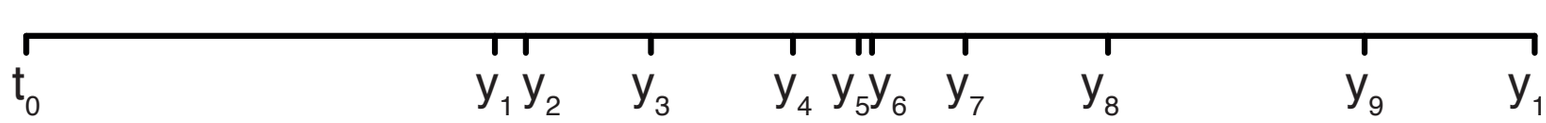


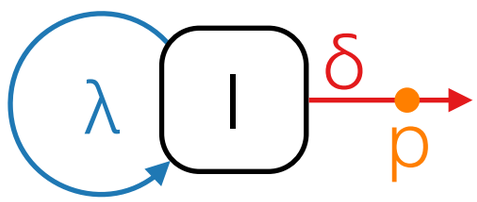
- Assume Wright-Fisher like population dynamics
- Given effective population size ($\mathbf{N_e}$)
- Calculate the probability for **2** nodes to coalesce in time **t**
- Calculate the probability of observing a given **tree** for a particular $\mathbf{N_e}$

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

Birth-death model



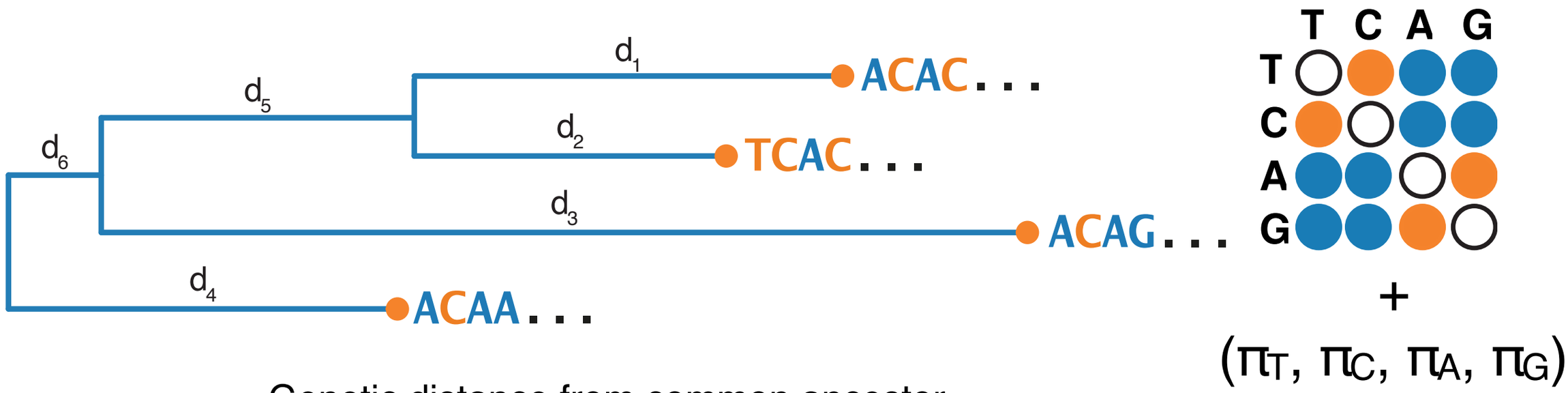




- λ — infection rate
- δ — becoming-noninfectious rate
- p — sampling probability

- Forward-in-time branching process
- Events happen at different rates (speciation/infection, recovery/extinction etc.)
- Calculate the probability of a series of events happening at specific times to generate a tree

Site model



Genetic distance from common ancestor

- Links the genome sequences to the genealogy
- We observe sequences at the tips, not their histories
- Multiple substitutions at the same site means not all substitutions are observed
- To infer the evolutionary history we need to take **all possible evolutionary trajectories** into account!

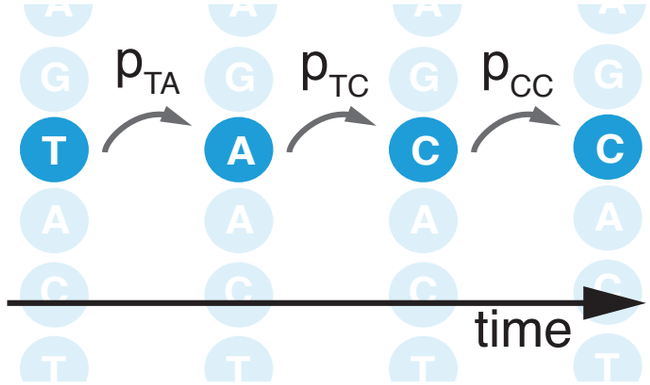
Substitutions as a Markov process

(courtesy of Carsten Magnus)

- Assume every site is evolving independently
- Assume nucleotide substitutions at each site is governed by a Markov process

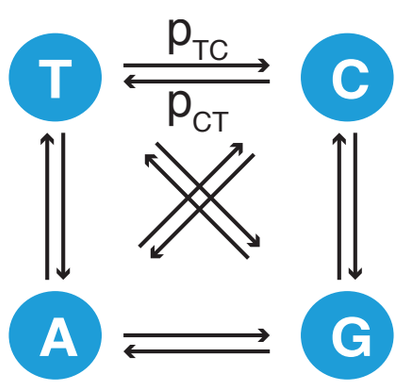
Markov process

Stochastic process



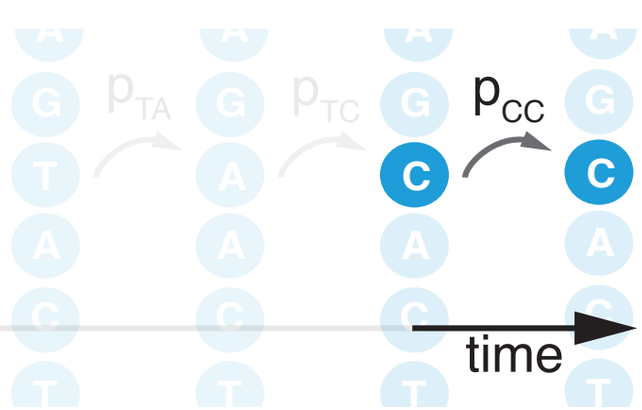
Series of random experiments through time

Finite state space



Lives on a state space and jumps between different states

Memoryless



Probability of jumping to a state only depends on the current state

Probabilities and rates

(courtesy of Carsten Magnus)

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

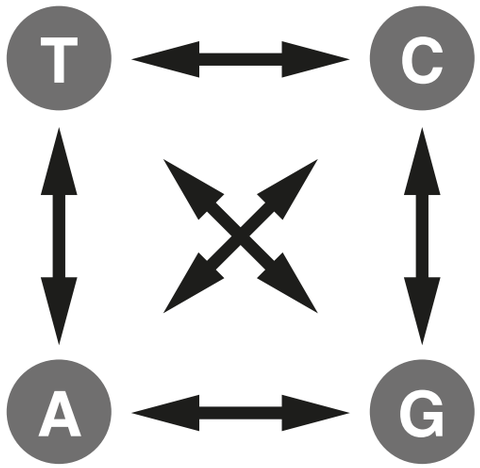
- Transition probabilities in **P(t)** take into account every possible evolutionary trajectory at each site (Chapman-Kolmogorov theorem)
- For convenience we work with the rate matrix **Q** where $\mathbf{q_{ij}}$ is the relative rate of substitutions from state **i** to **j**

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

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

Jukes-Cantor model (JC69)

(courtesy of Carsten Magnus)



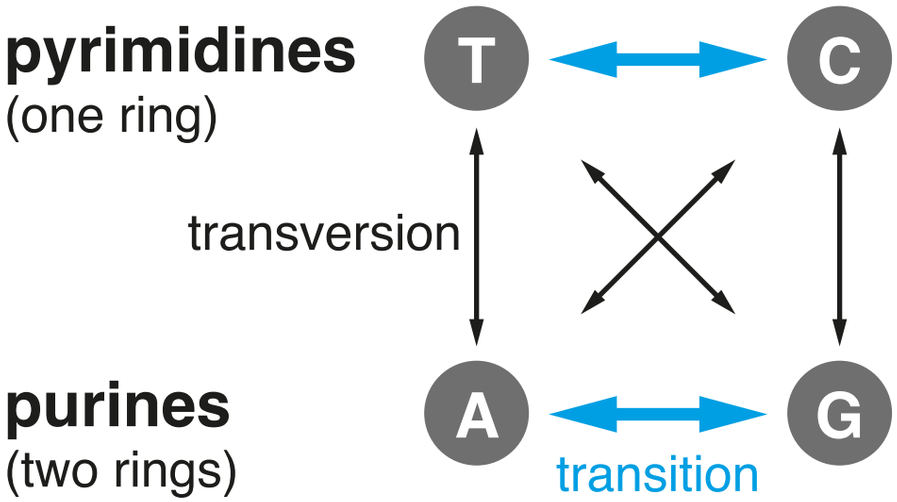
$$\begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} \cdot & \lambda & \lambda & \lambda \\ \lambda & \cdot & \lambda & \lambda \\ \lambda & \lambda & \cdot & \lambda \\ \lambda & \lambda & \lambda & \cdot \end{pmatrix} \end{array}$$

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

- Simplest model
- All rates and frequencies are equal!

Kimura 2-parameter model (K80)

(courtesy of Carsten Magnus)



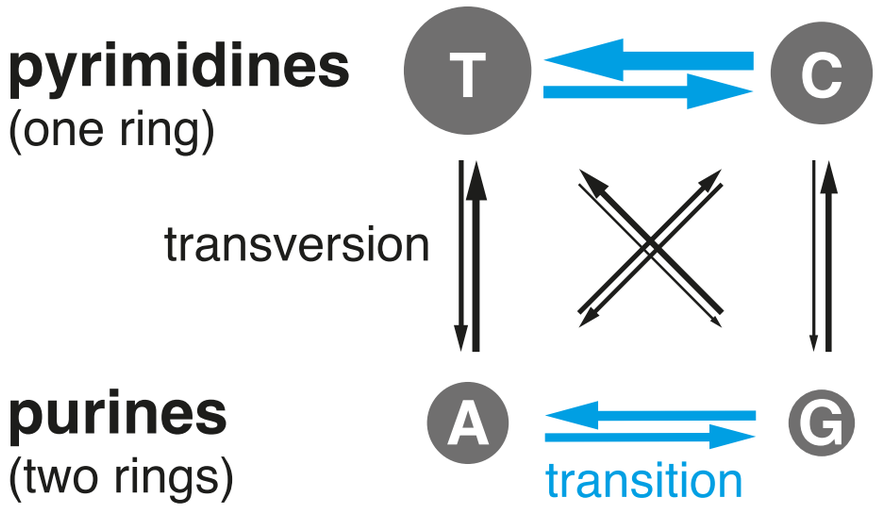
$$\begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} \cdot & \alpha & \beta & \beta \\ \alpha & \cdot & \beta & \beta \\ \beta & \beta & \cdot & \alpha \\ \beta & \beta & \alpha & \cdot \end{pmatrix} \end{array}$$

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

- 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)

(courtesy of Carsten Magnus)



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

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

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

- Accounts for transition/transversion bias
- 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)

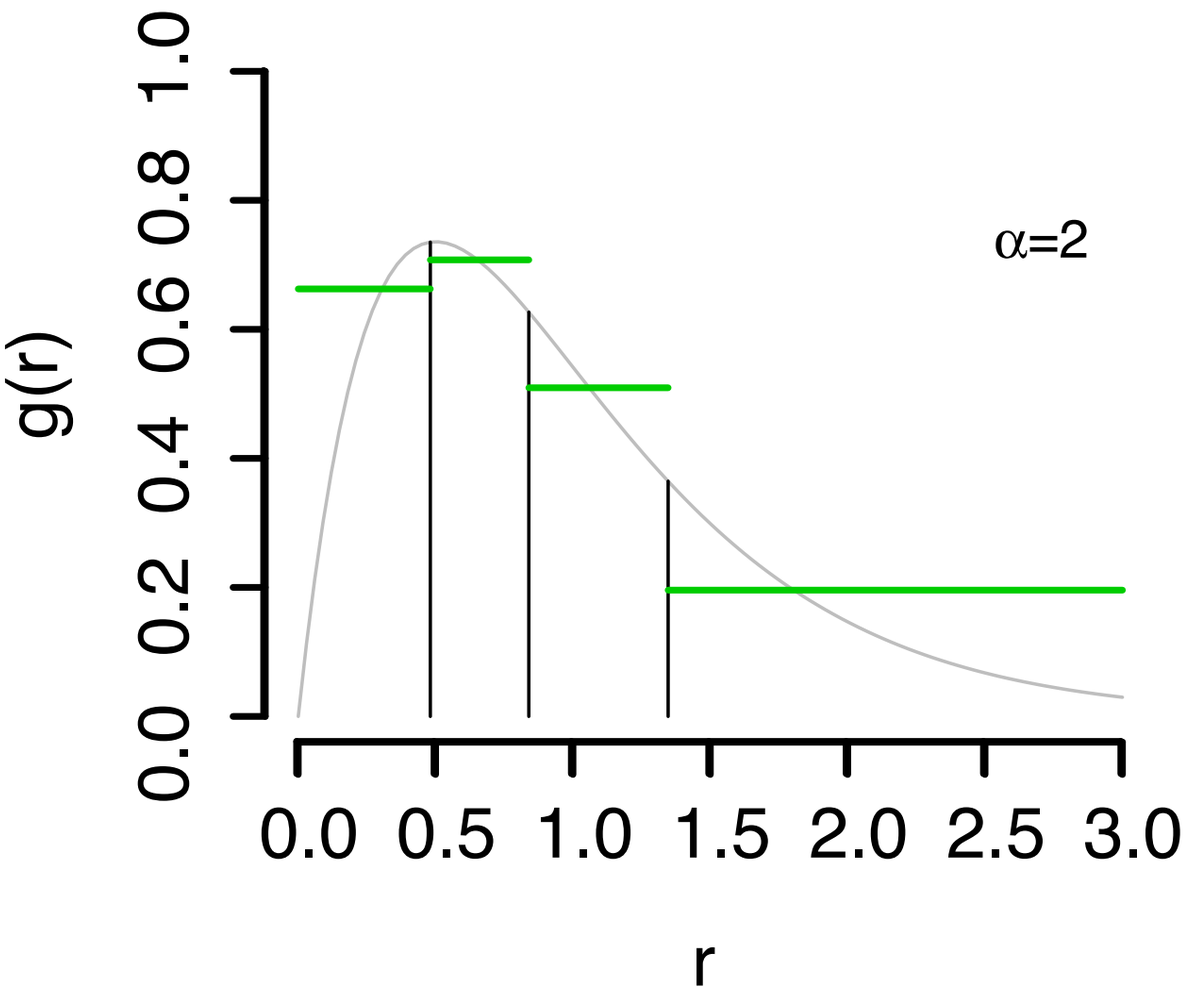
$$
\begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array}
\begin{array}{c}
\begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\
\begin{pmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot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

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

Gamma rate heterogeneity

(courtesy of Carsten Magnus)

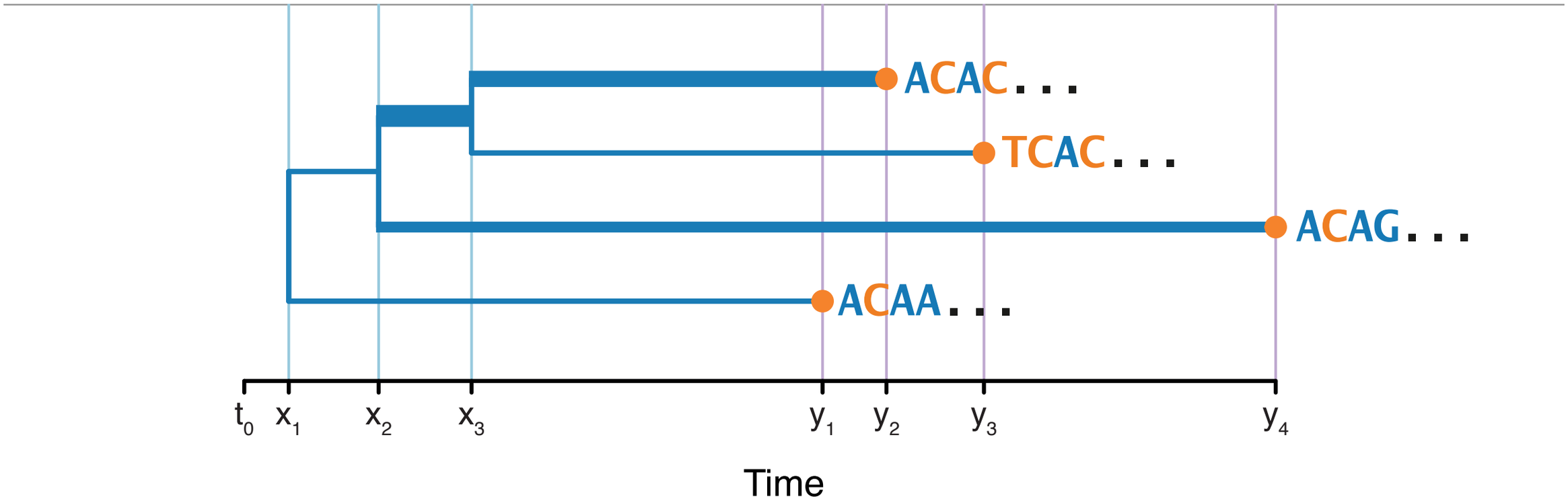
- Not all sites evolve at the same rate
- Assume rate heterogeneity is Γ-distributed
- Discretise Γ-distribution to n discrete rate categories for computational reasons



Multi-locus models

- Γ-distributed rate variation is not 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

Molecular clock model



- Scales branch lengths to calendar time
- How long does it take for substitutions to appear?
- Different branches may have different clock rates
- Priors on internal nodes can help to calibrate the clock

Putting it all together

genetic sequences

genealogy

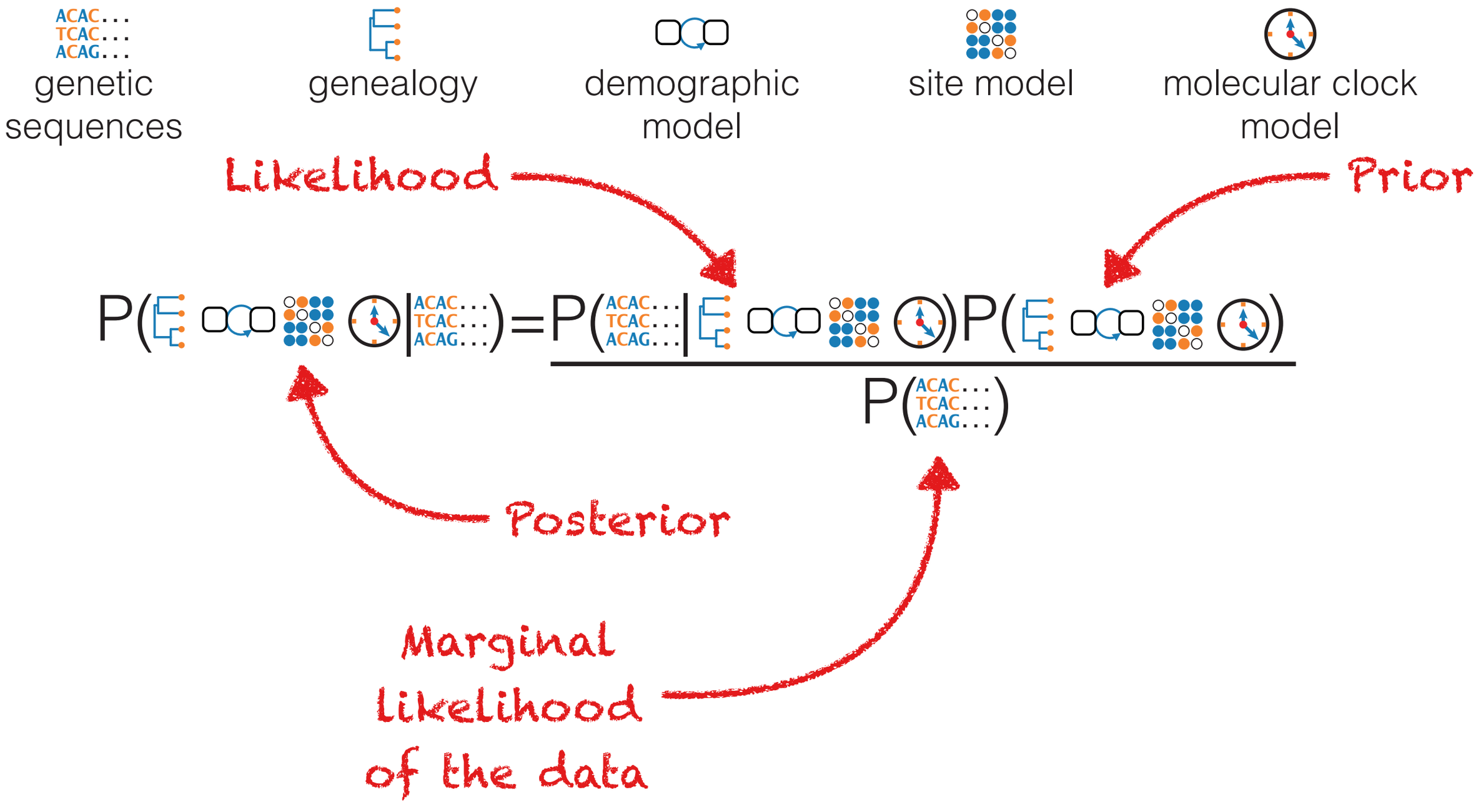
demographic model

site model

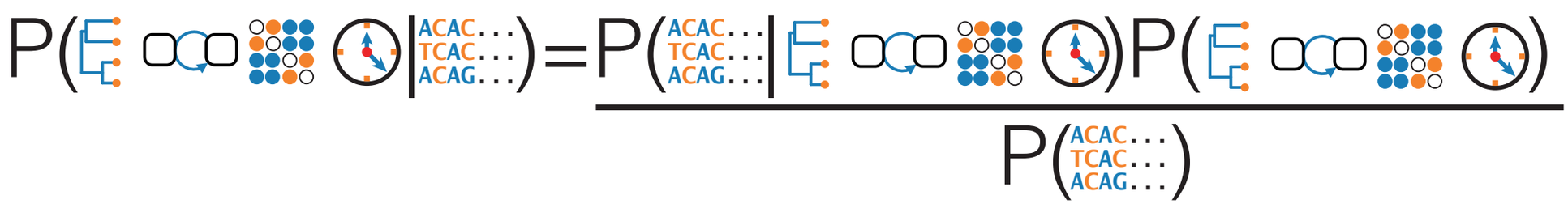
molecular clock model

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

Putting it all together

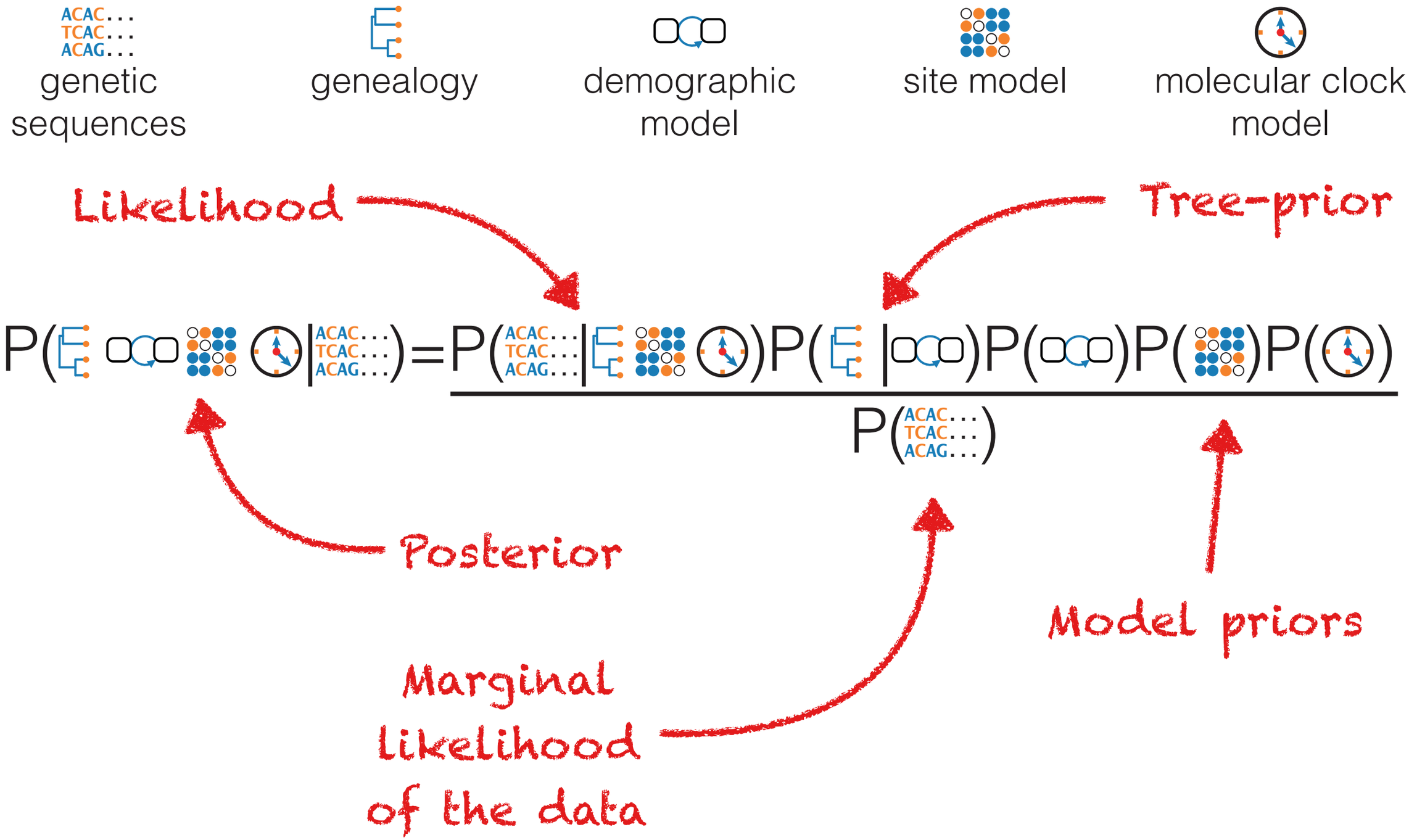


Putting it all together



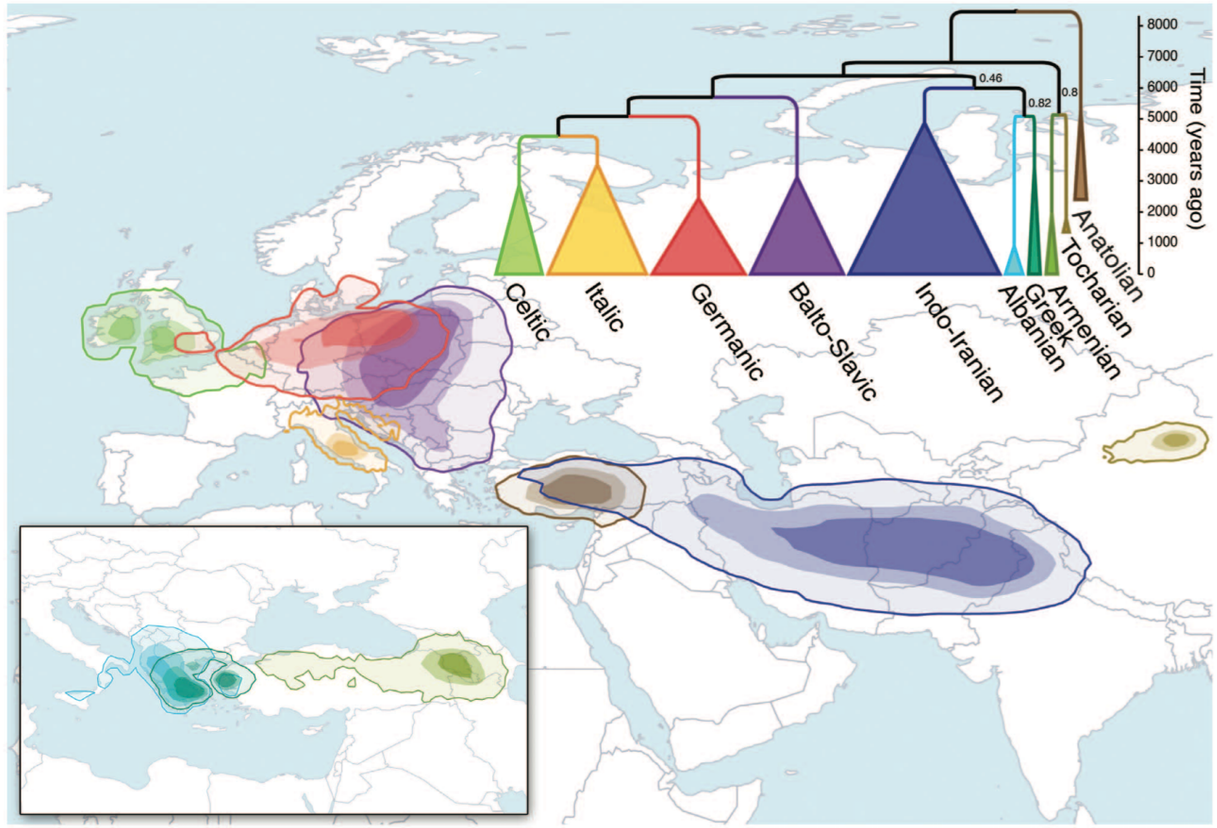
Assume independence

Posterior distribution in BEAST2



Exceptions I

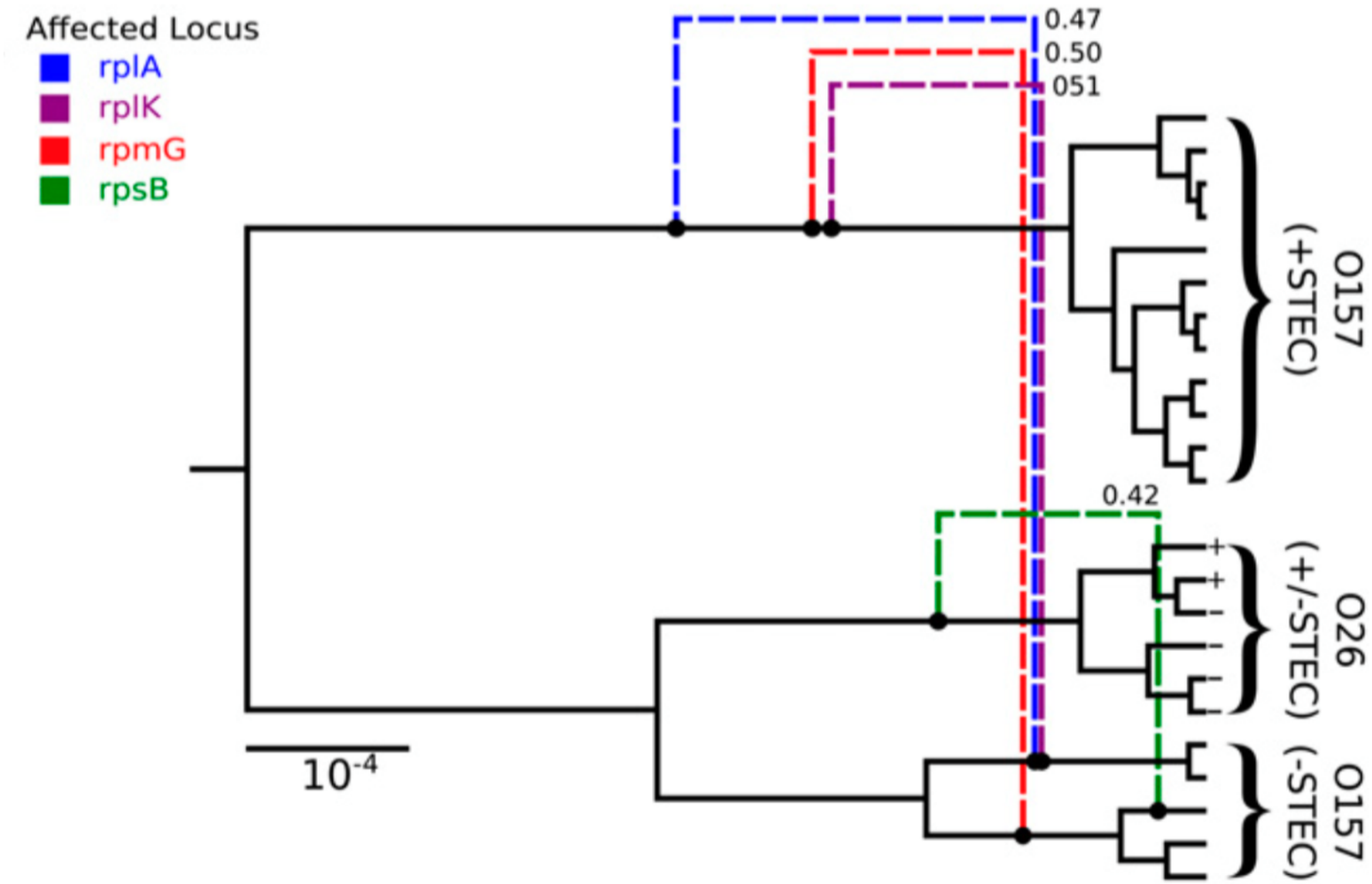
- Site models don’t have to be on nucleotides
- Could be on amino acids, morphological traits, roots of words etc.



Bouckaert *et al.* **Science 2012**

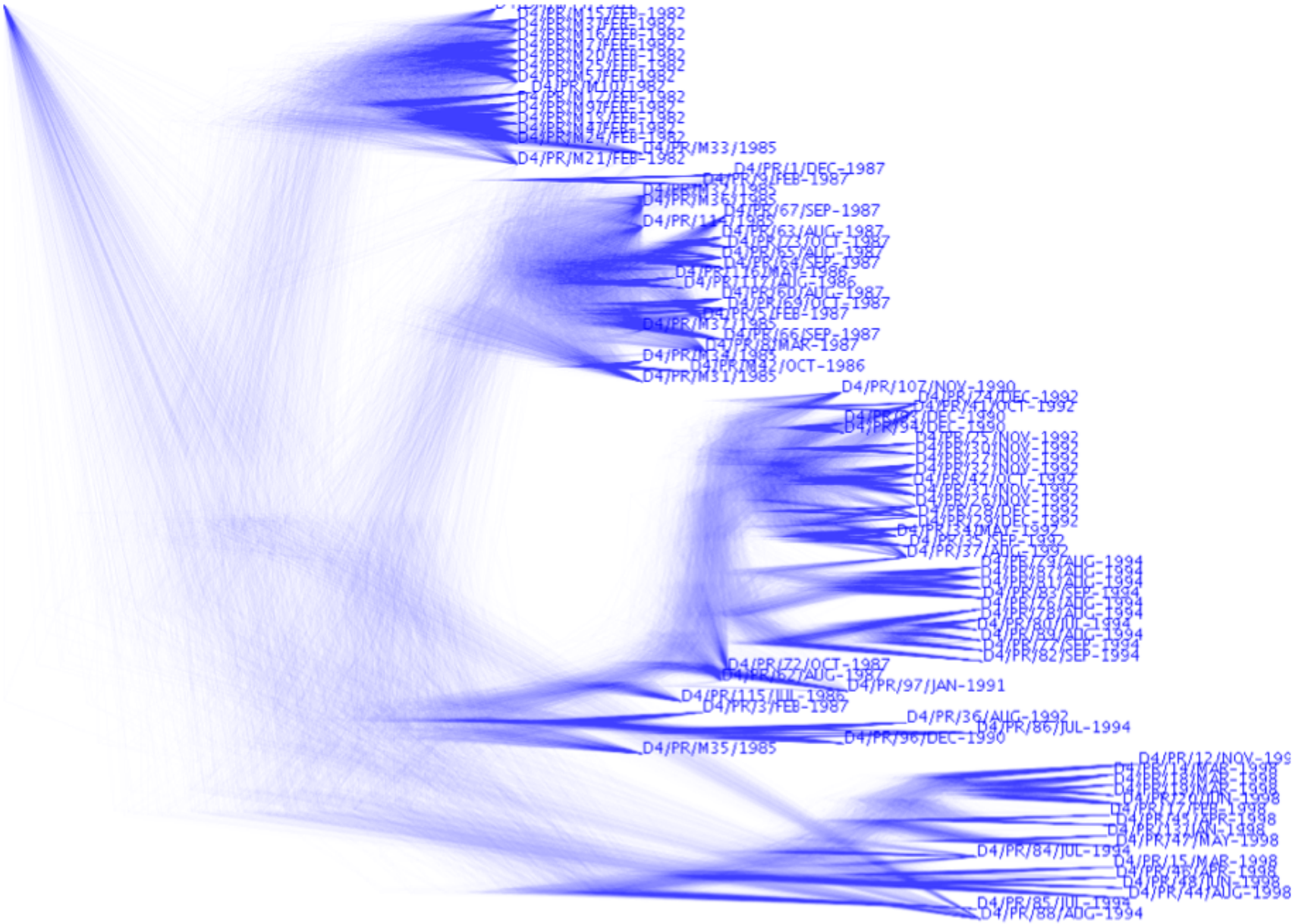
Exceptions II

BEAST2 doesn't always use trees…



Vaughan *et al.* **Genetics 2017**

The posterior distribution for larger trees



How can we find the posterior?

- We want to calculate the posterior distribution

- But we cannot easily calculate the marginal likelihood
  → use **MCMC**! (Markov-chain Monte Carlo)

Markov-chain

- Stochastic process
- Jumps between different states
- Memoryless

Monte Carlo algorithm

- Randomized algorithm
- Deterministic runtime
  (it **will** finish)
- Output may **not** be correct
  (with some small probability)

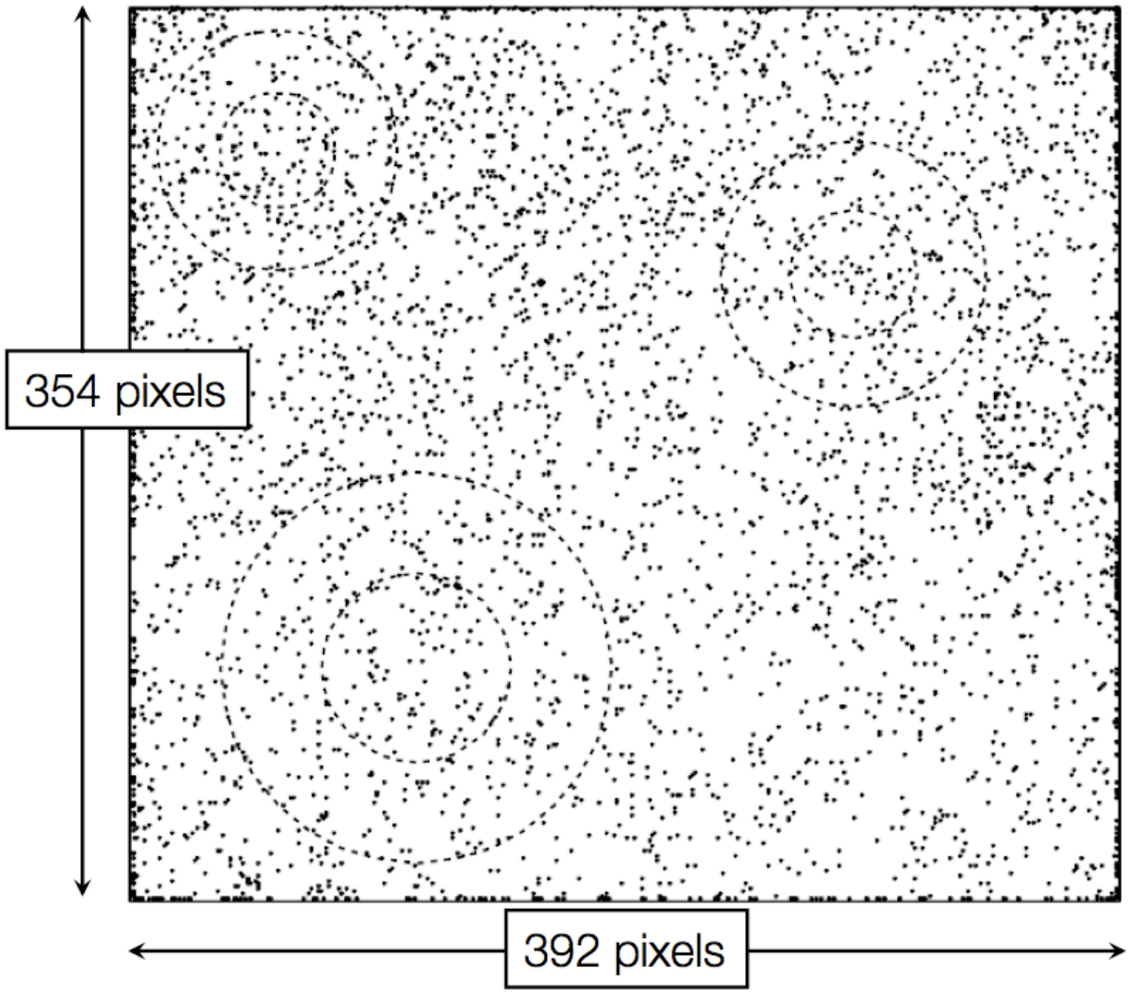
MCMC

(Markov-chain Monte Carlo)

- MCMC performs a random walk on the posterior, preferentially sampling high-density areas
- 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!

$$\frac{P(\textbf{model}_1 \mid \textbf{data})}{P(\textbf{model}_2 \mid \textbf{data})} = \frac{\frac{P(\textbf{data} \mid \textbf{model}_1)P(\textbf{model}_1)}{P(\cancel{\textbf{data}})}}{\frac{P(\textbf{data} \mid \textbf{model}_2)P(\textbf{model}_2)}{P(\cancel{\textbf{data}})}}$$

Pure random walk (courtesy of Paul Lewis)



Random walk

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

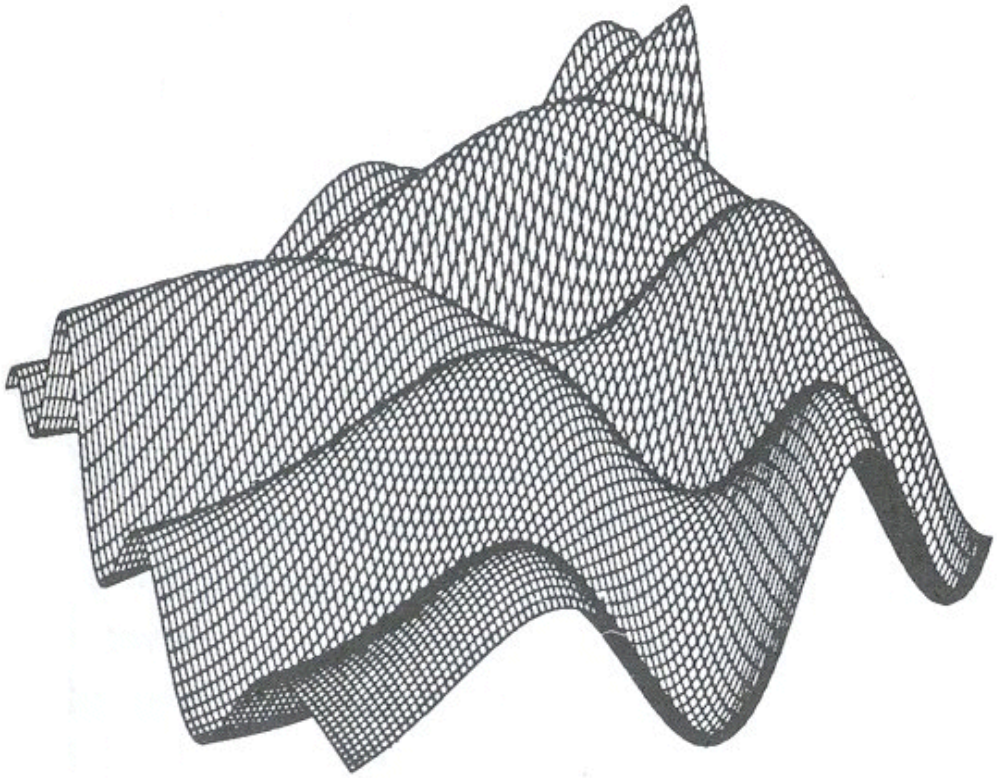
Target distribution

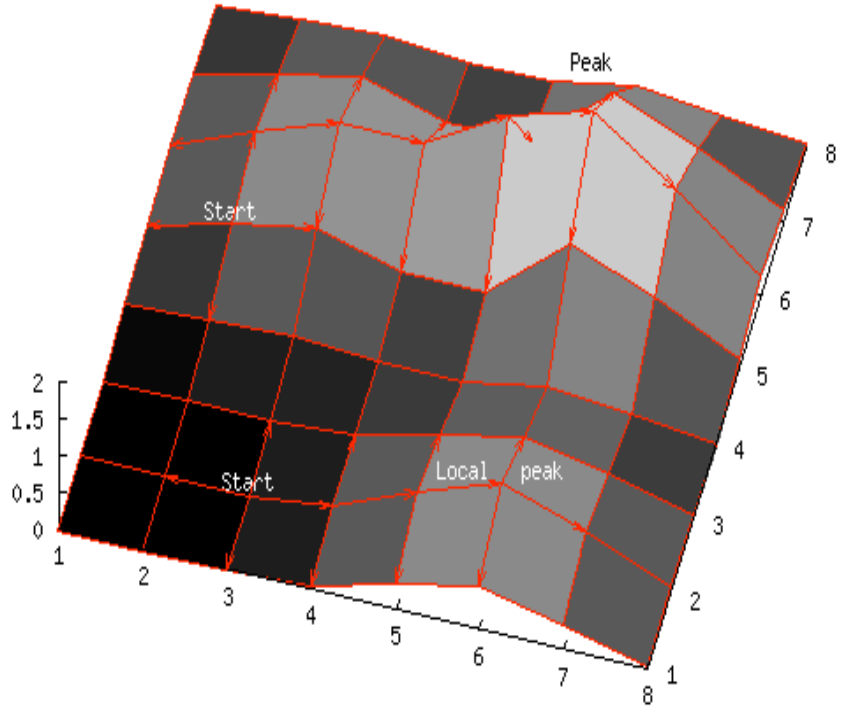
- Equal mixture of 3 bivariate normal hills
- Inner contours: 50%
- Outer contours: 95%

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

Tree space as a hilly landscape

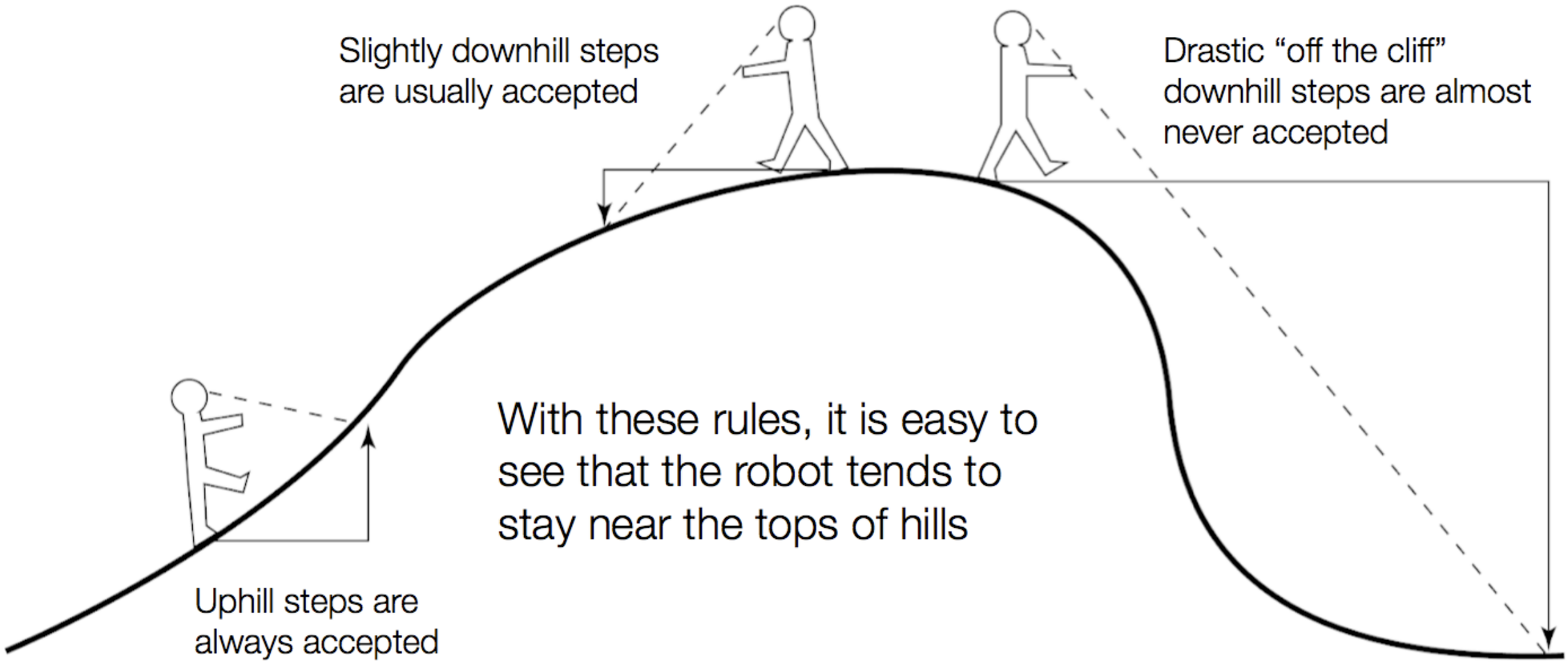
The space of all possible trees can be visualized as a hilly landscape. Nearby points in this landscape represent similar trees, and the height of the landscape is the probability of the tree at that point.



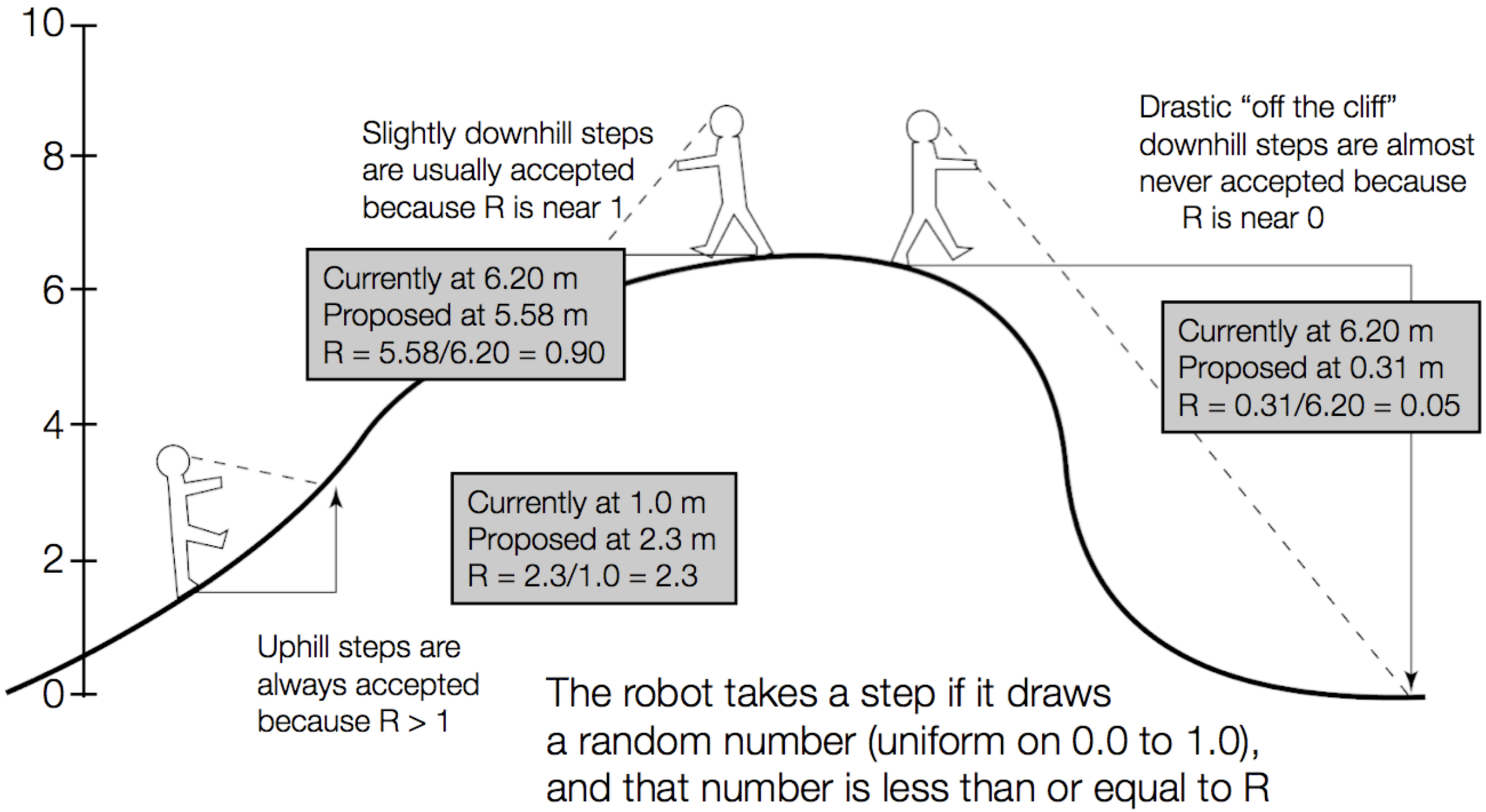


- This space can be **sampled** in a Bayesian analysis with MCMC
- The peak can be identified by a **search algorithm** in the context of maximum likelihoods

MCMC robot (courtesy of Paul Lewis)

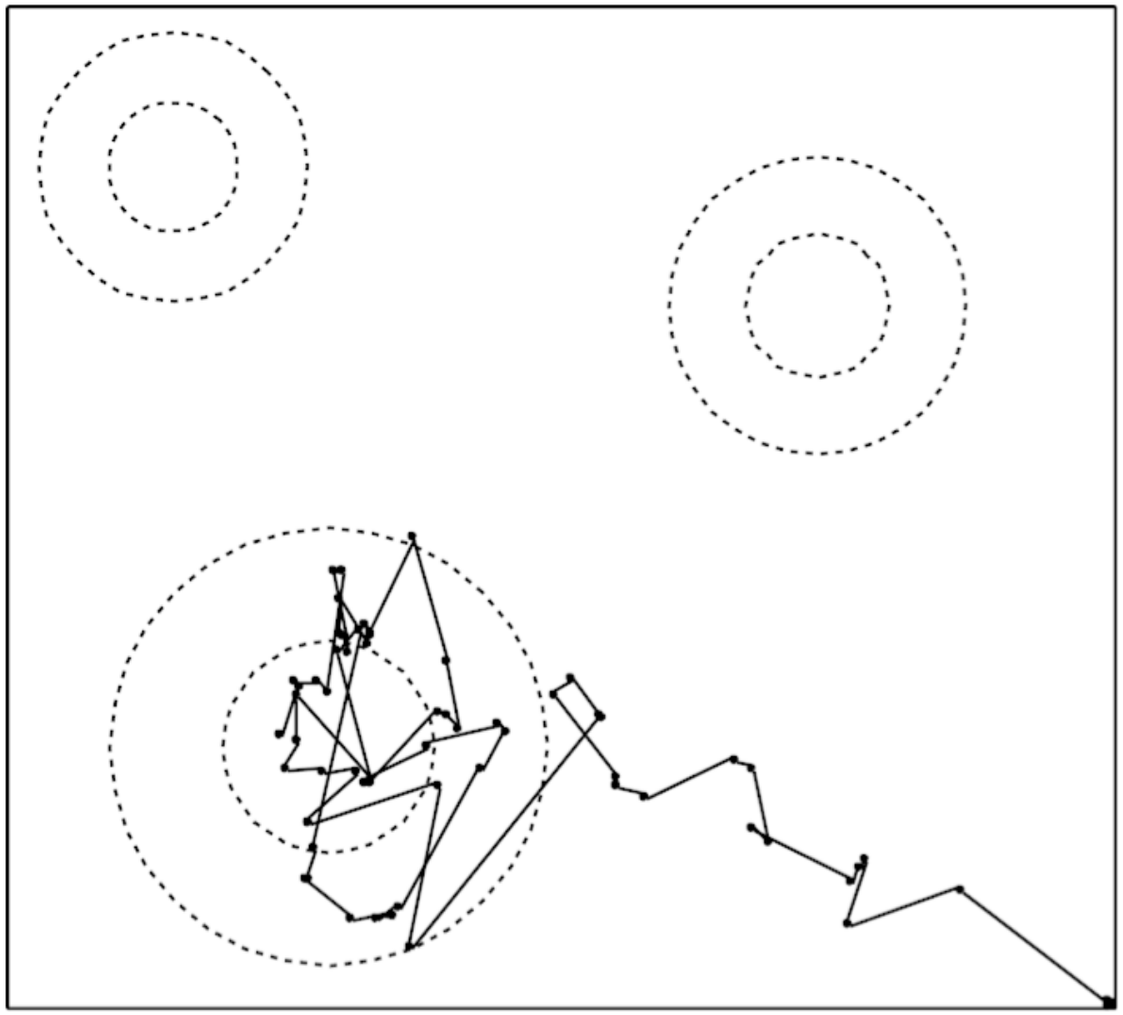


MCMC robot (courtesy of Paul Lewis)



(R is the ratio between the posterior densities)

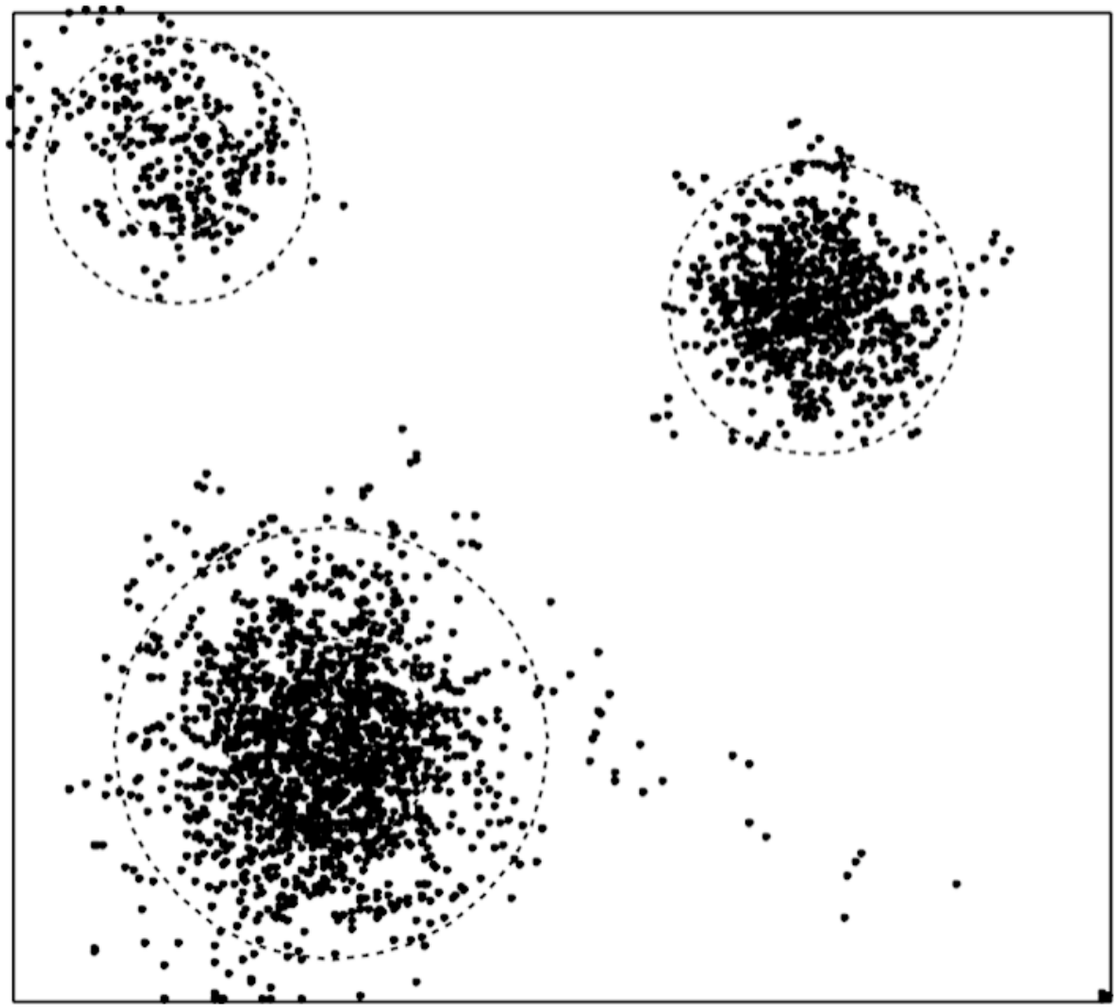
Burn in (courtesy of Paul Lewis)



- 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

Marginal distributions

- We only have the joint posterior: $P($  $)$

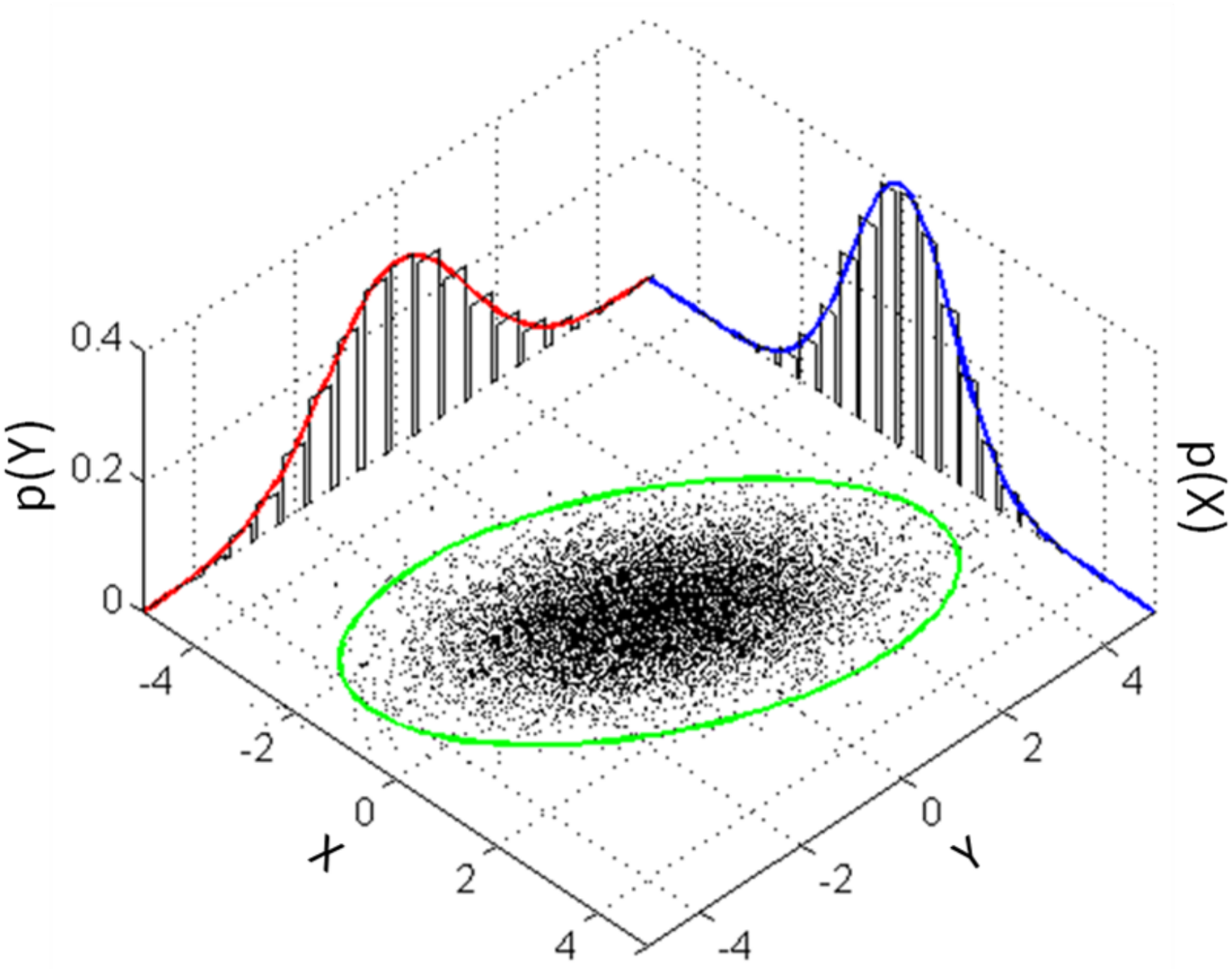
- But we want distributions for each of the parameters we are interested in → marginalize

$$P(\phi) = \int_{\theta} P(\phi|\theta)P(\theta)d\theta$$

Margin

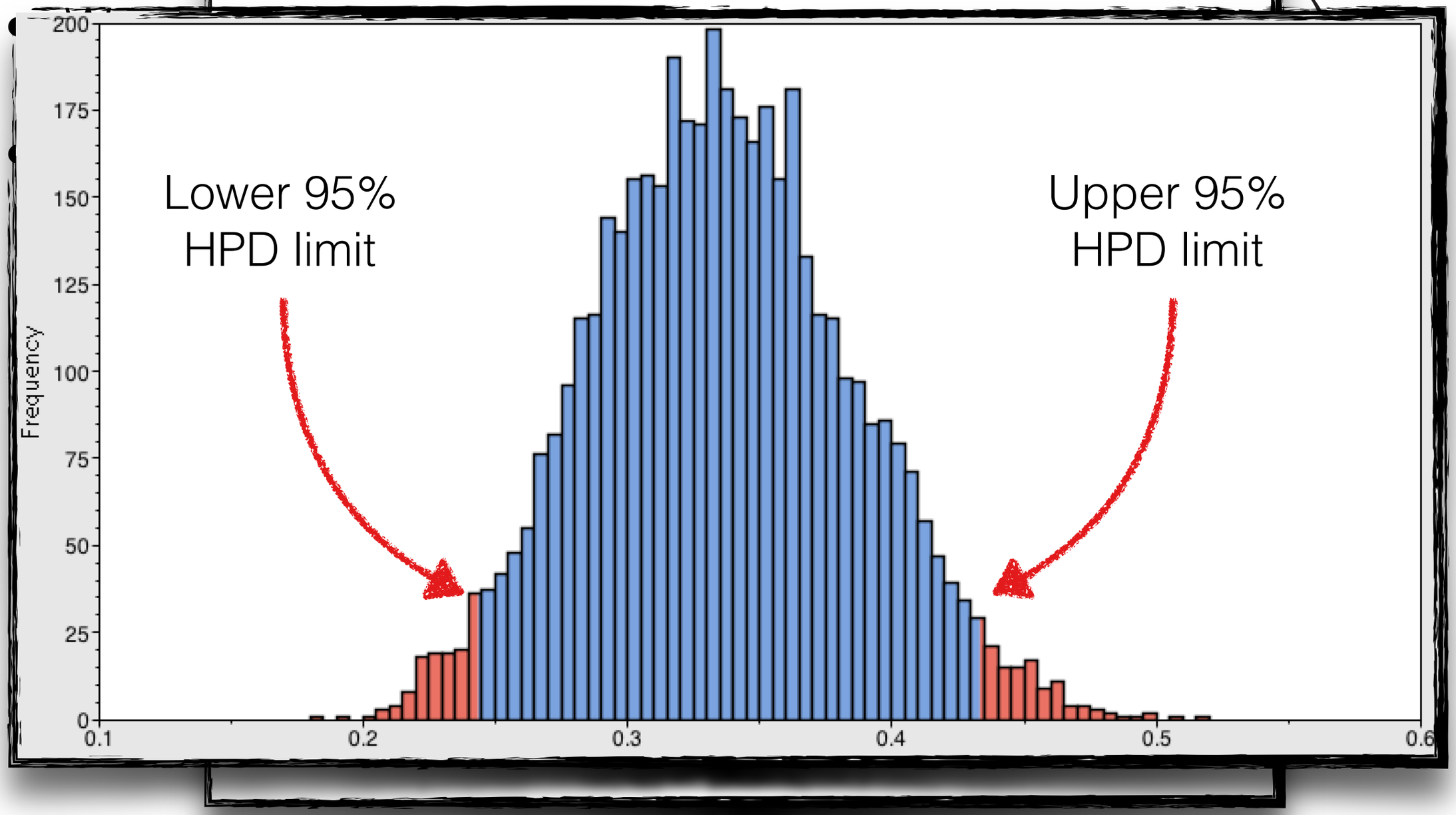
- We on ::)
- But we s
  we are

In practice



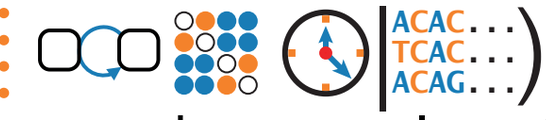
Margin

In practice



Operators

Target distribution

- This is the **posterior** in BEAST2: P()
- MCMC steps through the state space and samples the target distribution

Proposal distribution

- Used to decide where to step to next
- The choice only affects the **efficiency** of the algorithm
- In BEAST and BEAST2 operators are used to propose the next step
- Operators are a part of the MCMC **algorithm**, not the **model**
- Tuning operators can help to improve efficiency, but should not change the results

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)

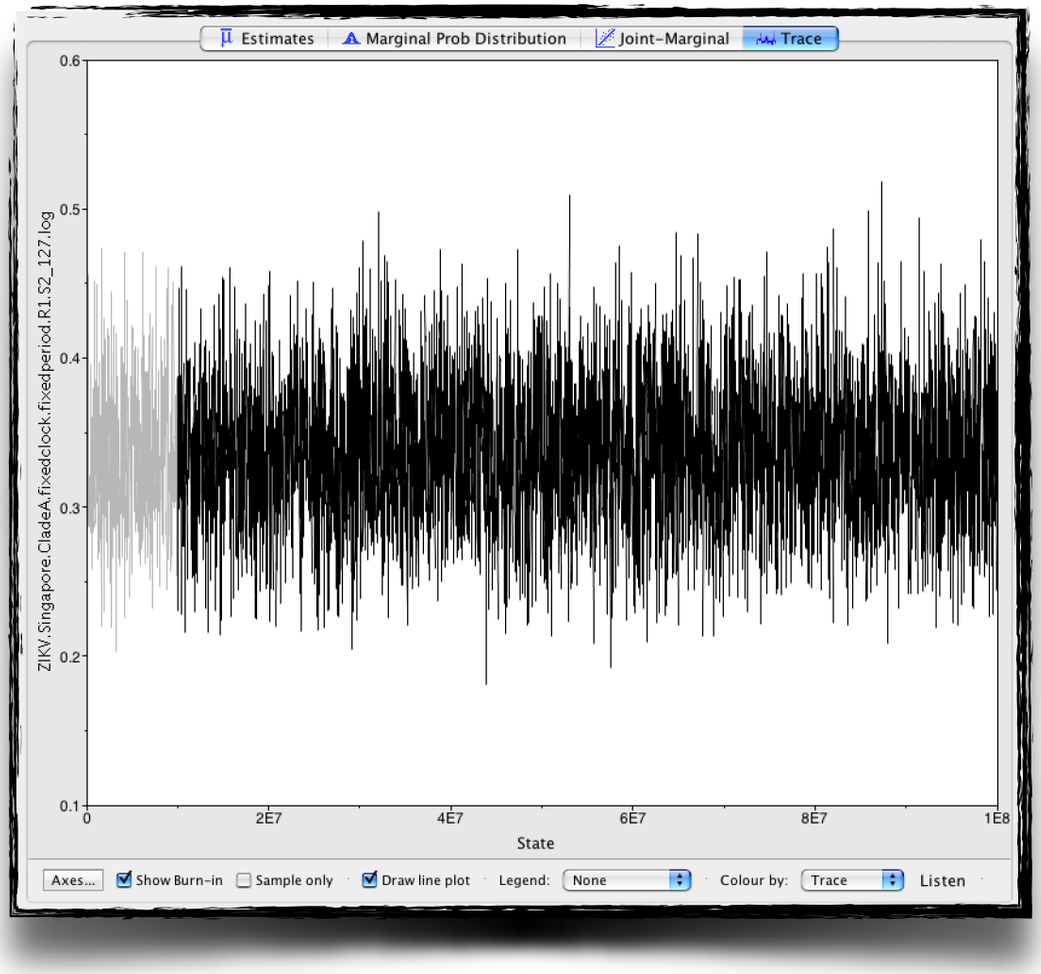
After

- Discard burn-in
  (until stationary state is reached)
- Assess convergence and mixing

More than 10,000 samples is a waste of space
(but need to sample at the right frequency)

What we hope will happen

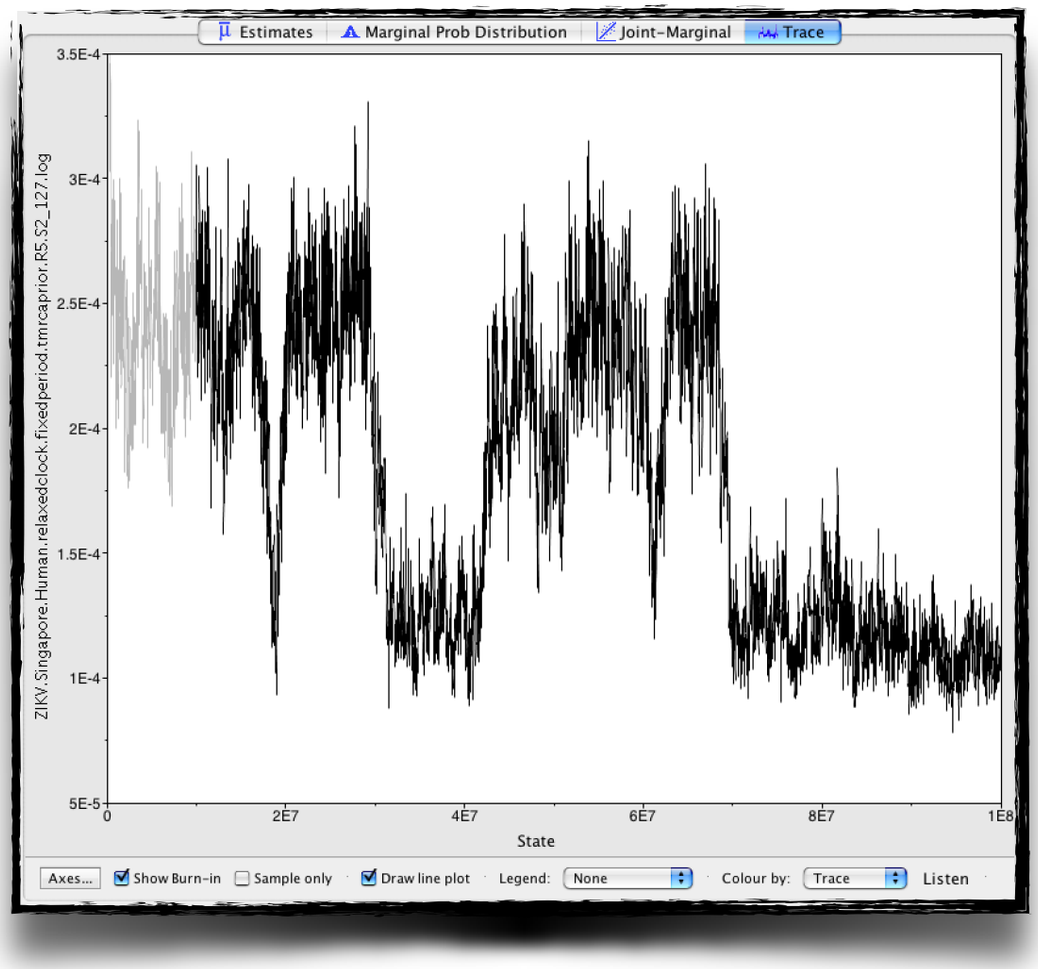
- The MCMC algorithm samples efficiently from high density areas of the posterior distribution
- We end up with a **good** approximation of the posterior distribution in **finite** time
- Appearance of white noise
- Everything is awesome!



Mixing well! 😄

Questions to ask…

- Is the chain **mixing** well?
- Are samples uniformly drawn from all over the stationary distribution?
- “Sticky chain”



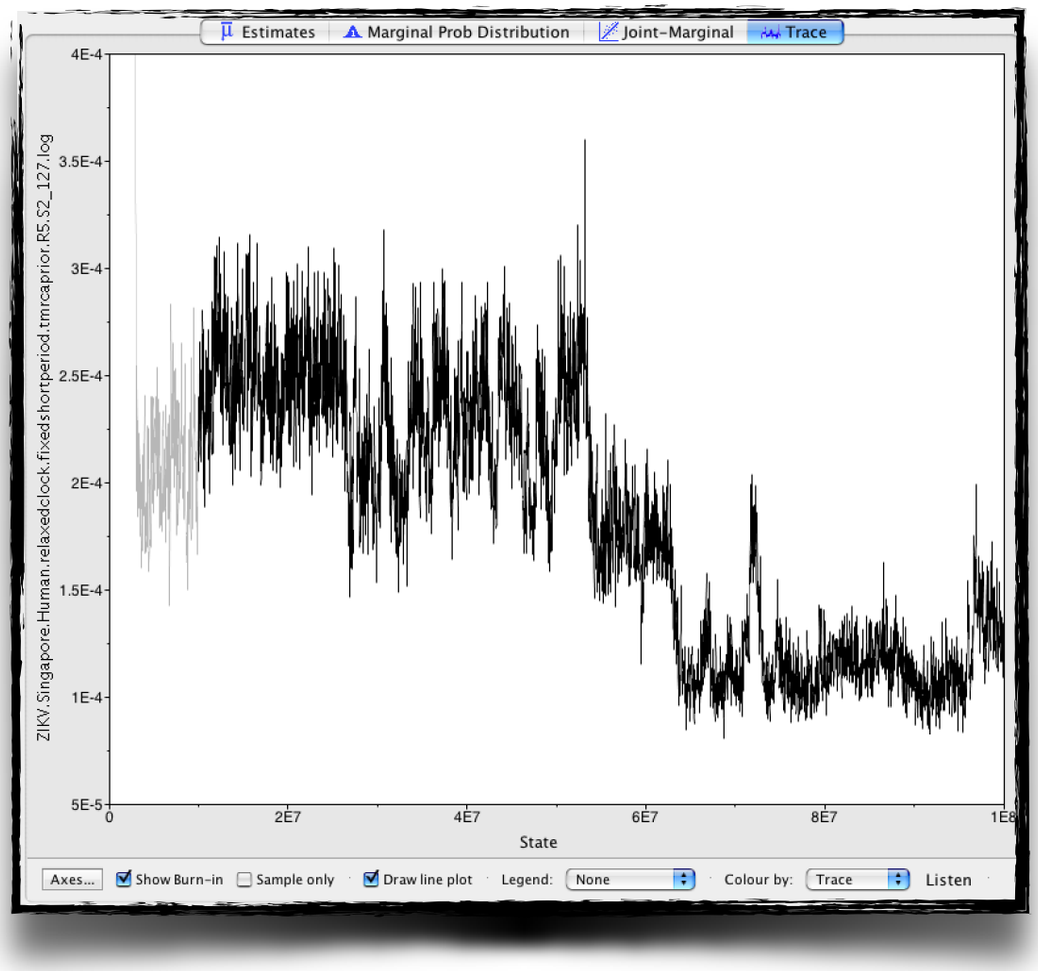
Not mixing! 😨

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?

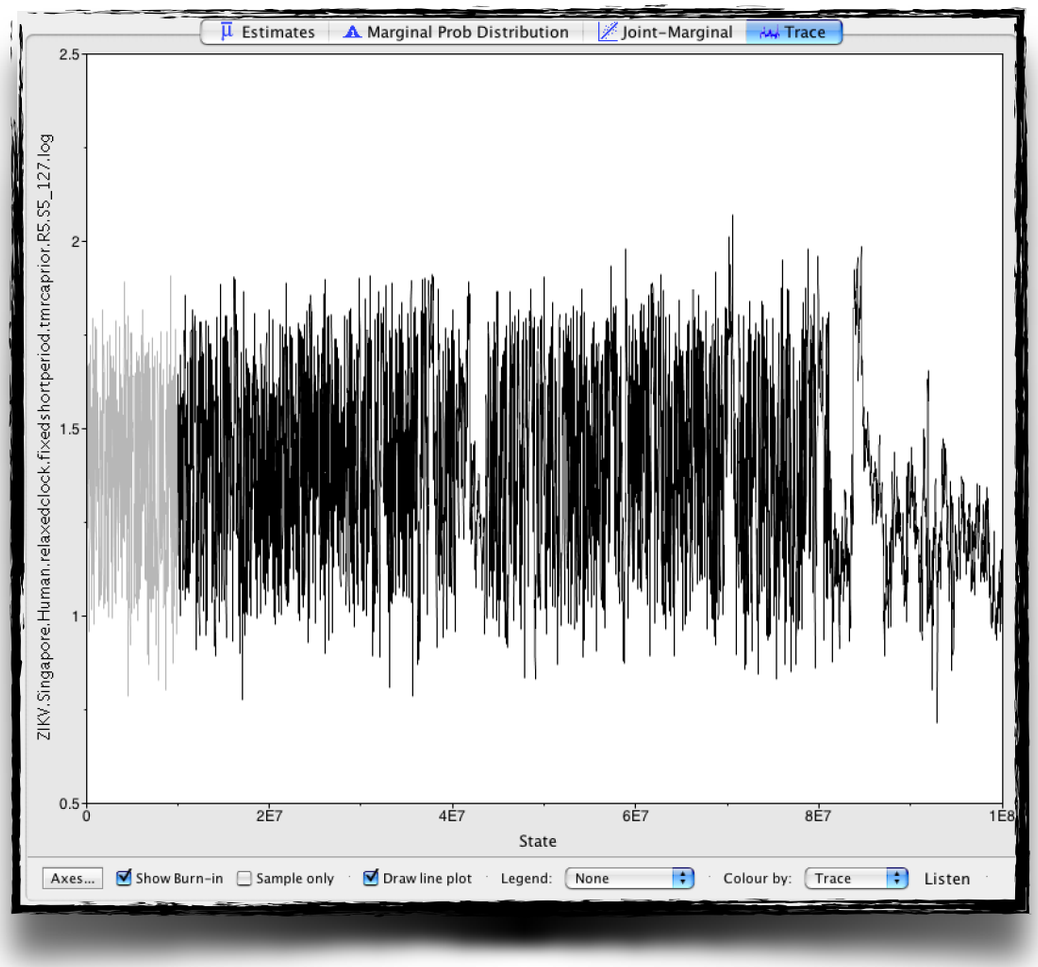


Not converged! 😓

Solution: Run for longer

Questions to ask…

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



Still not converged! 😵

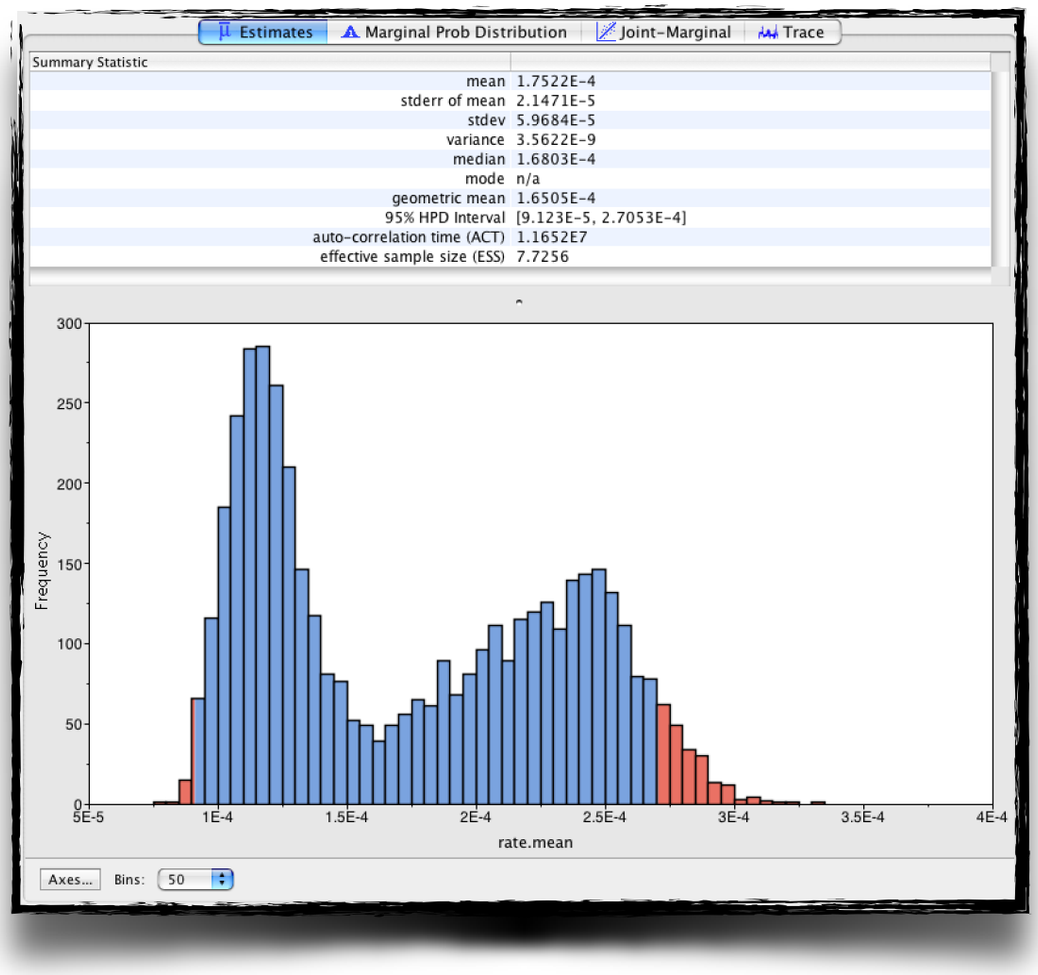
Solution

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

What if the answer is not what we wanted?

What is happening here?

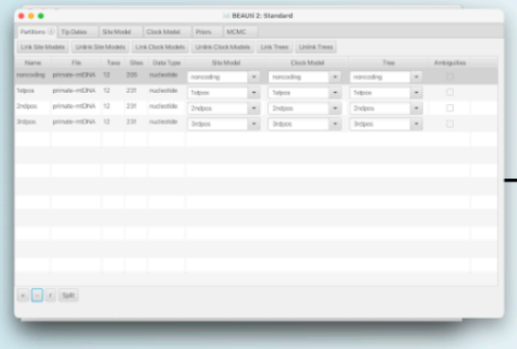
- If the chain converged and mixed well then this is due to the data and model choice
- The model supports a bimodal posterior distribution
- May not be the answer we wanted but it may be the truth
- Should we change the model or parameterisation?



Is this a problem? 🤔

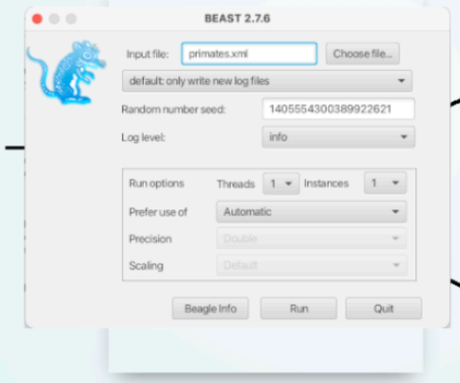
Solution: Be more open-minded

COMMON USAGE



BEAUti

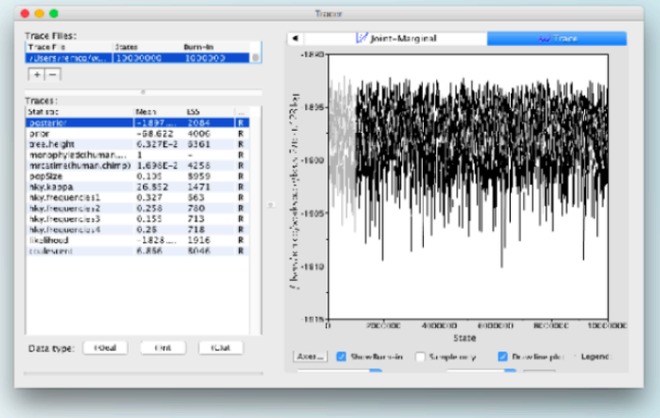
XML File



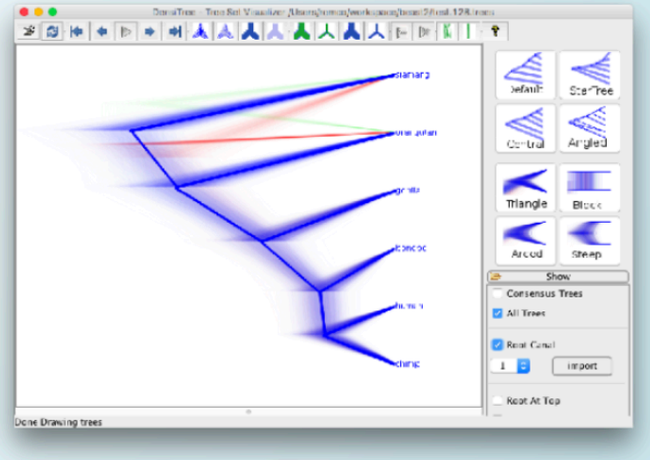
BEAST

Log file

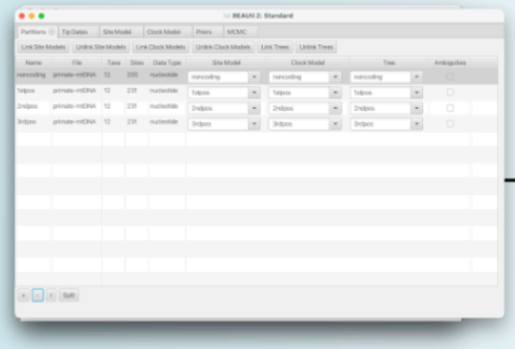
Trees file



Tracer



DensiTree



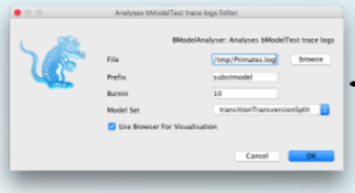
BEAUti

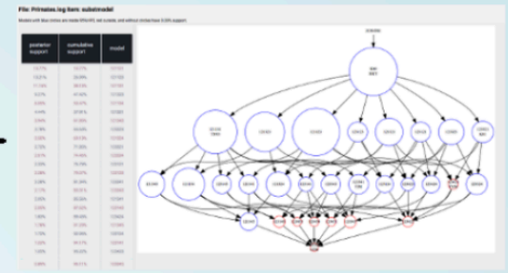
XML File

BEAST

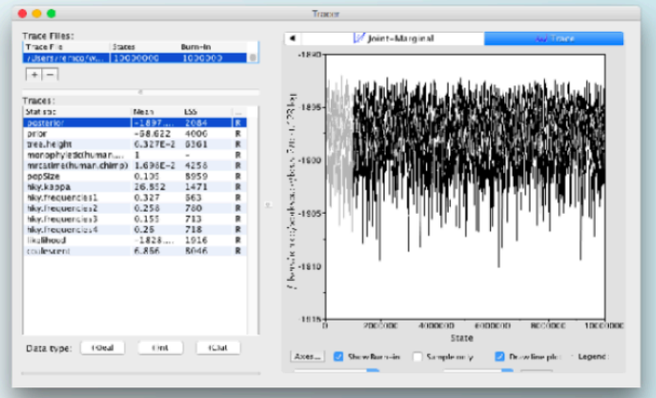
Log file

Trees file

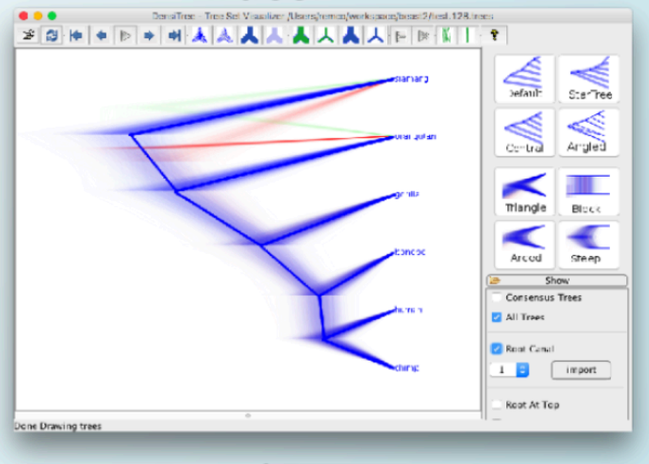




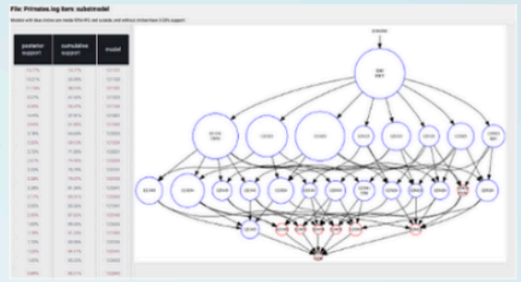
BModelAnalyser



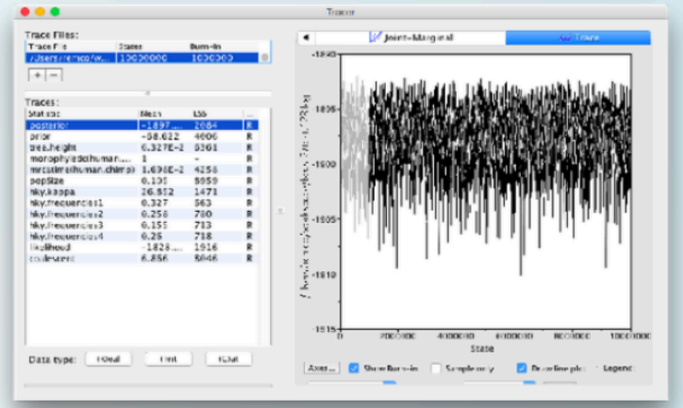
Tracer



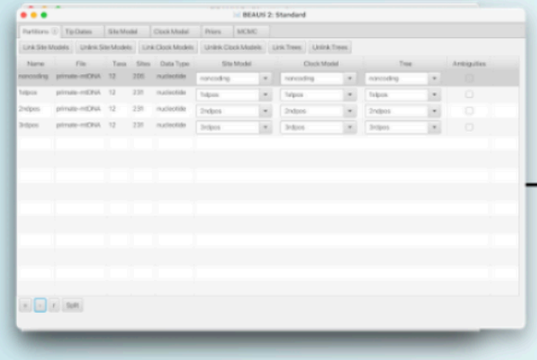
DensiTree



BModelAnalyser

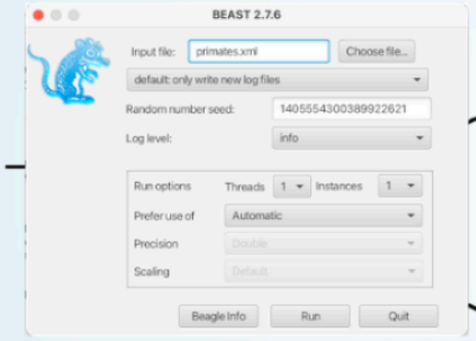


Tracer



BEAUti

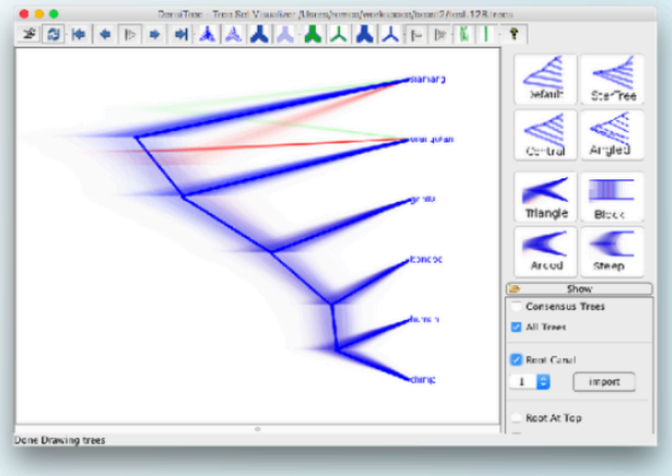
XML File



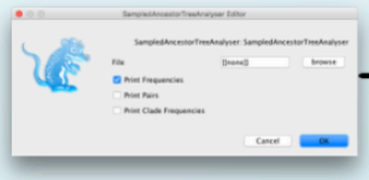
BEAST

Log file

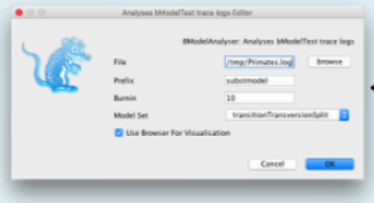
Trees file

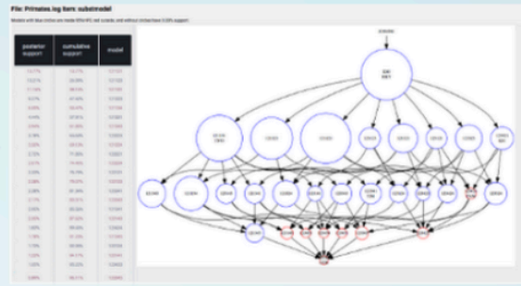


DensiTree

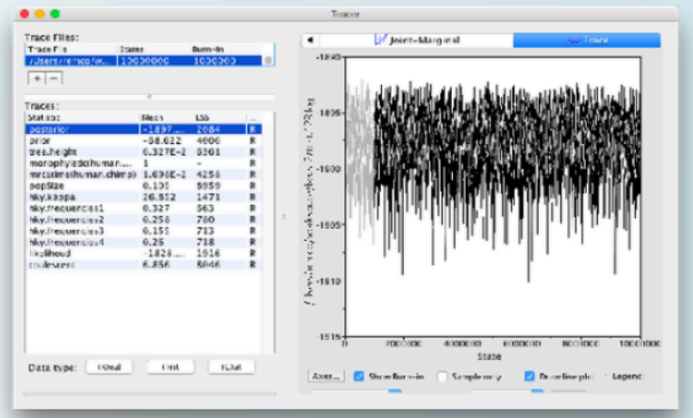


SATreeAnalyser



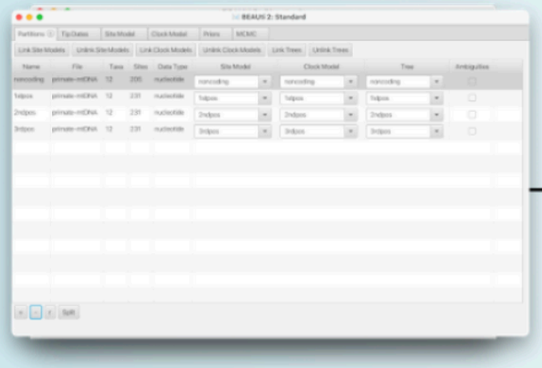


BModelAnalyser



Tracer

Log file

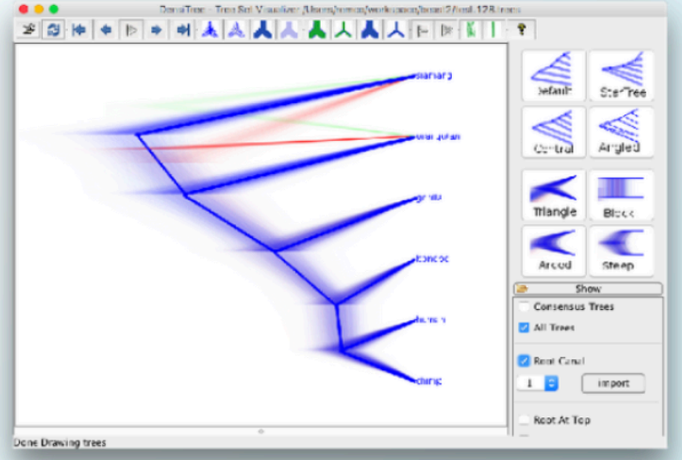


BEAUti

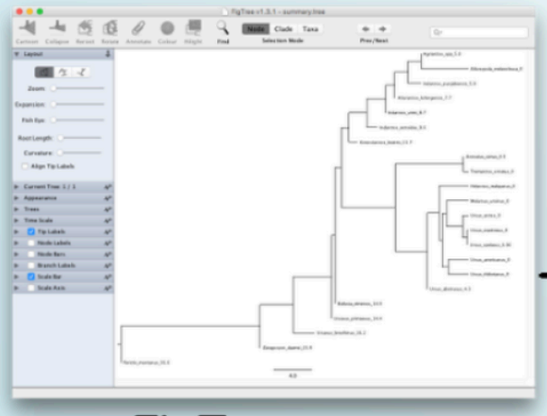
XML File

BEAST

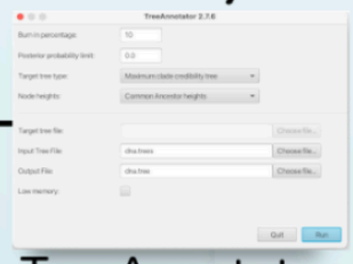
Trees file



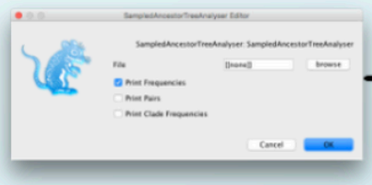
DensiTree



FigTree



TreeAnnotator



SATreeAnalyser

BEAUti2

(http://beast2.org)

GUI for setting up BEAST2 input file in xml format

Input:

- Sequence alignment

Output:

- BEAST2 configuration file (xml file)

BEAUti 2: Standard

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	Ambiguities
noncoding	primate-mtDNA	12	205	nucleotide	noncoding	noncoding	noncoding	☐
1stpos	primate-mtDNA	12	231	nucleotide	1stpos	1stpos	1stpos	☐
2ndpos	primate-mtDNA	12	231	nucleotide	2ndpos	2ndpos	2ndpos	☐
3rdpos	primate-mtDNA	12	231	nucleotide	3rdpos	3rdpos	3rdpos	☐

+ | - | r | Split

BEAUti 2: Standard

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

clockRate.c:1stpos	Uniform[0.0,Infinity]	initial = [1.0] [0.0,∞]	substitution rate of partition c:1stpos
clockRate.c:2ndpos	Uniform[0.0,Infinity]	initial = [1.0] [0.0,∞]	substitution rate of partition c:2ndpos
clockRate.c:3rdpos	Uniform[0.0,Infinity]	initial = [1.0] [0.0,∞]	substitution rate of partition c:3rdpos
clockRate.c:noncoding	Uniform[0.0,Infinity]	initial = [1.0] [0.0,∞]	substitution rate of partition c:noncoding

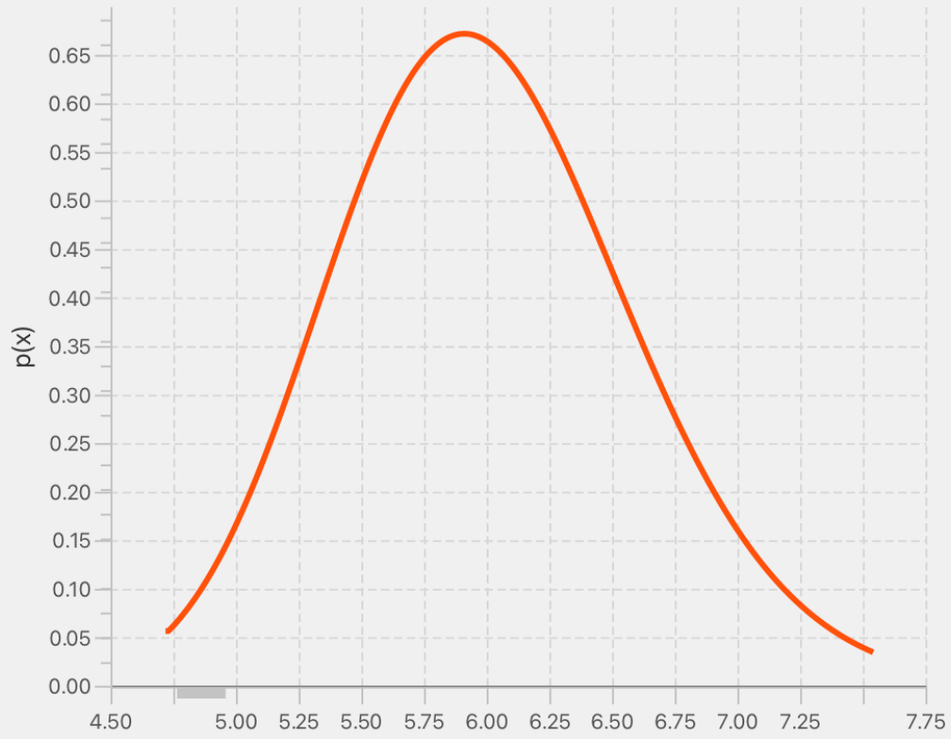
human-chim... | Log Normal | monophyletic | -

M 6 estimate

S 0.1 estimate

Mean In Real Space

Offset 0.0



2.5% Quantile	4.91
5% Quantile	5.06
Median	5.97
95% Quantile	7.04
97.5% Quantile	7.26

mean 6.00

primates.xml

Free Mode

/private/tmp/primates.xml

(no function selected)

```
    <map name="Gamma" >beast.base.inference.distribution.Gamma</map>

    <map name="LaplaceDistribution" >beast.base.inference.distribution.LaplaceDistribution</map>

    <map name="prior" >beast.base.inference.distribution.Prior</map>

    <map name="InverseGamma" >beast.base.inference.distribution.InverseGamma</map>

    <map name="OneOnX" >beast.base.inference.distribution.OneOnX</map>

    <run id="mcmc" spec="MCMC" chainLength="10000000">
        <state id="state" spec="State" storeEvery="5000">
            <tree id="Tree.t:tree" spec="beast.base.evolution.tree.Tree" name="stateNode">
                <taxonset id="TaxonSet.noncoding" spec="TaxonSet">
                    <alignment id="noncoding" spec="FilteredAlignment" filter="1,458-659,897-898">
                        <data idref="primate-mtDNA"/>
                    </alignment>
                </taxonset>
            </tree>
            <parameter id="birthRate.t:tree" spec="parameter.RealParameter" lower="0.0" name="stateNode">1.0</parameter>
            <parameter id="clockRate.c:3rdpos" spec="parameter.RealParameter" lower="0.0" name="stateNode">1.0</parameter>
            <parameter id="clockRate.c:2ndpos" spec="parameter.RealParameter" lower="0.0" name="stateNode">1.0</parameter>
            <parameter id="clockRate.c:1stpos" spec="parameter.RealParameter" lower="0.0" name="stateNode">1.0</parameter>
            <parameter id="clockRate.c:noncoding" spec="parameter.RealParameter" lower="0.0" name="stateNode">1.0</parameter>
        </state>
        <init id="RandomTree.t:1stpos" spec="RandomTree" estimate="false" initial="@Tree.t:tree">
            <taxa id="1stpos" spec="FilteredAlignment" data="@primate-mtDNA" filter="2-457\3,660-896\3"/>
            <populationModel id="ConstantPopulation0.t:1stpos" spec="ConstantPopulation">
                <parameter id="randomPopSize.t:1stpos" spec="parameter.RealParameter" name="popSize">1.0</parameter>
            </populationModel>
        </init>
        <distribution id="posterior" spec="CompoundDistribution">
            <distribution id="prior" spec="CompoundDistribution">
                <distribution id="YuleModel.t:tree" spec="beast.base.evolution.speciation.YuleModel" birthDiffRate="@birthRate.t:tree"
                <prior id="YuleBirthRatePrior.t:tree" name="distribution" x="@birthRate.t:tree">
                    <Uniform id="Uniform.4" name="distr" upper="Infinity"/>
                </prior>
```

L: 1 C: 1 XML Unicode (UTF-8) Unix (LF) Saved: 1:35:02 PM 30,569 / ... / 266 - 100%

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 (epidemiological/speciation) model

- Similar to BEAST 1.8 but completely separate

- BEAST2 has most of the functionality of BEAST 1.8 and a lot more

- BEAST2 has a modular design that makes it easy to extend

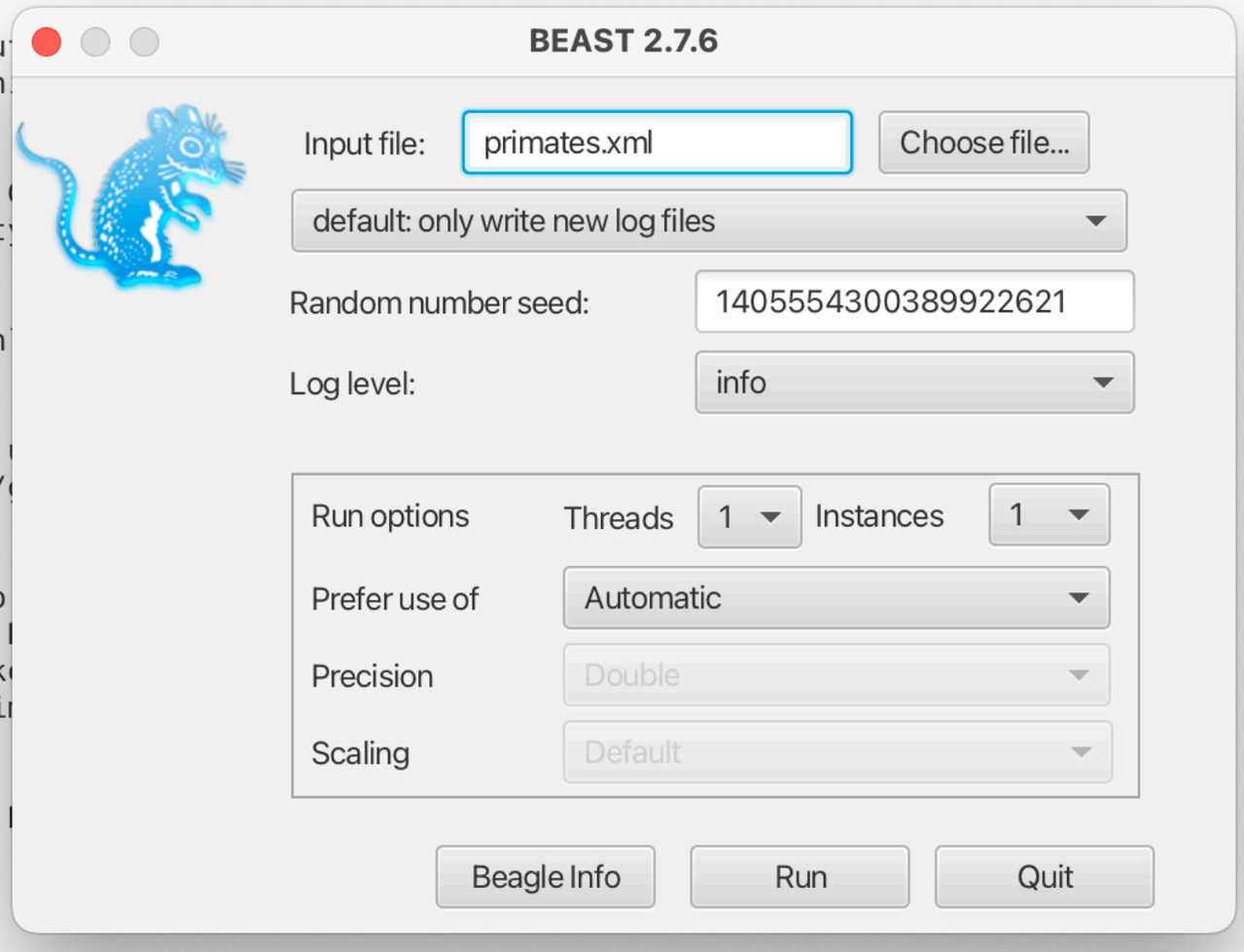
Input:

- xml file

Outputs:

- log file
- trees file
- state file

```
                     BEAST v2.7.6, 2002-2023
            Bayesian Evolutionary Analysis Sampling Trees
                        Designed and developed by
Remco Bouckaert, Alexei J. Drummond, Andrew Rambaut & Marc A. Suchard

                   Centre for Computational Evolution
                         University of Auckland
                       r.bouckaert@auckland.ac.nz
                        alexei@cs.auckland.ac.nz

                   Institu
                        Un

                    David
                  University

                      Down

   Source code distributed
                      http://

    Alex Alekseyenko, Trevo
  Sebastian Hoehna, Denise
 Gerton Lunter, Sidney Mark
            Oliver Pybus, Ti

            Roald Forsberg,
```



BEAST2 packages

- Independent researchers can easily develop their own BEAST2 packages

- Packages can be frequently updated without waiting for the next BEAST2 release

- Packages add new models or completely new functionality

- Phylogeography, bacterial ARG inference, morphological models, model selection and averaging, stochastic simulations etc.

- Install new packages through BEAUti

BEAST 2 Package Manager

List of available packages for BEAST v2.7.*

Name	Installed	Latest	Dependencies	Link	
BEAST.base	2.7.6	2.7.6			BEAST base
BEAST.app	2.7.6	2.7.6	BEAST.base		BEAST base applications
Babel		0.4.2	BEAST.app, BEASTLabs, BEAST....		BABEL = BEAST analysis backing effective linguistics
bacter		3.0.1	feast, BEAST.app, BEAST.base		Bacterial ARG inference.
BADTRIP		2.0.0	BEAST.base		Infer transmission time for non-haplotype data and epi data
BASTA		4.0.0	BEAST.app, BEAST.base		Bayesian structured coalescent approximation
bdmm		2.0.0	BEAST.base, MultiTypeTree, MA...		Multitype birth-death model (aka birth-death-migration model)
BDSKY		1.5.0	BEAST.app, BEAST.base		birth death skyline - handles serially sampled tips, piecewise con
bdtree		0.0.1	BEAST.base, BEAST.app		Birth-death sequential sampling
BEAST_CLASSIC		1.6.3	BEAST.base, BEAST.app		BEAST classes ported from BEAST 1 in wrappers
BEASTLabs	2.0.2	2.0.2	BEAST.base, BEAST.app		BEAST utilities, such as Script, multi monophyletic constraints
BEASTvntr		0.2.0	BEAST.app, BEAST.base		Variable Number of Tandem Repeat data, such as microsattelites
BICEPS		1.1.2	BEAST.app, BEAST.base		Bayesian Integrated Coalescent Epoch PlotS + Yule Skyline
bModelTest		1.3.3	BEAST.app, BEAST.base		Bayesian model test for nucleotide subst models, gamma rate he
BREAK_AWAY		1.2.0	BEASTLabs, BEAST.base, BEAST...		break-away model of phylogegraphy
CA		2.1.0	BEAST.app, BEAST.base		Bayesian estimation of clade ages based on probabilities of fossil
ClaDS		2.0.3	BEAST.base, BEAST.app		Implementation of the ClaDS birth-death tree prior
CoalRe		1.0.4	BEAST.app, BEAST.base		Inference of Recombination networks
CodonSubstModels		2.0.0	BEAST.base, BEAST.app		Codon substitution models for DNA
contacTrees		1.2.0	BEAST.base, BICEPS, BEASTLab...		Phylogenetic model with horizontal transfer for linguistics
contraband		1.0.1	BEAST.app, BEAST.base		Scalable brownian models for continuous trait evolution
CoupledMCMC		1.2.1	BEAST.base, BEAST.app		Adaptive coupled MCMC (adaptive parallel tempering or MC3)
DENIM		1.1.1	BEAST.app, BEAST.base		Divergence Estimation Notwithstanding ILS and Migration

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

Tracer

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

- Analyse log files from BEAST2 runs
- Check mixing, ESS, ACT, parameter correlations
- Overview of posterior parameter estimates
- Comparisons of several analyses

Input:

- log file

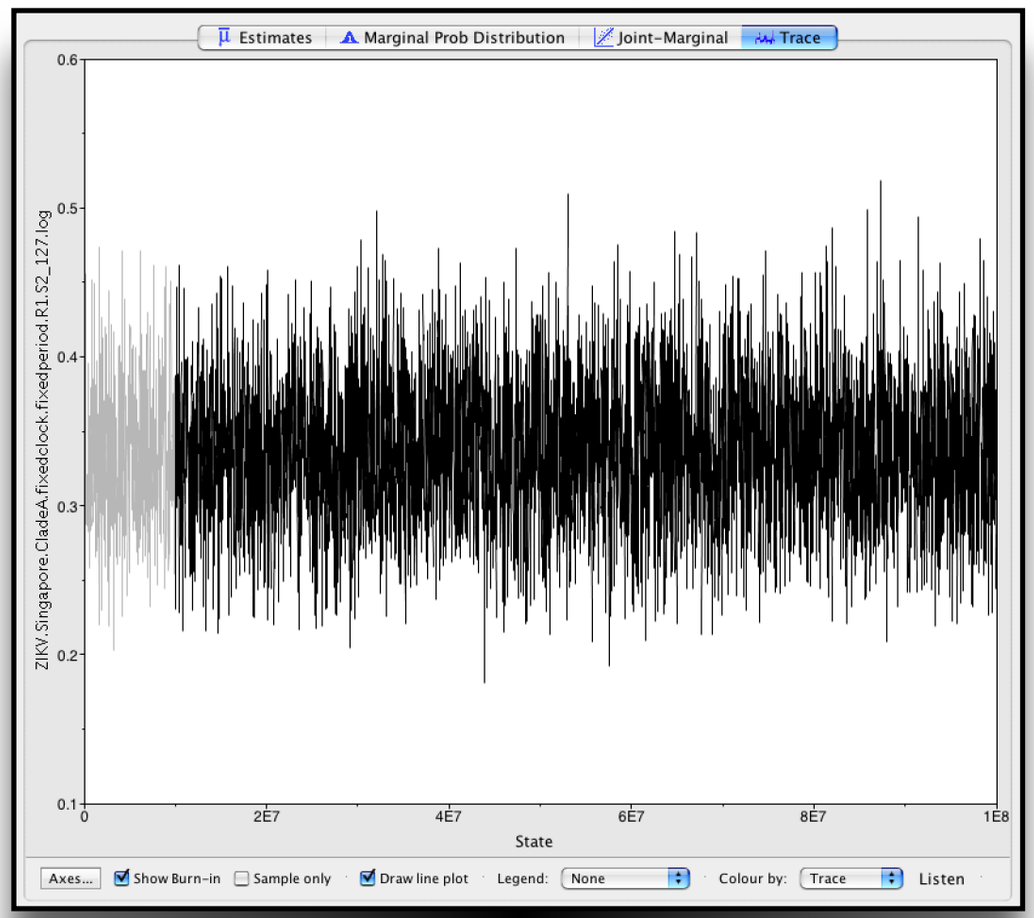
Output:

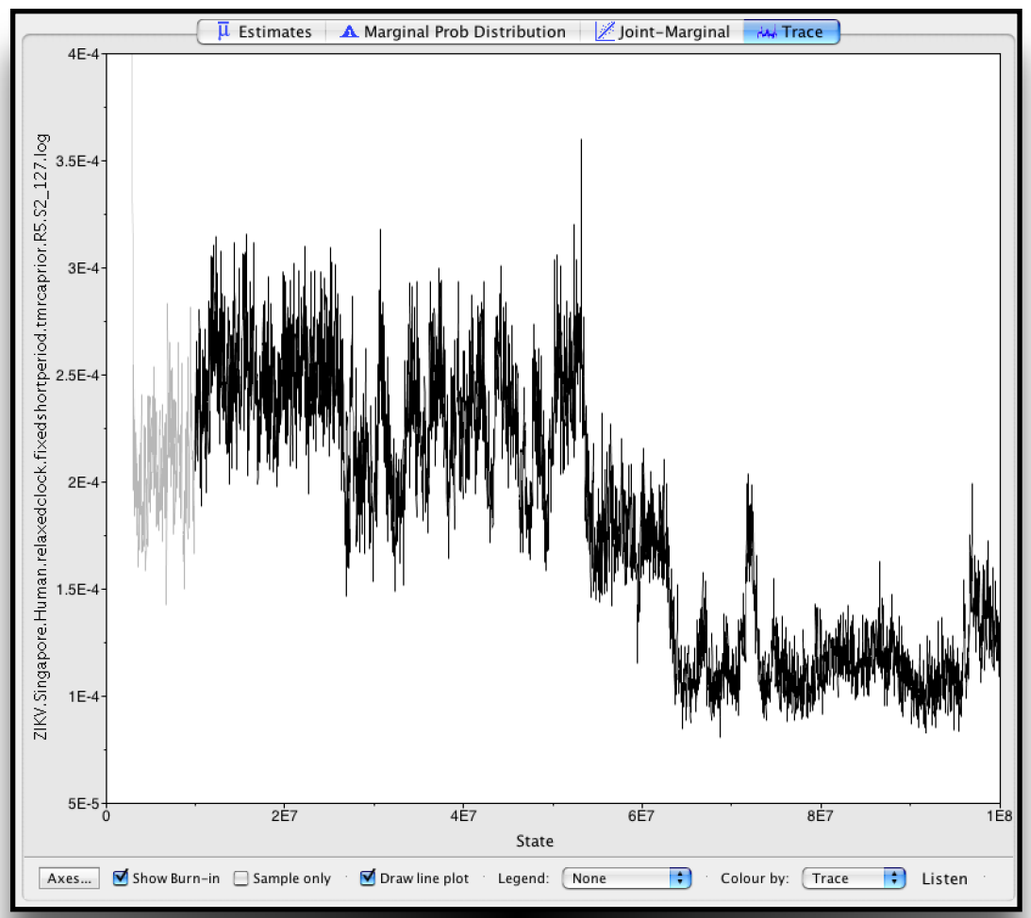
- Gain insight

Tracer

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

Look at the chains first!



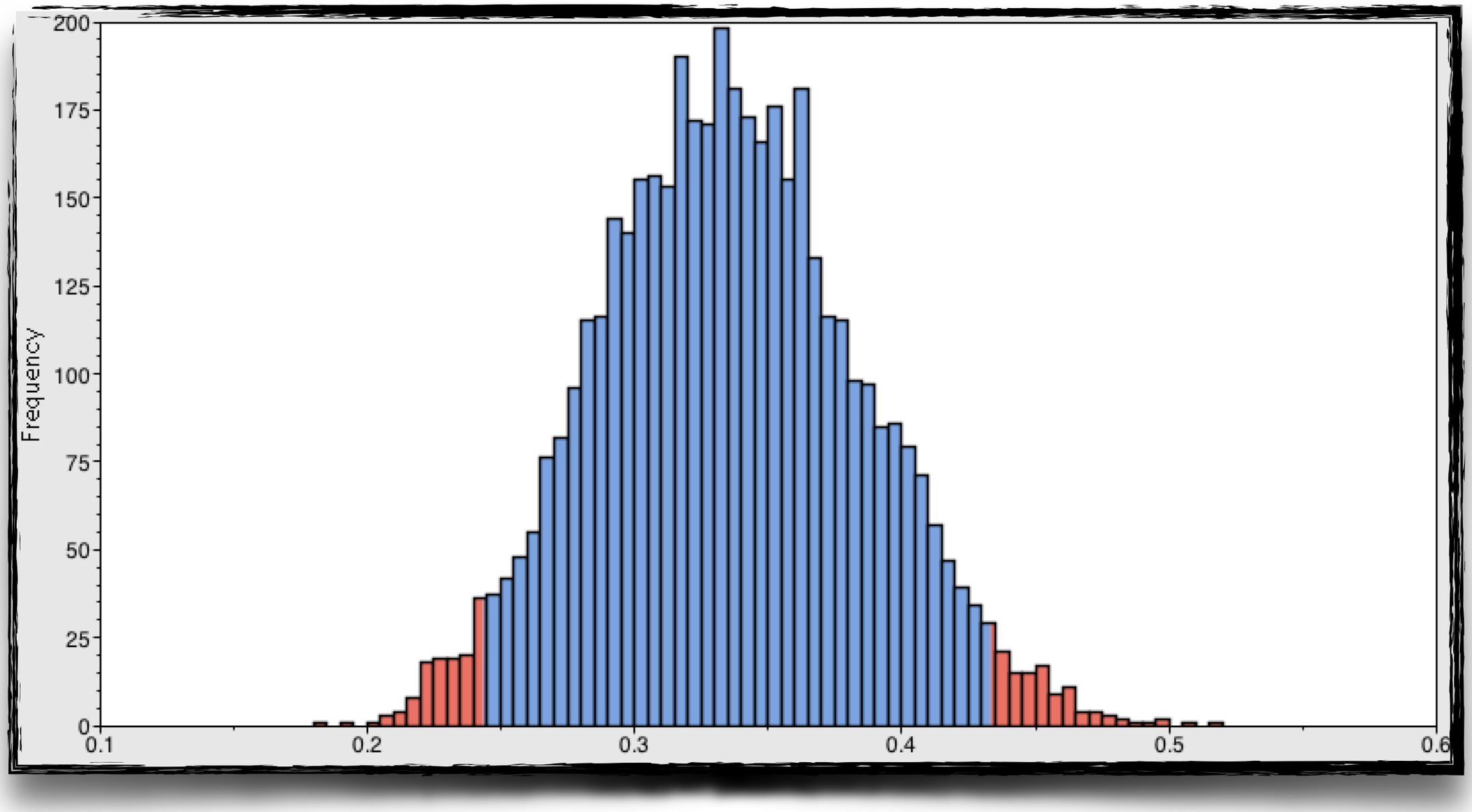


Mixing well! 😄

Not mixing! 😭

Tracer

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



TreeAnnotator

(Included with BEAST2)

- Analyse trees file from BEAST2 runs
- Produces MCC tree with node annotations (posterior probability)
- Note that the MCC tree is just a summary and may never actually appear in the trees file!

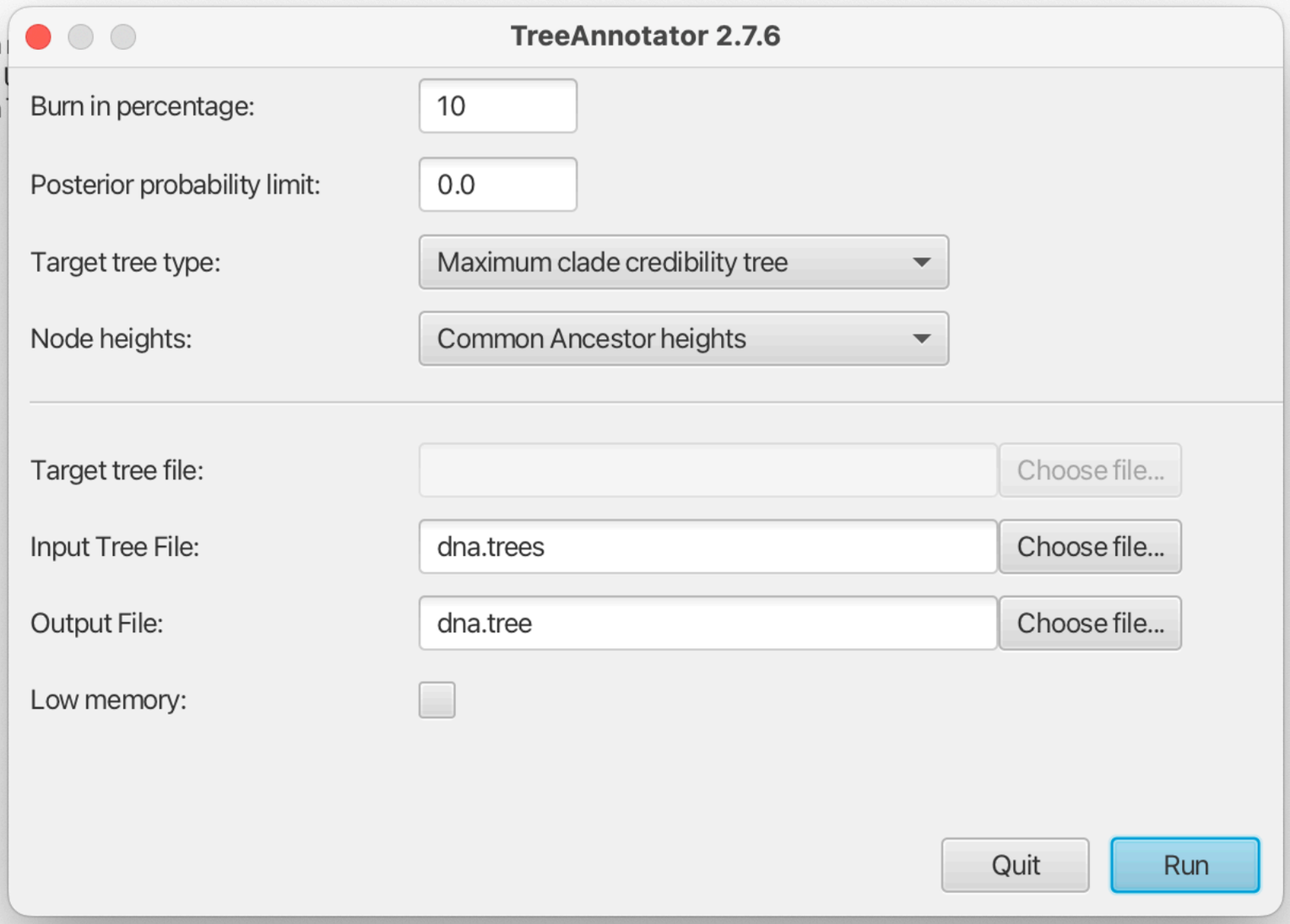
Input:

- trees file (many trees)

Output:

- MCC tree (one tree)

```
   TreeAnnotator v2.7.6, 2002-2023
         MCMC Output analysis
                   by
Andrew Rambaut and Alexei J. Drummond

  Institute of Evolutionary Biology
      University of Edinburgh
         a.rambaut@ed.ac.uk

Depa
     a
```



FigTree

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

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

Input:

- trees file

Output:

- Insight
- Figures

DensiTree

★ Comes with BEAST 2

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

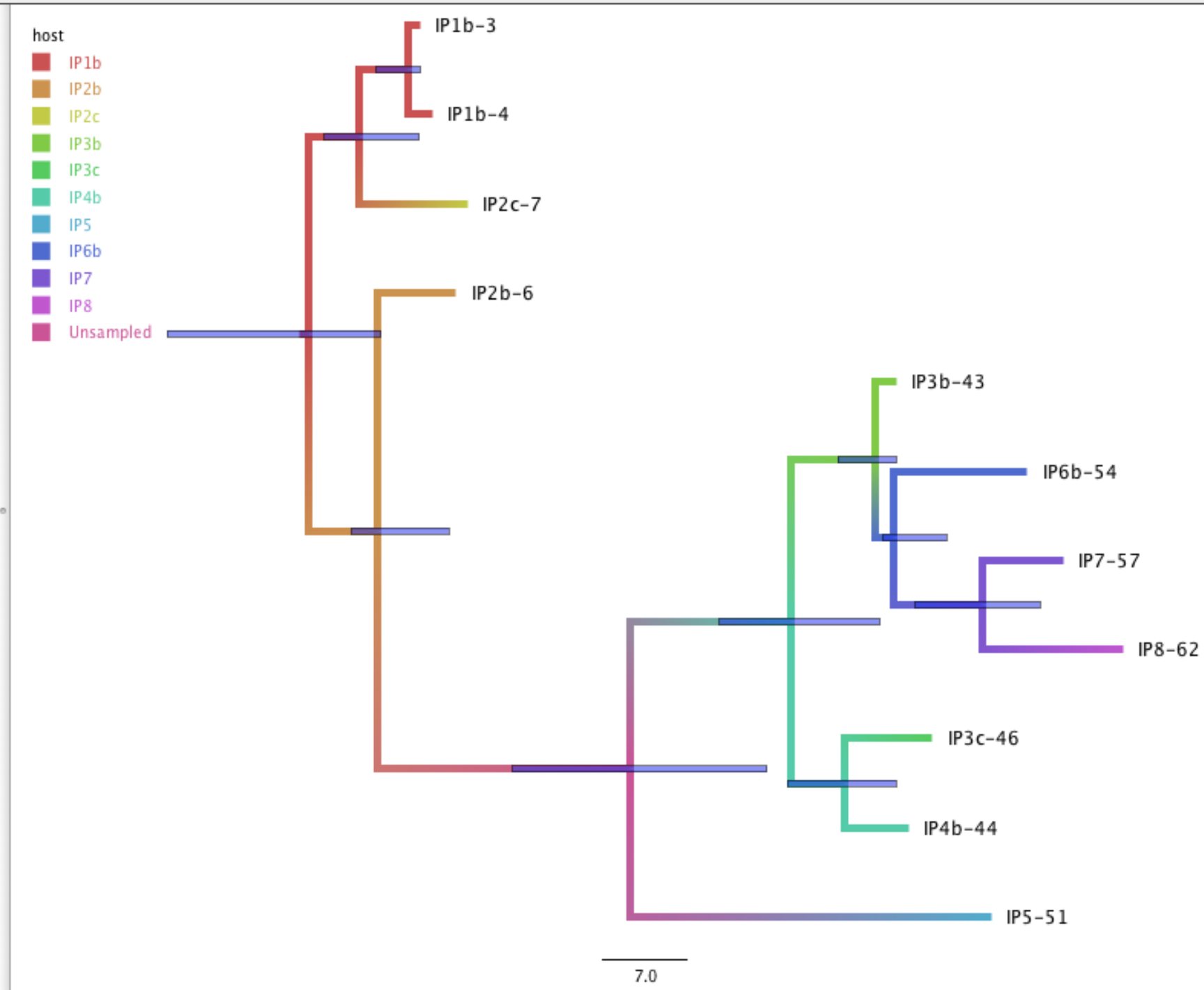
Node Bars

Branch Labels

Scale Bar

Scale Axis

Legend



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

DensiTree - Tree Set Visualizer H1.caching_tree_with_trait.trees

"Moderate"
"None"
"Specialised"

Default
StarTree
Central
Angled
Triangle
Block
Arced
Steep

Show
Grid
Label
Geography
Line Width
Line Color

location

Show legend

Multi color cons-trees

categorical

pattern:

.*location="([^"]*).*

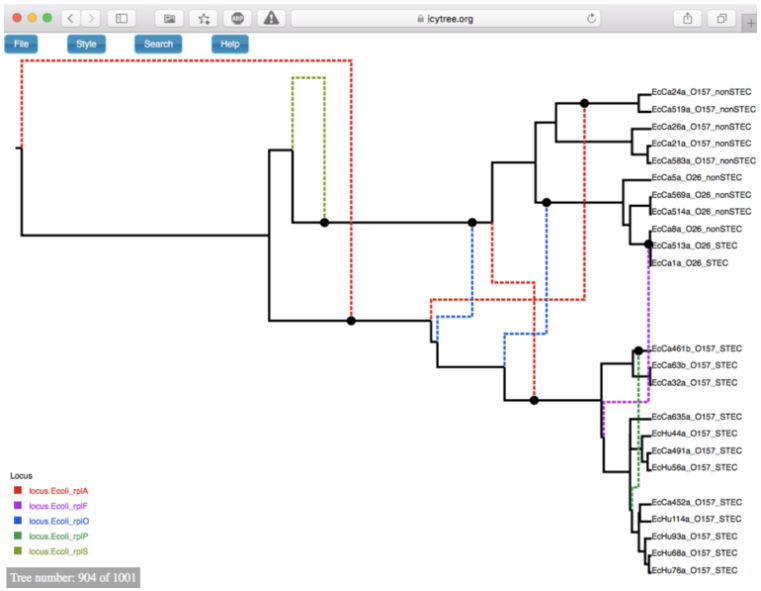
Line colors

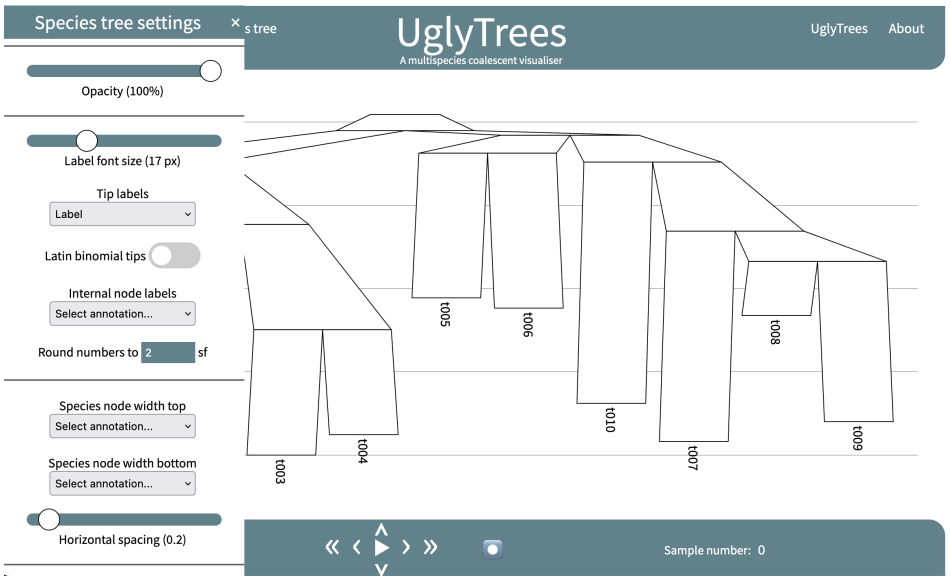
Burn in

height=0.1701855

Tree visualisation

- UgyTrees for multi species coalescent (StarBeast)
  `https://uglytrees.nz/`
- IcyTree for for structured models and ancestral recombination graphs (ARGs)
  `https://icytree.org/`





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 and BEAST2 output files (.log)

TreeAnnotator

Analysis of BEAST2 output files (.trees)

DensiTree, FigTree, IceyTree

Visualisation of trees (.trees)

BEAST best practice

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

Before you begin

1) Know your data
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

Let's do some work

Introduction to BEAST2 tutorial (divergence dating)
`https://taming-the-beast.github.io/tutorials/Introduction-to-BEAST2/`