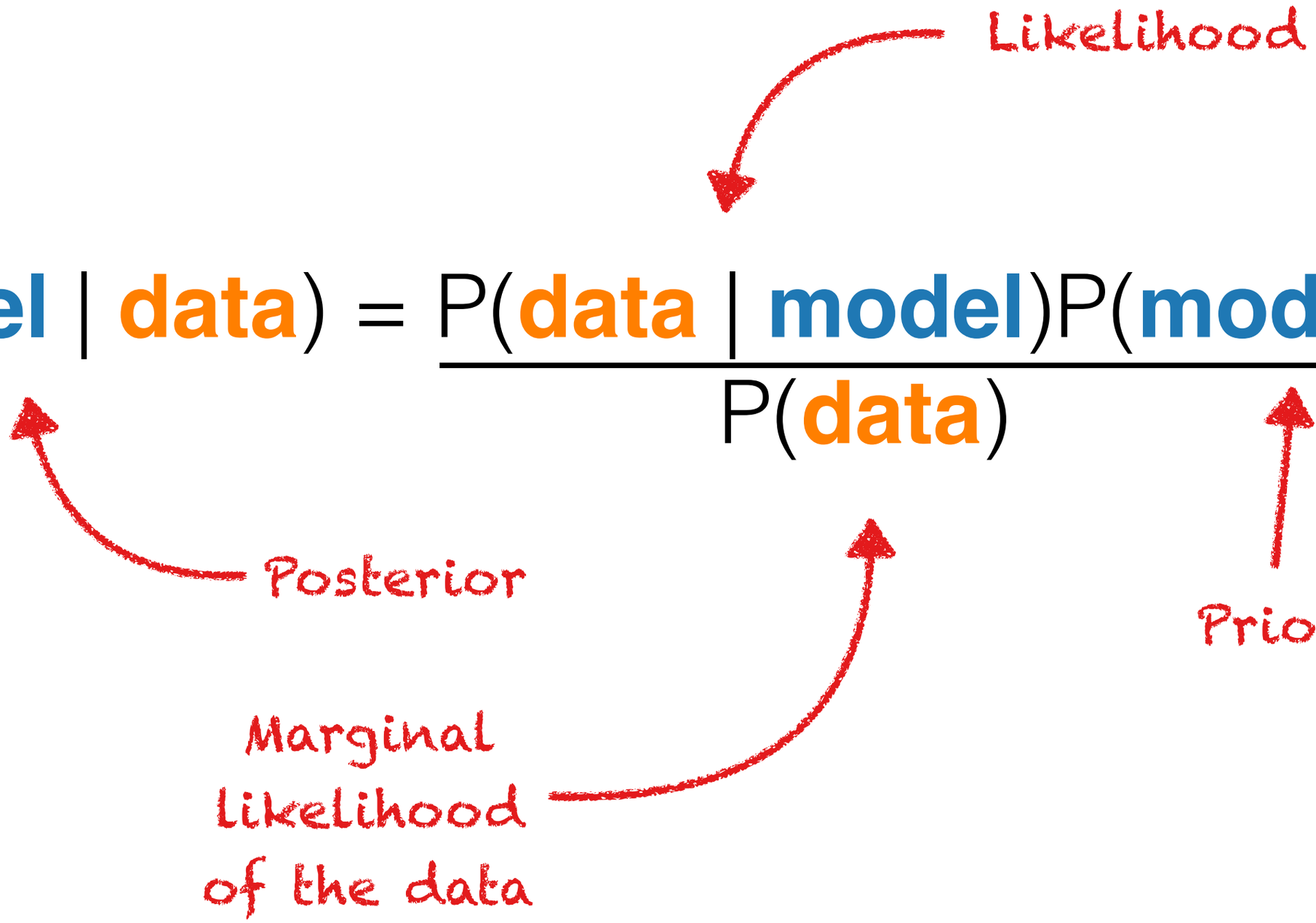
Viral phylodynamics using BEAST2

Louis du Plessis

1. Bayesian inference basics
2. What goes into a BEAST model?
3. MCMC inference
4. BEAST2 practical

http://taming-the-beast.github.io

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



Data

- Typically one or more alignments of **genomic sequencing** data
- Sampled at **one** or **many** time points
- Timescale of days to years
- May contain sampling **location** or phenotypic **trait** data
- Assume that the data are correct
- Realisation of a stochastic process

Model and hypothesis

- **Model** is a mathematical description of the **process** that generated the data
- Each model has a number of **model parameters**
- Parameters are **random variables**
- Several types of evolutionary models are often used in combination

- **Hypothesis** is some statement about the model parameters
- The model may have **many parameters** but our hypothesis may only concern **some of them**. The rest are called **nuisance parameters**.

Model vs. Hypothesis

- Hypothesis assigns values to the model parameters

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

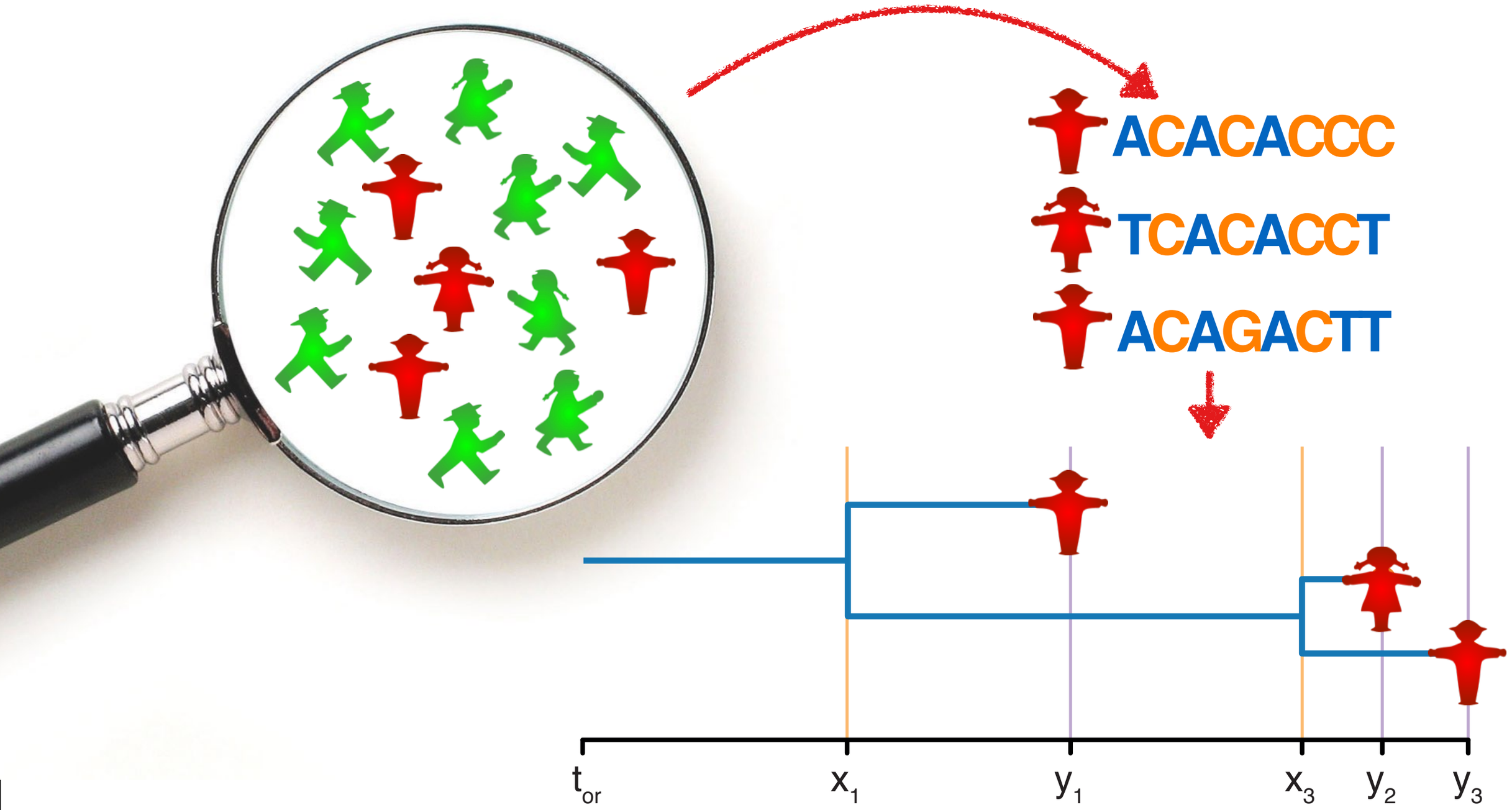
Likelihood → P(data | model)

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

Posterior → P(model | data)

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

Transmission dynamics from sequencing data

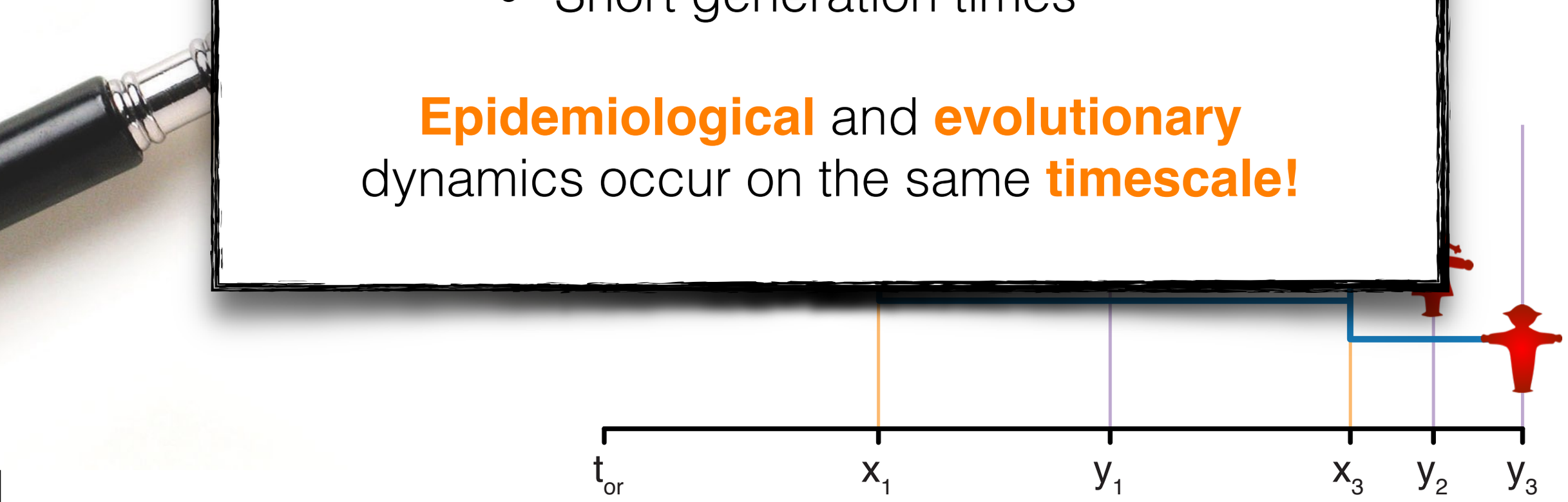


Transmission dynamics from sequencing data

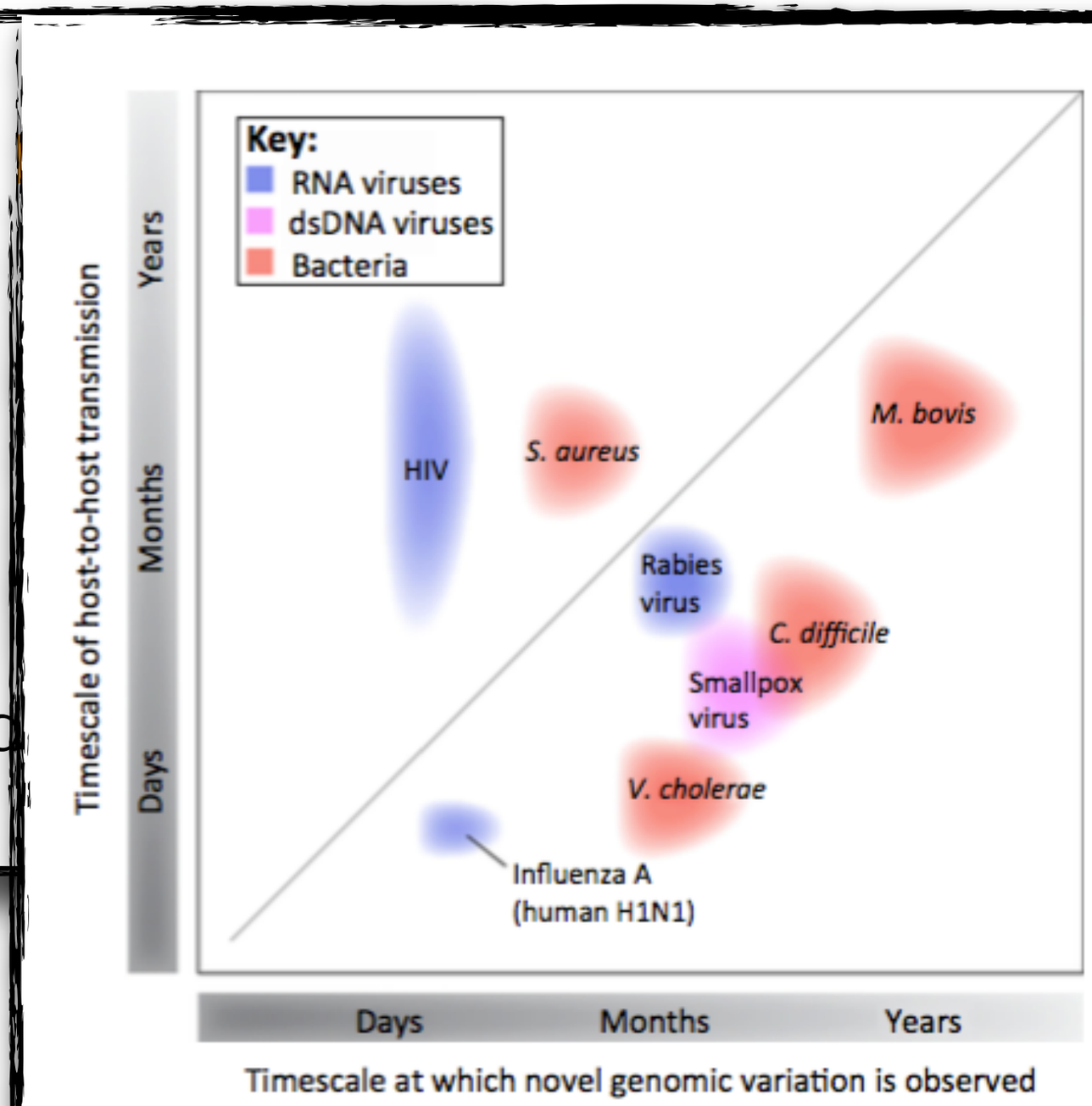
Need a measurably evolving population!

- Large population size
- High mutation rate
- Short generation times

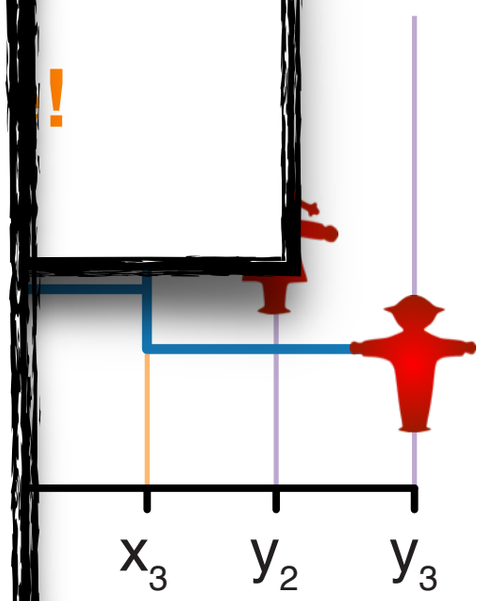
Epidemiological and **evolutionary** dynamics occur on the same **timescale!**



Transmission dynamics from sequencing data

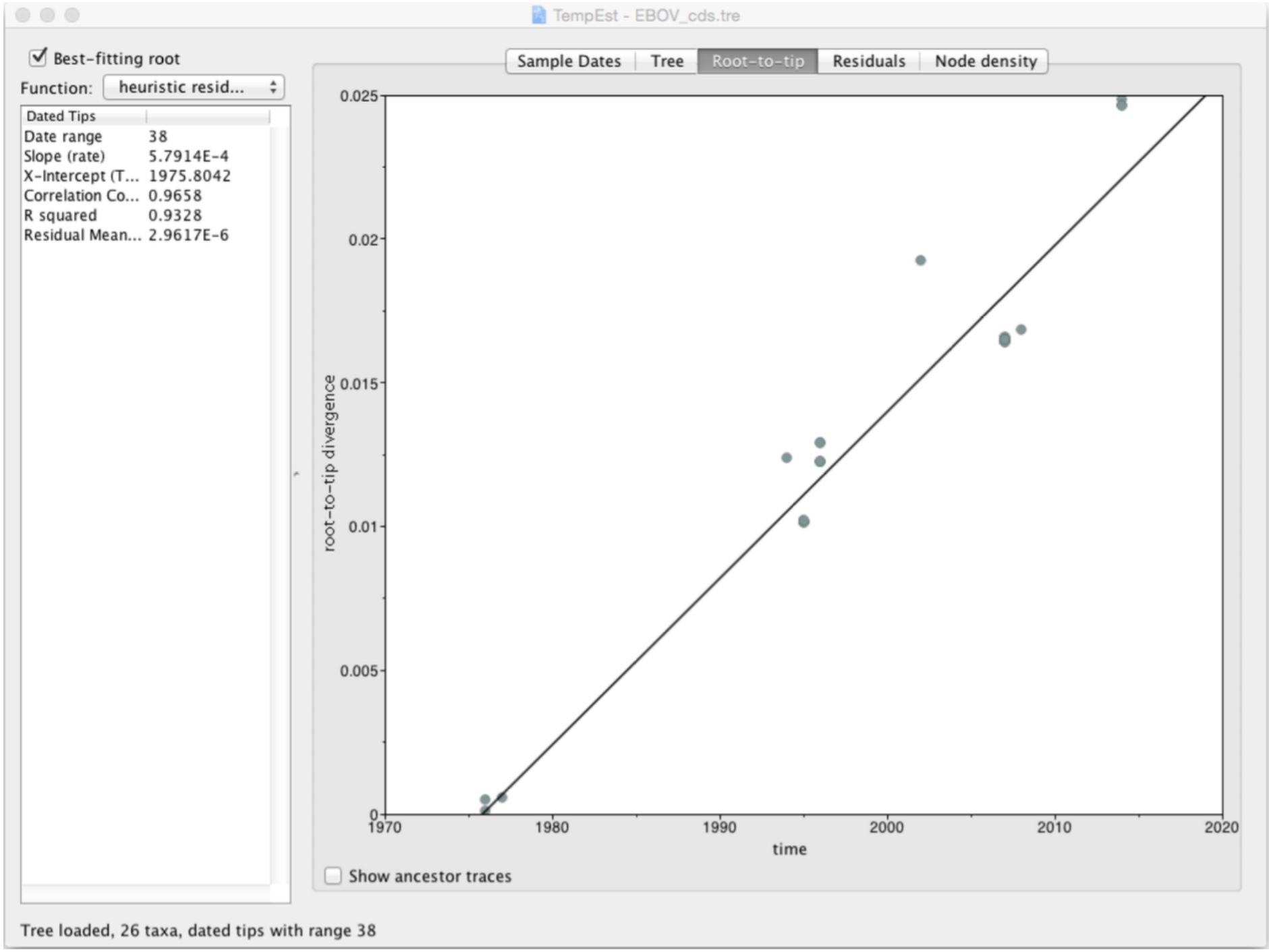


Biek *et al.*Trends in Ecology & Evolution **2015**



Tran

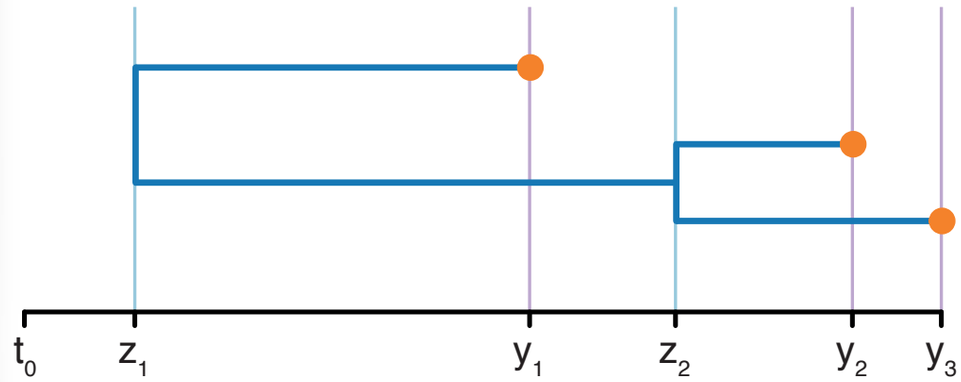
If unsure use TemPest to assess clock signal

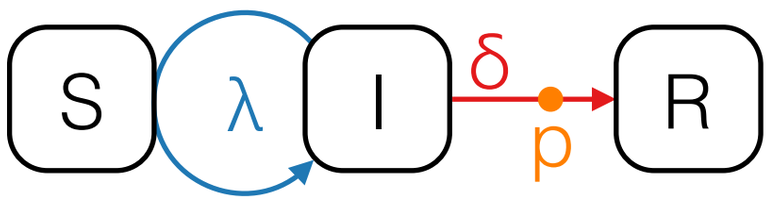


What is phylodynamics?

Phylogenetics

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





Phylodynamics

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

Phylodynamic 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?

What goes into a BEAST model?

genealogy

demographic model

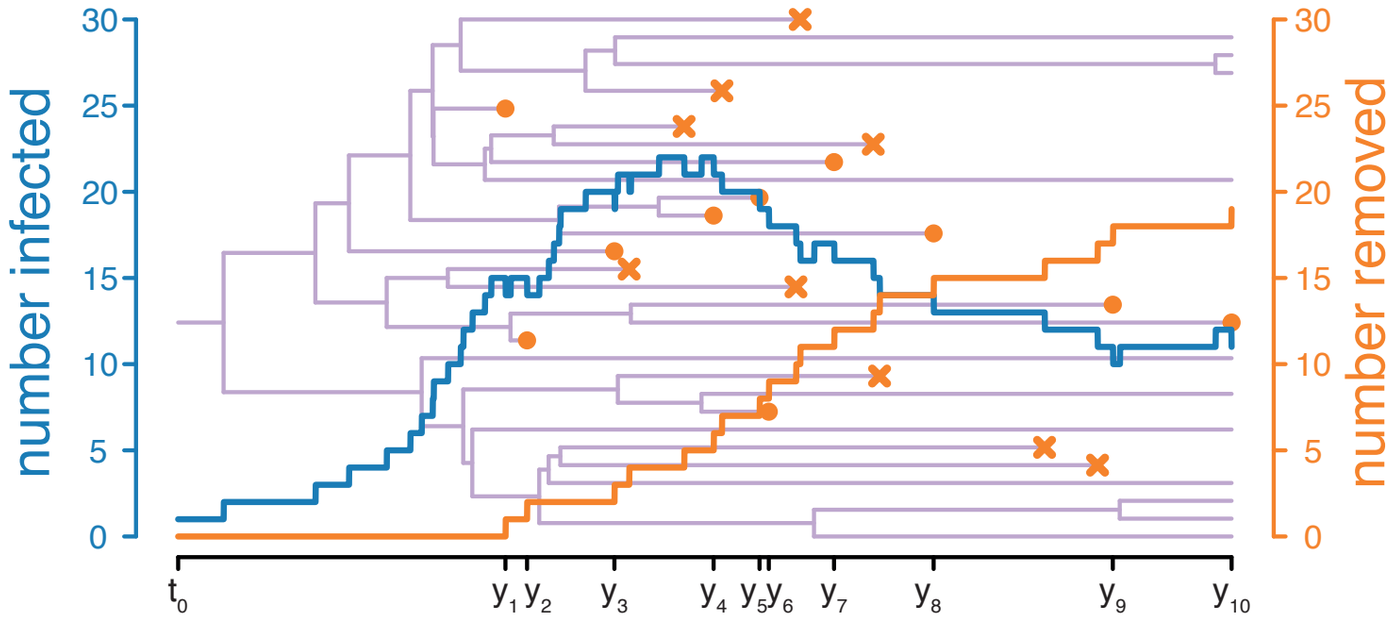
site model

molecular clock model

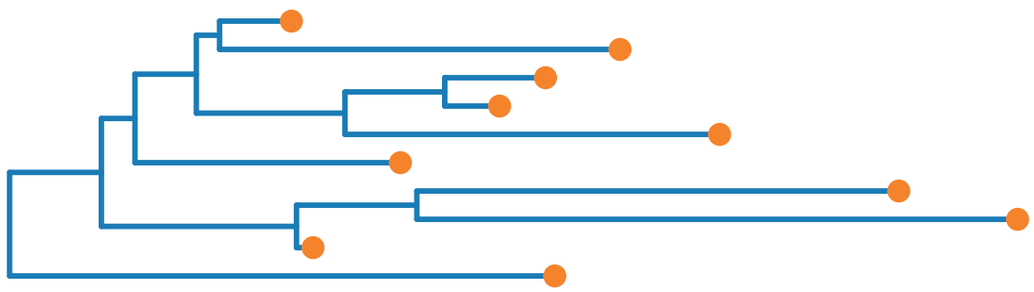
Estimate posterior distributions!

The genealogy (tree)

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



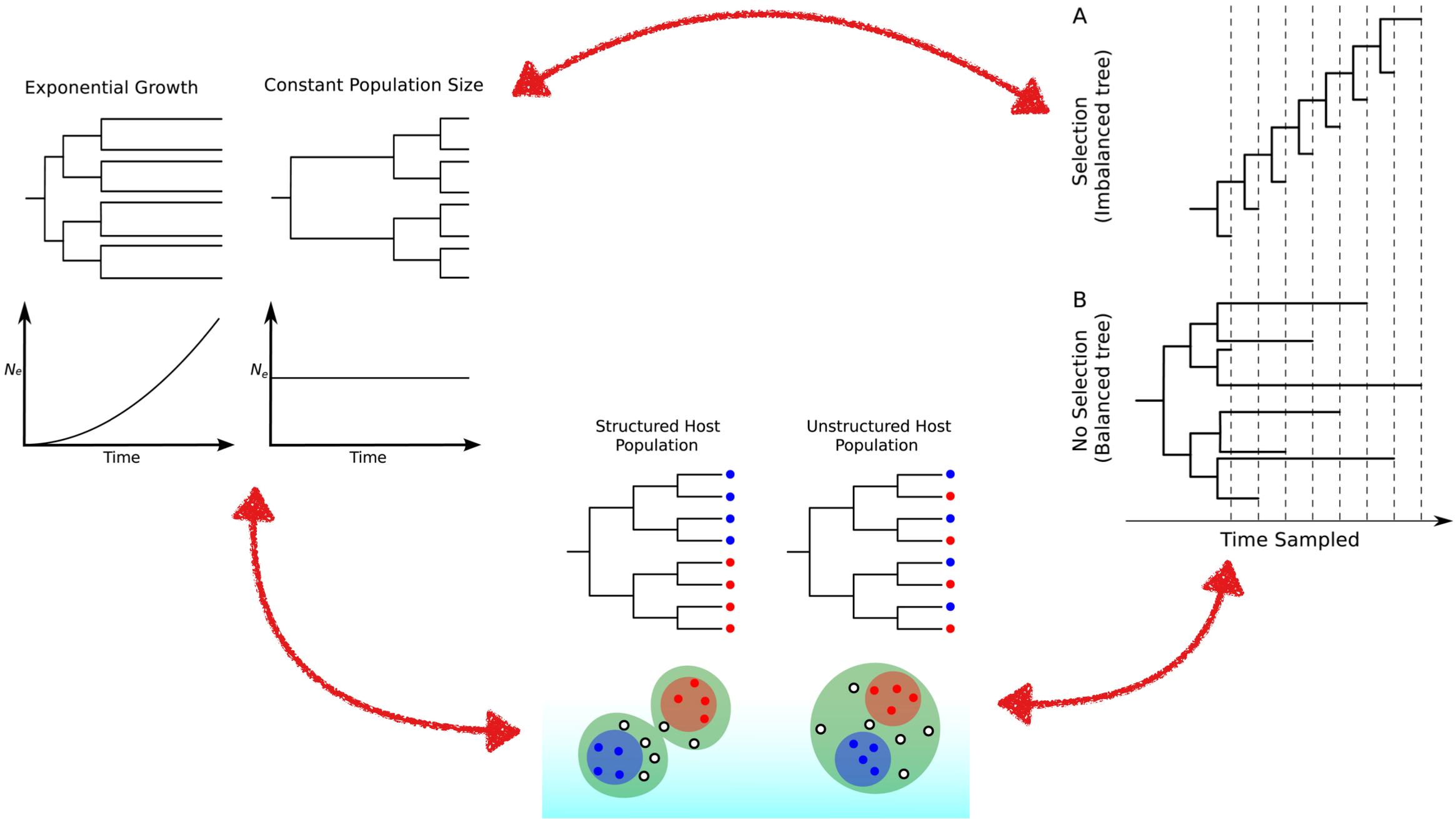
Full transmission tree



Sampled tree

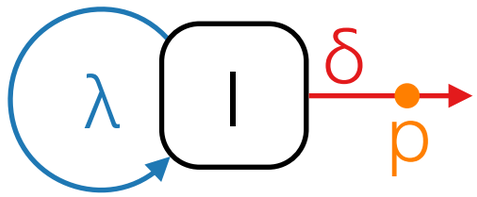
- Only the relationships between the **sampled** sequences!

Different population dynamics generate different trees



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

Demographic model

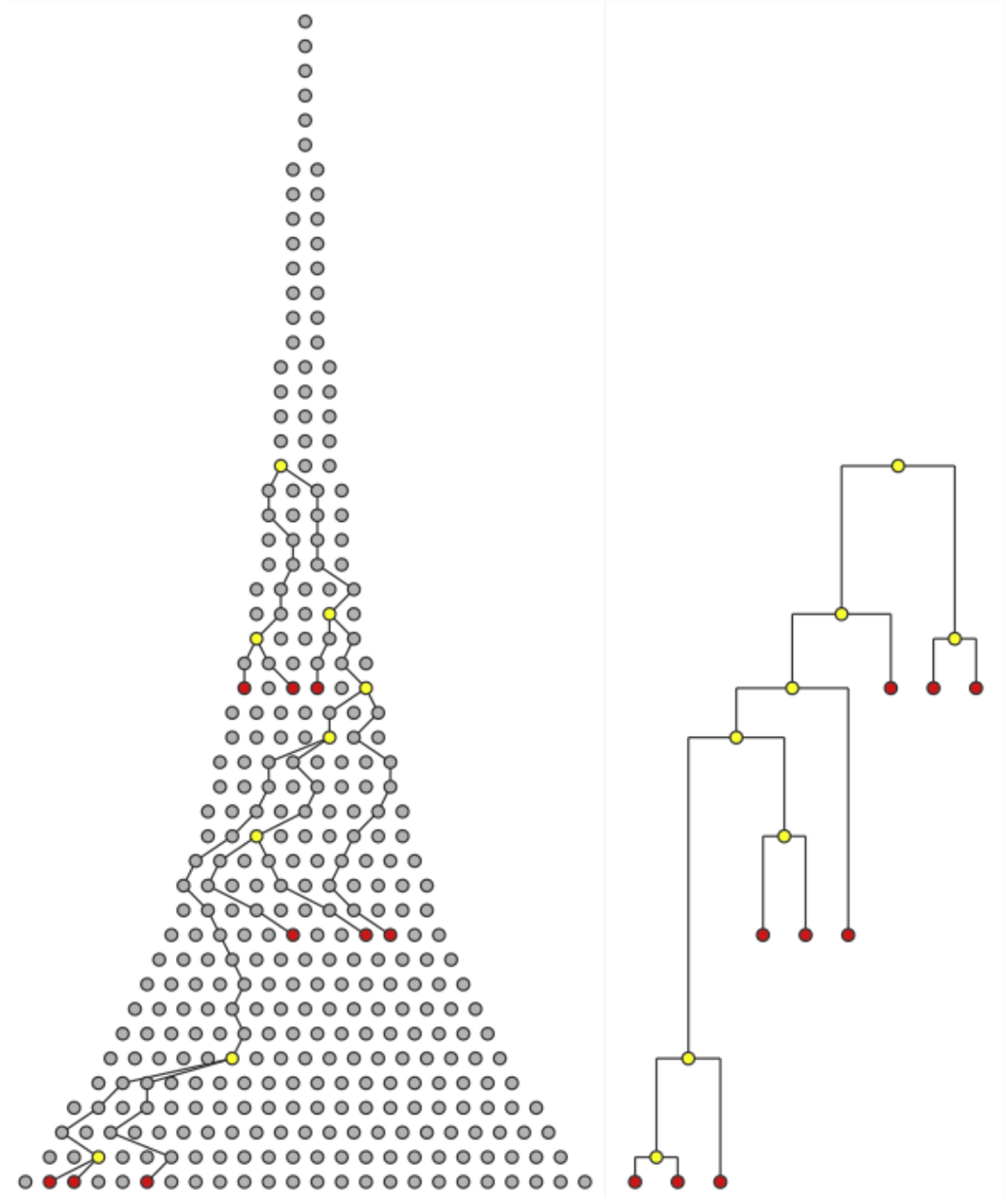


- λ — infection rate
- δ — becoming-noninfectious rate
- p — sampling probability

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

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

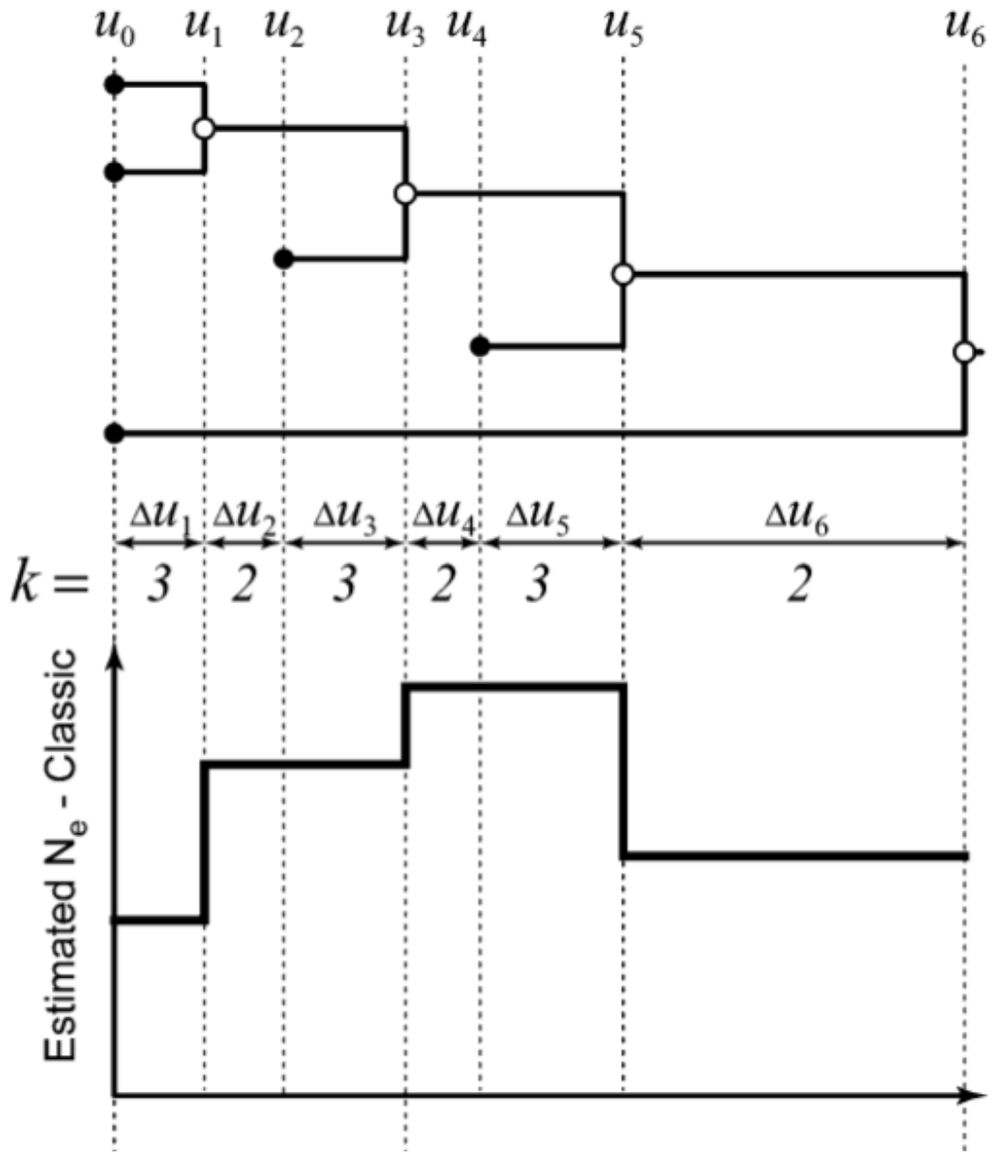
Coalescent model



- Assume Wright-Fisher like population dynamics
- Given effective population size (**N_e**)
- Calculate the probability for **2** nodes to coalesce in time **t**
- What if we don't have a good intuition for how **N_e** is changing?

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

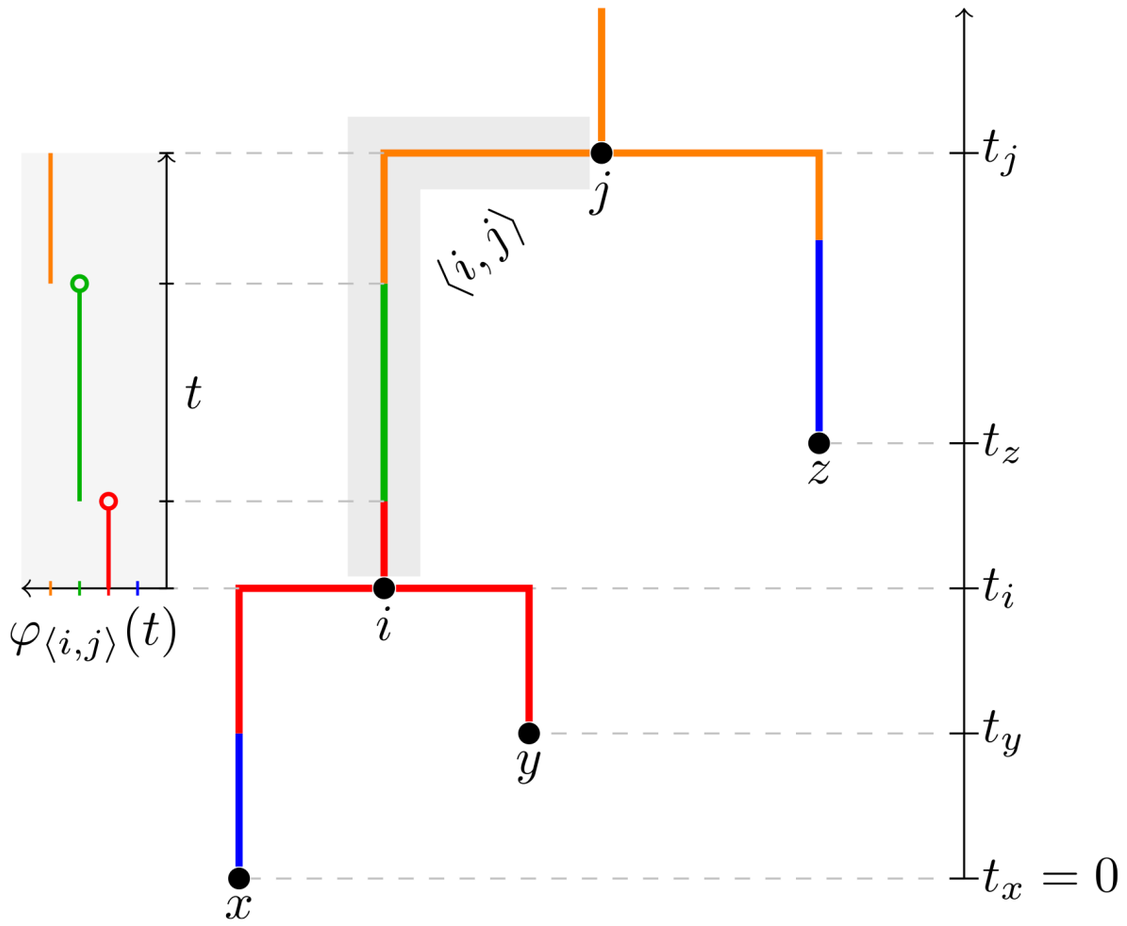
Nonparametric models



- Skylines model time-dependent demographic parameters
- Nonparametric piecewise constant approximation

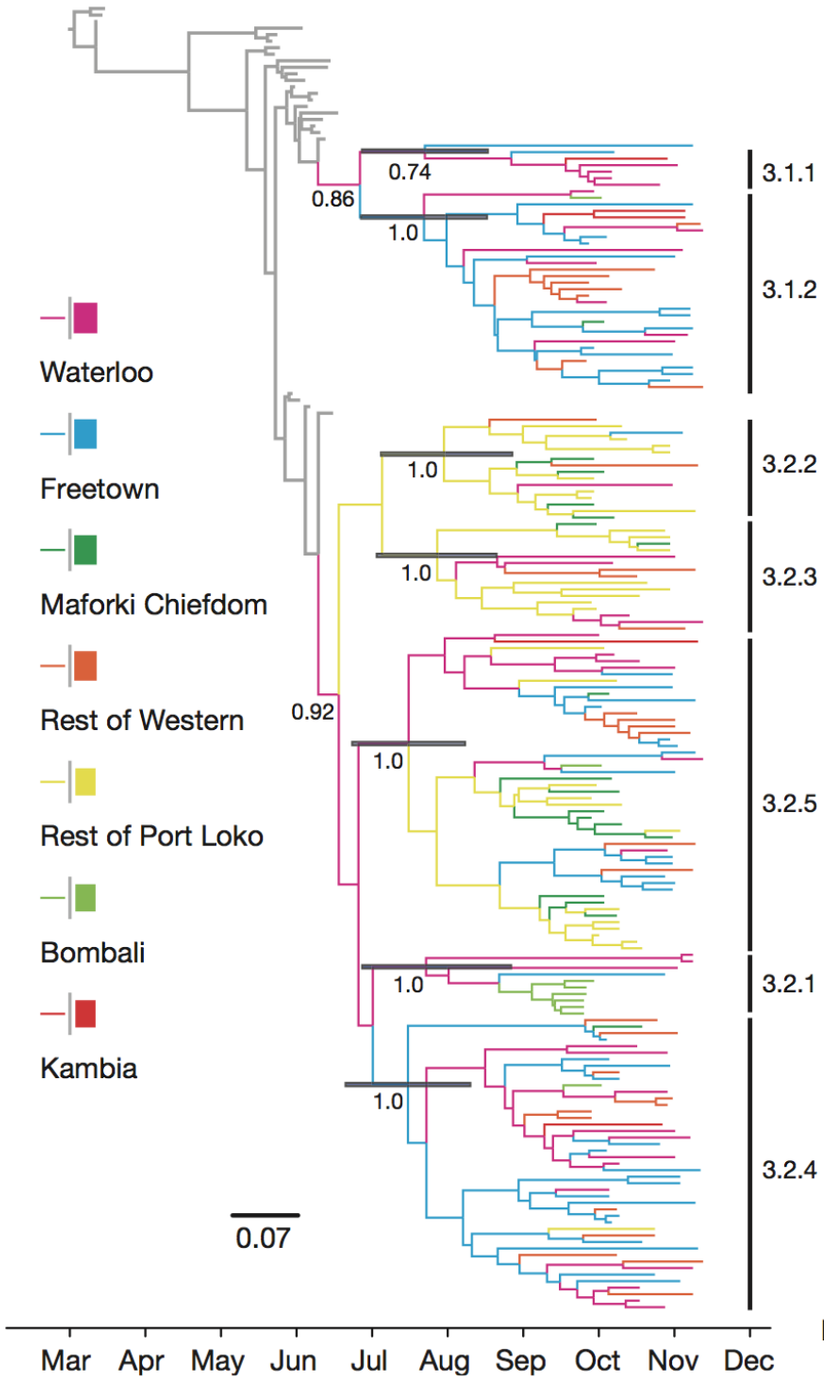
Drummond *et al.* **MBE 2005**

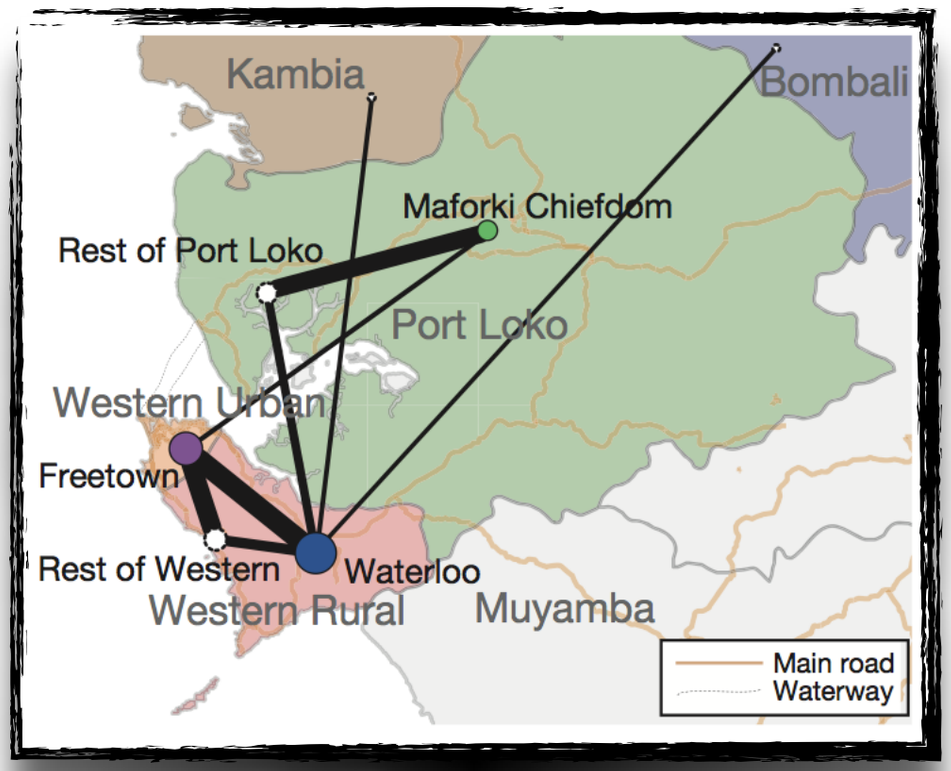
Population structure



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

Phylogeography: Ebola in West Africa

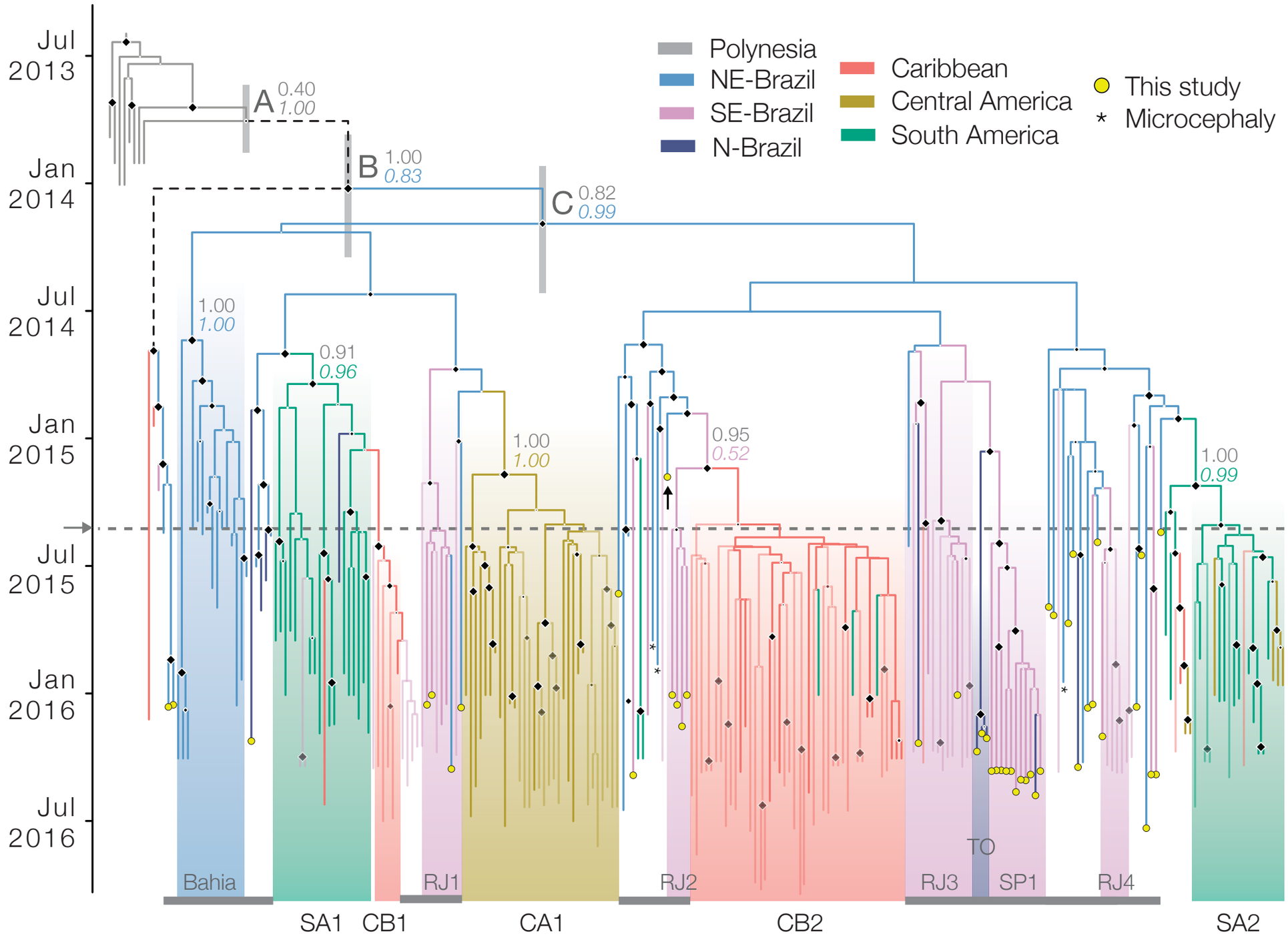




- Infer how the virus spread across the region
- Infer the most probable ancestral states
- Infer migration rates between different regions

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

Phylogeography: Zika in the Americas

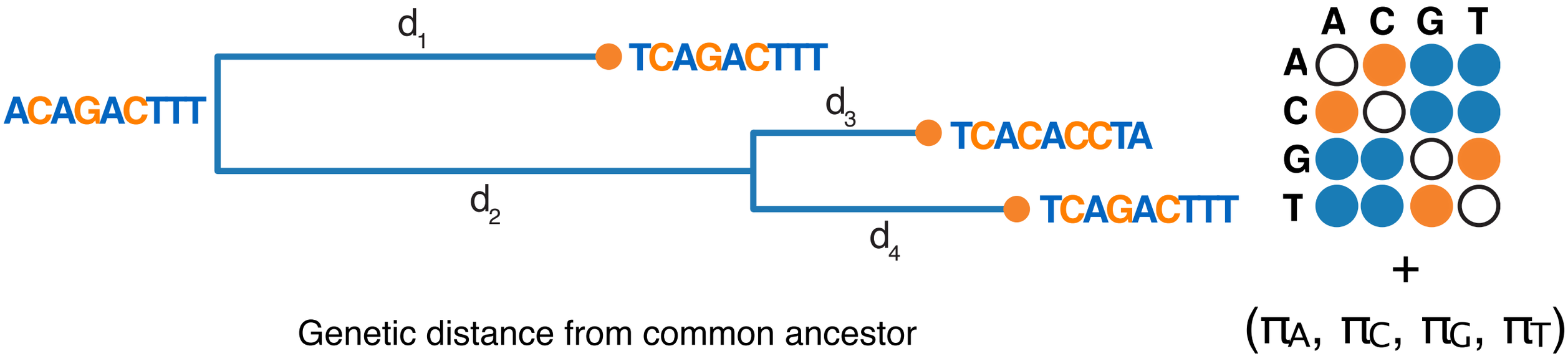


Demographic model

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

- How likely is the genealogy given a demographic model?
- Usually a birth-death or a coalescent model
- Parametric or nonparametric models for changing dynamics over time
- Structured models for different subpopulations and phylogeography

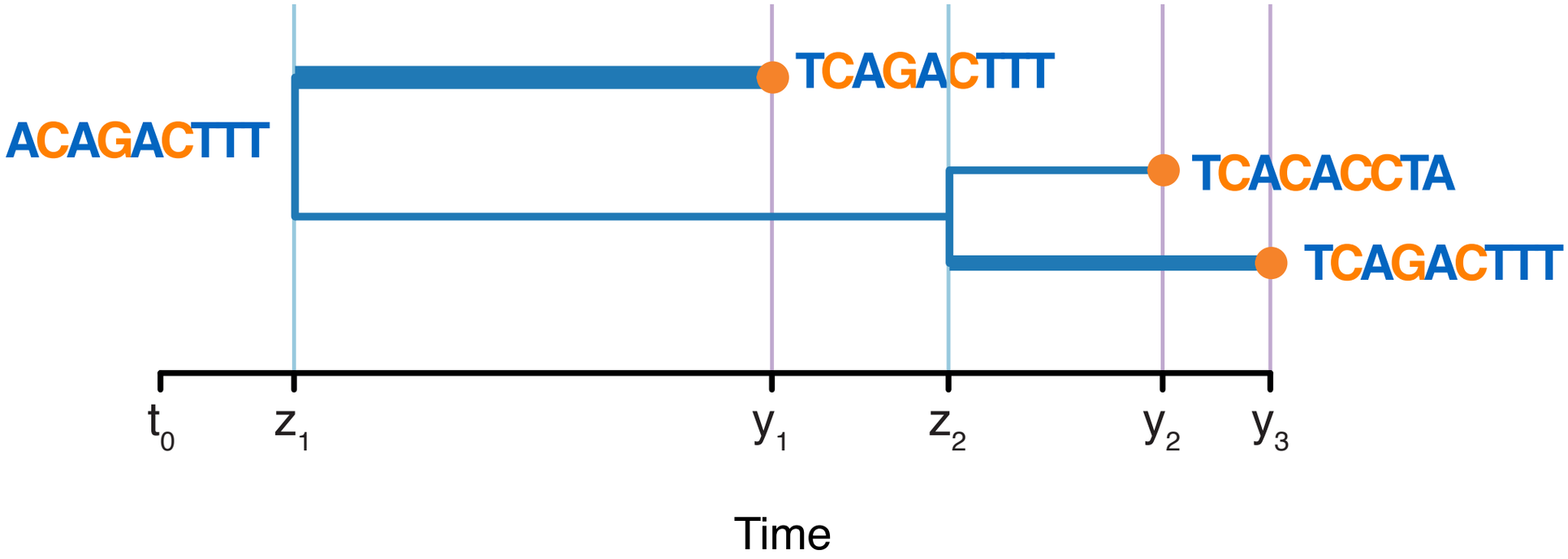
Site model



Genetic distance from common ancestor

- Links the genome sequences to the genealogy
- Relative rate of substitution from one nucleotide to another
- Can also model site-to-site variation
- Separate site model for each data partition
  - Multiple genes/loci
  - Model 1st, 2nd, 3rd codon positions differently

Molecular clock model



- How long does it take for substitutions to appear?
- Scales branch lengths to calendar time
- Different branches may have different clock rates

Putting it all together

genetic sequences

genealogy

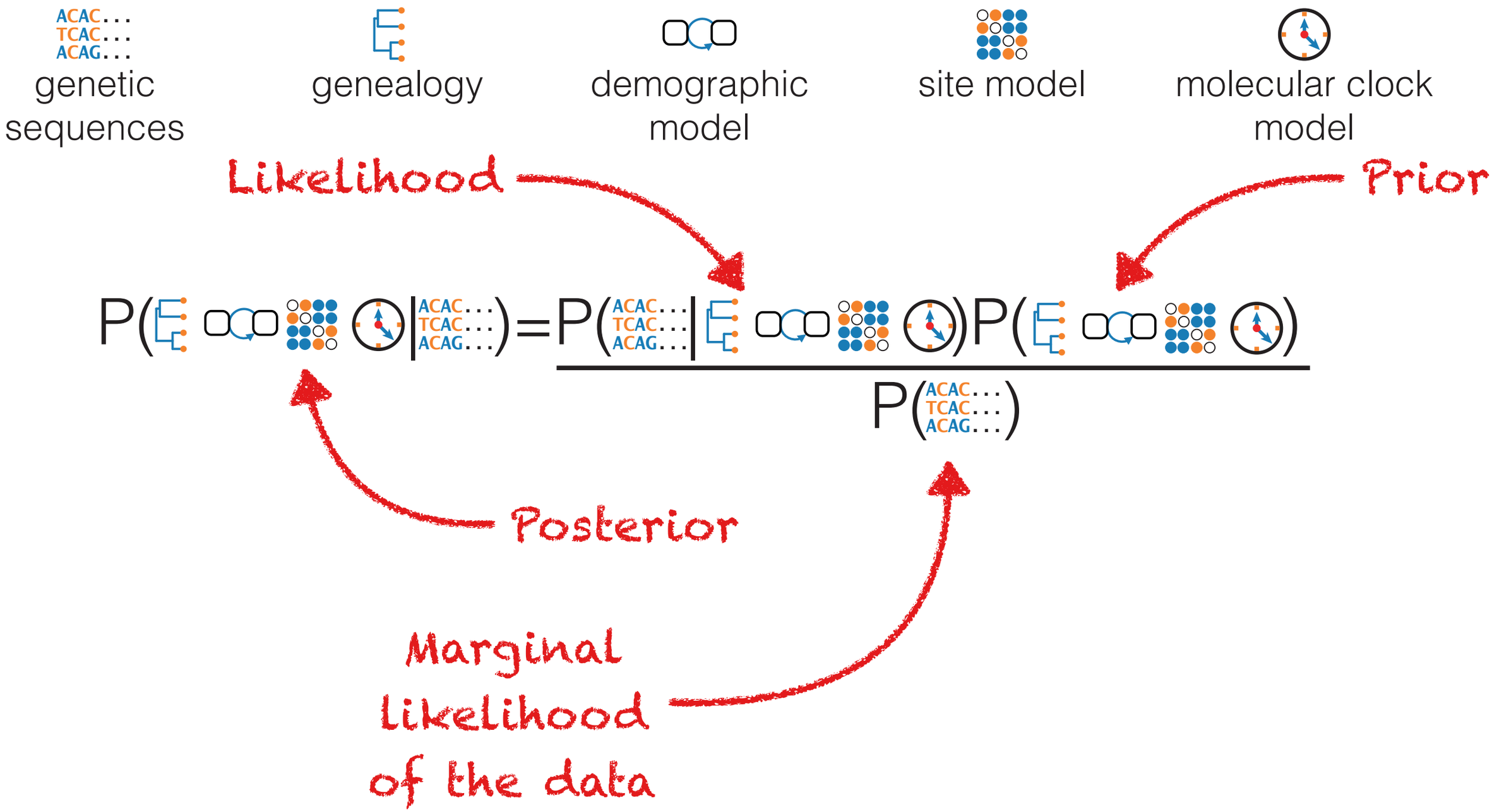
demographic model

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

genetic sequences


genealogy


demographic model


site model


molecular clock model

Assume independence

Posterior distribution in BEAST2

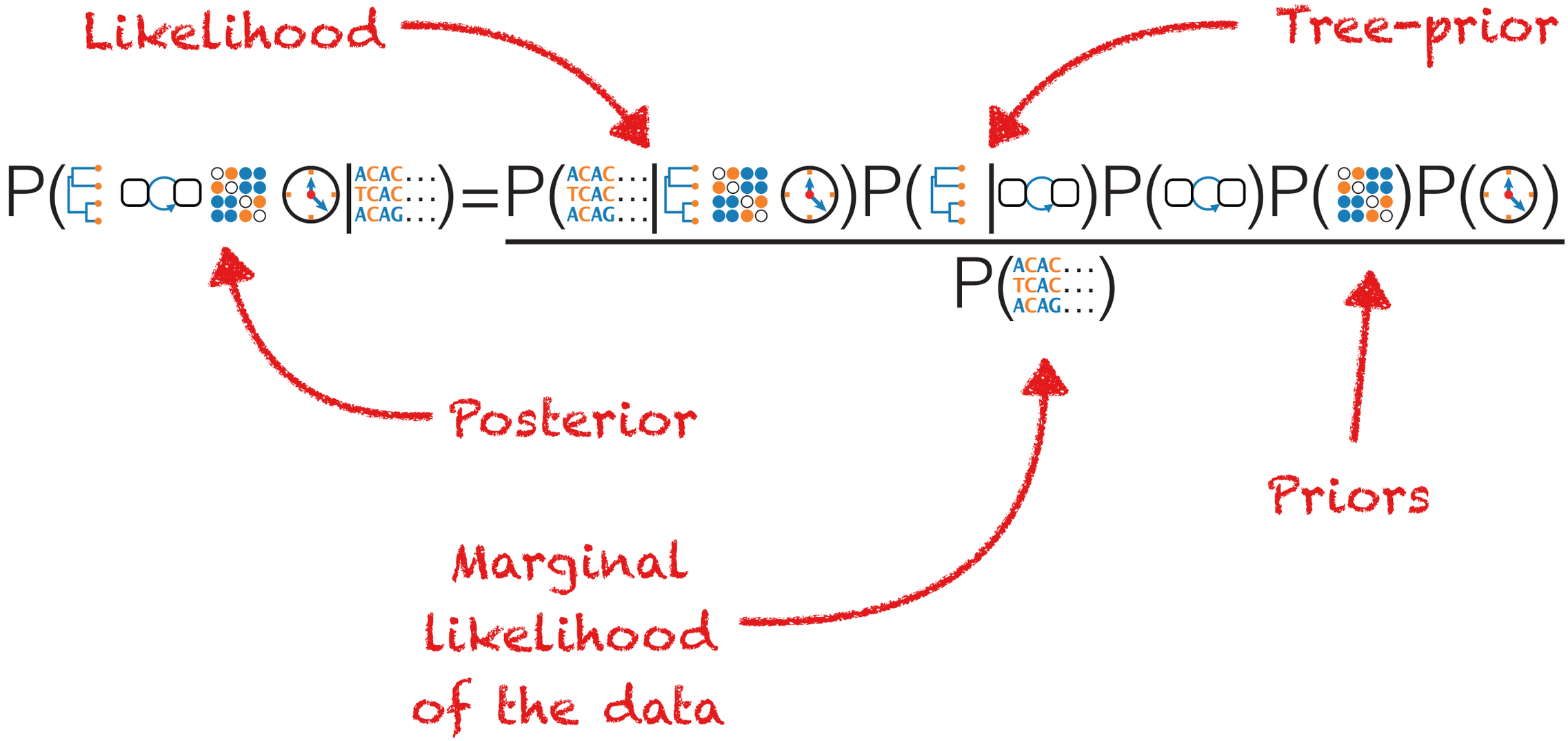
genetic sequences

genealogy

demographic model

site model

molecular clock model



MCMC

(Markov-chain Monte Carlo)

- We want to calculate the posterior
- But we cannot easily calculate the marginal likelihood
  → use MCMC!

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)

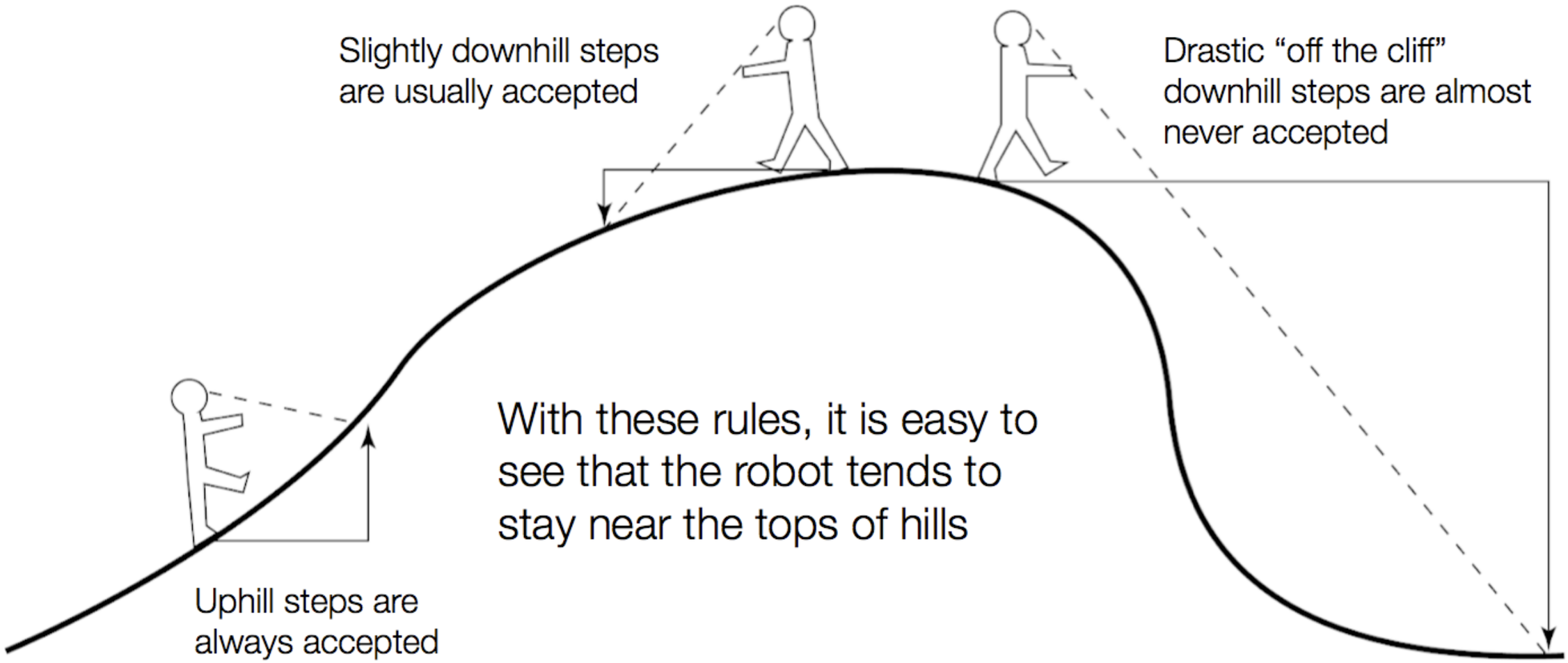
MCMC

(Markov-chain Monte Carlo)

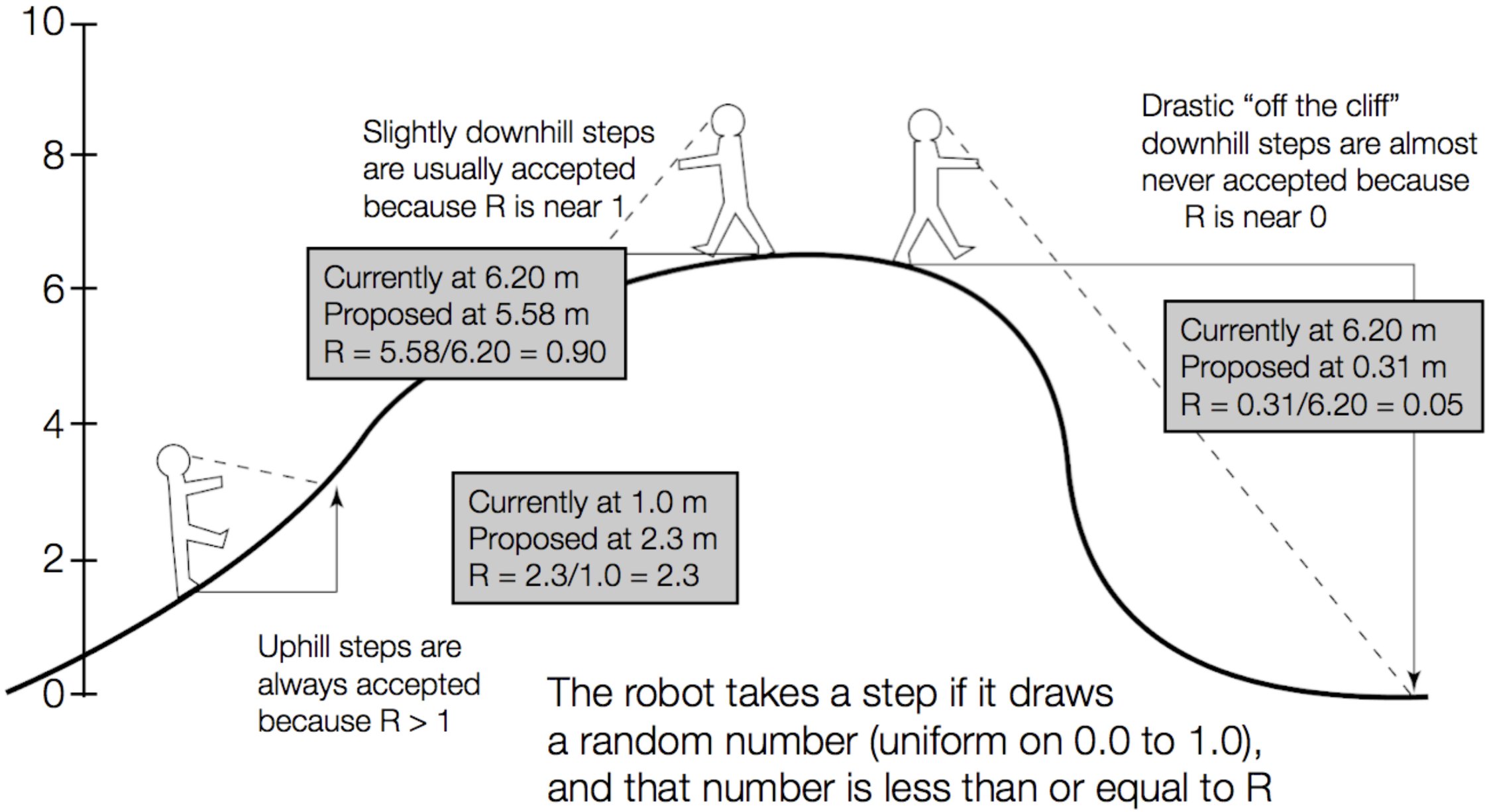
- MCMC performs a random walk on the posterior, preferentially sampling high-density areas
- Only need to compare which posterior density is higher
- So we only need the ratio of posteriors and the marginal likelihoods cancel out

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

MCMC robot (courtesy of Paul Lewis)

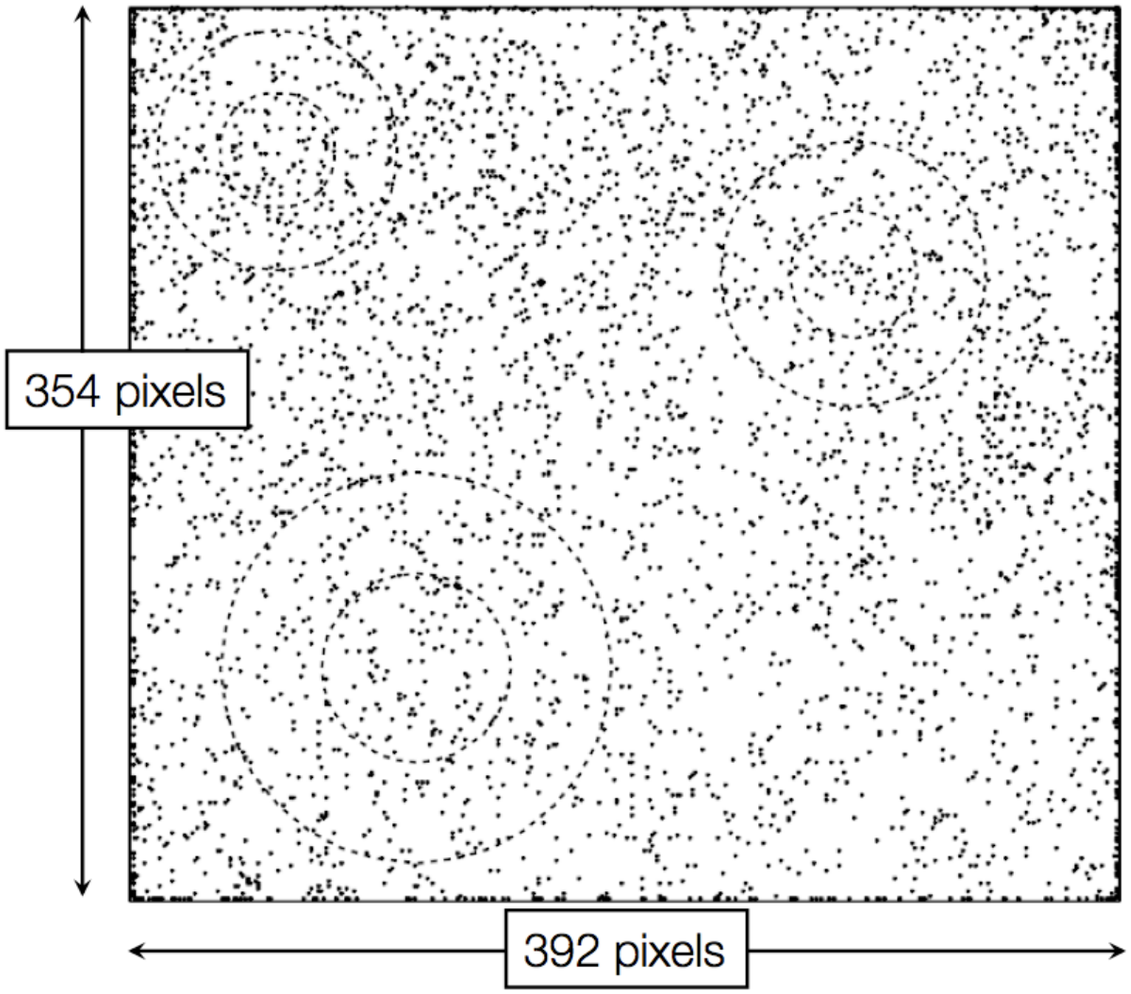


MCMC robot (courtesy of Paul Lewis)



(R is the ratio between the posterior densities)

Pure random walk (courtesy of Paul Lewis)



Random walk

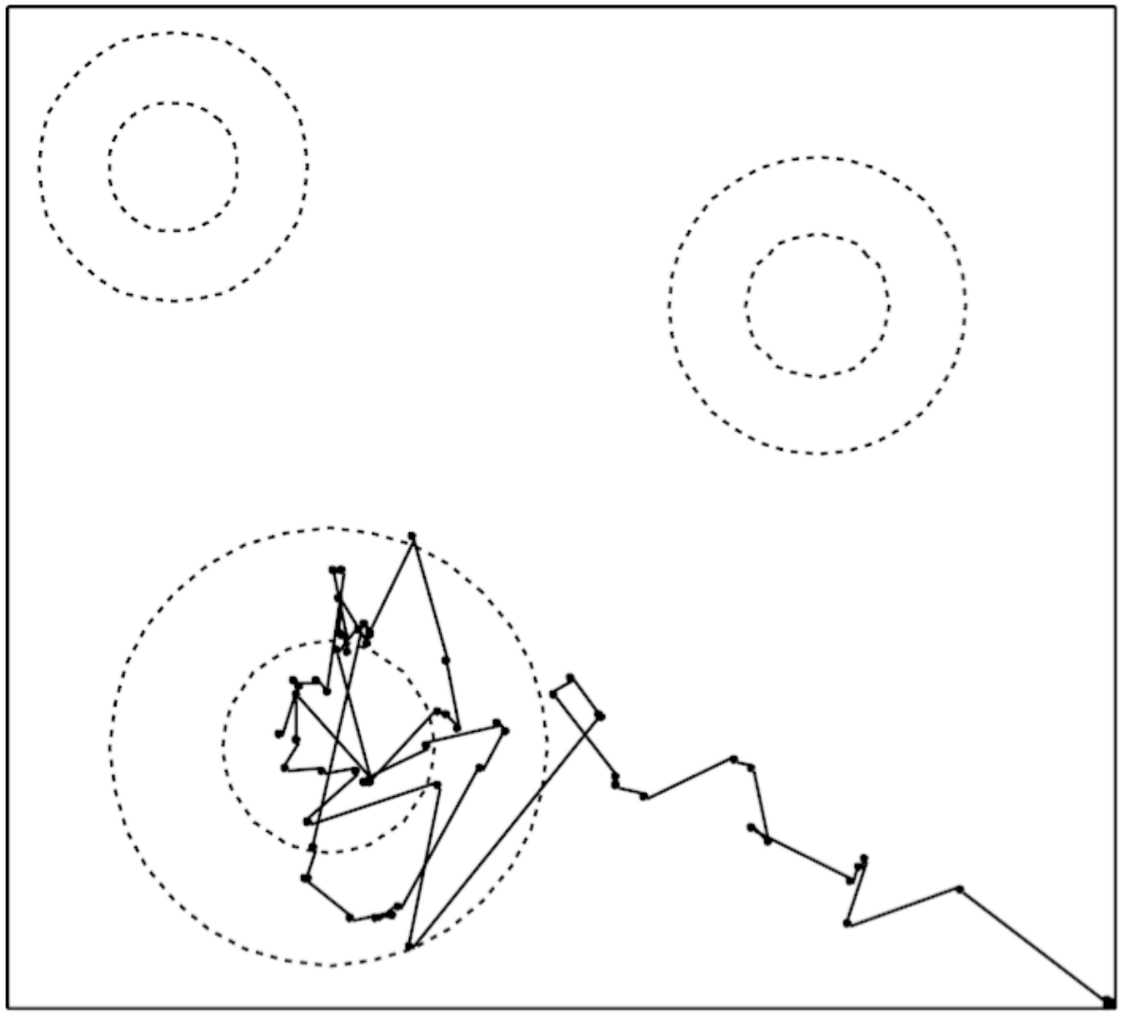
- Random direction
- Gamma distributed step size
- Reflection at edges

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!

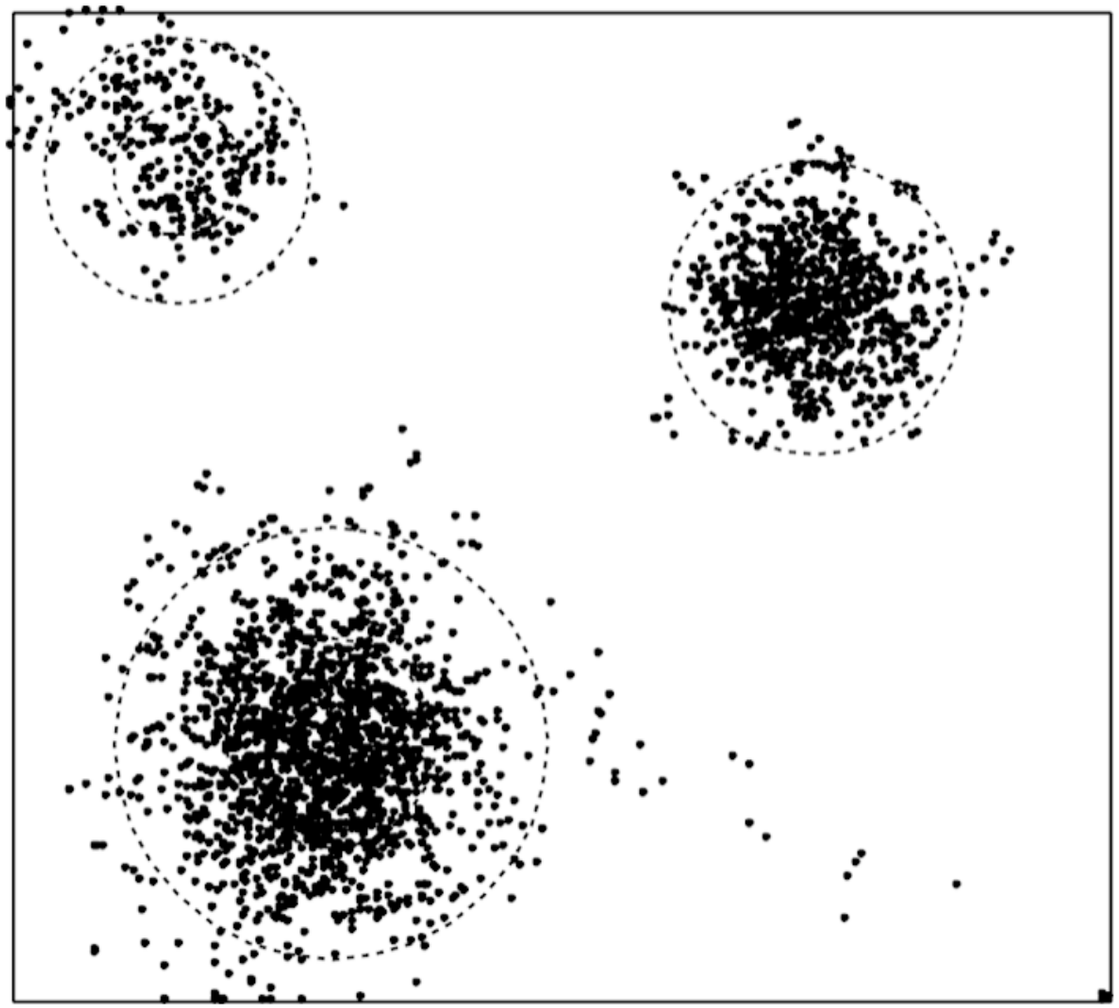
Burn in (courtesy of Paul Lewis)



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

100 steps by the robot

MCMC approximation (courtesy of Paul Lewis)



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

5000 steps by the robot

Summary: MCMC inference

Target distribution

- This is the **posterior** in BEAST2: $P($  $)$

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 mixing, but should not change the results

Marginal distributions

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

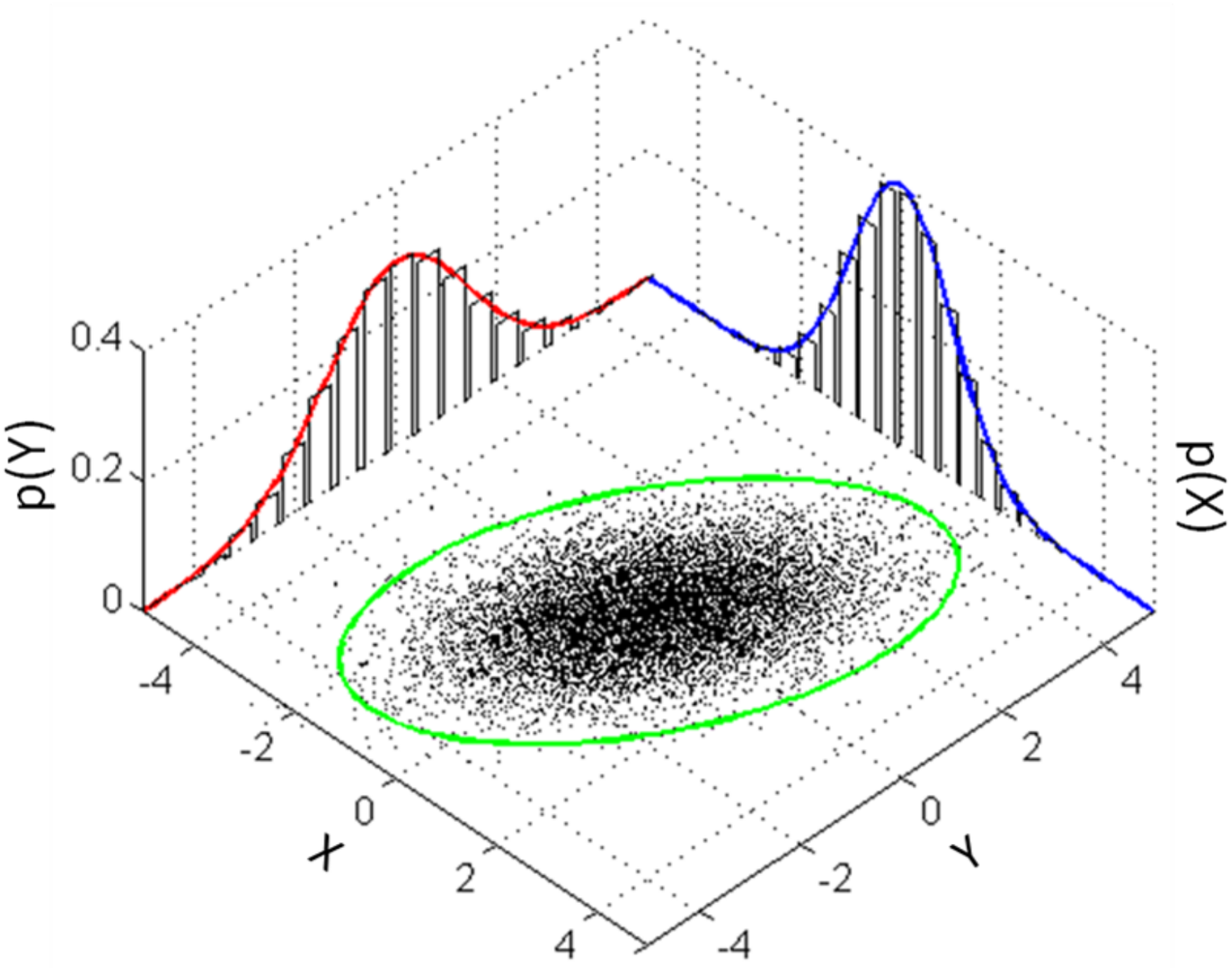
- But we want distributions for each of the parameters we are interested in → marginalize

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

Margin

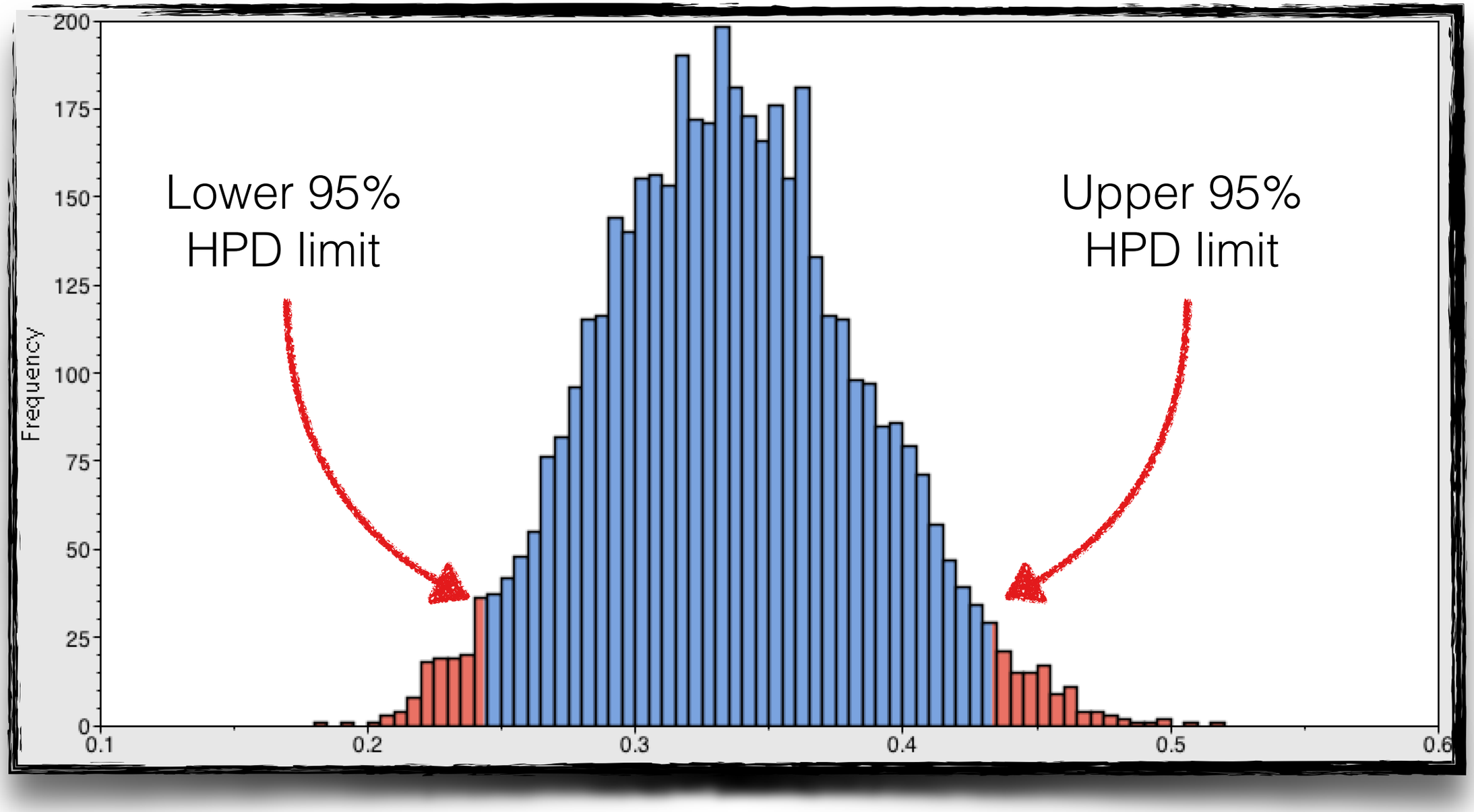
- We on ::)
- But we s
  we are

In practice



HPD intervals

(Highest Posterior Density)



Smallest region that contains 95% of the posterior probability

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)

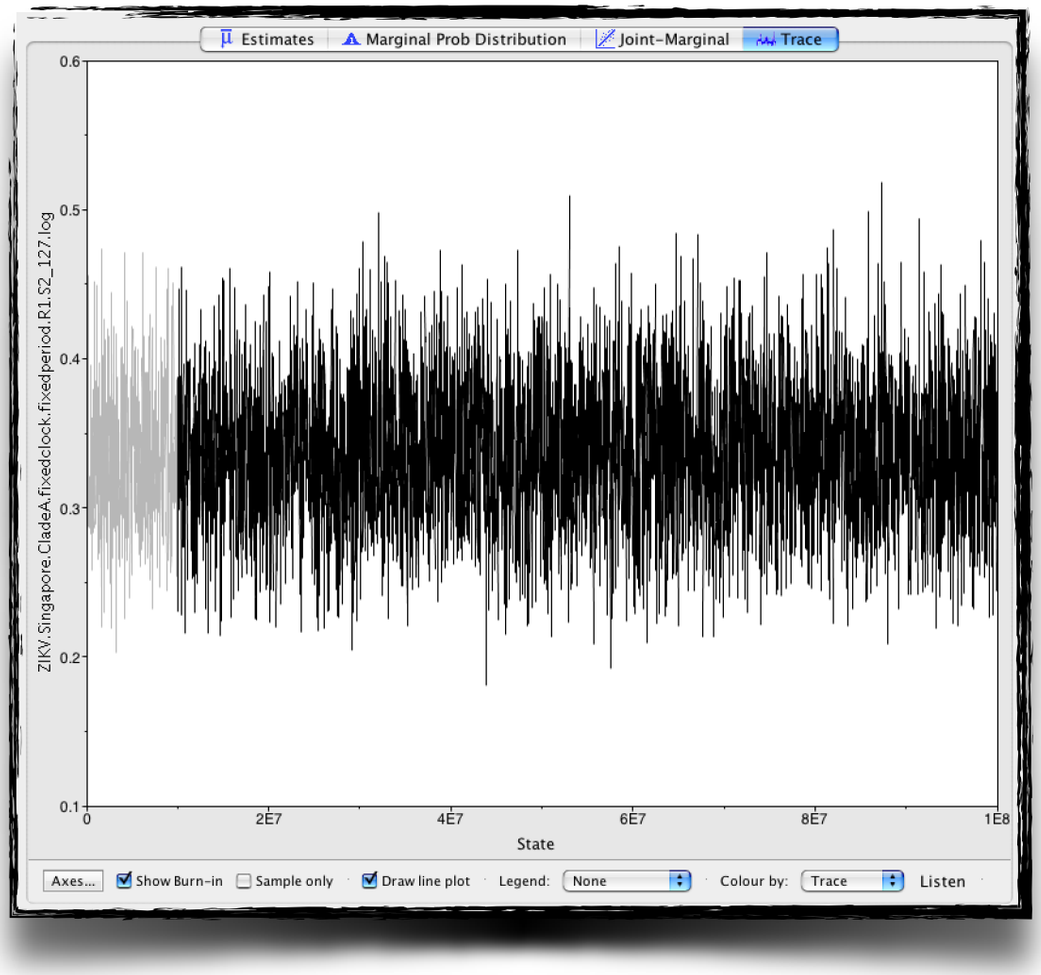
After

- 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

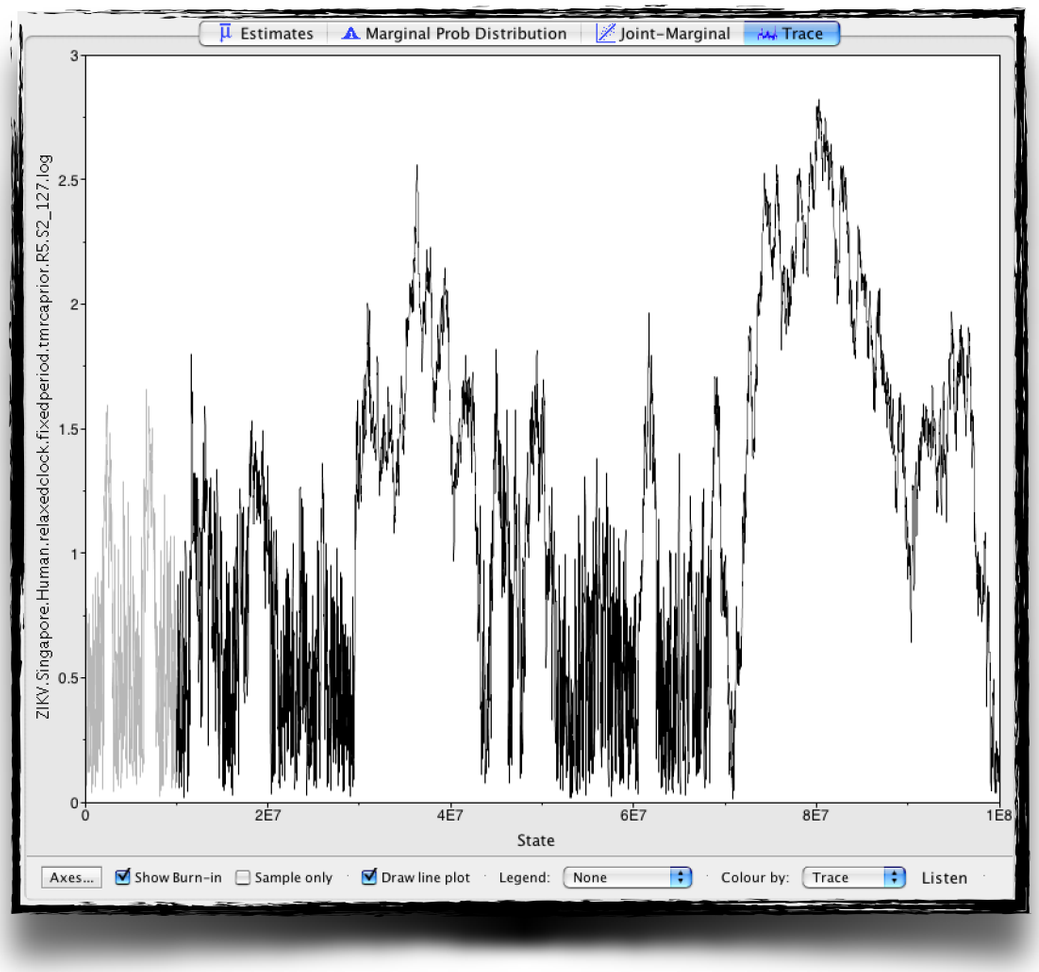
- The MCMC algorithm samples efficiently from high density areas of the posterior distribution
- We end up with a **good** approximation of the posterior distribution in **finite** time
- Everything is awesome!



Mixing well! 😄

What often happens in practice…

- Not so much…



Not mixing! 😭

What o

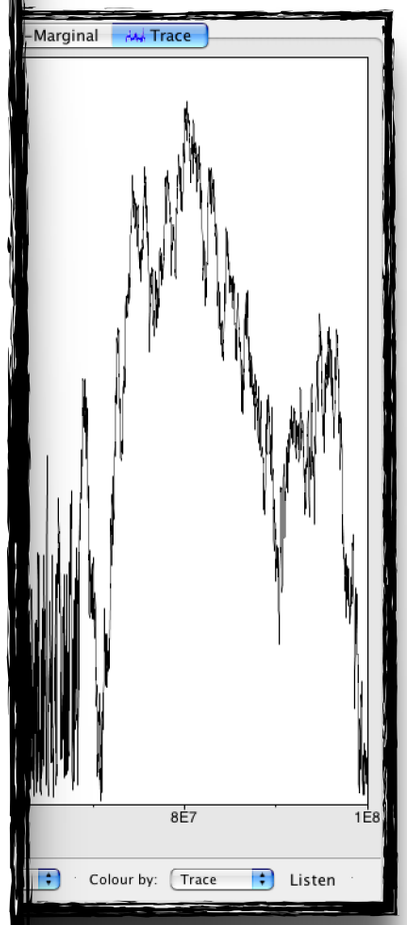
- Not so

What went wrong?

- Recall that MCMC is a Monte Carlo algorithm — it is not **guaranteed** to find the correct solution in finite time!

How can we make it work better?

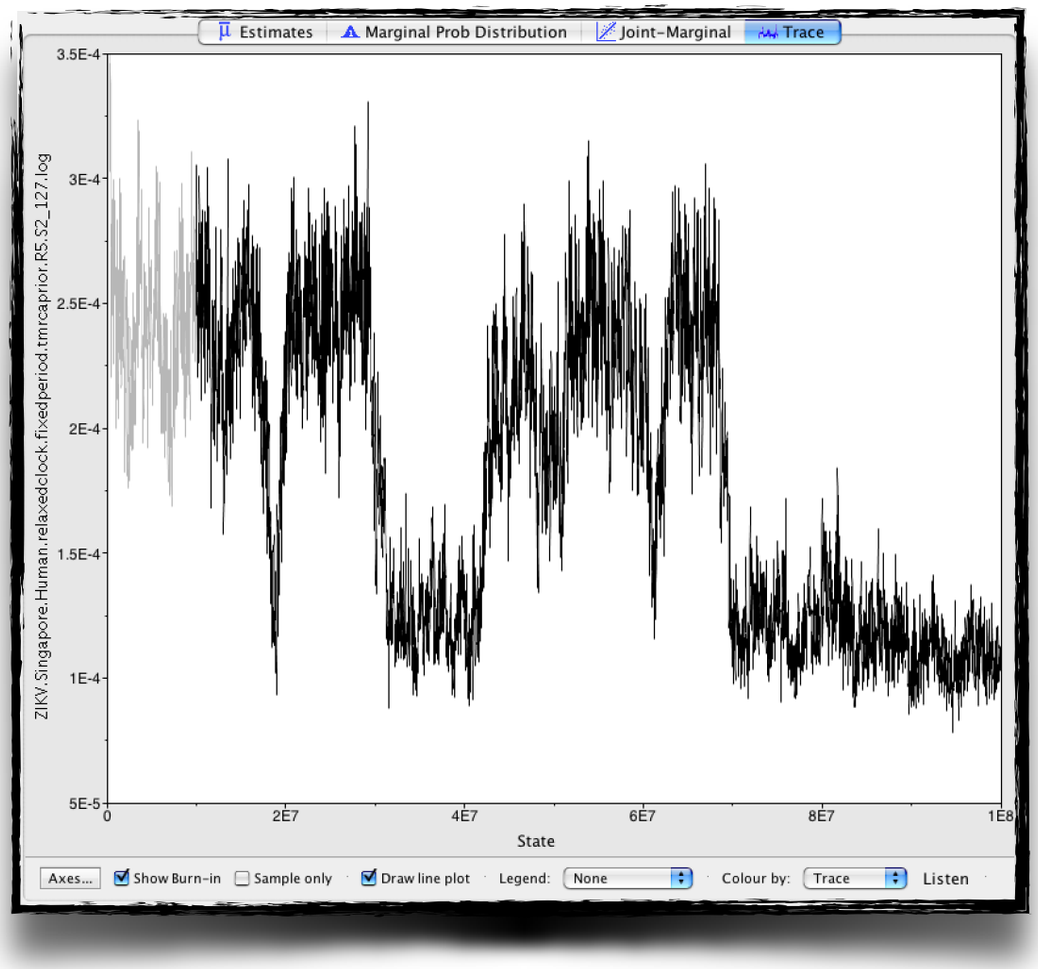
- Tweak the operators to make good proposals
  (increase operator efficiency)
- Fine tune specifics about the MCMC run
  (sampling frequency, length etc.)
- Poor model choice?
- Poor choice of parameterization?



😭

Questions to ask

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



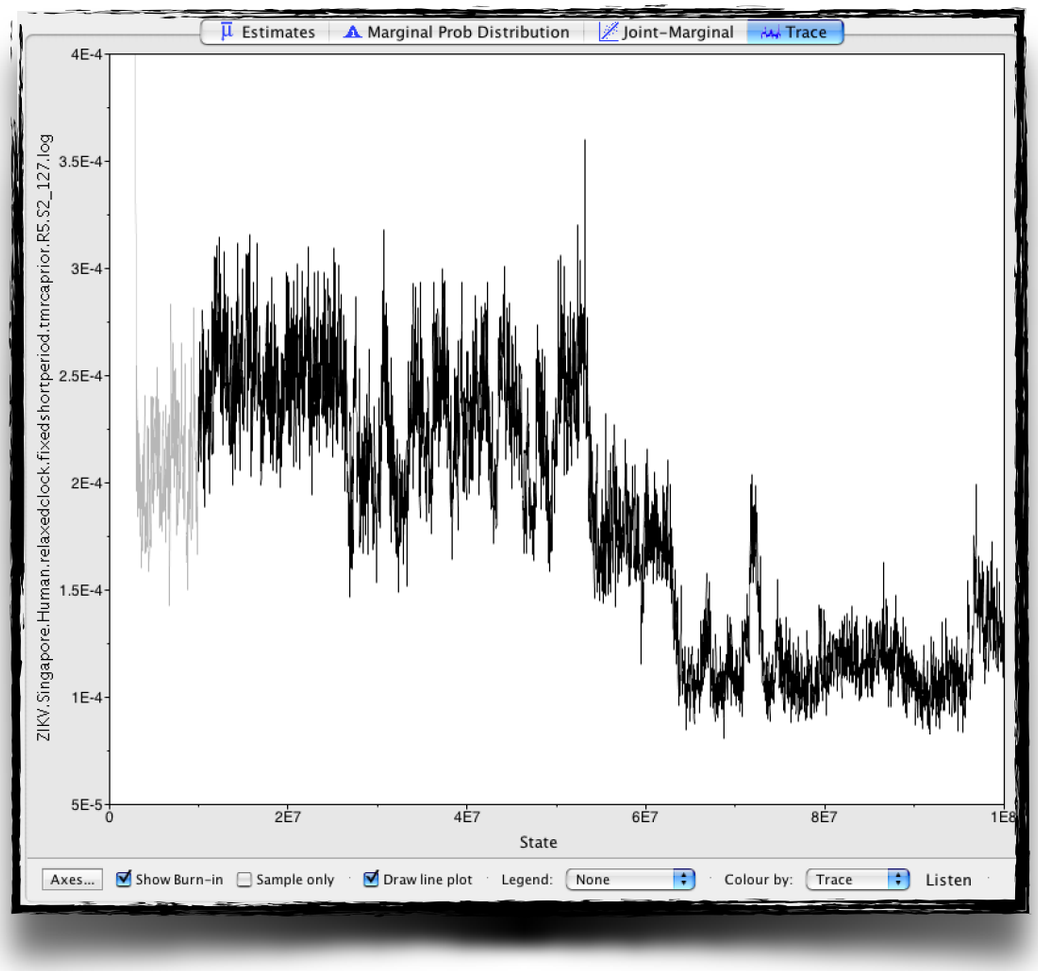
Not mixing! 😨

Solutions

- Tune operators to make better proposals
- Change sampling frequency

Questions to ask

- Has the chain reached the stationary distribution?

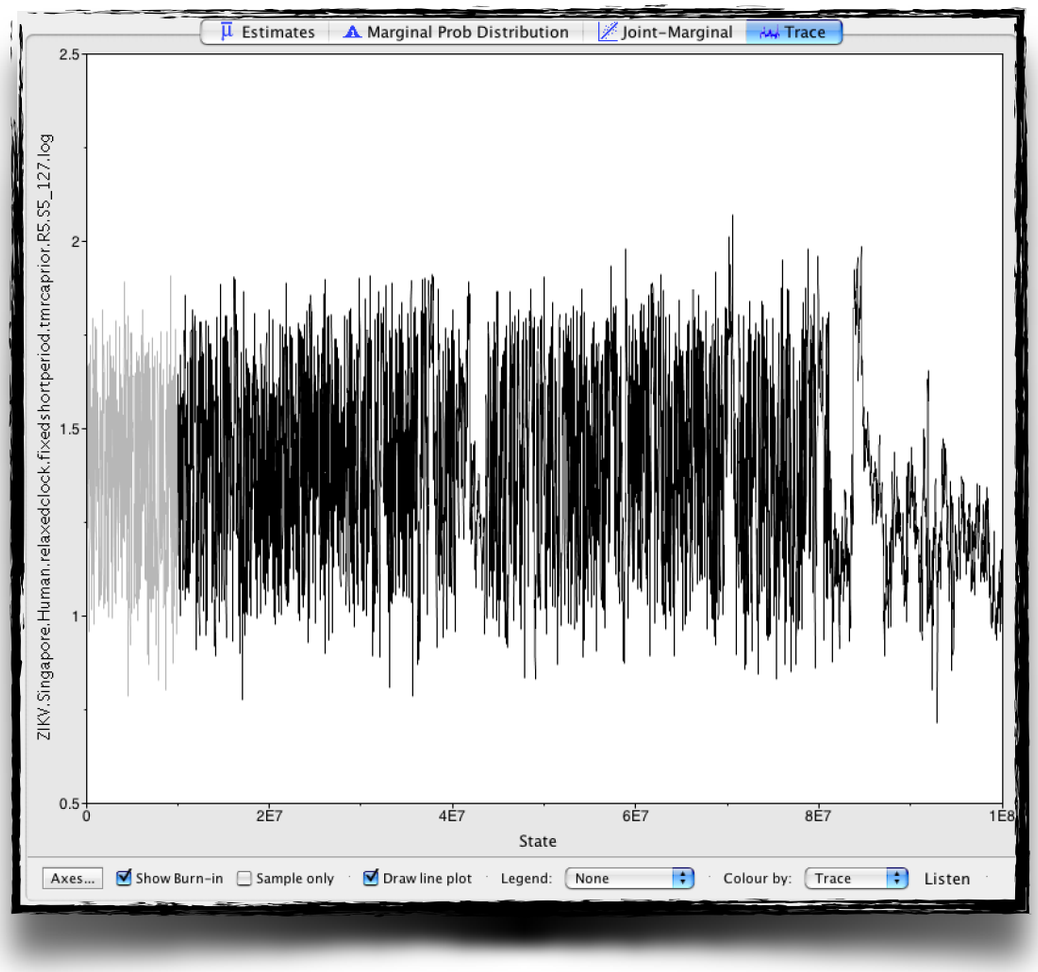


Not converged! 😓

Solution: Run for longer

Questions to ask

- When to stop?
- How to know if the chain is long enough?



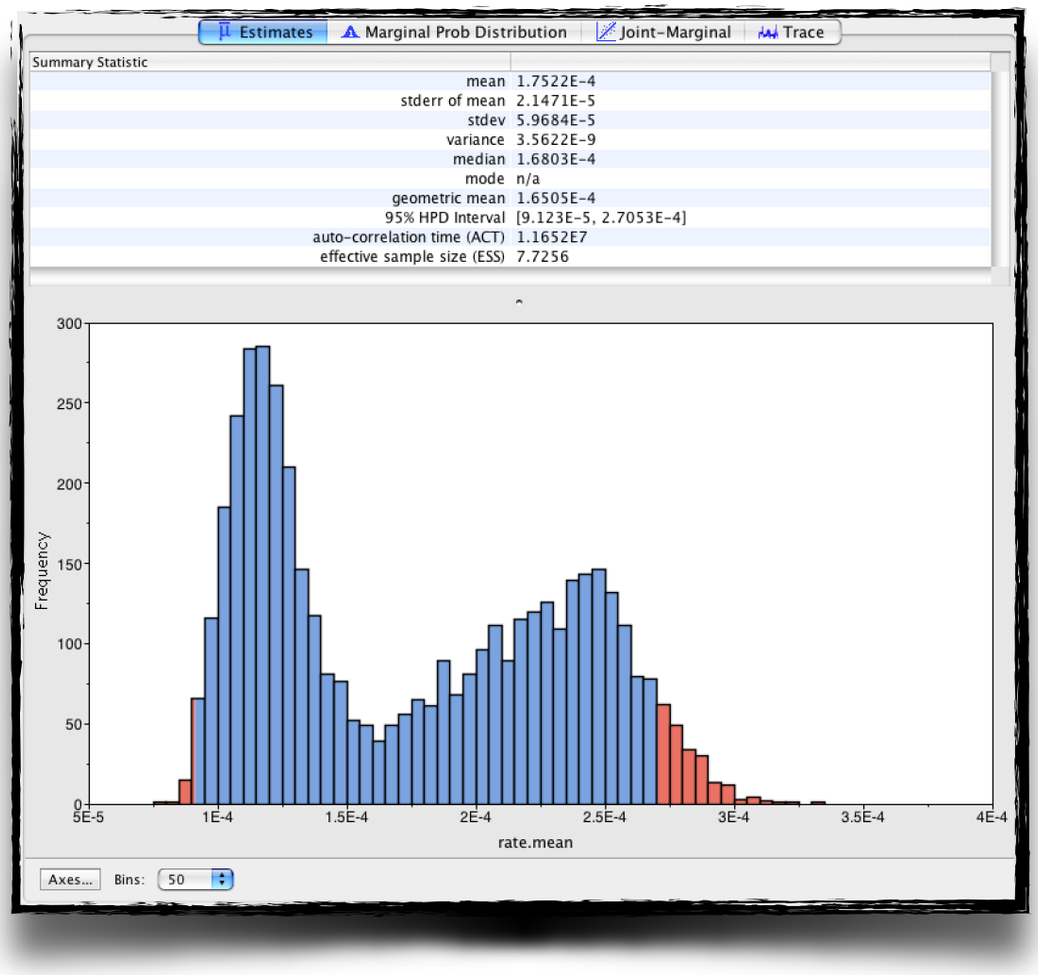
Still not converged! 😵

Solution: Run multiple chains

What if the answer is not what we wanted?

What is happening here?

- If the chain converged then this has something to do with the model
- The model supports a bimodal posterior distribution
- May not be the answer we want but it may be the truth
- Change model
- Change parameterisation

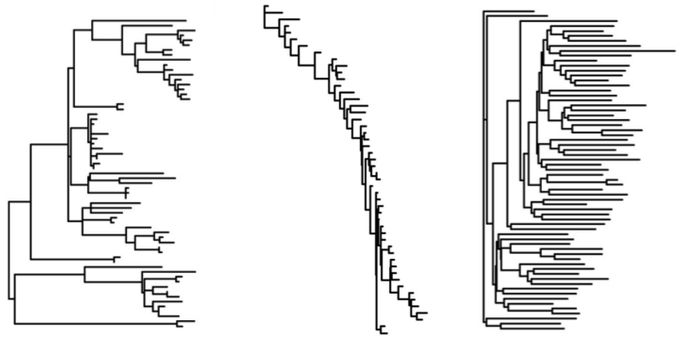


Is this a problem? 🤔

Solution: Be more open-minded

Summary

- Epidemiological dynamics leaves a signal on the branching pattern in the phylogeny



- Phylodynamic models can extract this signal by calculating the probability of the observed phylogeny conditioned on a demographic model

- Using MCMC we can infer the posterior distributions of epidemiological parameters while accounting for phylogenetic uncertainty

- May need to fine-tune operators or MCMC parameters for efficient inference

Tools needed

BEAST2

Software implementing MCMC for model parameter and tree inference

BEAUti

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

Tracer

Analysis of BEAST2 output files (.log)

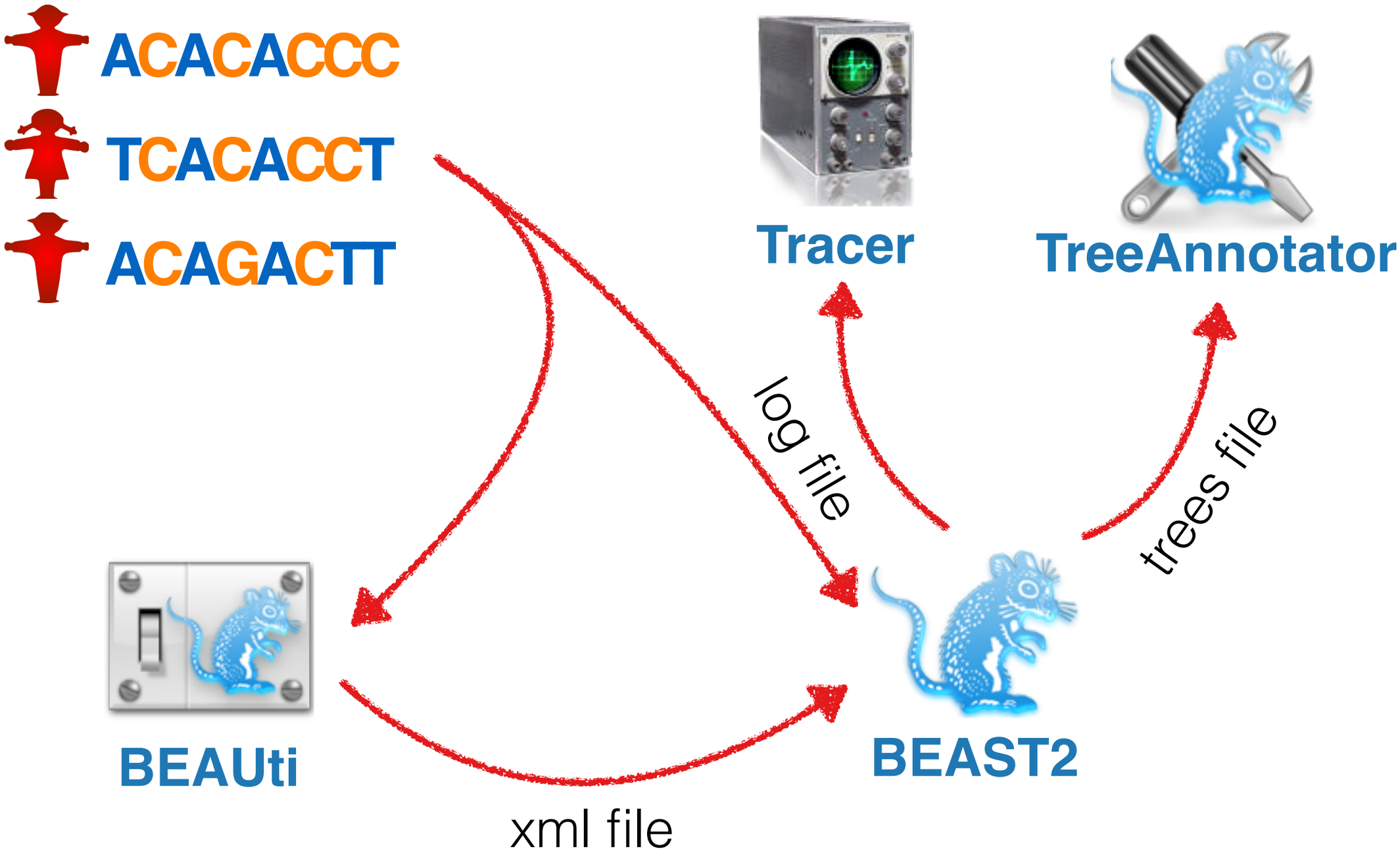
Tree Annotator

Analysis of BEAST2 output files (.trees)

FigTree

Visualisation of trees (.trees)

BEAST2 workflow



BEAUti

GUI for setting up BEAST2 input file in xml format

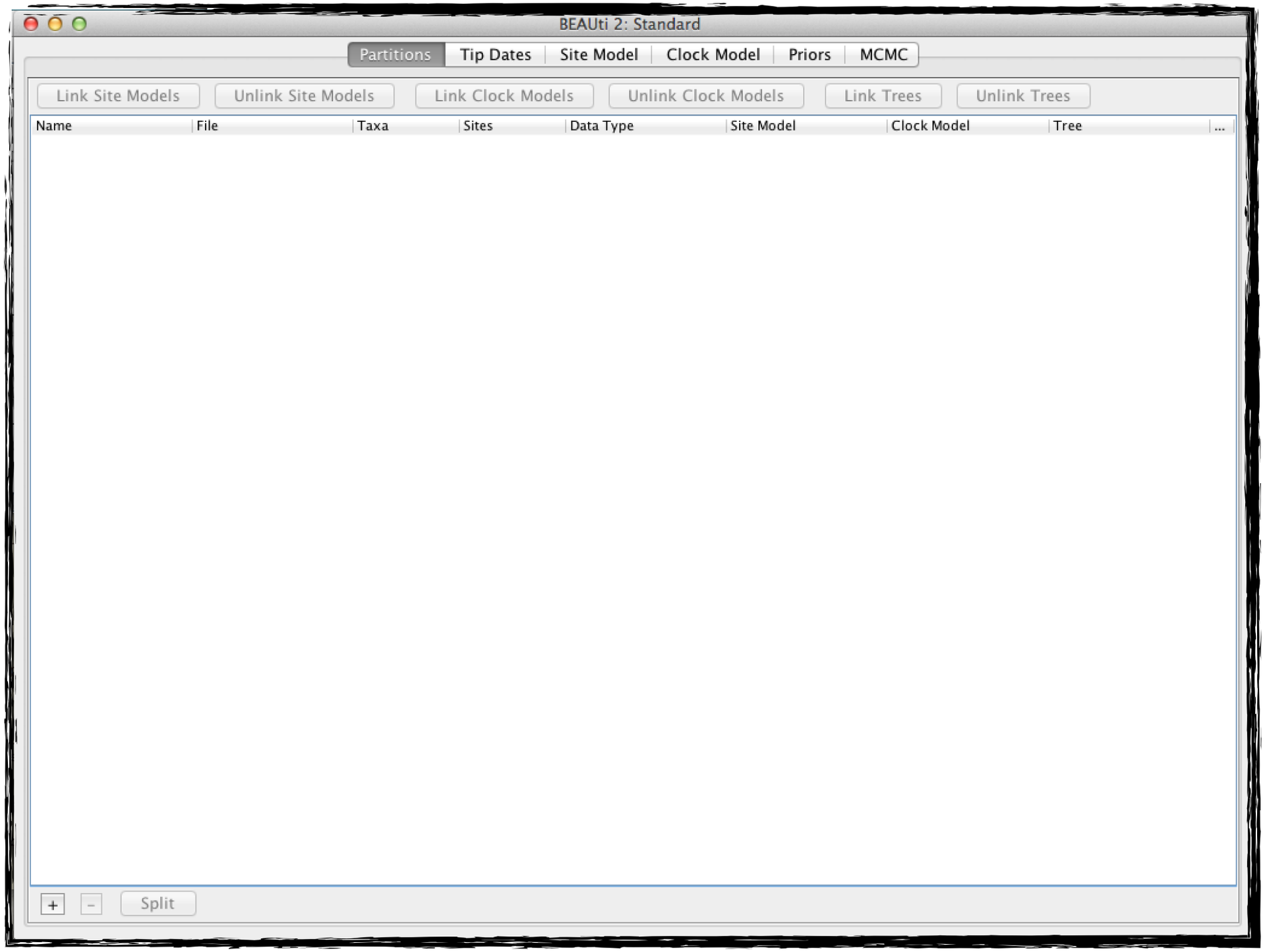
Input:

- Sequence alignment

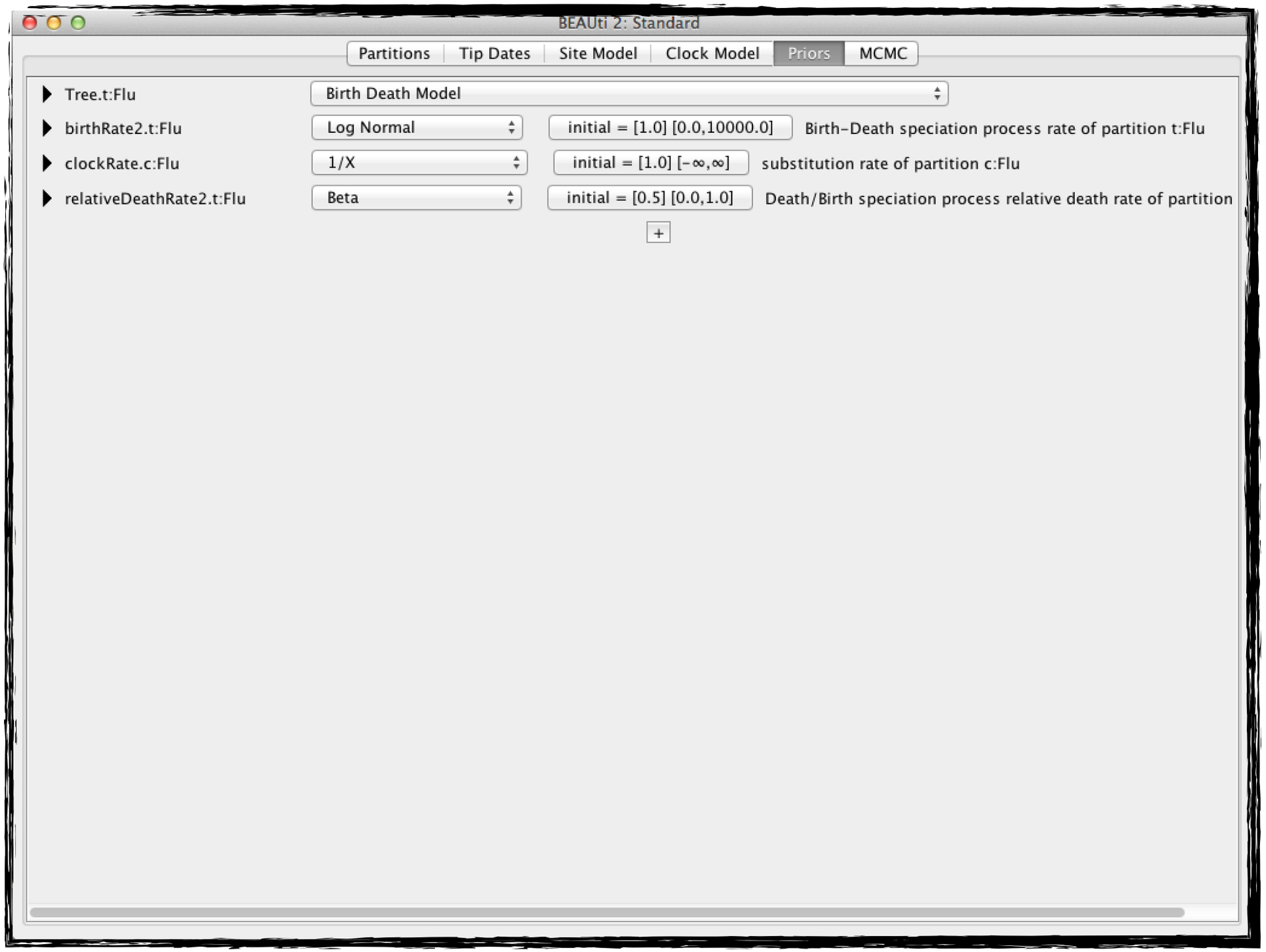
Output:

- xml file

BEAUti


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
...
+
–
Split

BEAUti


BEAUti 2: Standard
Partitions
Tip Dates
Site Model
Clock Model
Priors
MCMC
Tree.t:Flu
Birth Death Model
birthRate2.t:Flu
Log Normal
initial = [1.0] [0.0,10000.0]
Birth-Death speciation process rate of partition t:Flu
clockRate.c:Flu
1/X
initial = [1.0] [-∞,∞]
substitution rate of partition c:Flu
relativeDeathRate2.t:Flu
Beta
initial = [0.5] [0.0,1.0]
Death/Birth speciation process relative death rate of partition
+

BEAUti

GUI for setting up BEAST2 input file in xml format

Input:

- Sequence alignment

Output:

- xml file

BEAUti

testHKY.xml

<beast version='2.0' namespace='beast.evolution.alignment:beast.core:beast.evolution.tree.coalescent:beast.core.ut…

```
<beast version='2.0'
       namespace='beast.evolution.alignment:beast.core:beast.evolution.tree.coalescent:beast.core.util:
           beast.evolution.nuc:beast.evolution.operators:beast.evolution.sitemodel:
           beast.evolution.substitutionmodel:beast.evolution.likelihood'>


    <!-- The sequence alignment                                              -->
    <!-- ntax=6 nchar=768                                                    -->
    <!-- npatterns=69                                                        -->
    <data id="alignment" dataType="nucleotide">
        <sequence taxon="human">

                AGAAATATGTCTGATAAAAGAGTTACTTTGATAGAGTAAATAATAGGAGCTTAAACCCCCTTATTTCTACTAGGACTATGAGAATCGAACCCAT
                CCCTGAGAATCCAAAATTCTCCGTGCCACCTATCACACCCCATCCTAAGTAAGGTCAGCTAAATAAGCTATCGGGCCCATACCCCGAAAATGTT
                GGTTATACCCTTCCCGTACTAAGAAATTTAGGTTAAATACAGACCAAGAGCCTTCAAAGCCCTCAGTAAGTTG-
                CAATACTTAATTTCTGTAAGGACTGCAAAACCCCACTCTGCATCAACTGAACGCAAATCAGCCACTTTAATTAAGCTAAGCCCTTCTAGACCAA
                TGGGACTTAAACCCACAAACACTTAGTTAACAGCTAAGCACCCTAATCAAC-TGGCTTCAATCTAAAGCCCCGGCAGG-
                TTTGAAGCTGCTTCTTCGAATTTGCAATTCAATATGAAAA-
                TCACCTCGGAGCTTGGTAAAAAGAGGCCTAACCCCTGTCTTTAGATTTACAGTCCAATGCTTCA-
                CTCAGCCATTTTACCACAAAAAAGGAAGGAATCGAACCCCCCAAAGCTGGTTTCAAGCCAACCCCATGGCCTCCATGACTTTTTCAAAAGGTAT
                TAGAAAAACCATTTCATAACTTTGTCAAAGTTAAATTATAGGCT-AAATCCTATATATCTTA-
                CACTGTAAAGCTAACTTAGCATTAACCTTTTAAGTTAAAGATTAAGAGAACCAACACCTCTTTACAGTGA
        </sequence>
        <sequence taxon="chimp">

                AGAAATATGTCTGATAAAAGAATTACTTTGATAGAGTAAATAATAGGAGTTCAAATCCCCTTATTTCTACTAGGACTATAAGAATCGAACTCAT
                CCCTGAGAATCCAAAATTCTCCGTGCCACCTATCACACCCCATCCTAAGTAAGGTCAGCTAAATAAGCTATCGGGCCCATACCCCGAAAATGTT
                GGTTACACCCTTCCCGTACTAAGAAATTTAGGTTAAGCACAGACCAAGAGCCTTCAAAGCCCTCAGCAAGTTA-
                CAATACTTAATTTCTGTAAGGACTGCAAAACCCCACTCTGCATCAACTGAACGCAAATCAGCCACTTTAATTAAGCTAAGCCCTTCTAGATTAA
                TGGGACTTAAACCCACAAACATTTAGTTAACAGCTAAACACCCTAATCAAC-TGGCTTCAATCTAAAGCCCCGGCAGG-
                TTTGAAGCTGCTTCTTCGAATTTGCAATTCAATATGAAAA-
                TCACCTCAGAGCTTGGTAAAAAGAGGCTTAACCCCTGTCTTTAGATTTACAGTCCAATGCTTCA-
                CTCAGCCATTTTACCACAAAAAAGGAAGGAATCGAACCCCCTAAAGCTGGTTTCAAGCCAACCCCATGACCTCCATGACTTTTTCAAAAGATAT
                TAGAAAAACTATTTCATAACTTTGTCAAAGTTAAATTACAGGTT-AACCCCCGTATATCTTA-
                CACTGTAAAGCTAACCTAGCATTAACCTTTTAAGTTAAAGATTAAGAGGACCGACACCTCTTTACAGTGA
        </sequence>
        <sequence taxon="bonobo">

                AGAAATATGTCTGATAAAAGAATTACTTTGATAGAGTAAATAATAGGAGTTTAAATCCCCTTATTTCTACTAGGACTATGAGAGTCGAACCCAT
                CCCTGAGAATCCAAAATTCTCCGTGCCACCTATCACACCCCATCCTAAGTAAGGTCAGCTAAATAAGCTATCGGGCCCATACCCCGAAAATGTT
                GGTTATACCCTTCCCGTACTAAGAAATTTAGGTTAAACACAGACCAAGAGCCTTCAAAGCTCTCAGTAAGTTA-
                CAATACTTAATTTCTGTAAGGACTGCAAAACCCCACTCTGCATCAACTGAACGCAAATCAGCCACTTTAATTAAGCTAAGCCCTTCTAGATTAA
                TGGGACTTAAACCCACAAACATTTAGTTAACAGCTAAACACCCTAATCAGC-TGGCTTCAATCTAAAGCCCCGGCAGG-
                TTTGAAGCTGCTTCTTTGAATTTGCAATTCAATATGAAAA-
                TCACCTCAGAGCTTGGTAAAAAGAGGCTTAACCCCTGTCTTTAGATTTACAGTCCAATGCTTCA-
```

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

- Initially planned to be an extension to BEAST1
  → now two separate software packages

- BEAST2 has most of the functionality of BEAST1

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

Input:
- xml file

Outputs:
- log file
- trees file
- state file

BEAST2 (http://beast2.org)

- **B**ayesian **ev**

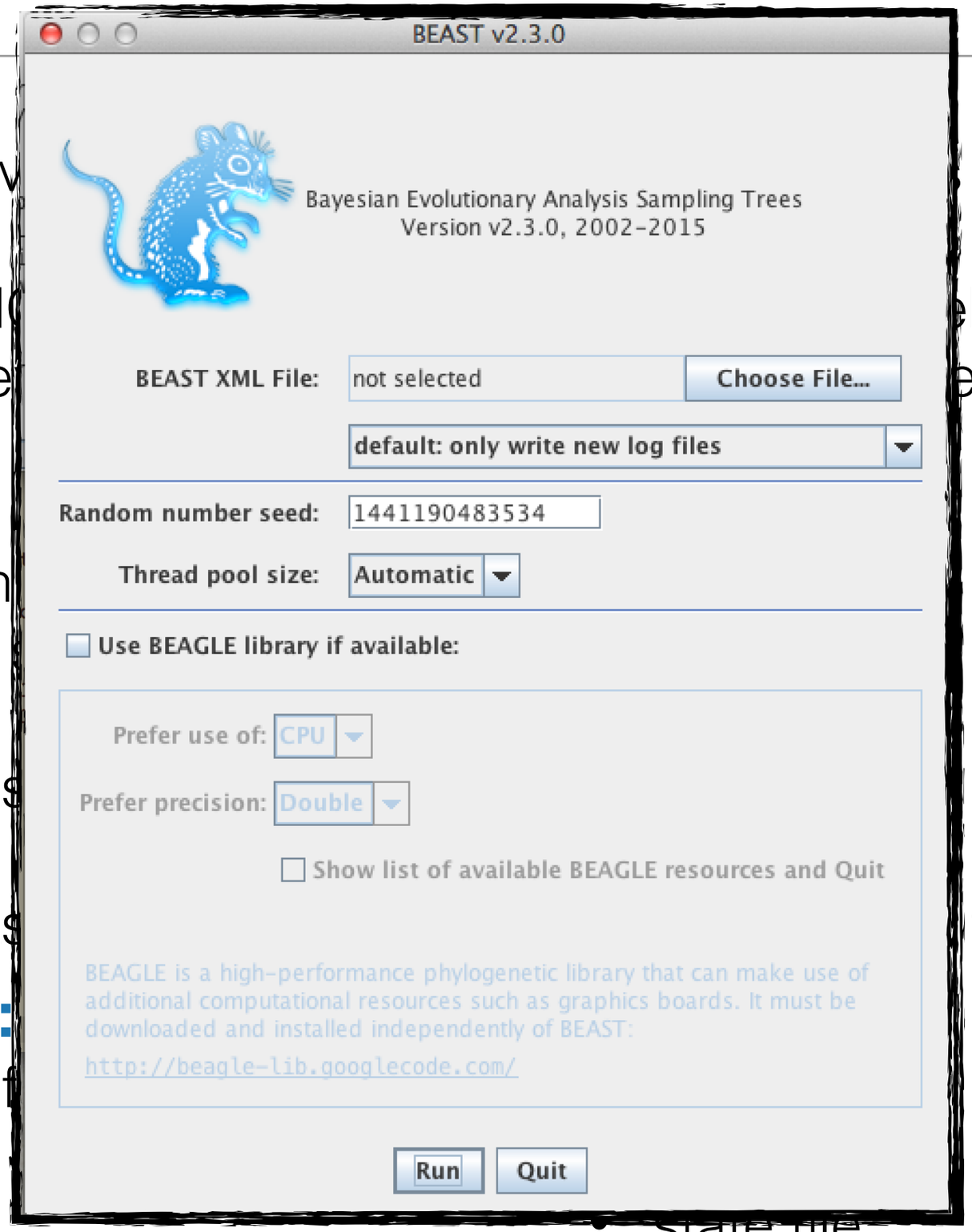
- Performs M elected sequence e eciation) model

- Initially plan
  → now two

- BEAST2 has

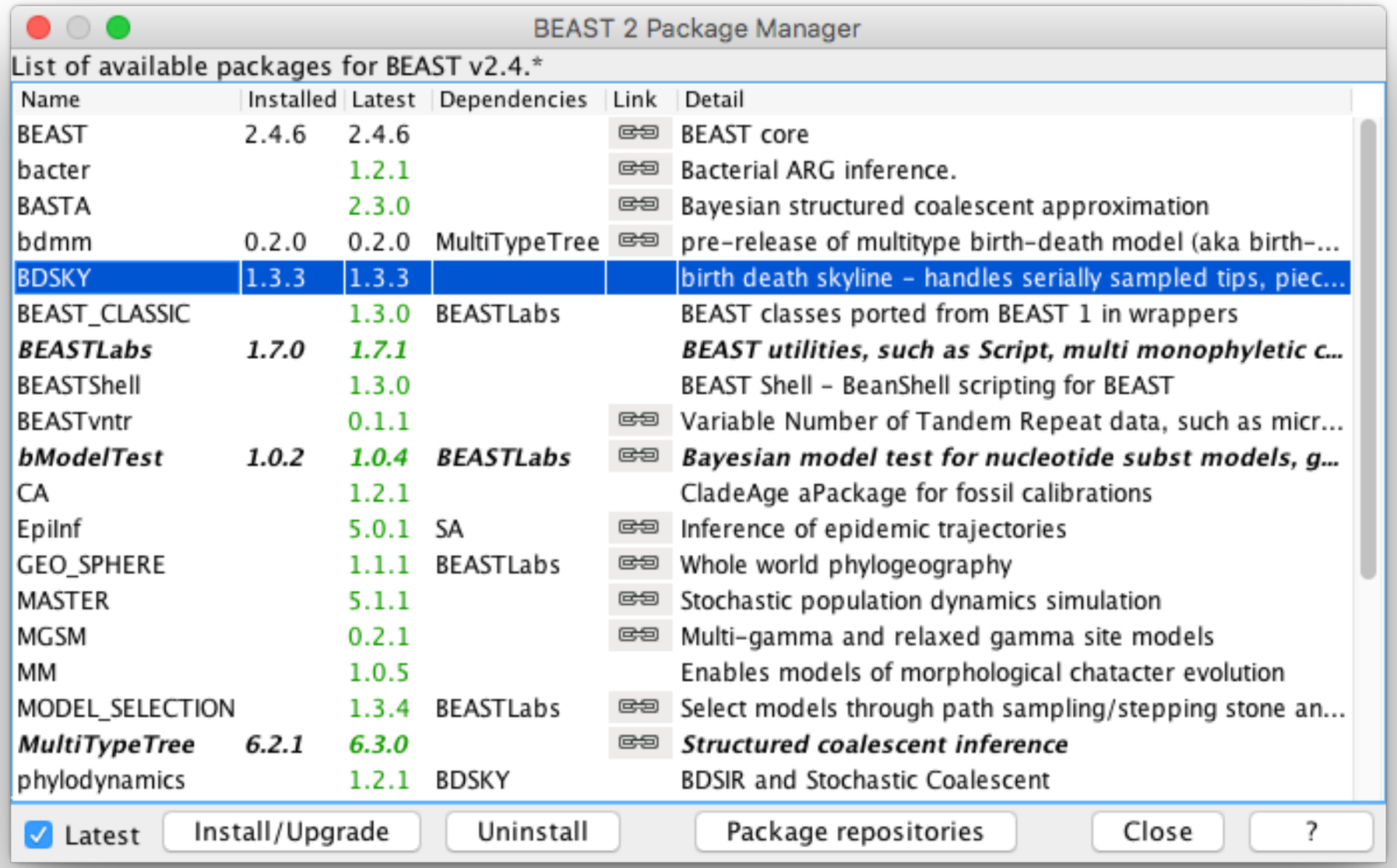
- BEAST2 has to extend

  Input:
  - xml f
  - state file



Installing new BEAST2 packages

Install packages from **BEAUti**



Tracer

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

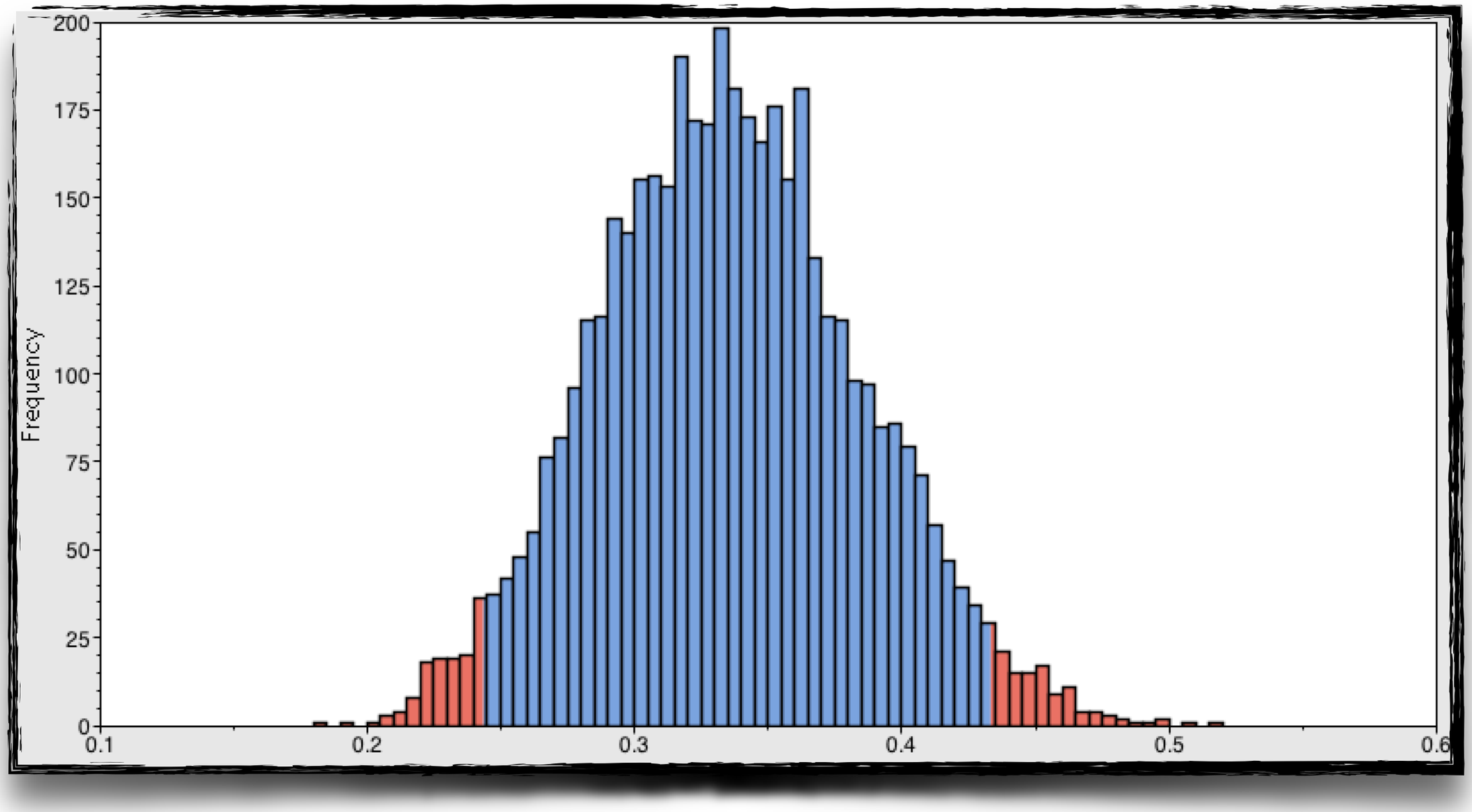
Input:

- log file

Output:

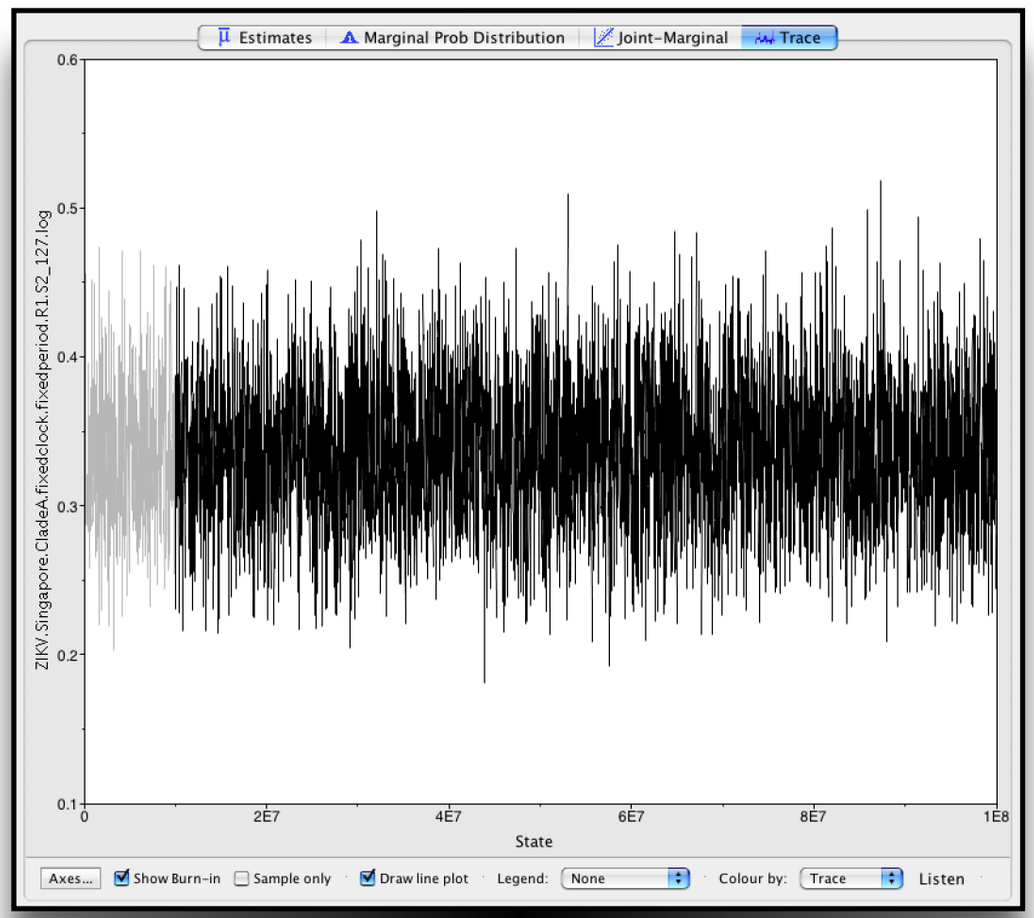
- Gain insight

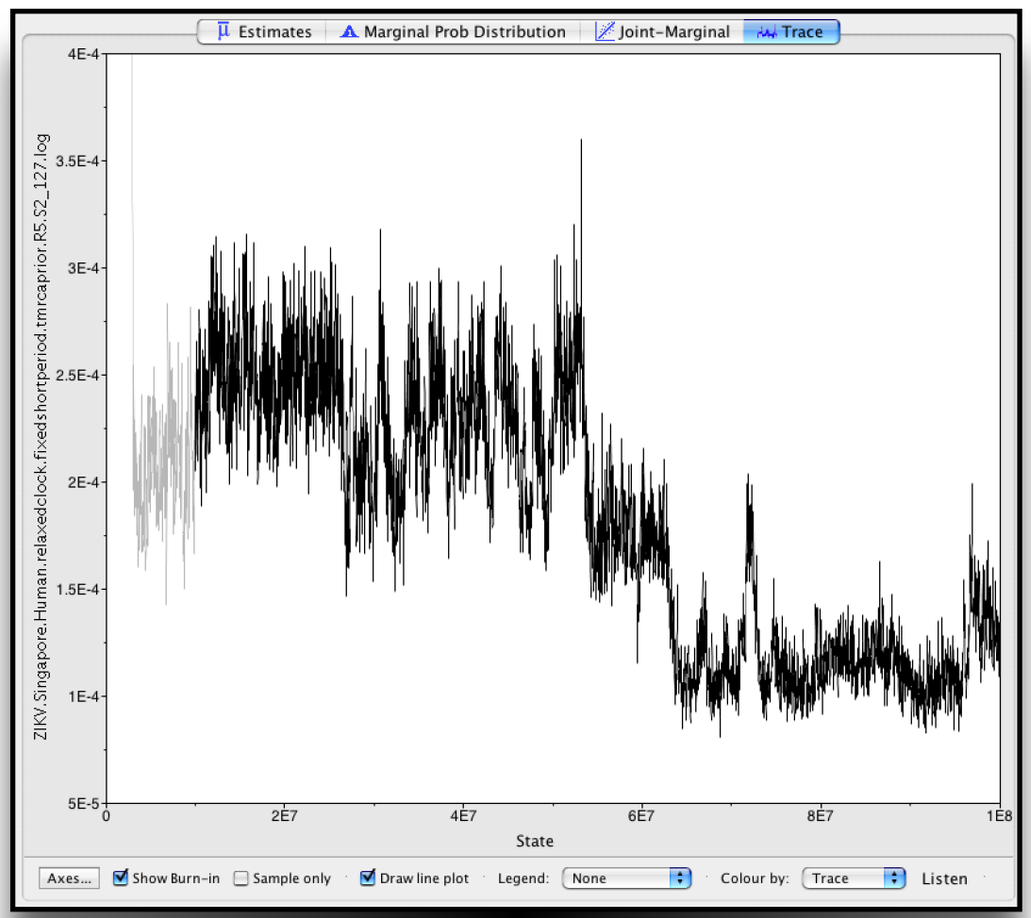
Tracer



Tracer

Look at the chains first!





Mixing well! 😄

Not mixing! 😭

TreeAnnotator

- Analyze trees file from BEAST2 runs
- Produces MCC tree with node annotations (posterior probability)
- Note that the MCC tree may never actually appear in the trees file

Input:

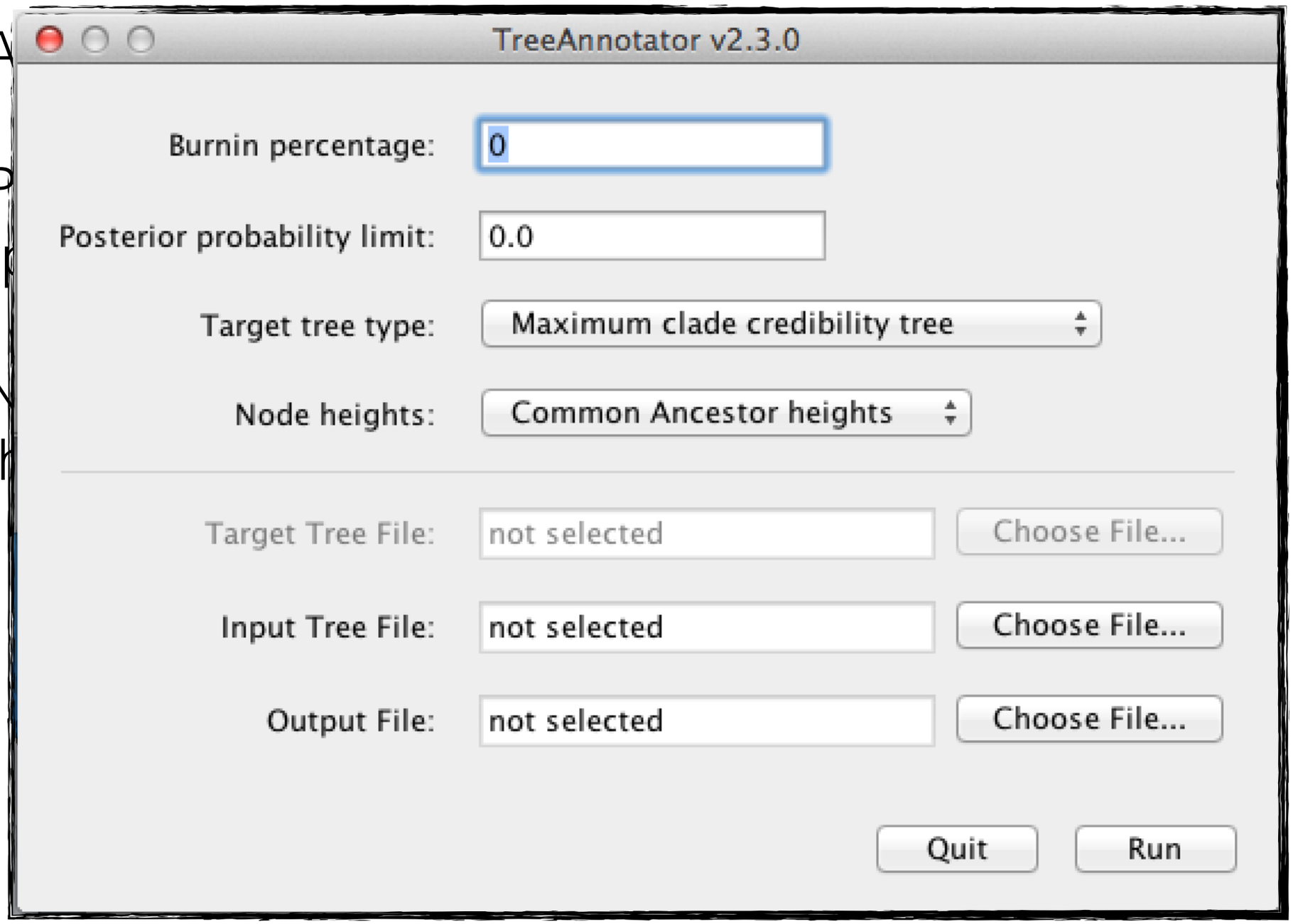
- trees file

Output:

- MCC tree

TreeAnnotator

- A
- P
(
- N
th



n

FigTree

- Analyze trees file from BEAST2 runs
- Visual analysis only

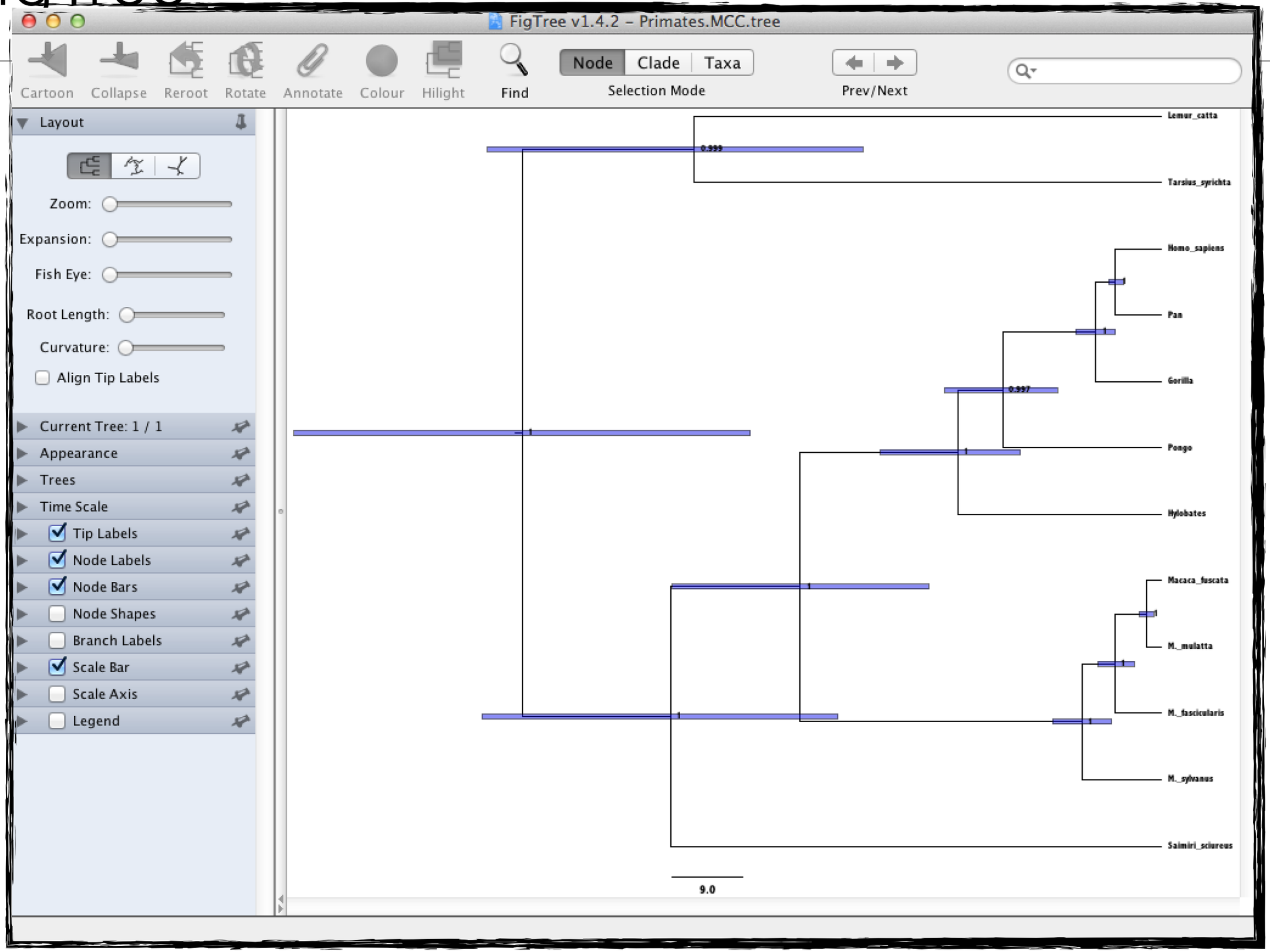
Input:

- trees file

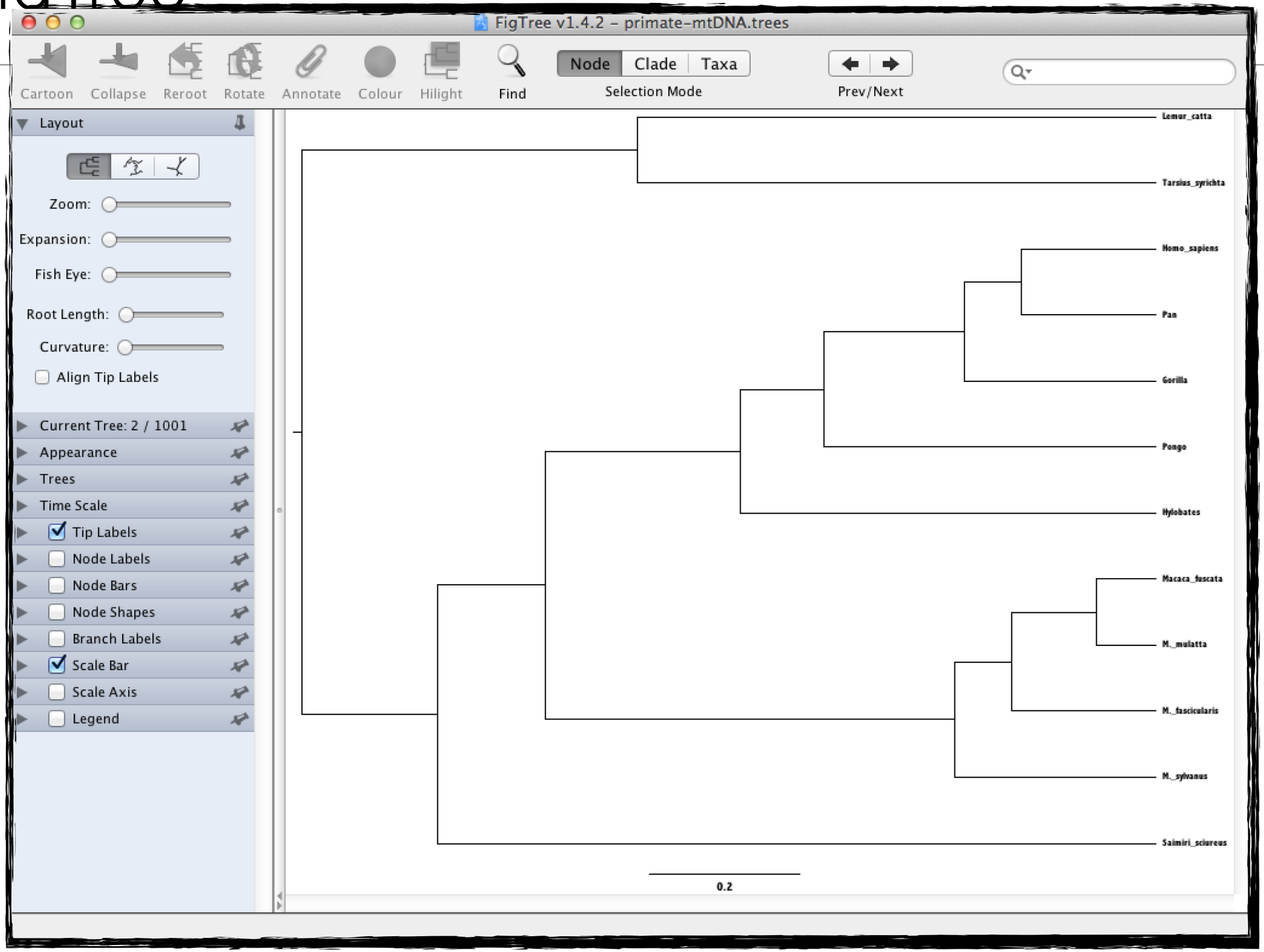
Output:

- Gain insight

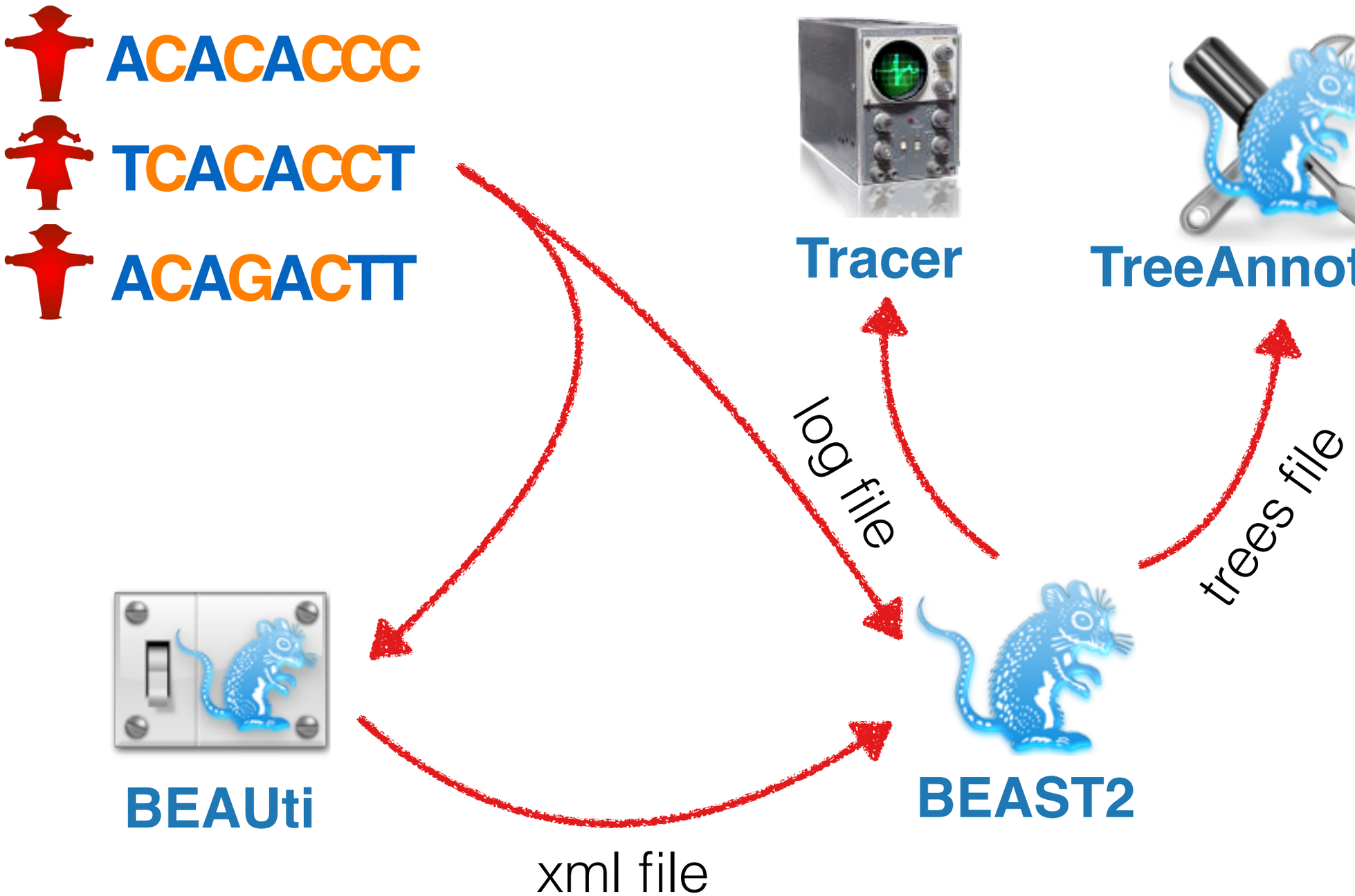
FigTree


FigTree v1.4.2 – Primates.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
Node Labels
Node Bars
Node Shapes
Branch Labels
Scale Bar
Scale Axis
Legend
0.999
0.997
Lemur_catta
Tarsius_syrichta
Homo_sapiens
Pan
Gorilla
Pongo
Hylobates
Macaca_fascata
M._mulatta
M._fascicularis
M._sylvanus
Saimiri_sciureus
9.0

FigTree


FigTree v1.4.2 - primate-mtDNA.trees
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: 2 / 1001
Appearance
Trees
Time Scale
Tip Labels
Node Labels
Node Bars
Node Shapes
Branch Labels
Scale Bar
Scale Axis
Legend
Lemur_catta
Tarsius_syrichta
Homo_sapiens
Pan
Gorilla
Pongo
Hylobates
Macaca_fuscata
M._mulatta
M._fascicularis
M._sylvanus
Saimiri_sciureus
0.2

Workflow revisited



Help

BEAST 1.8 and TemPest tutorials

- http://beast.bio.ed.ac.uk/tutorials
- https://perswww.kuleuven.be/~u0036765/SISMID/practicals.html

BEAST 2 tutorials

- https://taming-the-beast.github.io
- https://www.beast2.org/tutorials/
- http://groups.google.com/group/beast-users

BEAST users discussion group

- http://groups.google.com/group/beast-users

Thanks

Some slides inspired by (or shamelessly copied from) slides by Paul Lewis, Jūlija Pečerska & Veronika Bošková and Oliver Pybus

Thank you for listening!