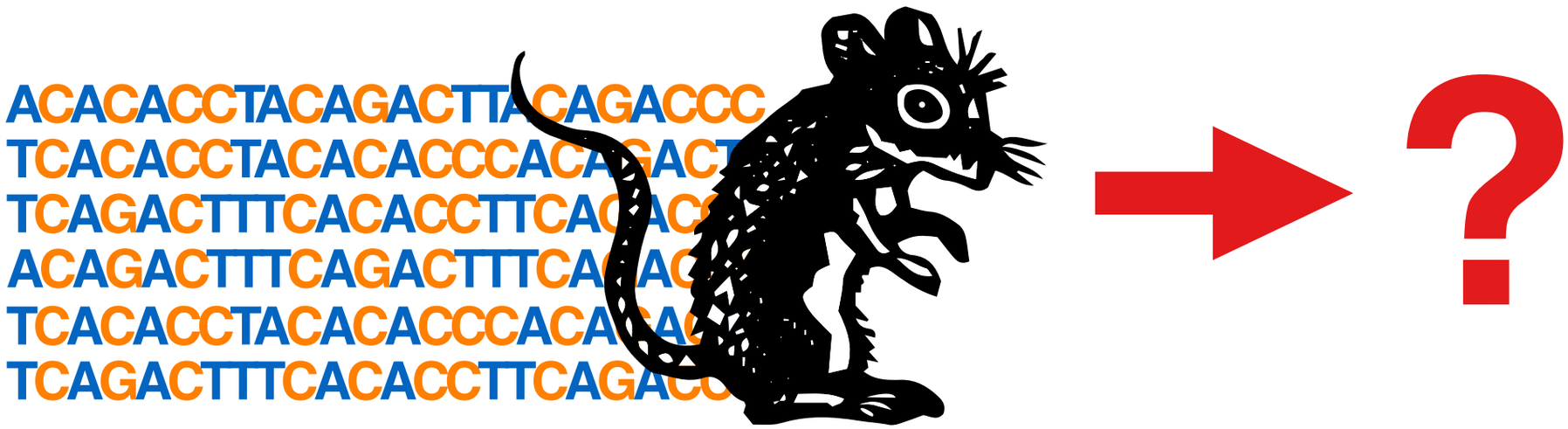
Taming the BEAST

Bayesian Evolutionary Analyses by Sampling Trees

Louis du Plessis

1. What goes into a BEAST analysis?
2. BEAST2 workflow
3. Tutorial introduction and demo

We all have one thing in common…



BEAST2

We all use (or want to use) **BEAST2** to answer questions about our data

but *how*?

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 *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

Bayesian **inference** recap

Likelihood

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

Posterior

Prior

Model evidence

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

Prior → P(model)

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

Likelihood → P(data | model)

- Probability of data given parameters (defined by model)

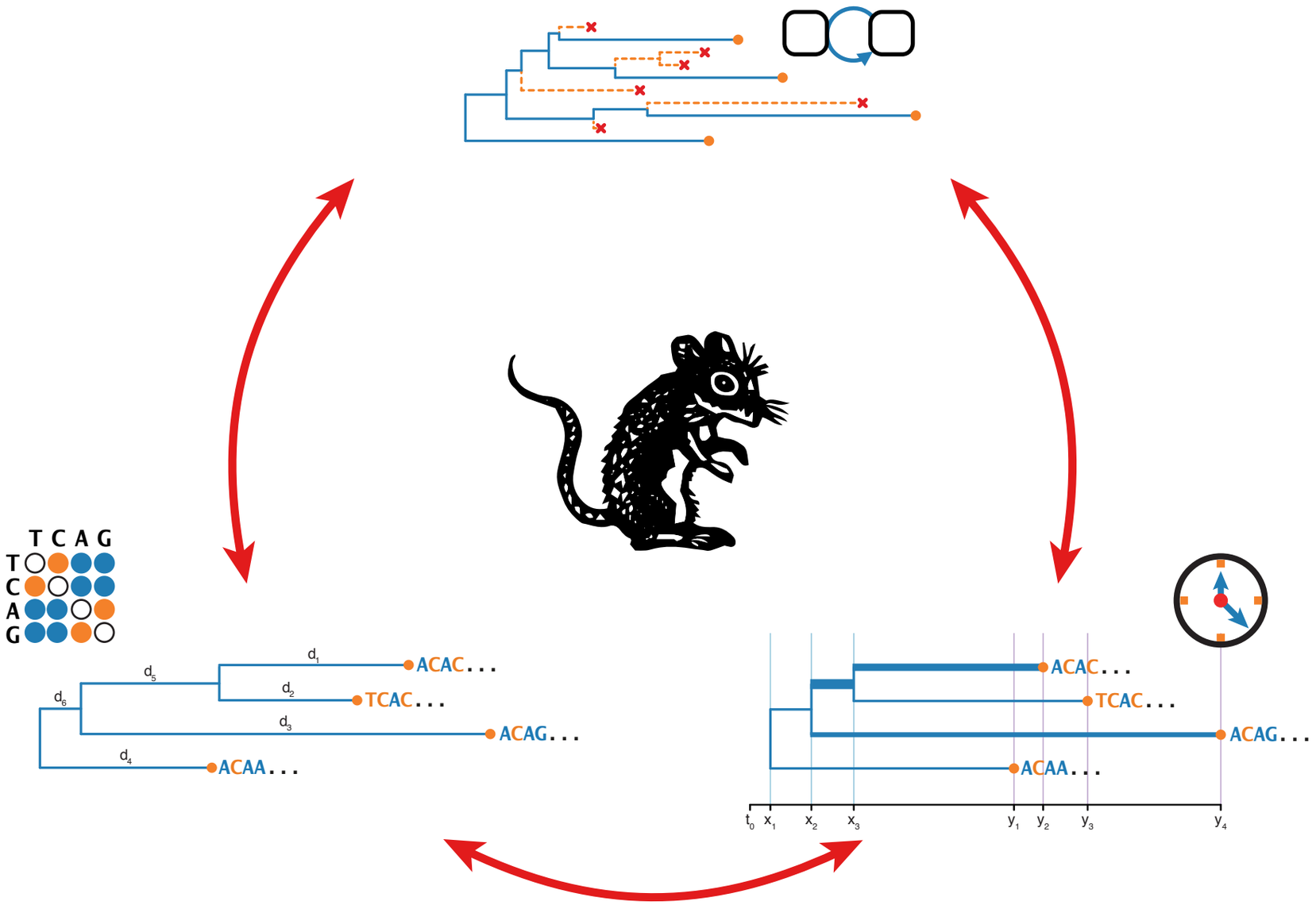
Posterior → P(model | data)

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

Model evidence → P(data)

- Probability for data given model (any combination of parameters)
- Used for Bayesian model selection

What goes into a **BEAST** analysis?



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

genealogy

demographic model

site model

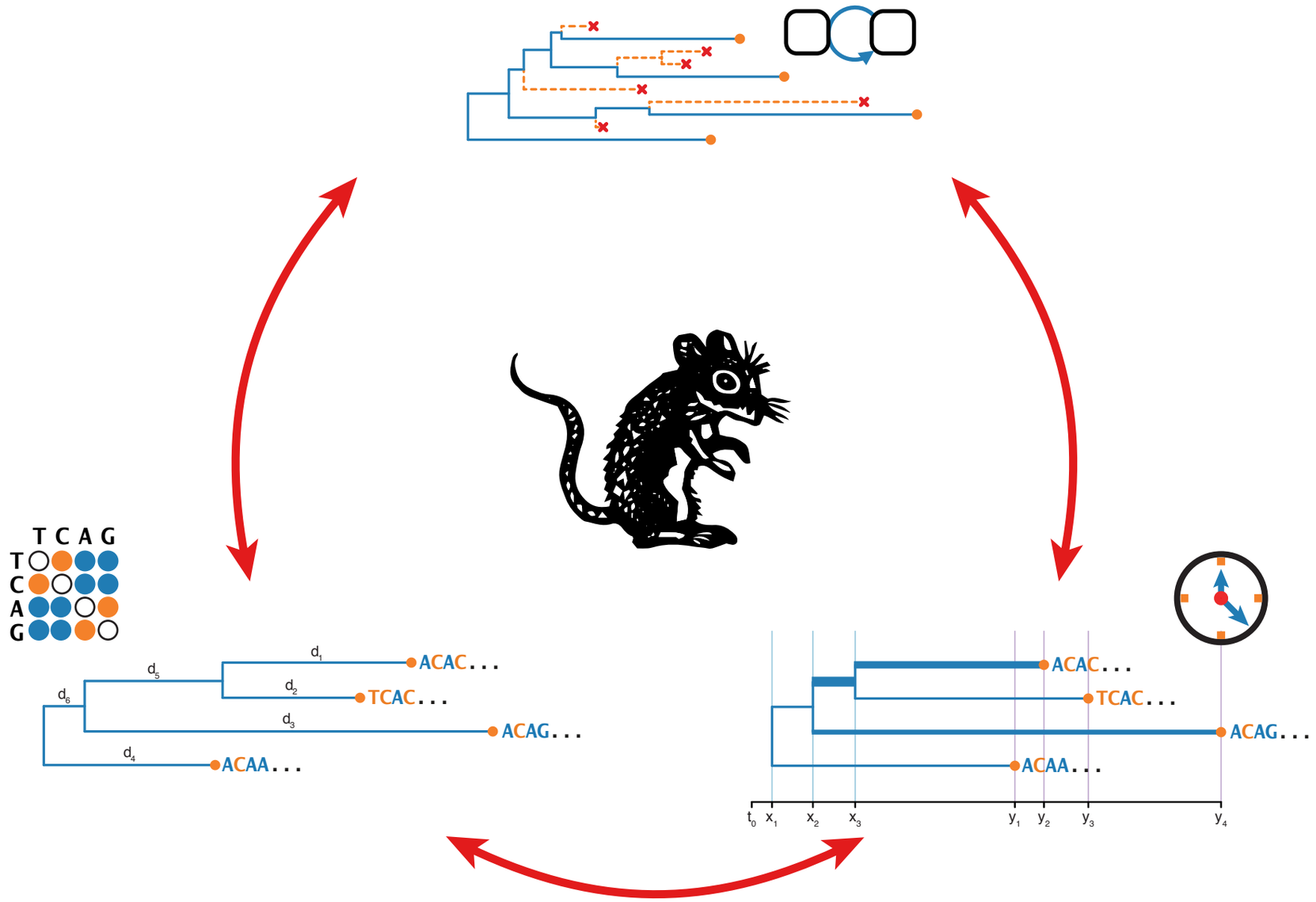
molecular clock model

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 contain sampling location or phenotypic trait data

What goes into a **BEAST** model?



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

genealogy

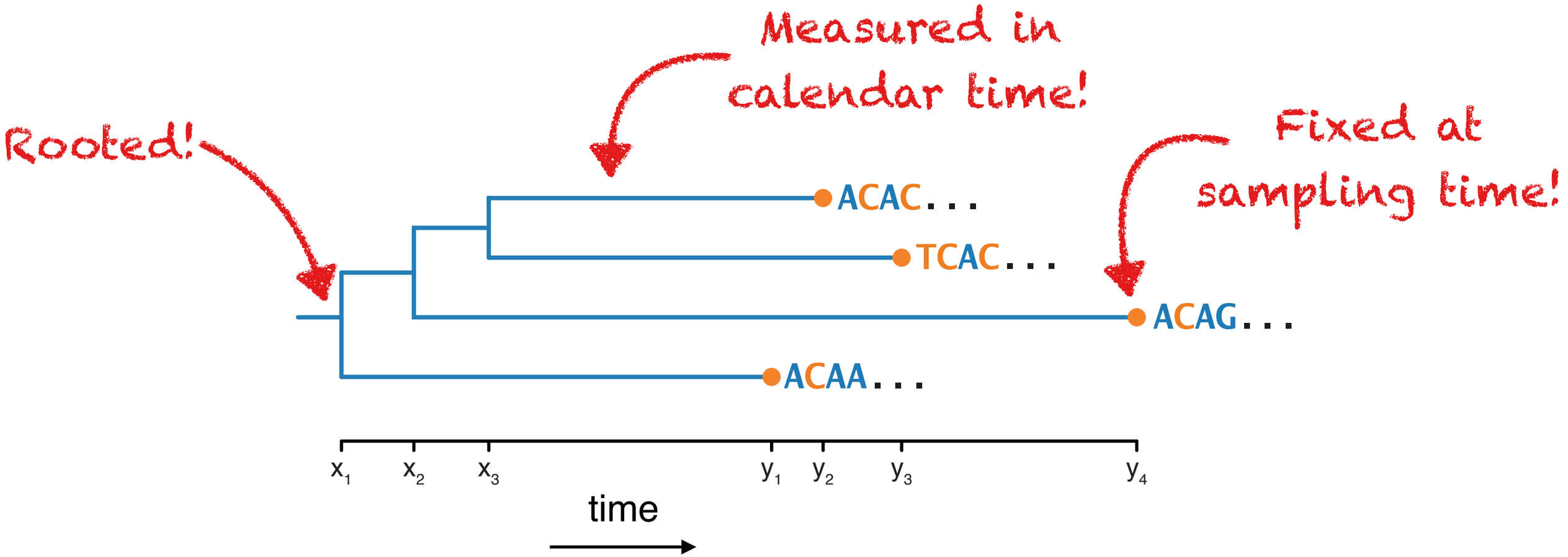
demographic model

site model

molecular clock model

The genealogy (tree)

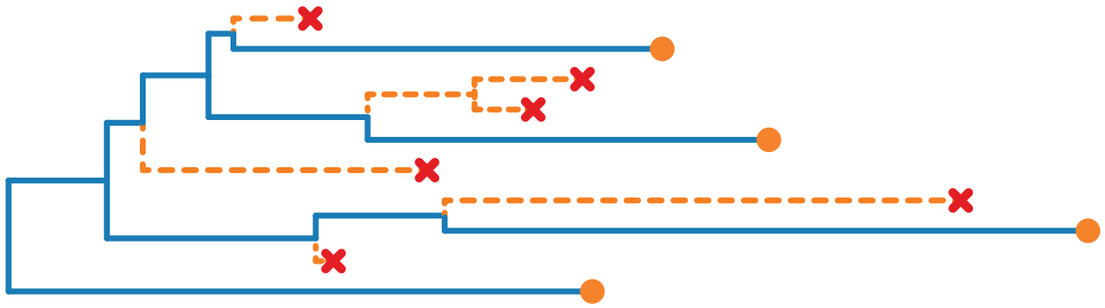
The fundamental genealogical structure
in **BEAST2** is the **rooted time-tree**

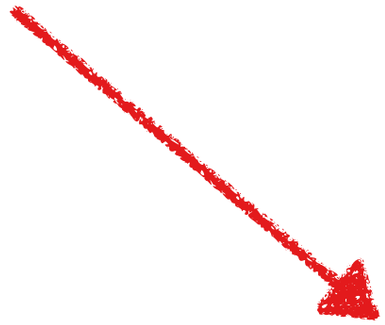


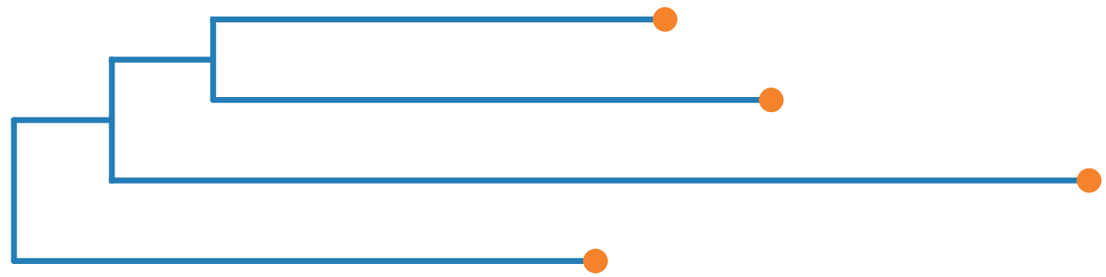
- This tree is a "sampled" or "reconstructed" tree
- Displays ancestral relationships between **sampled sequences** (individuals/taxa)

The genealogy (tree)

full tree



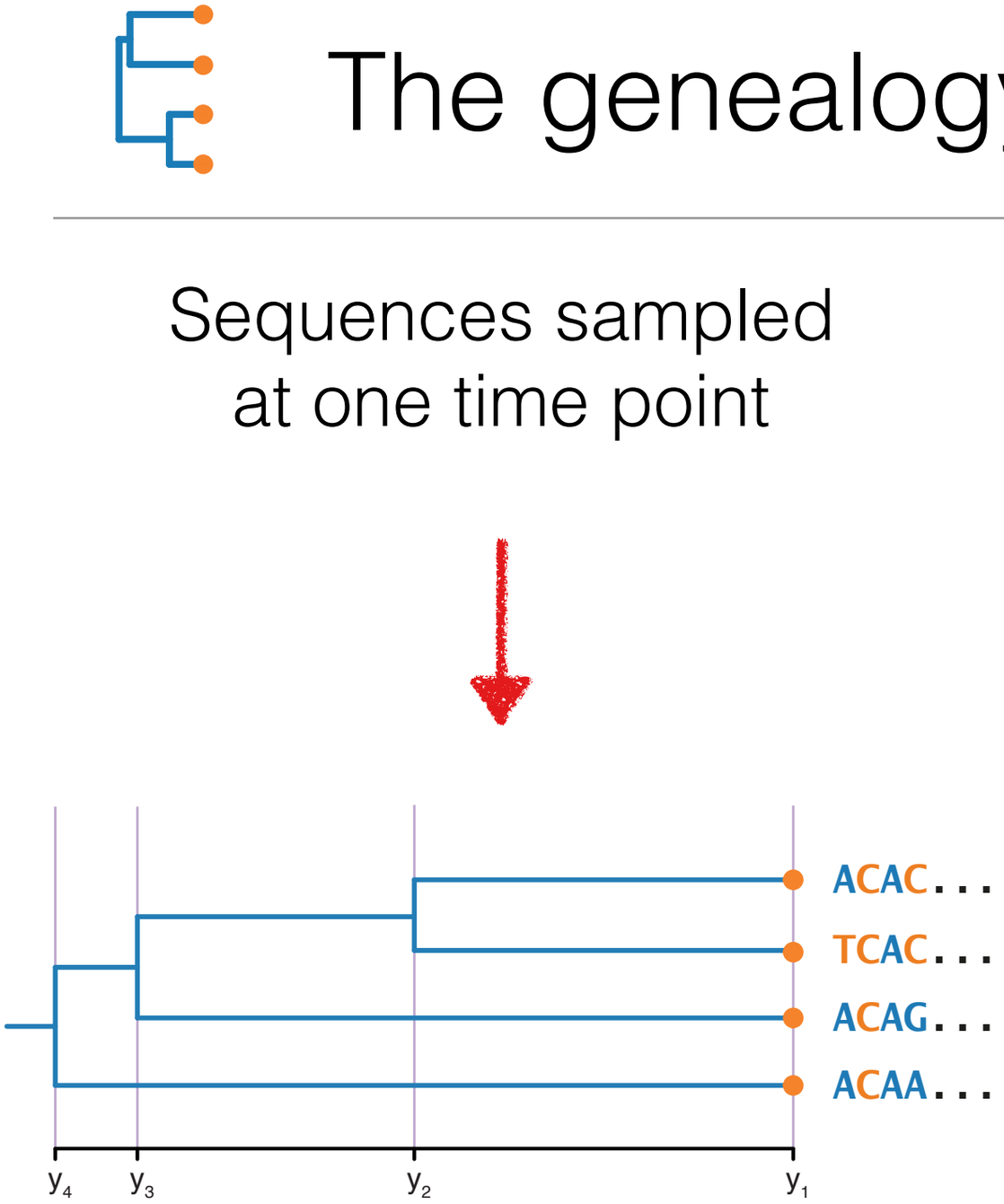




sampled/reconstructed tree

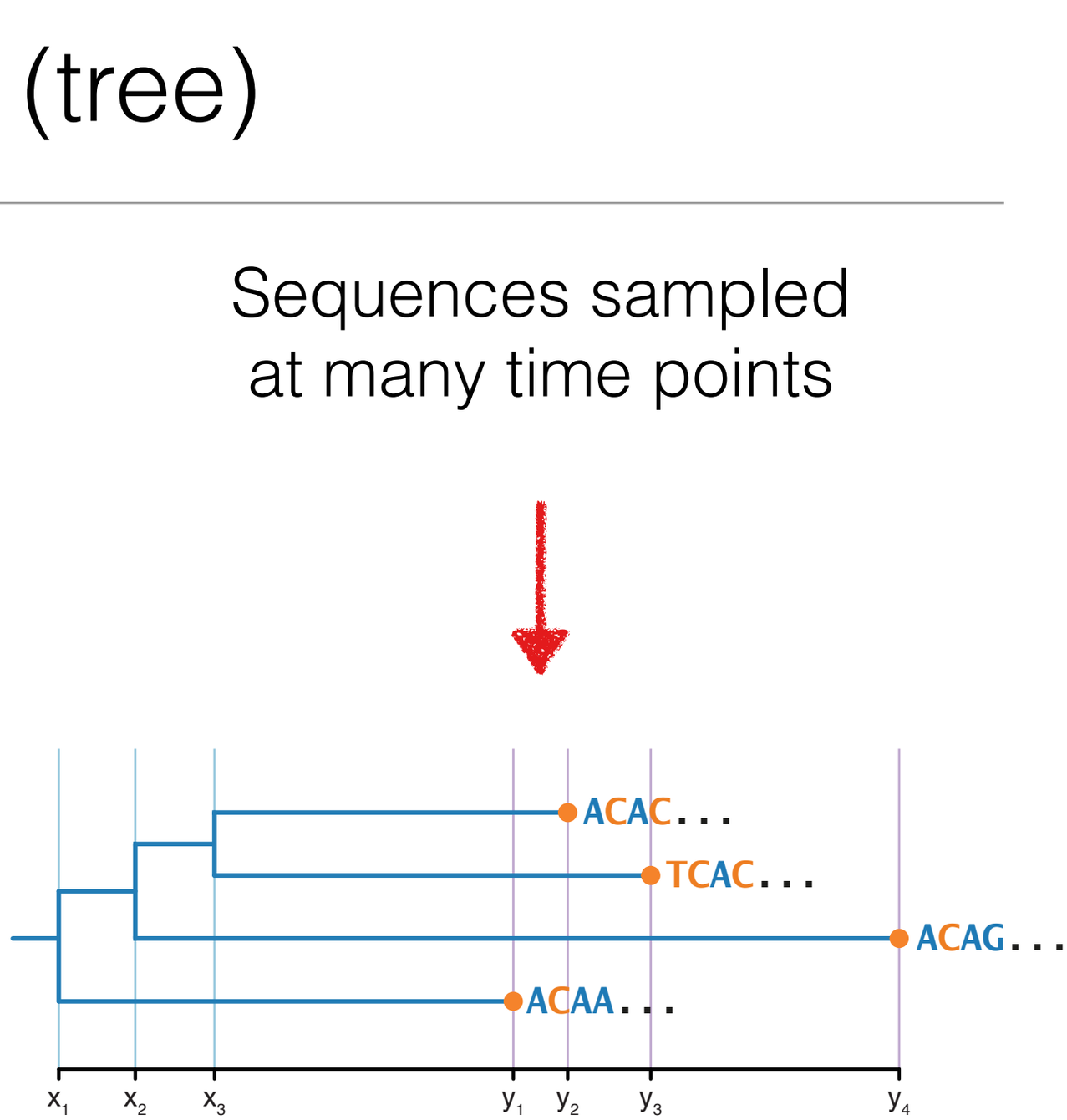
The genealogy (tree)

Sequences sampled at one time point



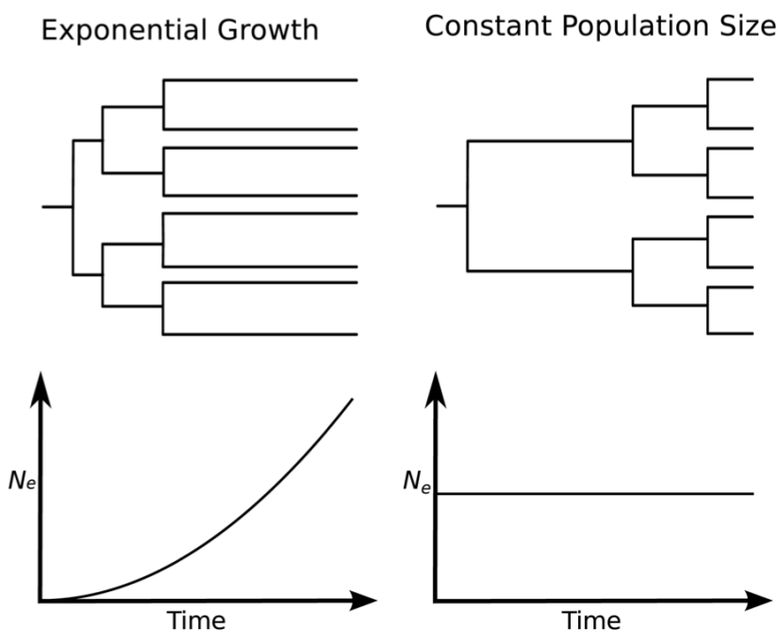
homochronous tree

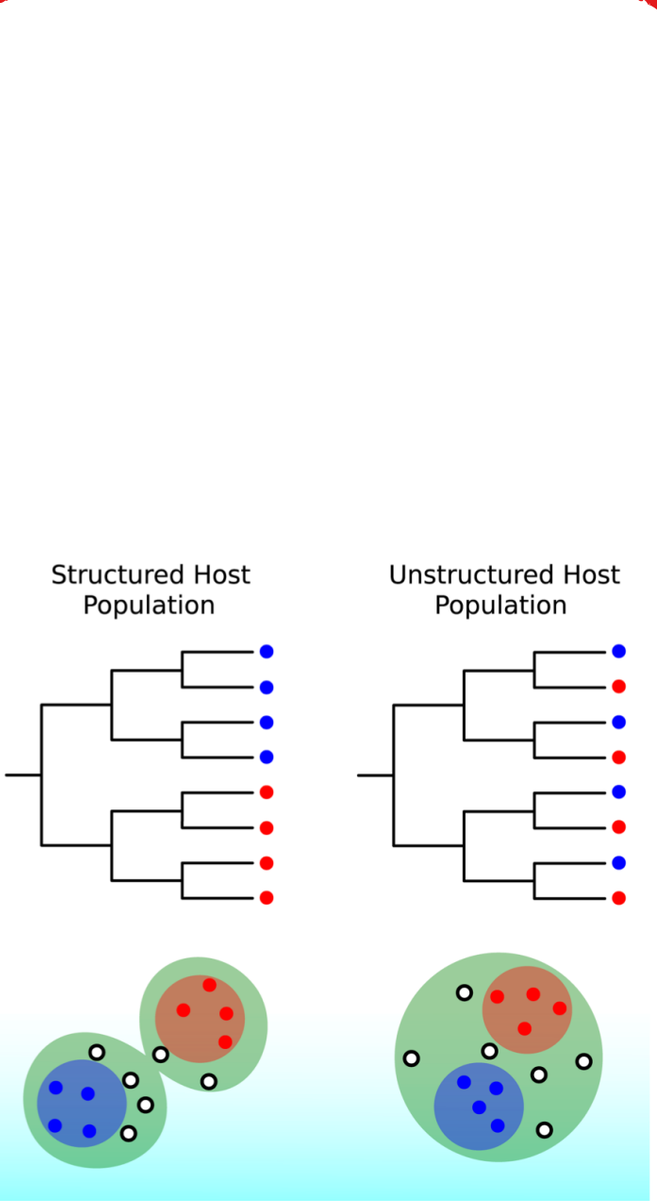
Sequences sampled at many time points

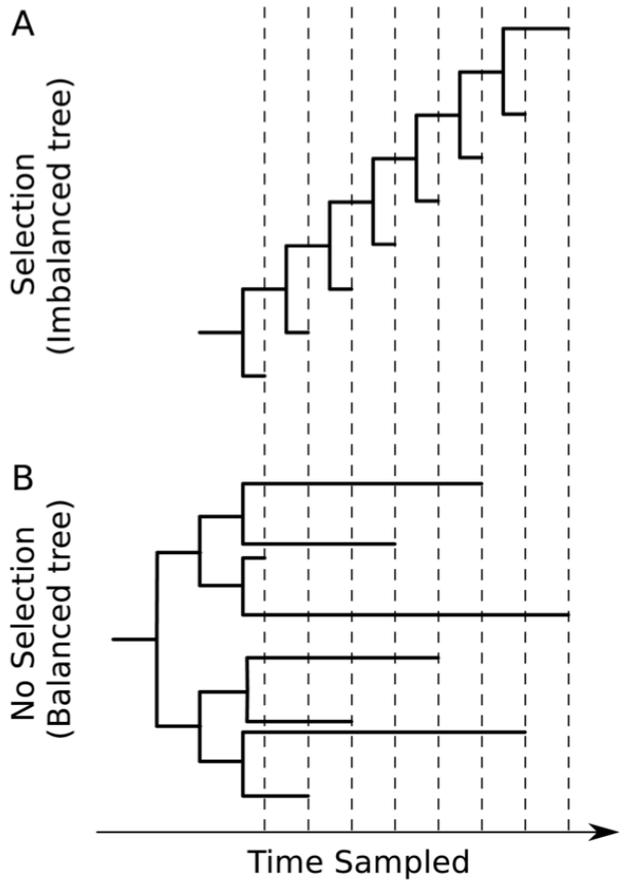


heterochronous tree

Different population dynamics generate trees that look different



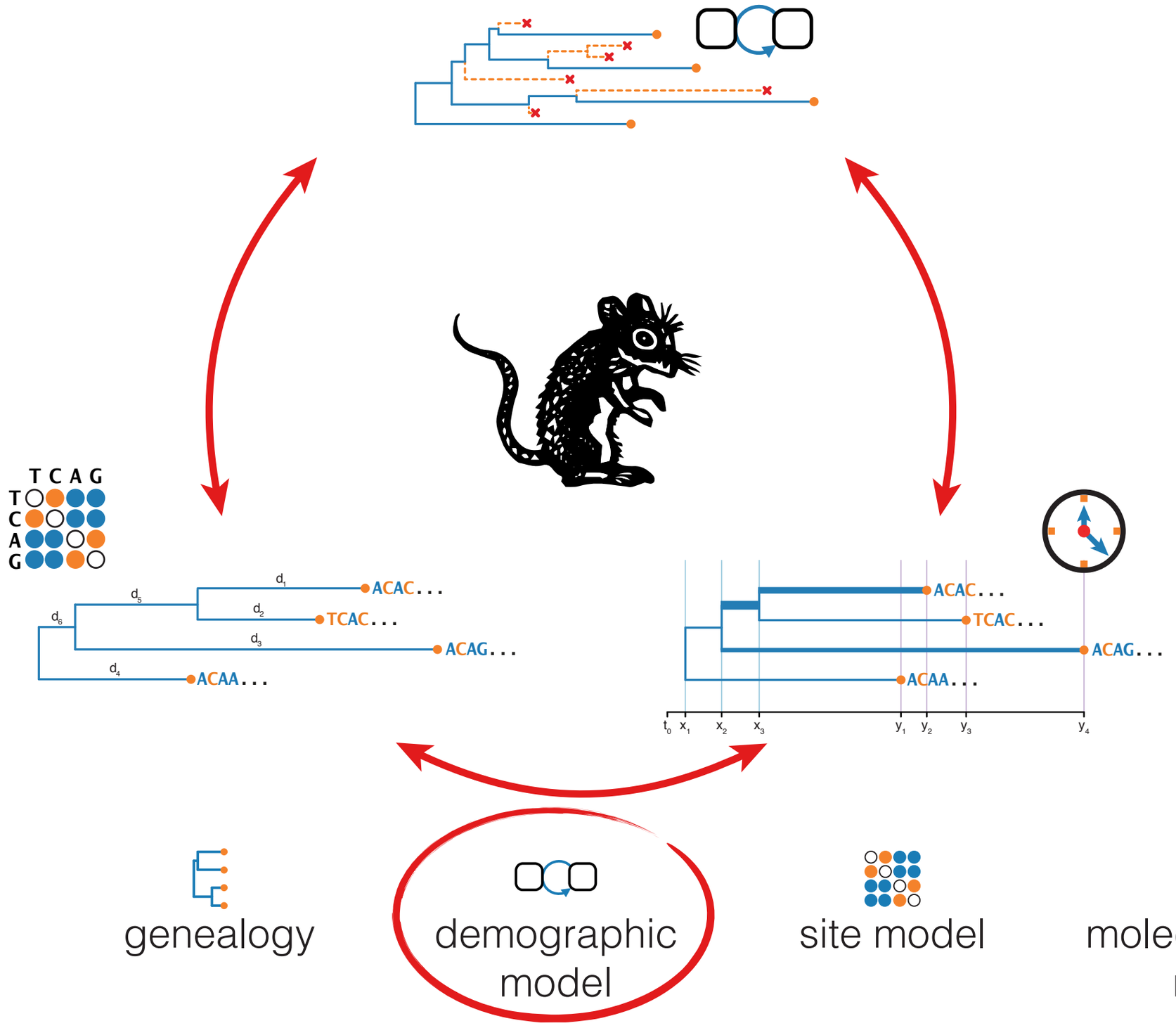




Volz *et al.* **PLoS Comp Biol 2013**

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

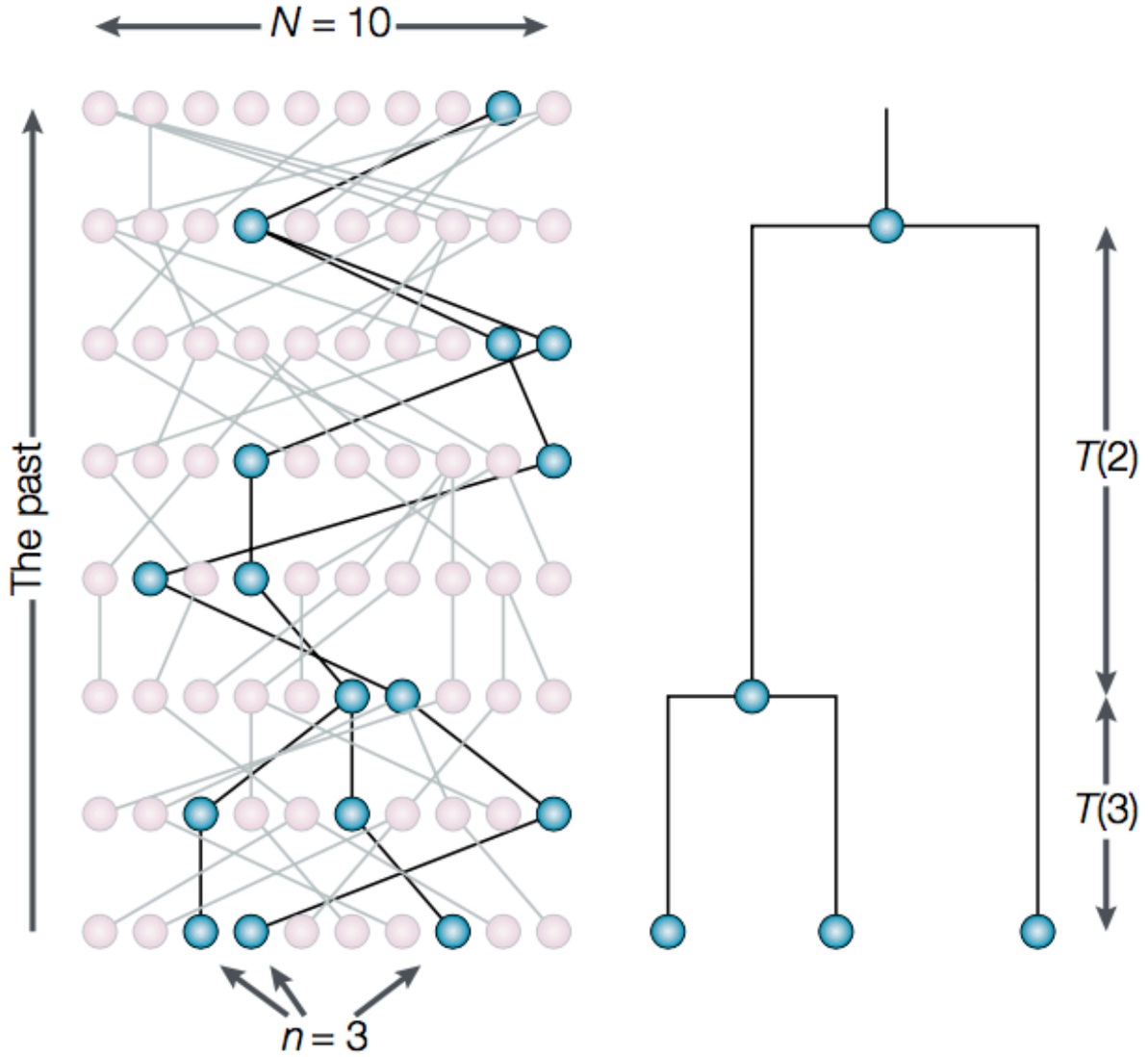
$P(\text{tree} \mid \text{demographic model})$

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

Coalescent models

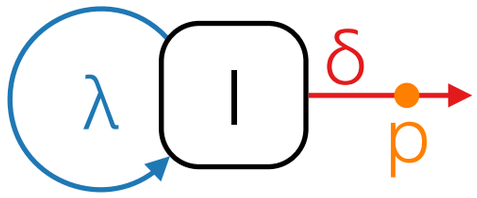
Full population **Sampled tree**



- Approximation to Wright-Fisher population dynamics (with large **N**)
- 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**
- Calculate the probability of observing a given **tree** for a particular **N**
  → estimate **N** (**Ne** in practice)
- Easy to extend to time changing **N(t)**

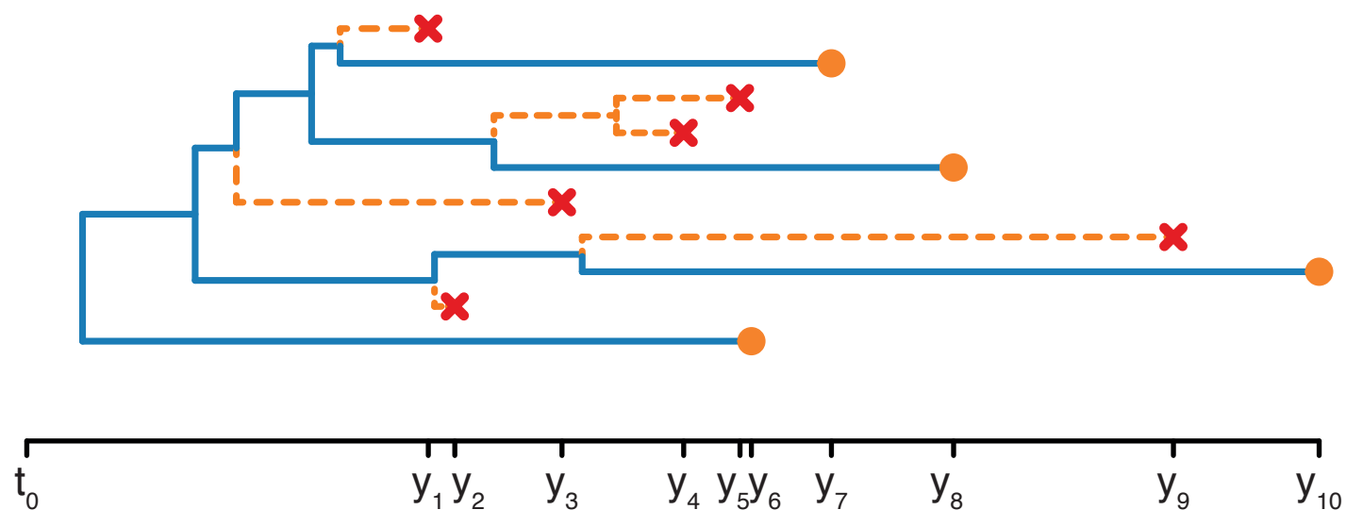
Rosenberg and Nordborg **Nature Reviews Genetics 2002**

Birth-death models



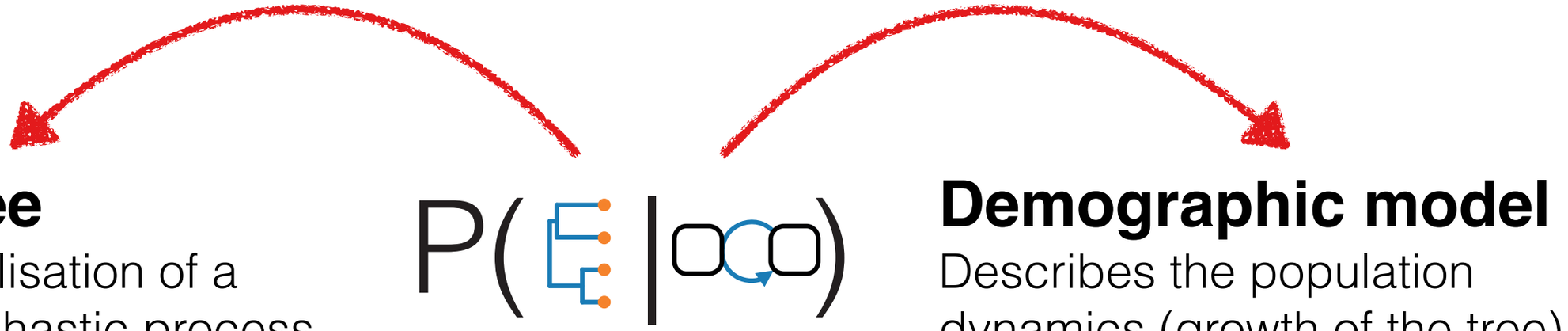
- λ — birth rate (lineages added to **full** tree)
- δ — death rate (lineages removed from the **full** tree)
- p — sampling probability (samples added to **sampled** tree)

- 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

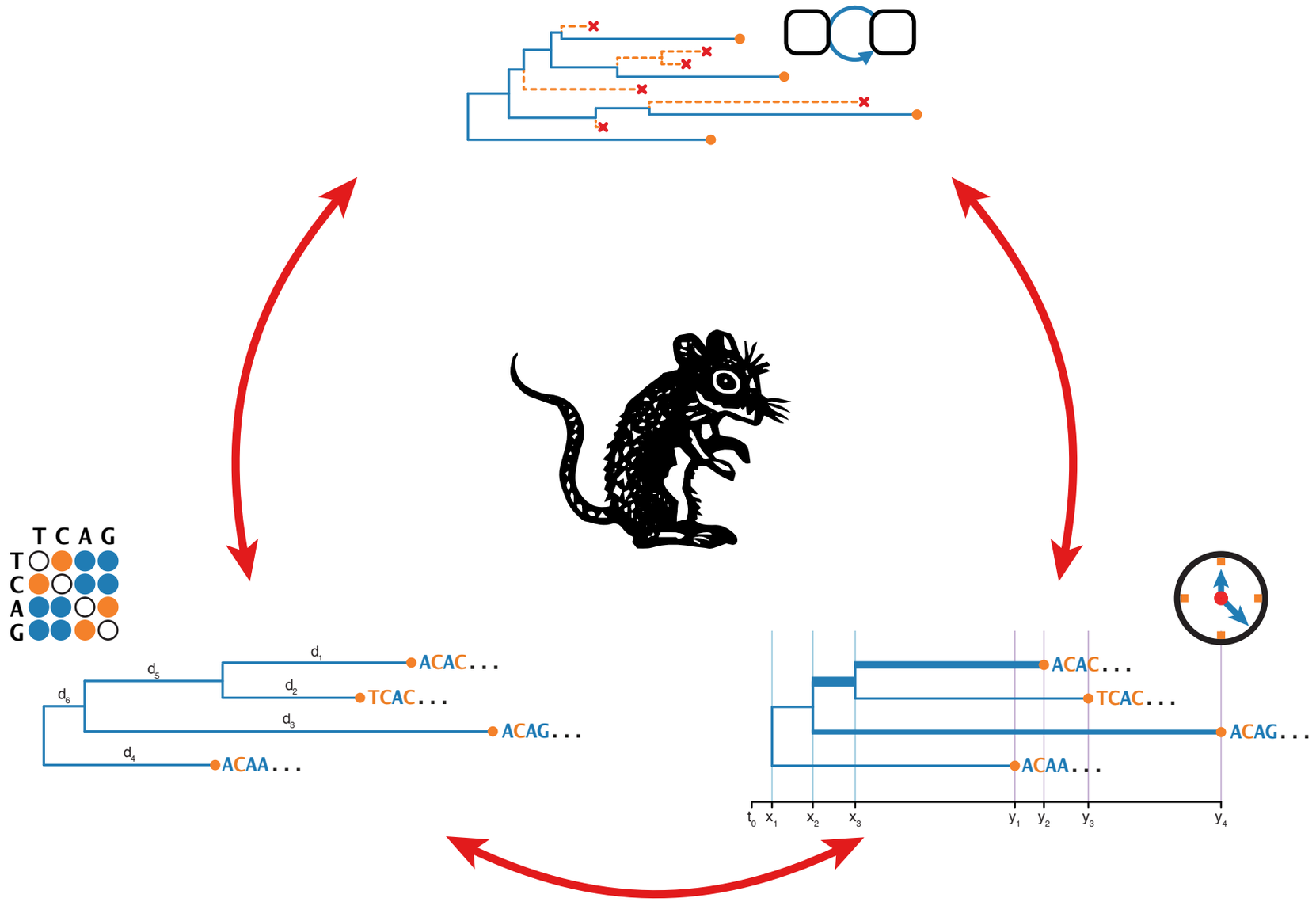
Demographic model



- **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?

What goes into a **BEAST** model?



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

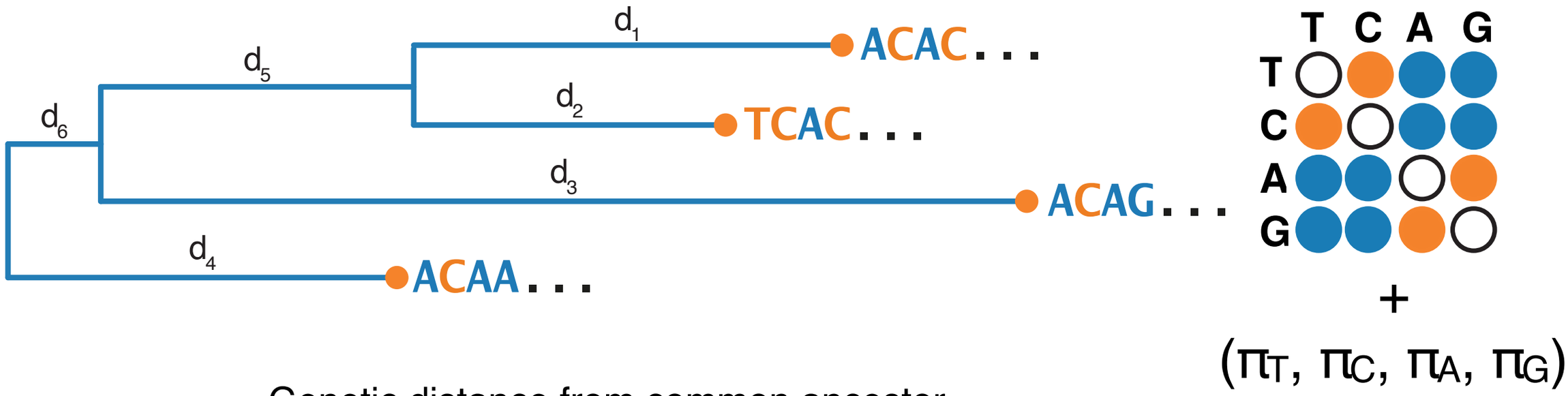
genealogy

demographic model

site model

molecular clock model

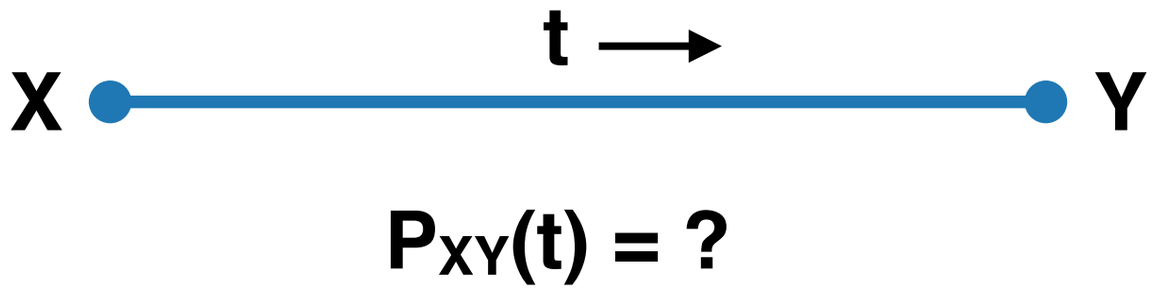
Site model



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

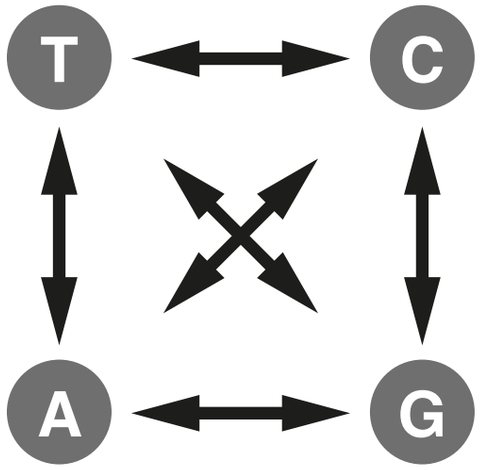
Assume:

- 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)

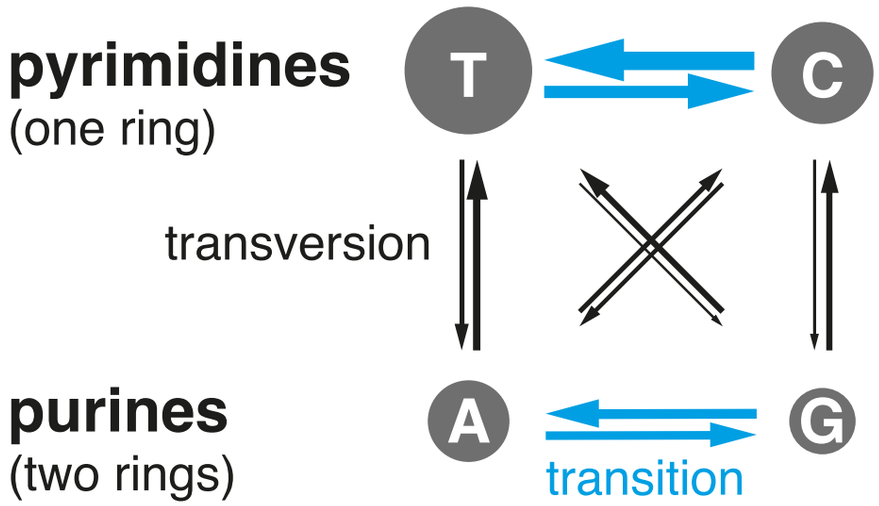


$$\begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} \cdot & \lambda & \lambda & \lambda \\ \lambda & \cdot & \lambda & \lambda \\ \lambda & \lambda & \cdot & \lambda \\ \lambda & \lambda & \lambda & \cdot \end{pmatrix} \end{array}$$

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

- Simplest model
- All rates and frequencies are equal!

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)

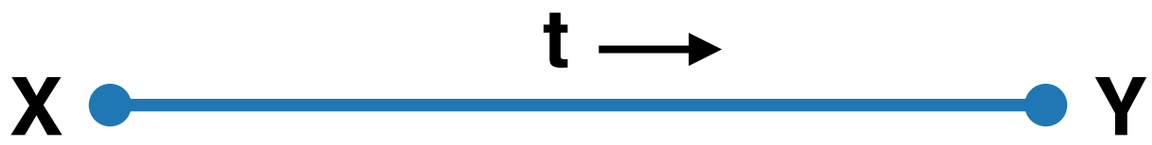
$$
\begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array}
\begin{array}{c}
\begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\
\begin{pmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot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

- 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)

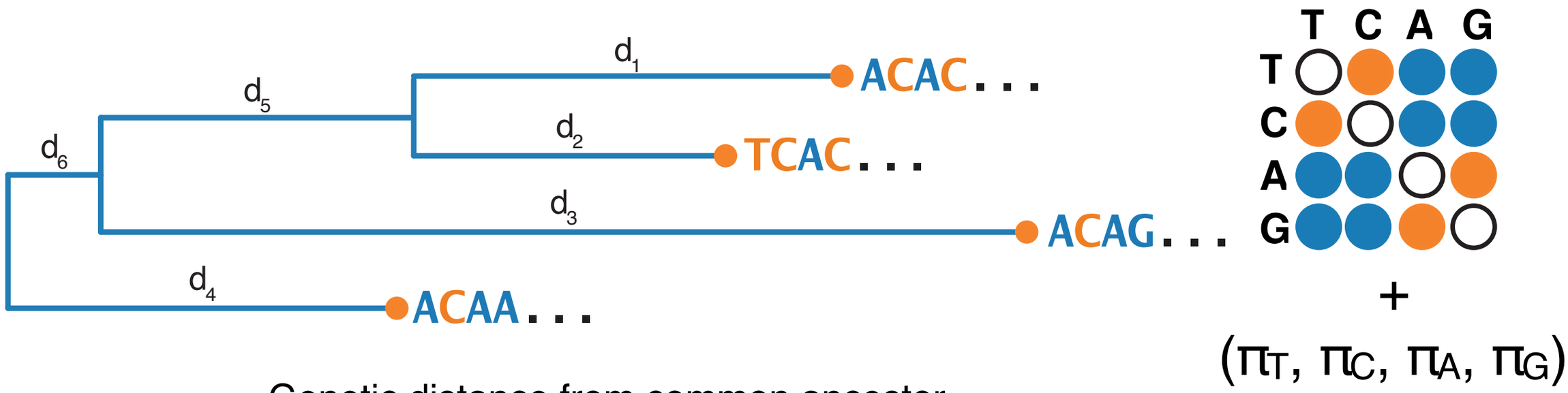


$$P_{XY}(t) = ?$$

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

⇒ distance is measured in expected number of substitutions per site

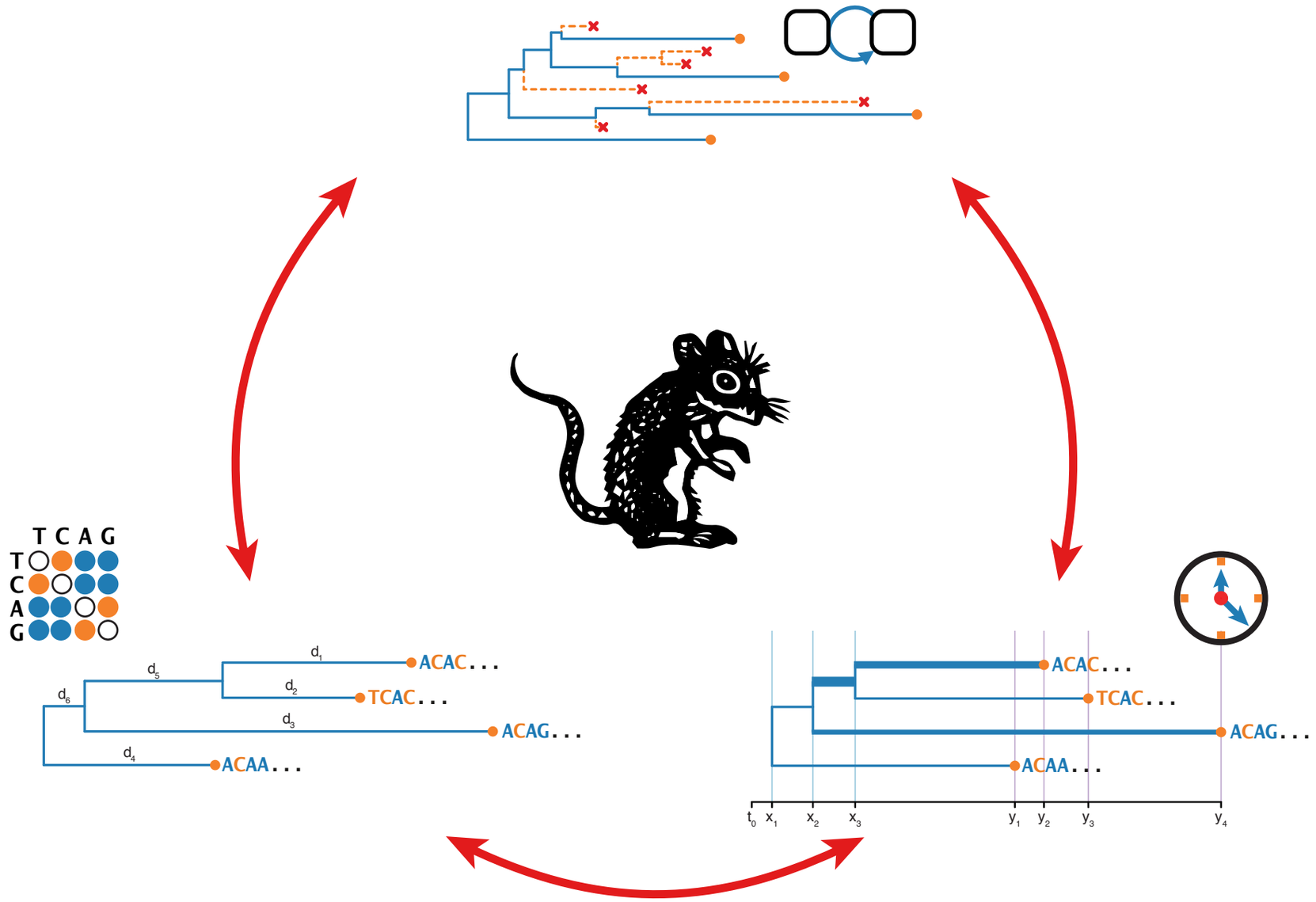
Site model



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
  - Discrete Γ model
  - Multiple partitions with different rates

What goes into a **BEAST** model?



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

genealogy

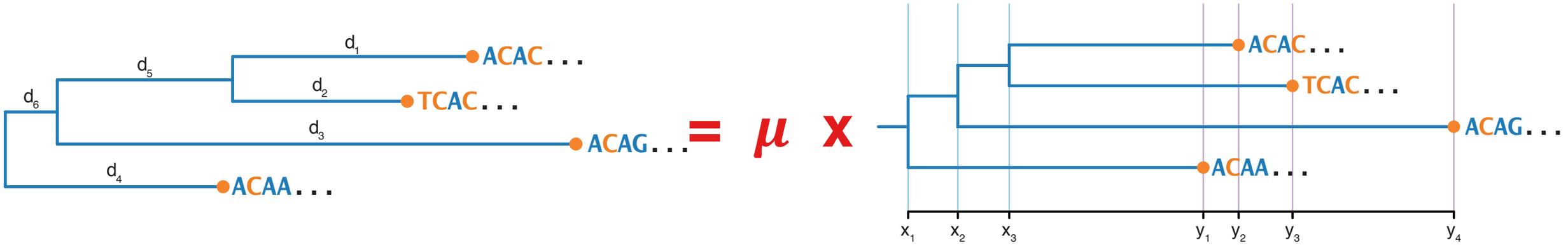
demographic model

site model

molecular clock model

Molecular clock model

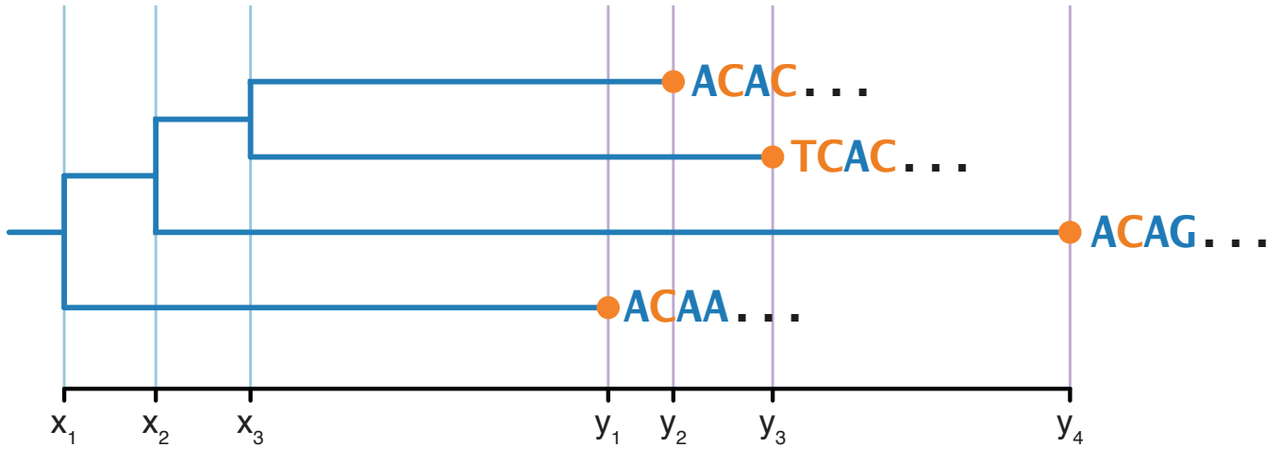
genetic distance tree (subst/site) **clock rate** (subst/site/year) **time tree** (years)



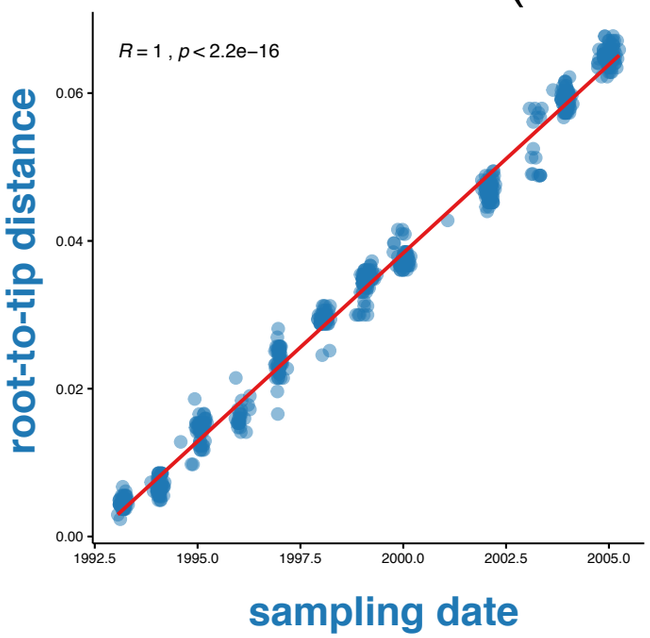
- Determines how quickly sequences are evolving along the tree
- **Genetic distance = Rate x Time**

Molecular clock model

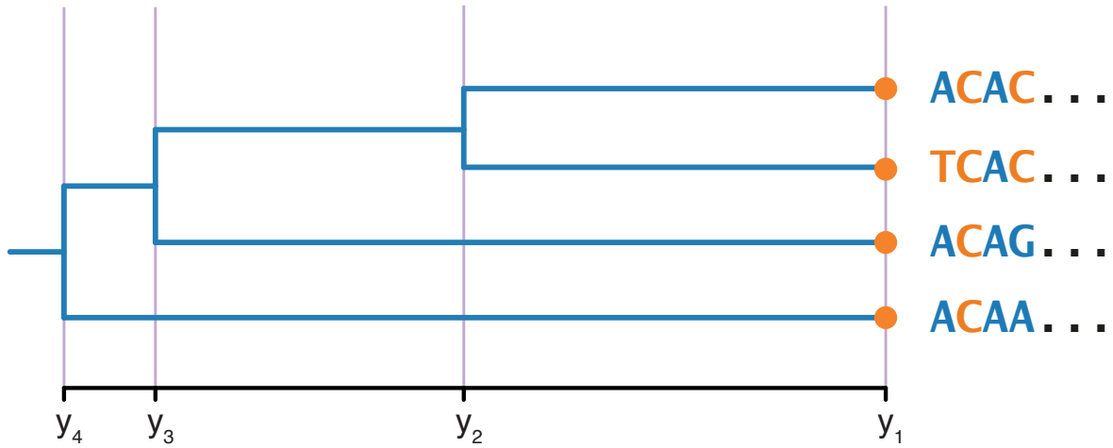
heterochronous tree



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

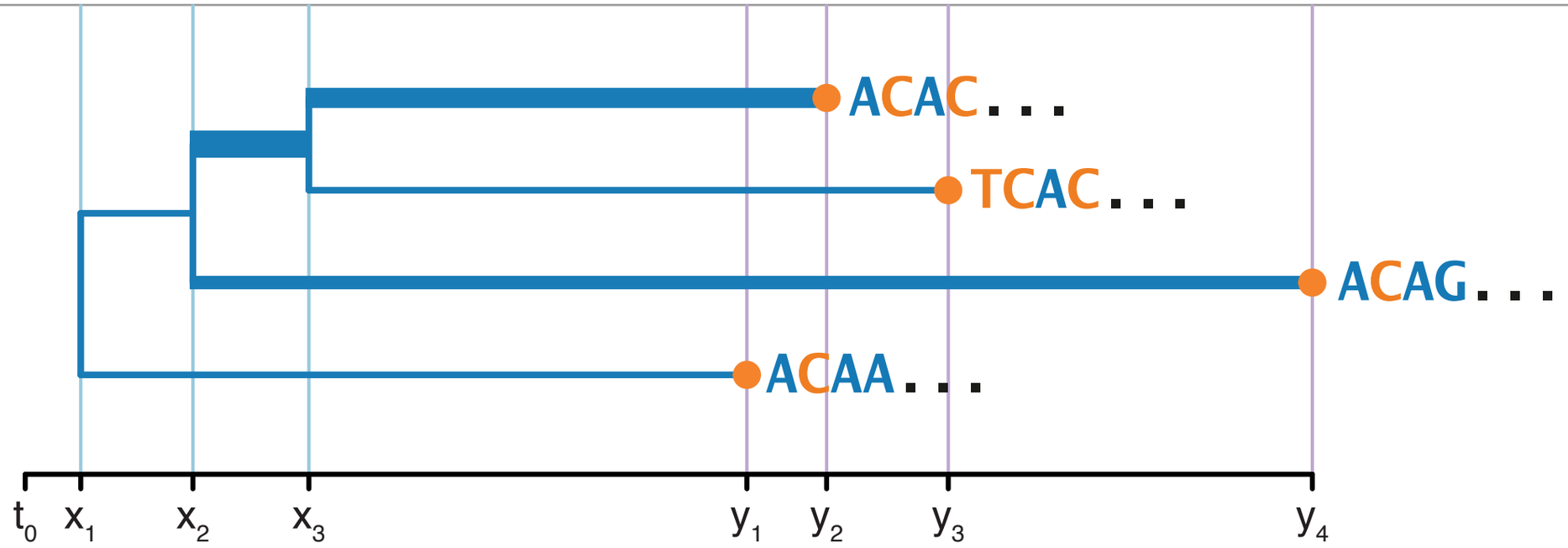


homochronous tree



- Need external information to calibrate the clock
- Fix clock rate
- Use node calibration

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

Likelihood

Prior

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

Posterior

Model evidence

Putting it all together

genetic sequences


genealogy

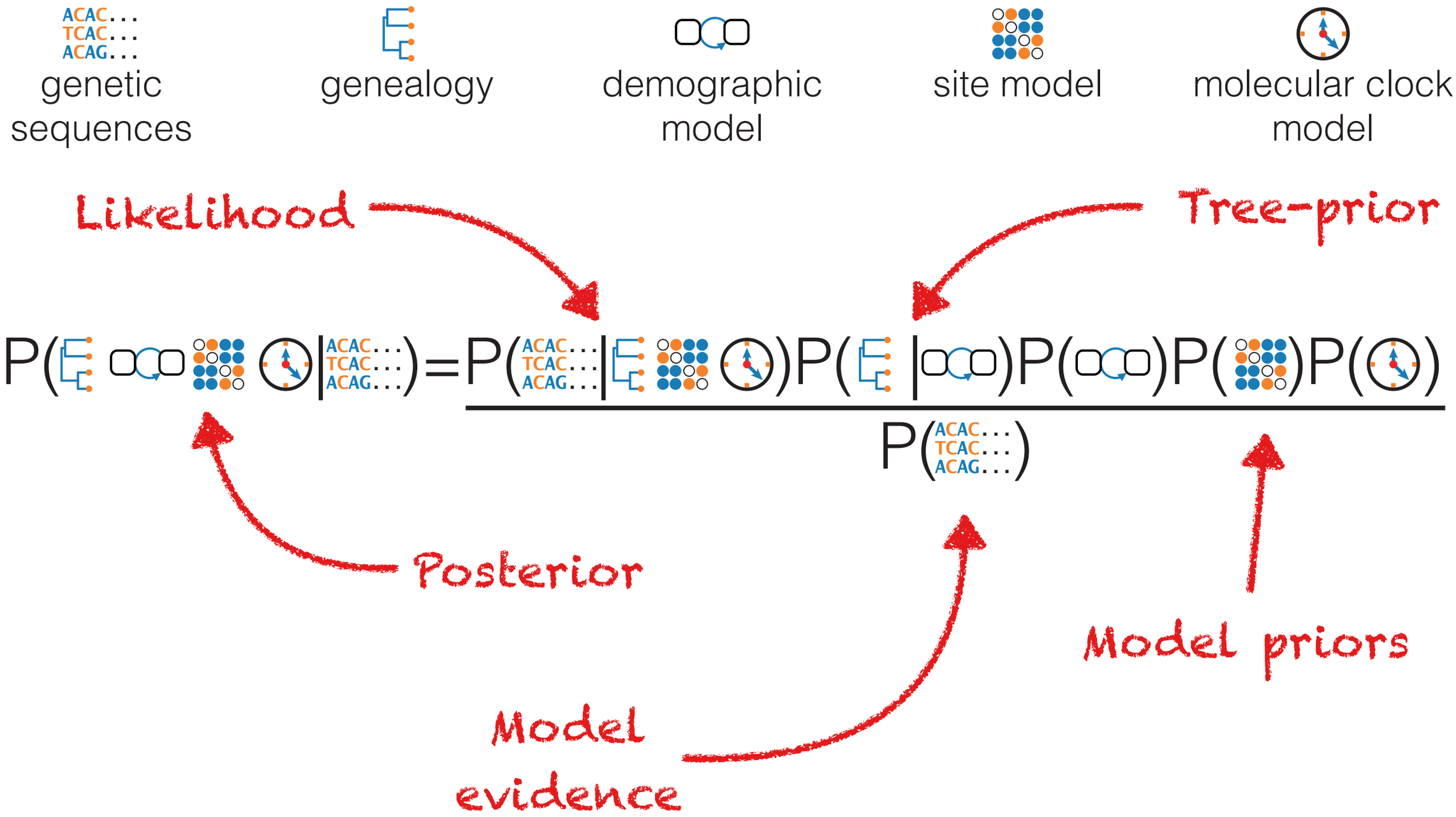

demographic model


site model


molecular clock model

Assume independence

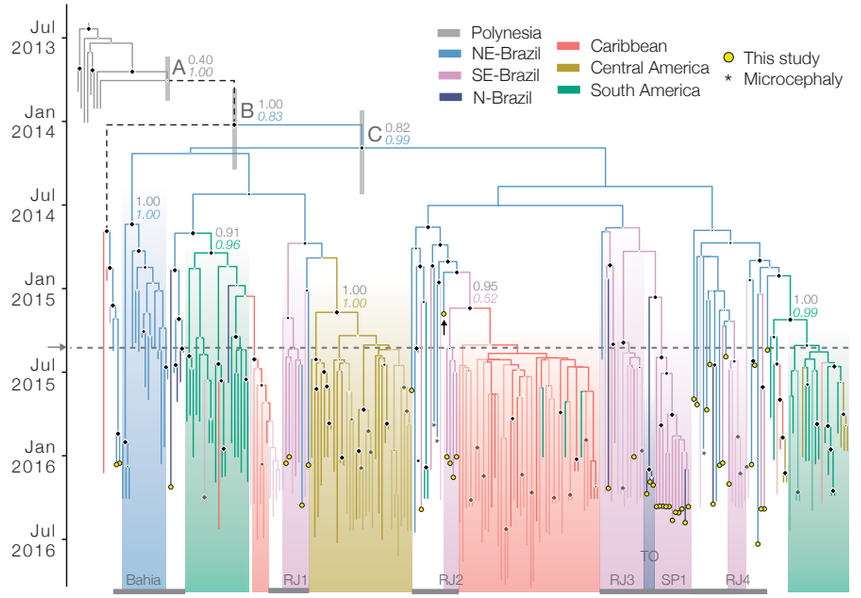
Posterior distribution in BEAST2



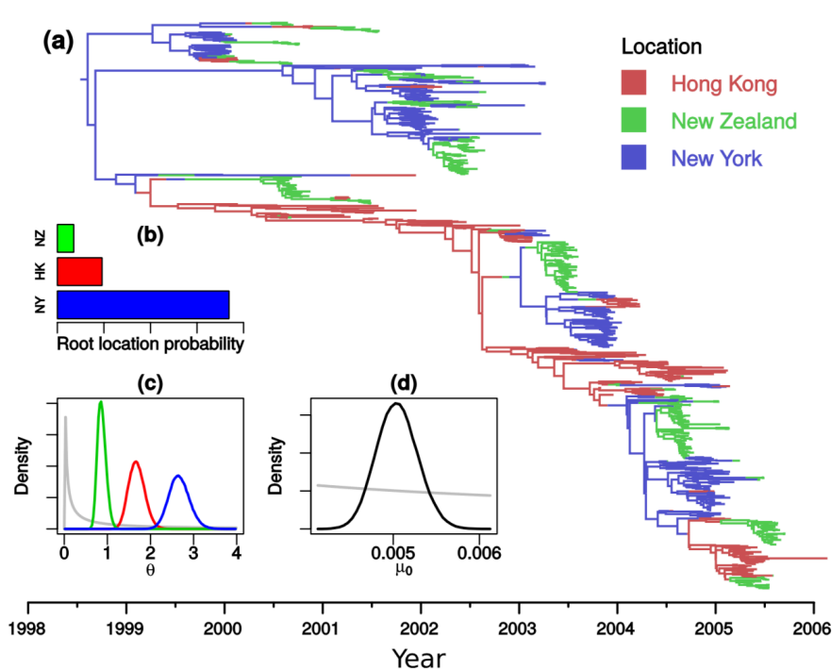
What about population structure?

Phylogeography, trait-evolution, host-types etc.

- "Mugration" model model it as an independent migration process
  - Discrete models are like a substitution model on locations/traits
  - Continuous models have a stochastic process (e.g. Brownian motion) evolving along a tree
- Migration and trait evolution is independent of the tree structure (conditions on the tree)
- Assumes the same sampling intensity in each location/type



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



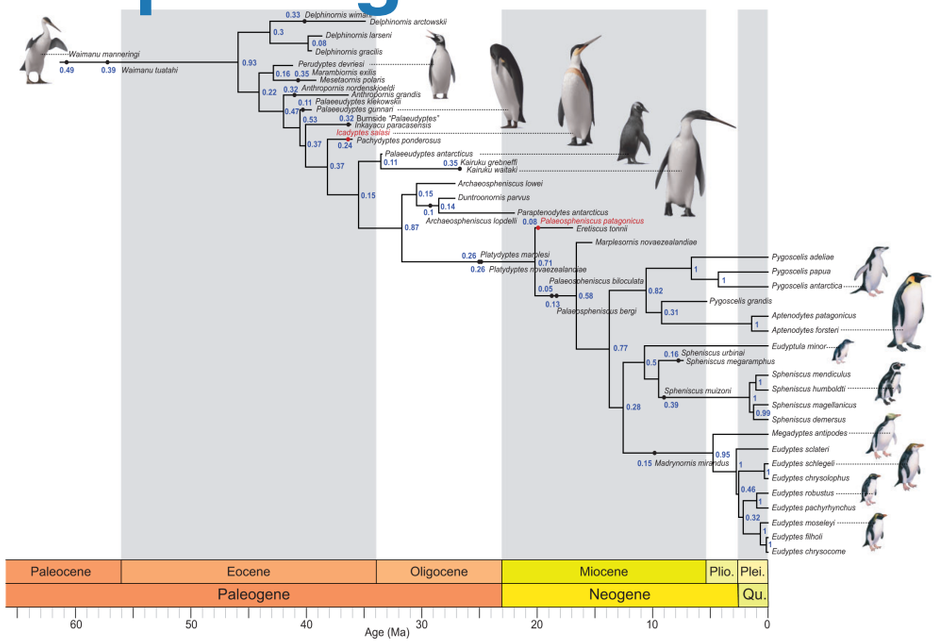
Vaughan *et al.* **Bioinformatics 2015**

- True structured models (structured coalescent, multi type birth-death model etc.) model it as part of the tree-prior
- Migration and trait evolution directly affect the structure of the tree
- More complex and computationally intensive
- No continuous analogue (yet…)

Exceptions I

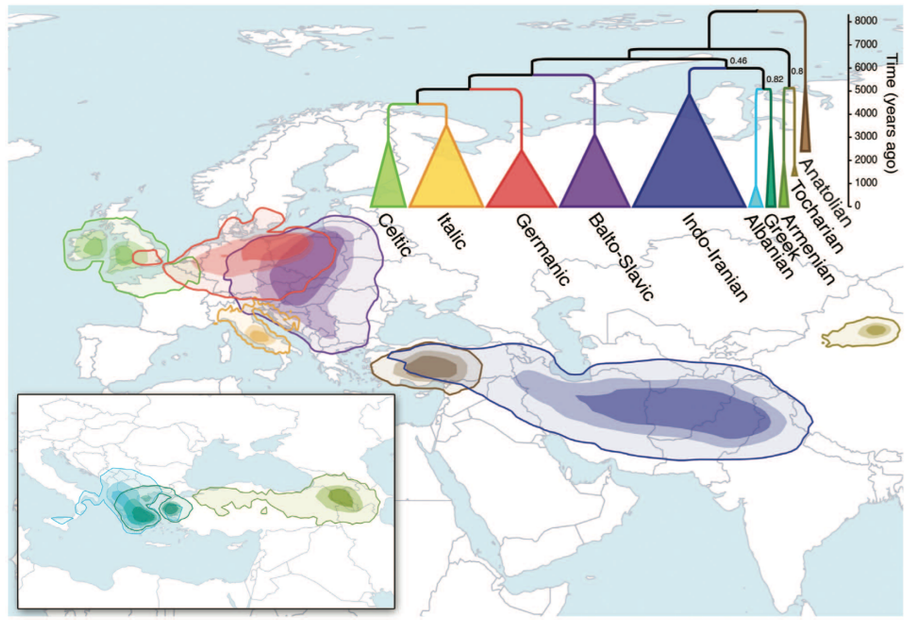
- Site models don’t have to be on nucleotides
- Could be on morphological traits, roots of words etc.

Morphological traits



Gavryushkina *et al.* **Systematic Biology 2017**

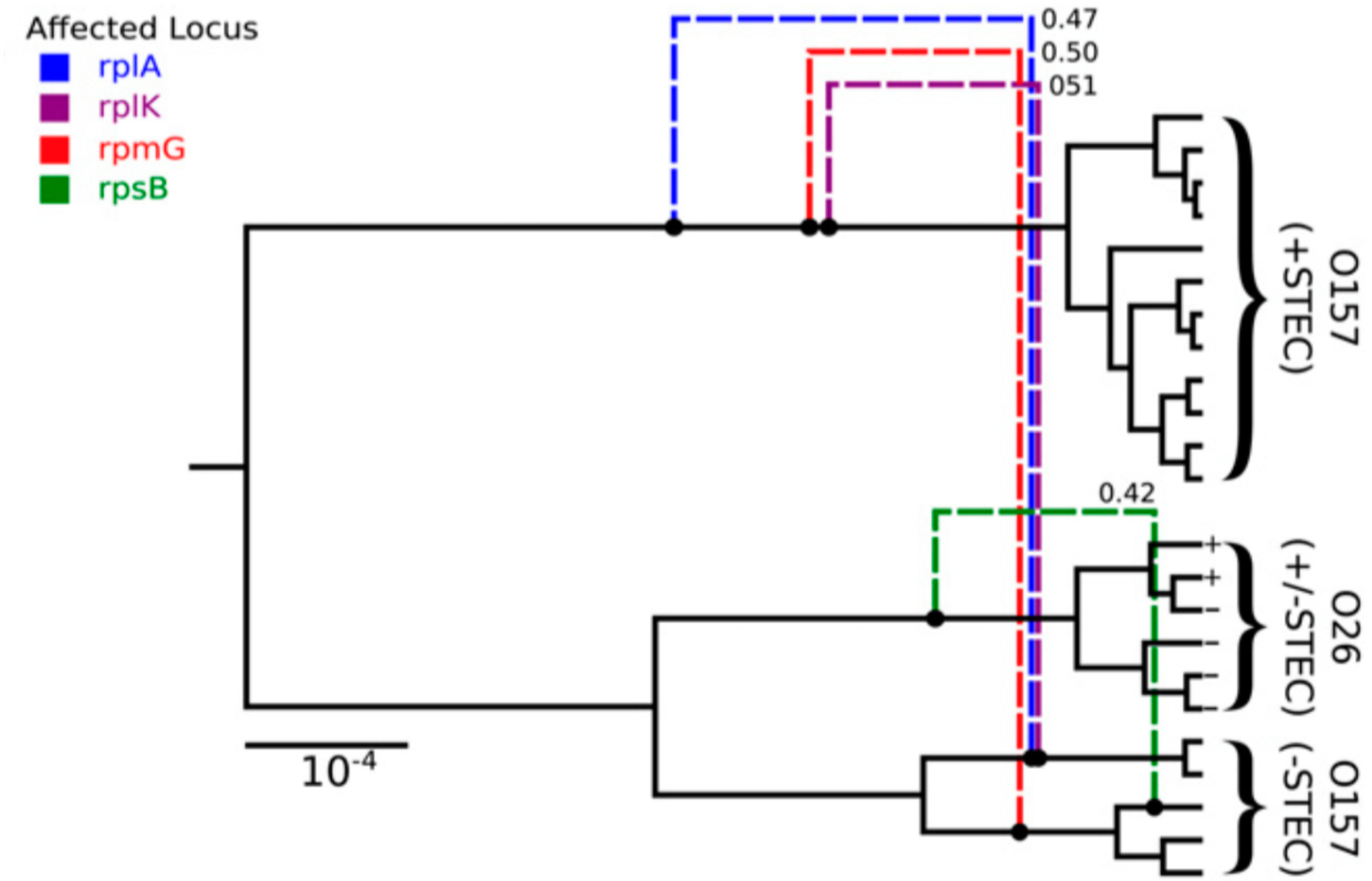
Roots of words



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

Exceptions II

BEAST2 doesn't always use trees…
e.g. ARG inference with BACTER

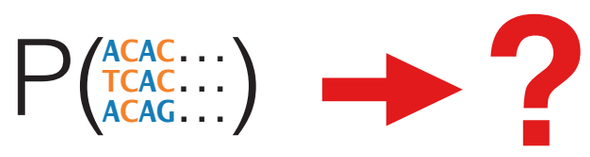


Vaughan *et al.* **Genetics 2017**

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!)

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

Operators

Target distribution

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

)
- MCMC steps 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 distribution

- This is the **posterior** in BEAST2: P()
- M[illegible] the
- [illegible]

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

Pro[illegible]

- 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

MCMC in practice

Before

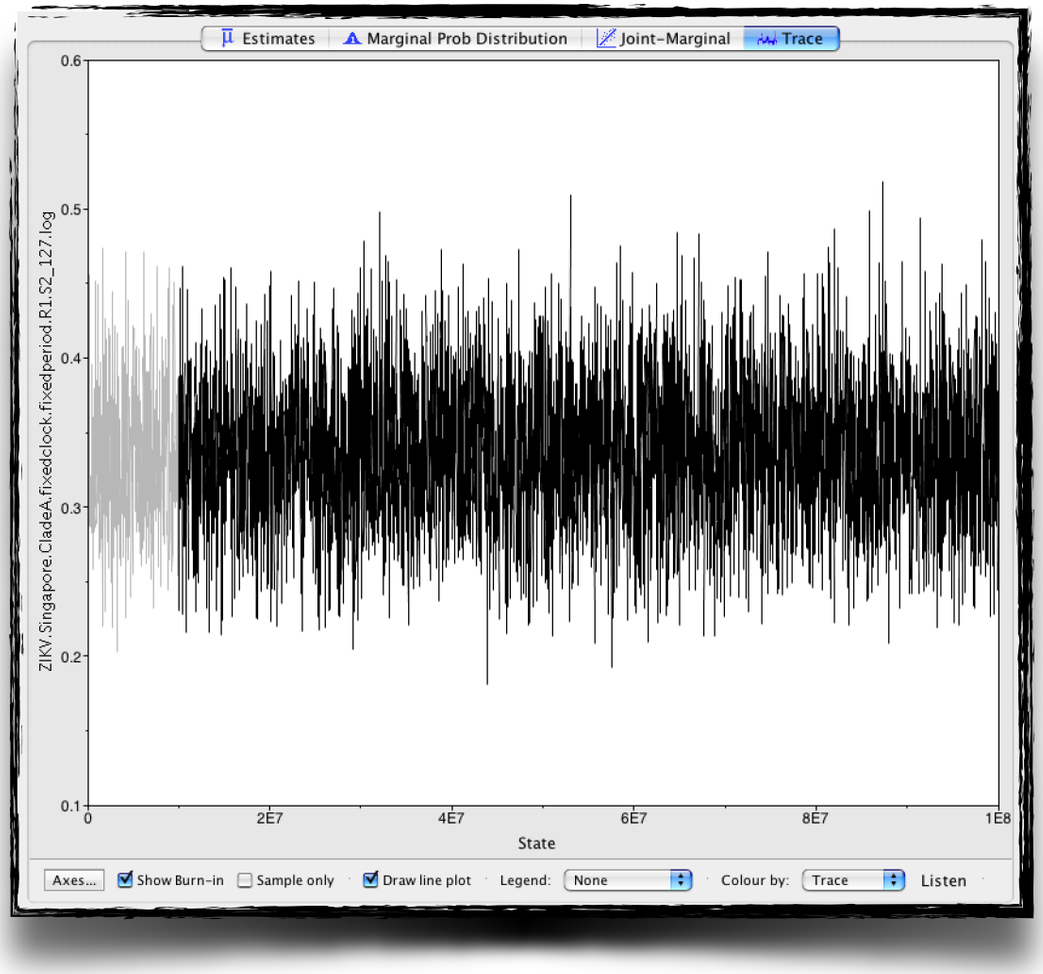
- Decide on the length of the chain
  (total number of steps to take)
- Decide on the sampling frequency
  (how often to record samples so
  that they are uncorrelated)

After

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

What we hope will happen

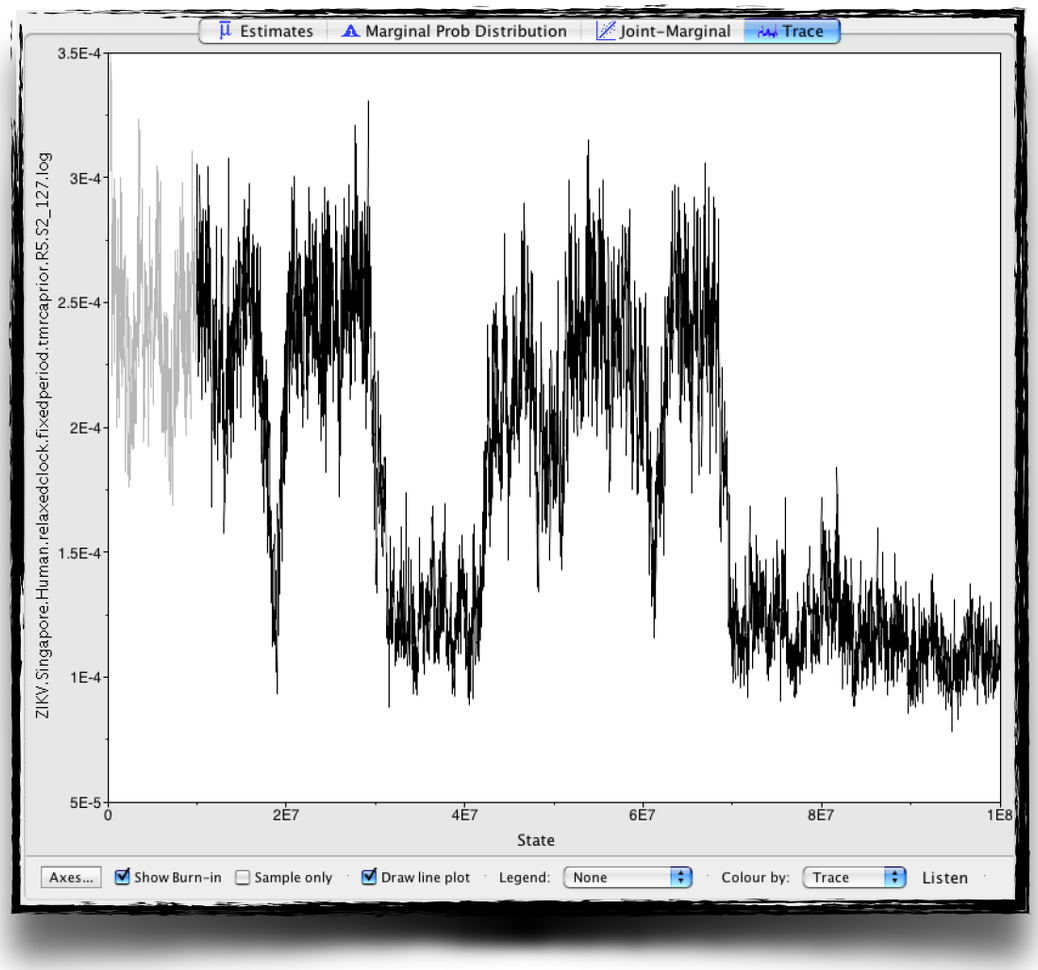
- The MCMC algorithm samples efficiently from high density areas of the posterior distribution
- We end up with a **good** approximation of the posterior distribution in **finite** time
- Appearance of white noise
- Everything is awesome!



Mixing well! 😄

Questions to ask…

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



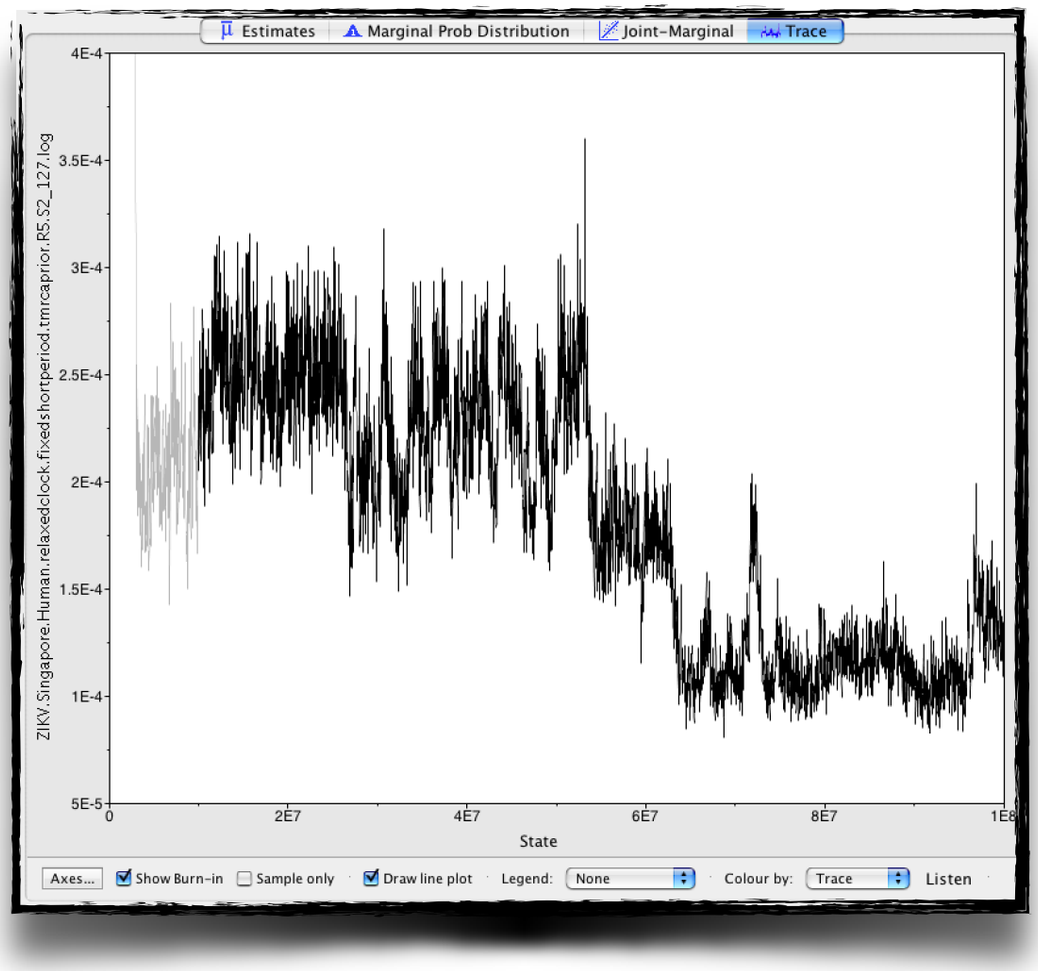
Not mixing! 😨

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?

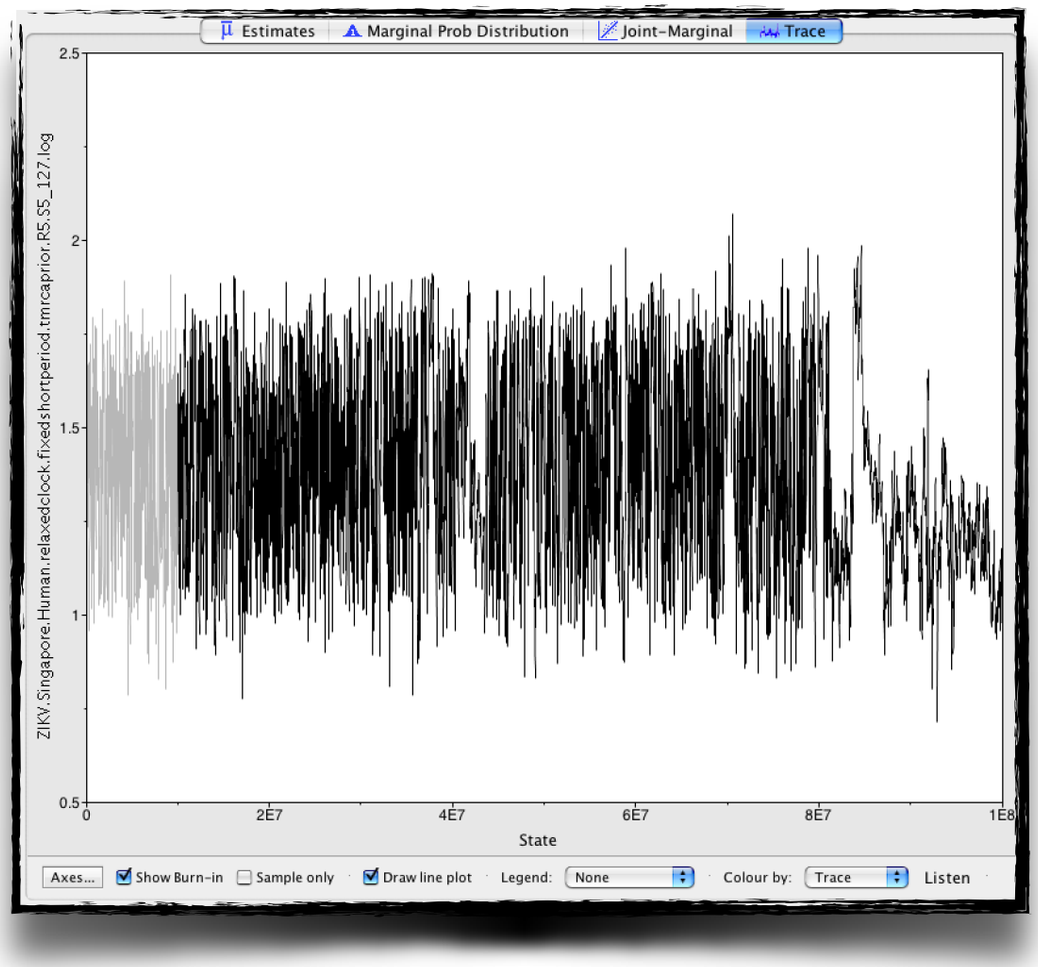


Not converged! 😓

Solution: Run for longer

Questions to ask…

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



Still not converged! 😵

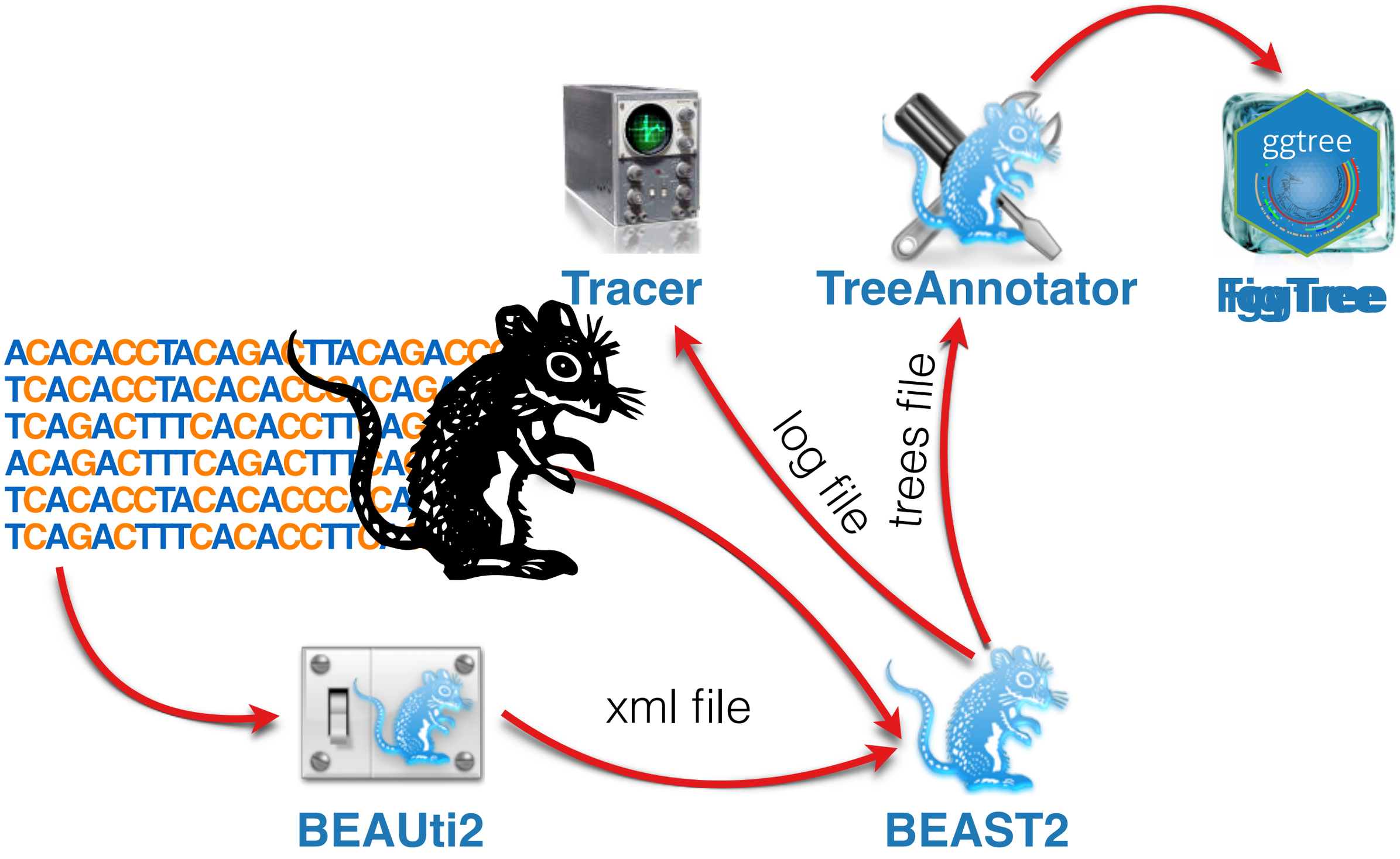
Solution

- 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

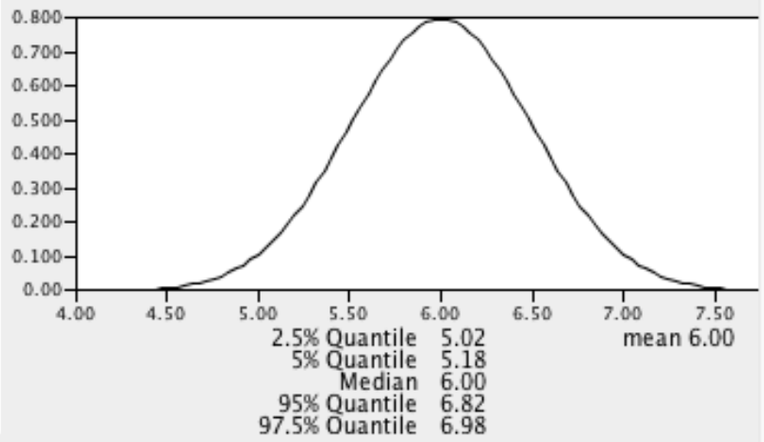
BEAUti 2: Standard /Users/louis/Documents/Taming_the_BEAST/Tutorials-Git/Introduction-to-BEAST2/xml/Primates.xml

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 1.8.4/1.10 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

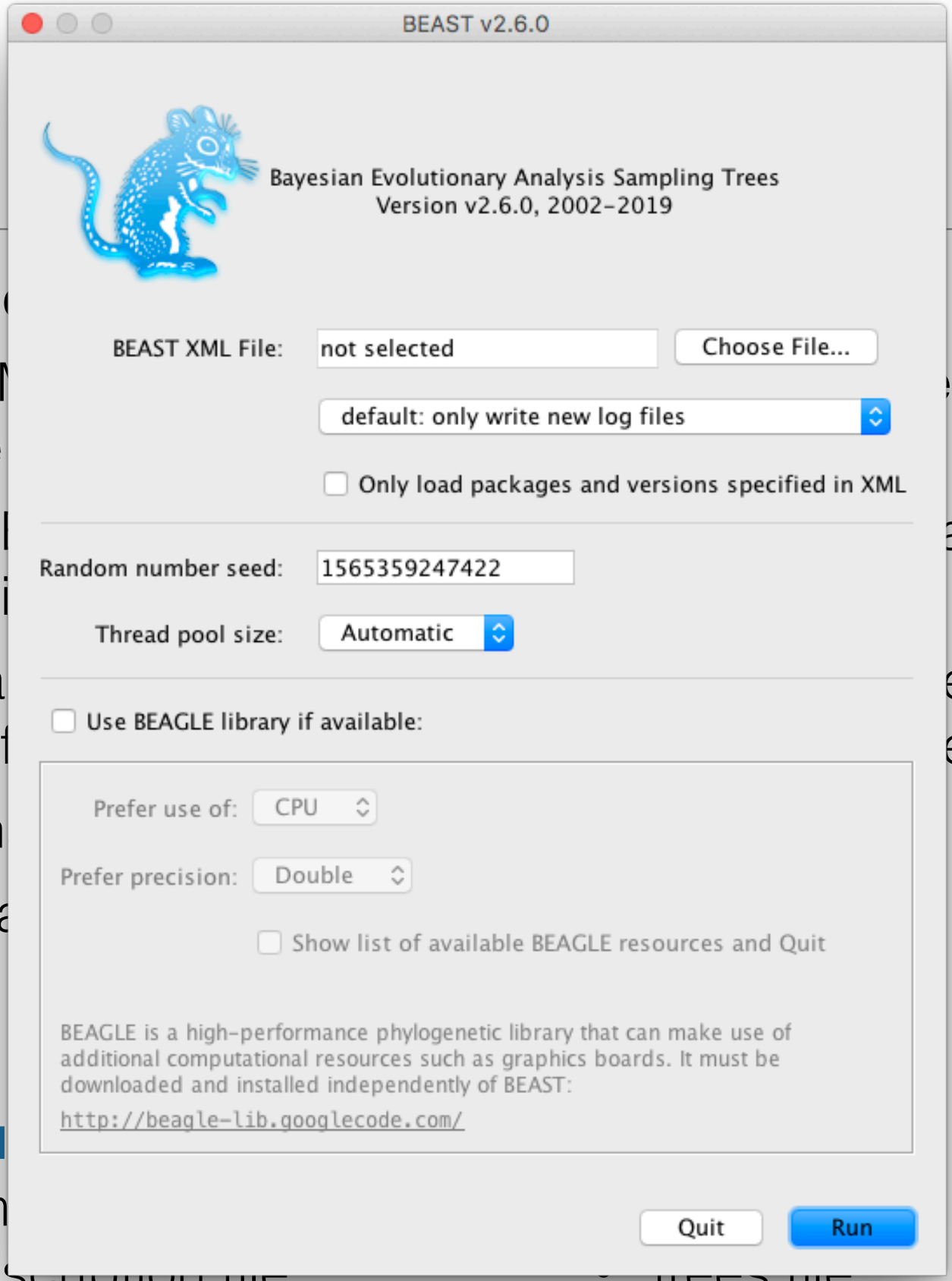
Outputs:

- log file
- trees file
- state file

BEAST2

(http://beast2.org)

- **B**ayesian
- Performs M … lected sequence
- Similar to I … ate and generally i
- BEAST2 a … e much of the same f … e
- BEAST2 h … to extend
- GUI interfa … line (e.g. on a



Inpu

- xm … 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

Tracer

Trace Files:

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

+ − Reload

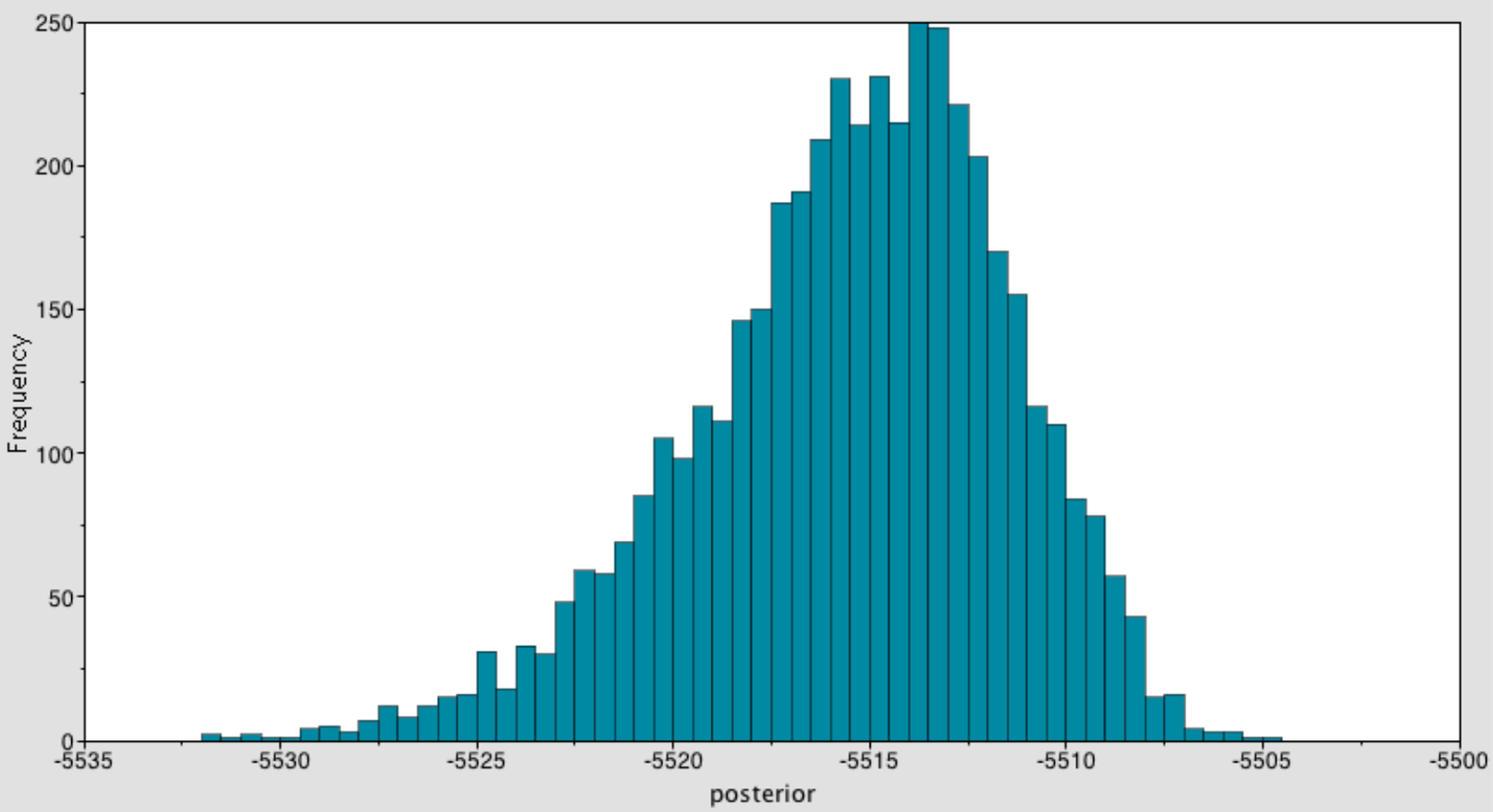
Traces:

Statistic	Mean	ESS	...
posterior	-5515....	261	R
likelihood	-5442....	335	R
prior	-73.432	113	R
treeLikelihood.1stpos	-1383....	502	R
treeLikelihood.2ndpos	-952.555	321	R
treeLikelihood.3rdpos	-2148....	202	R
treeLikelihood.noncod...	-957.464	192	R
TreeHeight	85.181	235	R
mutationRate.1stpos	0.451	120	R
mutationRate.2ndpos	0.179	124	R
mutationRate.3rdpos	2.955	96	R
mutationRate.noncoding	0.34	180	R
gammaShape.1stpos	0.477	105	R
gammaShape.2ndpos	0.553	76	R
gammaShape.3rdpos	2.998	98	R
gammaShape.noncodi...	0.249	88	R
kappa.1stpos	6.424	86	R
kappa.2ndpos	8.681	79	R
kappa.3rdpos	29.35	42	R
kappa.noncoding	13.619	69	R
CalibratedYuleModel	-47.452	320	R
birthRateY	2.547E-2	731	R
logP(mrca(human-chi...	-0.744	4203	R
mrcatime(human-chi...	5.95	2567	R
clockRate	1.165E-2	391	R

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

Estimates | Marginal Density | Joint-Marginal | Trace

Summary Statistic	posterior
mean	-5515.4884
stderr of mean	0.2487
stdev	4.0176
variance	16.141
median	-5515.0556
value range	[-5531.8494, -5504.8478]
geometric mean	n/a
95% HPD interval	[-5523.3461, -5508.1268]
auto-correlation time (ACT)	3449.4716
effective sample size (ESS)	261
number of samples	4501



Setup... Bins: 50

- Demographic reconstructions

Tracer

Trace Files:

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

+ − Reload

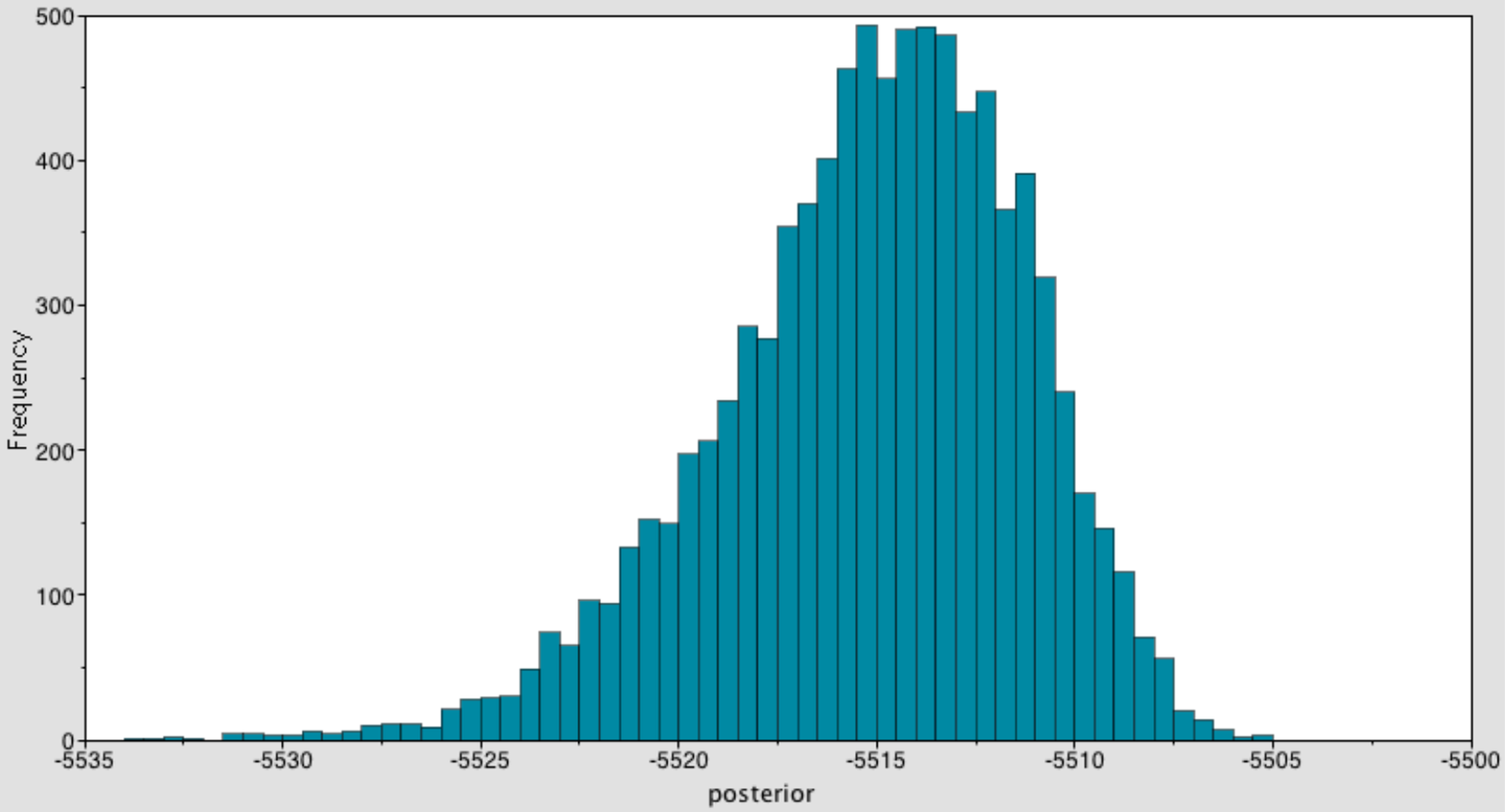
Traces:

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

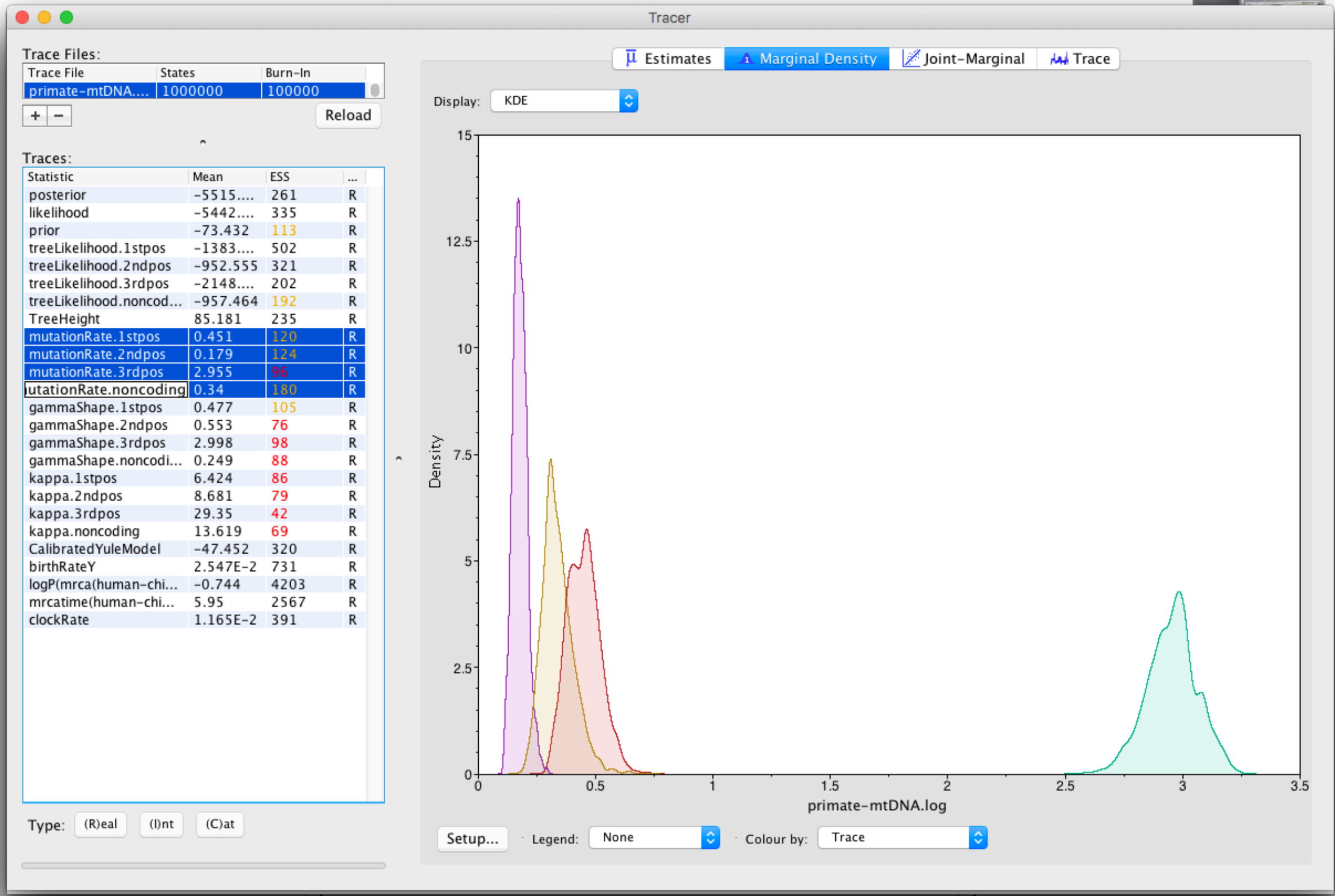
Estimates | Marginal Density | Joint-Marginal | Trace

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

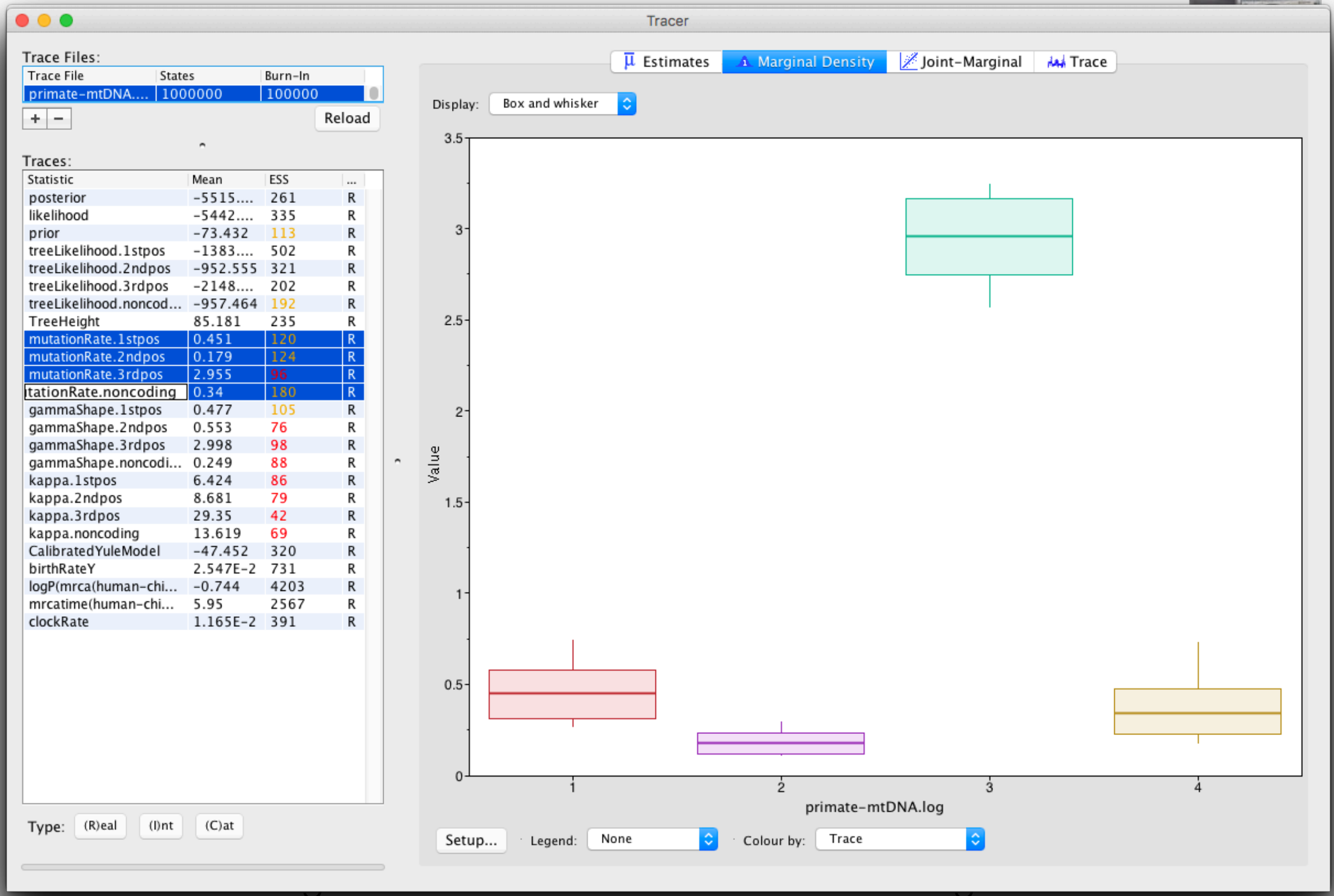


Setup... Bins: 50

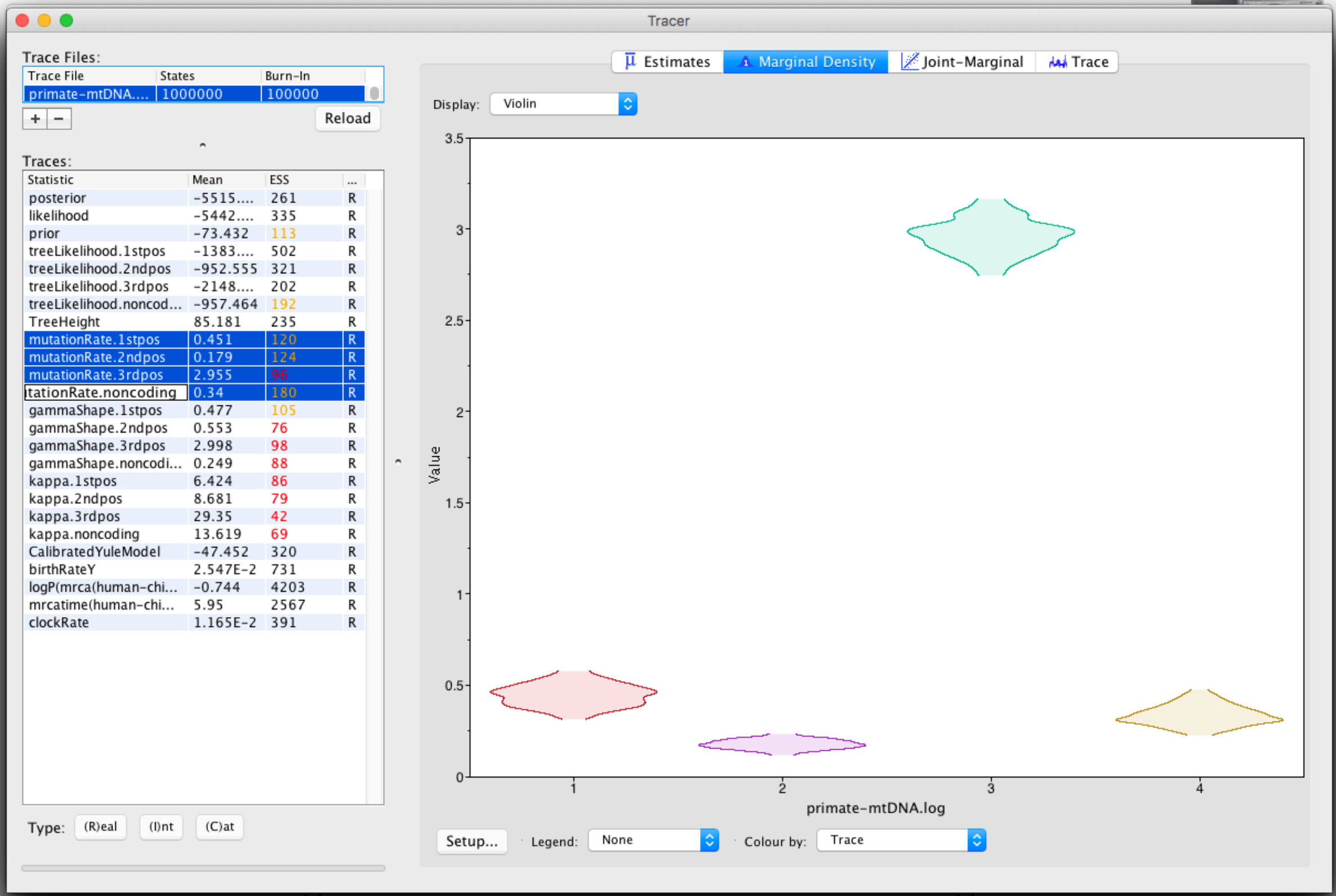
- Demographic reconstructions



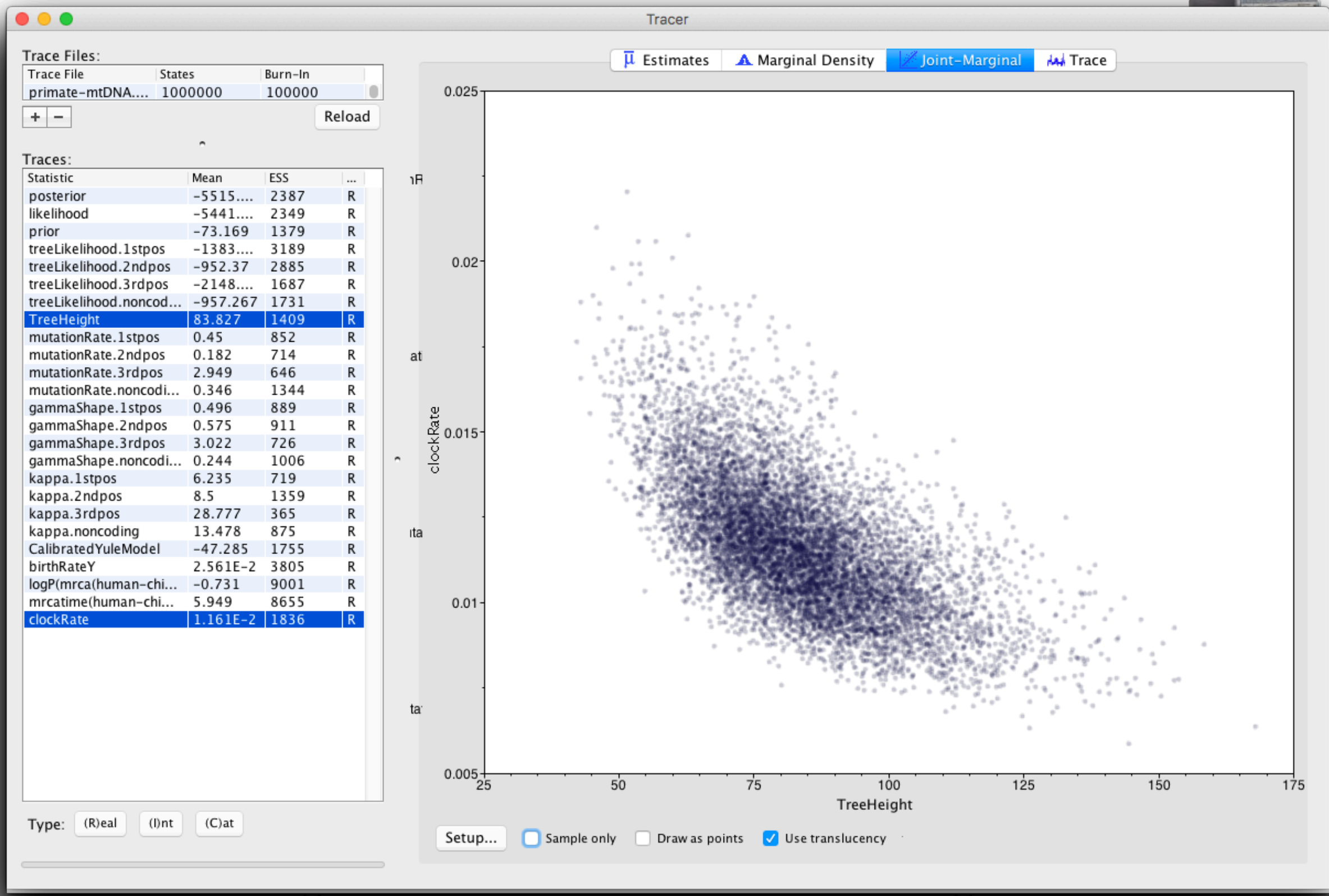
- Demographic reconstructions



- Demographic reconstructions



- Demographic reconstructions



- Demographic reconstructions

Tracer

Trace Files:

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

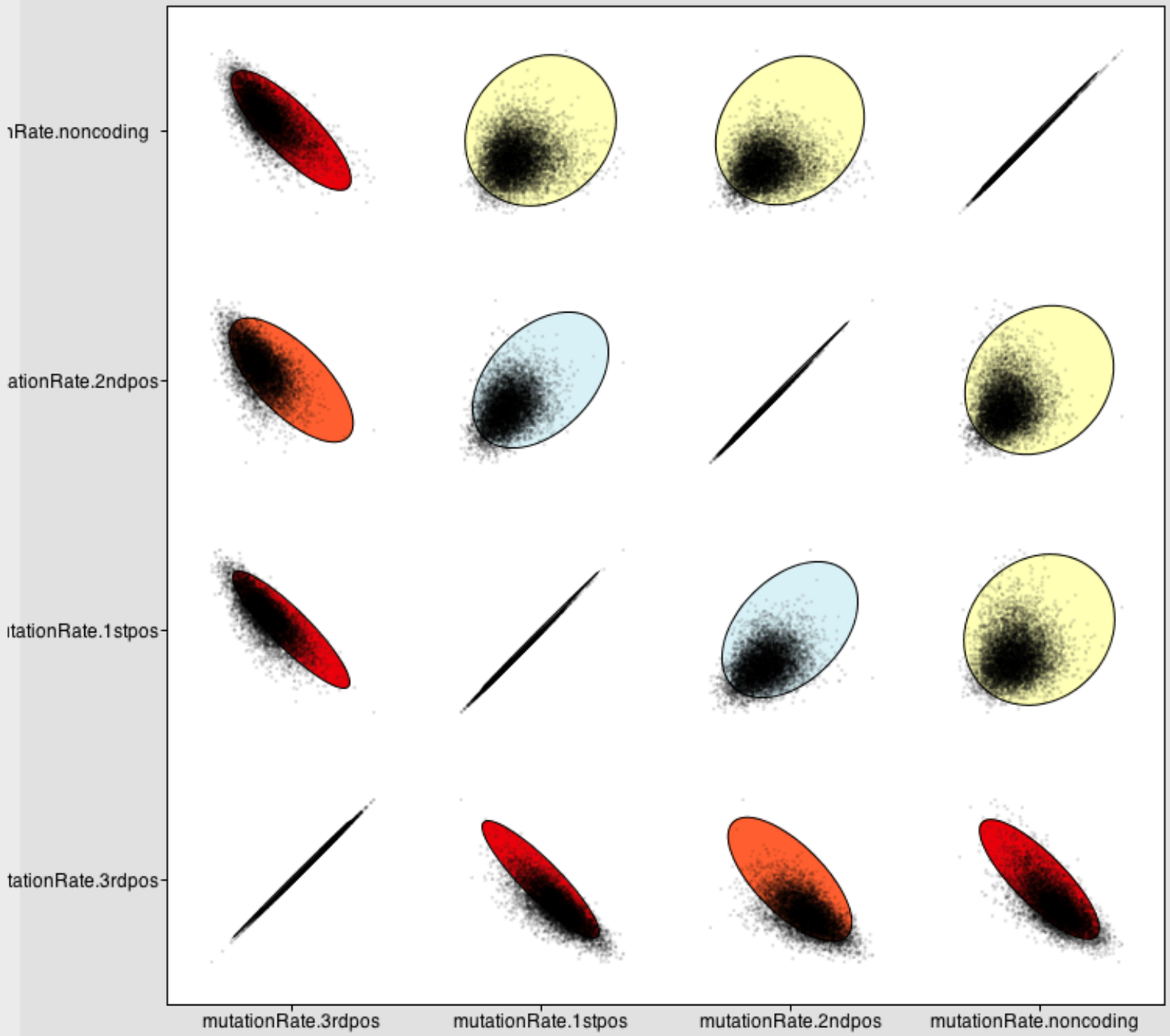
+ − Reload

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

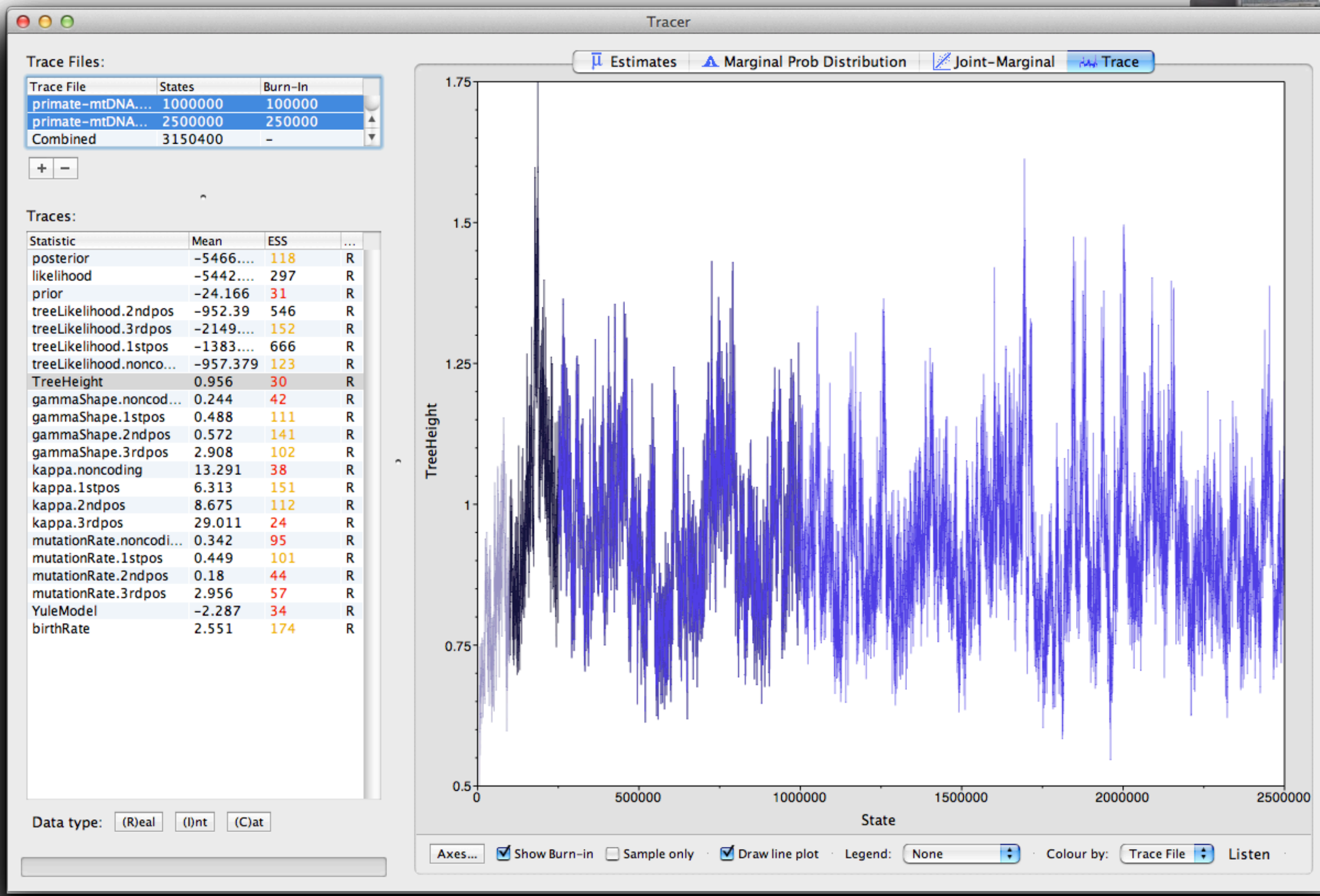
Type: (R)eal (I)nt (C)at

Estimates | Marginal Density | Joint-Marginal | Trace



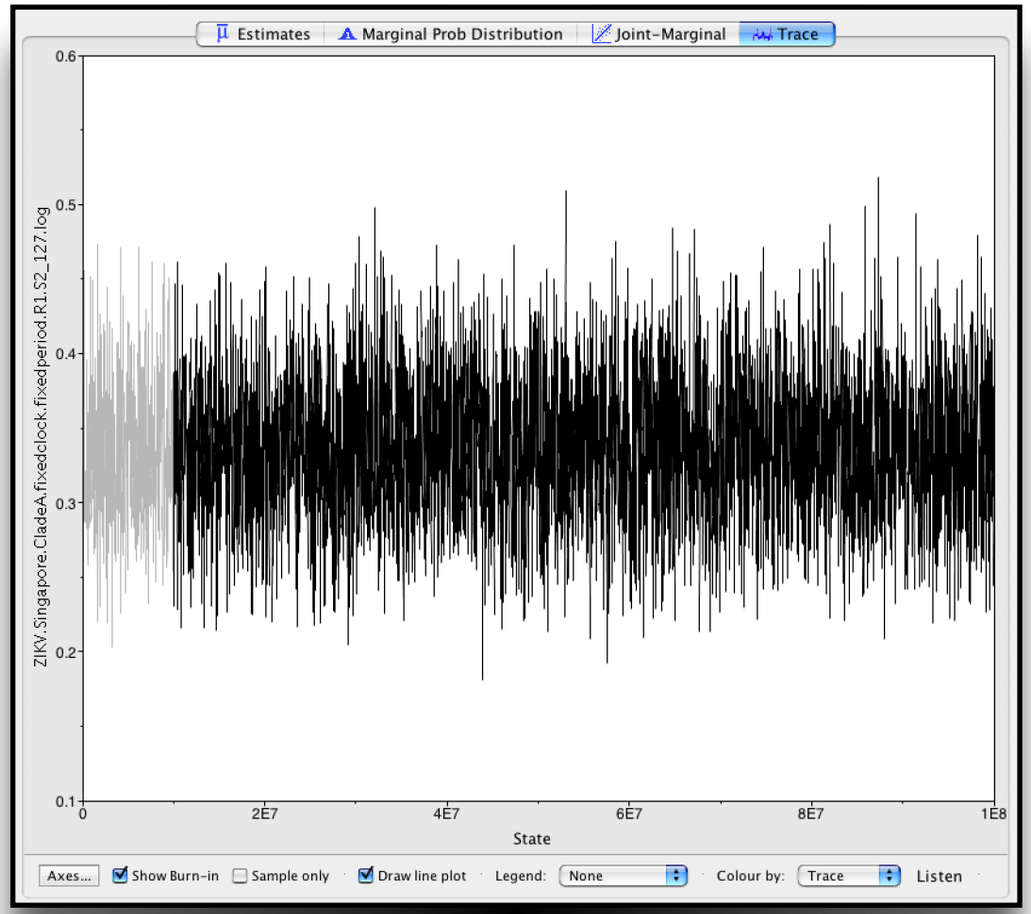
Show points

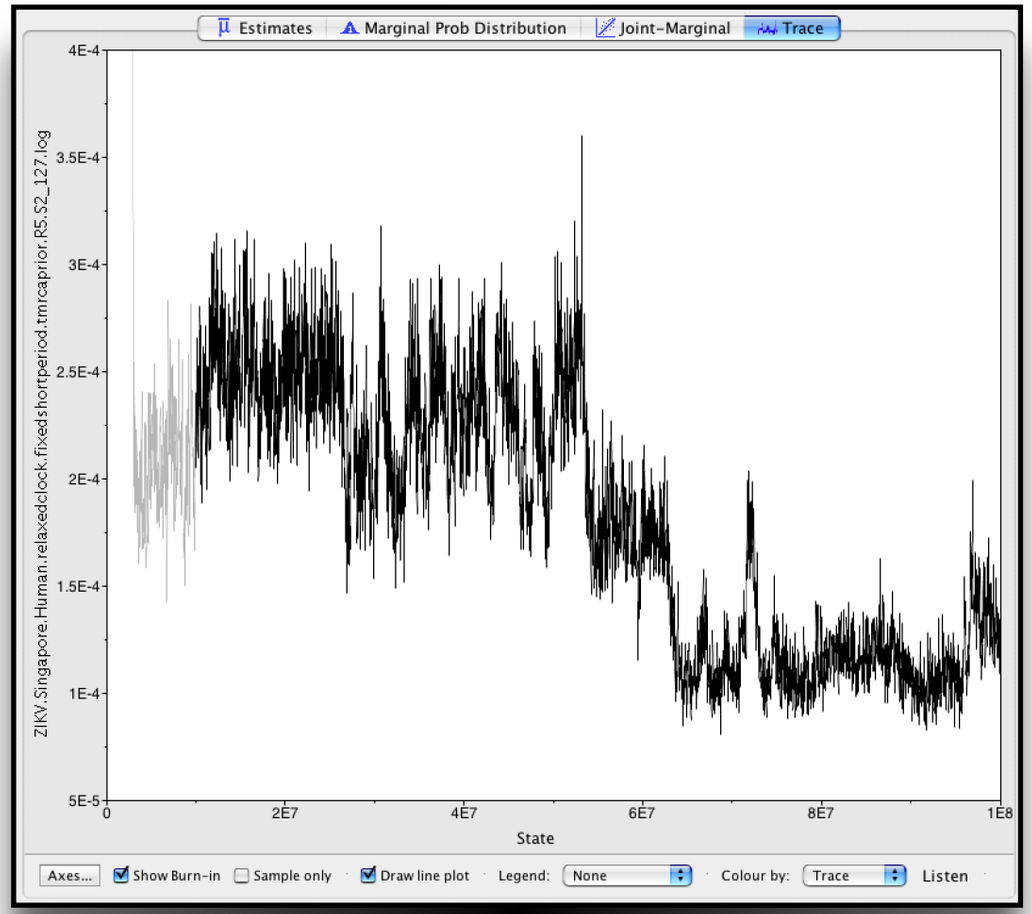
- Demographic reconstructions



- 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)

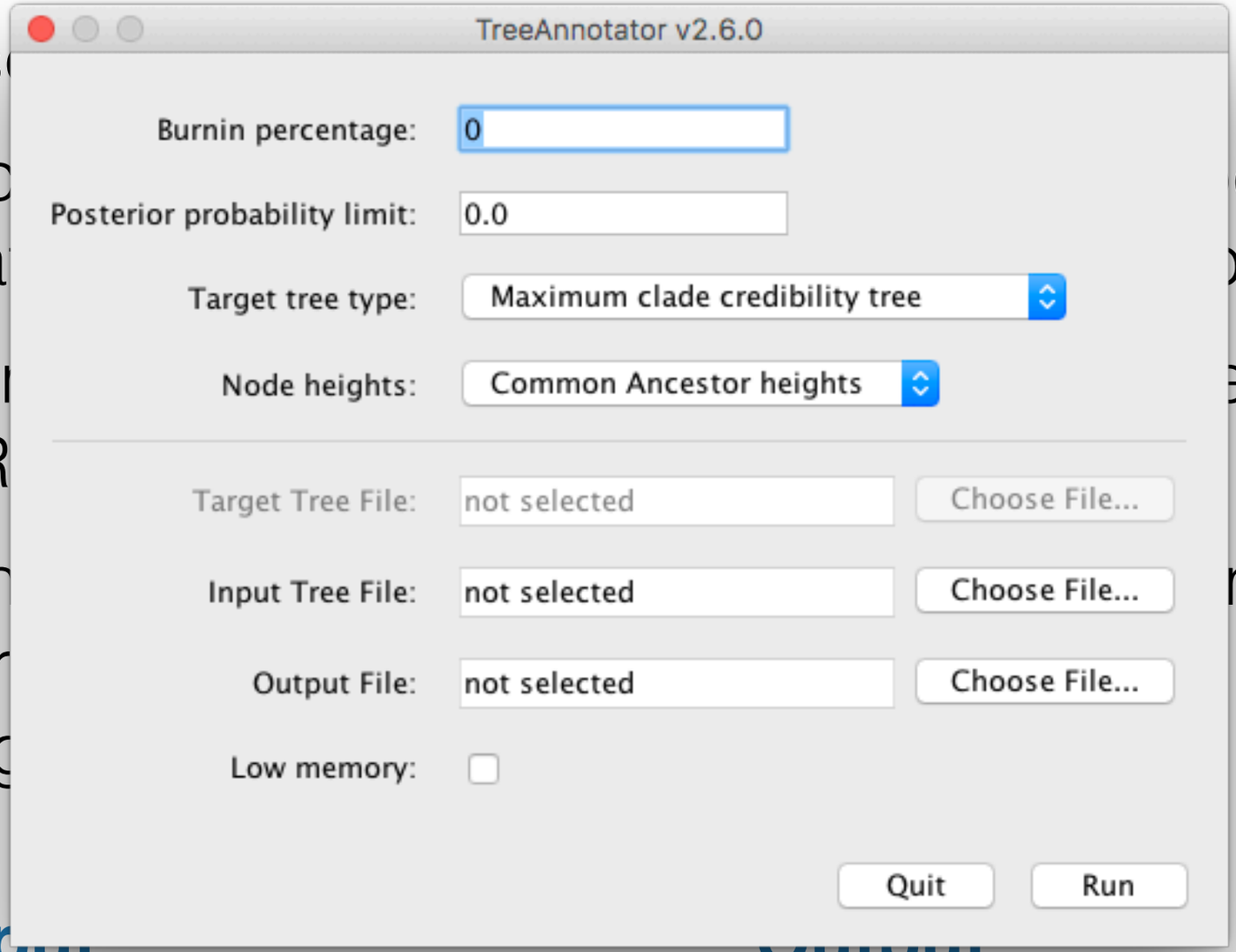
Output:

- 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

F
(h

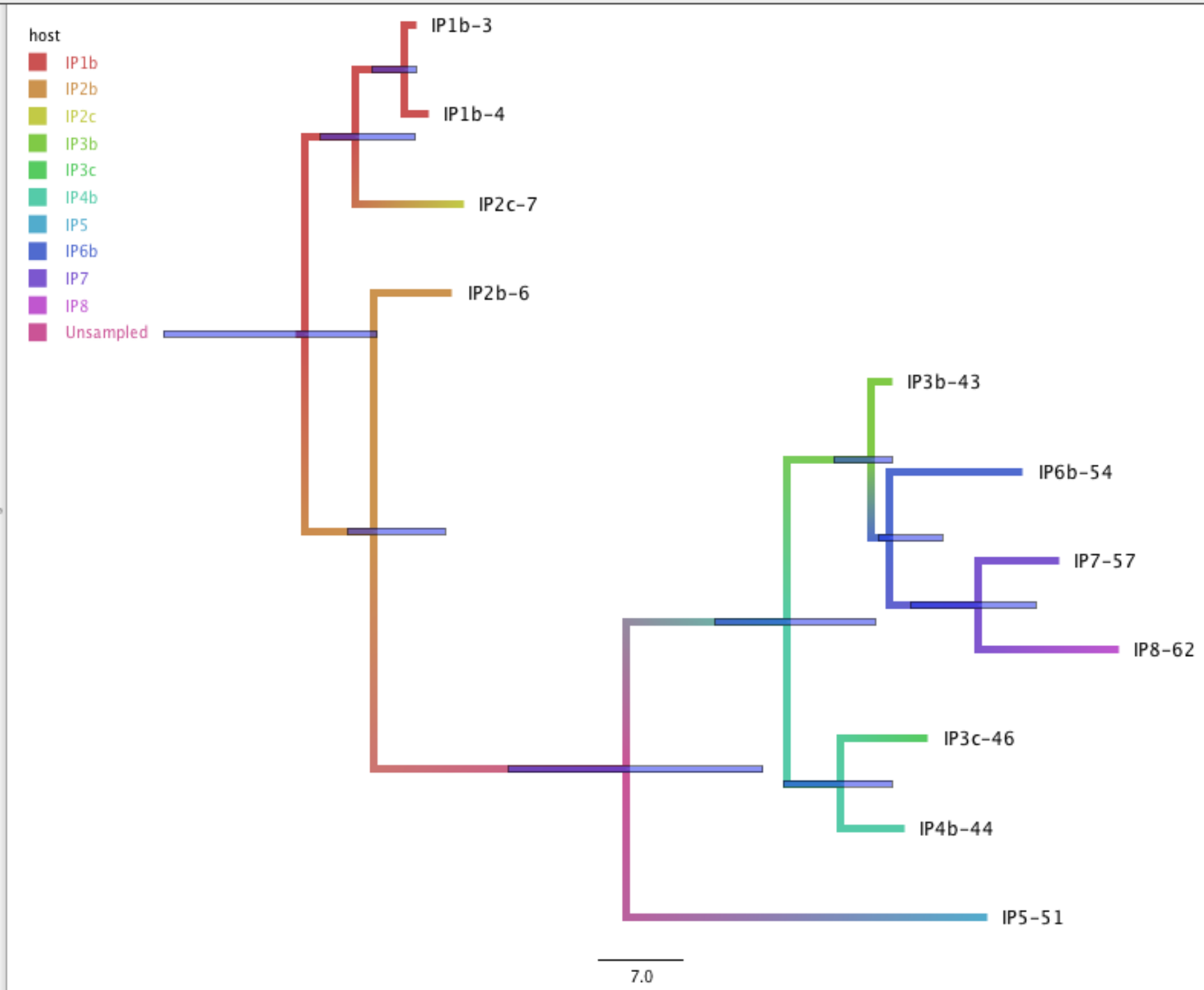
FigTree v1.4.3 - FMDV.MCC.tree

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]

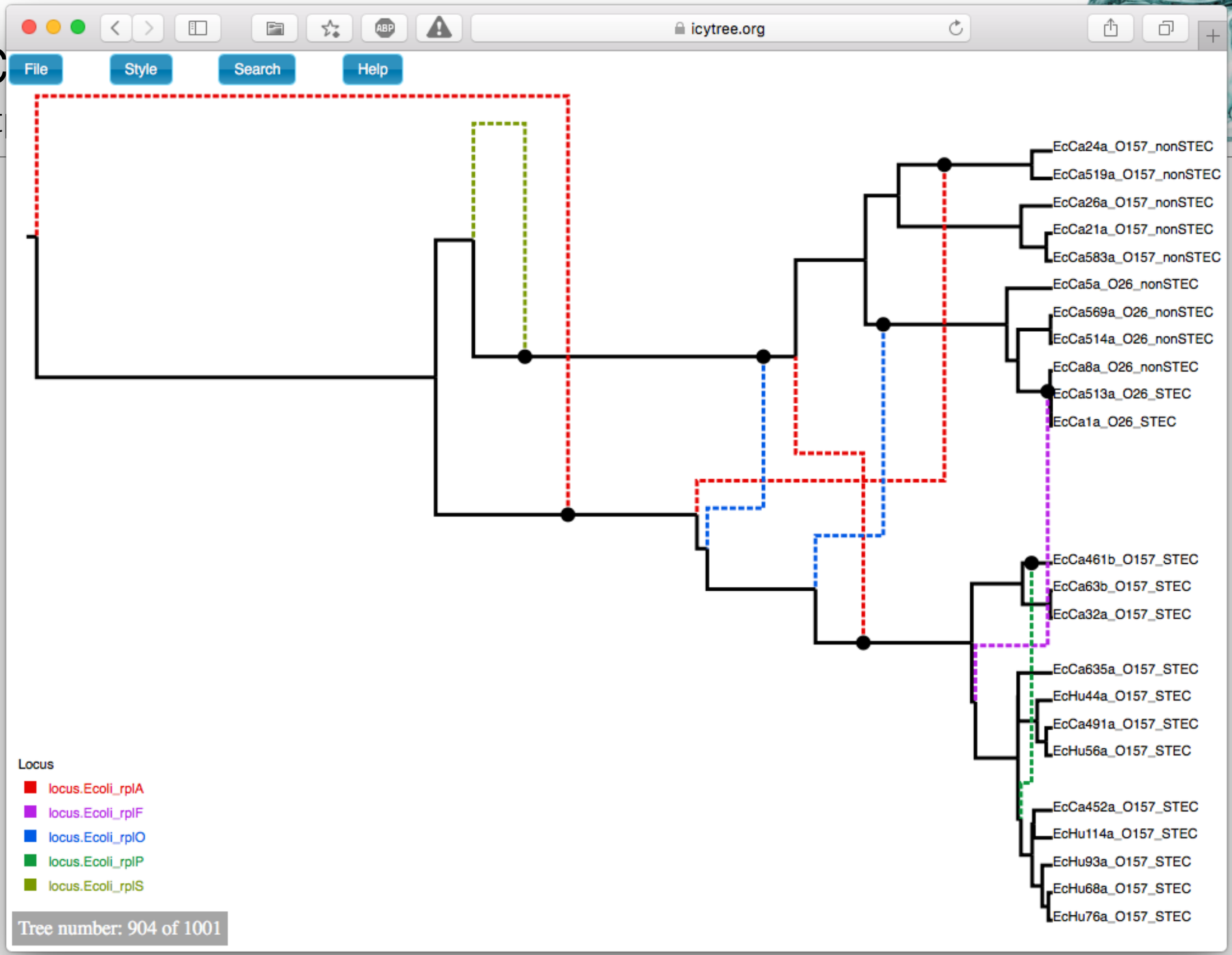
IcyTree

(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

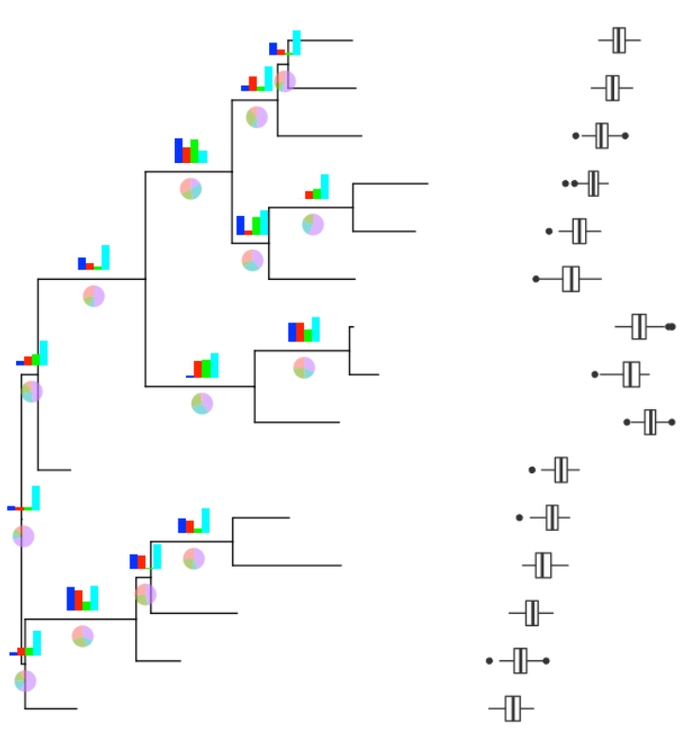
(htt

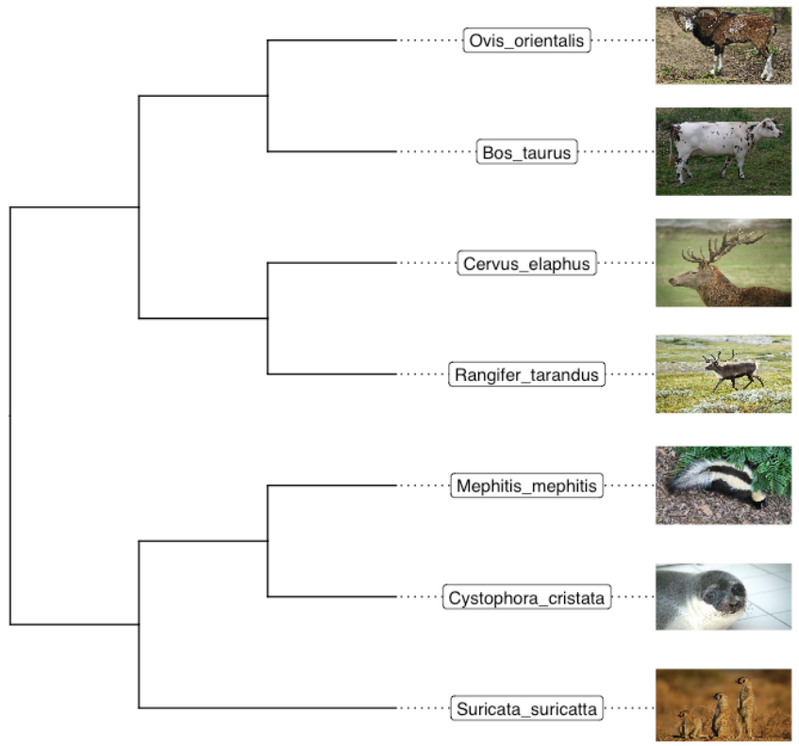

icytree.org
File
Style
Search
Help
EcCa24a_O157_nonSTEC
EcCa519a_O157_nonSTEC
EcCa26a_O157_nonSTEC
EcCa21a_O157_nonSTEC
EcCa583a_O157_nonSTEC
EcCa5a_O26_nonSTEC
EcCa569a_O26_nonSTEC
EcCa514a_O26_nonSTEC
EcCa8a_O26_nonSTEC
EcCa513a_O26_STEC
EcCa1a_O26_STEC
EcCa461b_O157_STEC
EcCa63b_O157_STEC
EcCa32a_O157_STEC
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

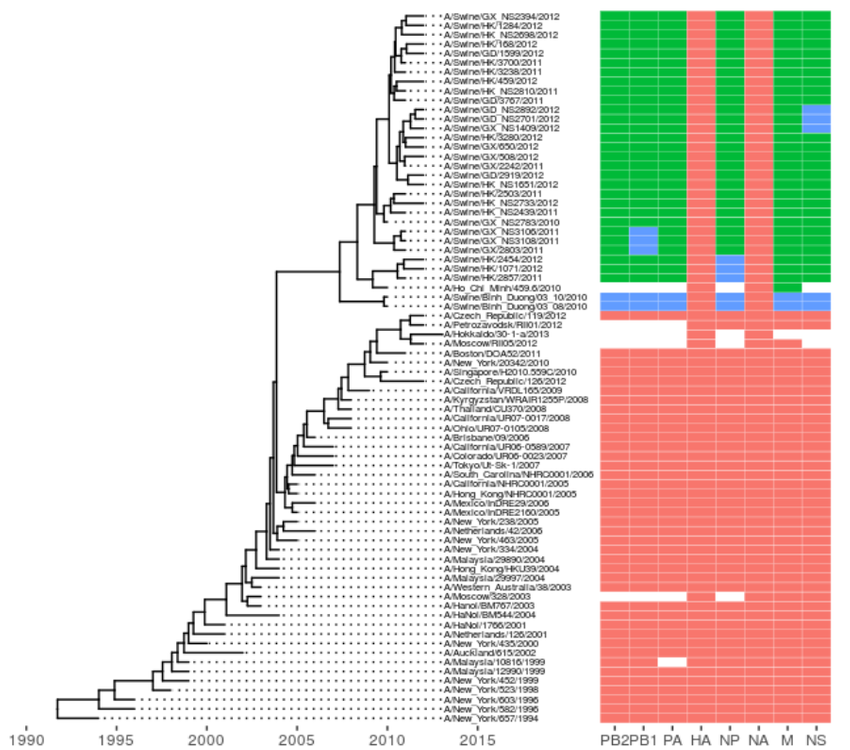
ggtree

(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

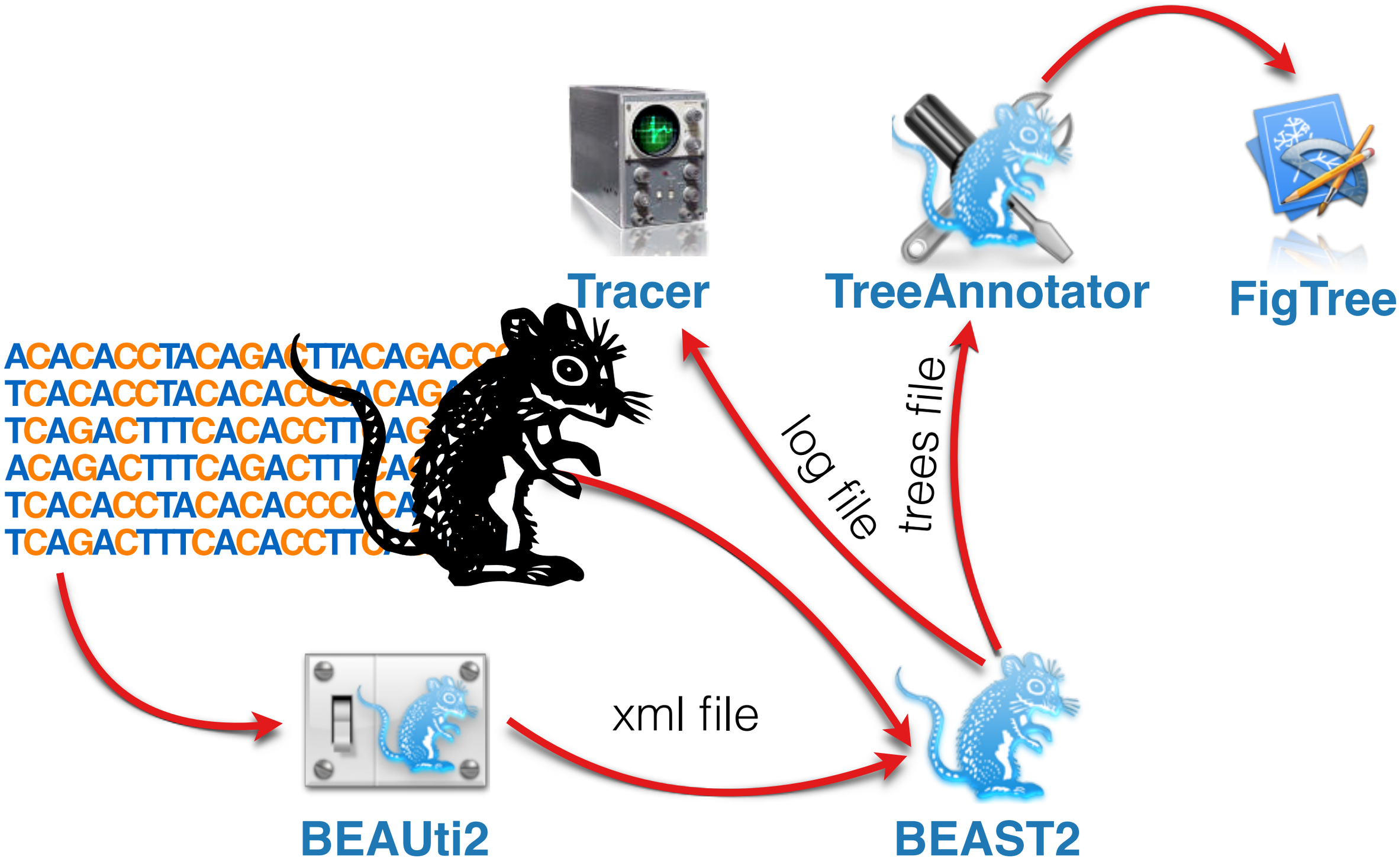






Website demo…

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
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