

Introduction to BEAST 2

Nídia Sequeira Trovão, PhD

nidia.trovao@nih.gov

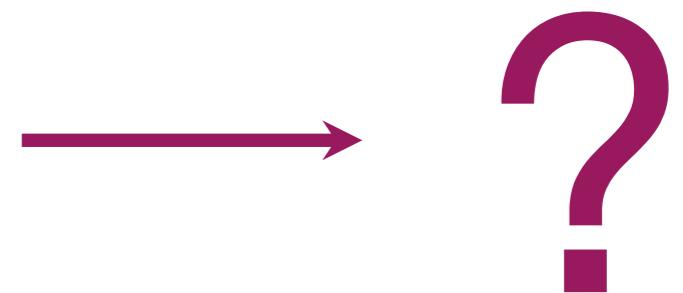
National Institutes of Health, U.S.



- 1.What goes into a BEAST model?
- 2.BEAST2 workflow
- 3.BEAST best practice
- 4.Introduction to BEAST2 tutorial

We all have one thing in common...

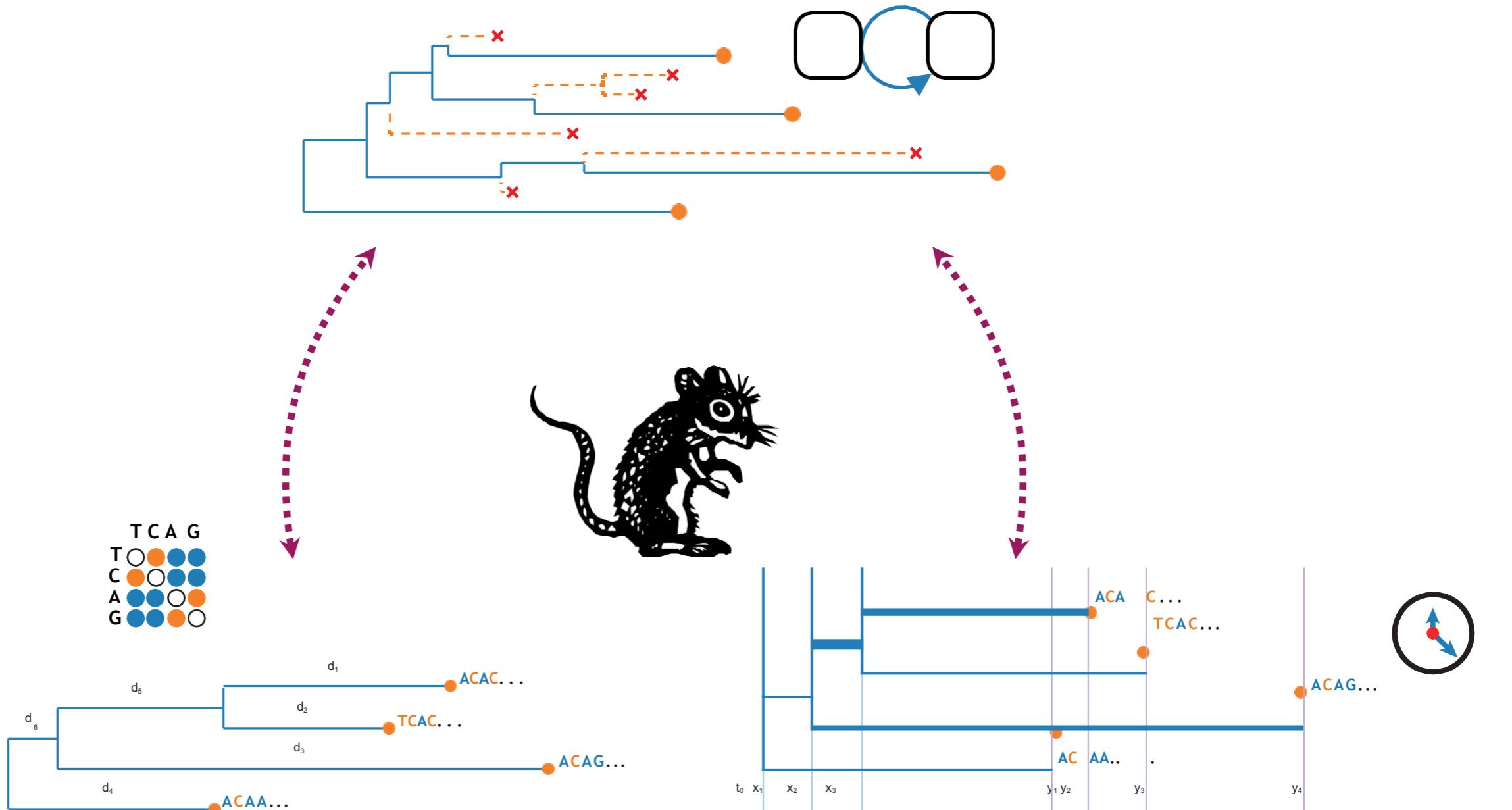
ACACACCTACAGACTTACAGACCC
TCACACCTACACACCCCACAGACTT
TCAGACTTTCACACCTTCAGACCT
ACAGACTTTCAGACTTTCAGACCC
TCACACCTACACACCCCACAGACTT
TCAGACTTTCACACCTTCAGACCTT



BEAST2

We all use **BEAST** to answer questions about our data

What goes into a BEAST model?



ACAC...
TCAC...
ACAG...
ACAA...

Genetic
sequences

Site model

Genealogy

Molecular clock
model

Demographic
model

Genetic data



NCBI Resources How To

Nucleotide Nucleotide Advanced Search Help

Nucleotide

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

GISAID © 2008 - 2020 | Terms of Use | Privacy Notice | Contact
You are logged in as **Nidia Trovao** - [logout](#)

Registered Users EpiFlu™ **EpiCoV™** My profile

EpiCoV™ Browse Downloads Upload My Unreleased

Search

Accession ID Virus name complete high coverage
Location Host low coverage excl w/Patient status
Collection Submission Reset Fulltext ▲

<input type="checkbox"/>	Virus name	Passage de	Accession ID	Collection da	Submission d	Length	Host	Location	Originating lab
<input type="checkbox"/>	hCoV-19/India/AFMC_5306/2020	Original	EPI_ISL_496602	2020-05-21	2020-07-24	29,889	Human	Asia / India / Mah	Armed Forces
<input type="checkbox"/>	hCoV-19/India/NCCS_NR570/2020	Original	EPI_ISL_496587	2020-05-08	2020-07-24	29,901	Human	Asia / India / Mah	National Centre
<input type="checkbox"/>	hCoV-19/India/NCCS_NR342/2020	Original	EPI_ISL_496586	2020-05-03	2020-07-24	29,873	Human	Asia / India / Mah	National Centre
<input type="checkbox"/>	hCoV-19/India/NCCS_NR2527/2020	Original	EPI_ISL_496585	2020-05-24	2020-07-24	29,873	Human	Asia / India / Mah	National Centre
<input type="checkbox"/>	hCoV-19/India/NCCS_NR2523/2020	Original	EPI_ISL_496584	2020-05-24	2020-07-24	29,892	Human	Asia / India / Mah	National Centre
<input type="checkbox"/>	hCoV-19/India/NCCS_NR2517/2020	Original	EPI_ISL_496583	2020-05-24	2020-07-24	29,848	Human	Asia / India / Mah	National Centre
<input type="checkbox"/>	hCoV-19/India/NCCS_NR2513/2020	Original	EPI_ISL_496582	2020-05-24	2020-07-24	29,901	Human	Asia / India / Mah	National Centre
<input type="checkbox"/>	hCoV-19/India/NCCS_NR2359/2020	Original	EPI_ISL_496581	2020-05-22	2020-07-24	29,837	Human	Asia / India / Mah	National Centre
<input type="checkbox"/>	hCoV-19/India/NCCS_NR2347/2020	Original	EPI_ISL_496580	2020-05-22	2020-07-24	29,873	Human	Asia / India / Mah	National Centre
<input type="checkbox"/>	hCoV-19/India/NCCS_NR2329/2020	Original	EPI_ISL_496579	2020-05-22	2020-07-24	29,661	Human	Asia / India / Mah	National Centre
<input type="checkbox"/>	hCoV-19/India/NCCS_NR2328/2020	Original	EPI_ISL_496578	2020-05-22	2020-07-24	29,872	Human	Asia / India / Mah	National Centre
<input type="checkbox"/>	hCoV-19/India/NCCS_NR2327/2020	Original	EPI_ISL_496577	2020-05-22	2020-07-24	29,887	Human	Asia / India / Mah	National Centre

Total: 71,581 viruses << first < prev 1 2 3 4 5 6 7 8 9 10 next > last >> Analysis Download

Important note: In the [GISAID EpiFlu™ Database Access Agreement](#), you have accepted certain terms and conditions for viewing and using data regarding influenza viruses. To the extent the Database contains data relating to non-influenza viruses, the viewing and use of these data is subject to the same terms and conditions, and by viewing or using such data you agree to be bound by the terms of the [GISAID EpiFlu™ Database Access Agreement](#) in respect of such data in the same manner as if they were data relating to influenza viruses.

Sequence alignment

MUSCLE

[Input form](#) [Web services](#) [Help & Documentation](#)

[Feedback](#) [Share](#)

Tools > Multiple Sequence Alignment > MUSCLE

Service Retirement

We remind you that it is not long until the EBI's Wise2DBA and Promoterwise services are retired on 15th April 2018. Alternatives can be found at [Exonerate](#), [BWA](#) or [BLAT](#). If you have any concerns, please contact us via [support](#).

Multiple Sequence Alignment

MUSCLE stands for **M**Ultiple **S**equence **C**omparison by **L**og- **E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.

Important note: This tool can align up to 500 sequences or a maximum file size of 1 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of sequences in any supported format:

Or upload a file: No file chosen

STEP 2 - Set your Parameters

OUTPUT FORMAT:

ClustalW

The default settings will fulfill the needs of most users.

(Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

Be notified by email (Tick this box if you want to be notified by email when the results are available)

www.genome.jp/tools-bin/clustalw



Multiple Sequence Alignment by CLUSTALW

[ETE3](#) [MAFFT](#) [CLUSTALW](#) [PRRN](#)

[Help](#)

General Setting Parameters:

Output Format:

Pairwise Alignment: FAST/APPROXIMATE SLOW/ACCURATE

Enter your sequences (with labels) below (copy & paste): PROTEIN DNA

Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

Or give the file name containing your query

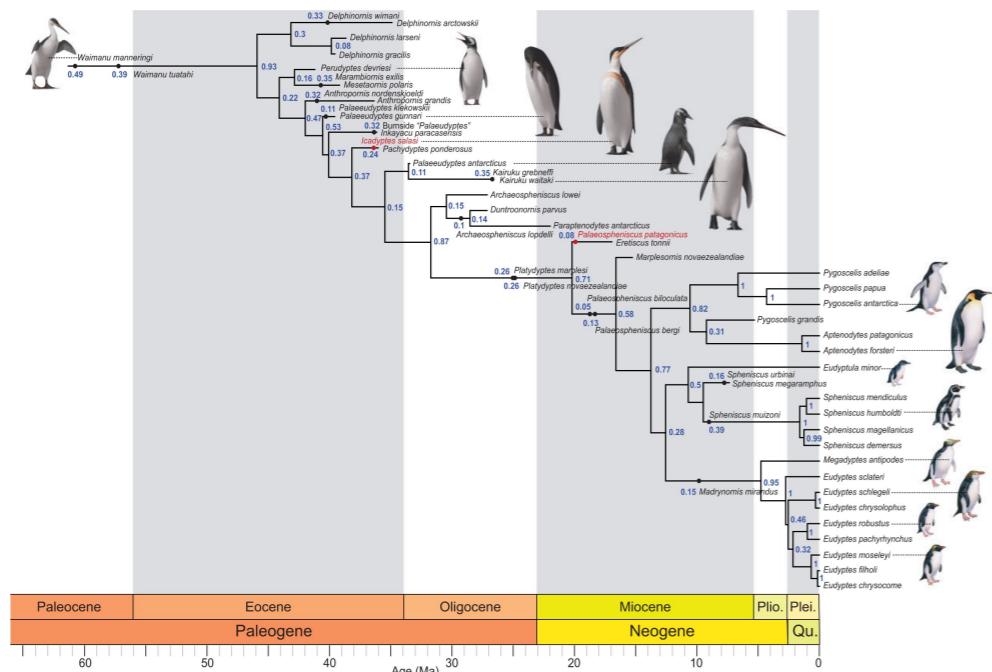
No file chosen

The Data!

- Typically an alignment of DNA or RNA sequences
- Can also be amino acids or codons
- Often split data into multiple partitions
 - Multiple genes
 - 1st, 2nd and 3rd codon positions
- Does **not** have to be genetic sequences!

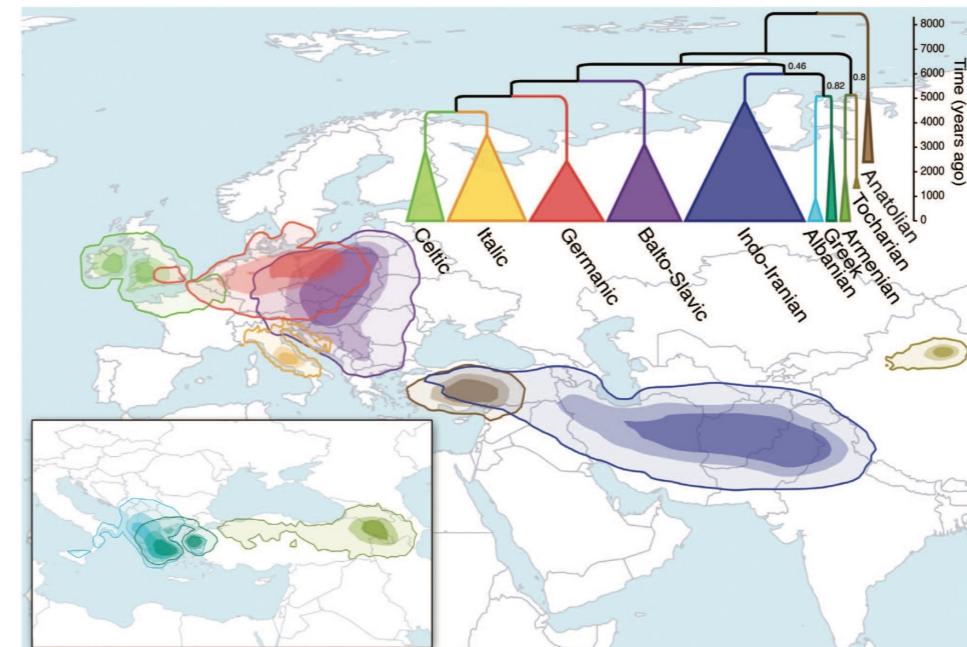
ACAC...
 TCAC...
 ACAG...
 ACAC...
 TCAC...
 ACAG...
 ACAC...
 TCAC...
 ACAG...
 ACAC...
 TCAC...
 ACAG...

Morphological traits



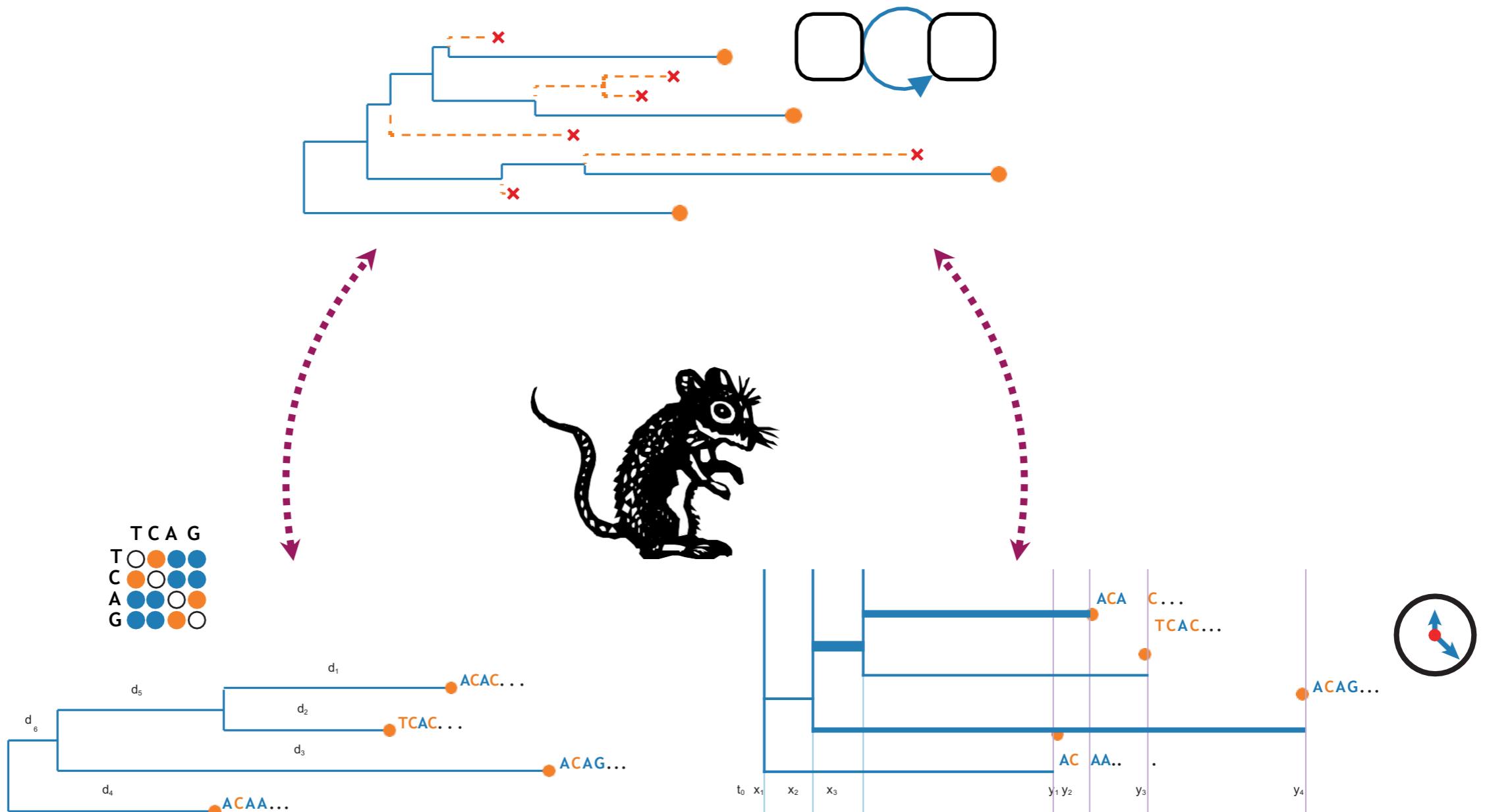
Gavryushkina et al. Systematic Biology 2017

Roots of words



Bouckaert et al. Science 2012

What goes into a BEAST model?



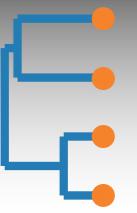
ACAC...
TCAC...
ACAG...
ACAA...
Genetic sequences

Genealogy

Demographic model

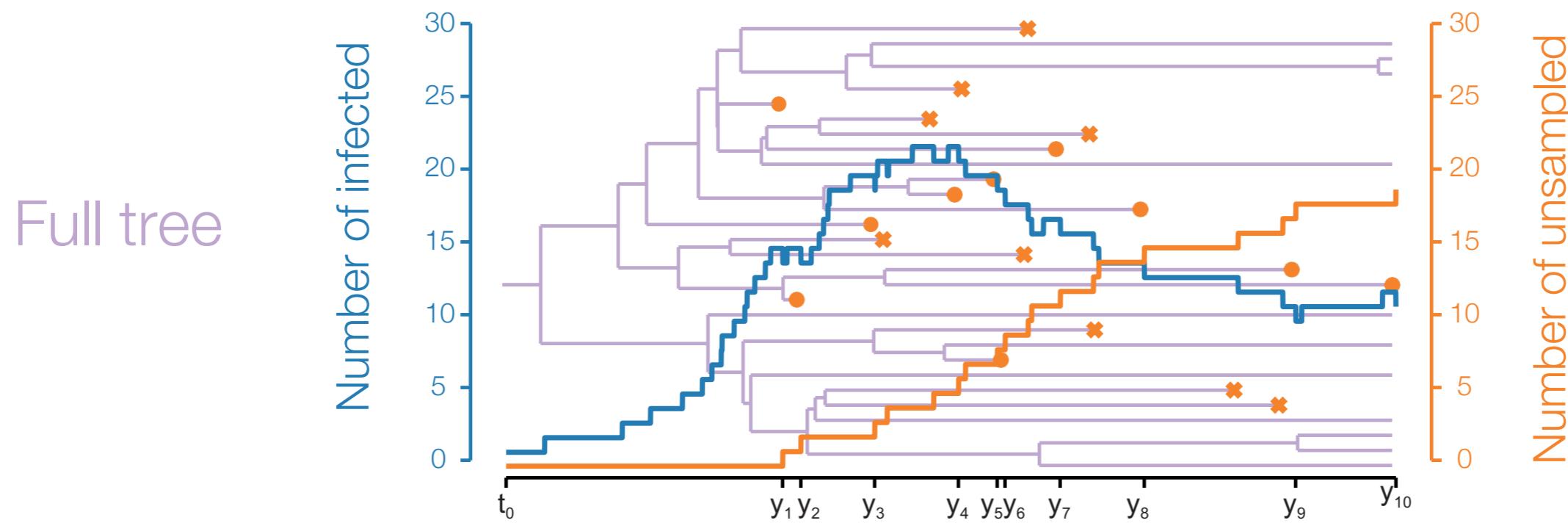
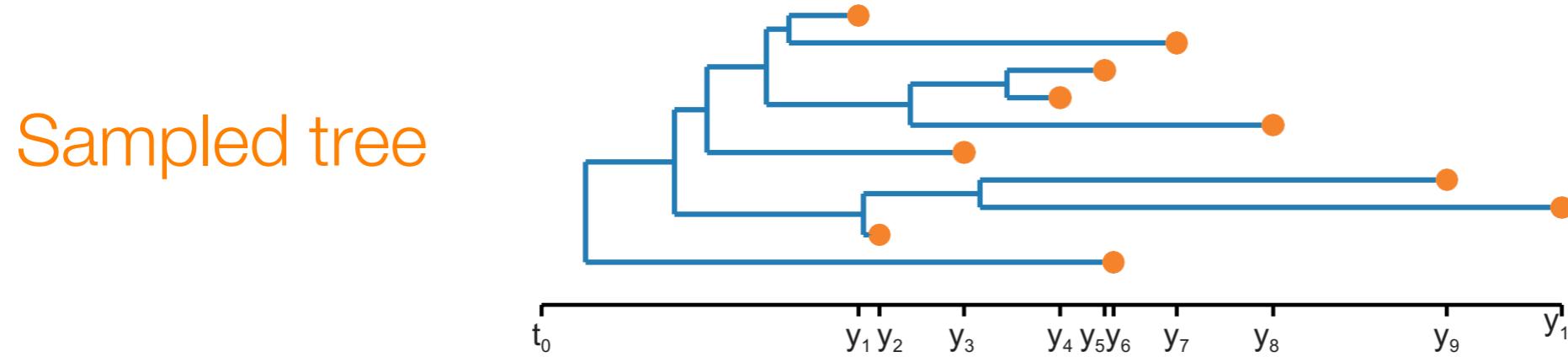
Site model

Molecular clock model



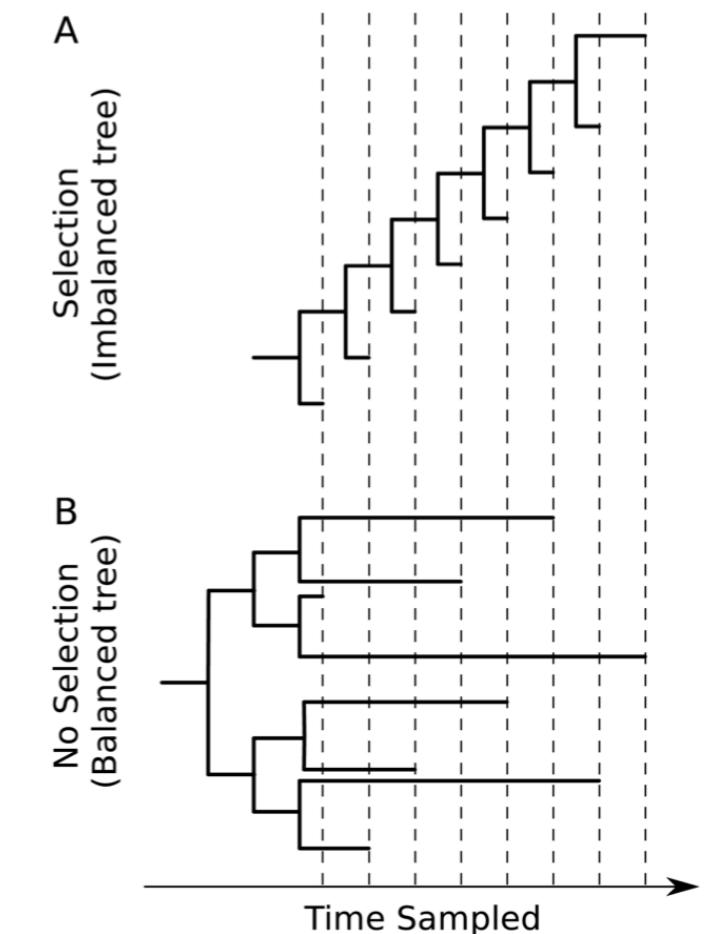
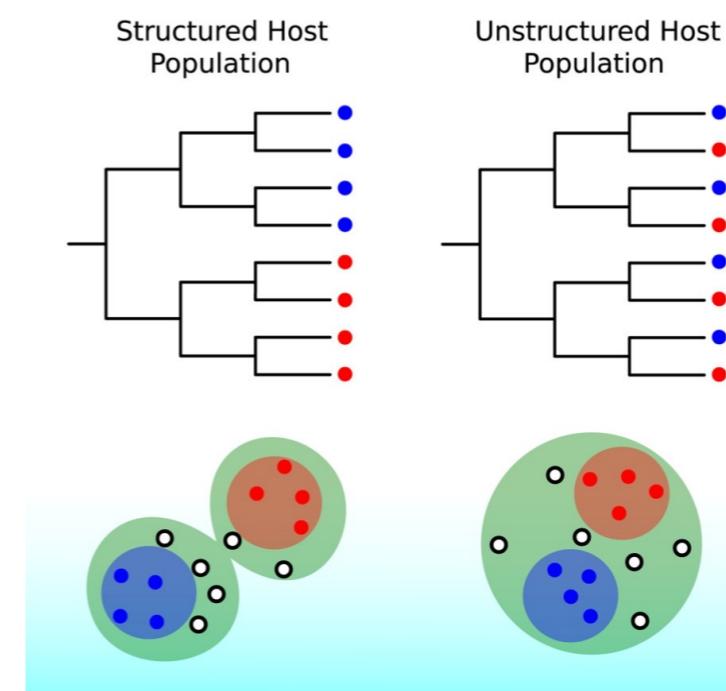
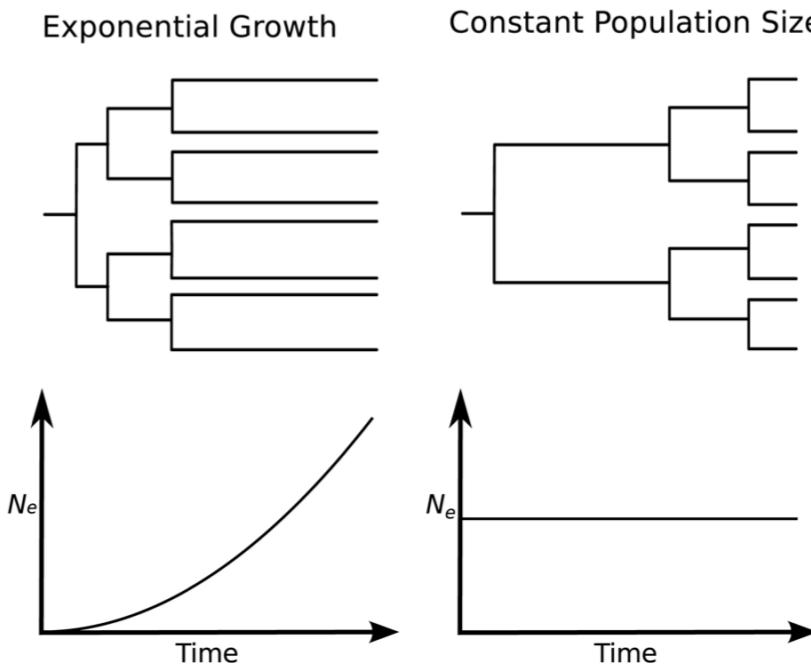
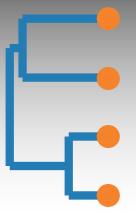
The genealogy (phylogenetic tree)

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

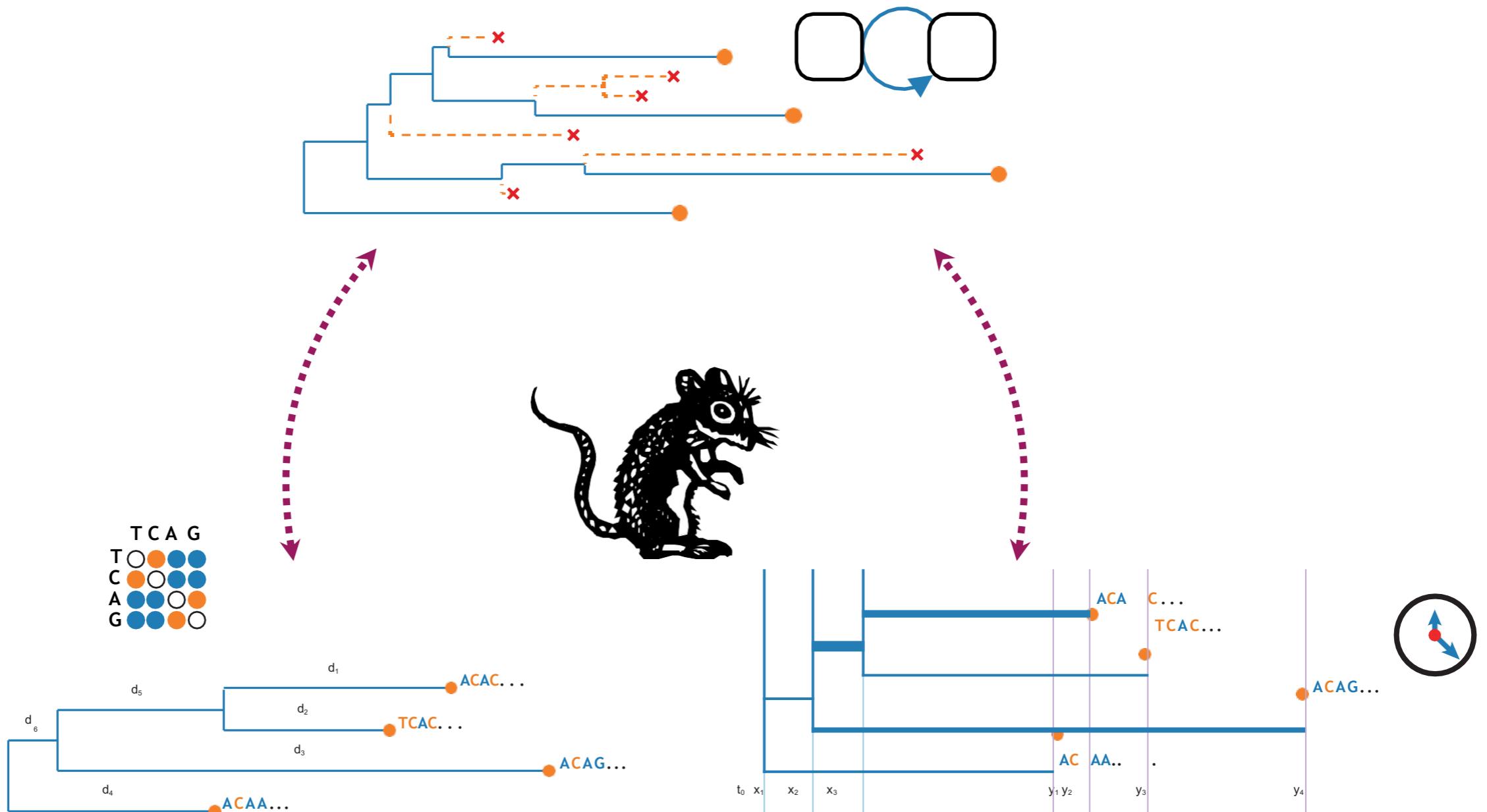


- Only the relationships between the sampled sequences!

Different population dynamics generate different trees



What goes into a BEAST model?



ACAC...
TCAC...
ACAG...
Genetic
sequences

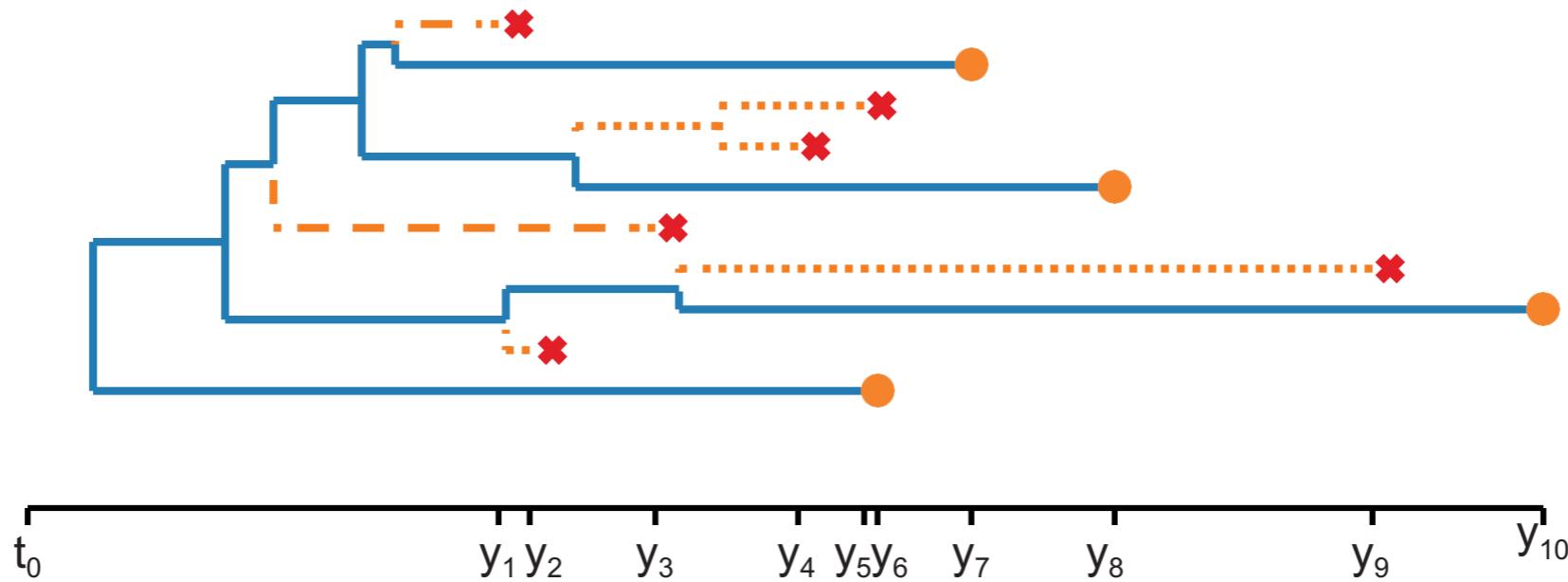
E
Genealogy

Demographic
model

Site model

ACAC...
TCAC...
ACAG...
Molecular clock
model

Demographic model

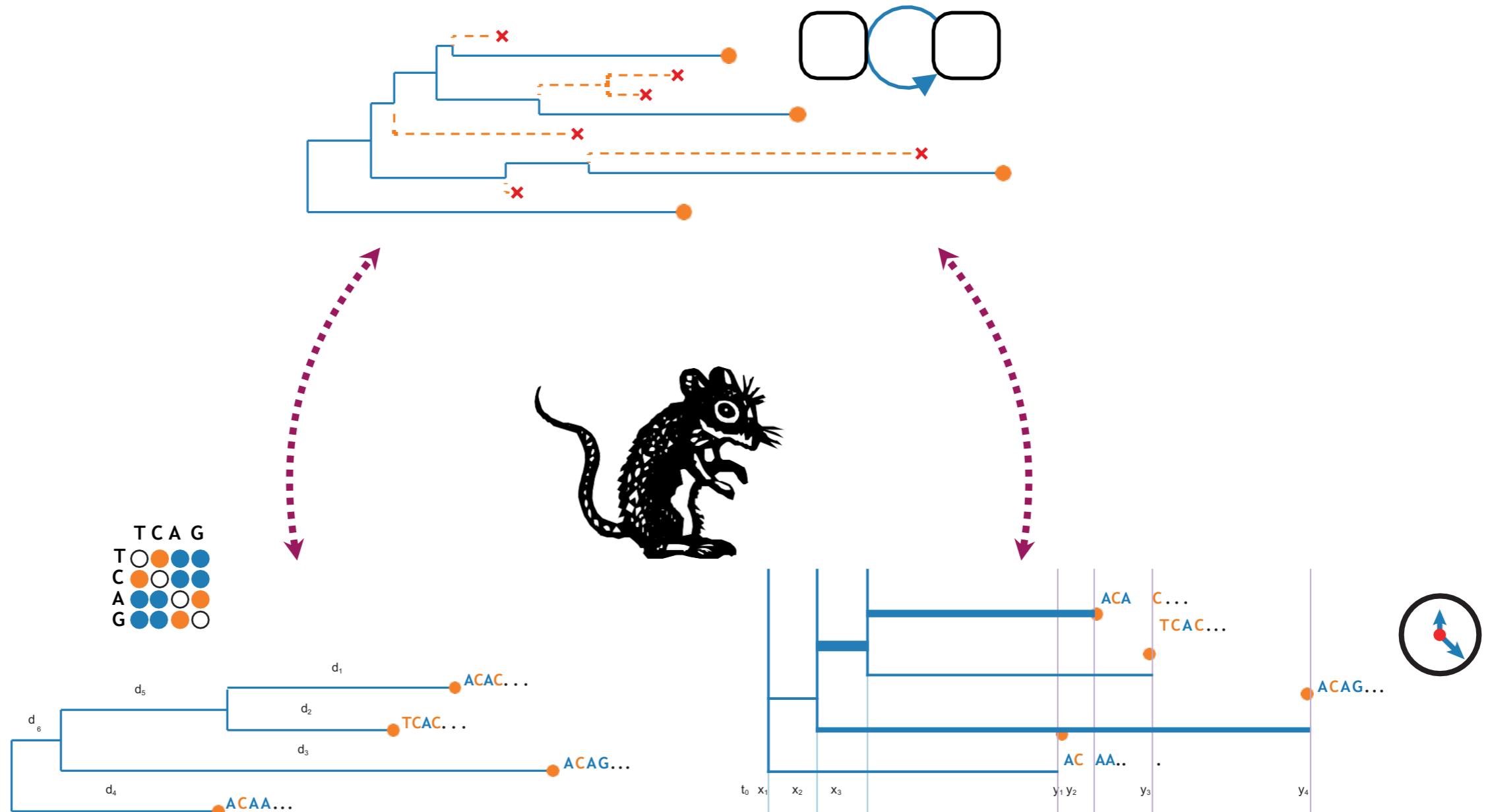


- Describes the population/speciation dynamics
- How does the population grow over time?
- How does the species diversity change over time?

$$P(\mathcal{E} \mid \text{Tree-Prior})$$

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

What goes into a BEAST model?



ACAC...
TCAC...
ACAG...
ACAA...

Genetic
sequences

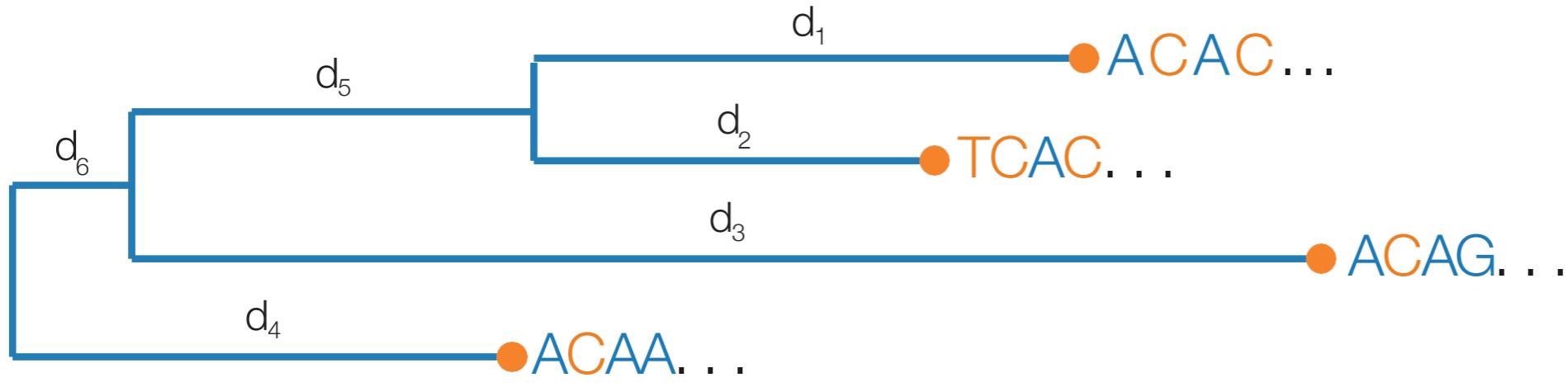
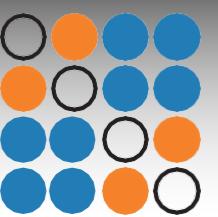
Genealogy

Demographic
model

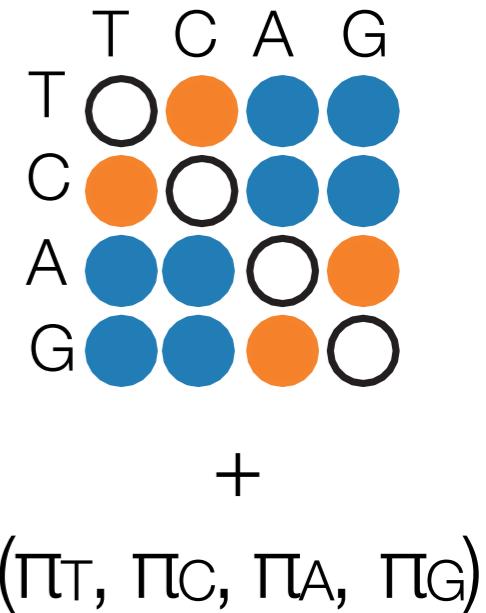
Site model

Molecular clock
model

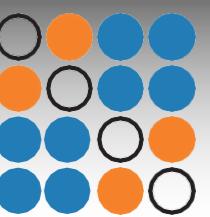
Site model



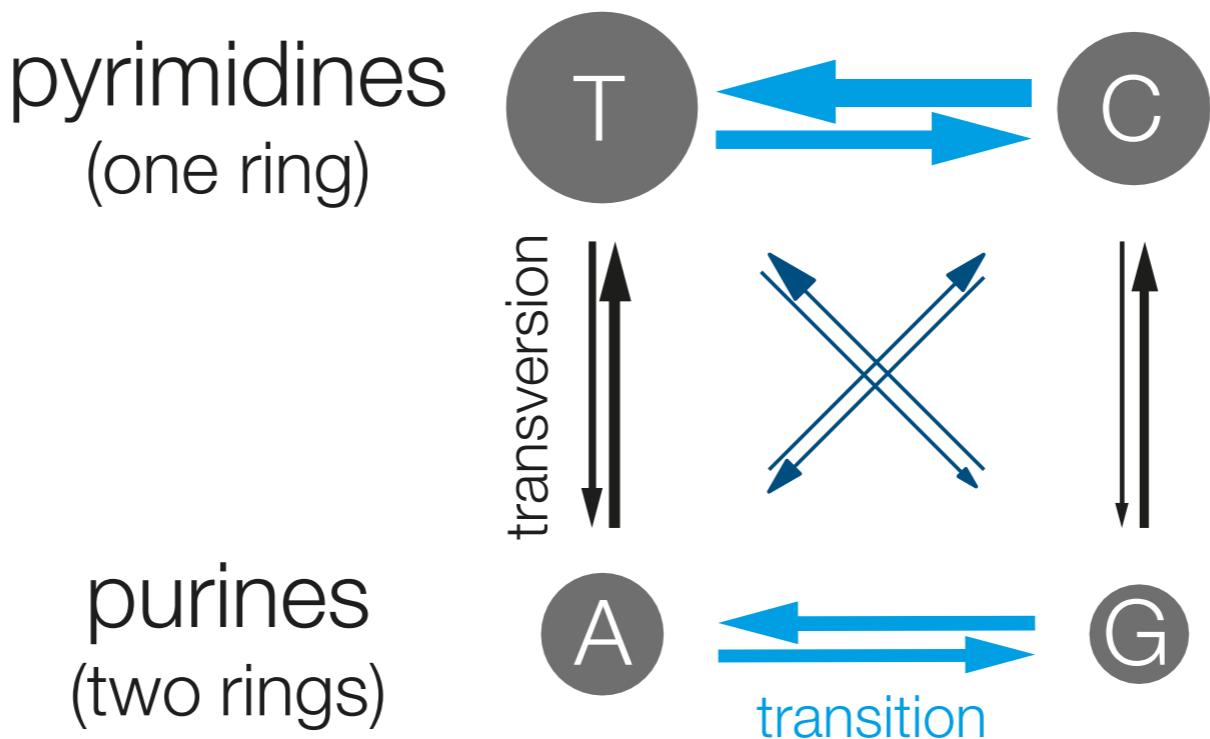
d: Genetic distance from common ancestor



- 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!

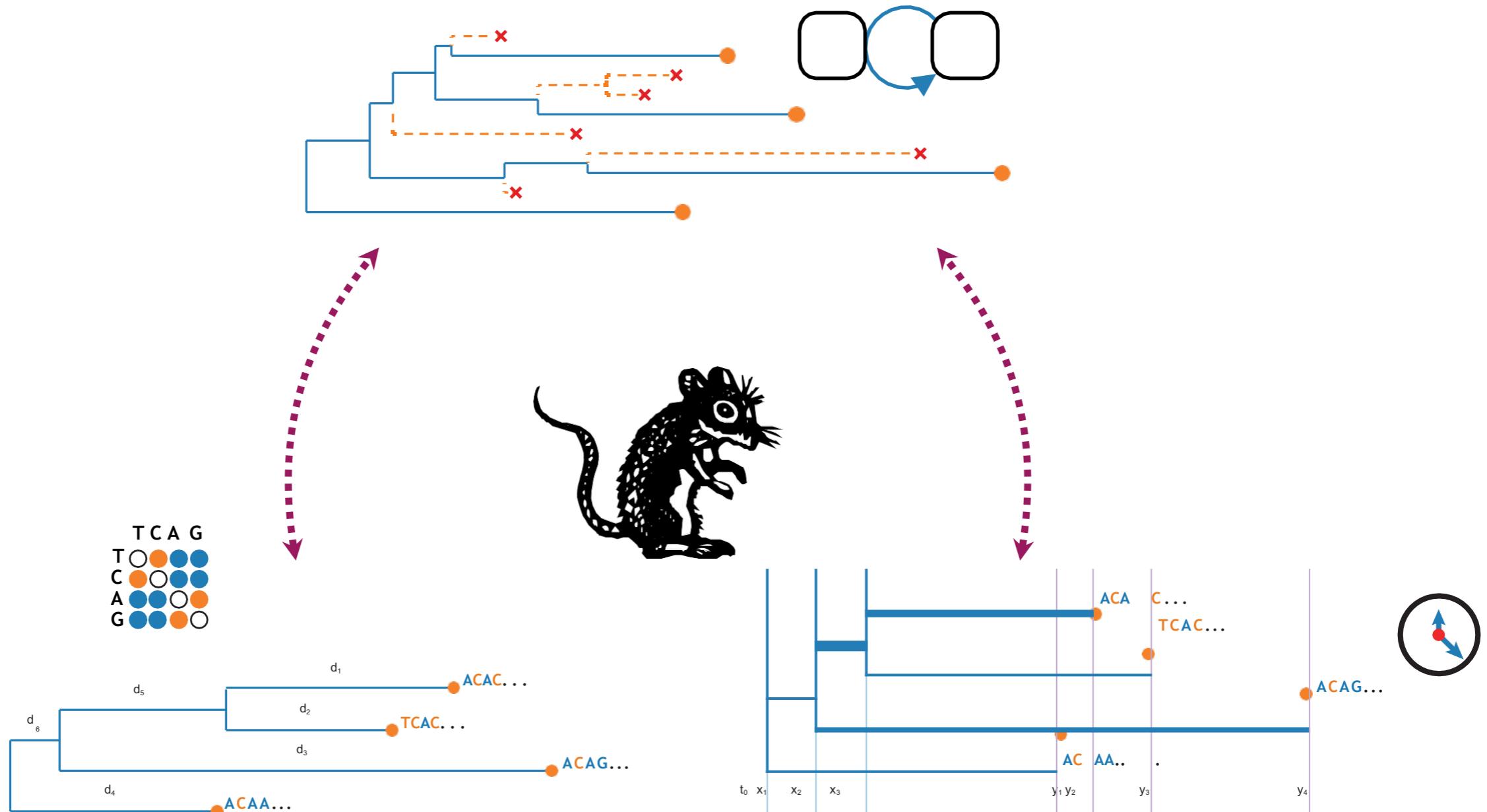


Substitutions as a Markov process



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

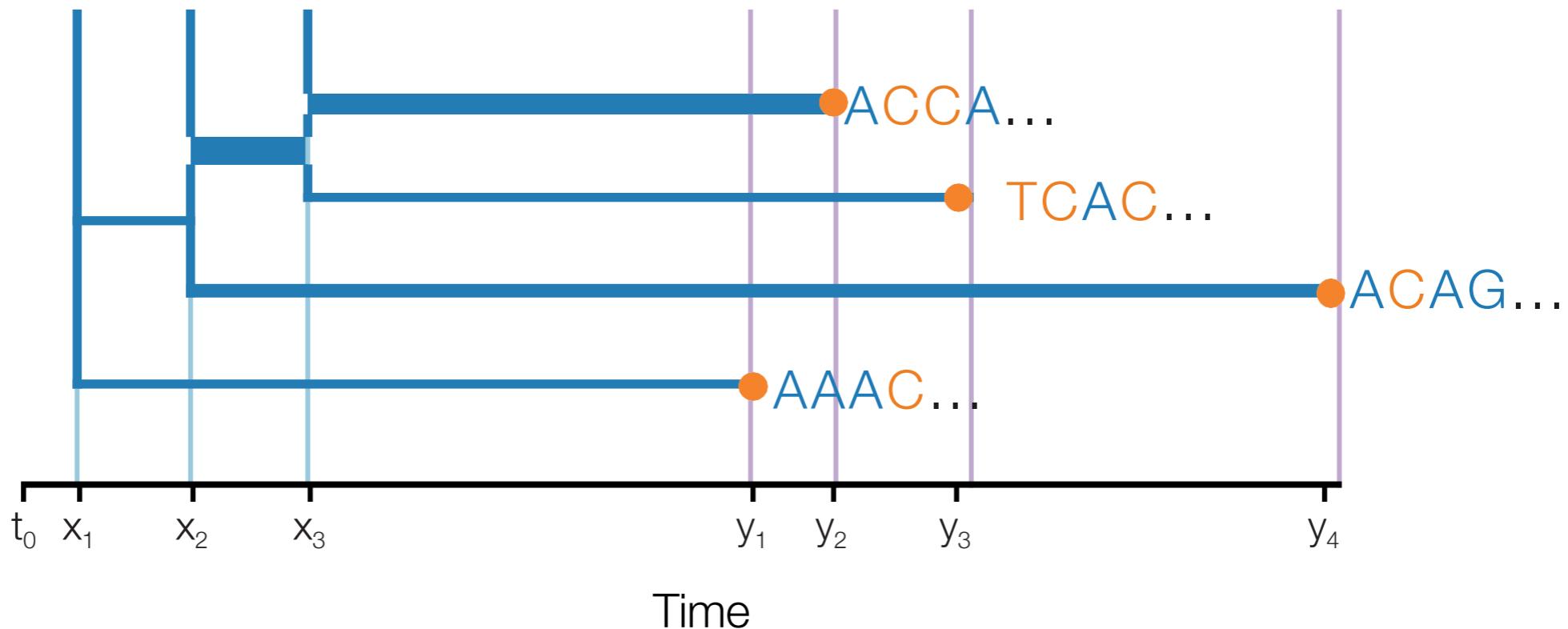
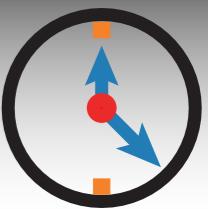
Genealogy

Demographic
model

Site model

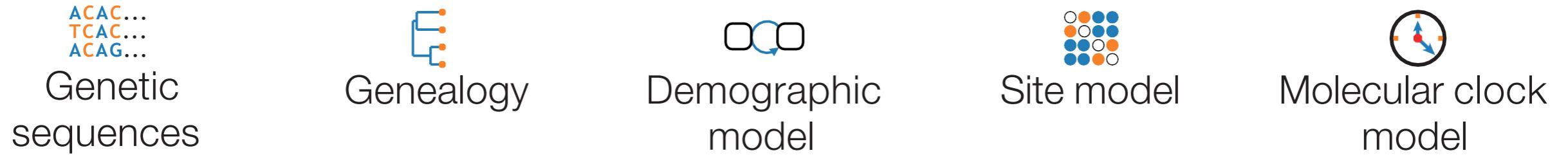
Molecular clock
model

Molecular clock model



- 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...



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

Posterior Likelihood Prior

$$P\left(\begin{array}{c} \text{E} \\ \text{O} \\ \text{S} \end{array} \mid \text{ACAC...} \right) = \frac{P\left(\text{ACAC...} \mid \begin{array}{c} \text{E} \\ \text{O} \\ \text{S} \end{array}\right) \times P\left(\begin{array}{c} \text{E} \\ \text{O} \\ \text{S} \end{array}\right)}{P(\text{ACAC...})}$$

Marginal likelihood of the data

How can we find the posterior?

We want to calculate the posterior distribution

$$P(\text{EcoO restriction enzyme} \mid \text{ACAC...TCAC...ACAG...}) = \text{Normal Distribution}$$

But we cannot easily calculate the marginal likelihood

$$P(\text{ACAC...TCAC...ACAG...}) \longrightarrow ?$$

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

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

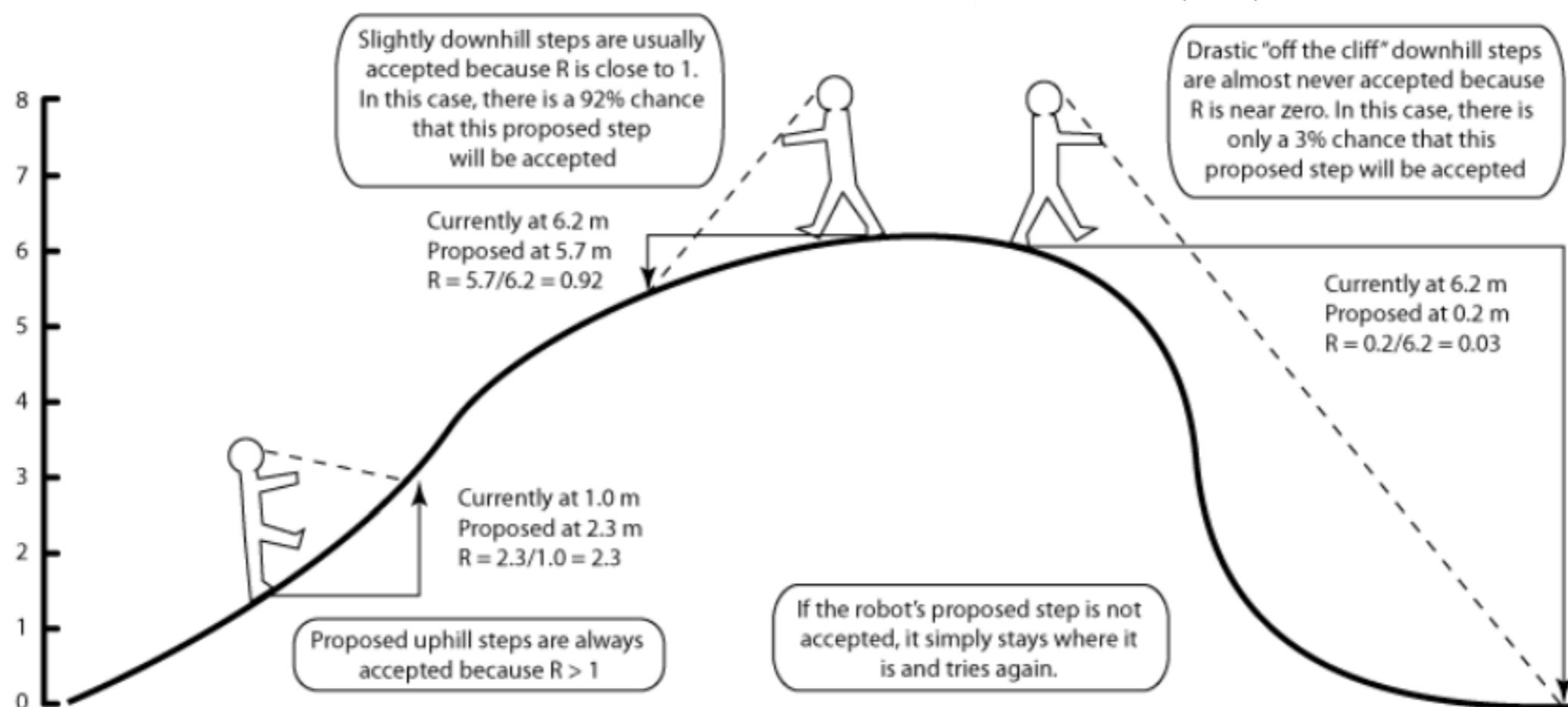
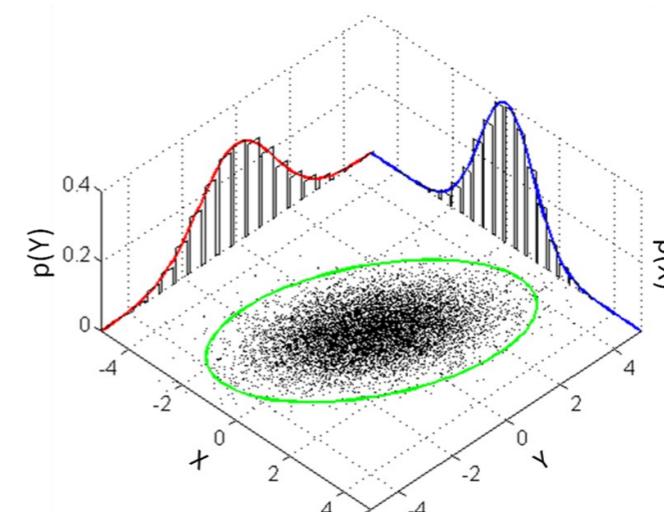
MCMC (Markov-chain Monte Carlo)

- 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!)

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

Marginal distributions

- We only have the joint posterior: $P(E \text{ } o_o \text{ } o_c \text{ } | \text{ACAC...}, \text{TCAC...}, \text{ACAG...})$
- But we want distributions for each of the parameters we are interested in → marginalize



Target distribution

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 are a part of the MCMC **algorithm**, not the model!

Tuning operators can help to improve efficiency, but should not change the results.

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)

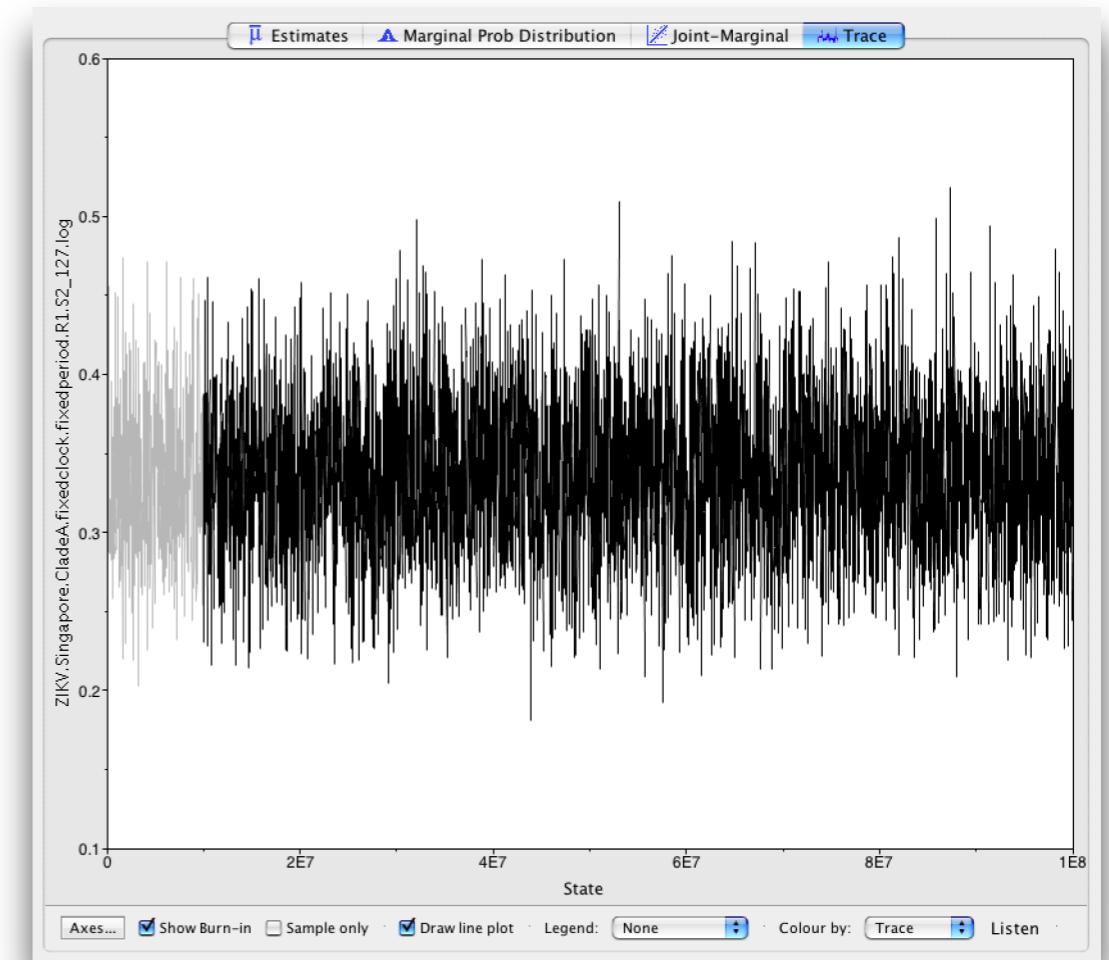
After

- 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

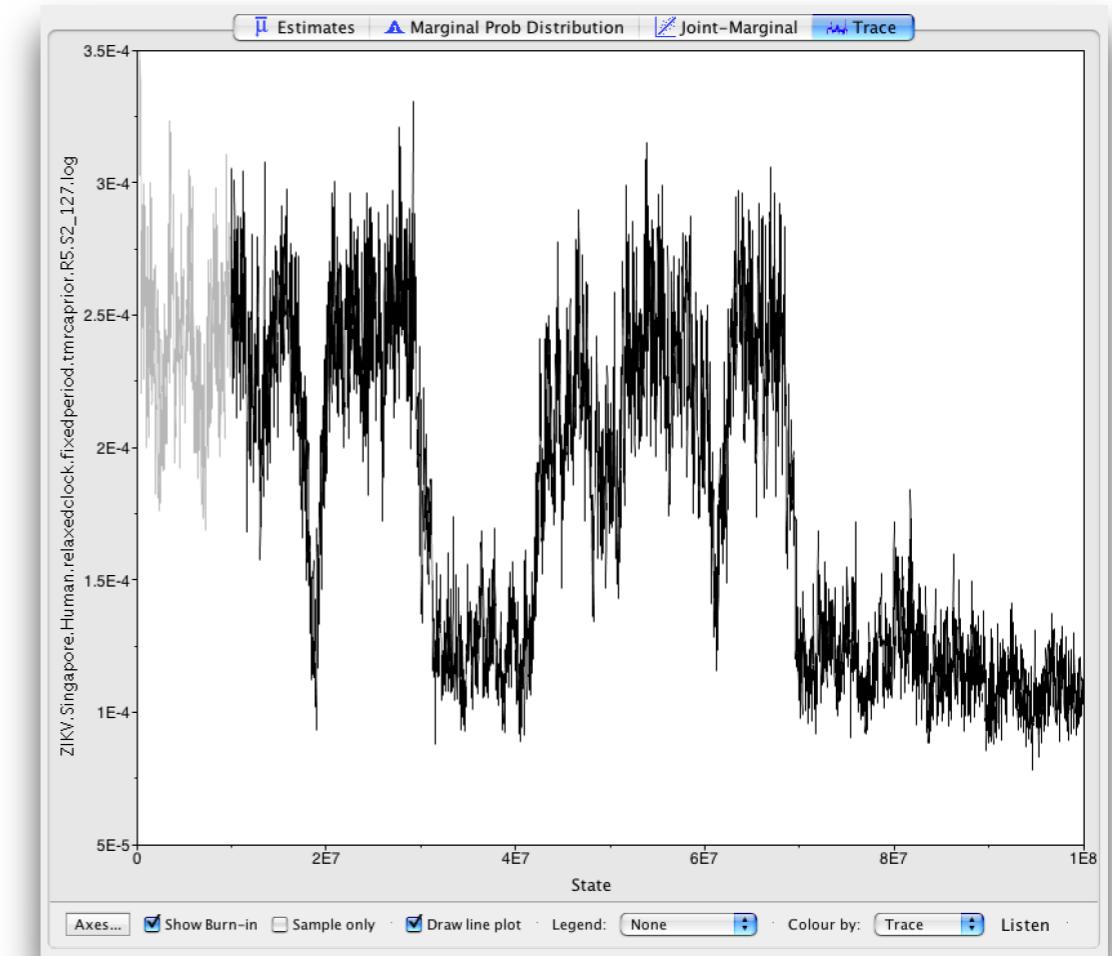
- The MCMC algorithm samples efficiently from high density areas of the posterior distribution
- We end up with a good approximation of the posterior distribution in finite time
- Appearance of white noise
- Everything is awesome!



Mixing well! 😊

Questions to ask...

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



Solutions

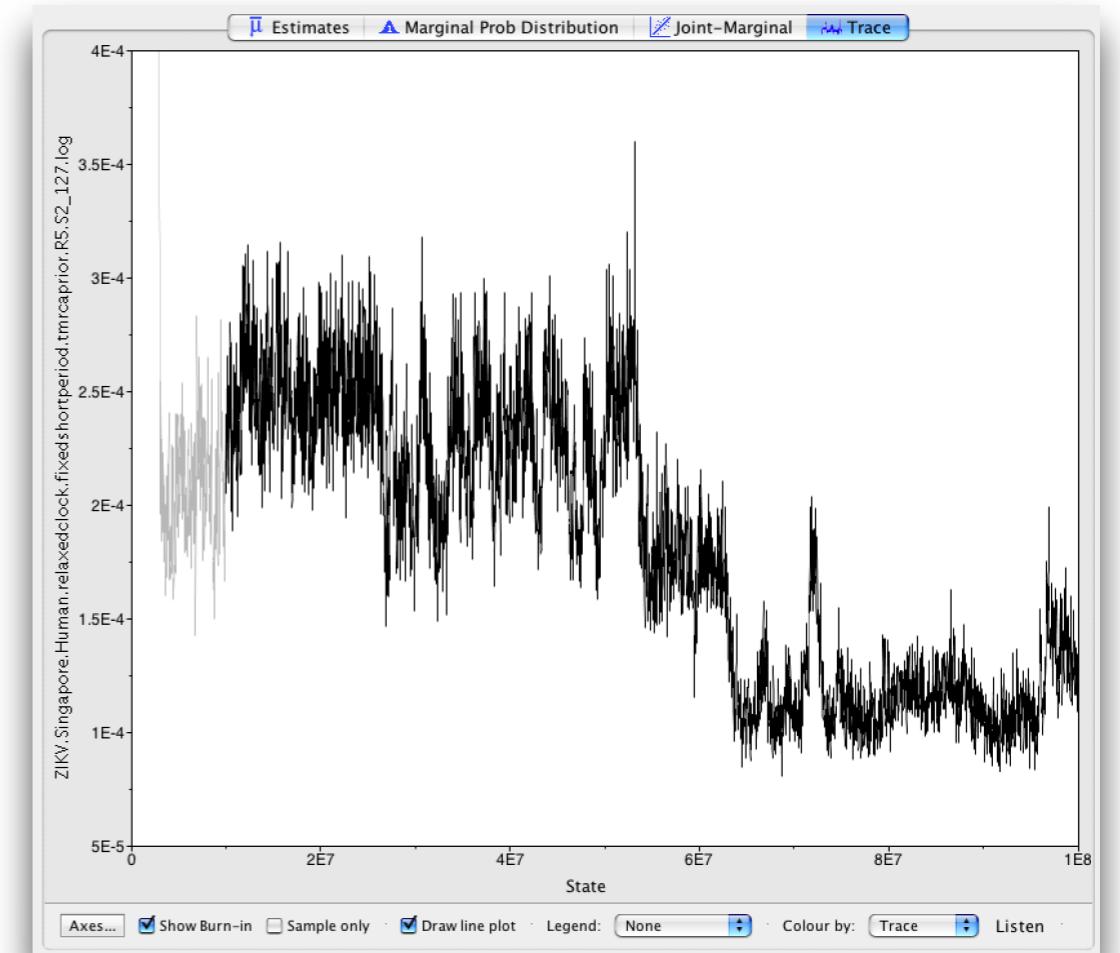
- MCMC gets stuck in some states for long times
- Tune operators to make better proposals

Not mixing! 😞

Questions to ask...

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

Solution: Run for longer



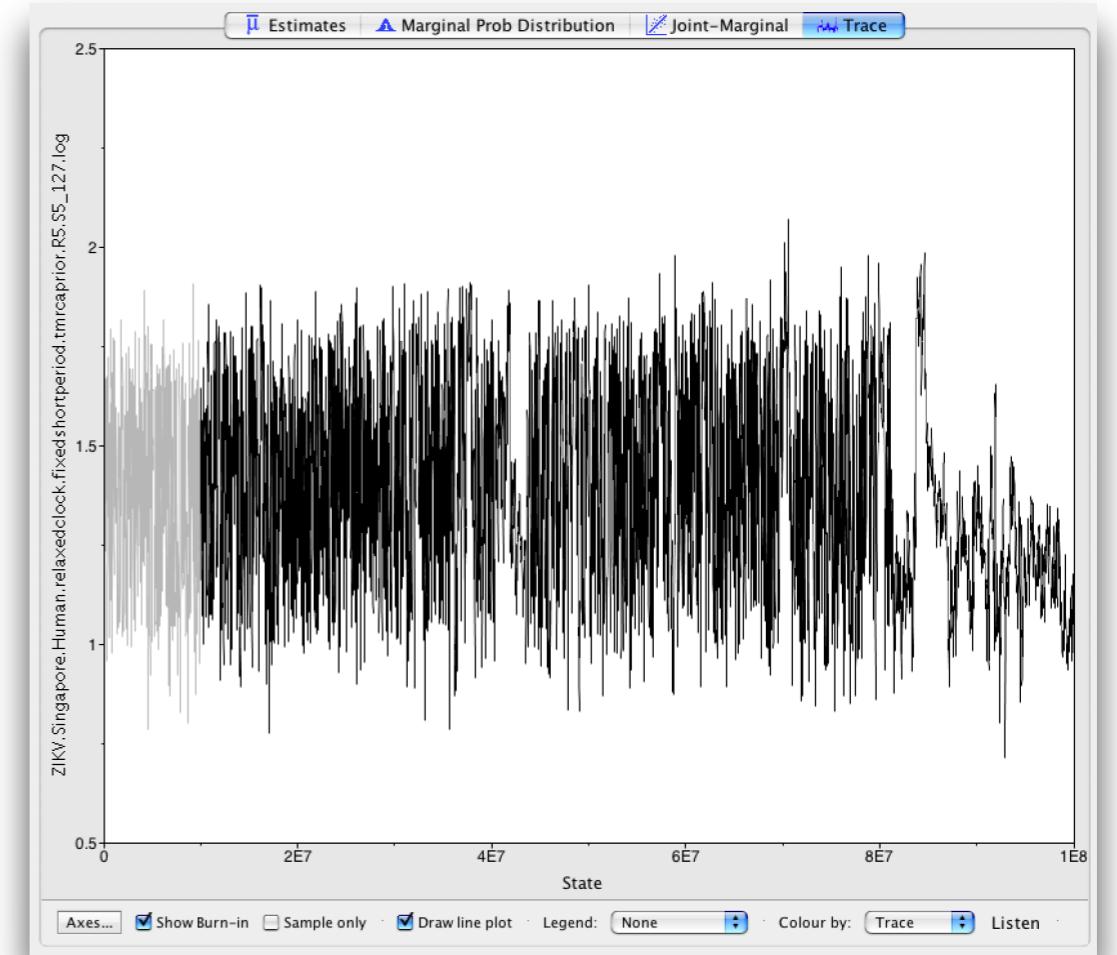
Not converged! 😞

Questions to ask...

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

Solution

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

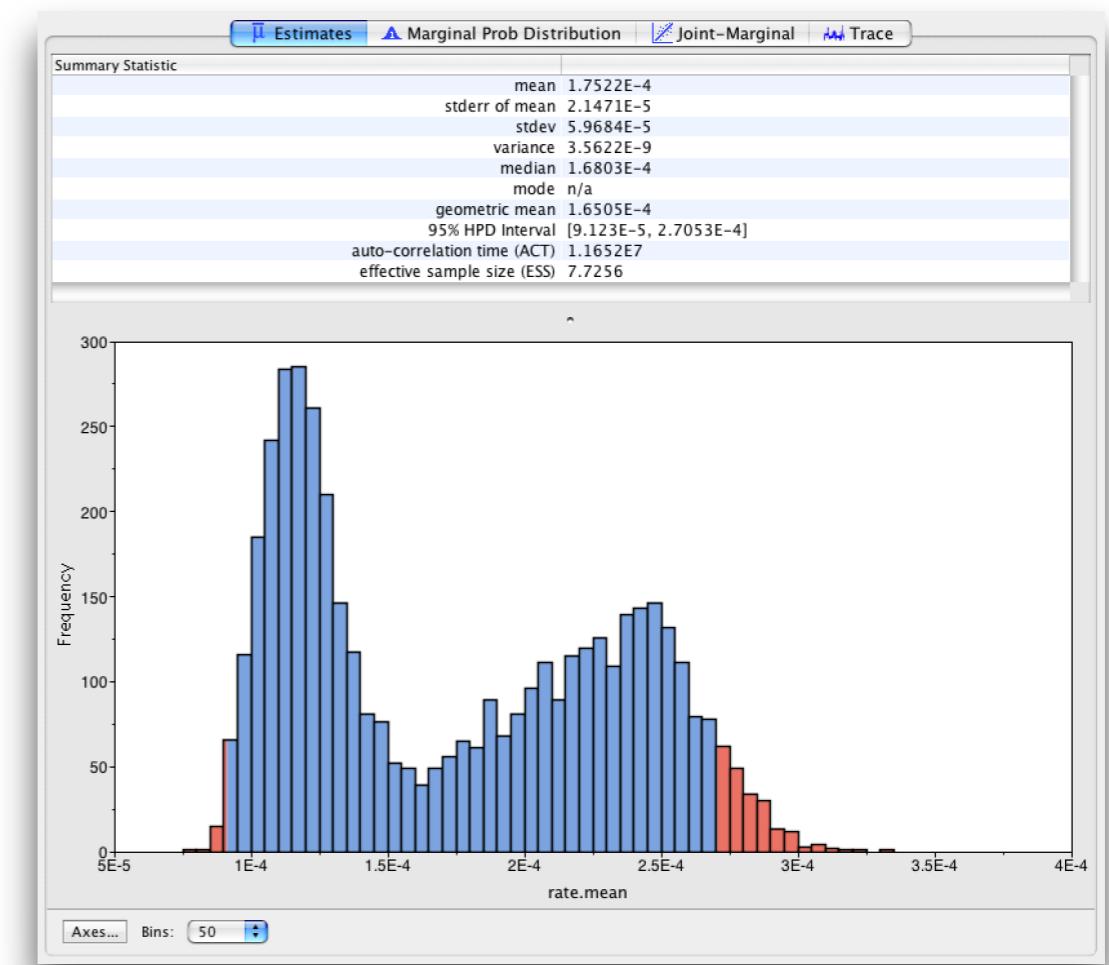


Still not converged! 😵

What if the answer is not what we wanted?

What is happening here?

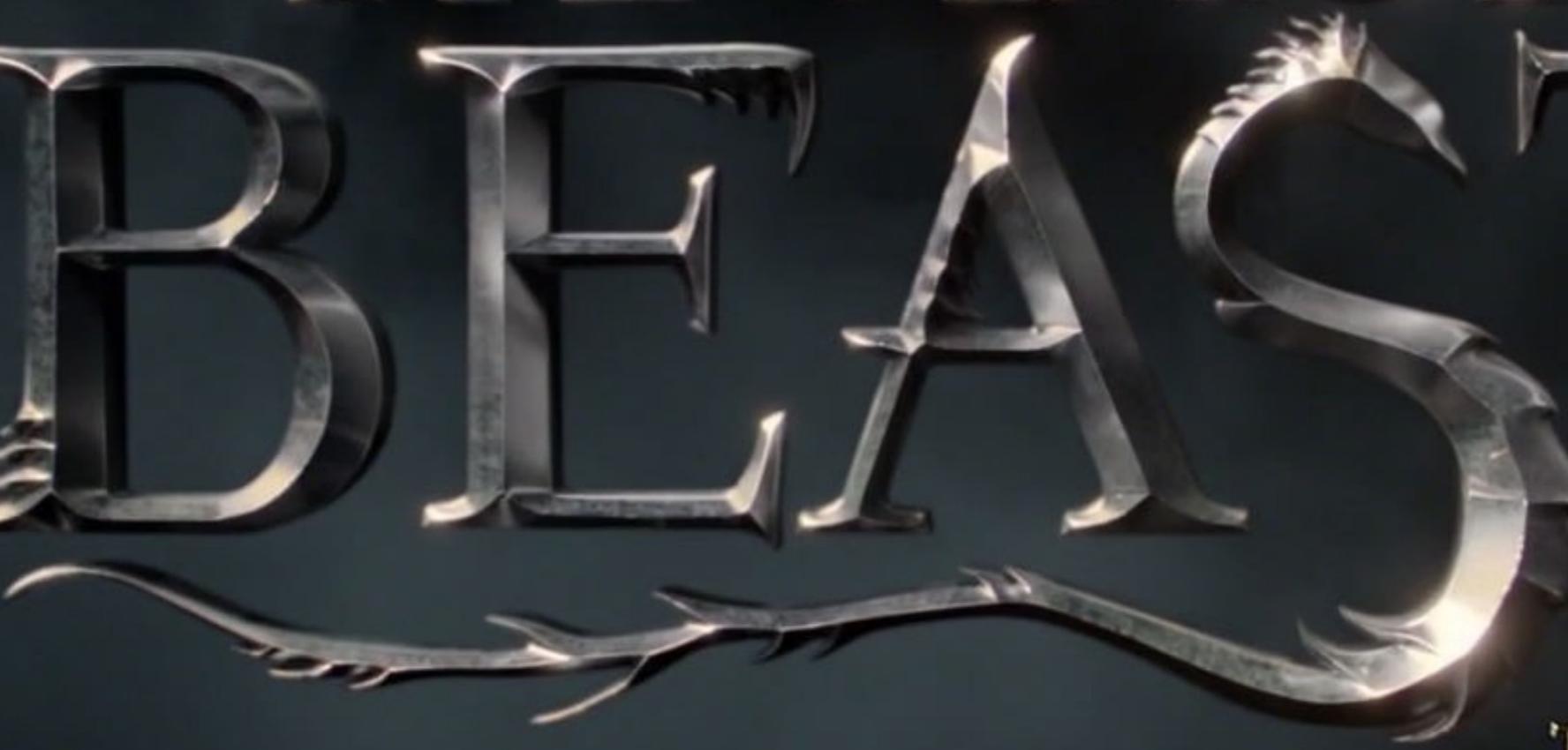
- If the chain converged and mixed well then this is due to the data and model choice
The model supports a bimodal posterior distribution
May not be the answer we wanted but it may be the truth
Should we change the model or parameterization?
- How would we know if another model fits better?



Is this a problem? 🤔

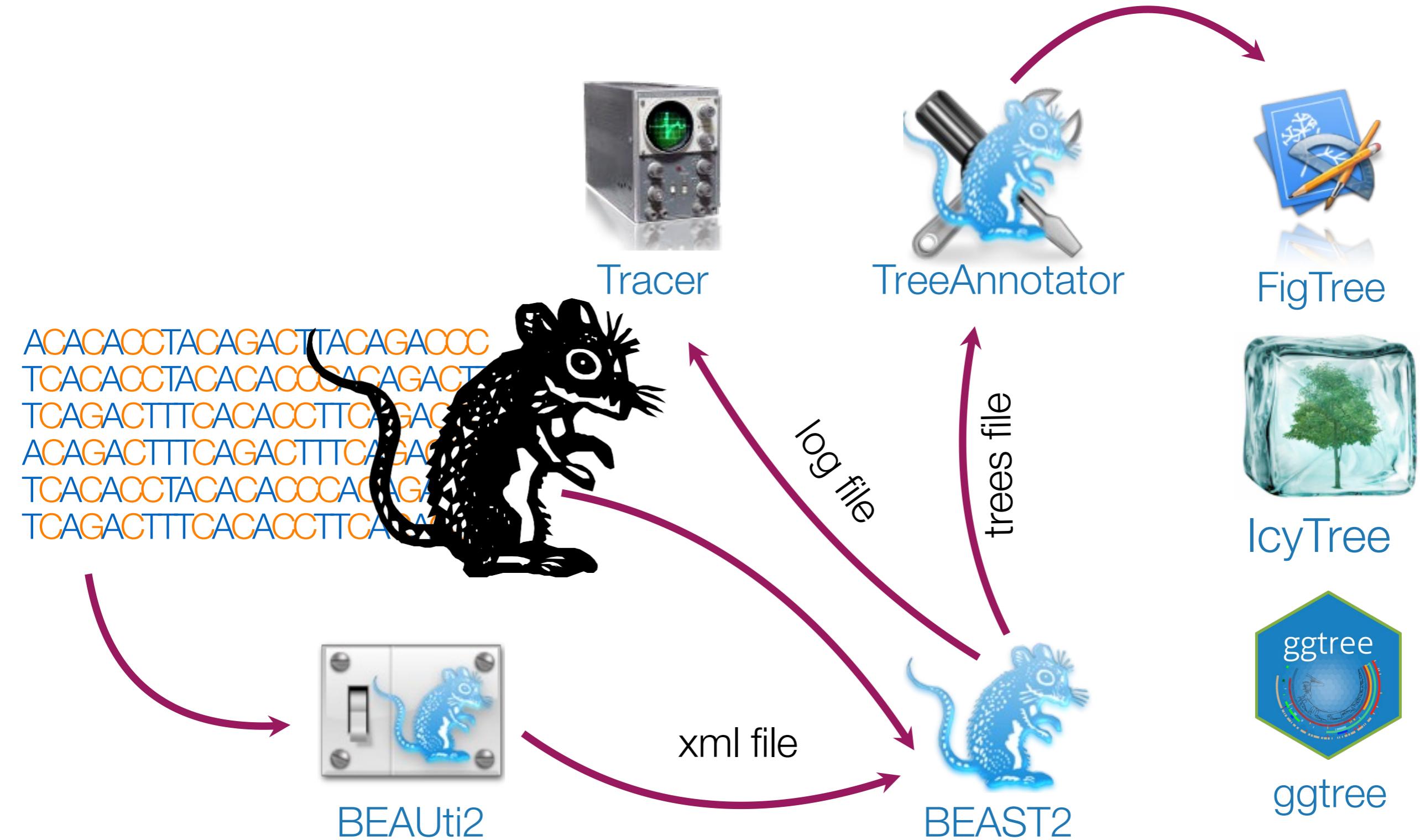
Solution: Be more open-minded

FANTASTIC BEASTS



AND WHERE
TO FIND THEM

BEAST2 workflow



BEAUti2 (<http://beast2.org>)



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	<input type="checkbox"/>
1stpos	primate-mtDNA	12	231	nucleotide	1stpos	clock	tree	<input type="checkbox"/>
2ndpos	primate-mtDNA	12	231	nucleotide	2ndpos	clock	tree	<input type="checkbox"/>
3rdpos	primate-mtDNA	12	231	nucleotide	3rdpos	clock	tree	<input type="checkbox"/>

+ - r Split

BEAUti2 (<http://beast2.org>)



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] $[-\infty, \infty]$ Calibrated Yule speciation process birth rate for t:3rdpos

▶ clockRate.c:clock Uniform initial = [1.0] $[-\infty, \infty]$ substitution rate of partition c:3rdpos

▶ gammaShape.s:1stpos Exponential initial = [1.0] $[-\infty, \infty]$ Prior on gamma shape for partition s:1stpos

▶ gammaShape.s:2ndpos Exponential initial = [1.0] $[-\infty, \infty]$ Prior on gamma shape for partition s:2ndpos

▶ gammaShape.s:3rdpos Exponential initial = [1.0] $[-\infty, \infty]$ Prior on gamma shape for partition s:3rdpos

▶ gammaShape.s:noncoding Exponential initial = [1.0] $[-\infty, \infty]$ Prior on gamma shape for partition s:noncoding

▶ kappa.s:1stpos Log Normal initial = [2.0] $[0.0, \infty]$ HKY transition-transversion parameter of partition s:1stpos

▶ kappa.s:2ndpos Log Normal initial = [2.0] $[0.0, \infty]$ HKY transition-transversion parameter of partition s:2ndpos

▶ kappa.s:3rdpos Log Normal initial = [2.0] $[0.0, \infty]$ HKY transition-transversion parameter of partition s:3rdpos

▶ kappa.s:noncoding Log Normal initial = [2.0] $[0.0, \infty]$ HKY transition-transversion parameter of partition s:noncoding

▼ human-chimp.prior Normal monophyletic

Mean: 6.0 estimate

Sigma: 0.5 estimate

Offset: 0.0

0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.00
4.00 4.50 5.00 5.50 6.00 6.50 7.00 7.50 mean 6.00
2.5% Quantile 5.02
5% Quantile 5.18
Median 6.00
95% Quantile 6.82
97.5% Quantile 6.98

Tipsonly

Use Originate



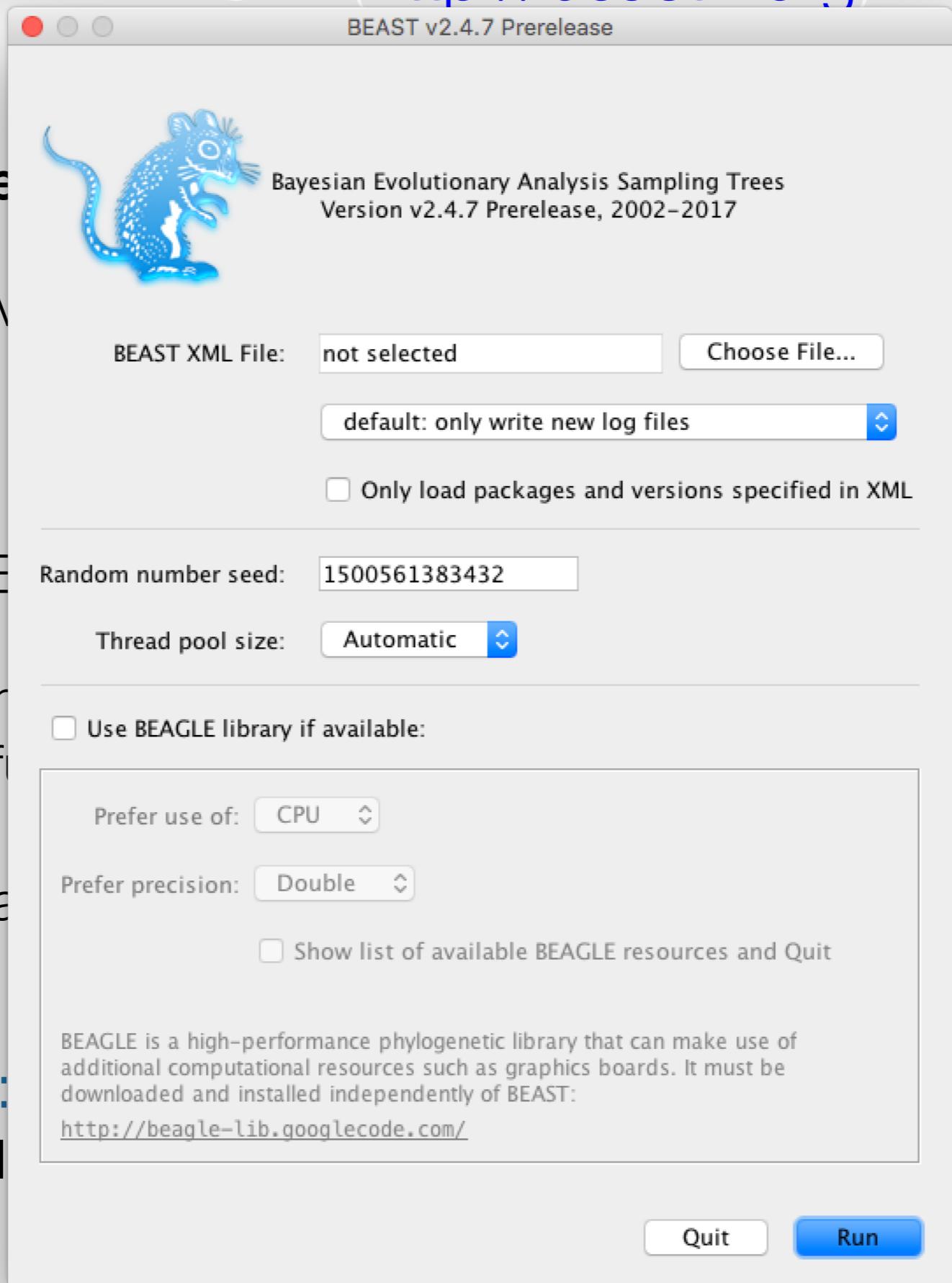
BEAUti2 (<http://beast2.org>)

Primates_long.xml UNREGISTERED

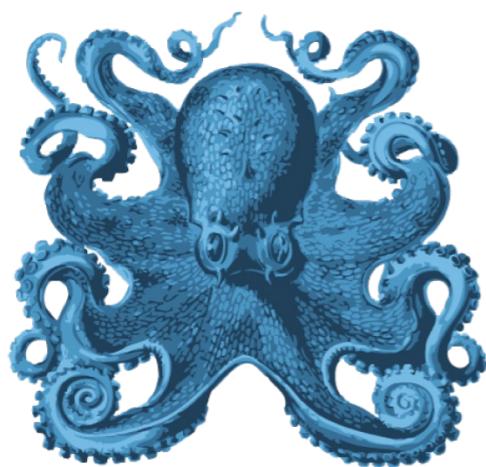
```
39
40 <run id="mcmc" spec="MCMC" chainLength="2500000">
41   <state id="state" storeEvery="5000">
42     <tree id="Tree.t:tree" name="stateNode">
43       <taxonset id="TaxonSet.noncoding" spec="TaxonSet">
44         <alignment id="noncoding" spec="FilteredAlignment" filter="1,458-659,897-898">
45           <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>
83     </distribution>
84   </distribution>
85 </run>
```

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

BEAST2 (<http://beast2.org>)



BEAST v.1 versus BEAST v.2



BEAST

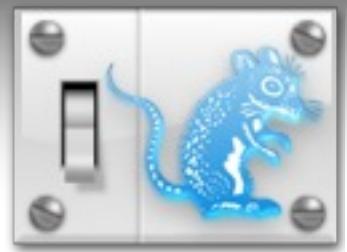
Bayesian Evolutionary Analysis Sampling Trees



Beast2

Bayesian evolutionary analysis by sampling trees

BEAST2 packages



Ind... es
Pack... BE...
Pack... Phy...
mo...
Inst...

BEAST 2 Package Manager

List of available packages for BEAST v2.4.*

Name	Installed	Latest	Dependencies	Link	Detail
BEAST	2.4.7	2.4.7			BEAST core
bacter	1.2.1	1.2.1			Bacterial ARG inference.
BASTA		2.3.1			Bayesian structured coalescent approximation
bdmm	0.2.0	0.2.0	MultiTypeTree		pre-release of multitype birth-death model (aka birth-... birth death skyline - handles serially sampled tips, piec...
BDSKY	1.3.3	1.3.3			
BEAST_CLASSIC	1.3.0	1.3.0	BEASTLabs		BEAST classes ported from BEAST 1 in wrappers
BEASTLabs	1.7.0	1.7.1			BEAST utilities, such as Script, multi monophyletic c...
BEASTShell		1.3.0			BEAST Shell - BeanShell scripting for BEAST
BEASTvntr		0.1.1			Variable Number of Tandem Repeat data, such as micr...
bModelTest	1.0.4	1.0.4	BEASTLabs		Bayesian model test for nucleotide subst models, gamm...
CA		1.2.1			CladeAge aPackage for fossil calibrations
DENIM		0.3.0			Divergence Estimation Notwithstanding ILS and Migration
Epilnf		5.0.1	SA		Inference of epidemic trajectories
GEO_SPHERE		1.1.2	BEASTLabs		Whole world phylogeography
Mascot		0.0.2			Marginal approximation of the structured coalescent
MASTER		5.1.1			Stochastic population dynamics simulation
MGSM		0.2.1			Multi-gamma and relaxed gamma site models
MM		1.0.5			Enables models of morphological character evolution
MODEL_SELECTION		1.3.4	BEASTLabs		Select models through path sampling/stepping stone an...

Latest [Install/Upgrade](#) [Uninstall](#) [Package repositories](#) [Close](#) [?](#)

Tracer (<http://beast.community>)



- Analyze log files from BEAST2 runs
- Check mixing, ESS, parameter correlations
- Overview of posterior parameter estimates
- Tracer is a diagnostic tool **only** — perform final analyses in a statistical package like R!
- Comparisons of several analyses

Input:

- log file

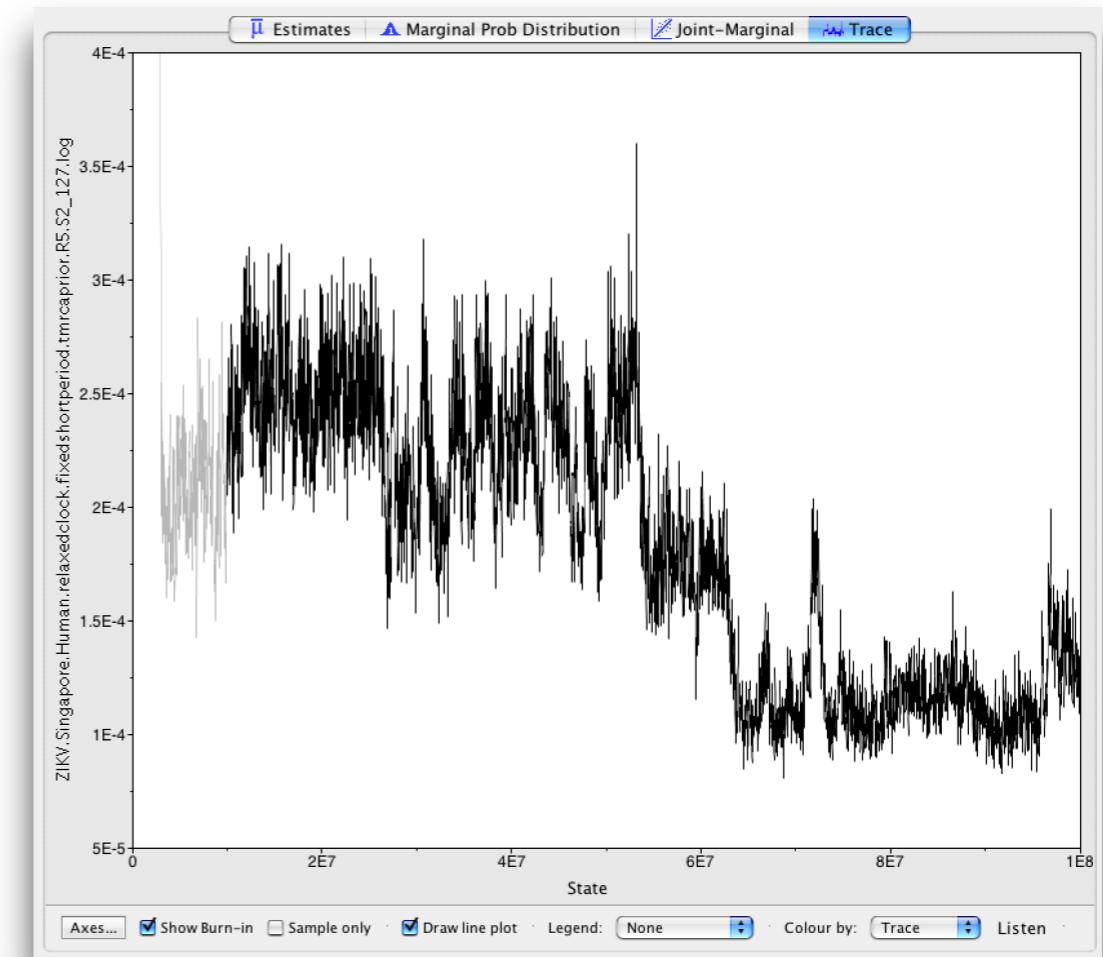
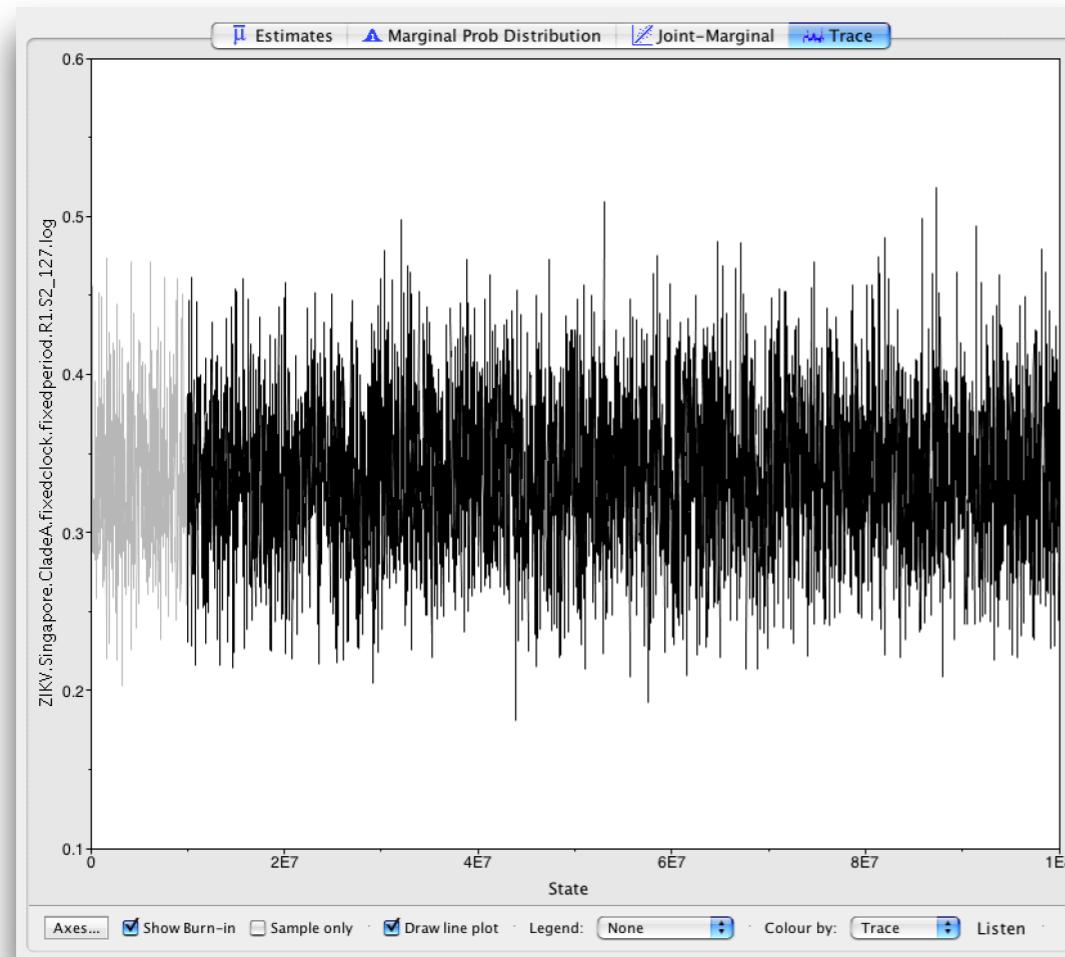
Output:

- Gain insight

Tracer (<http://beast.community>)

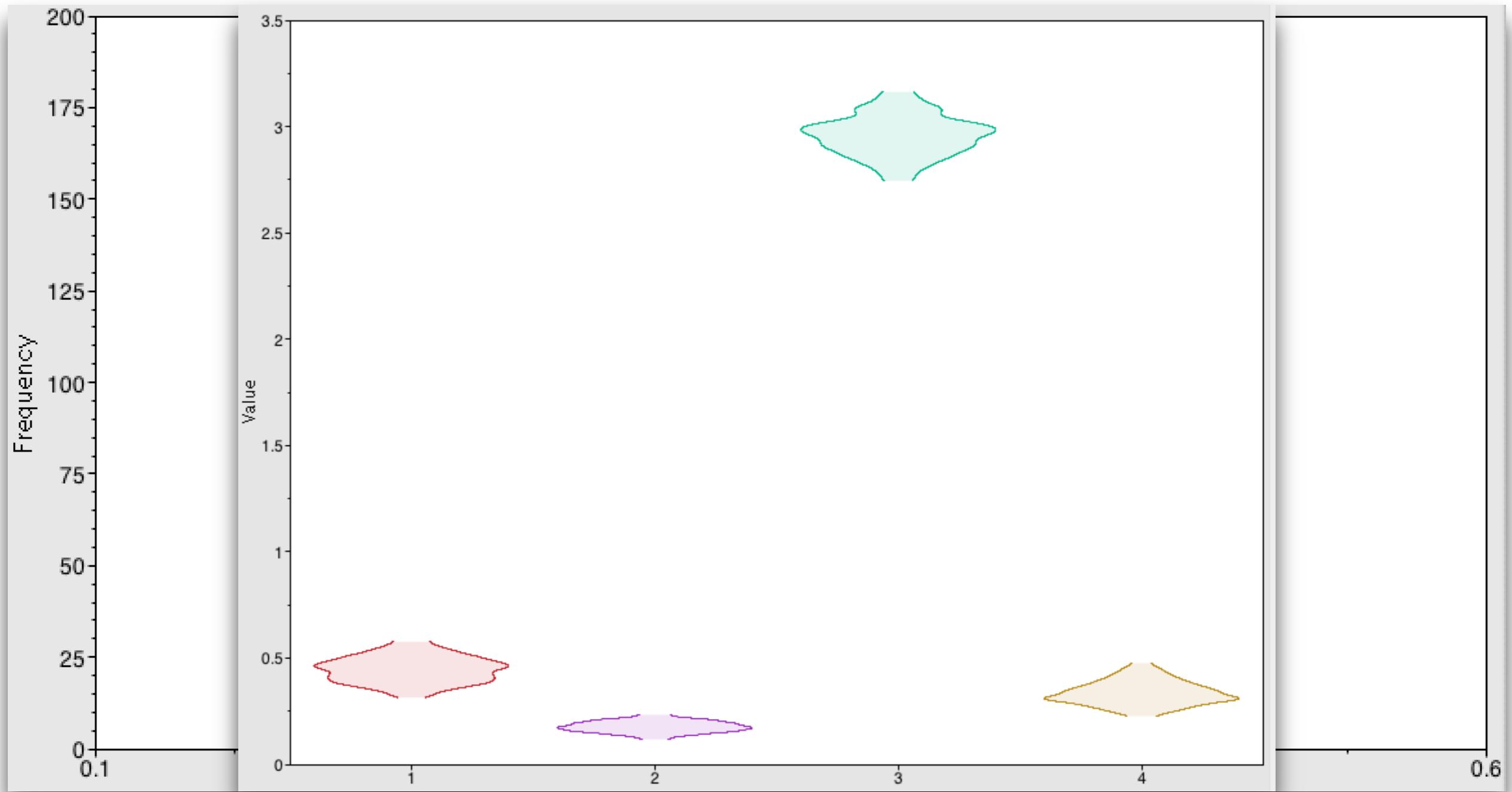


Look at the chains first!



Mixing well! 😊

Not mixing! 😢



TreeAnnotator (Included with BEAST2)



TreeAnnotator v2.4.6

Burnin percentage:

Posterior probability limit:

Target tree type:

Node heights:

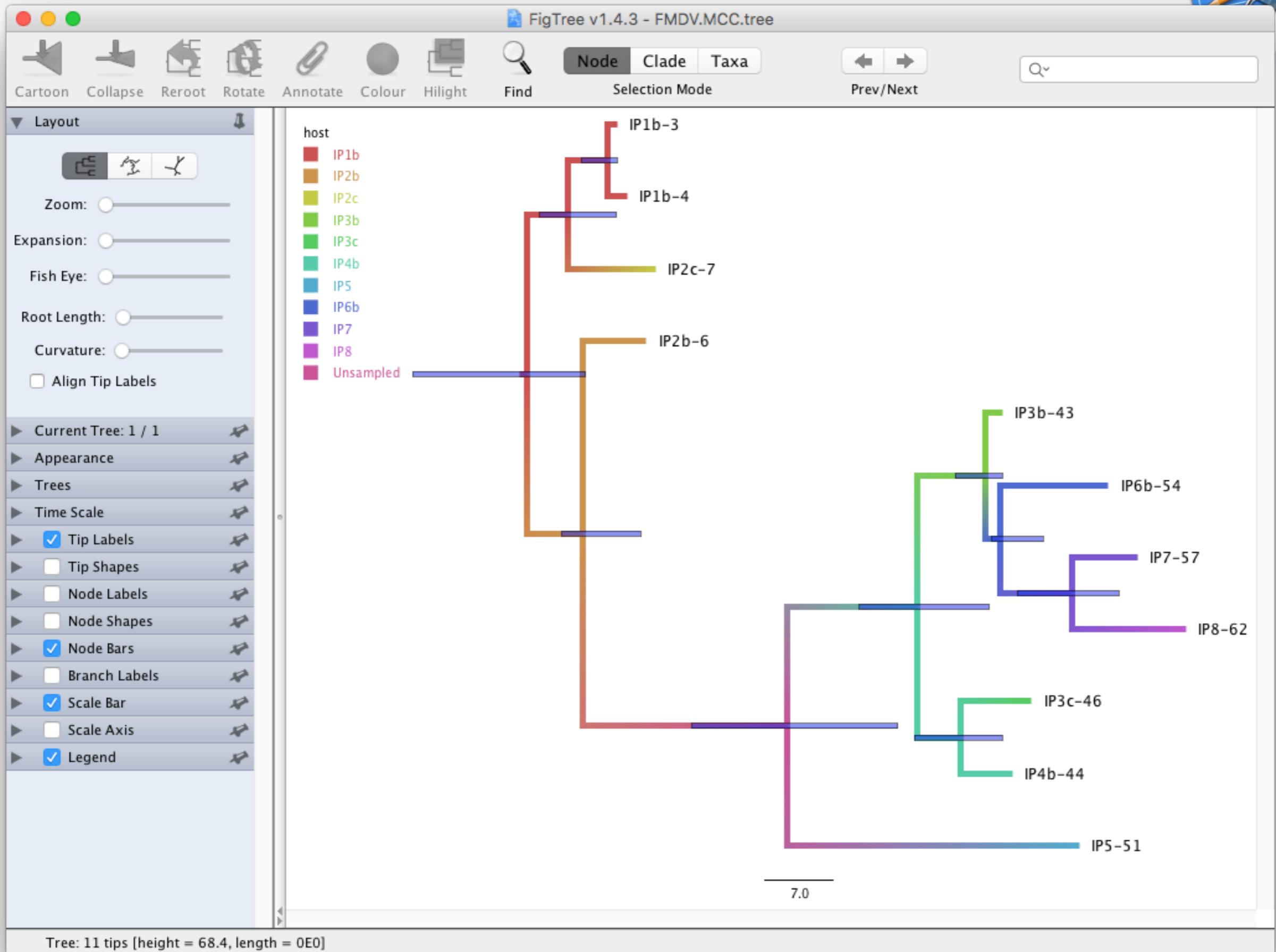
Target Tree File:

Input Tree File:

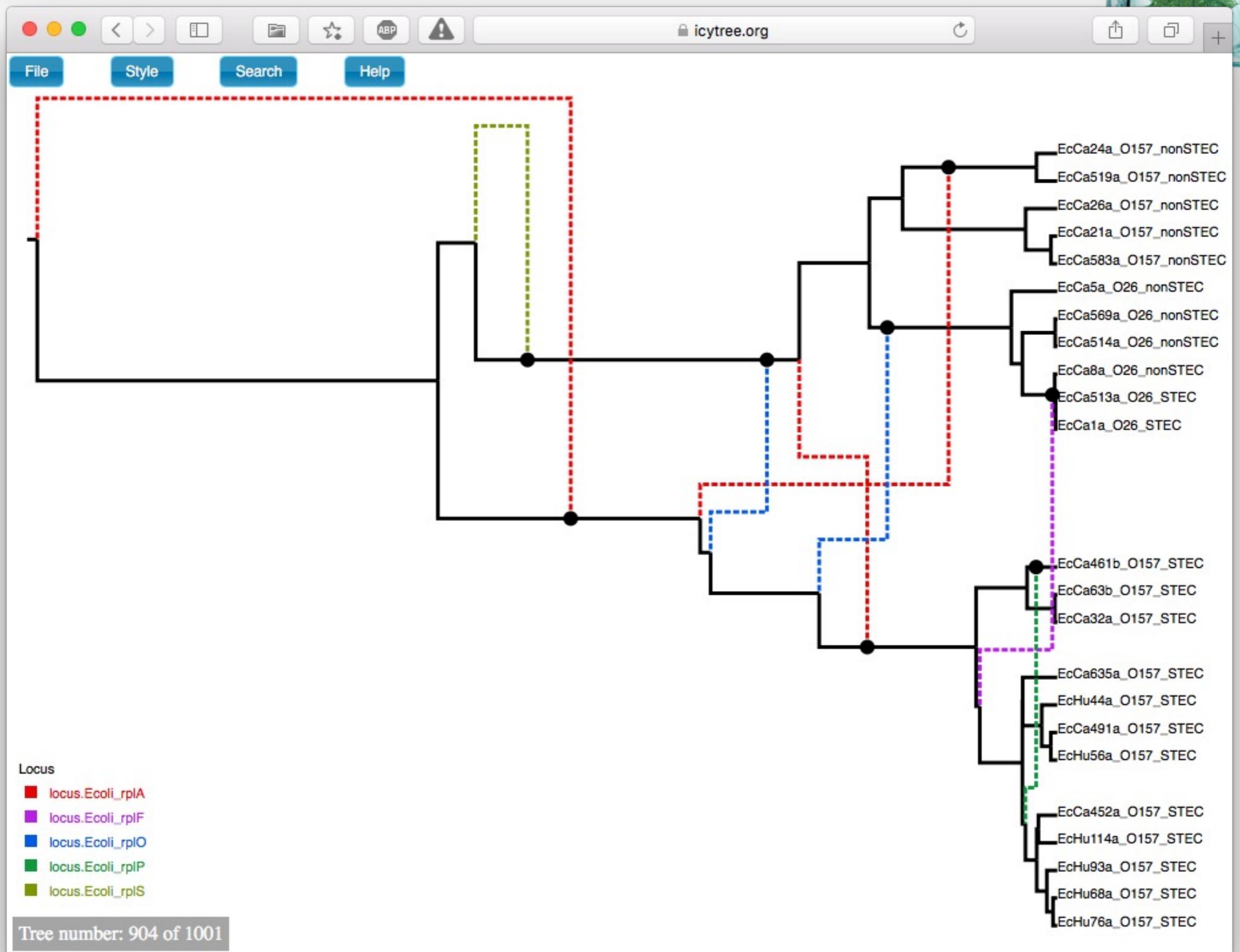
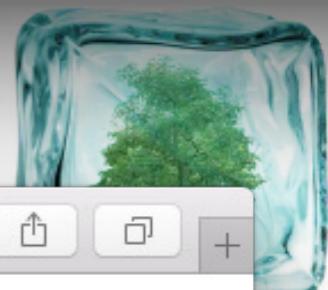
Output File:

Low memory:

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



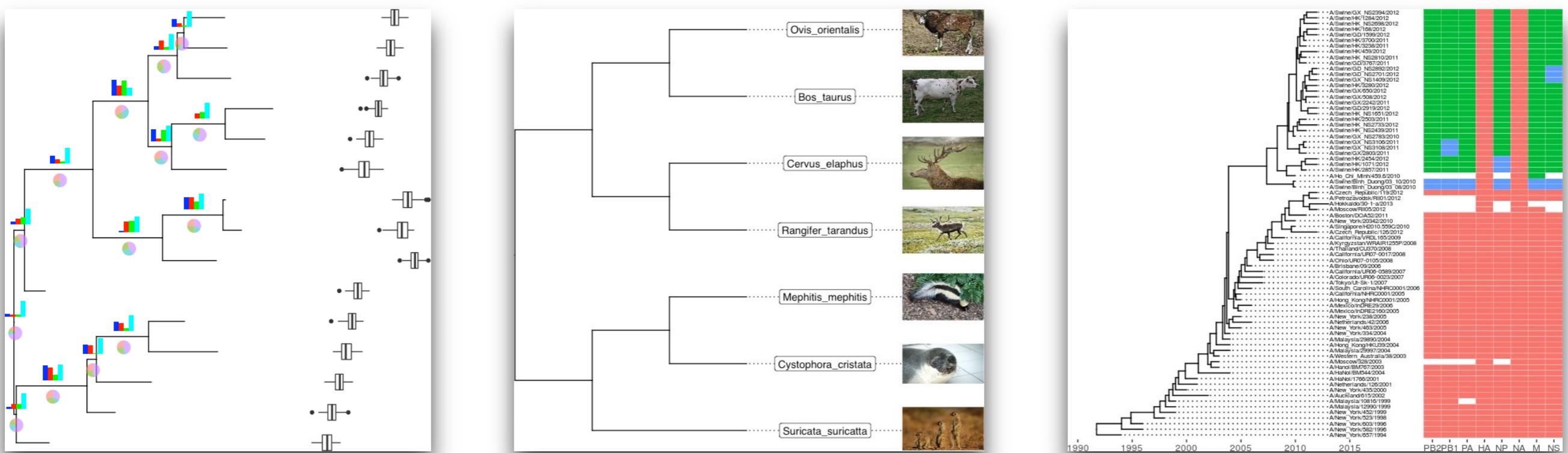
IcyTree (<https://icytree.org>)



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



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



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

