BEAST

Bayesian Evolutionary Analysis Sampling Trees

Xiang Ji
xji3@ucla.edu
University of California, Los Angeles
& Tulane University

http://beast.community/

Download and install the software

- BEAST: http://beast.community/installing
- BEAGLE: https://github.com/beagle-dev/beagle-lib
- Tracer: https://github.com/beast-dev/tracer/releases/tag/v1.7.1
- Tutorial link:
  - http://beast.community/phylodynamics_of_seasonal_influenza
- Tutorial Material:
  - http://beast.community/tutorials/phylodynamics_of_seasonal_influenza/files/NewYork.HA.2000-2003.nex
  - http://beast.community/tutorials/phylodynamics_of_seasonal_influenza/files/Long_Run_H3N2_SkyGrid.zip

Phylodynamics: unifying principle

“ Rapidly evolving pathogens are unique in that their ecological and evolutionary dynamics occur on the same timescale and can therefore potentially interact. ”

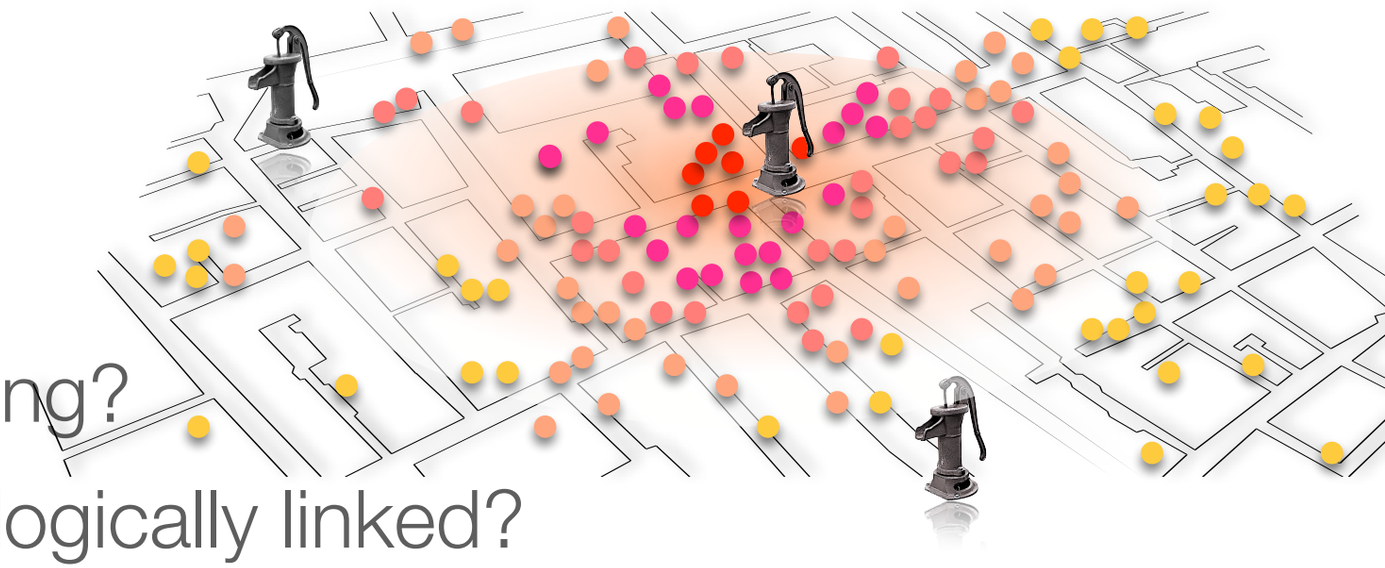
Pybus & Rambaut (2009) Nat. Rev. Genetics 10:540-50

Fundamental Phylodynamic Questions

- How genetically diverse is a pathogen population?
- How does pathogen genetic diversity vary through time and space?
- How do pathogen genomes change through time?
- What are the effects of pathogen genetic diversity on virulence, transmissibility, resistance to treatment, etc.

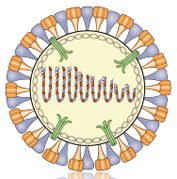
Specific questions

- When did a 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?
- Of how many strains is the epidemic composed?
- Are strains associated with particular transmission routes?
- What adaptations has it accrued?



SAMPLES FROM INFECTED INDIVIDUALS

Clinical Virology
Genomics



PATHOGEN GENOME SEQUENCES

Alignment Methods

ALIGNMENT

1 *Sequence Evolution Models*
Phylogenetic Methods

Recombination Analysis 5

PHYLOGENY

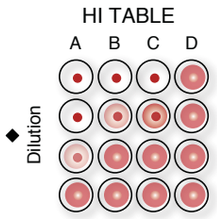
(time scale = genetic distance)

2 *Molecular Clocks*

TIMED EVOLUTIONARY HISTORY

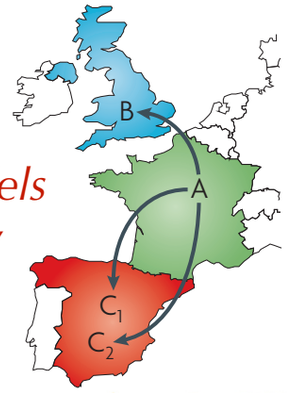
(time scale = years)

3 *Coalescent Theory*
Birth-death models



Adaptation Analysis 6

4 *Phylogeographic models*
& coalescent theory

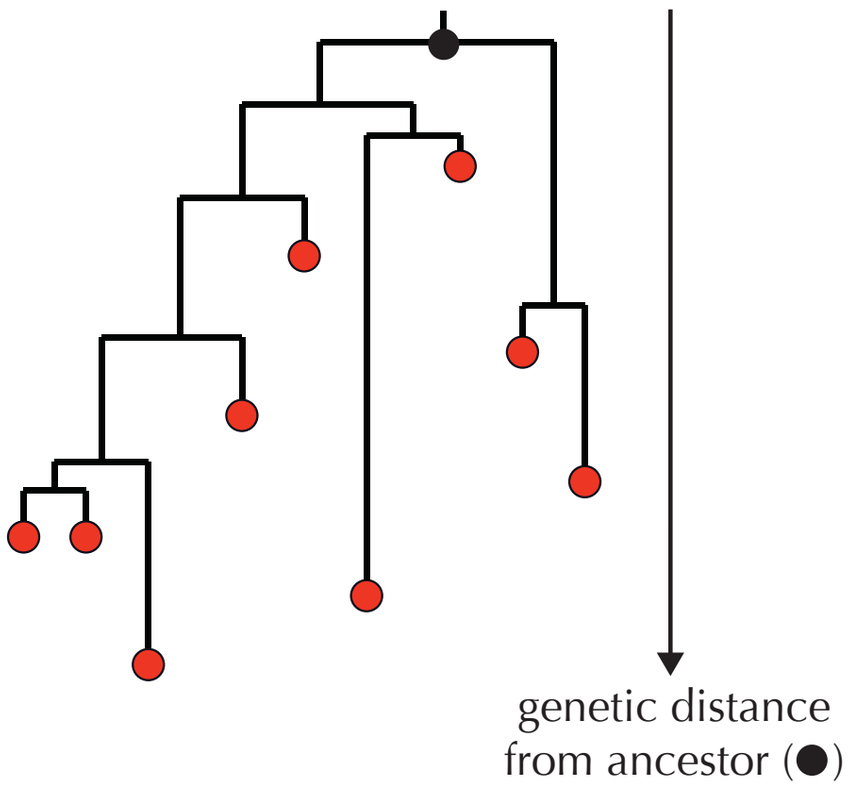


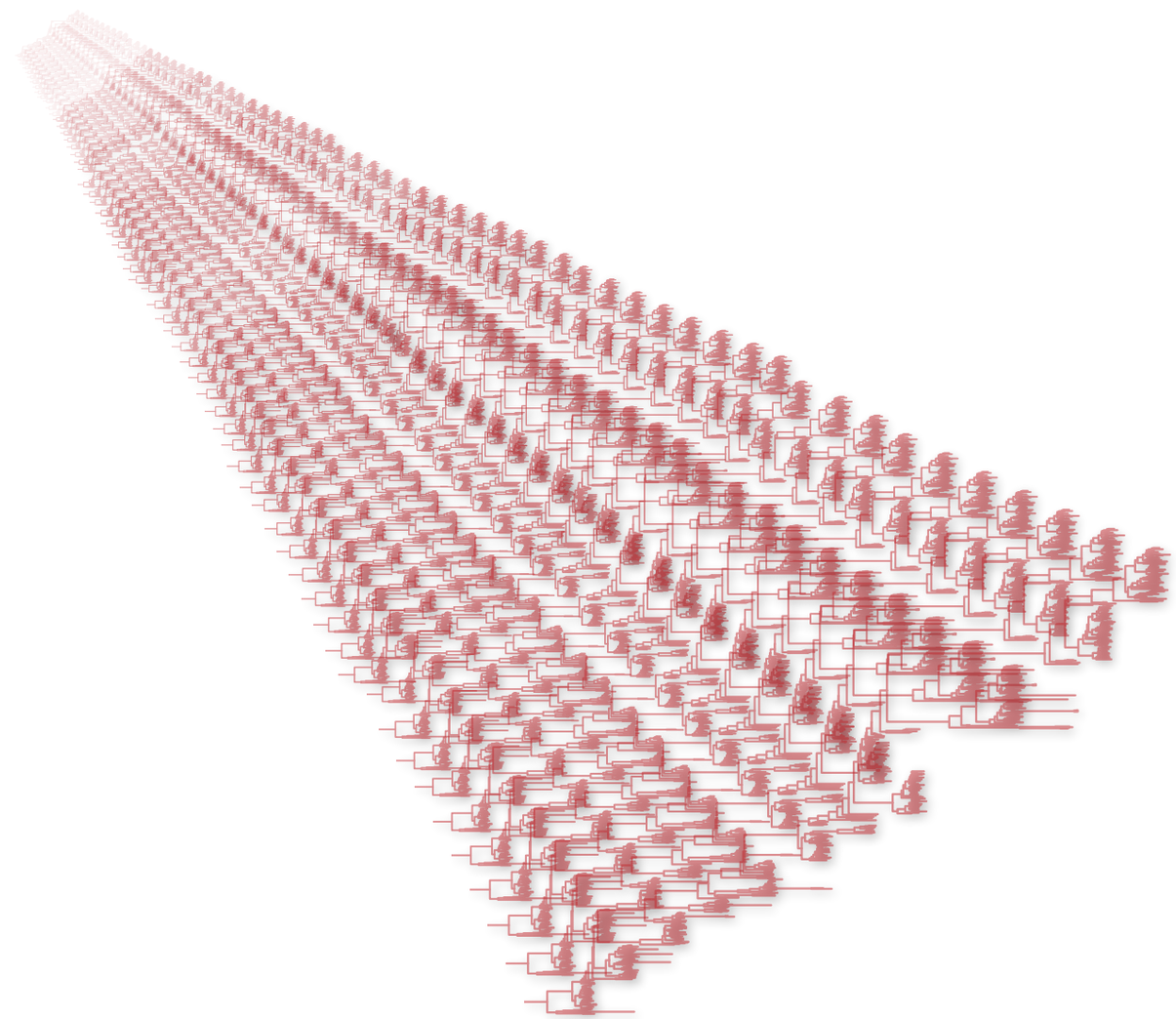
POPULATION DYNAMICS

ADAPTATION DYNAMICS

SPATIAL & HOST DYNAMICS

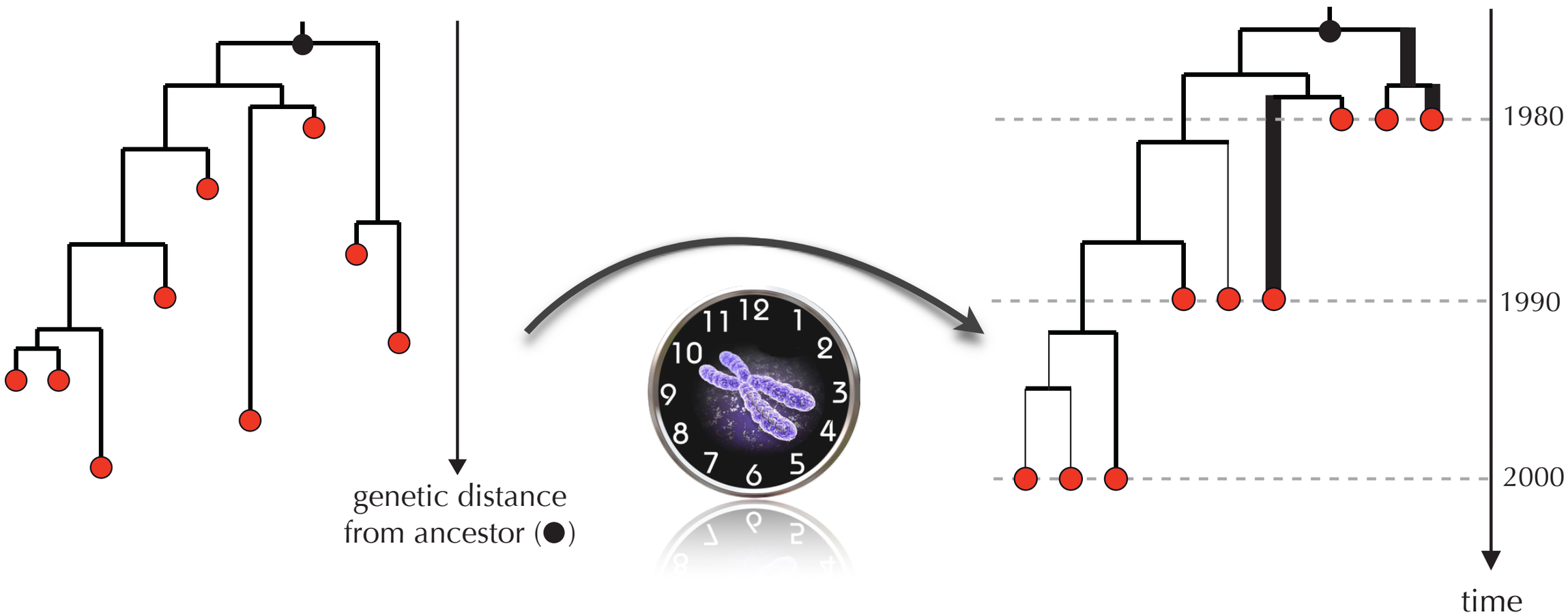
Phylodynamics toolbox





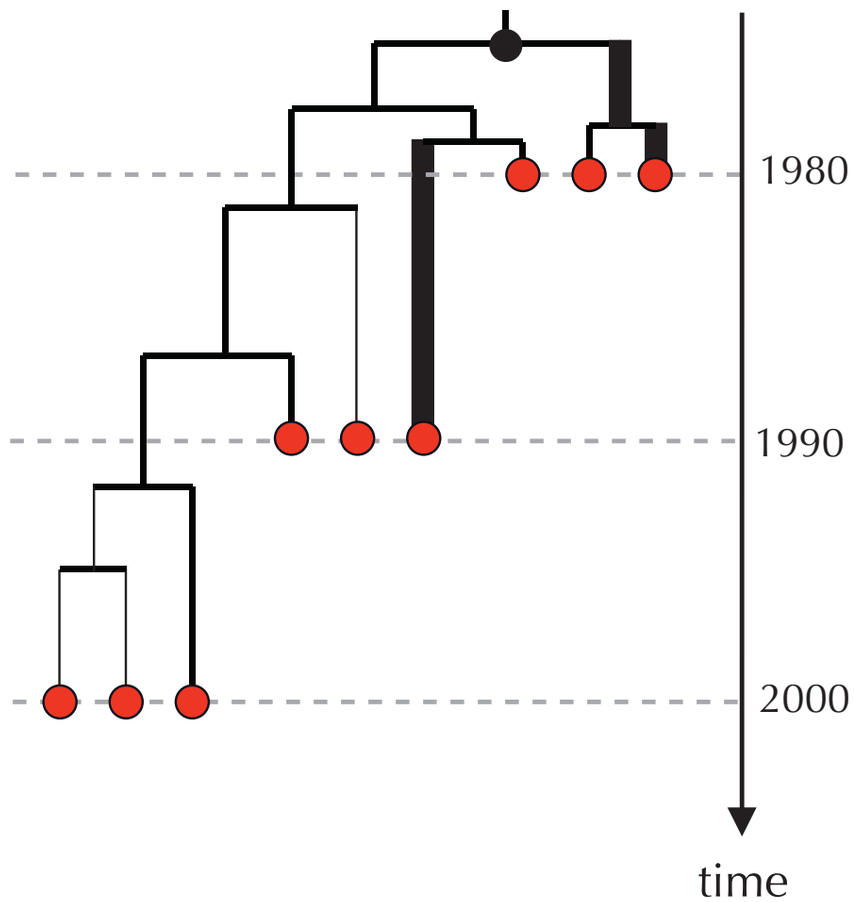
1 Phylogenetics

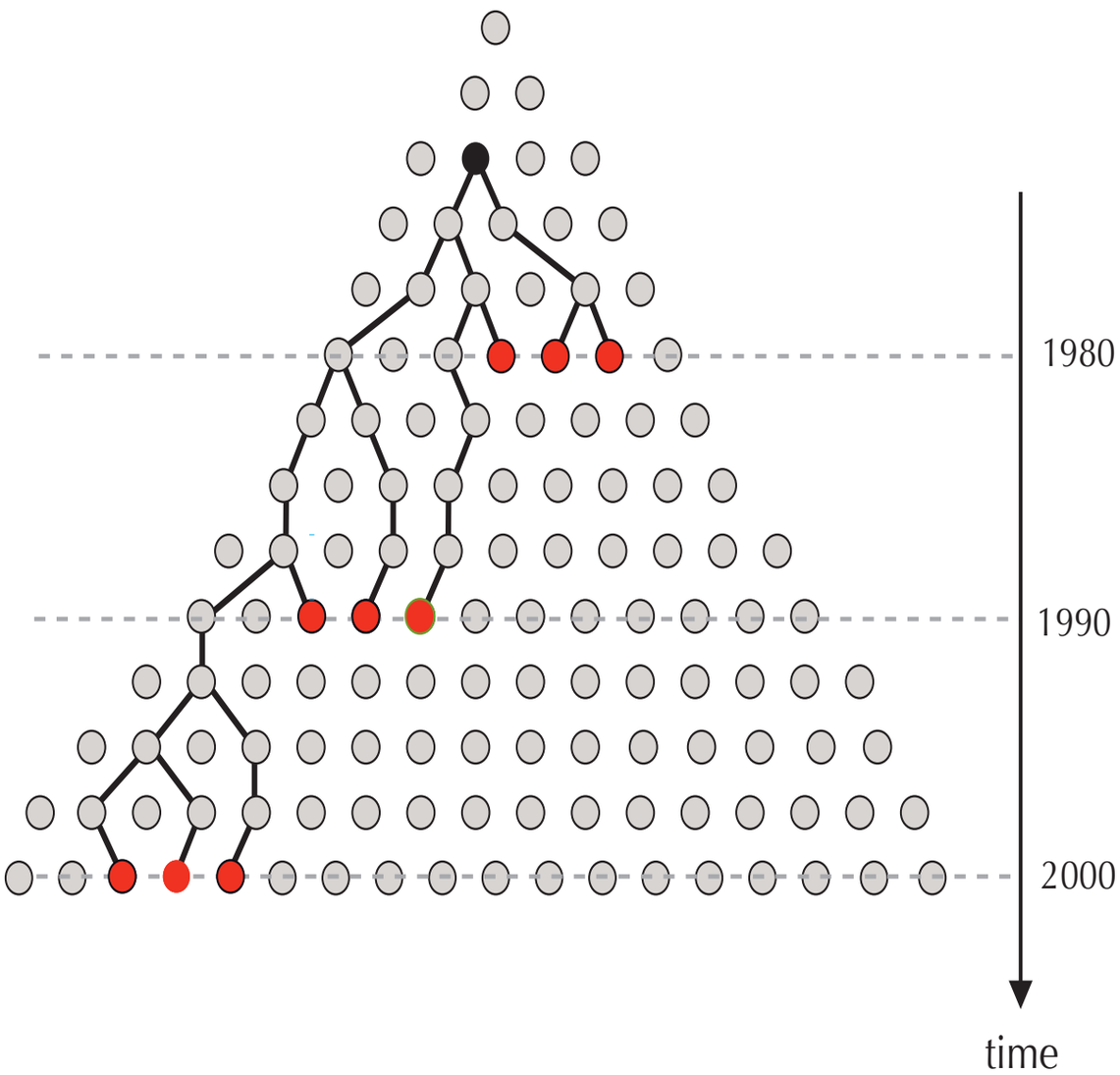
Phylodynamics toolbox



2 Molecular clocks

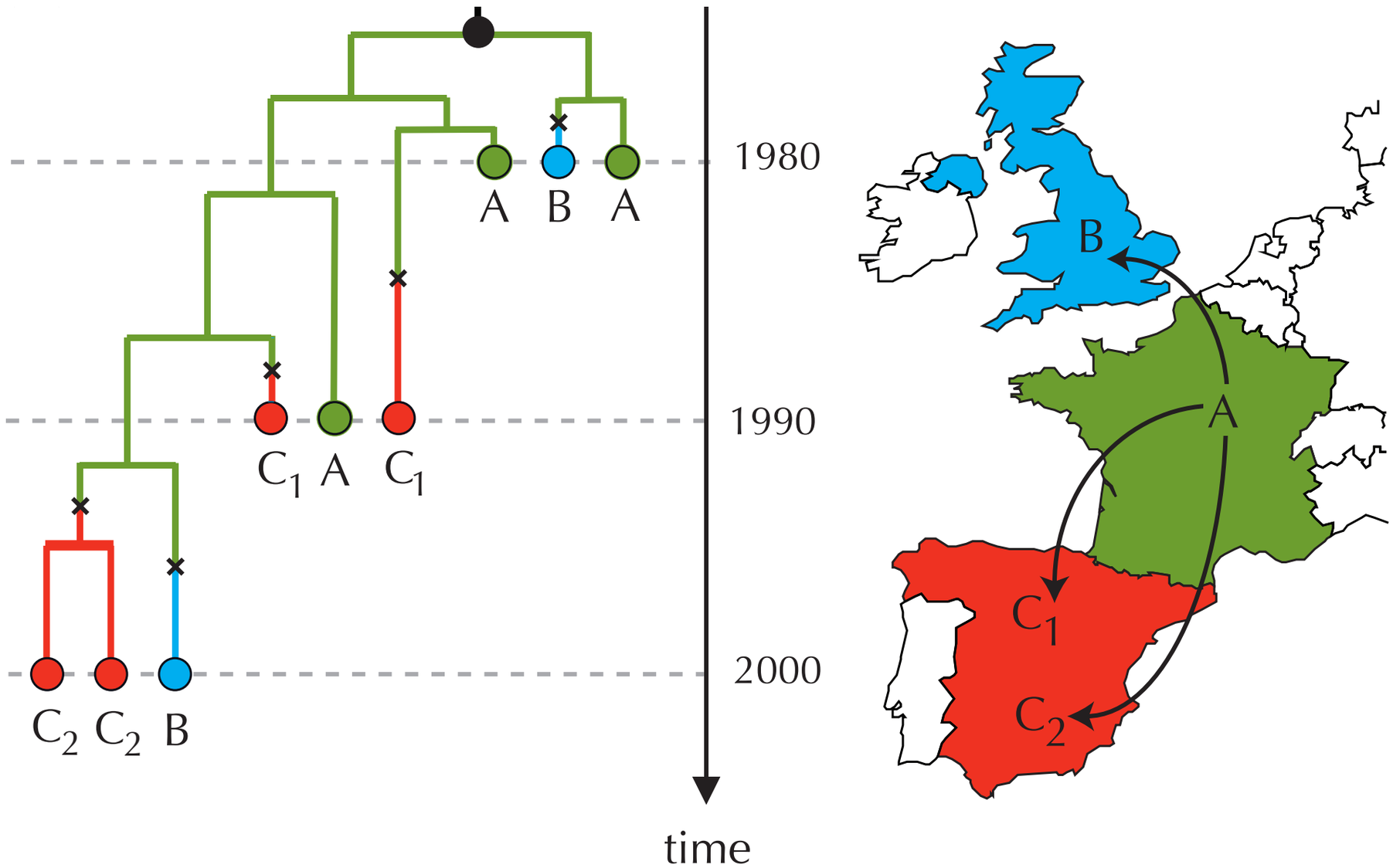
Phylodynamics toolbox





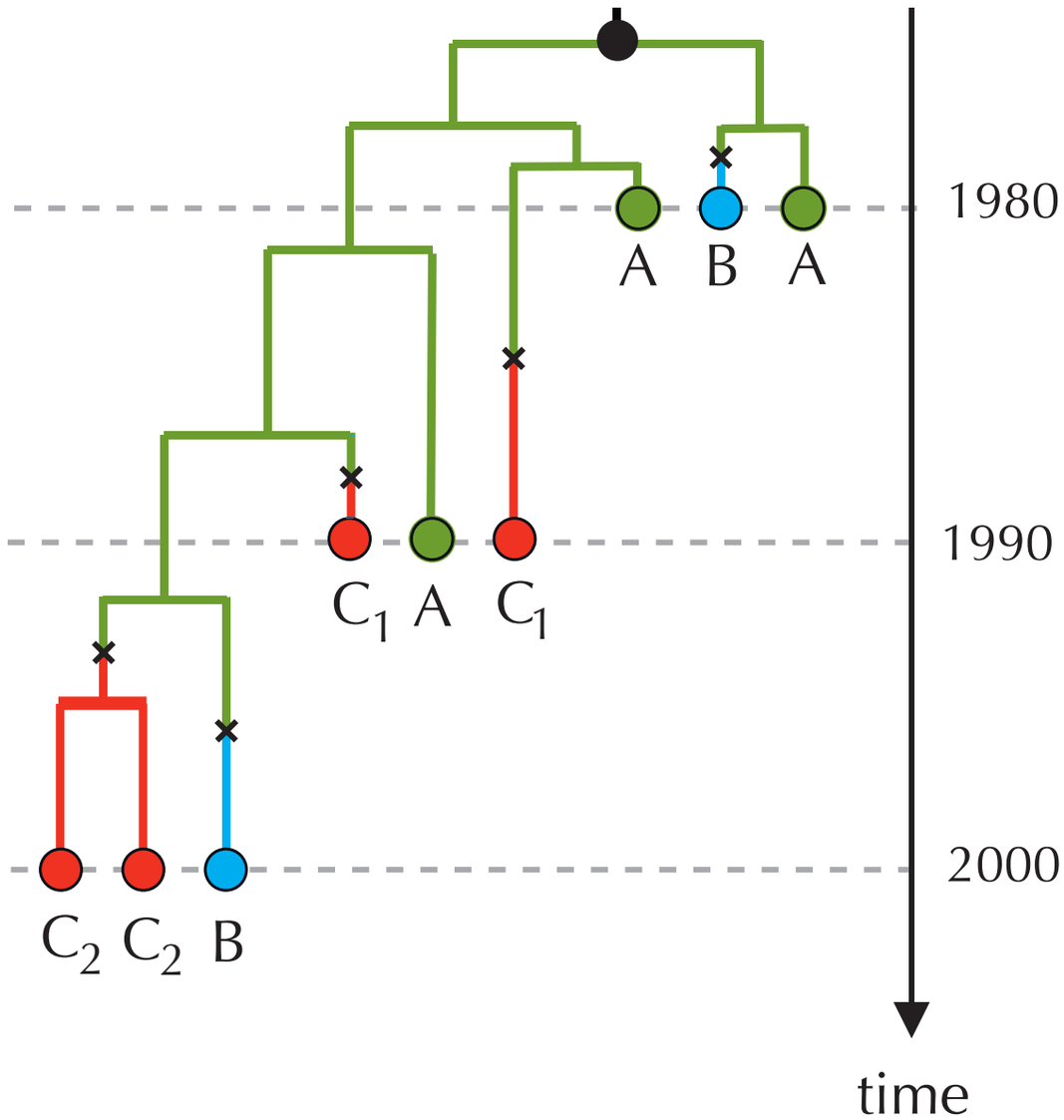
3 Coalescent theory

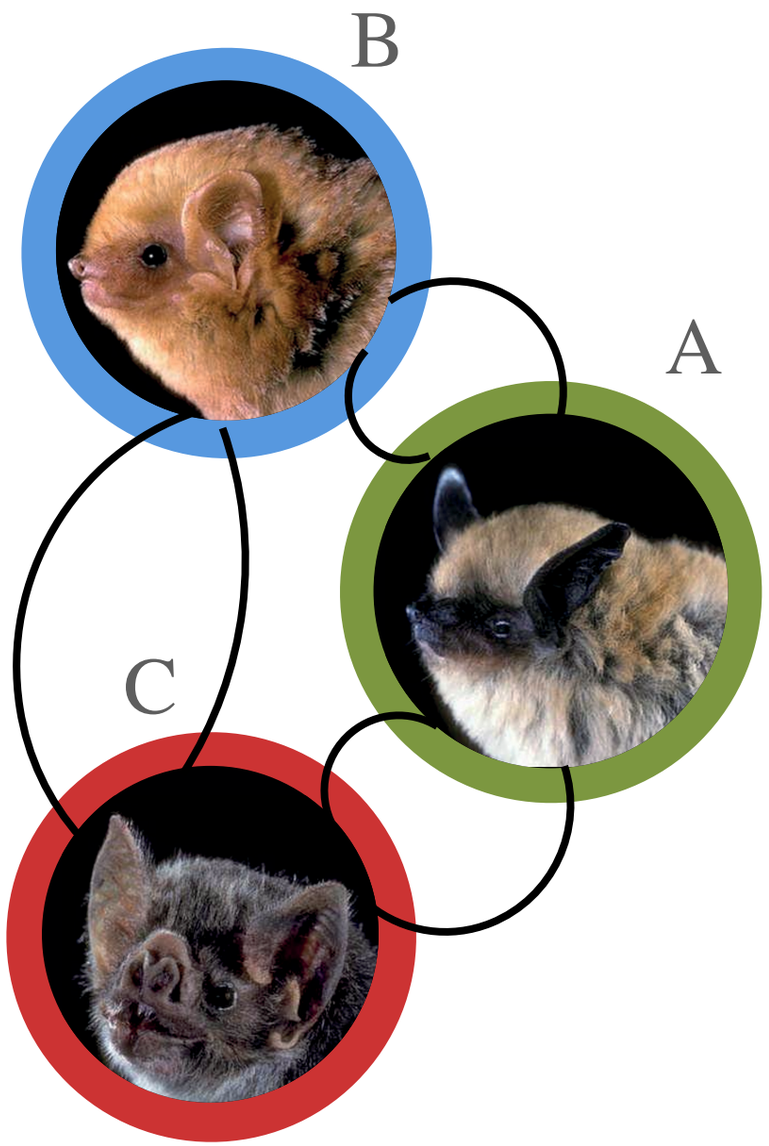
Phylodynamics toolbox



4 Phylogeography

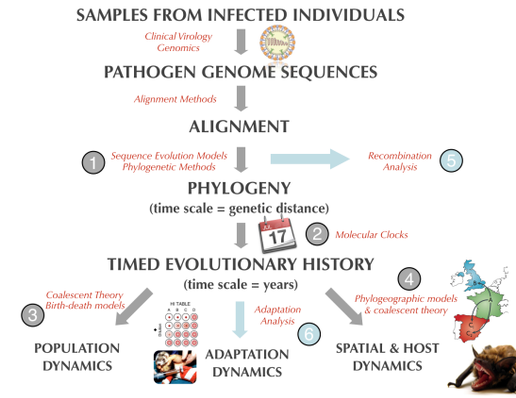
Phylodynamics toolbox

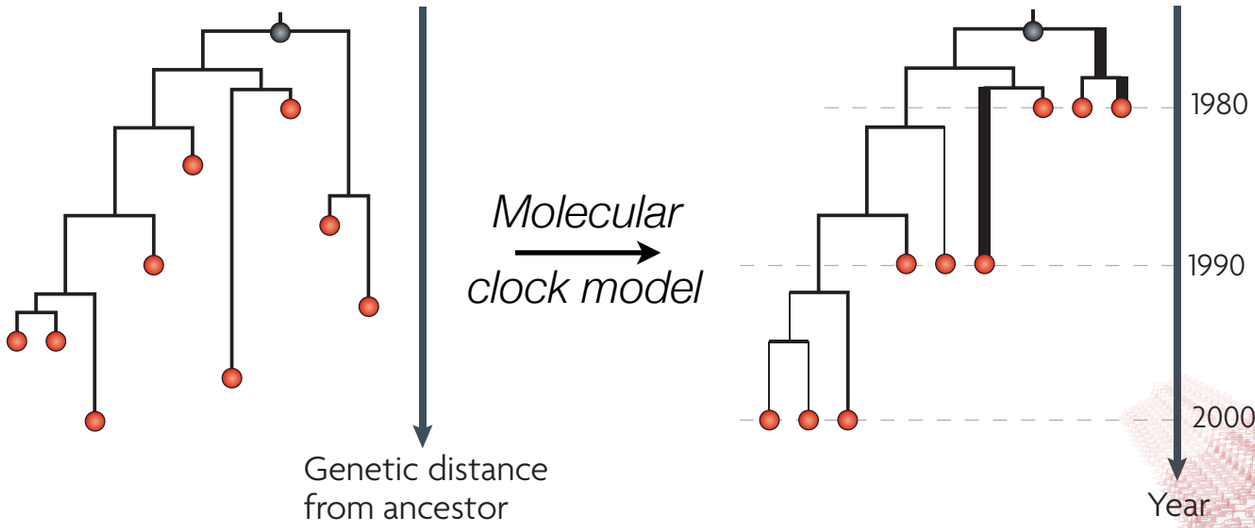


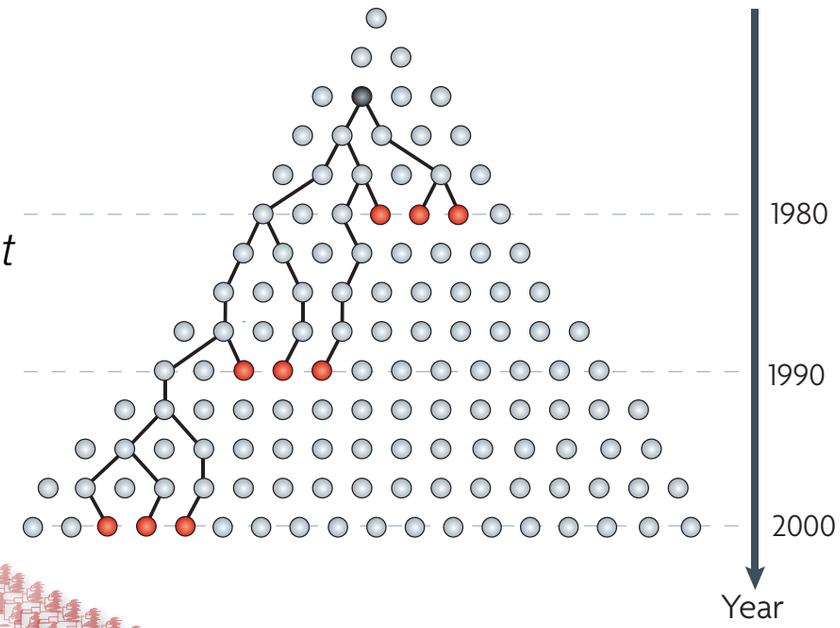


4 Phylogenetic diffusion

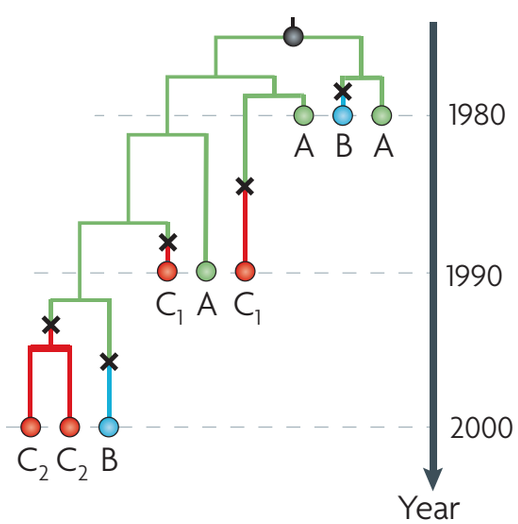
Phylodynamics integration

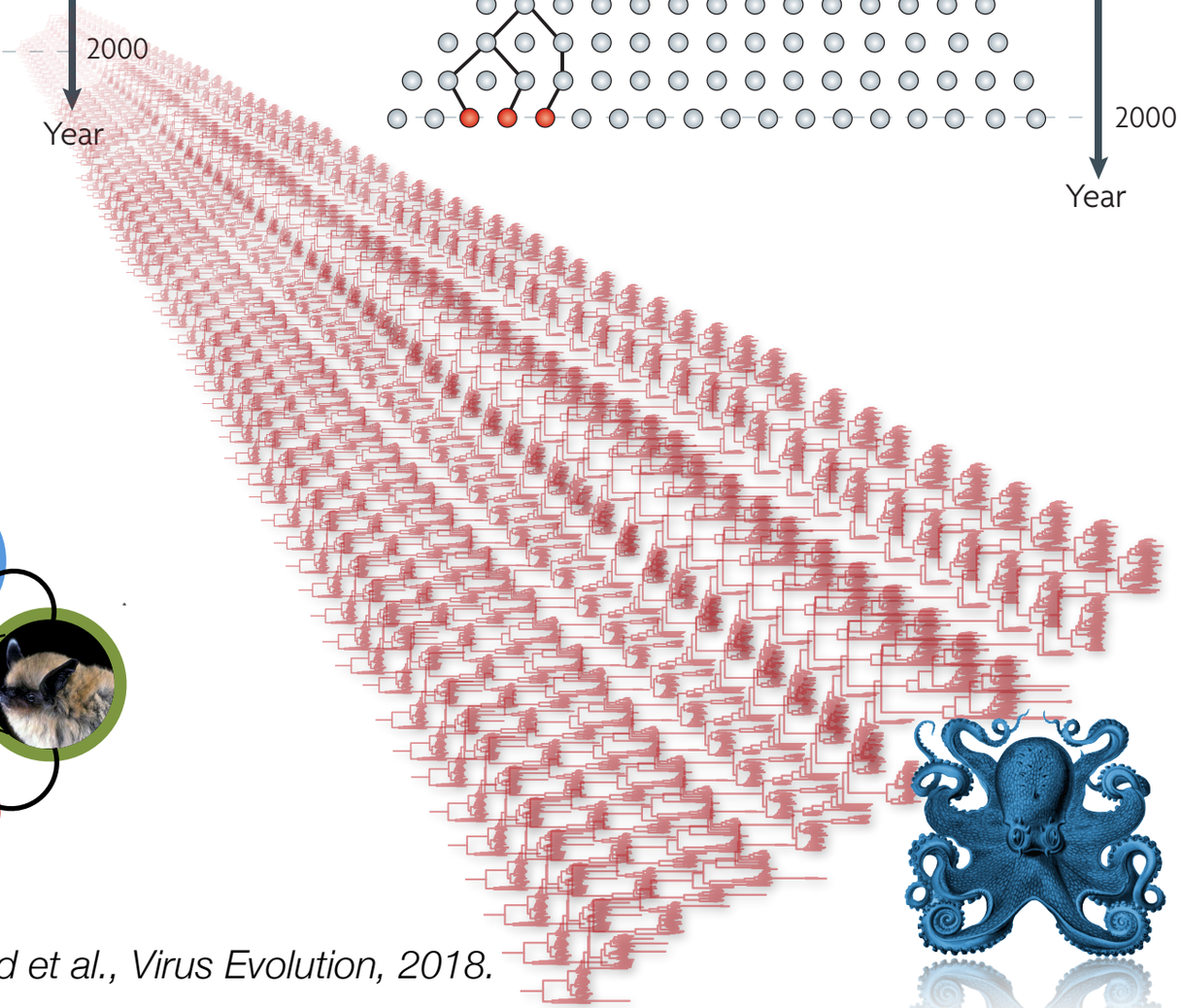






Pybus and Rambaut, NRG, 2009.

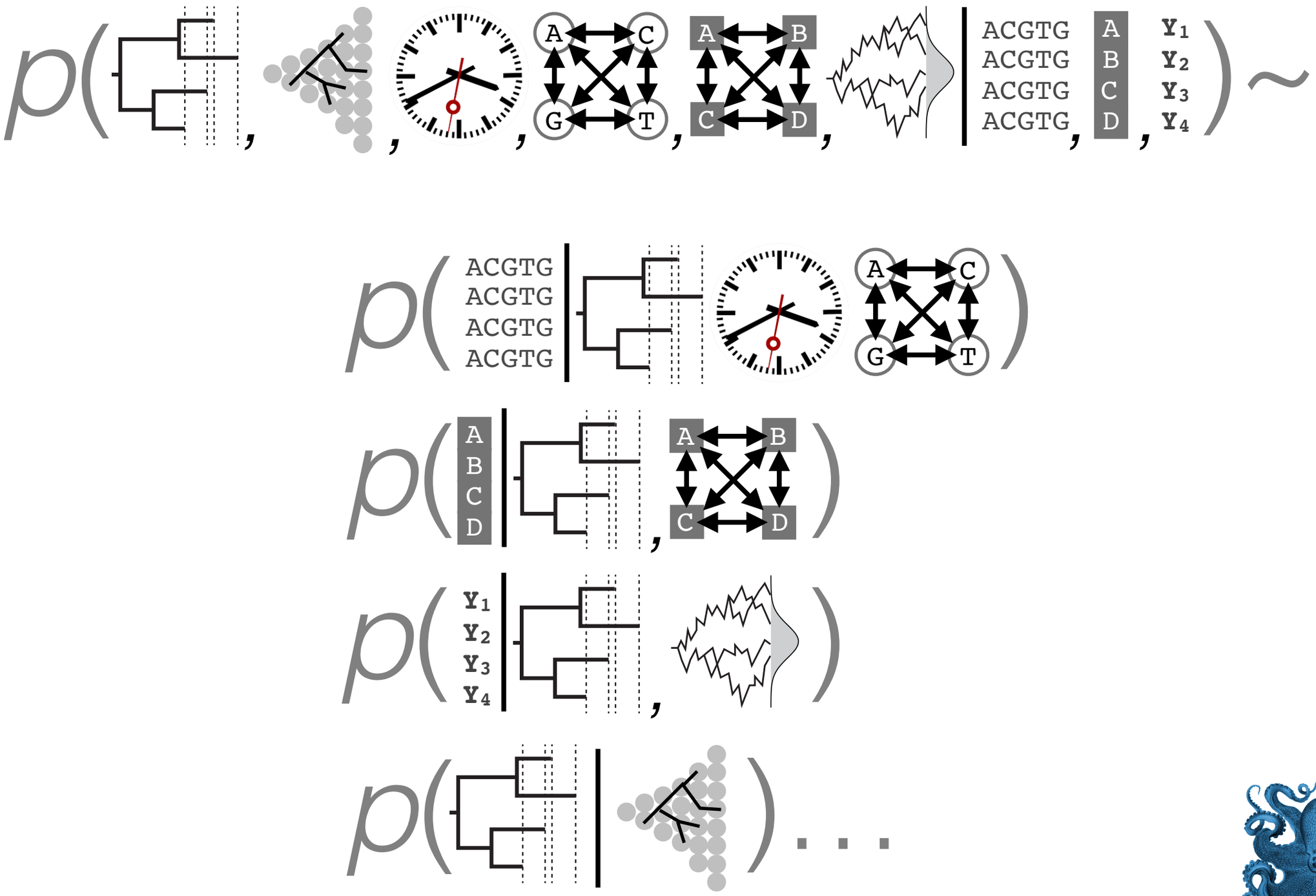




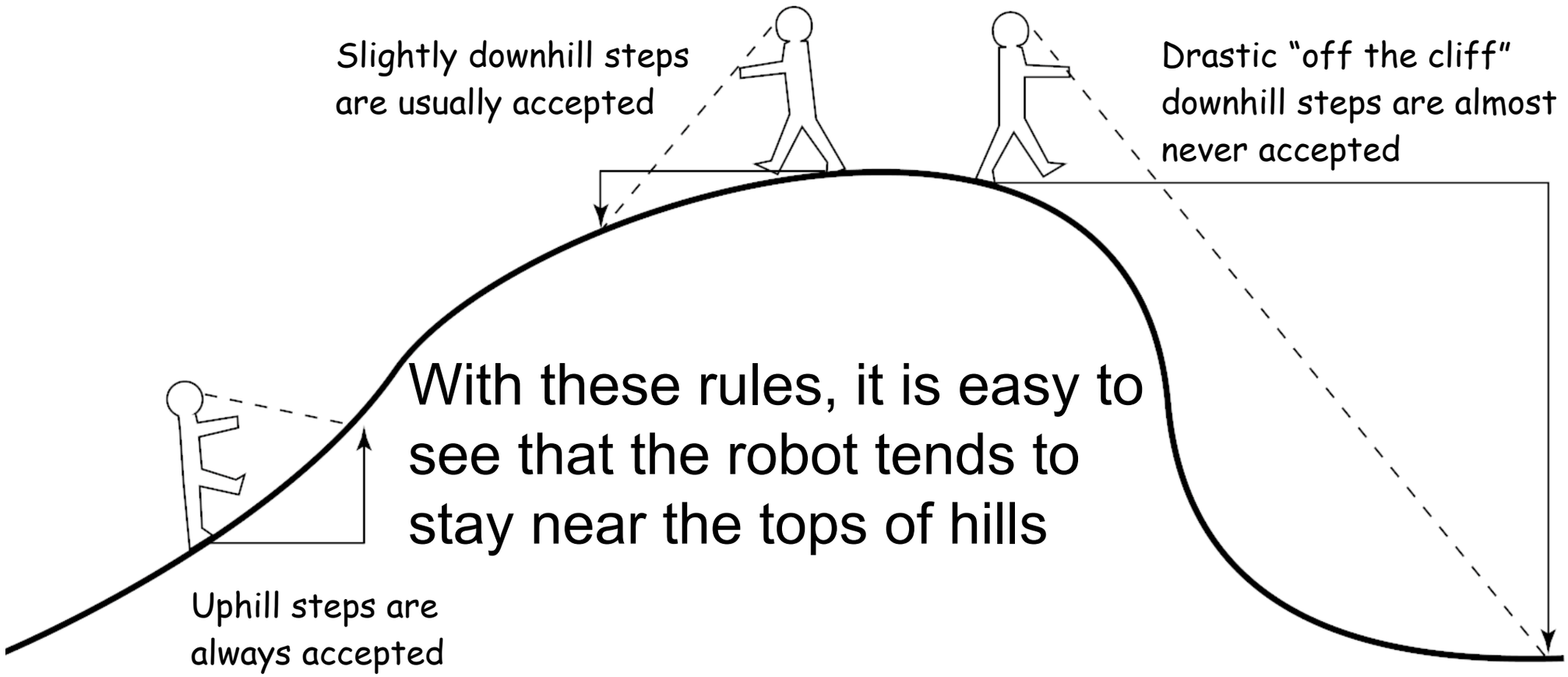
Suchard et al., Virus Evolution, 2018.

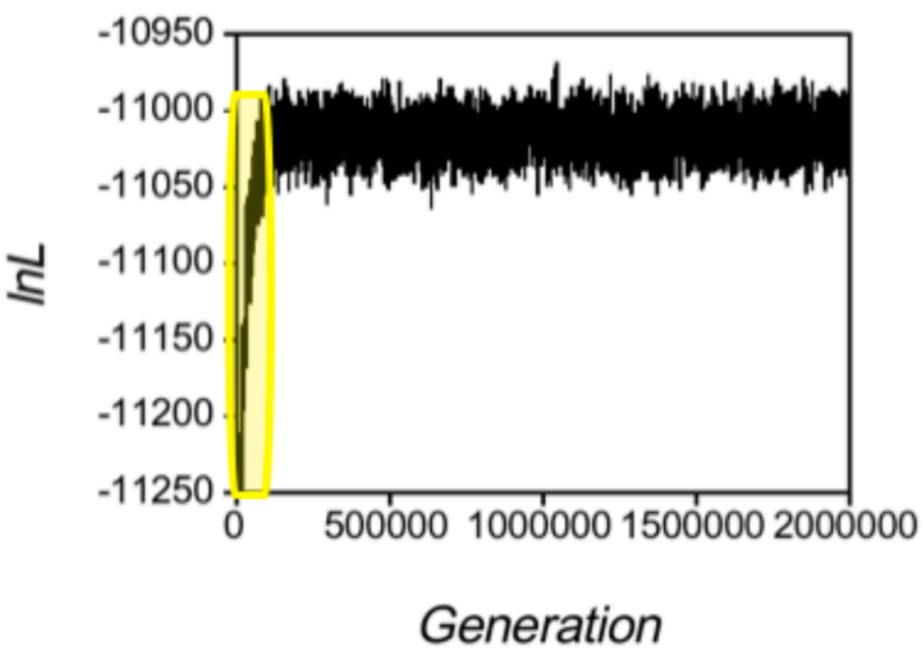
Bayesian Evolutionary Analysis

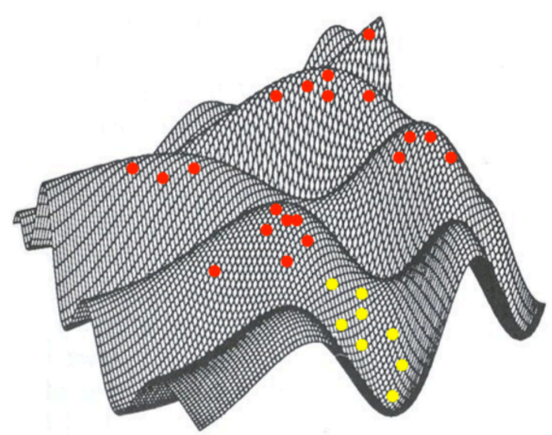
Sampling Trees



MCMC robot rules

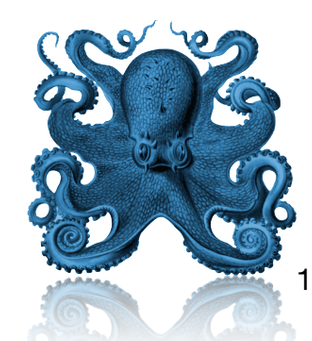






Although correlated, the Markov chain samples are valid draws from the posterior;

However, initial sampling (burn-in) is often discarded due to correlation with chain's starting point (≠posterior)



What is BEAST ?

BEAST is a Bayesian MCMC analysis package

- Unlike MrBayes we have concentrated on rooted trees with a **timescale**
- Numerous models are included and can be used in many different combinations

What is BEAST ?

Free software package written by Alexei Drummond, Marc Suchard & Andrew Rambaut

- With help from many people including (alphabetically):

  Alex Alekseyenko, Daniel Ayres, Guy Baele, Trevor Bedford, Peter Beerli, Filip Bielejec, Erik Bloomquist, Aaron Darling, Gytis Dudas, Roald Forsberg, Matthew Hall, Jess Hedge, Joseph Heled, Simon Ho, Sebastian Hoehna, Mark Holder, Denise Kuehnert, Philippe Lemey, Paul Lewis, Wai Lok Sibon Li, Gerton Lunter, Sam Lycett, Sidney Markowitz, Vladimir Minin, Michael Defoin Platel, Oliver Pybus, Tulio de Oliveira, Jayna Ragwani, Beth Shapiro, Korbinian Strimmer, David Swofford, Chieh-Hsi Wu, Walter Xie, Ziheng Yang

- Cross platform so it runs on Macintosh, Linux & other operating systems

Website and documentation

- http://beast.community/

Open source project

- https://github.com/beast-dev/beast-mcmc

What is BEAST ?

Comes with a number of utility programs to help diagnose and summarize the output

- A basic user-interface allows the selection of a range of commonly-used models
- More complicated models can be selected by editing a configuration file

What can we estimate?

Population genetics:

- Population sizes
- Growth/decline in population
- Bottlenecks/transition points
- Gene trees/species trees

Rates of evolution:

- Dates of MRCAs
- Times of bottlenecks/ transitions
- Rates for genes
- Rates for lineages

Phylogenetic:

- Founder-effects
- Epidemiological tracking
- Phylogeography
- Trait evolution
- Virus transmission dynamics

Of what does BEAST comprise?

the various bits of BEAST:

- BEAUti
- BEAST
- LogCombiner
- TreeAnnotator
- Tracer

BEAUti v1.10.4

BEAST v1.10.4

LogCombiner v1.10.4

TreeAnnotator v1.10.4

Tracer v1.7.1

FigTree v1.4.4

TreeStat v1.10.4

TempEst v1.5.3

SpreaD3 v0.9.6

PiBUSS v1.3.8

Setting up the model in BEAUti...

BEAUti

Utility program for creating BEAST input files

Accepts alignment files as input

- imports NEXUS or FASTA format

Allows ages of samples to be input

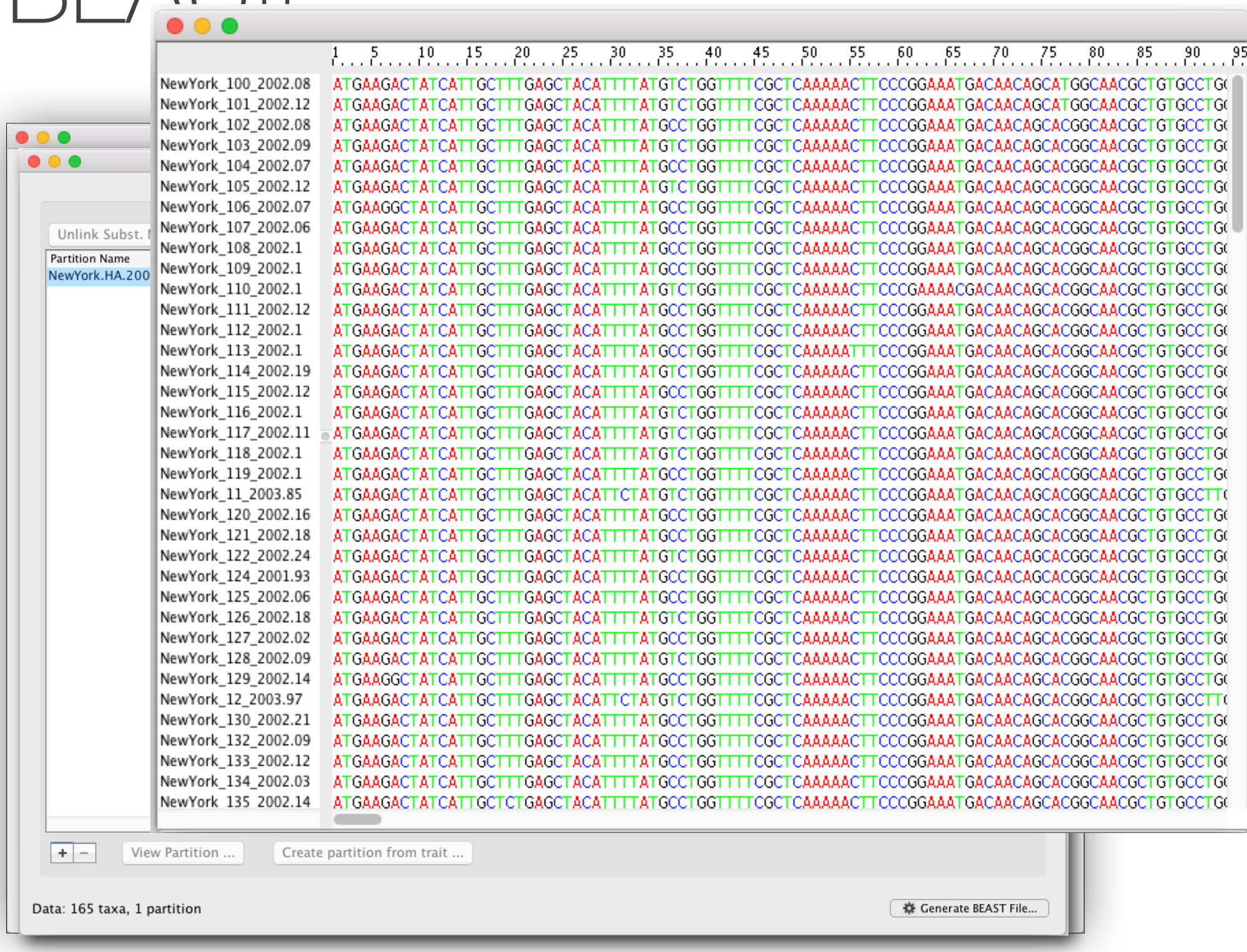
Allows selection of strict and relaxed clocks

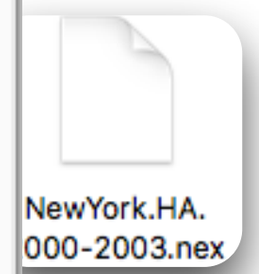
Allows selection of nucleotide substitution models

Allows manipulation of priors, operators, and MCMC chain settings

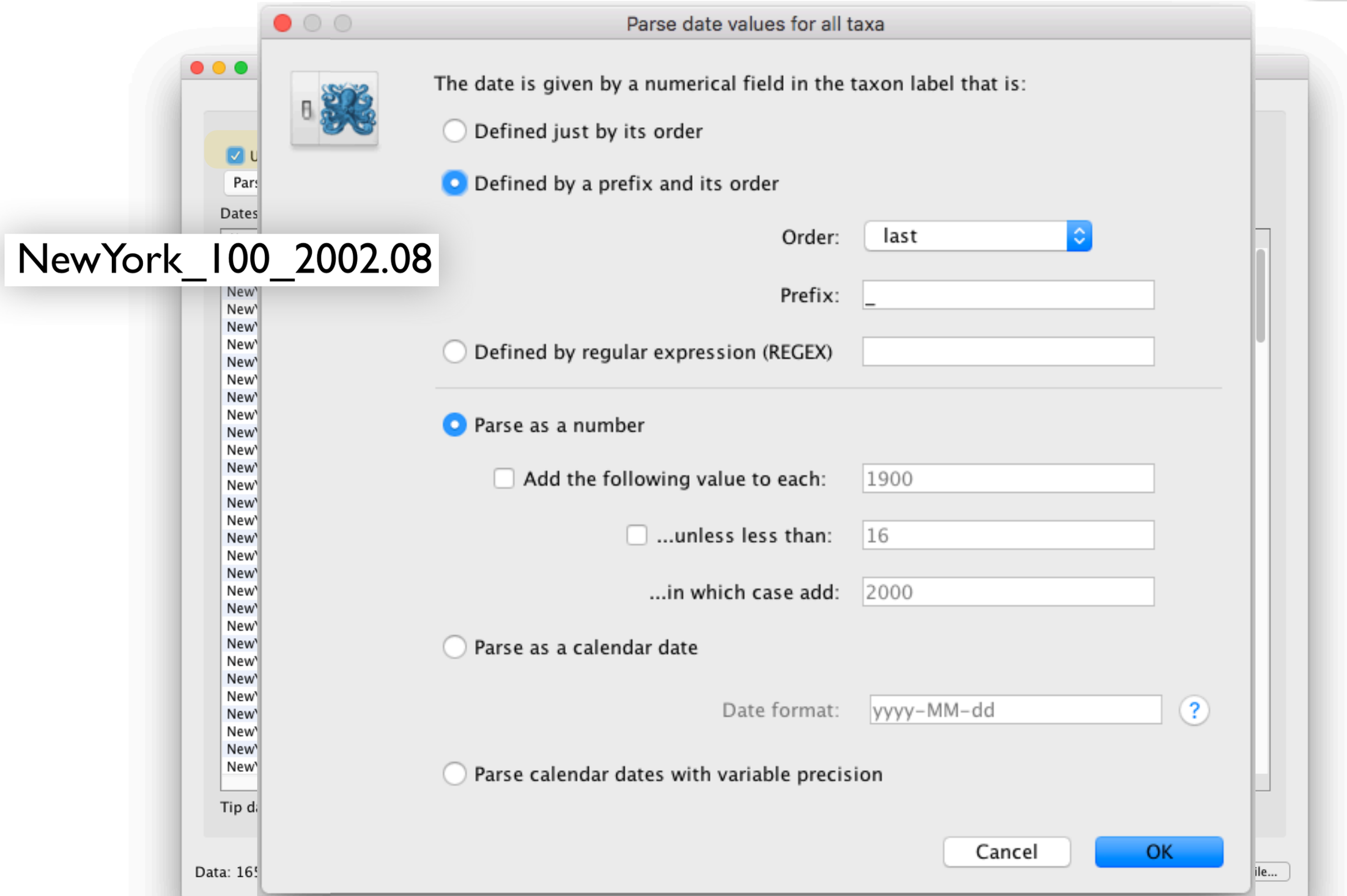
Create a BEAST XML file

BEAUti

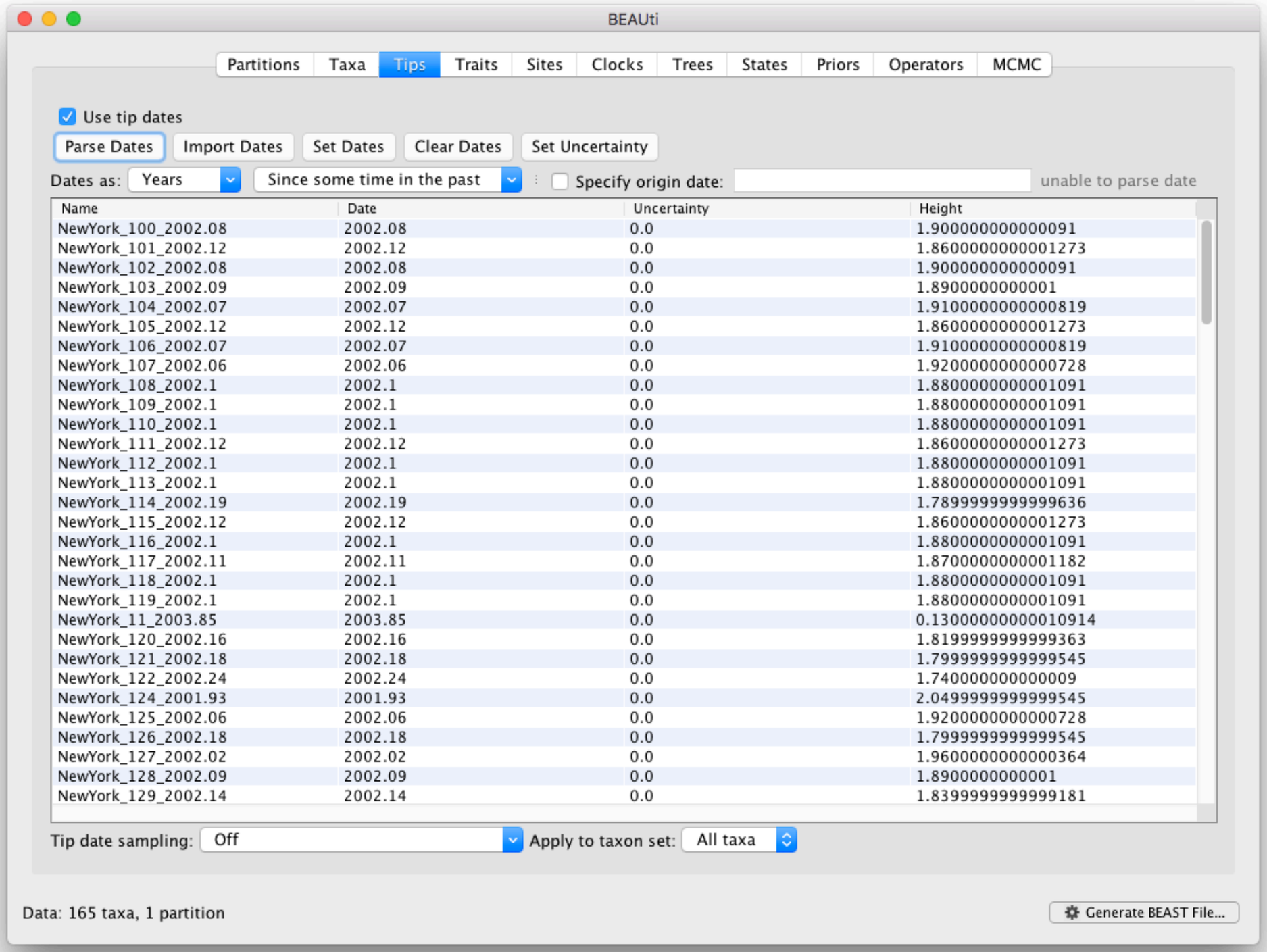

Unlink Subst. M
Partition Name
NewYork.HA.200
View Partition ...
Create partition from trait ...
Data: 165 taxa, 1 partition
Generate BEAST File...


NewYork.HA.
000-2003.nex

BEAUti

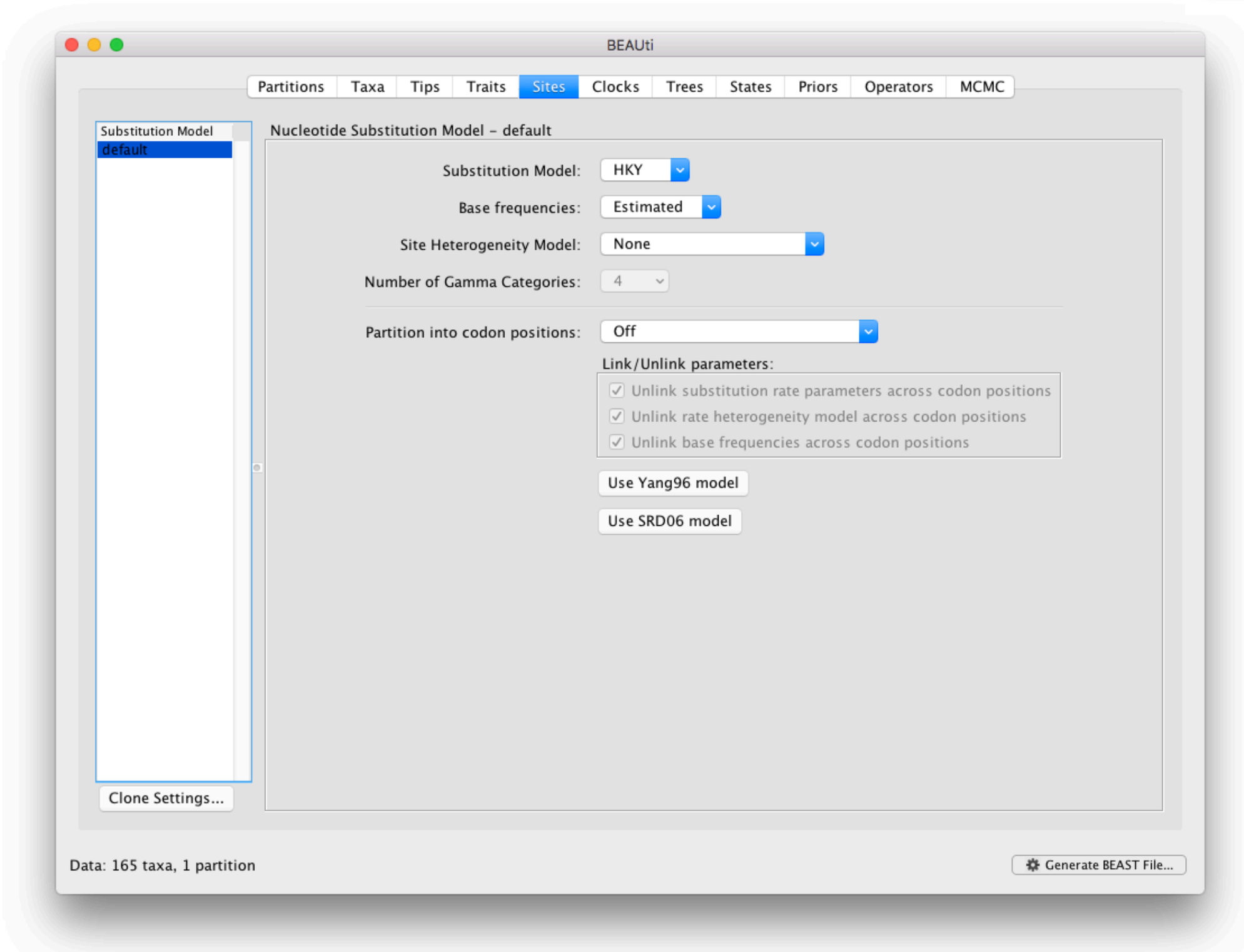

NewYork_100_2002.08
Parse date values for all taxa
The date is given by a numerical field in the taxon label that is:
Defined just by its order
Defined by a prefix and its order
Order:
last
Prefix:
_
Defined by regular expression (REGEX)
Parse as a number
Add the following value to each:
1900
...unless less than:
16
...in which case add:
2000
Parse as a calendar date
Date format:
yyyy-MM-dd
Parse calendar dates with variable precision
Cancel
OK
Dates
Data: 165

BEAUti

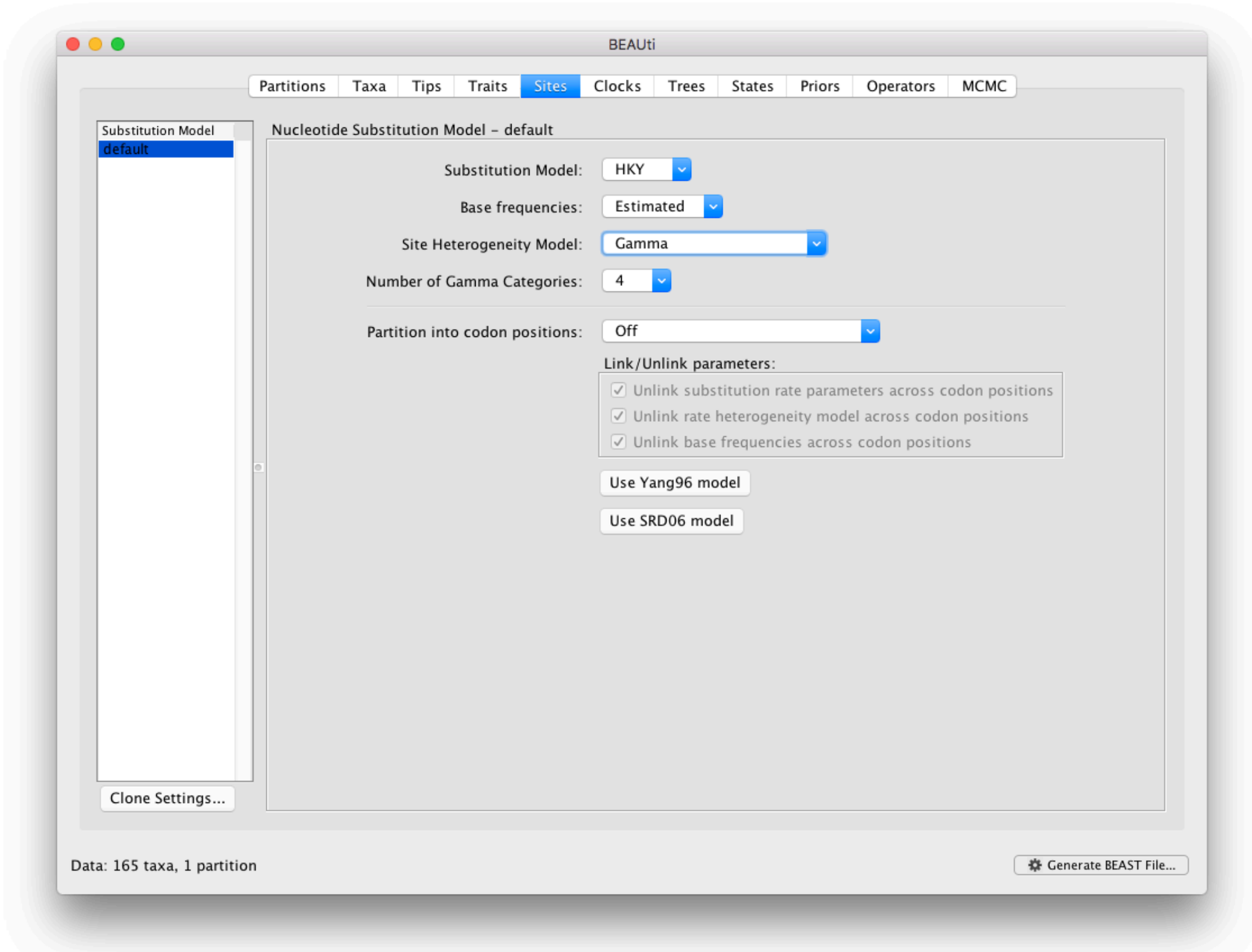


Name	Date	Uncertainty	Height
NewYork_100_2002.08	2002.08	0.0	1.900000000000091
NewYork_101_2002.12	2002.12	0.0	1.8600000000001273
NewYork_102_2002.08	2002.08	0.0	1.900000000000091
NewYork_103_2002.09	2002.09	0.0	1.8900000000001
NewYork_104_2002.07	2002.07	0.0	1.9100000000000819
NewYork_105_2002.12	2002.12	0.0	1.8600000000001273
NewYork_106_2002.07	2002.07	0.0	1.9100000000000819
NewYork_107_2002.06	2002.06	0.0	1.9200000000000728
NewYork_108_2002.1	2002.1	0.0	1.8800000000001091
NewYork_109_2002.1	2002.1	0.0	1.8800000000001091
NewYork_110_2002.1	2002.1	0.0	1.8800000000001091
NewYork_111_2002.12	2002.12	0.0	1.8600000000001273
NewYork_112_2002.1	2002.1	0.0	1.8800000000001091
NewYork_113_2002.1	2002.1	0.0	1.8800000000001091
NewYork_114_2002.19	2002.19	0.0	1.7899999999999636
NewYork_115_2002.12	2002.12	0.0	1.8600000000001273
NewYork_116_2002.1	2002.1	0.0	1.8800000000001091
NewYork_117_2002.11	2002.11	0.0	1.8700000000001182
NewYork_118_2002.1	2002.1	0.0	1.8800000000001091
NewYork_119_2002.1	2002.1	0.0	1.8800000000001091
NewYork_11_2003.85	2003.85	0.0	0.13000000000010914
NewYork_120_2002.16	2002.16	0.0	1.8199999999999363
NewYork_121_2002.18	2002.18	0.0	1.7999999999999545
NewYork_122_2002.24	2002.24	0.0	1.740000000000009
NewYork_124_2001.93	2001.93	0.0	2.0499999999999545
NewYork_125_2002.06	2002.06	0.0	1.9200000000000728
NewYork_126_2002.18	2002.18	0.0	1.7999999999999545
NewYork_127_2002.02	2002.02	0.0	1.9600000000000364
NewYork_128_2002.09	2002.09	0.0	1.8900000000001
NewYork_129_2002.14	2002.14	0.0	1.8399999999999181

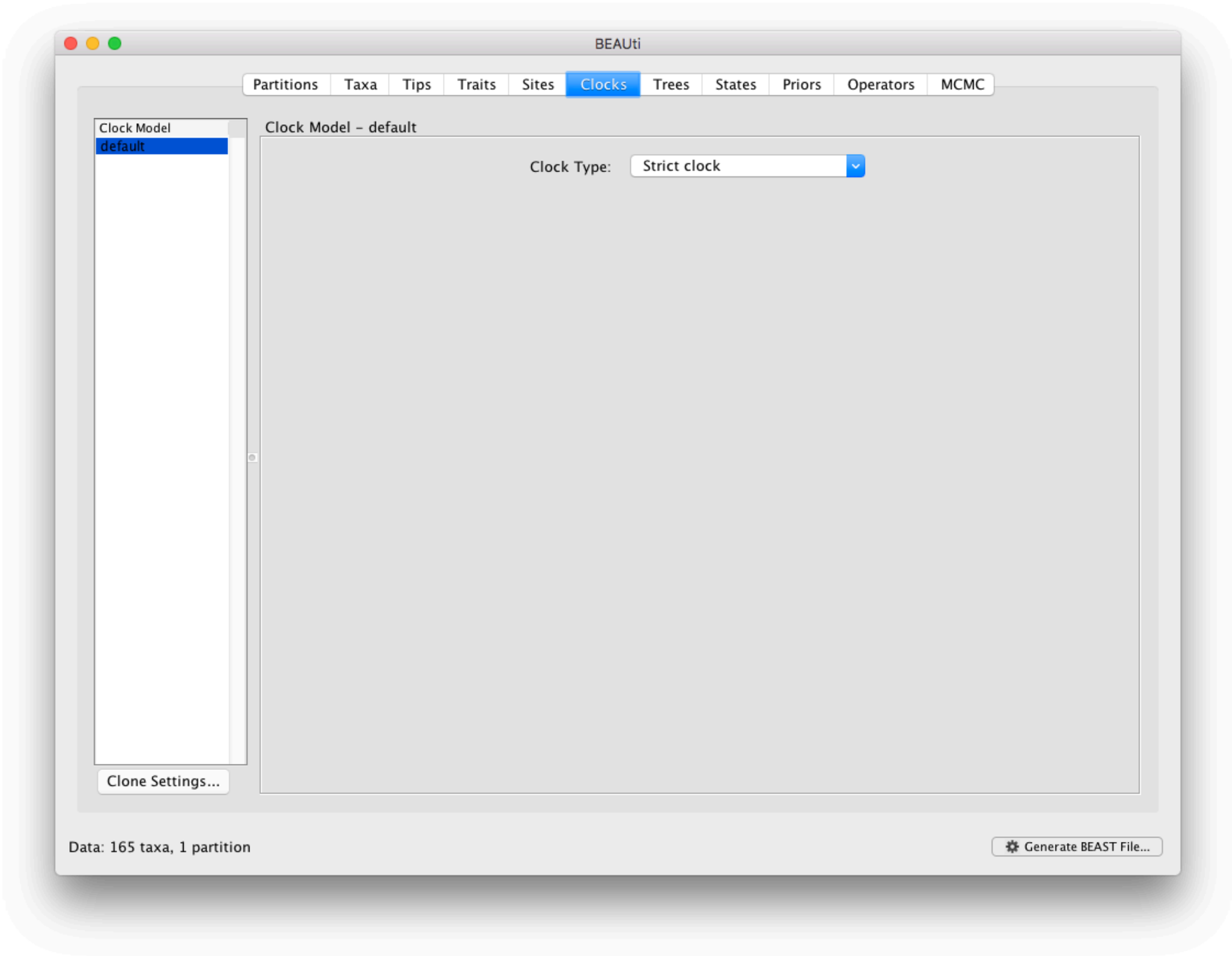
BEAUti


BEAUti
Partitions
Taxa
Tips
Traits
Sites
Clocks
Trees
States
Priors
Operators
MCMC
Substitution Model
default
Nucleotide Substitution Model – default
Substitution Model:
HKY
Base frequencies:
Estimated
Site Heterogeneity Model:
None
Number of Gamma Categories:
4
Partition into codon positions:
Off
Link/Unlink parameters:
Unlink substitution rate parameters across codon positions
Unlink rate heterogeneity model across codon positions
Unlink base frequencies across codon positions
Use Yang96 model
Use SRD06 model
Clone Settings...
Data: 165 taxa, 1 partition
Generate BEAST File...

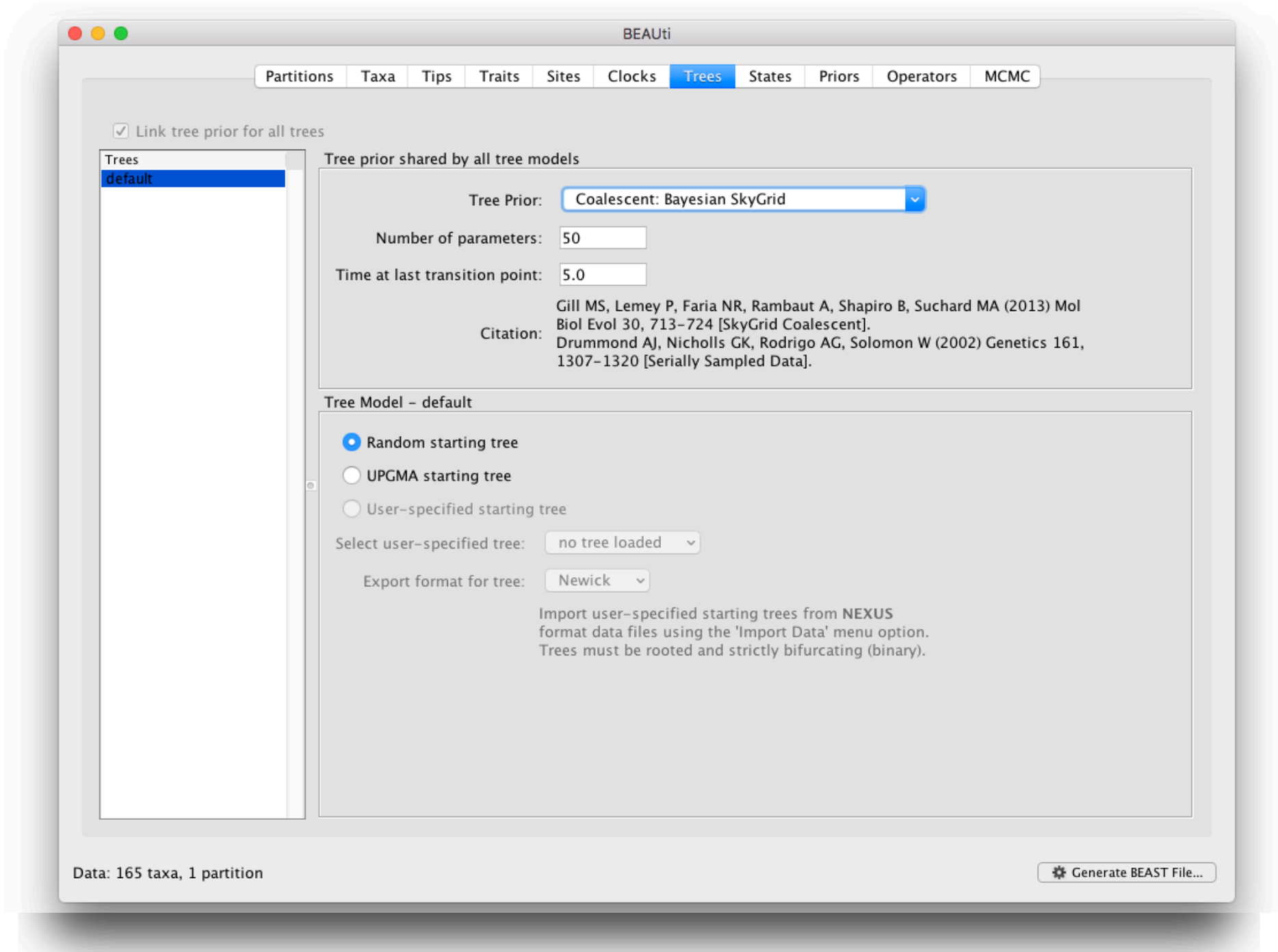
BEAUti


BEAUti
Partitions
Taxa
Tips
Traits
Sites
Clocks
Trees
States
Priors
Operators
MCMC
Substitution Model
default
Nucleotide Substitution Model – default
Substitution Model:
HKY
Base frequencies:
Estimated
Site Heterogeneity Model:
Gamma
Number of Gamma Categories:
4
Partition into codon positions:
Off
Link/Unlink parameters:
Unlink substitution rate parameters across codon positions
Unlink rate heterogeneity model across codon positions
Unlink base frequencies across codon positions
Use Yang96 model
Use SRD06 model
Clone Settings...
Data: 165 taxa, 1 partition
Generate BEAST File...

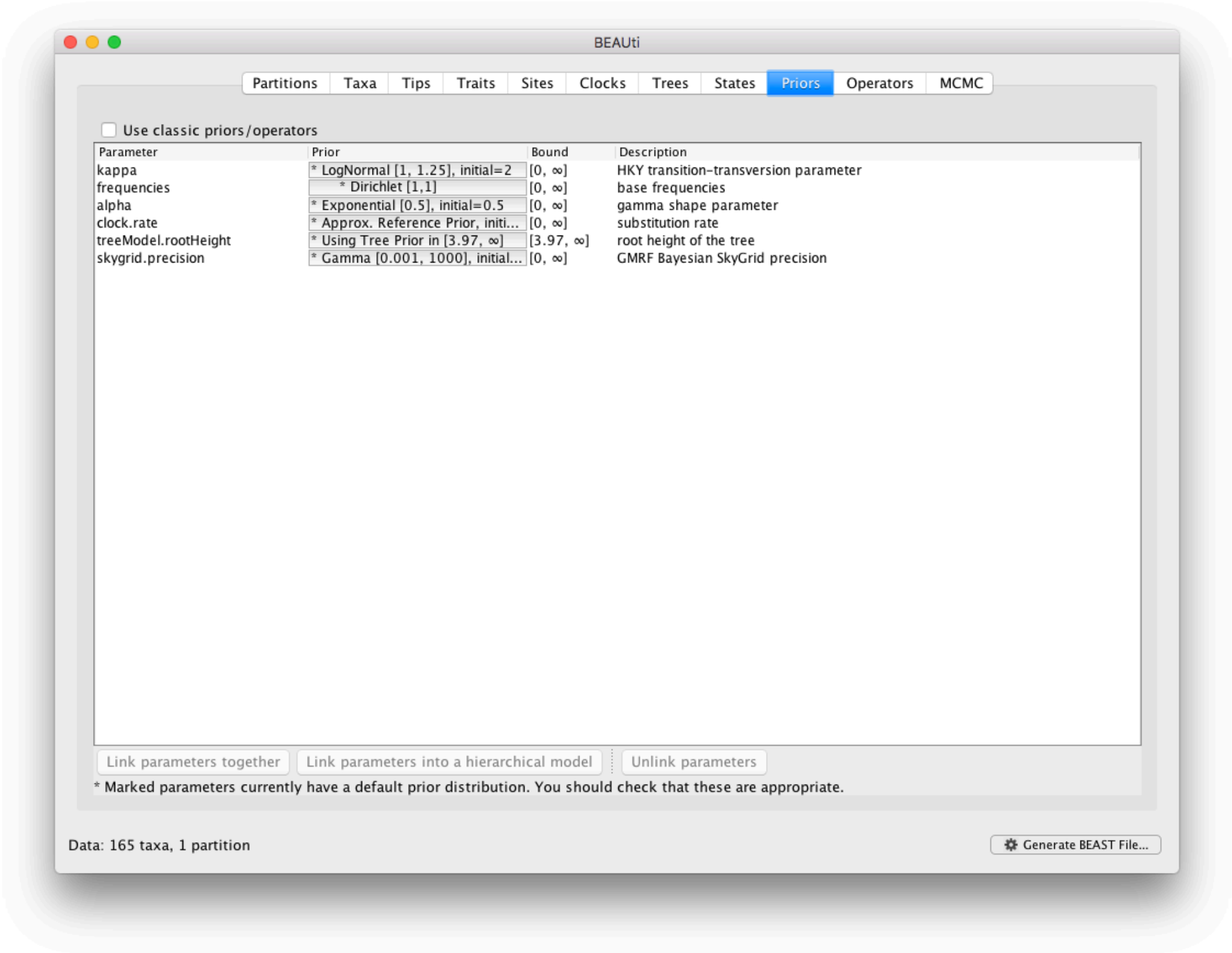
BEAUti


BEAUti
Partitions
Taxa
Tips
Traits
Sites
Clocks
Trees
States
Priors
Operators
MCMC
Clock Model
default
Clock Model – default
Clock Type:
Strict clock
Clone Settings...
Data: 165 taxa, 1 partition
Generate BEAST File...

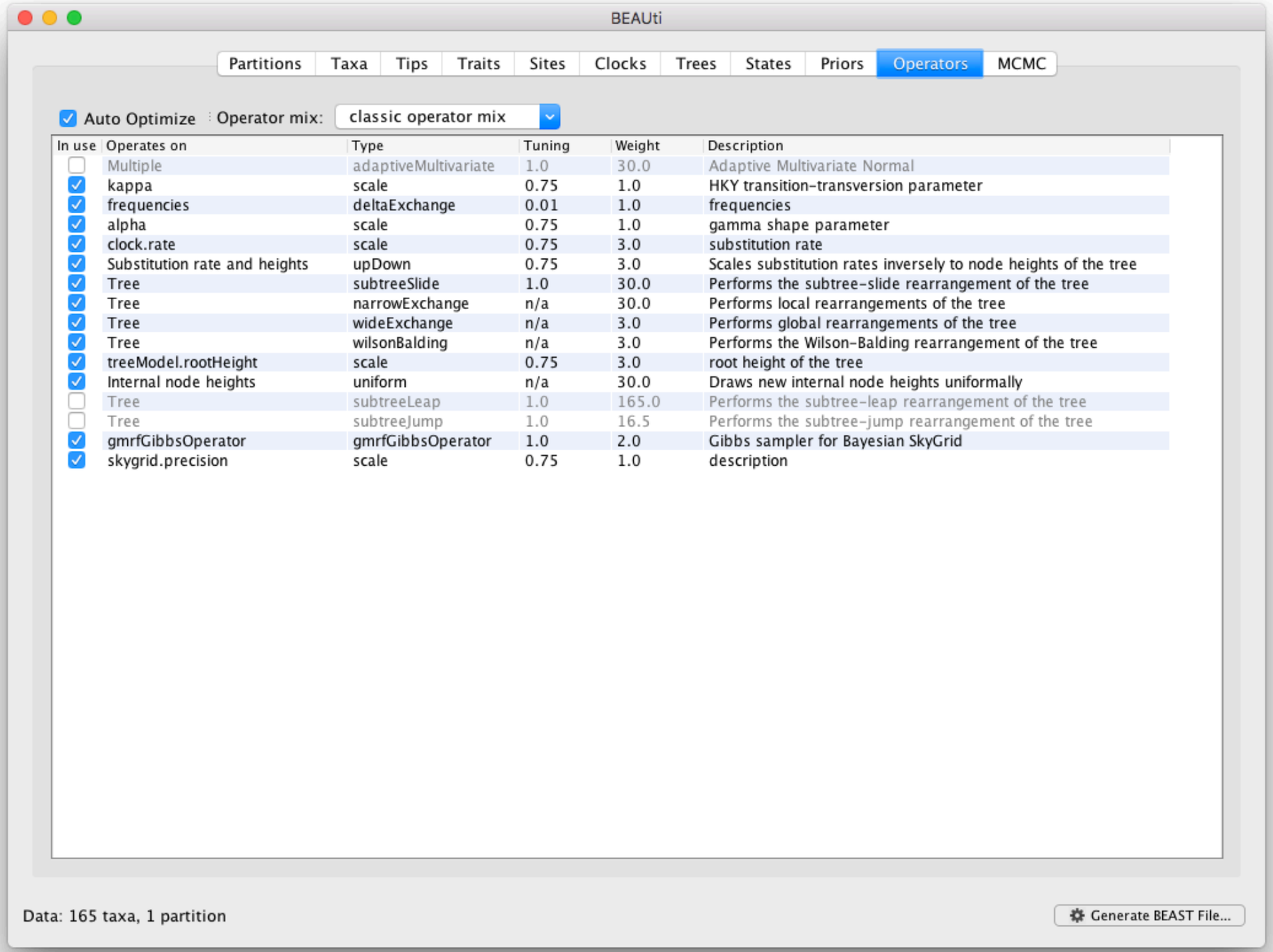
BEAUti


BEAUti
Partitions
Taxa
Tips
Traits
Sites
Clocks
Trees
States
Priors
Operators
MCMC
Link tree prior for all trees
Trees
default
Tree prior shared by all tree models
Tree Prior:
Coalescent: Bayesian SkyGrid
Number of parameters:
50
Time at last transition point:
5.0
Citation:
Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA (2013) Mol Biol Evol 30, 713-724 [SkyGrid Coalescent].
Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Genetics 161, 1307-1320 [Serially Sampled Data].
Tree Model - default
Random starting tree
UPGMA starting tree
User-specified starting tree
Select user-specified tree:
no tree loaded
Export format for tree:
Newick
Import user-specified starting trees from NEXUS format data files using the 'Import Data' menu option.
Trees must be rooted and strictly bifurcating (binary).
Data: 165 taxa, 1 partition
Generate BEAST File...

BEAUti

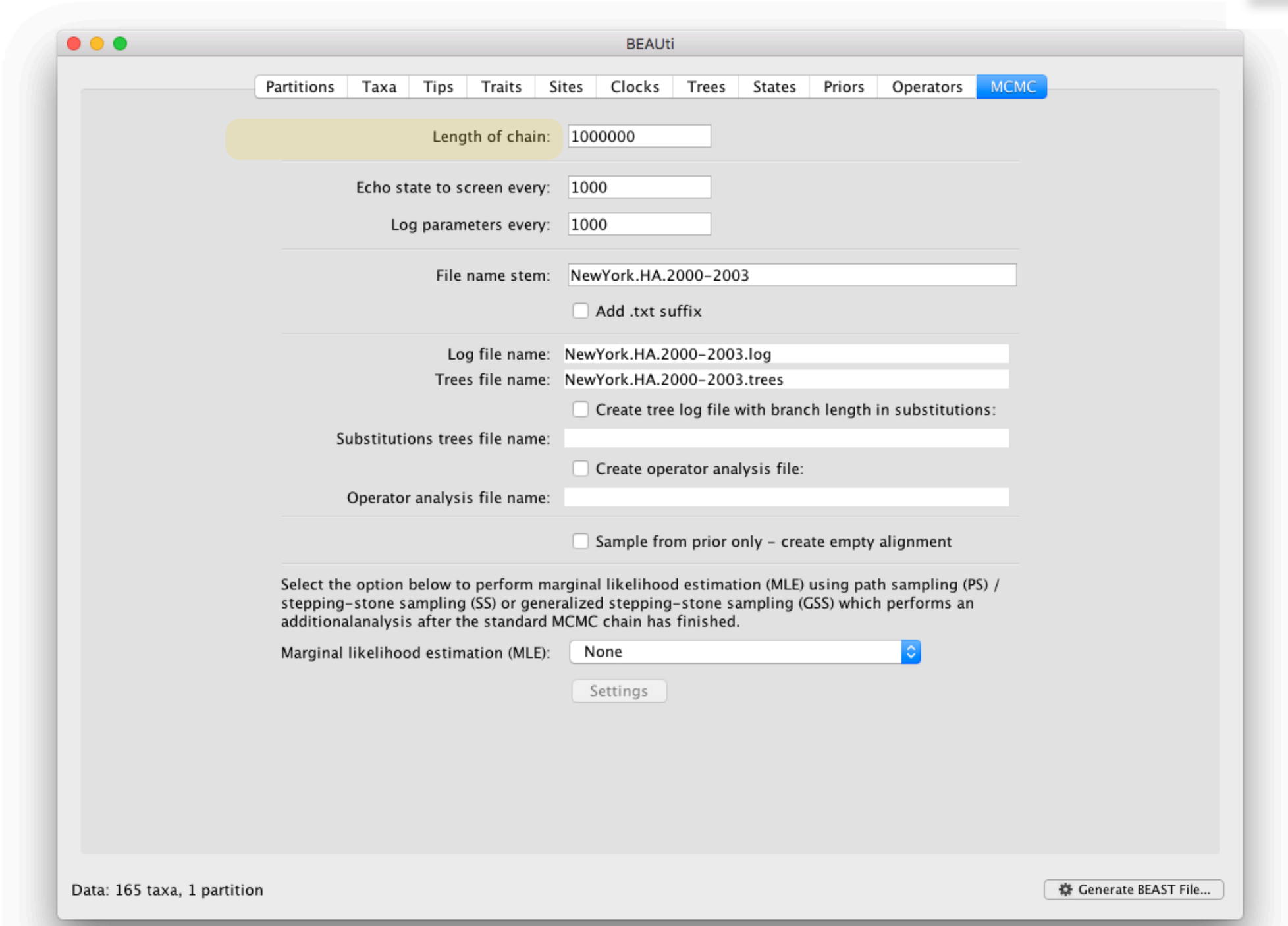

BEAUti
Partitions
Taxa
Tips
Traits
Sites
Clocks
Trees
States
Priors
Operators
MCMC
Use classic priors/operators
Parameter
Prior
Bound
Description
kappa
* LogNormal [1, 1.25], initial=2
[0, ∞]
HKY transition-transversion parameter
frequencies
* Dirichlet [1,1]
[0, ∞]
base frequencies
alpha
* Exponential [0.5], initial=0.5
[0, ∞]
gamma shape parameter
clock.rate
* Approx. Reference Prior, initi...
[0, ∞]
substitution rate
treeModel.rootHeight
* Using Tree Prior in [3.97, ∞]
[3.97, ∞]
root height of the tree
skygrid.precision
* Gamma [0.001, 1000], initial...
[0, ∞]
GMRF Bayesian SkyGrid precision
Link parameters together
Link parameters into a hierarchical model
Unlink parameters
* Marked parameters currently have a default prior distribution. You should check that these are appropriate.
Data: 165 taxa, 1 partition
Generate BEAST File...

BEAUti



In use	Operates on	Type	Tuning	Weight	Description
☐	Multiple	adaptiveMultivariate	1.0	30.0	Adaptive Multivariate Normal
☑	kappa	scale	0.75	1.0	HKY transition-transversion parameter
☑	frequencies	deltaExchange	0.01	1.0	frequencies
☑	alpha	scale	0.75	1.0	gamma shape parameter
☑	clock.rate	scale	0.75	3.0	substitution rate
☑	Substitution rate and heights	upDown	0.75	3.0	Scales substitution rates inversely to node heights of the tree
☑	Tree	subtreeSlide	1.0	30.0	Performs the subtree-slide rearrangement of the tree
☑	Tree	narrowExchange	n/a	30.0	Performs local rearrangements of the tree
☑	Tree	wideExchange	n/a	3.0	Performs global rearrangements of the tree
☑	Tree	wilsonBalding	n/a	3.0	Performs the Wilson-Balding rearrangement of the tree
☑	treeModel.rootHeight	scale	0.75	3.0	root height of the tree
☑	Internal node heights	uniform	n/a	30.0	Draws new internal node heights uniformally
☐	Tree	subtreeLeap	1.0	165.0	Performs the subtree-leap rearrangement of the tree
☐	Tree	subtreeJump	1.0	16.5	Performs the subtree-jump rearrangement of the tree
☑	gmrfGibbsOperator	gmrfGibbsOperator	1.0	2.0	Gibbs sampler for Bayesian SkyGrid
☑	skygrid.precision	scale	0.75	1.0	description

BEAUti


BEAUti
Partitions
Taxa
Tips
Traits
Sites
Clocks
Trees
States
Priors
Operators
MCMC
Length of chain:
1000000
Echo state to screen every:
1000
Log parameters every:
1000
File name stem:
NewYork.HA.2000-2003
Add .txt suffix
Log file name:
NewYork.HA.2000-2003.log
Trees file name:
NewYork.HA.2000-2003.trees
Create tree log file with branch length in substitutions:
Substitutions trees file name:
Create operator analysis file:
Operator analysis file name:
Sample from prior only - create empty alignment
Select the option below to perform marginal likelihood estimation (MLE) using path sampling (PS) / stepping-stone sampling (SS) or generalized stepping-stone sampling (GSS) which performs an additionalanalysis after the standard MCMC chain has finished.
Marginal likelihood estimation (MLE):
None
Settings
Data: 165 taxa, 1 partition
Generate BEAST File...

Running BEAST...

BEAST

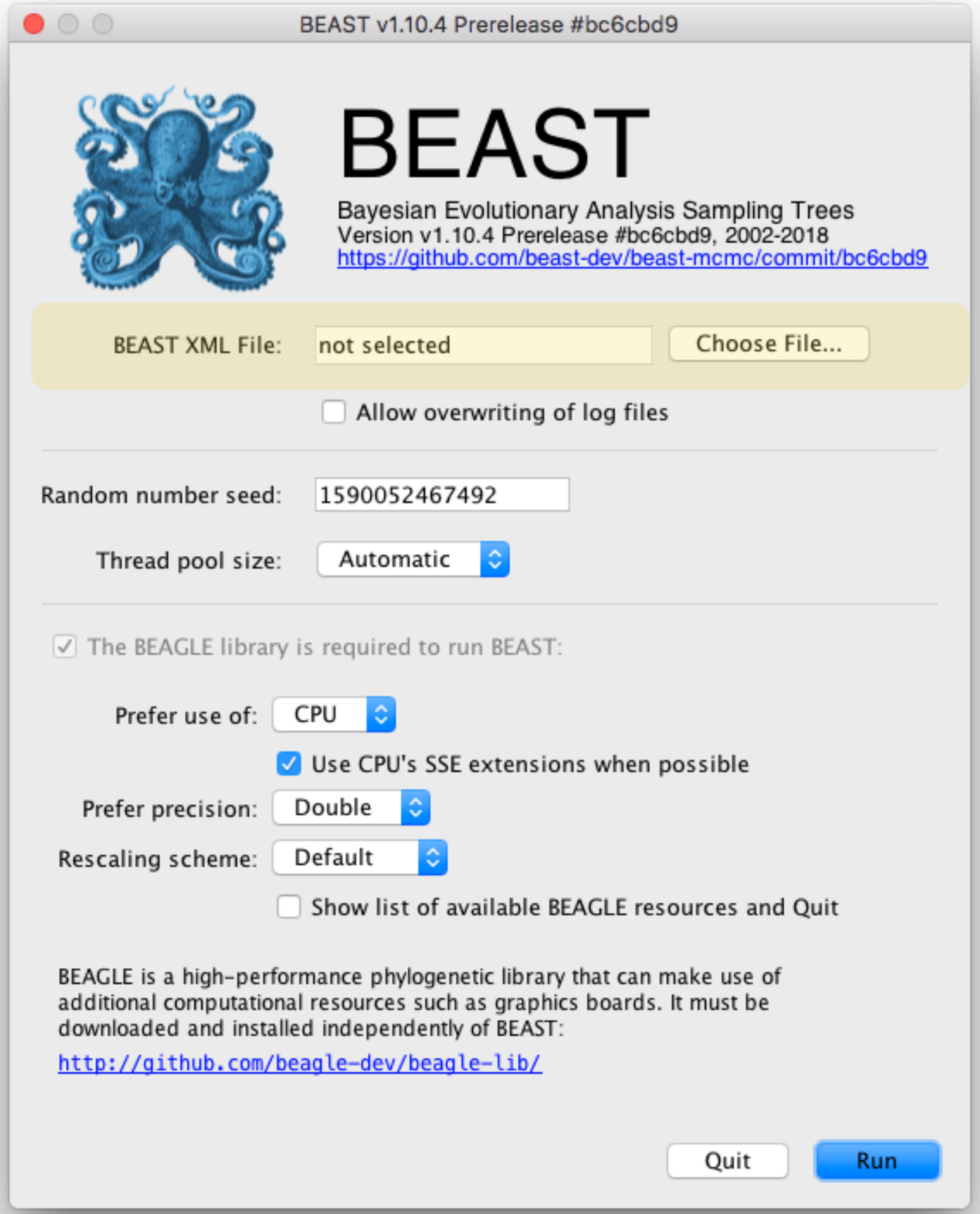
Reads instructions from the xml file

Performs the analysis

- go for a coffee while this happens
- or lunch
- or a vacation

Produces two files (with names specified in BEAUti):

- [analysisResults].log
- [analysisResults].trees



Input ‘XML’ file

Random number seed

BEAGLE settings

NewYork.HA.2000-2003.xml

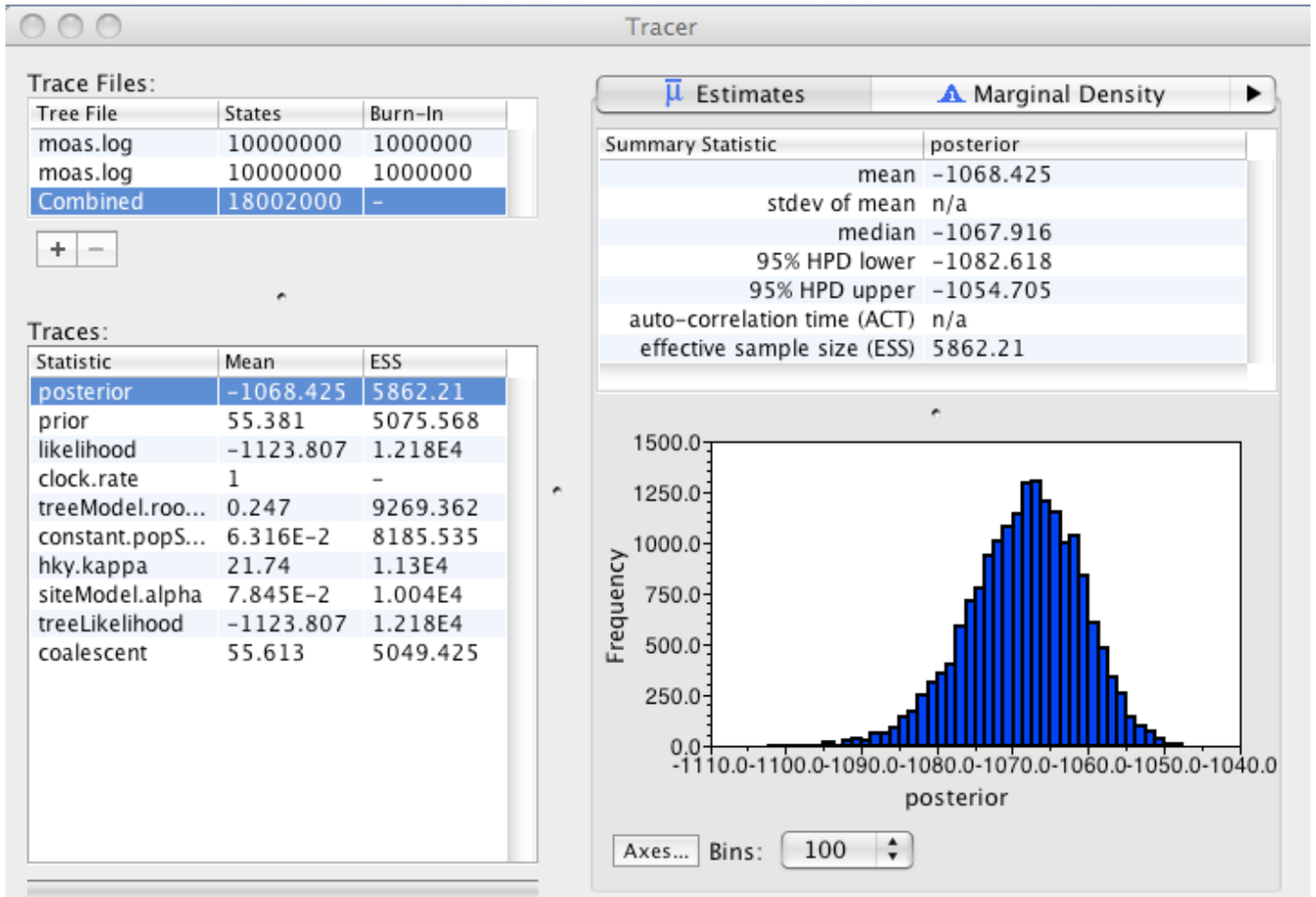
```
Creating a GMRF smoothed skyride model for multiple loci (SkyGrid)
        Population sizes: 50

Using strict molecular clock model.

Creating state frequencies model 'frequencies': Initial frequencies = {0.25, 0.25, 0.25, 0.25}

Creating HKY substitution model. Initial kappa = 2.0

Creating site rate model:
  4 category discrete gamma with initial shape = 0.5

Using BEAGLE DataLikelihood Delegate
  Using BEAGLE resource 0: CPU
    with instance flags:  PRECISION_DOUBLE COMPUTATION_SYNCH EIGEN_REAL SCALING_MANUAL SCALERS_RAW VECTOR_SSE THREADING_CPP PROCESSOR_CPU FRAMEWORK_CPU
  Ignoring ambiguities in tree likelihood.
  With 258 unique site patterns.
  Using rescaling scheme : dynamic (rescaling every 100 evaluations, delay rescaling until first overflow)

Using TreeDataLikelihood
  Branch rate model used: strictClockBranchRates
Optimization Schedule: log
Creating CTMC Scale Reference Prior model.
Acting on subtree of size 165

Likelihood computation is using an auto sizing thread pool.

Creating the MCMC chain:
  chain length = 1000000
  operator adaption = true
  adaptation delayed for 10000 steps

Citations for this analysis:

FRAMEWORK
BEAST primary citation:
        Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A (2018) Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution. vey016. D

TREE DENSITY MODELS
Skyride coalescent:
        Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA (2013) Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol Biol
        Minin VN, Bloomquist EW, Suchard MA (2008) Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol Biol Evol. 25, 1459-147

SUBSTITUTION MODELS
HKY nucleotide substitution model:
        Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol.. 22, 160-174
Discrete gamma-distributed rate heterogeneity model:
        Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol.. 39, 306-314

PRIOR MODELS
CTMC Scale Reference Prior model:
        Ferreira MAR, Suchard MA (2008) Bayesian analysis of elapsed times in continuous-time Markov chains. Canadian Journal of Statistics. 36, 355-368


# BEAST v1.10.4 Prerelease #bc6cbd9
# Generated Thu May 21 02:17:02 PDT 2020 [seed=1590052467492]
# -window -working -options
# keywords: skygrid
state   Joint           Prior           Likelihood      age(root)       clock.rate
0       -64986.4794     -7518.9541      -57467.5253     1878.76         1.00000         -
1000    -18347.2876     -1172.3641      -17174.9235     1907.13         0.16541         -
2000    -14640.0831     -1072.6676      -13567.4155     1919.73         2.82034E-2      -
3000    -13243.9754     -1060.1764      -12183.7990     1922.10         2.0908E-2       -
4000    -11786.1890     -1035.8060      -10750.3830     1942.83         8.2063E-3       -
5000    -11217.7892     -979.4798       -10238.3093     1946.27         6.24351E-3      -
6000    -10518.4077     -977.9483       -9540.4593      1953.14         4.7687E-3       -
7000    -9926.5493      -959.0115       -8967.5378      1954.61         2.83338E-3      -
8000    -9618.7770      -953.5021       -8665.2749      1960.66         2.42015E-3      -
9000    -9218.4714      -941.2982       -8277.1732      1963.73         1.65709E-3      -
10000   -8864.5327      -930.5937       -7933.9390      1961.76         1.27192E-3      -
11000   -8496.3266      -947.4680       -7548.8585      1963.34         9.19785E-4      3.72 minutes/million states
12000   -8227.9452      -958.9548       -7268.9904      1961.83         6.3804E-4       3.74 minutes/million states
13000   -7782.2685      -966.6446       -6815.6239      1965.09         5.50098E-4      3.66 minutes/million states
```

Processing results

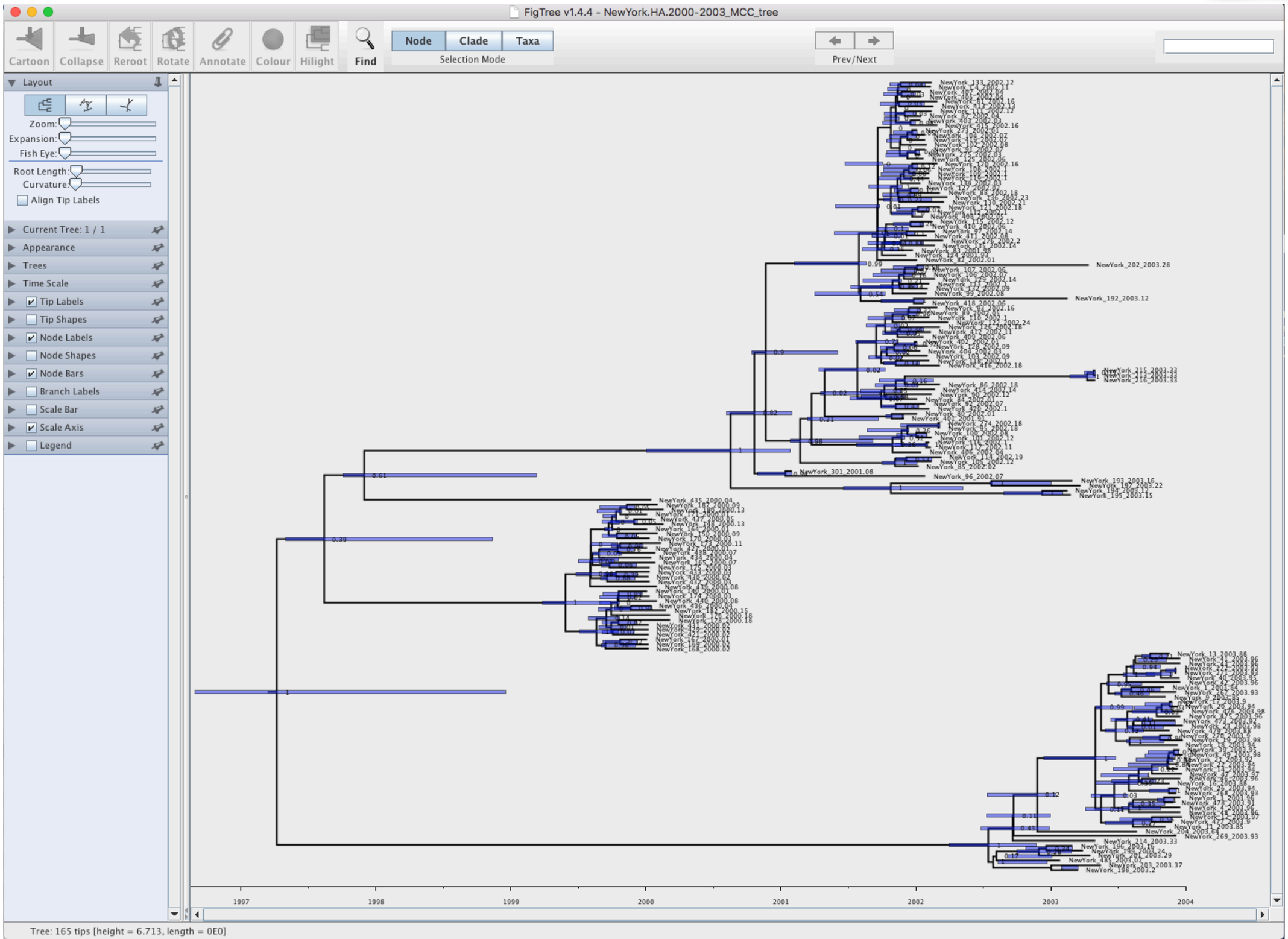
LogCombiner

Combines .log and .tree files from multiple (identical) runs

Also possible to combine log files in Tracer

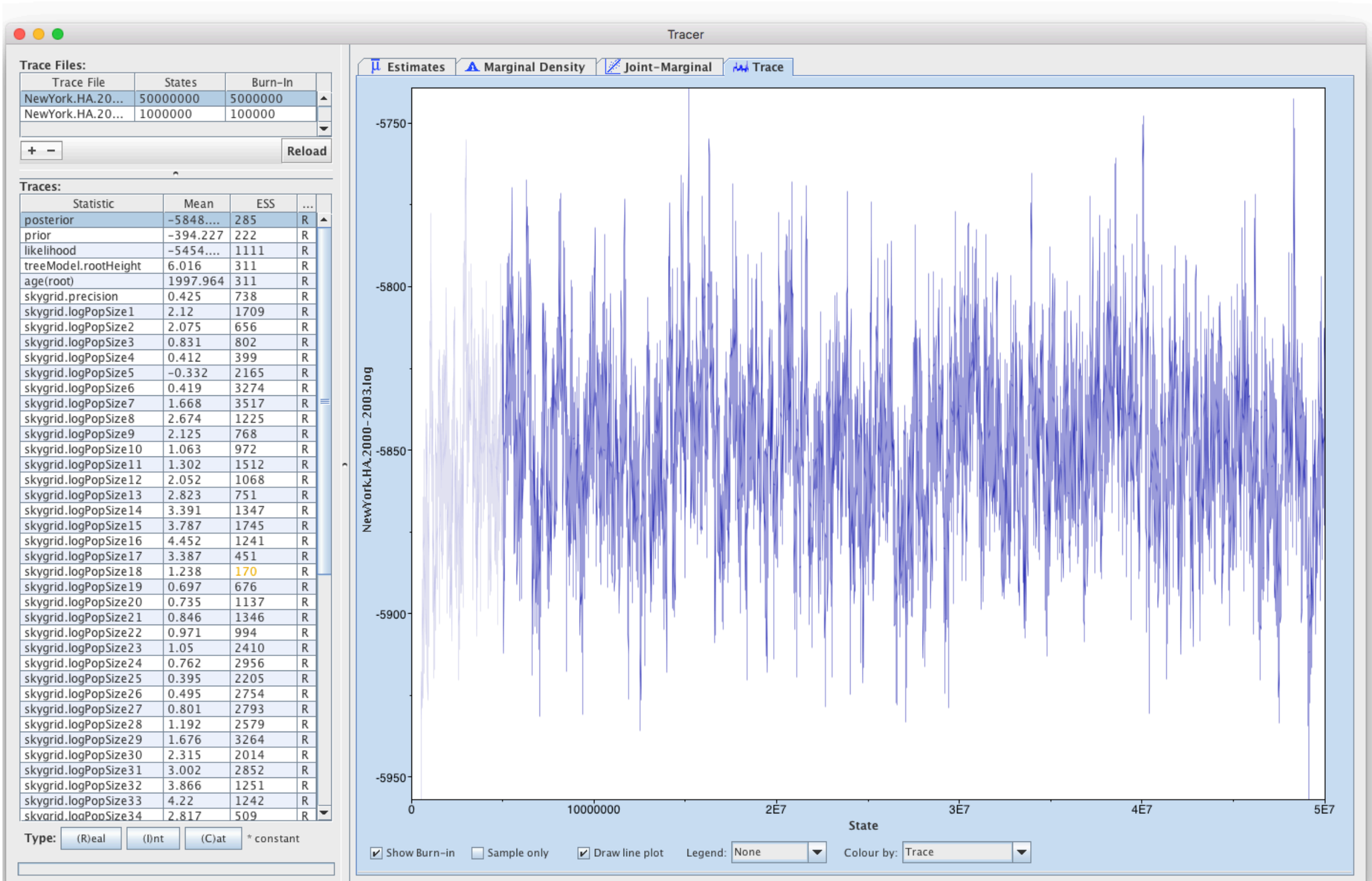


treeAnnotator



Summary of results and diagnostics

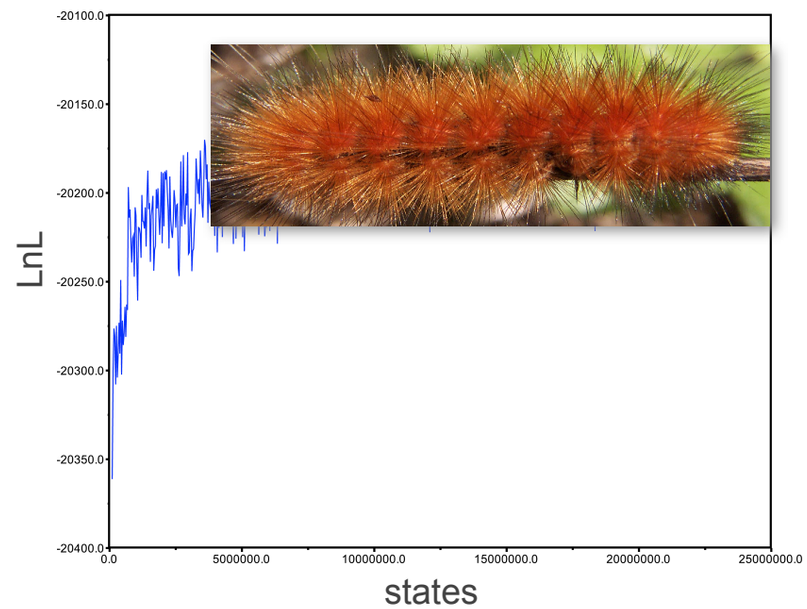
Summary of posterior distribution



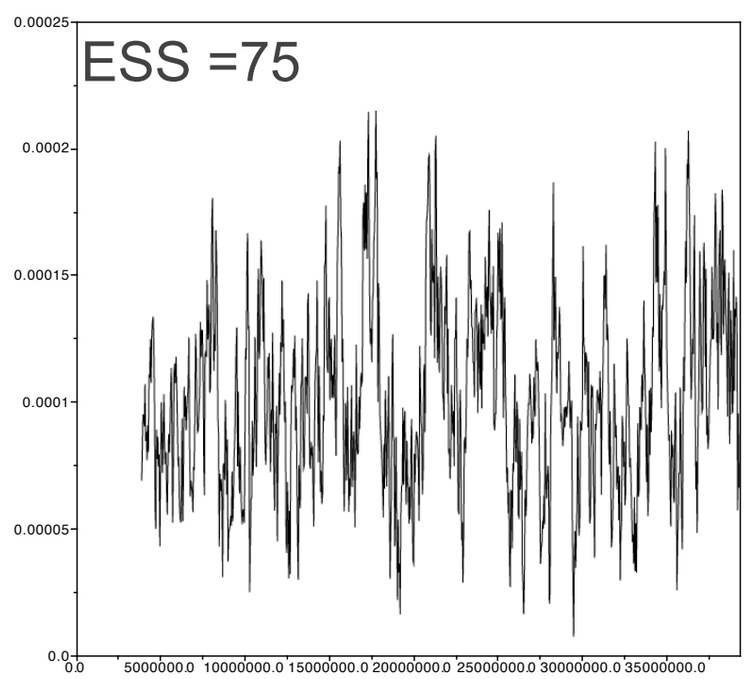
MCMC diagnostics

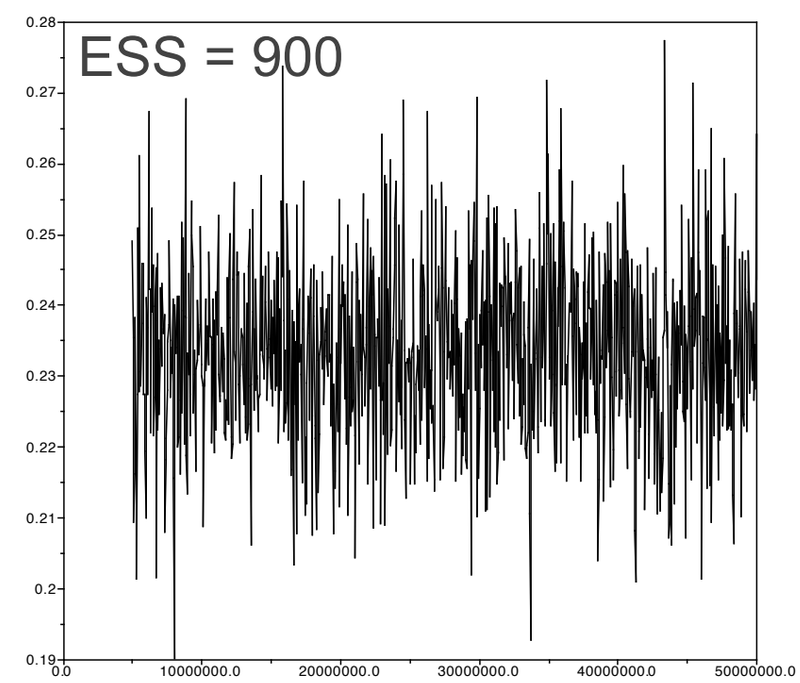
measure the behavior of a single chain

visually inspect output traces



measure autocorrelation within a chain: the effective sampling size (ESS)

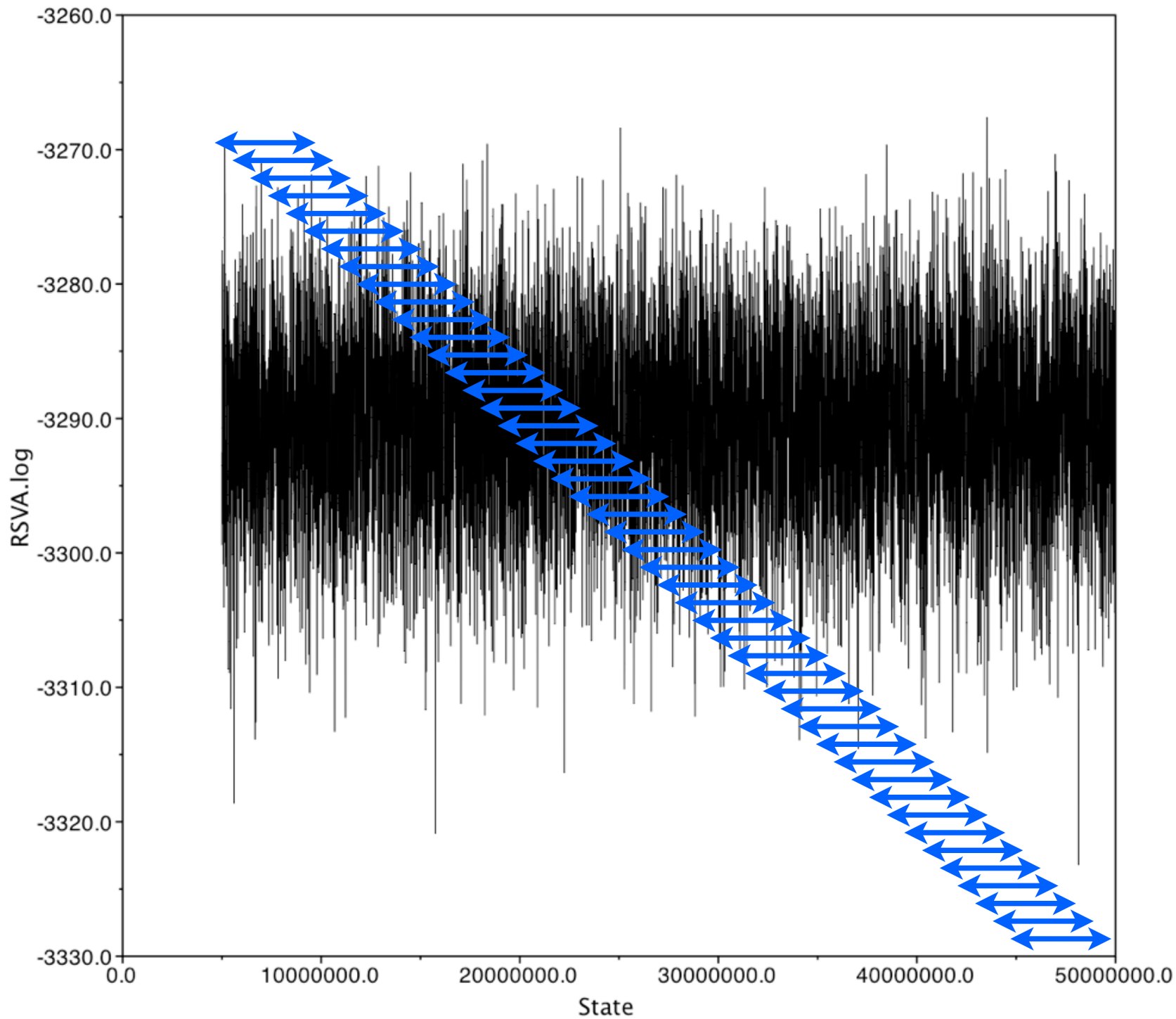




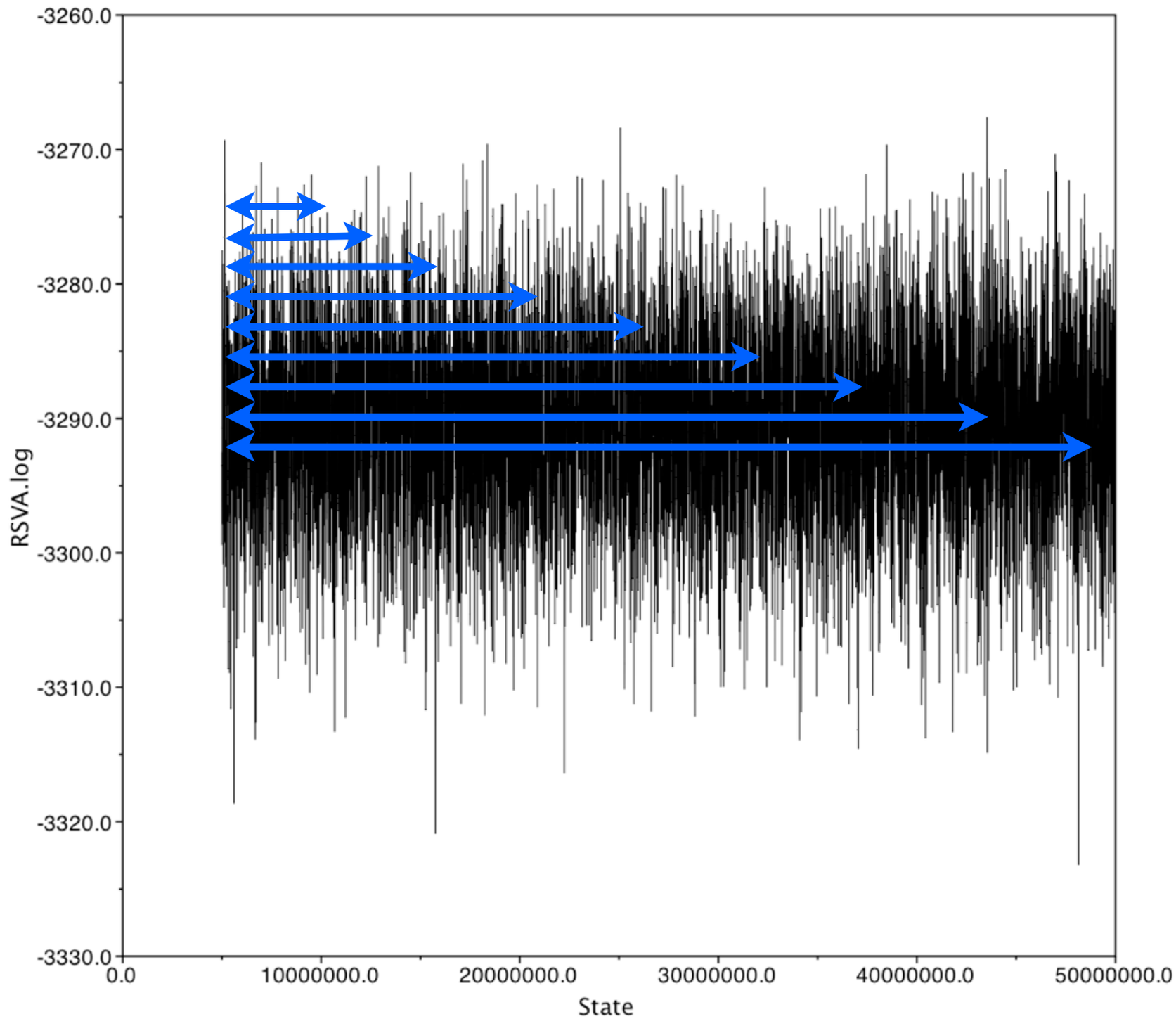
Summary of posterior distribution

- The effective sample size (ESS) is the number of independent draws from the (marginal) posterior distribution that the MCMC chain is equivalent to.
- Calculated as the number of samples divided by the Autocorrelation Time (ACT).

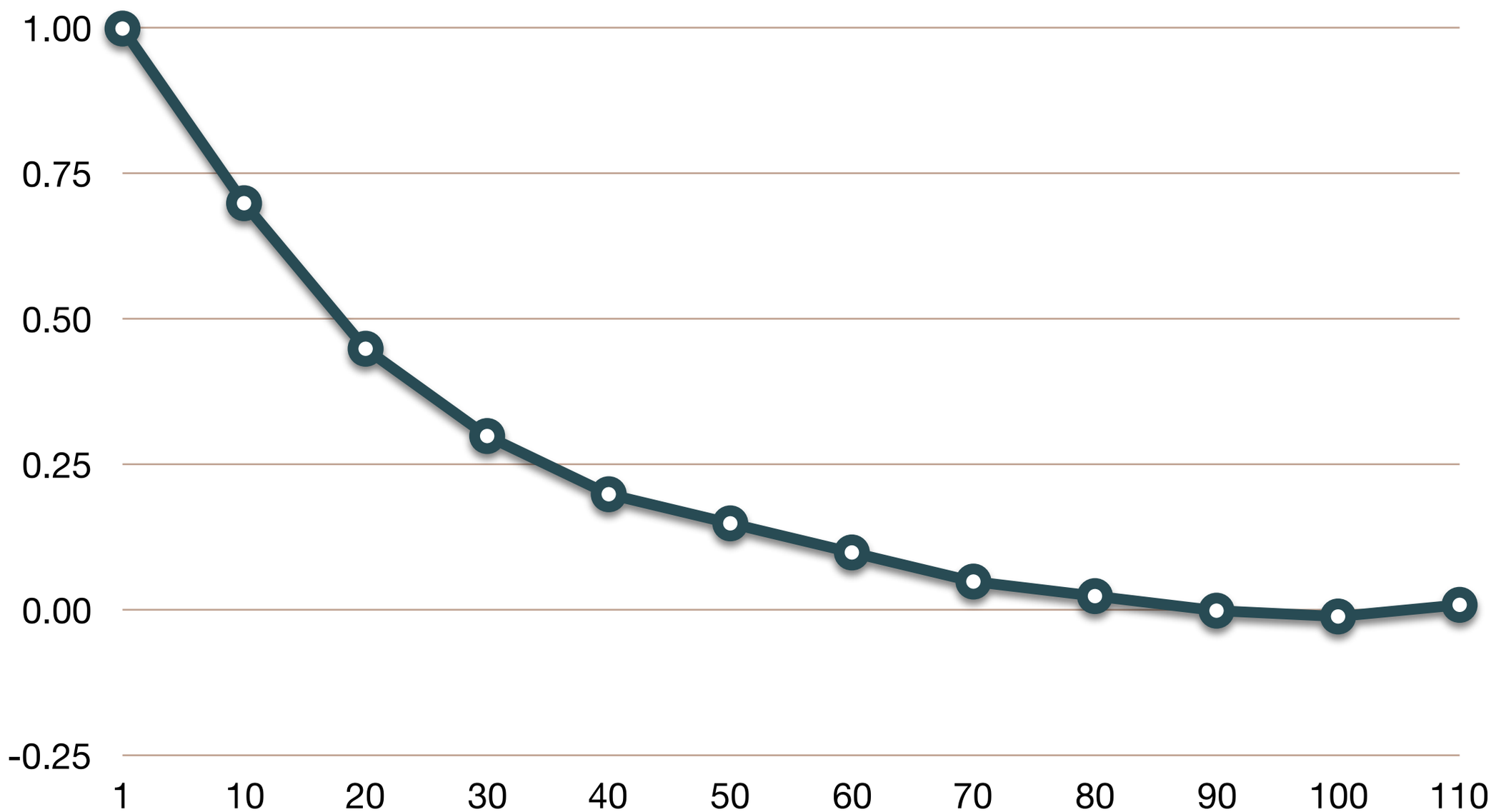
Autocorrelation in traces

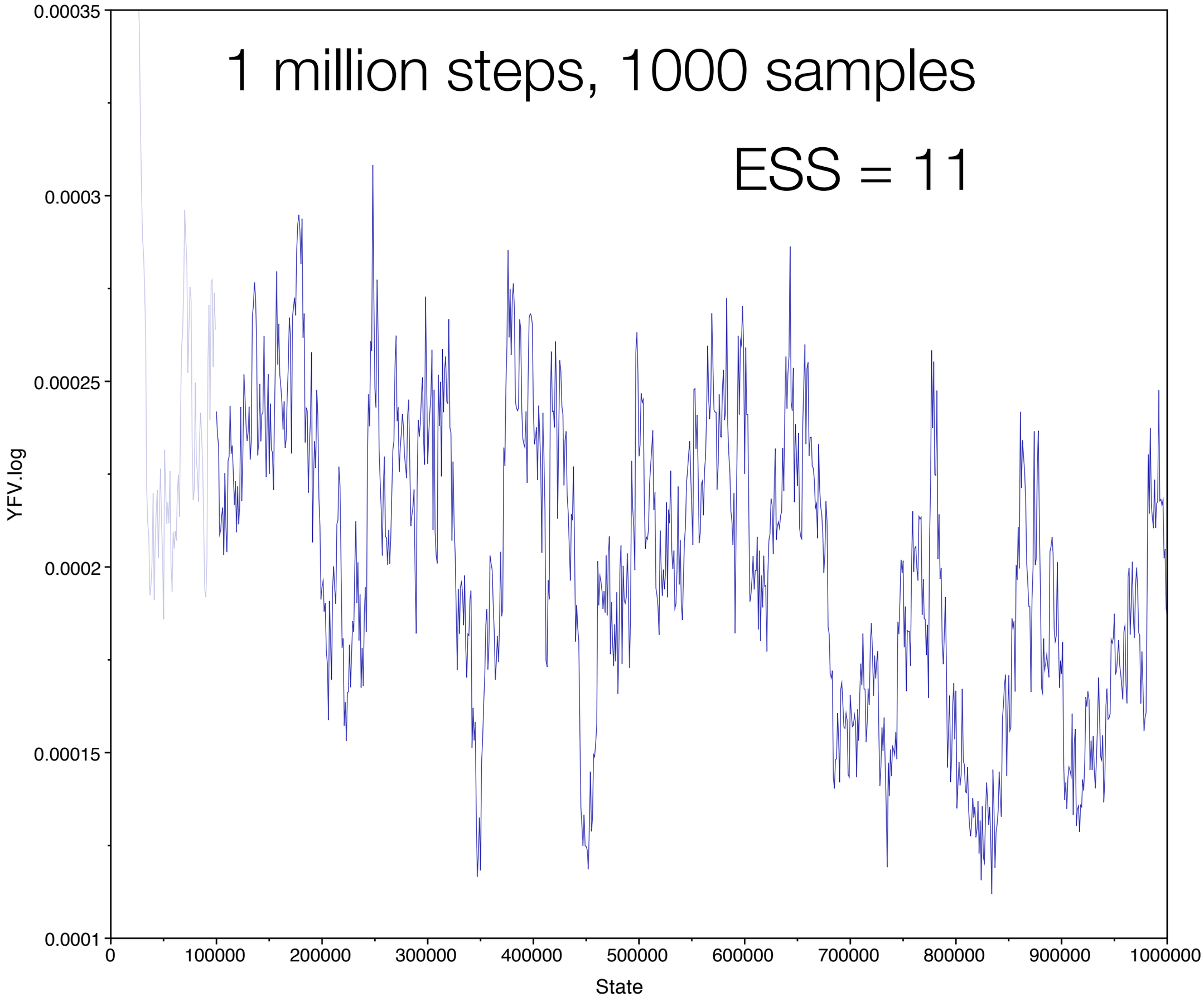


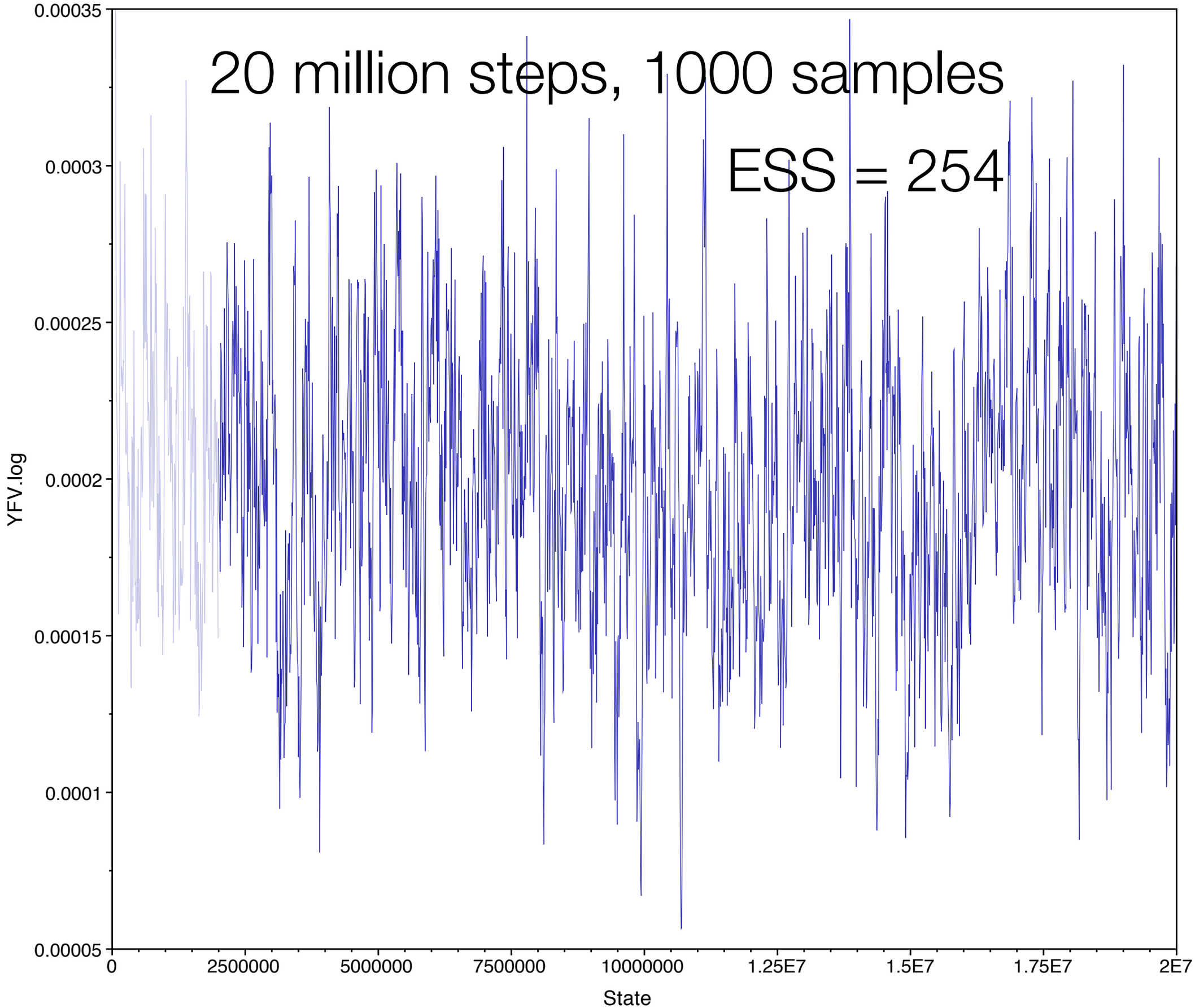
Autocorrelation in traces

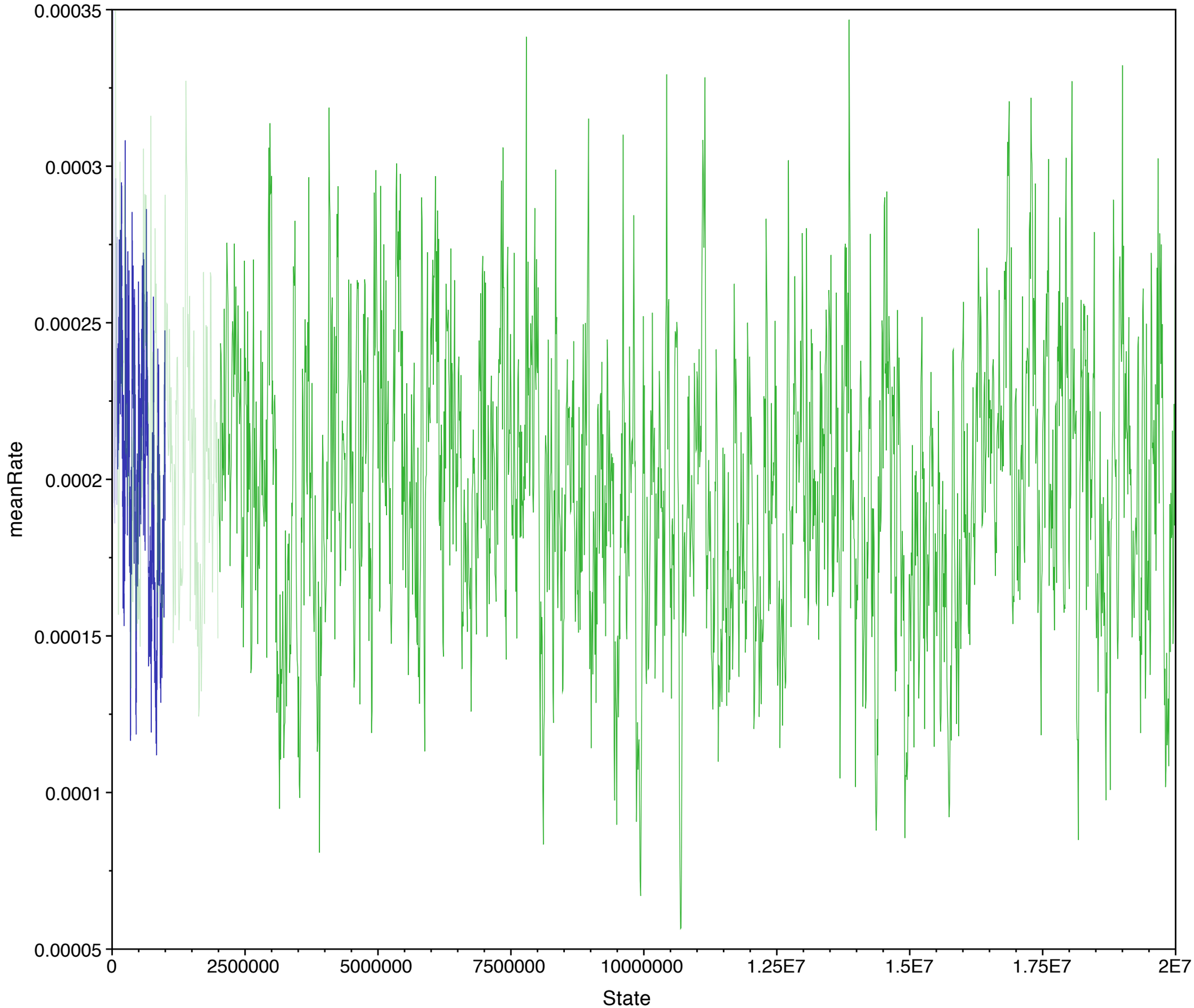


Autocorrelation time

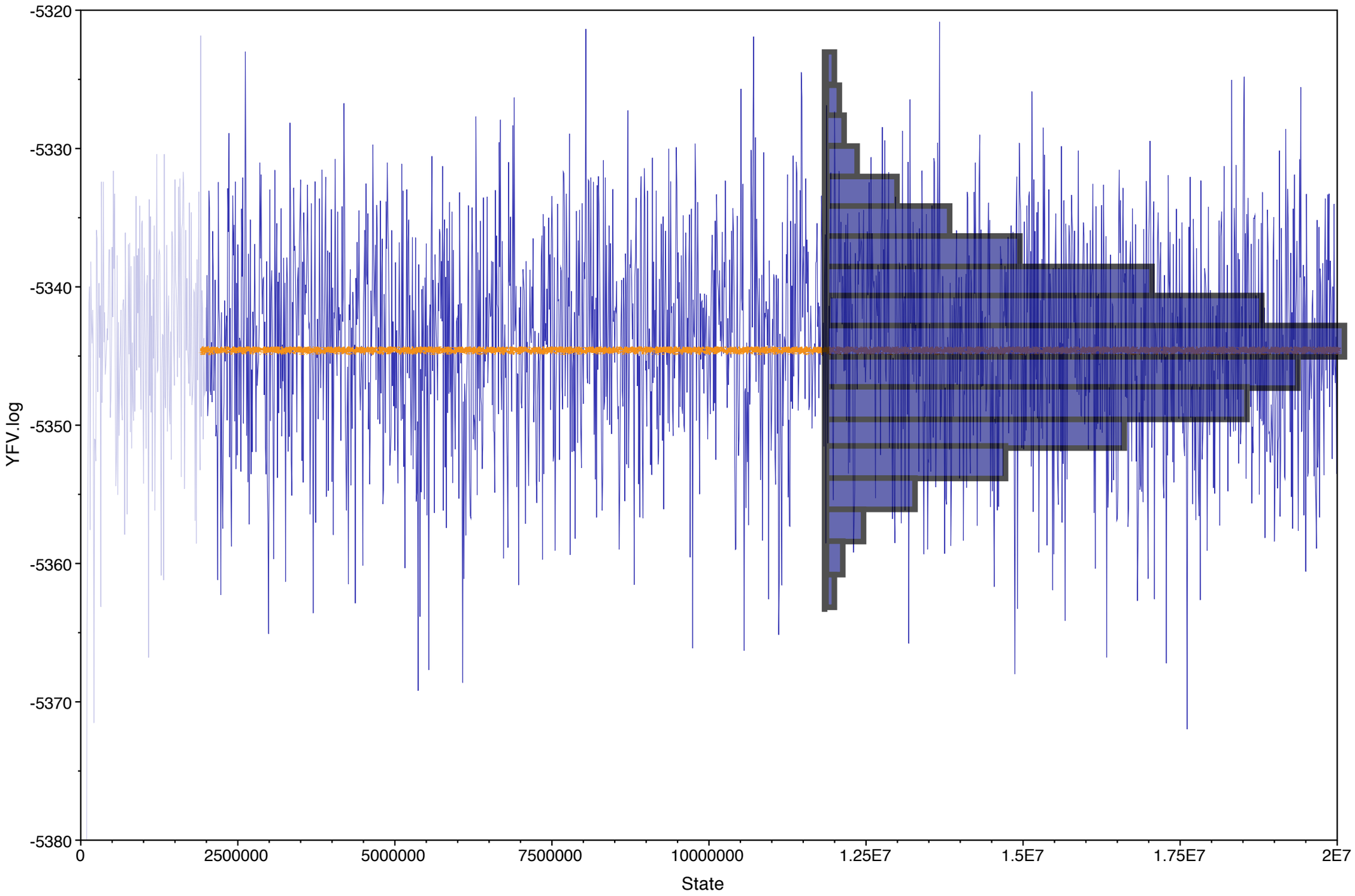



1 million steps, 1000 samples
ESS = 11
YFV.log
0.00035
0.0003
0.00025
0.0002
0.00015
0.0001
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
State

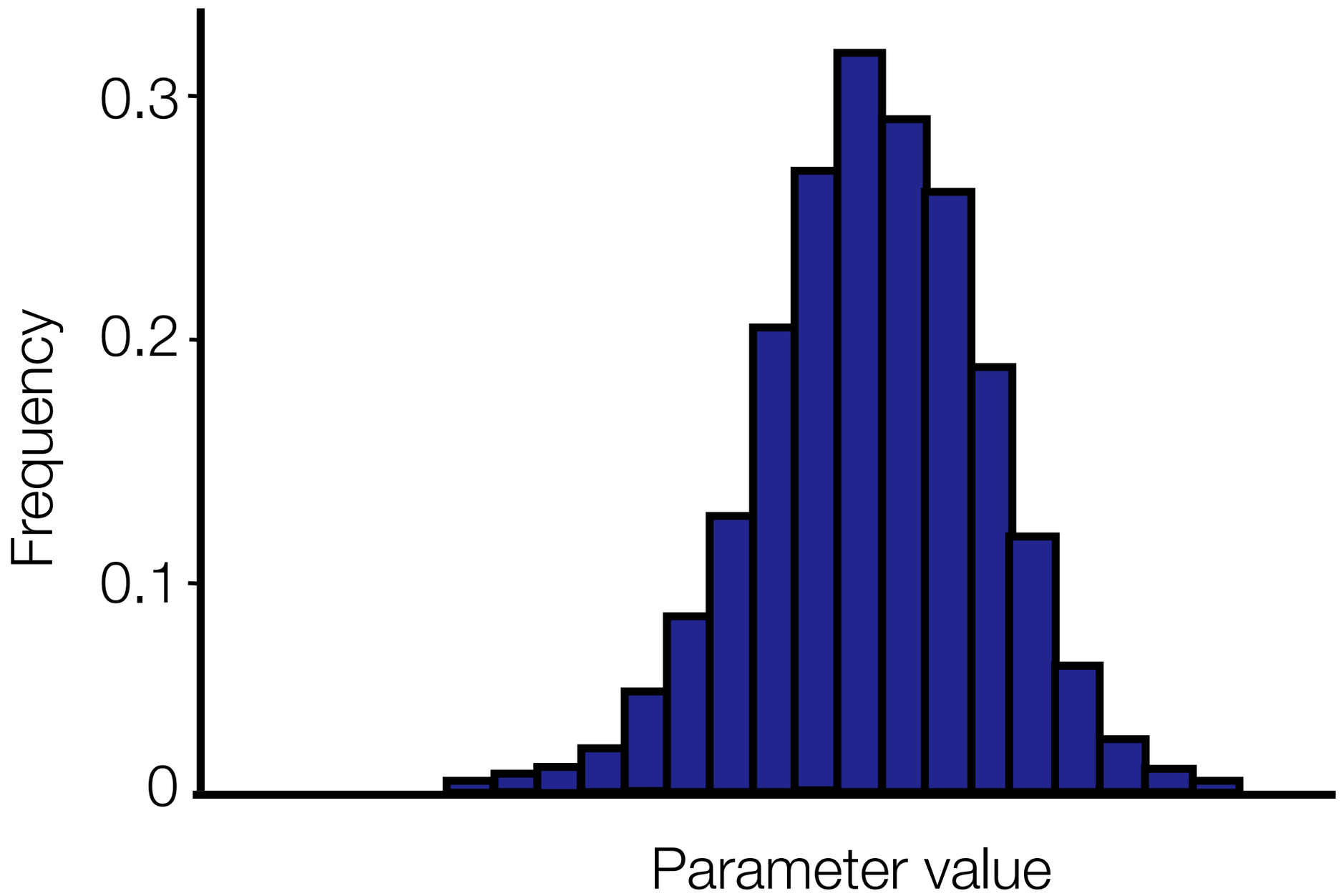

20 million steps, 1000 samples
ESS = 254
YFV.log
0.00035
0.0003
0.00025
0.0002
0.00015
0.0001
0.00005
0
2500000
5000000
7500000
10000000
1.25E7
1.5E7
1.75E7
2E7
State


0.00035
0.0003
0.00025
0.0002
0.00015
0.0001
0.00005
meanRate
0
2500000
5000000
7500000
10000000
1.25E7
1.5E7
1.75E7
2E7
State

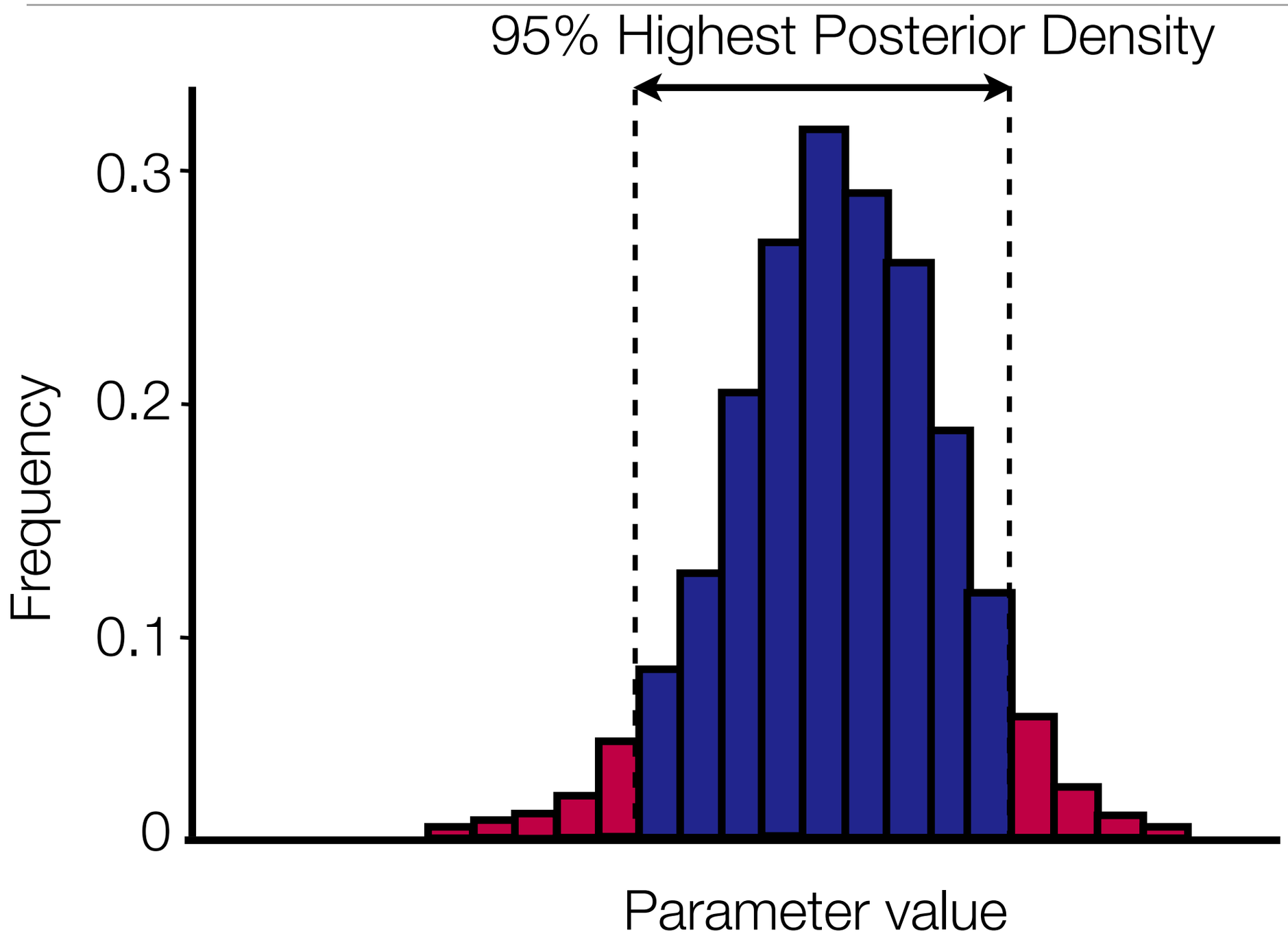
Markov chain of substitution rate

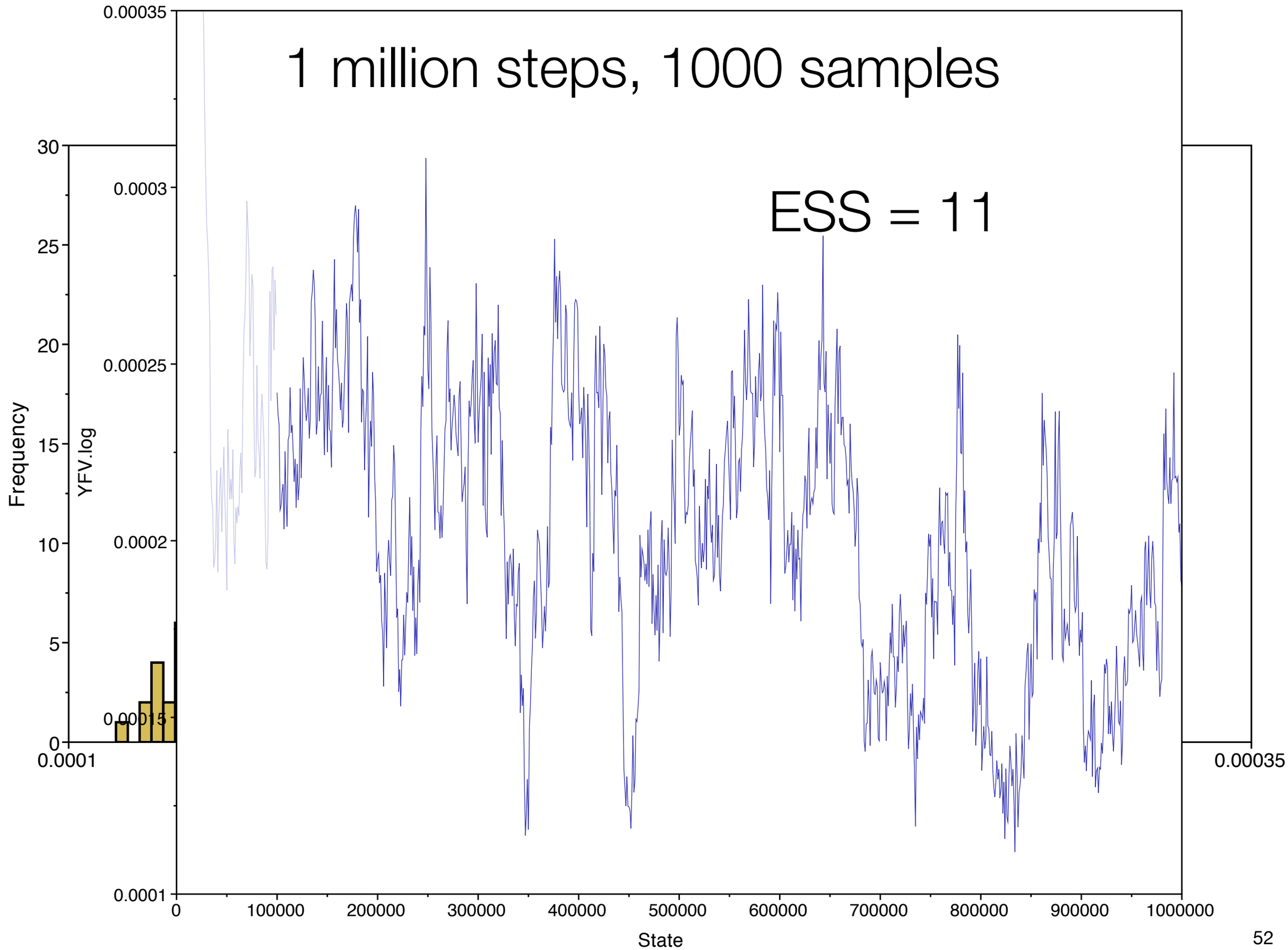


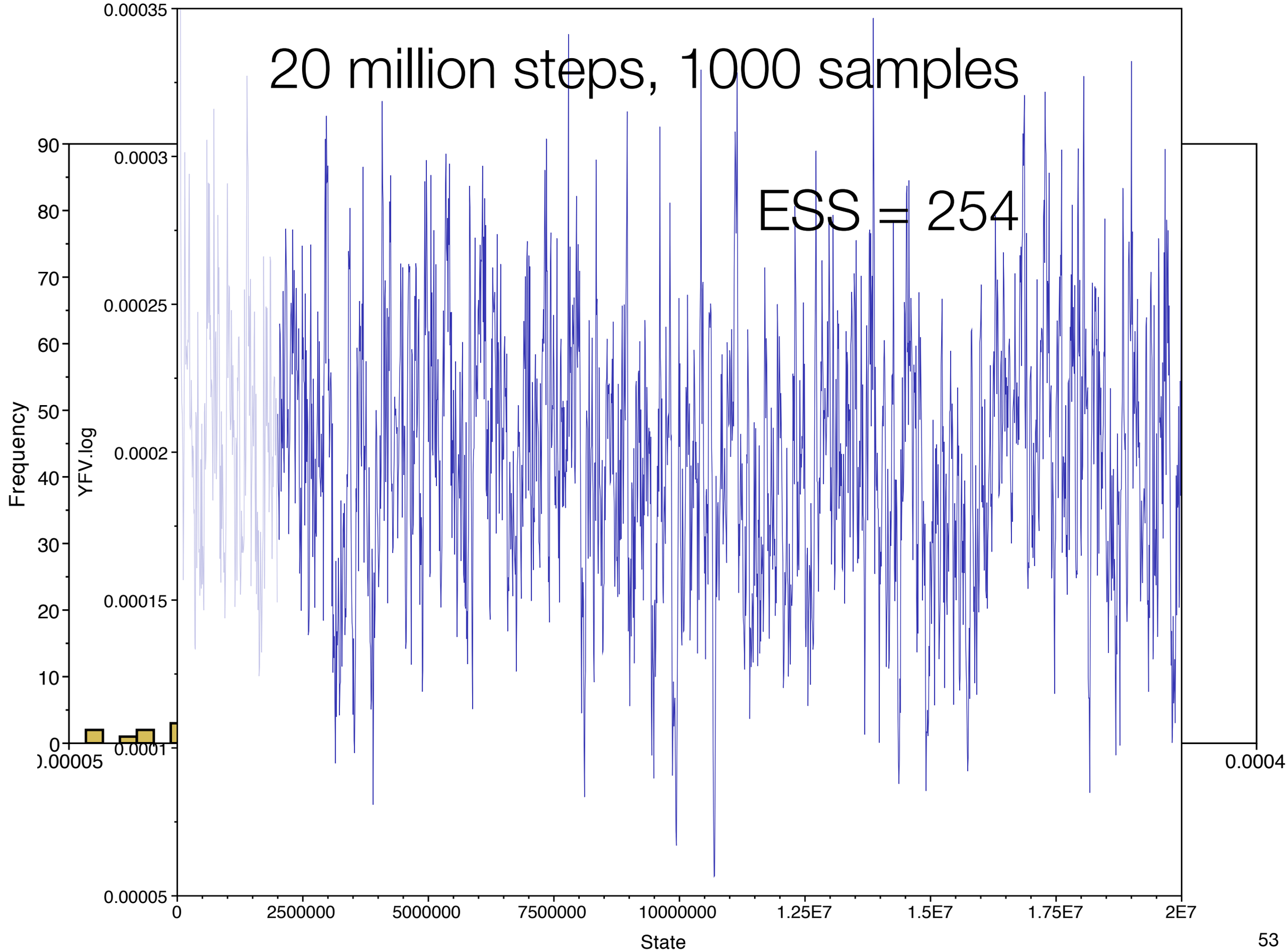
Summarizing parameters



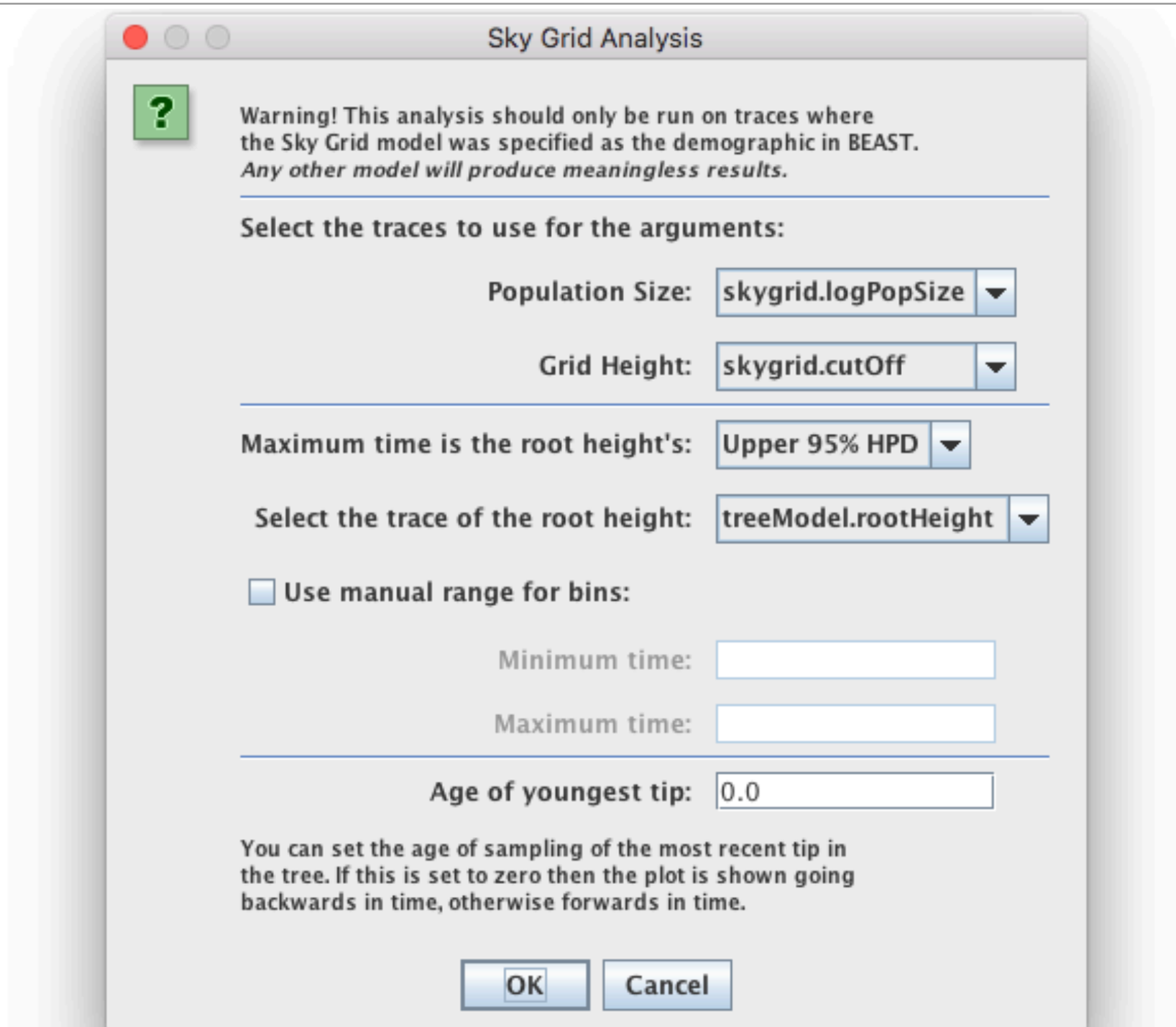
Summarizing parameters - credible interval



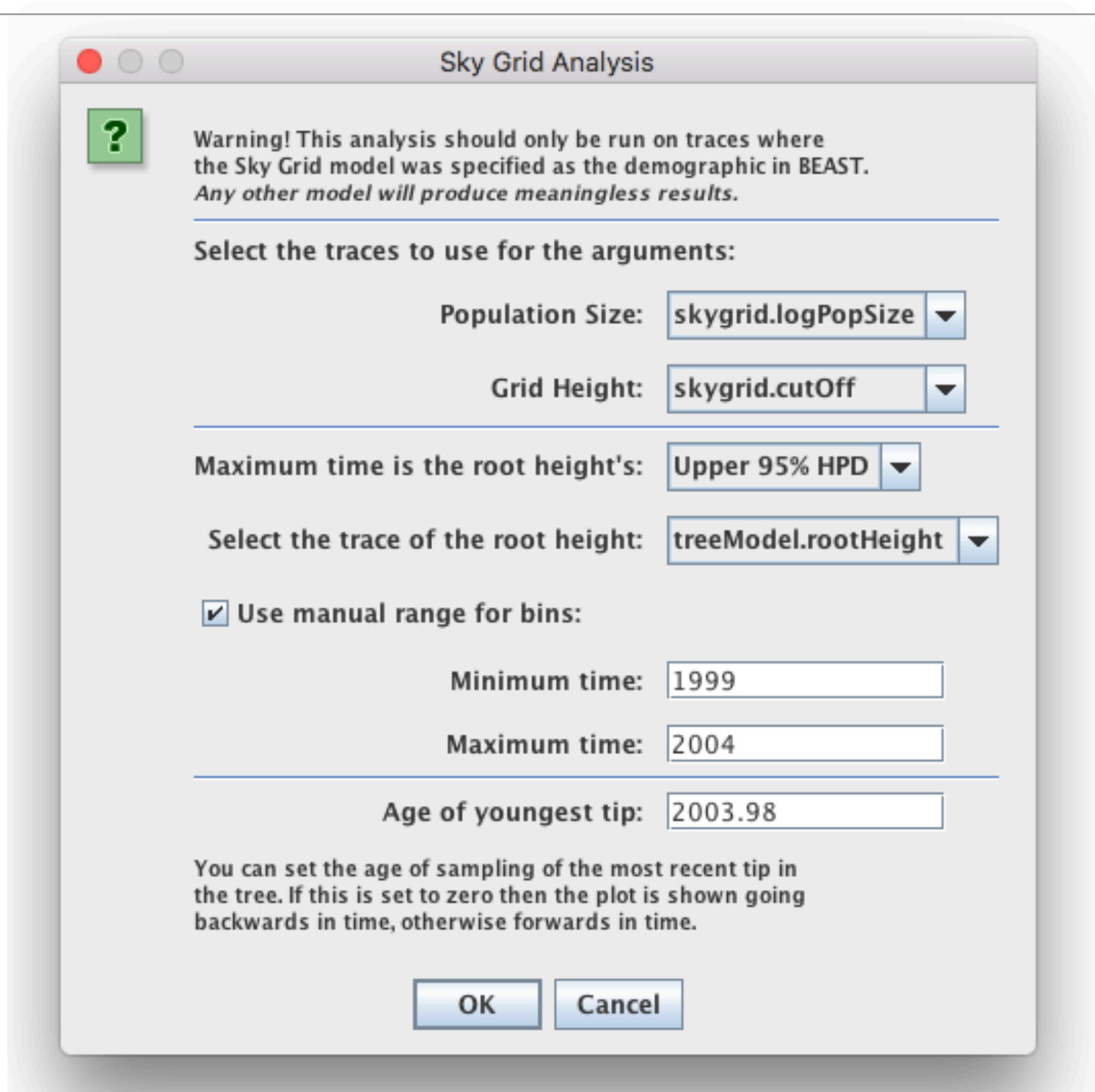
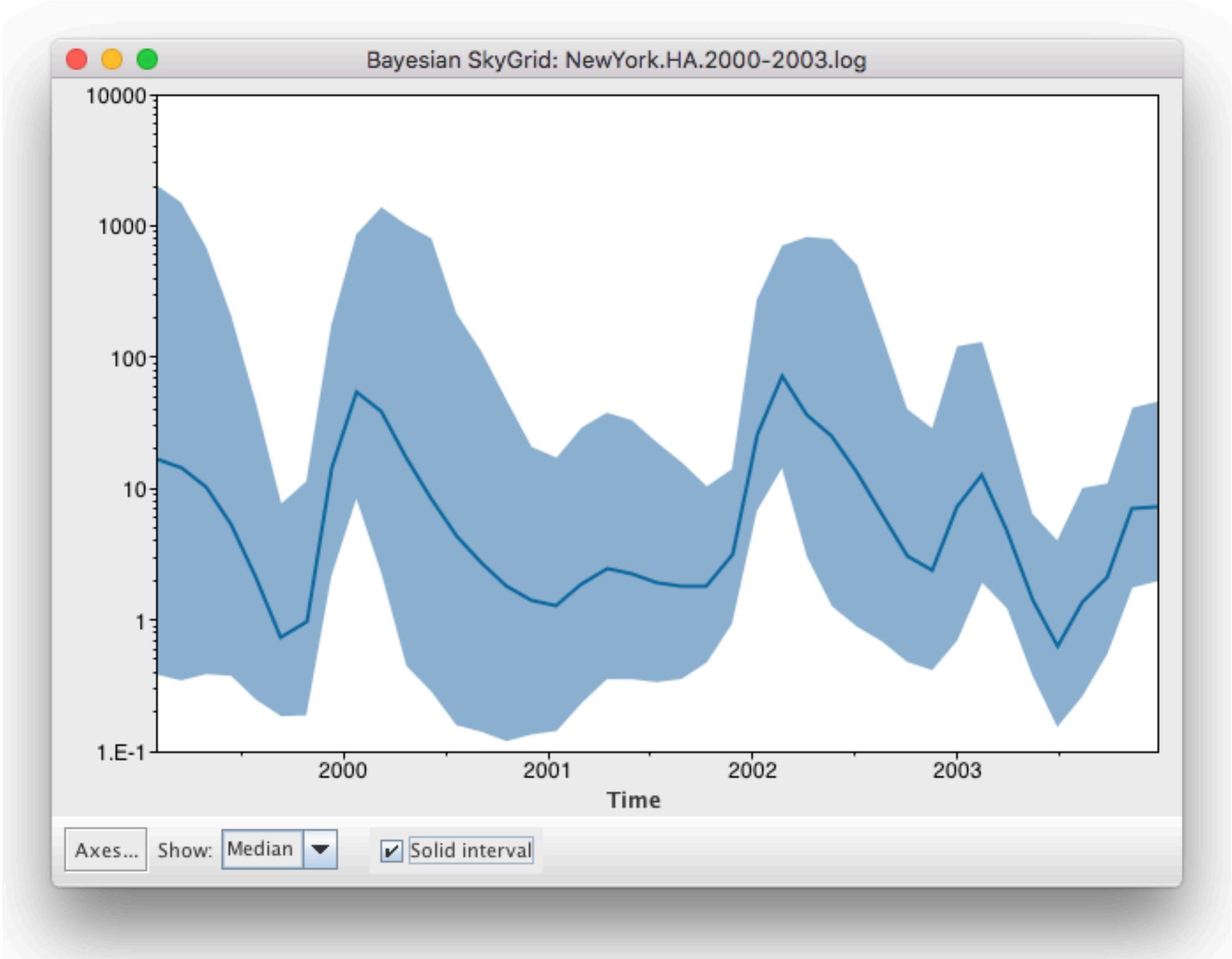

1 million steps, 1000 samples
ESS = 11
Frequency
YFV.log
State
0.00035
0.0003
0.00025
0.0002
0.00015
0.0001
30
25
20
15
10
5
0
0.0001
0.00035
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000


20 million steps, 1000 samples
ESS = 254
Frequency
YFV.log
State
0.00035
0.0003
0.00025
0.0002
0.00015
0.0001
0.00005
0
2500000
5000000
7500000
10000000
1.25E7
1.5E7
1.75E7
2E7
90
80
70
60
50
40
30
20
10
0
0.0004

Reconstruct the Bayesian skygrid plot



Reconstruct the Bayesian skygrid plot



Reconstruct the Bayesian skygrid plot