Phylogenetic tree construction and molecular clock analysis

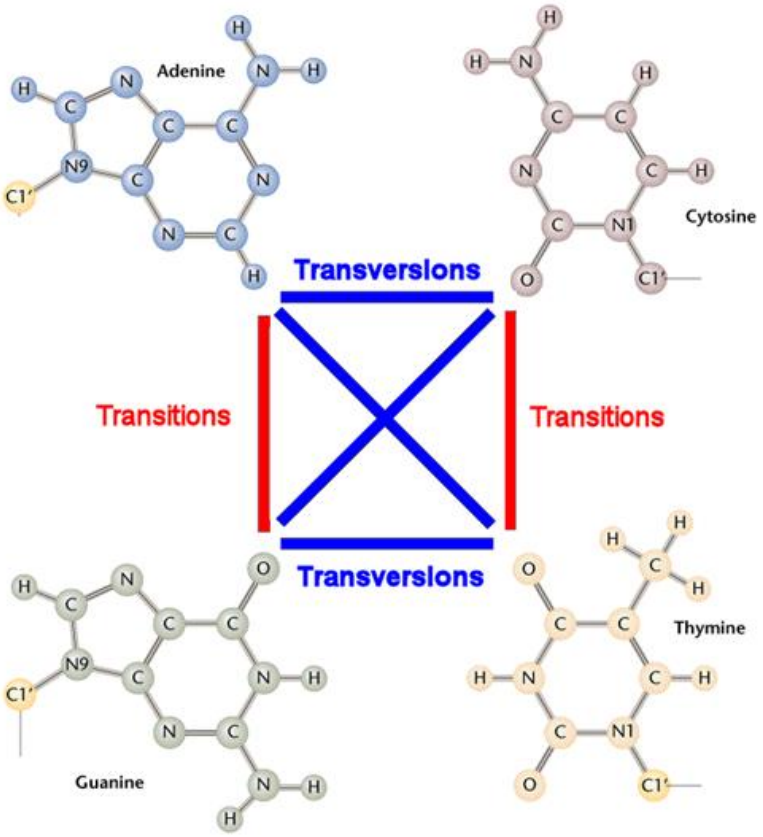
EDCTP ENNEA training on data management

November 14-16 2011, Muhimbili hospital, Dar Es Salaam

Irene Bontell

(irene.bontell@ki.se)

Nucleotide substitutions



Nucleotide substitution rates vary (a lot!) between genes and organisms and also depending on the position in the codon (synonymous / non-synonymous substitutions)

First Letter	Second Letter: T	Second Letter: C	Second Letter: A	Second Letter: G	Third Letter
T	TTT, TTC Phe; TTA, TTG Leu	TCT, TCC, TCA, TCG Ser	TAT, TAC Tyr; TAA Stop; TAG Stop	TGT, TGC Cys; TGA Stop; TGG Trp	T C A G
C	CTT, CTC, CTA, CTG Leu	CCT, CCC, CCA, CCG Pro	CAT, CAC His; CAA, CAG Gln	CGT, CGC, CGA, CGG Arg	T C A G
A	ATT, ATC, ATA Ile; ATG Met	ACT, ACC, ACA, ACG Thr	AAT, AAC Asn; AAA, AAG Lys	AGT, AGC Ser; AGA, AGG Arg	T C A G
G	GTT, GTC, GTA, GTG Val	GCT, GCC, GCA, GCG Ala	GAT, GAC Asp; GAA, GAG Glu	GGT, GGC, GGA, GGG Gly	T C A G

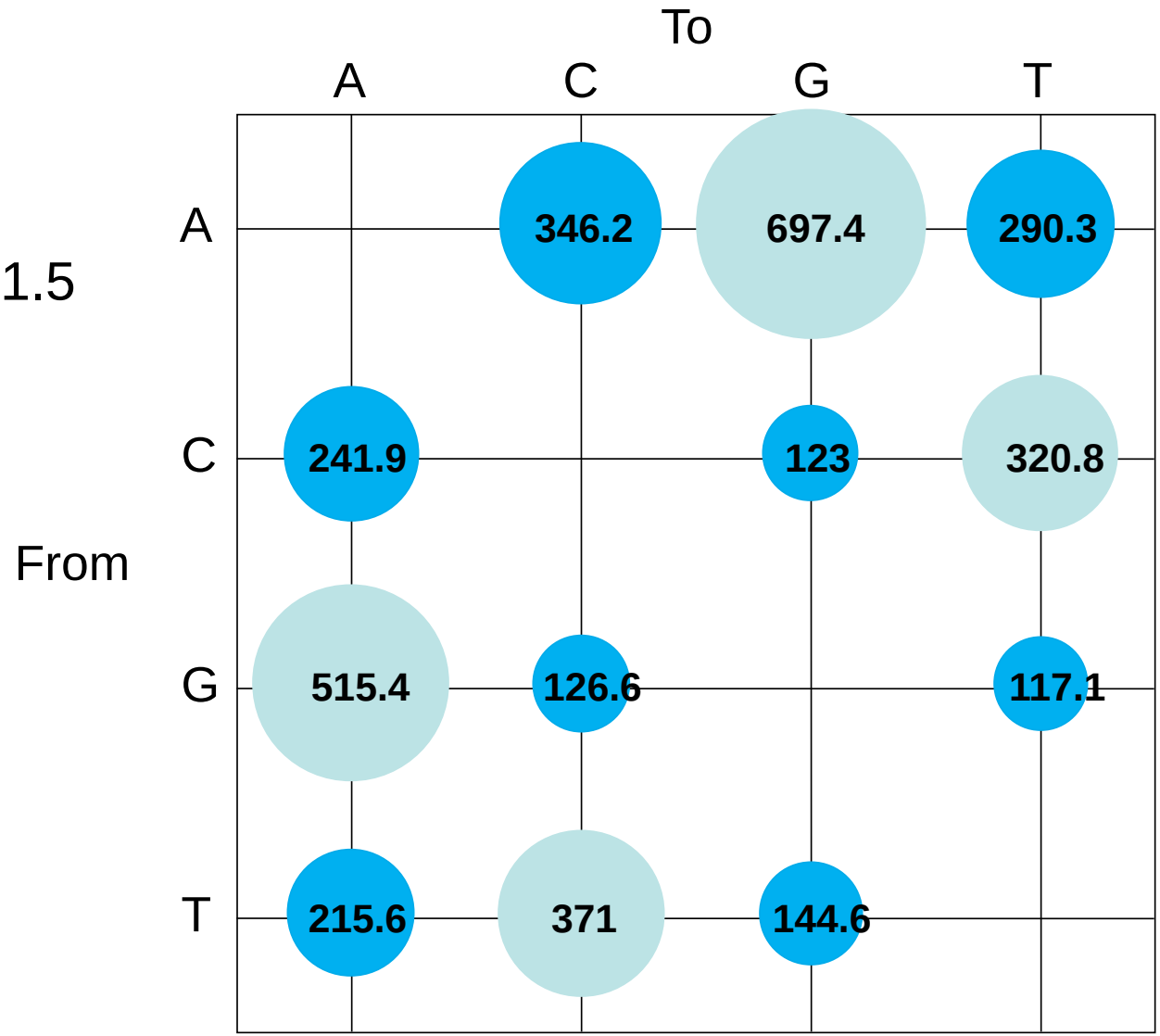
http://www.mun.ca/biology/scarr/Transitions_vs_Transversions.html

Average rates for HIV-1 env

(borrowed from presentation by Marco Salemi)

- Transitions
- Transversions

Average Ti/Tv=1.5



Nucleotide substitution models

Jukes-Cantor

	T	C	A	G
T	f_N	a	a	a
C	a	f_N	a	a
A	a	a	f_N	a
G	a	a	a	f_N

Kimura 2-parameter

	T	C	A	G
T	f_N	a	b	b
C	a	f_N	b	b
A	b	b	f_N	a
G	b	b	a	f_N

Hasegawa-Kishino-Yano

	T	C	A	G
T	f_T	a	b	b
C	a	f_C	b	b
A	b	b	f_A	a
G	b	b	a	f_G

General Time Reversible

	T	C	A	G
T	f_T	a	b	c
C	a	f_C	d	e
A	b	d	f_A	f
G	c	e	f	f_G

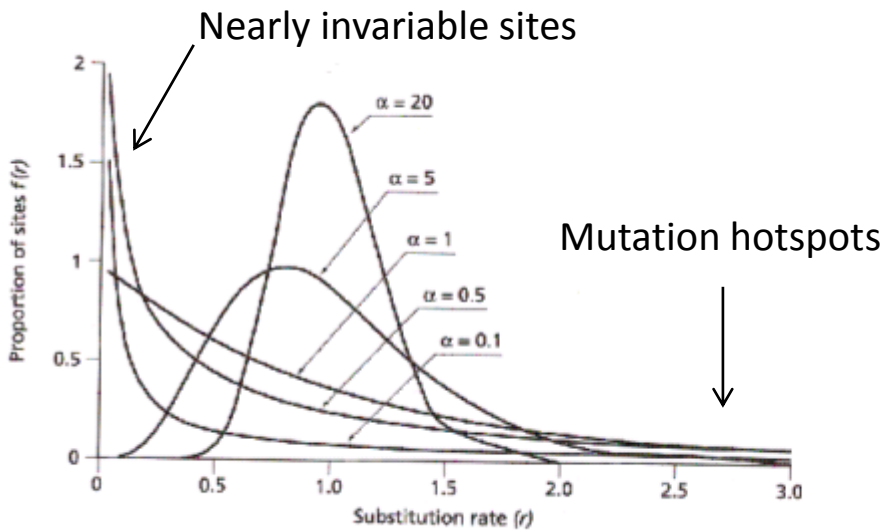


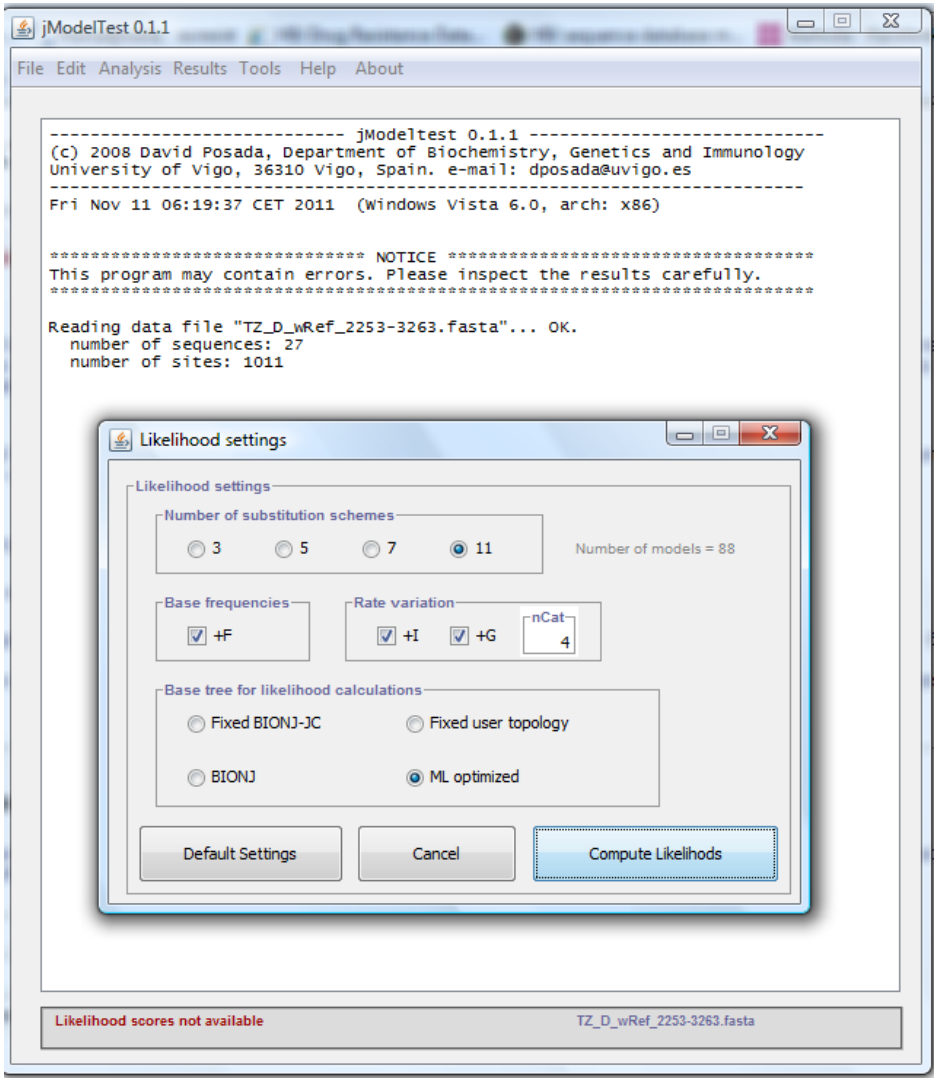
Fig. 2. Different forms of the gamma distribution are controlled by a single parameter alpha (the "shape parameter").

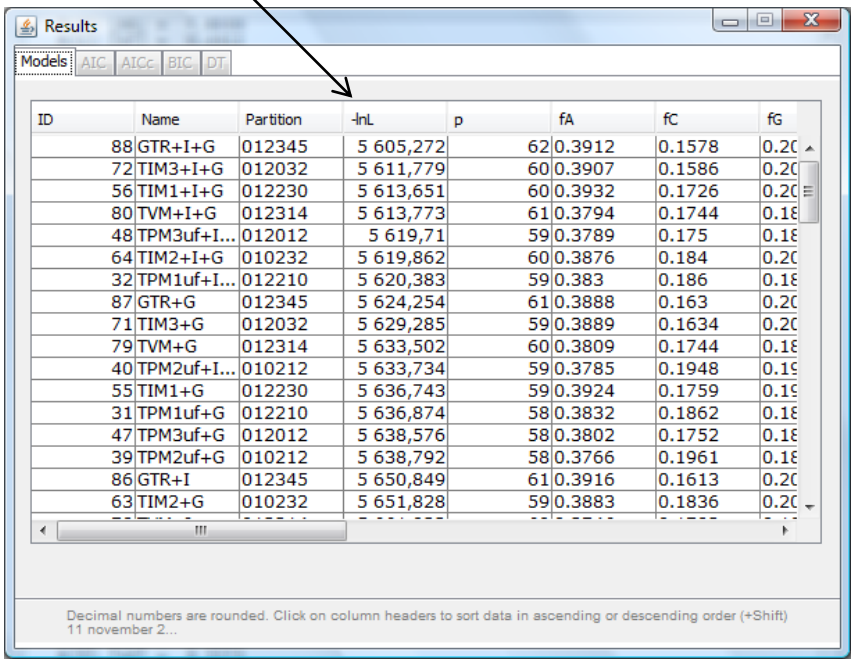
Rates of evolution are not equally distributed over sites. Gamma distributions with different values of alpha have been shown to be much more realistic. In addition a proportion of invariant sites can be added.
Tha most complex model is therefore GTR+Γ+I

http://www.hiv.lanl.gov/content/sequence/PHYML/showmodels.html

Modeltest

By running your data through jModeltest you get a maximum likelihood estimation of which nucleotid substitution model is most appropriate for your data





ID	Name	Partition	-lnL	p	fA	fC	fG
88	GTR+I+G	012345	5 605,272	62	0.3912	0.1578	0.20
72	TIM3+I+G	012032	5 611,779	60	0.3907	0.1586	0.20
56	TIM1+I+G	012230	5 613,651	60	0.3932	0.1726	0.20
80	TVM+I+G	012314	5 613,773	61	0.3794	0.1744	0.18
48	TPM3uf+I...	012012	5 619,71	59	0.3789	0.175	0.18
64	TIM2+I+G	010232	5 619,862	60	0.3876	0.184	0.20
32	TPM1uf+I...	012210	5 620,383	59	0.383	0.186	0.18
87	GTR+G	012345	5 624,254	61	0.3888	0.163	0.20
71	TIM3+G	012032	5 629,285	59	0.3889	0.1634	0.20
79	TVM+G	012314	5 633,502	60	0.3809	0.1744	0.18
40	TPM2uf+I...	010212	5 633,734	59	0.3785	0.1948	0.19
55	TIM1+G	012230	5 636,743	59	0.3924	0.1759	0.19
31	TPM1uf+G	012210	5 636,874	58	0.3832	0.1862	0.18
47	TPM3uf+G	012012	5 638,576	58	0.3802	0.1752	0.18
39	TPM2uf+G	010212	5 638,792	58	0.3766	0.1961	0.18
86	GTR+I	012345	5 650,849	61	0.3916	0.1613	0.20
63	TIM2+G	010232	5 651,828	59	0.3883	0.1836	0.20

```
Maximum likelihod estimation for the GTR+I+G model
  ML optimized tree topology
   Model = GTR+I+G
   partition = 012345
   -lnL = 5605.2715
   K = 62
   freqA = 0.3912
   freqC = 0.1578
   freqG = 0.2043
   freqT = 0.2466
   R(a) [AC] =  2.1800
   R(b) [AG] =  8.3273
   R(c) [AT] =  0.4242
   R(d) [CG] =  1.1252
   R(e) [CT] =  14.3022
   R(f) [GT] =  1.0000
   p-inv = 0.4670
   gamma shape = 1.0210
  Computation time = 00h:01:05:05  (00h:22:42:04)
```

FindModel

Purpose: FindModel analyzes your alignment to see which phylogenetic model best describes your data; this model can then be used to generate a better tree.

File size limits: Finding the best evolutionary model is computationally intensive, so we limit default runs to a reduced set of 12 models, excluding those that do not have an obvious biological interpretation. (If you know of any system where Modeltest consistently returns a model that we do not include, please let us know.) The full set of 28 models can be run by checking the checkbox. Currently, input files smaller than 6 Kb for the reduced set and 3 Kb for the full set are run immediately; if your input file exceeds the limit, your job will be run in batch, and you will receive an e-mail with a link to your results. Currently, input files larger than **500 Kb** for the **reduced** set and **350 Kb** for the **full** set are too large for our machine to process.

Formats: FindModel only accepts aligned nucleotide input. The program attempts to automatically recognize the format of your input file. See below for troubleshooting tips.

Input

Paste your input here
[Sample input]

or upload your file C:\Users\Irene\Desktop\Tanzania, 14-15 Nov\Practise Browse...

Options

Use all 28 models ☑

Construct initial tree using ◉ Neighbor ○ PAUP* ○ MrBayes

Always e-mail results ☐

Submit Reset

This is an alternative to jModelTest. FindModel is an online tool (no need to download and install). It does not test as many models, but is usually good enough

Findmodel parameter details

Parameters in the rate matrix (REV) (Yang 1994 J Mol Evol 39:105-111):

Rate parameters: 1.33969 0.04532 0.12403 0.23970 0.16154
Base frequencies: 0.23735 0.15797 0.39257 0.21211
Rate matrix Q, Average Ts/Tv = 4.3421
-0.778521 0.644268 0.054164 0.080089
0.967994 -1.358762 0.286461 0.104307
0.032748 0.115274 -0.793739 0.645717
0.089619 0.077685 1.195077 -1.362381

alpha (gamma, K=4) = 0.22449
r: 0.00113 0.04899 0.44452 3.50536
f: 0.25000 0.25000 0.25000 0.25000

Tree building methods

Distance matrix methods

Unweighted Pair Group Method with Arithmetic Mean (UPGMA): assumes constant rate of evolution, mostly used for guide trees

Neighbor-joining (NJ): Starts with starlike phylogeny with equal distance between all taxa. Distance matrix for all taxa, closest neighbours joined in a node, new matrix calculated, new node for closest neighbours and so on.

+Computationally efficient, can be used on very large datasets.
-Produce only one tree, no idea if there were other good trees.
-Not possible to estimate ancestral character states

Tree building methods

Character state methods

(evaluation of candidate trees “tree space” according to specific criteria.Scoring of each tree, trying to find the tree that optimizes the criteria (minimum evolution or maximum likelihood)

Maximum parsimony: the preferred tree is the one that requires the least evolutionary change. Often used for morphological data.

Maximum likelihood: the tree that has the highest probability of producing the observed data is the most likely tree. Often used for molecular data.

Bayesian: generates a posterior probability for each tree, which is based on the likelihood and incorporated prior knowledge.
(see http://www.whoi.edu/cms/files/stiwari/2006/8/mrbayesIntro_MBL05_12986.pdf)

The number of possible trees increase rapidly

For n taxa there are $(2n-3)!/[(2^{n-2})*(n-2)!]$ rooted, binary trees

Number of taxa	Number of possible tree topologies	
4	15	
5	105	
10	34 459 425	upper limit for exhaustive search
48	3.2e70	approximately the number of particles in the known universe

In LANL 10 nov 2011:		**No of possible trees**
Complete genomes	2832	6.4e7682
Pol 2253-3263	59508	too big for the calculator...
V3	133578	even bigger!

Not an option to include even 1% of all available sequences

It is necessary to make a careful choice when selecting reference sequences. Some subtype references are good as a framework (remember to include some outgroup sequences), otherwise include sequences from the subtype/geographic region etc that you are interested in, and if possible from a fairly large tine span for more accurate dating using the molecular clock.

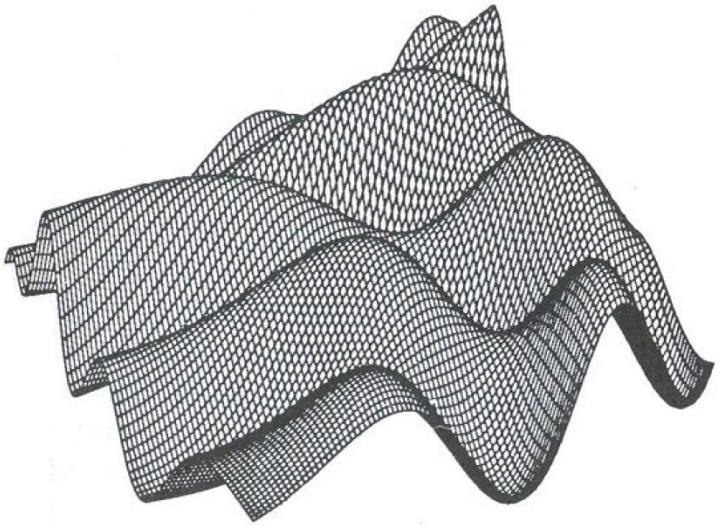
How do we find the best tree? (or at least a good one...)

Exhaustive search – only possible for very small data sets.

Branch-and-bound – finds the best tree by checking all possible trees (cutting of paths in the search tree that cannot possibly lead to optimal trees). Also very computationally demanding for large data sets

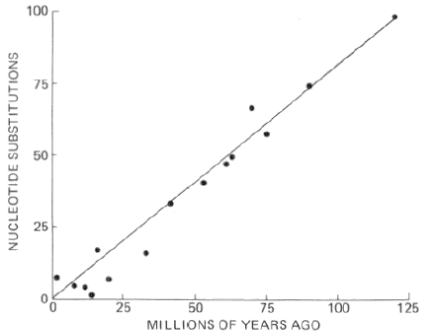
Heuristic – searches tree space by making small random changes to a starting tree, and accepting the new tree according to defined criteria. This method does not check all trees and cannot guarantee to find the best one.

There is always a risk of getting stuck on a local optimum rather than finding the global optimum. In Bayesian phylogenetics one does not search for a single optimal tree but a set of plausible trees.
Introduction of priors limit the search space



Molecular clocks

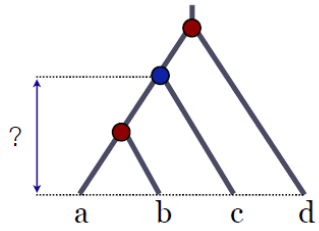
- Hypothesis: for any given macromolecule the rate of evolution is approximately constant over time in all evolutionary lineages (First proposed by Zuckerkandl and Pauling 1962)



from AC Wilson, 1976

- If this is true this can be used to estimate the time of the most recent common ancestor (tMRCA) using genetic sequence from organisms alive today
- However, there is no global clock. Homologous genes in species with different lifespans, metabolic rates etc have different mutation rates.
- Also problematic at very short and very long time-scales (due to non-fixed mutations / saturation)
- Nevertheless, the concept is very useful and local clocks can be applied within systems.

Molecular clock



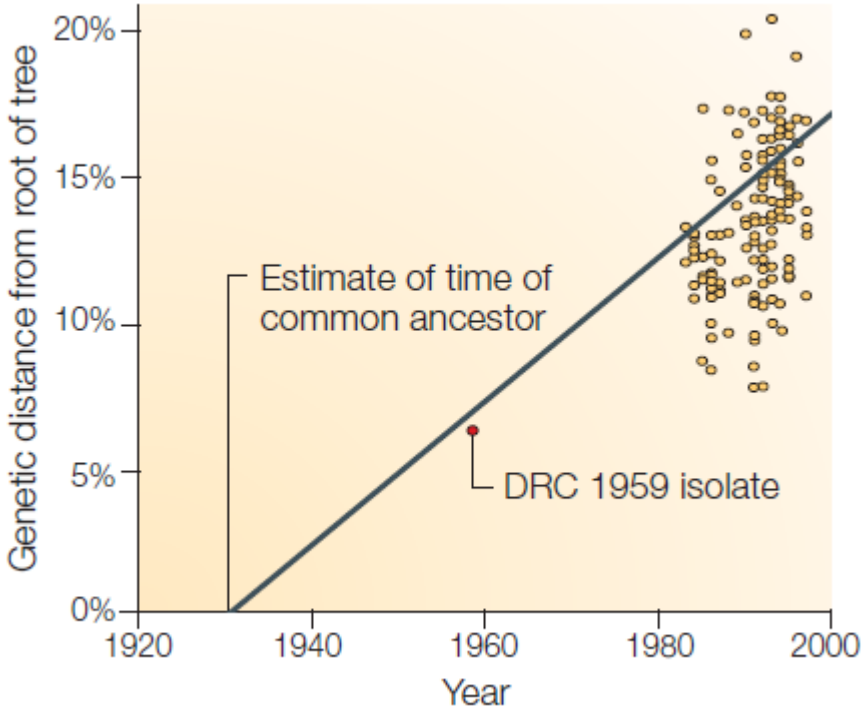
The **substitution/fixation rate (μ)** is the rate at which sequences in different *populations* diverge through time.
The **mutation rate (m)** is the rate at which *individuals incorporate errors are incorporated* during replication.
The **probability of fixation** determines the difference between them.

k = N μ p = substitution rate (per generation)
μ = mutation rate (per individual per generation)
p = probability of fixation
N = population size (2N for diploids)

Calibrating molecular clocks

RNA viruses are unusual since their extremely high mutation rate leads to measurably evolving populations.

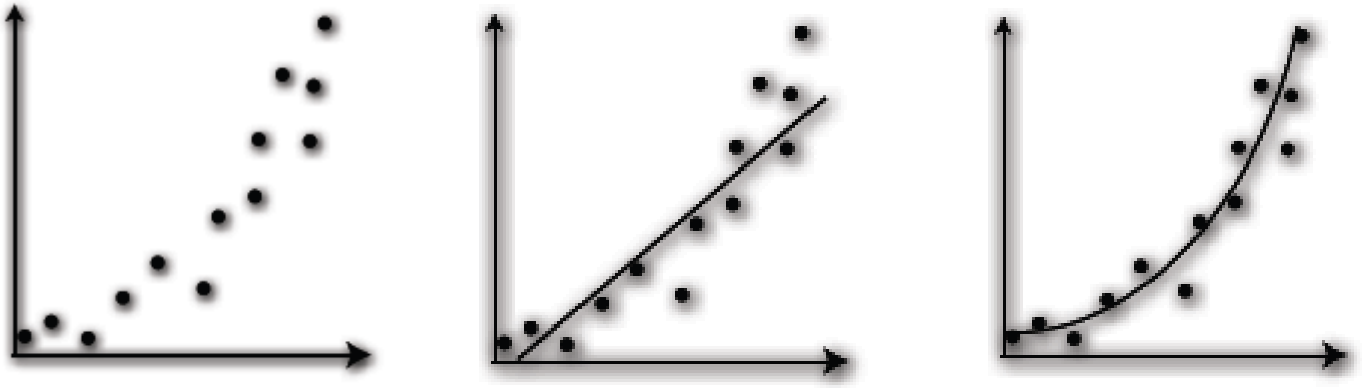
Therefore it is possible to use **tip dates** for calibration



Korber et al, 2000.Science

Relaxing the molecular clock

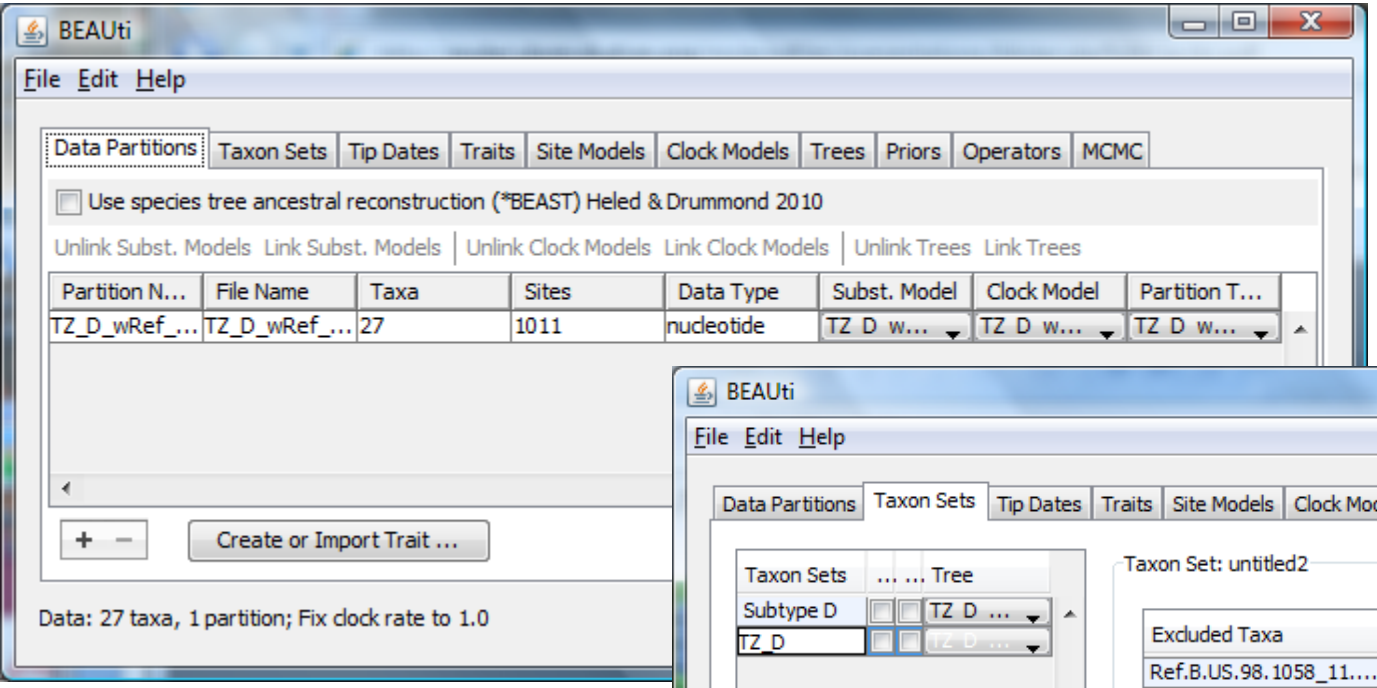
- Instead of assuming a constant rate as for the fixed clock, the relaxed clock draws the rate for each branch of the tree independently from an identical distribution (exponential or log-normal)

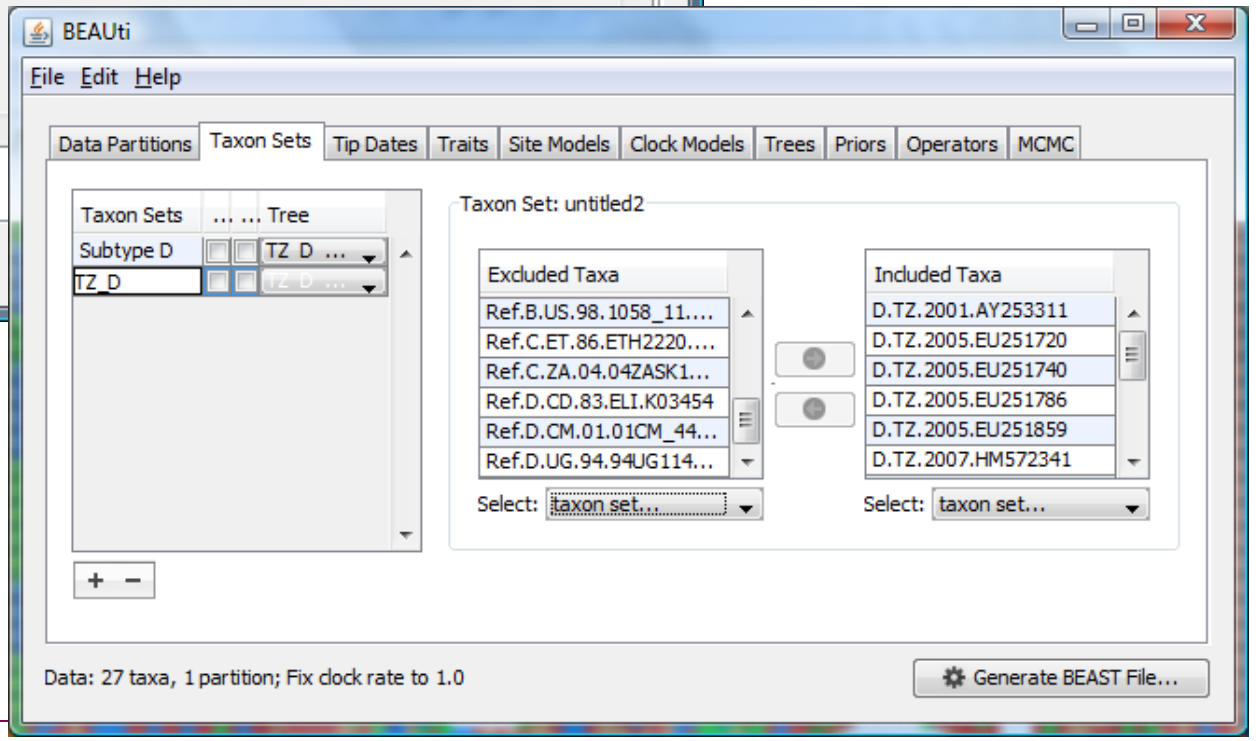


Pybus et al, 2006, Genome Biology

BEAUTi – a graphical user interface for BEAST

Import Nexus-file

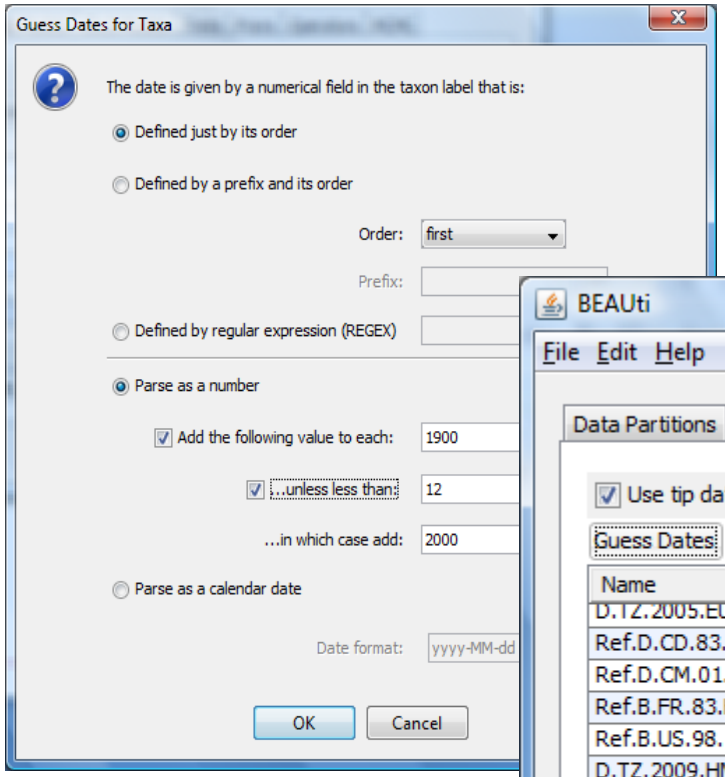




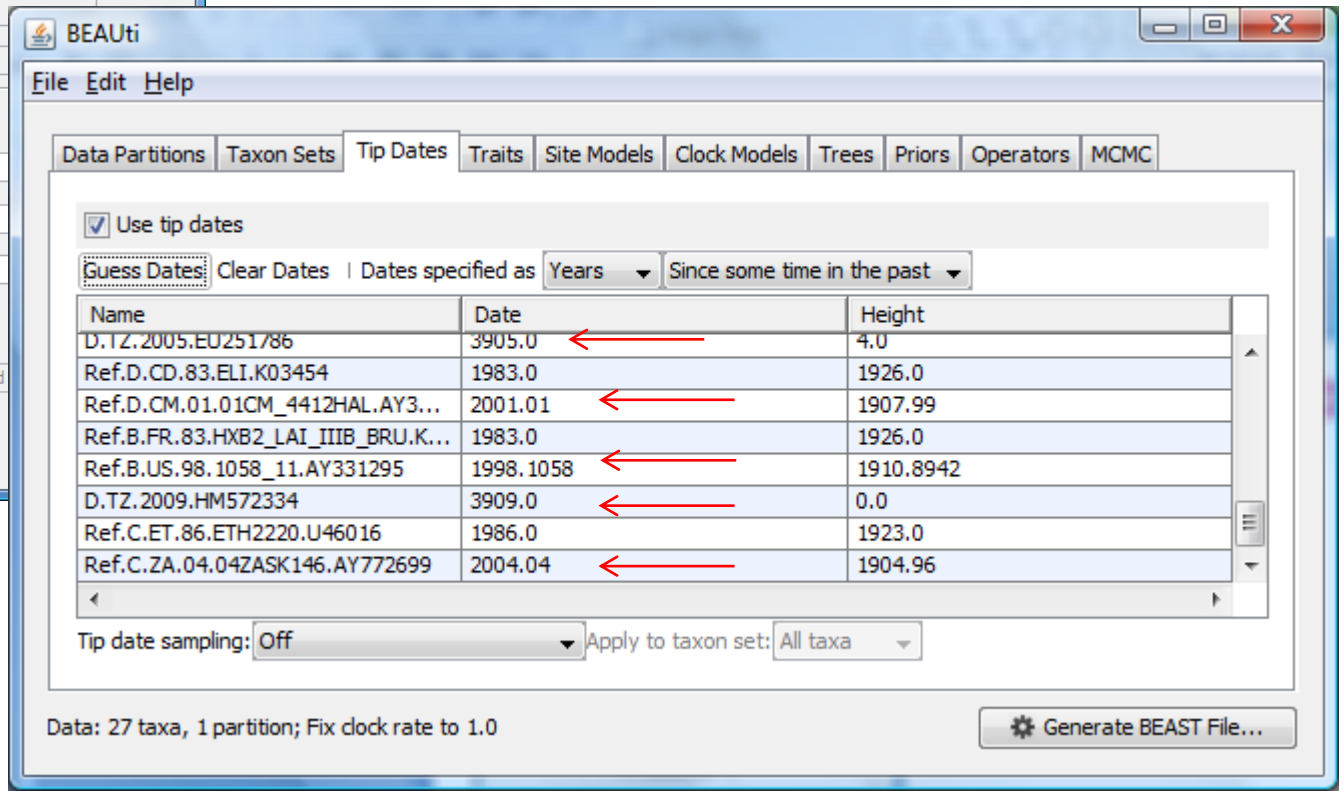
Specification of taxon sets for which we have or want a tMRCA estimate

BEAUTi

Use of tip dates for calibration

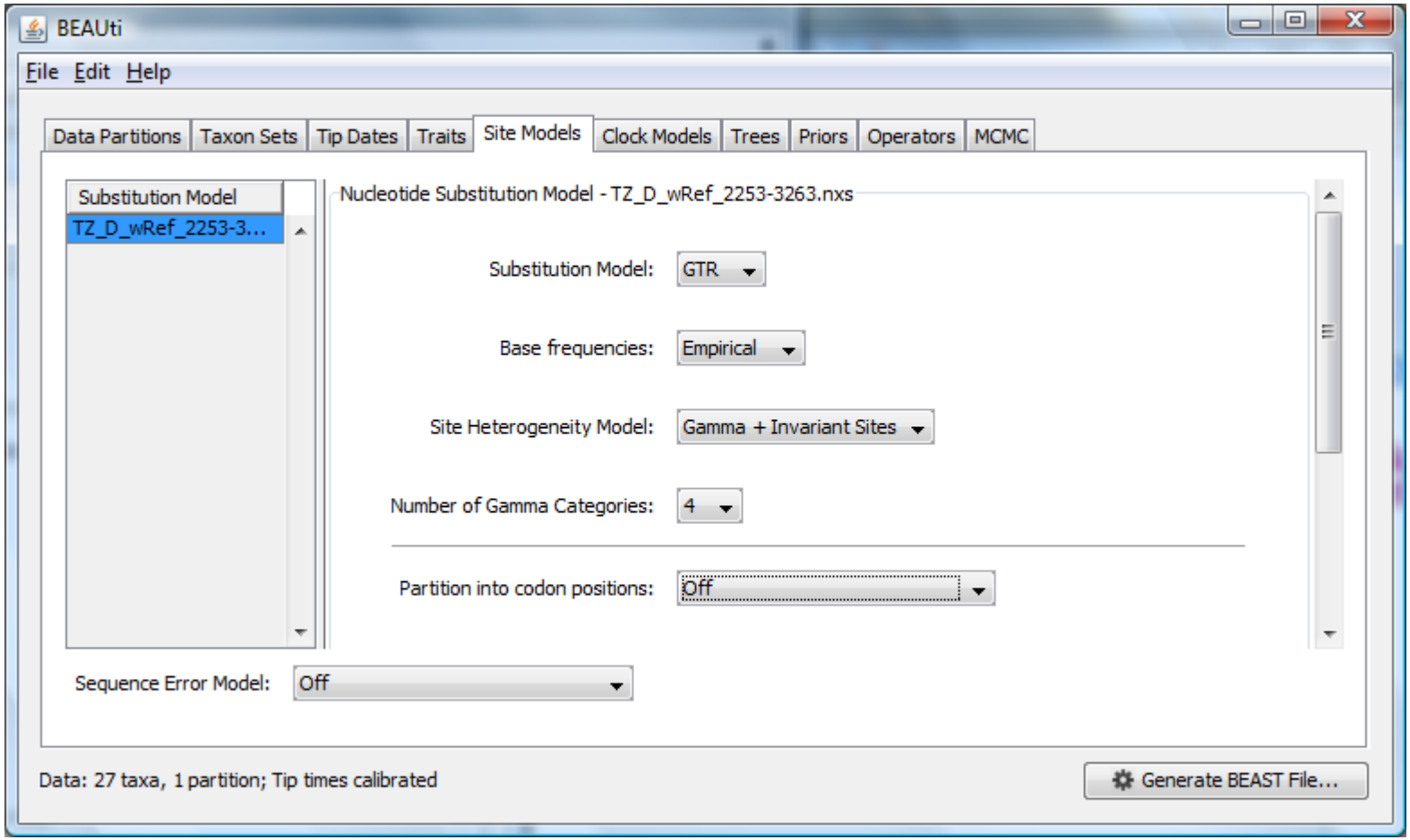


In order to avoid mistakes like those below (arrows) it is important to have a good nomenclature so that all dates are guessed correctly. Always check, and edit manually wherever necessary



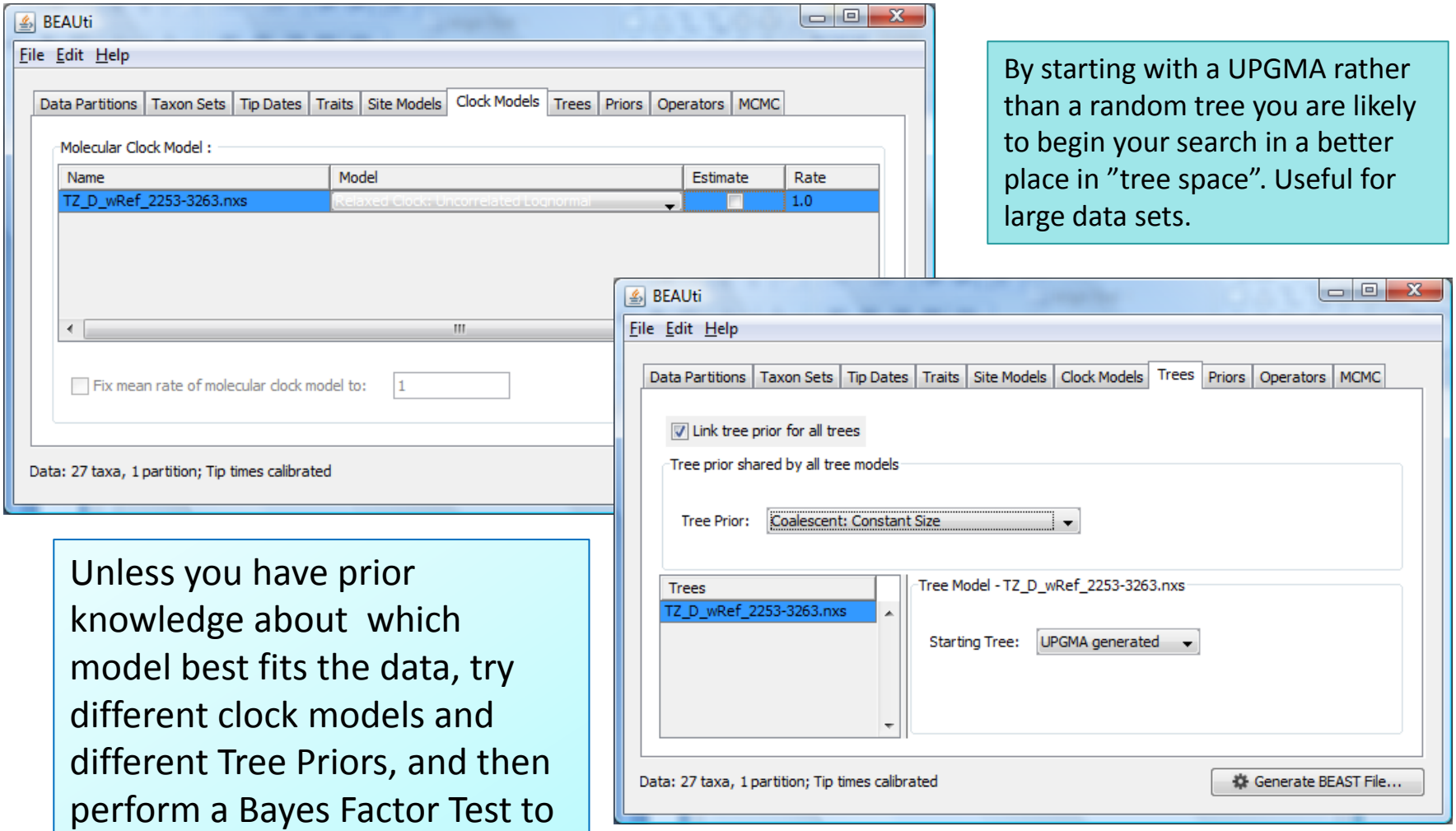
Height is calculated from the most recent date in the set, for example 2009 if that is when the latest sample was obtained

BEAUTi



We know from the jModeltest result that GTR+Γ+I fits this data best

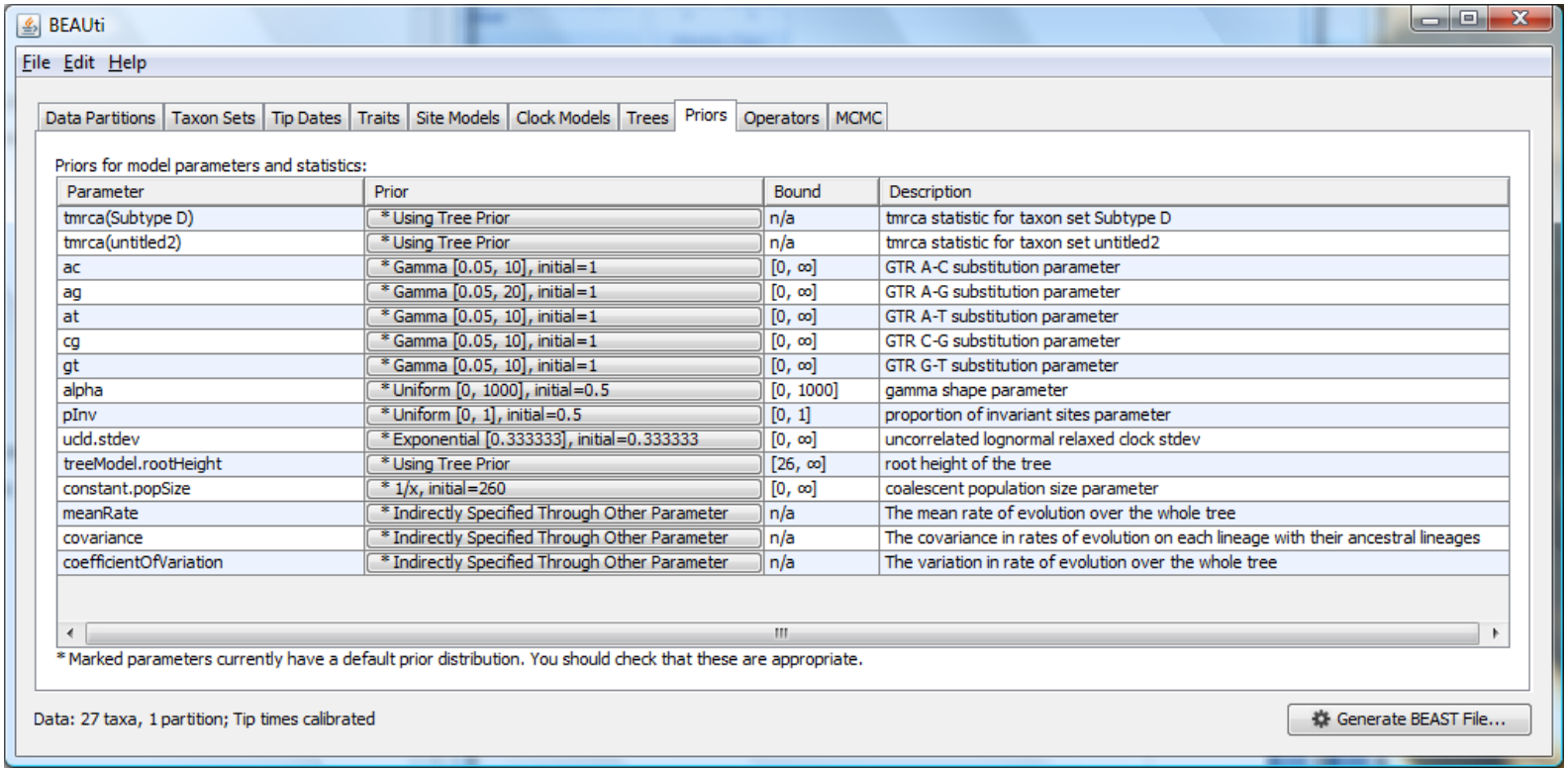
BEAUTi



By starting with a UPGMA rather than a random tree you are likely to begin your search in a better place in ”tree space”. Useful for large data sets.

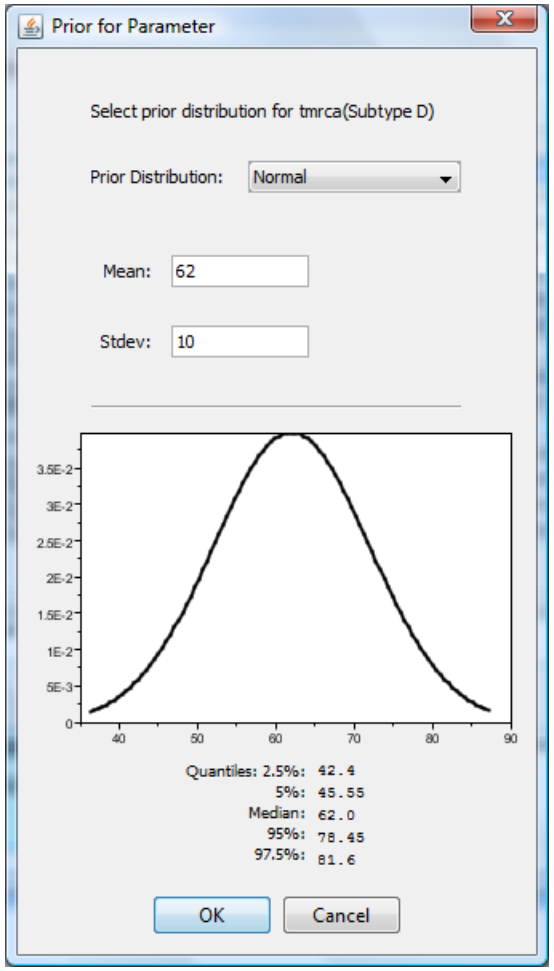
Unless you have prior knowledge about which model best fits the data, try different clock models and different Tree Priors, and then perform a Bayes Factor Test to see which is the best model

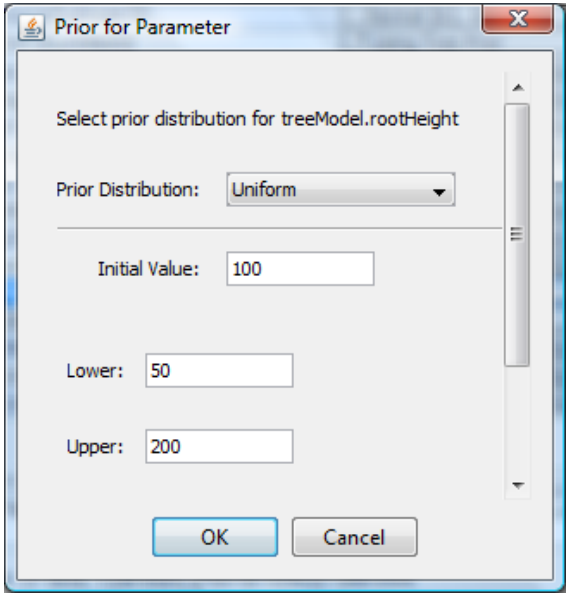
BEAUTi

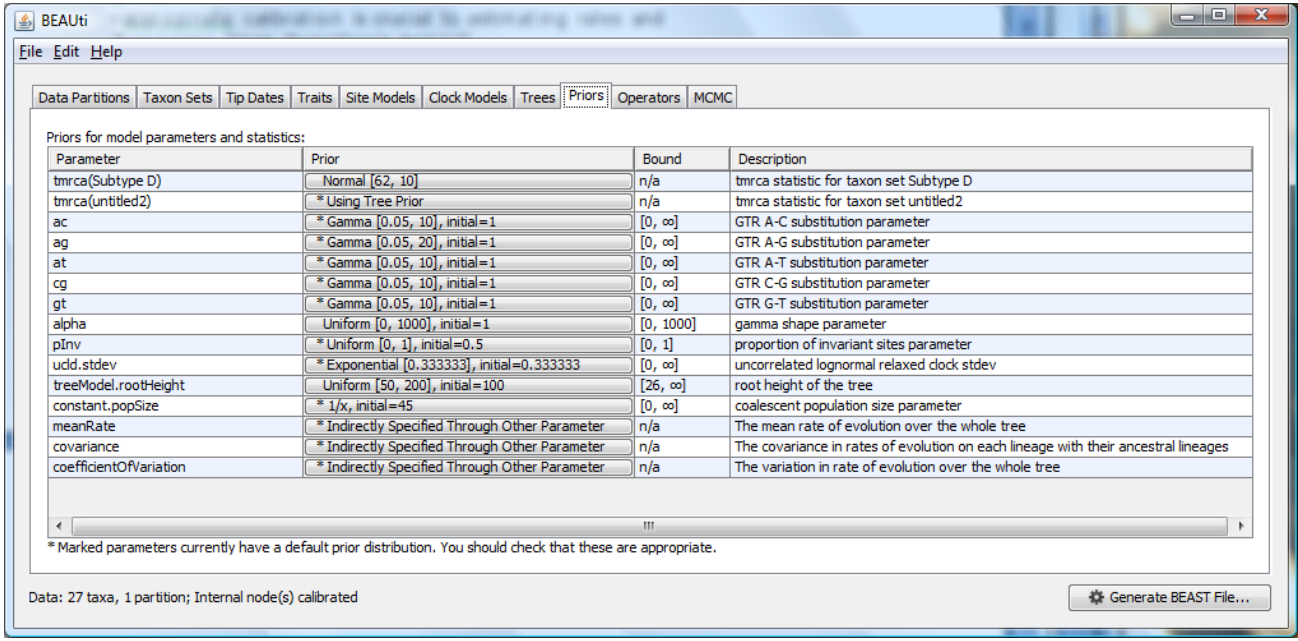


If you have prior knowledge, change the priors, otherwise leave the default values. For this dataset we have some parameter estimates from jModelTest (but be careful not to limit the search too much! Can guide the search by changing the initial value, but not the bounds). Also, in the litterature we can find a prior estimate for the tMRCA of subtype D in 1947 (Abecasis et al, 2009 J. Vir.), which is 62 years from the latest tip date, 2009.

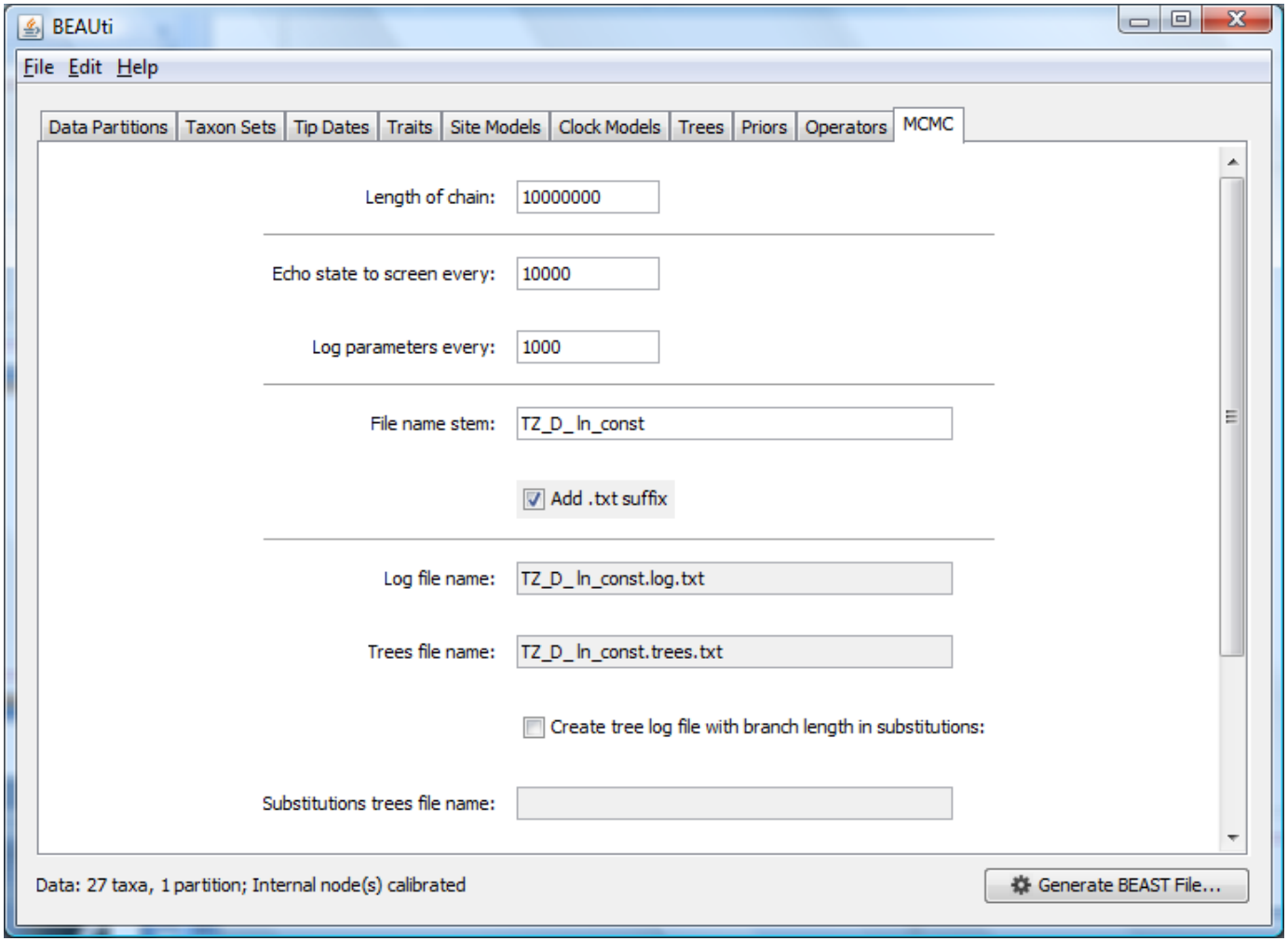
BEAUTi


Prior for Parameter
Select prior distribution for tmrca(Subtype D)
Prior Distribution: Normal
Mean: 62
Stdev: 10
3.5E-2
3E-2
2.5E-2
2E-2
1.5E-2
1E-2
5E-3
0
40
50
60
70
80
90
Quantiles: 2.5%: 42.4
5%: 45.55
Median: 62.0
95%: 78.45
97.5%: 81.6
OK
Cancel


Prior for Parameter
Select prior distribution for treeModel.rootHeight
Prior Distribution: Uniform
Initial Value: 100
Lower: 50
Upper: 200
OK
Cancel


BEAUti
File Edit Help
Data Partitions | Taxon Sets | Tip Dates | Traits | Site Models | Clock Models | Trees | Priors | Operators | MCMC
Priors for model parameters and statistics:
Parameter | Prior | Bound | Description
tmrca(Subtype D) | Normal [62, 10] | n/a | tmrca statistic for taxon set Subtype D
tmrca(untitled2) | * Using Tree Prior | n/a | tmrca statistic for taxon set untitled2
ac | * Gamma [0.05, 10], initial=1 | [0, ∞] | GTR A-C substitution parameter
ag | * Gamma [0.05, 20], initial=1 | [0, ∞] | GTR A-G substitution parameter
at | * Gamma [0.05, 10], initial=1 | [0, ∞] | GTR A-T substitution parameter
cg | * Gamma [0.05, 10], initial=1 | [0, ∞] | GTR C-G substitution parameter
gt | * Gamma [0.05, 10], initial=1 | [0, ∞] | GTR G-T substitution parameter
alpha | Uniform [0, 1000], initial=1 | [0, 1000] | gamma shape parameter
pInv | * Uniform [0, 1], initial=0.5 | [0, 1] | proportion of invariant sites parameter
ucld.stdev | * Exponential [0.333333], initial=0.333333 | [0, ∞] | uncorrelated lognormal relaxed clock stdev
treeModel.rootHeight | Uniform [50, 200], initial=100 | [26, ∞] | root height of the tree
constant.popSize | * 1/x, initial=45 | [0, ∞] | coalescent population size parameter
meanRate | * Indirectly Specified Through Other Parameter | n/a | The mean rate of evolution over the whole tree
covariance | * Indirectly Specified Through Other Parameter | n/a | The covariance in rates of evolution on each lineage with their ancestral lineages
coefficientOfVariation | * Indirectly Specified Through Other Parameter | n/a | The variation in rate of evolution over the whole tree
* Marked parameters currently have a default prior distribution. You should check that these are appropriate.
Data: 27 taxa, 1 partition; Internal node(s) calibrated
Generate BEAST File...

BEAUTi

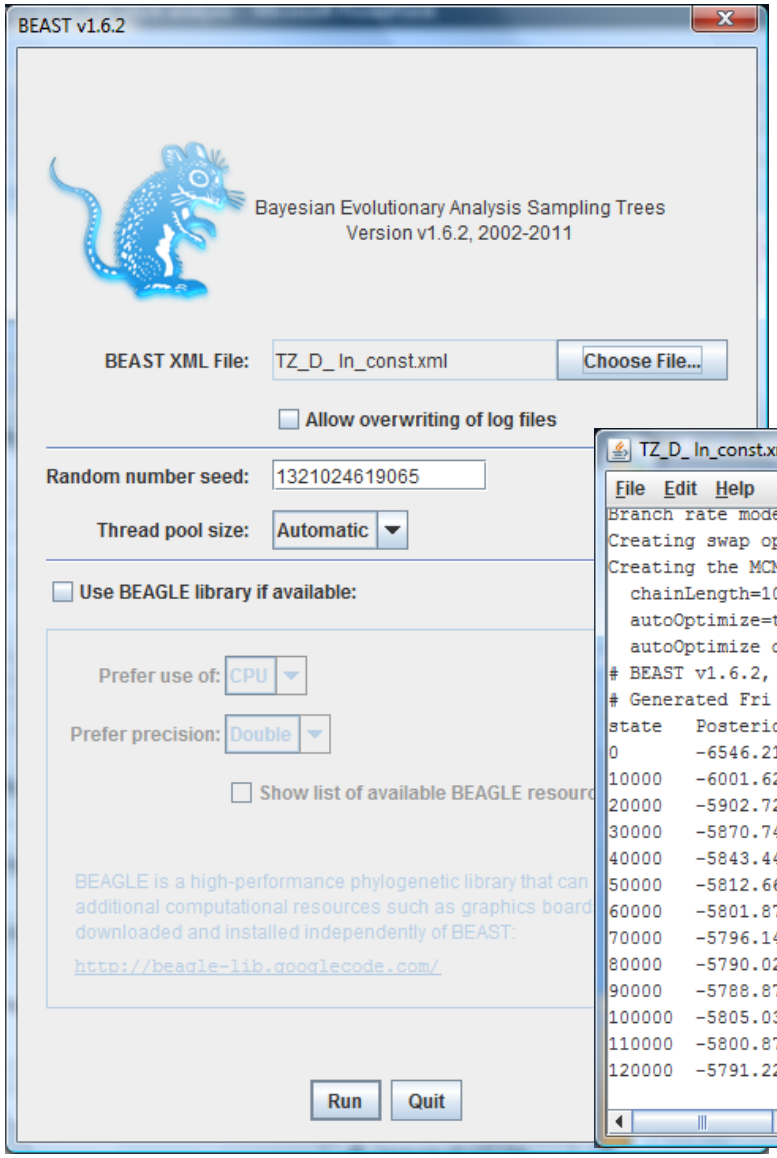


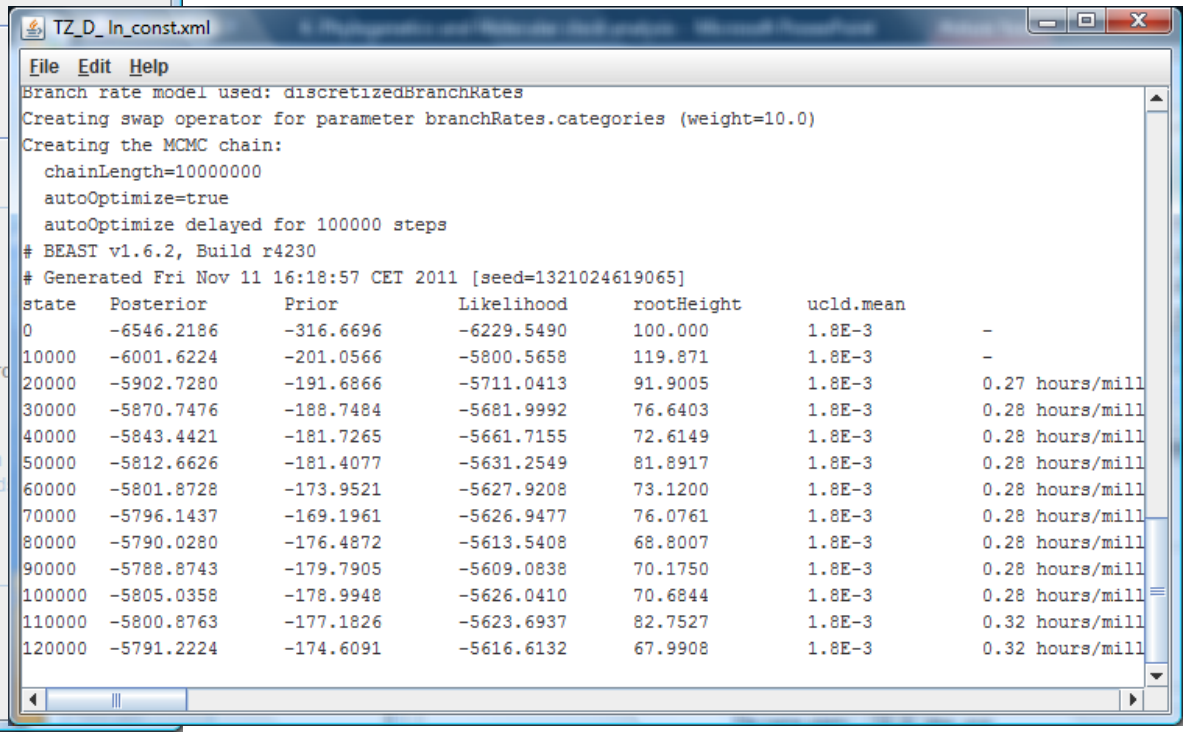
Do not close the BEAUTi file until you have seen that you can run it in BEAST. It is also easy now to go back and make minor changes in the model and save under different names for comparison.

Example: use the exponential and log-normal clocks in combination with the Tree Priors Constant growth and Bayesian Skyline to create the files
TZ_D_ln_const
TZ_D_ln_BSL
TZ_D_exp_const
TZ_D_exp_BSL

As these analyses are computationally demanding, start with ”short” runs while testing the models (around 10 million). Then, after you have performed Bayes factor analysis to decide on the best model you can run longer chains if necessary (can be >100 million), use the Tracer file to check for convergence and ESS-values to help decide what is needed for you analysis.

BEAST



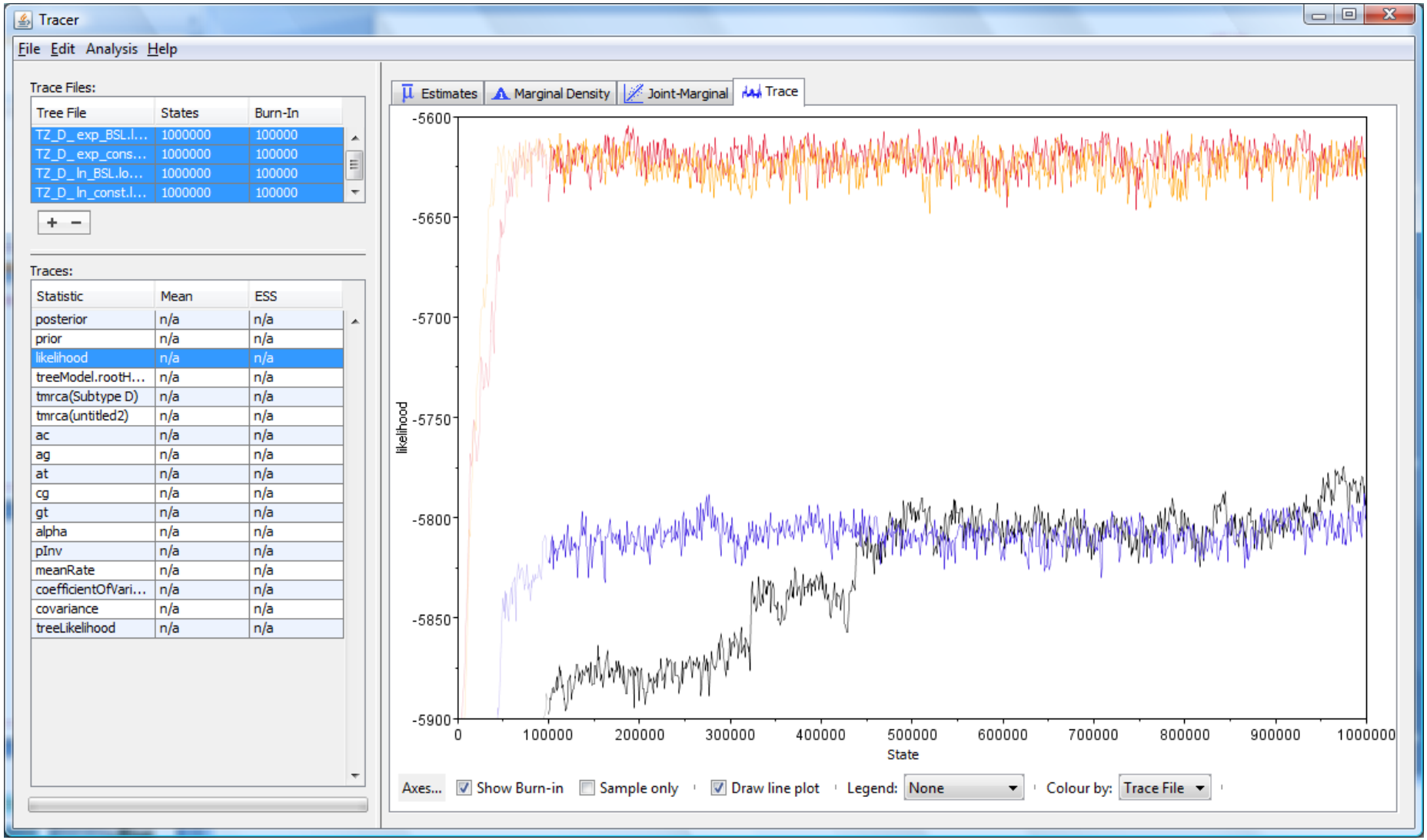


To make a BEAST run – just choose the .xml file created in BEAUTi and click Run.
BEAST generates two result files, one .log file which is analyzed in Tracer, and one .trees file which after annotation can be viewed in FigTree

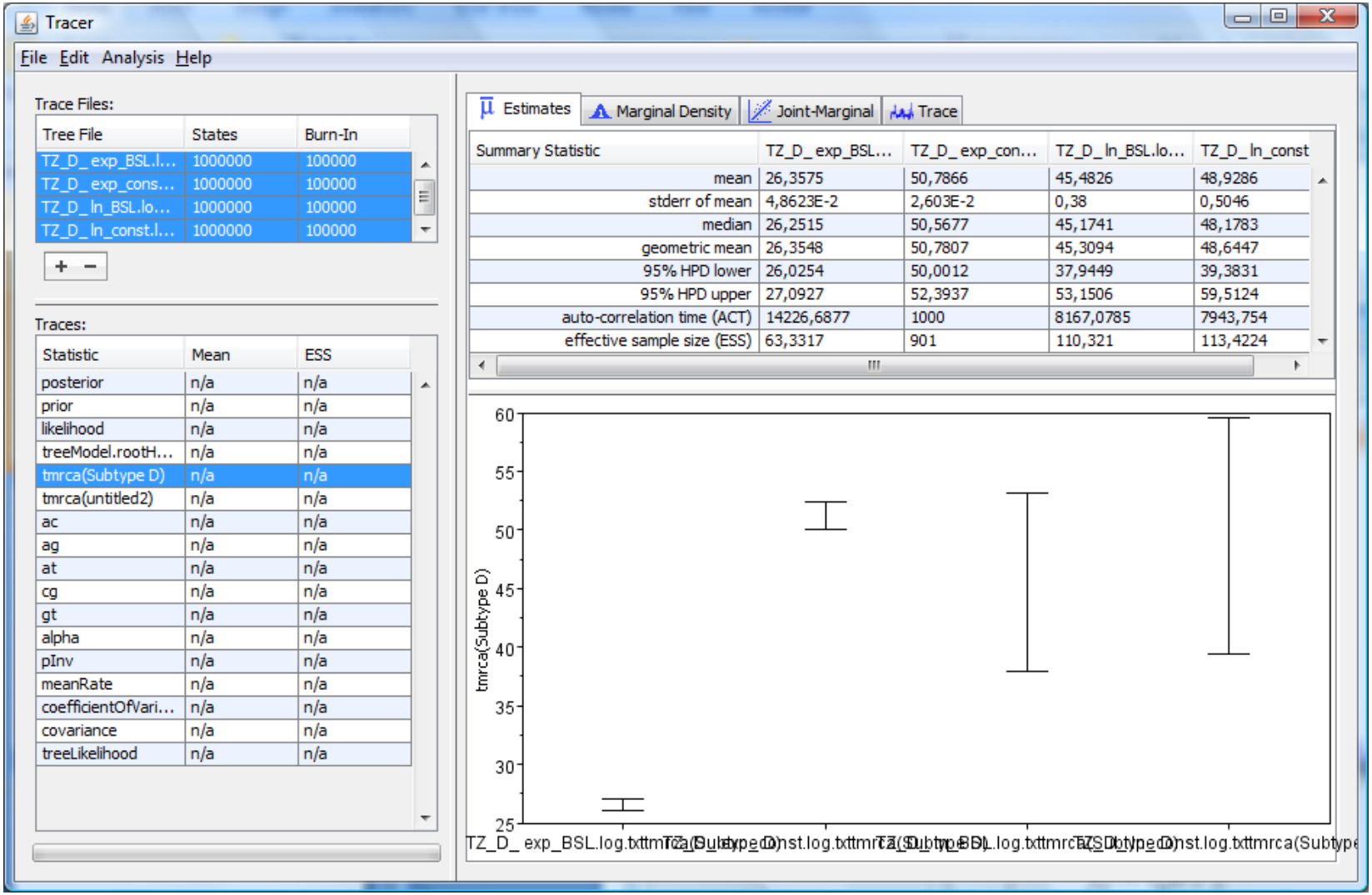
Tracer

It is possible to open and compare several log-files simultaneously in Tracer. Here it looks like the red, yellow and blue chains have ”converged” – they are fluctuating around a value. The black one looks like it may not have found the equilibrium yet.

Karolinska Institutet

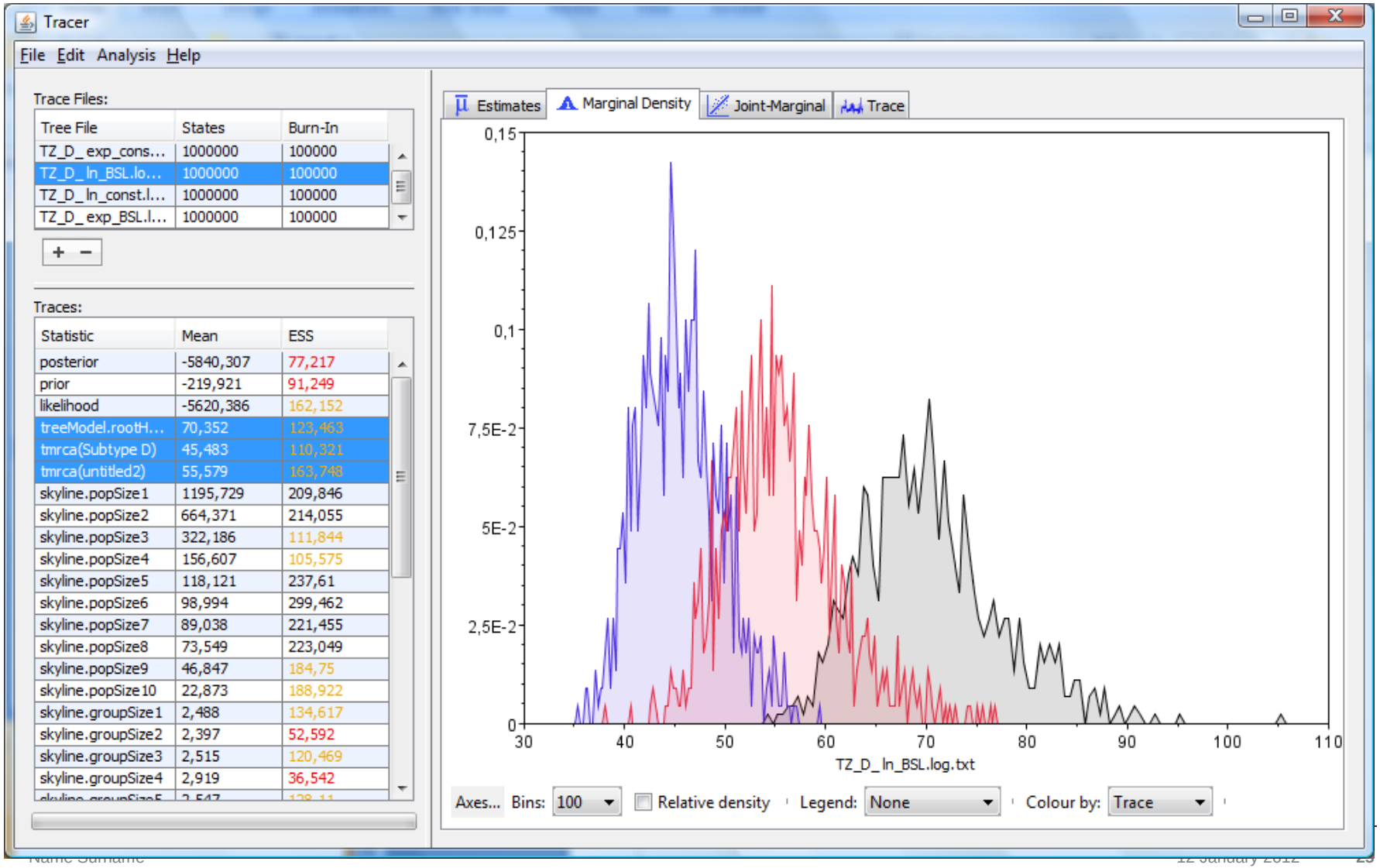


Tracer

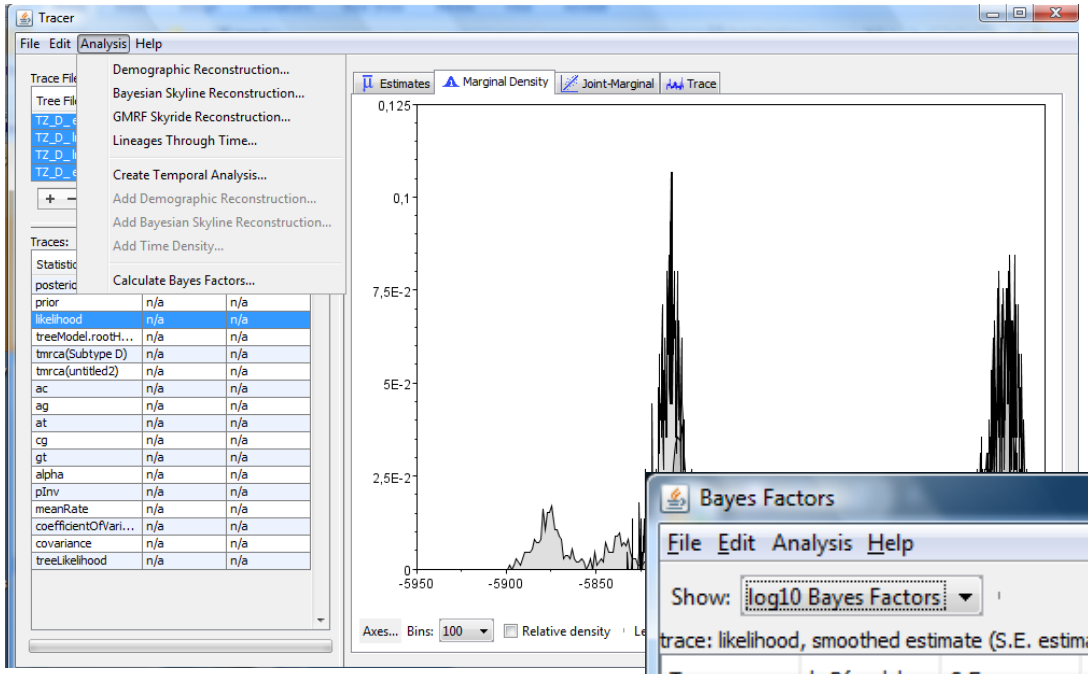

Tracer
File Edit Analysis Help
Trace Files:
Tree File | States | Burn-In
TZ_D_exp_BSL.l... | 1000000 | 100000
TZ_D_exp_cons... | 1000000 | 100000
TZ_D_ln_BSL.lo... | 1000000 | 100000
TZ_D_ln_const.l... | 1000000 | 100000
Traces:
Statistic | Mean | ESS
posterior | n/a | n/a
prior | n/a | n/a
likelihood | n/a | n/a
treeModel.rootH... | n/a | n/a
tmrca(Subtype D) | n/a | n/a
tmrca(untitled2) | n/a | n/a
ac | n/a | n/a
ag | n/a | n/a
at | n/a | n/a
cg | n/a | n/a
gt | n/a | n/a
alpha | n/a | n/a
pInv | n/a | n/a
meanRate | n/a | n/a
coefficientOfVari... | n/a | n/a
covariance | n/a | n/a
treeLikelihood | n/a | n/a
Estimates
Marginal Density
Joint-Marginal
Trace
Summary Statistic | TZ_D_exp_BSL... | TZ_D_exp_con... | TZ_D_ln_BSL.lo... | TZ_D_ln_const
mean | 26,3575 | 50,7866 | 45,4826 | 48,9286
stderr of mean | 4,8623E-2 | 2,603E-2 | 0,38 | 0,5046
median | 26,2515 | 50,5677 | 45,1741 | 48,1783
geometric mean | 26,3548 | 50,7807 | 45,3094 | 48,6447
95% HPD lower | 26,0254 | 50,0012 | 37,9449 | 39,3831
95% HPD upper | 27,0927 | 52,3937 | 53,1506 | 59,5124
auto-correlation time (ACT) | 14226,6877 | 1000 | 8167,0785 | 7943,754
effective sample size (ESS) | 63,3317 | 901 | 110,321 | 113,4224
60
55
50
45
40
35
30
25
tmrca(Subtype D)
TZ_D_ exp_BSL.log.txttmrca(Subtype D)

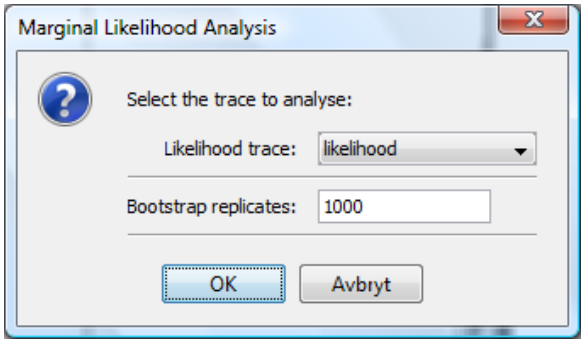
Tracer

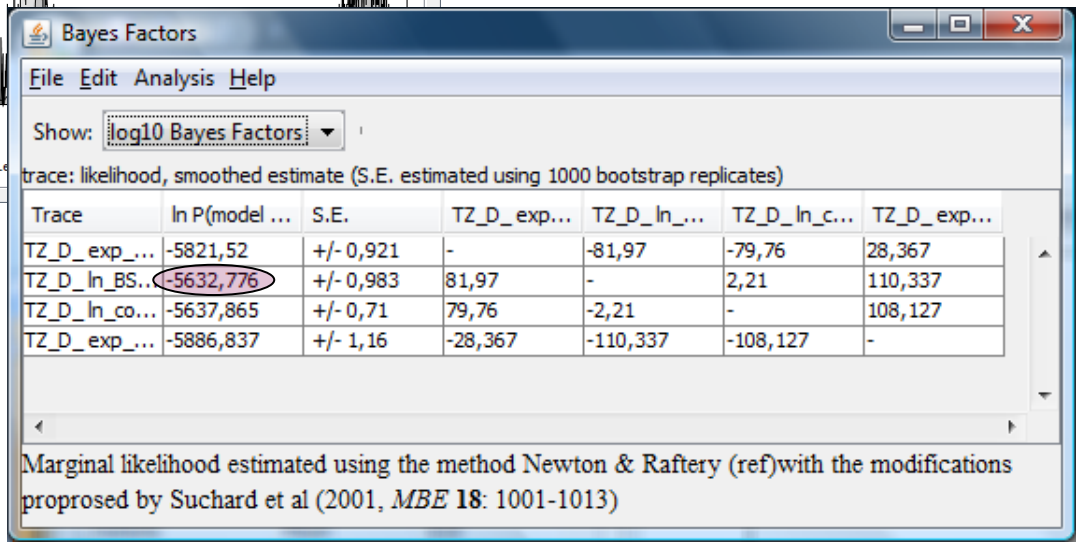
This analysis will probably improve if it would be run for a longer time – the ESS values are still low and the histograms are approaching normal distributions, but are not there yet



Bayes factor analysis





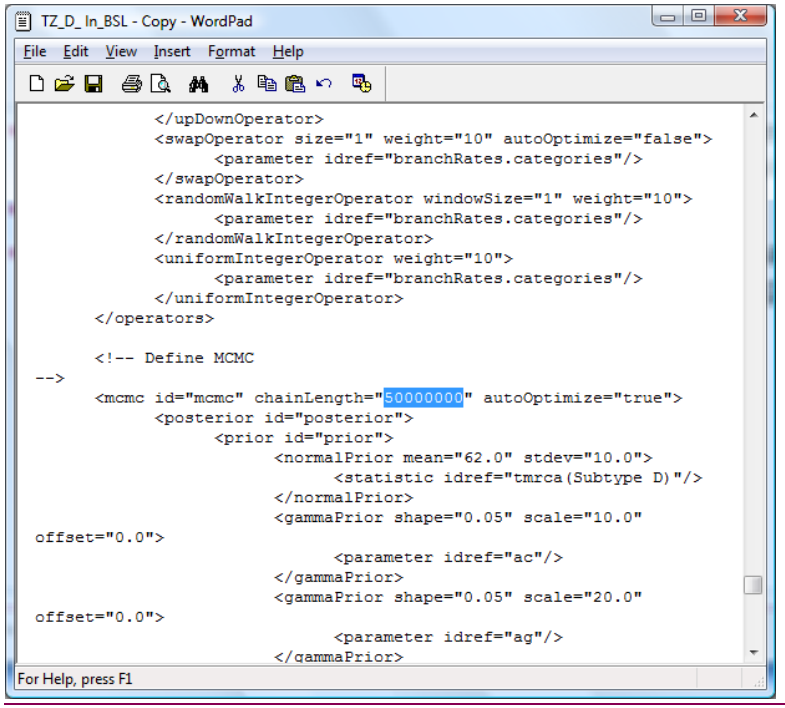


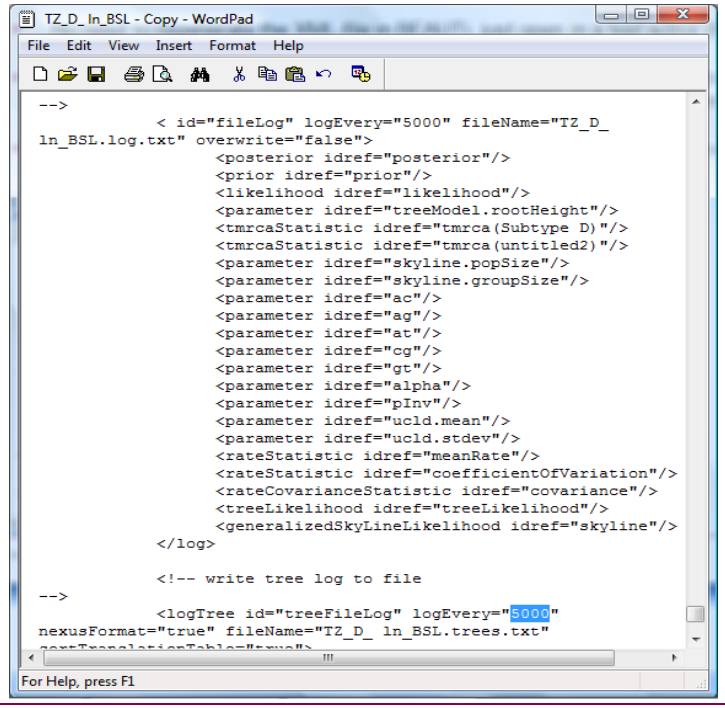
Bayes factor, rough interpretation guide:
0-3: not significant
3-10: uncertain significance
>10: important difference

TZ-D_ln_BSL is the model with the highest likelihood. The differences to both models with the exponential clock are high (81.97 and 110.337), but there is only a minor difference compared to TZ_D_ln_const (2.21) -> can continue with longer runs for both models

Perform longer runs using the selected model(s)

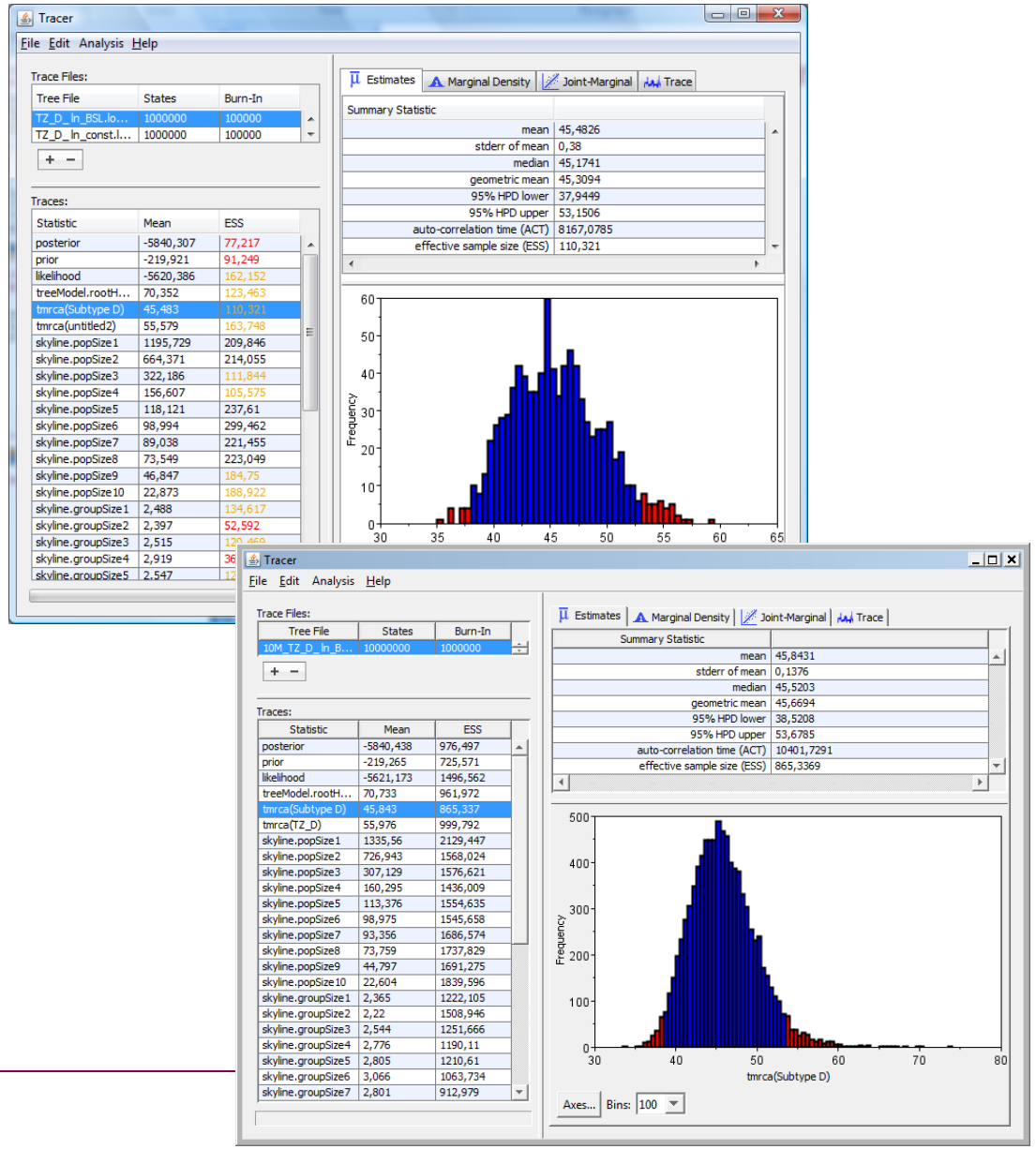
- No need to regenerate the XML-file in BEAUTi, just open in a text editor (for example WordPad), and change the chain length.
- It may also be useful to change the log frequencies (screenLog, fileLog, treeFileLog). Rule of thumb – obtain 10000 trees. Ex run of 50 million generations: fileLog and treeFileLog = logEvery 5000
- Save the file in a new folder since the output files will have the same name as from the previous run (unless you change this as well!)



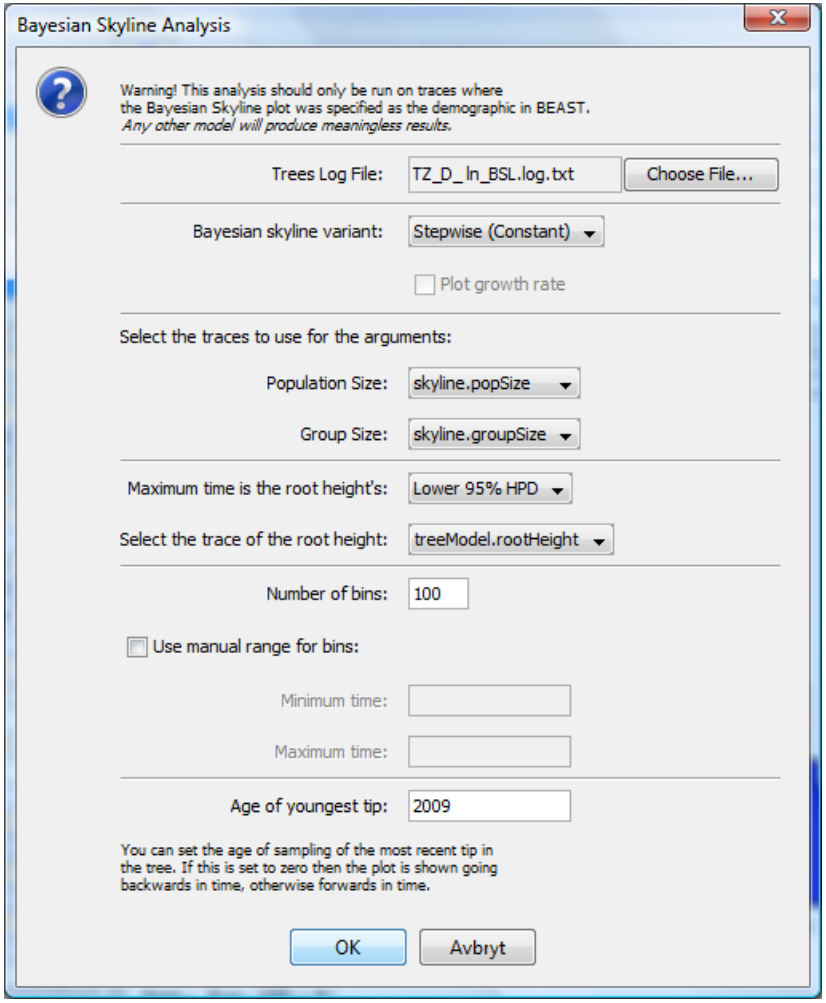


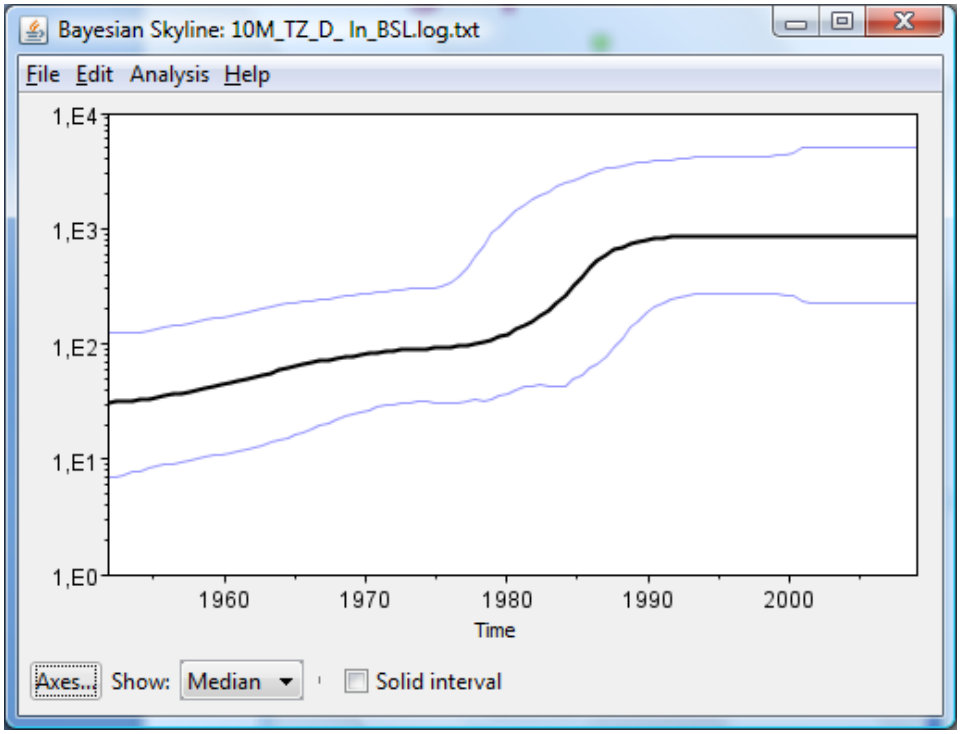
How do I know how long I need to run my analysis?

- Look at the Tracer log-file for
  - Chain convergence
  - ESS values
  - Shape of the distribution curve



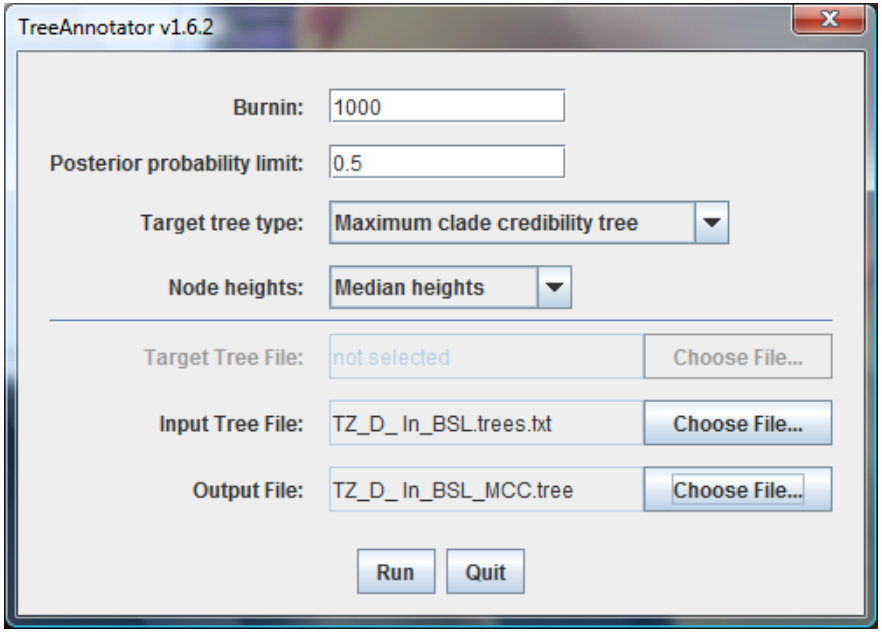
Demographic reconstruction, Tracer





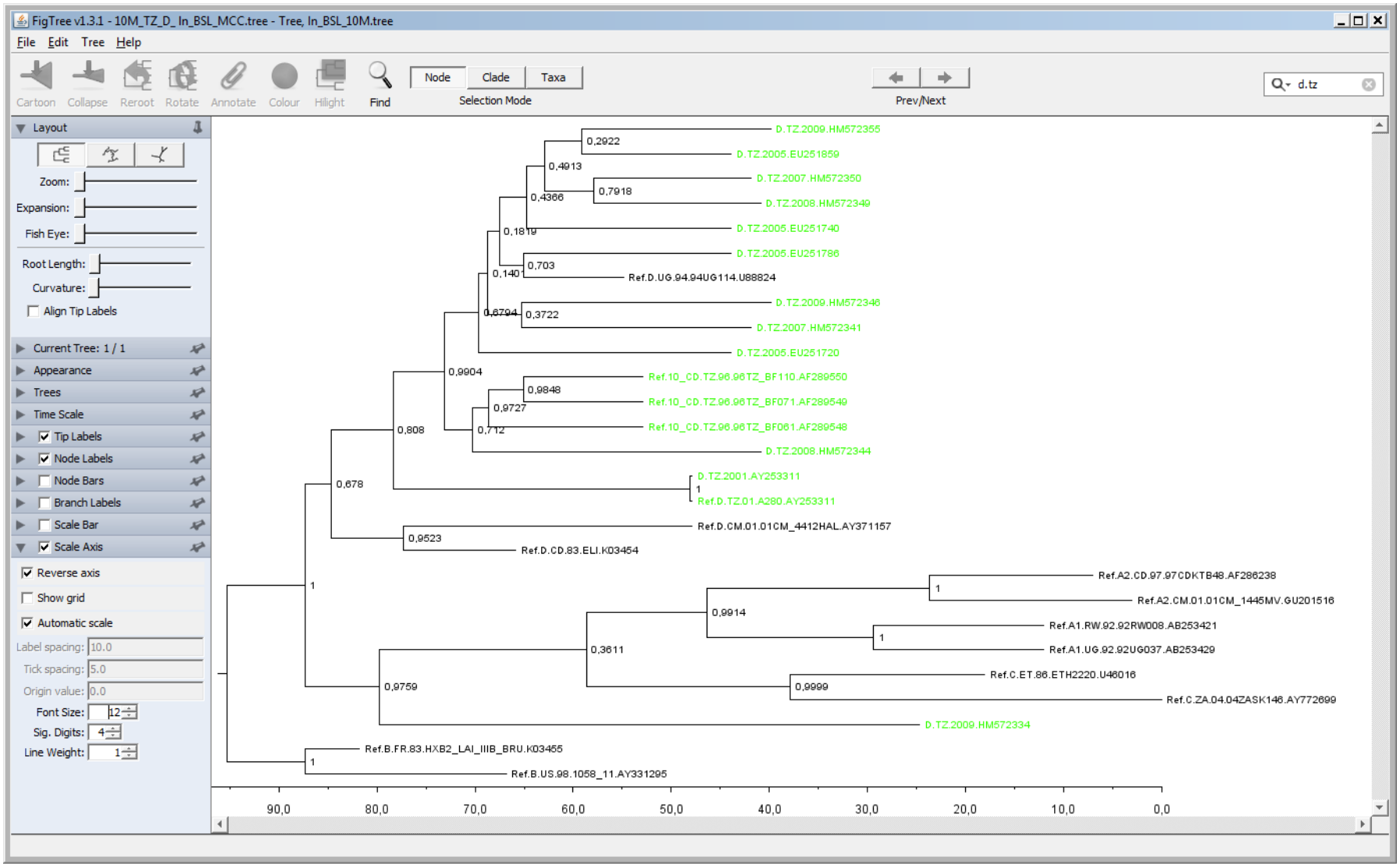
TreeAnnotator

- Distributed with the BEAST package
- Creates a consensus tree with posterior probabilities from the ≈10000 sampled trees
- Burnin approx 10% (1000 if you have sampled 10000 trees), but look at Tracer file and amend if necessary
- Save as a .tree file



FigTree

Good program for viewing and editing trees. Lots of functions, take the time to explore different options. Under Node Labels you can see the posterior for each node – gives the credibility for that clade and you can also see the age of the node – the tMRCA estimate for that branching event.



to learn more...

http://beast.bio.ed.ac.uk/Tutorials

http://regaweb.med.kuleuven.be/workshop/