



**Karolinska
Institutet**

Phylogenetic tree construction and molecular clock analysis

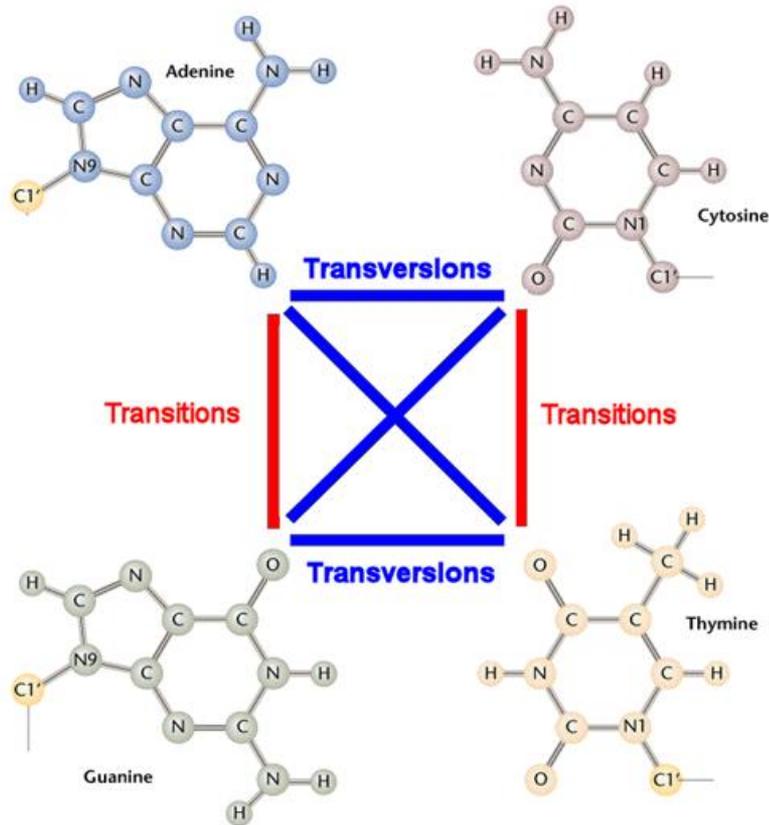
EDCTP ENNEA training on data management

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)

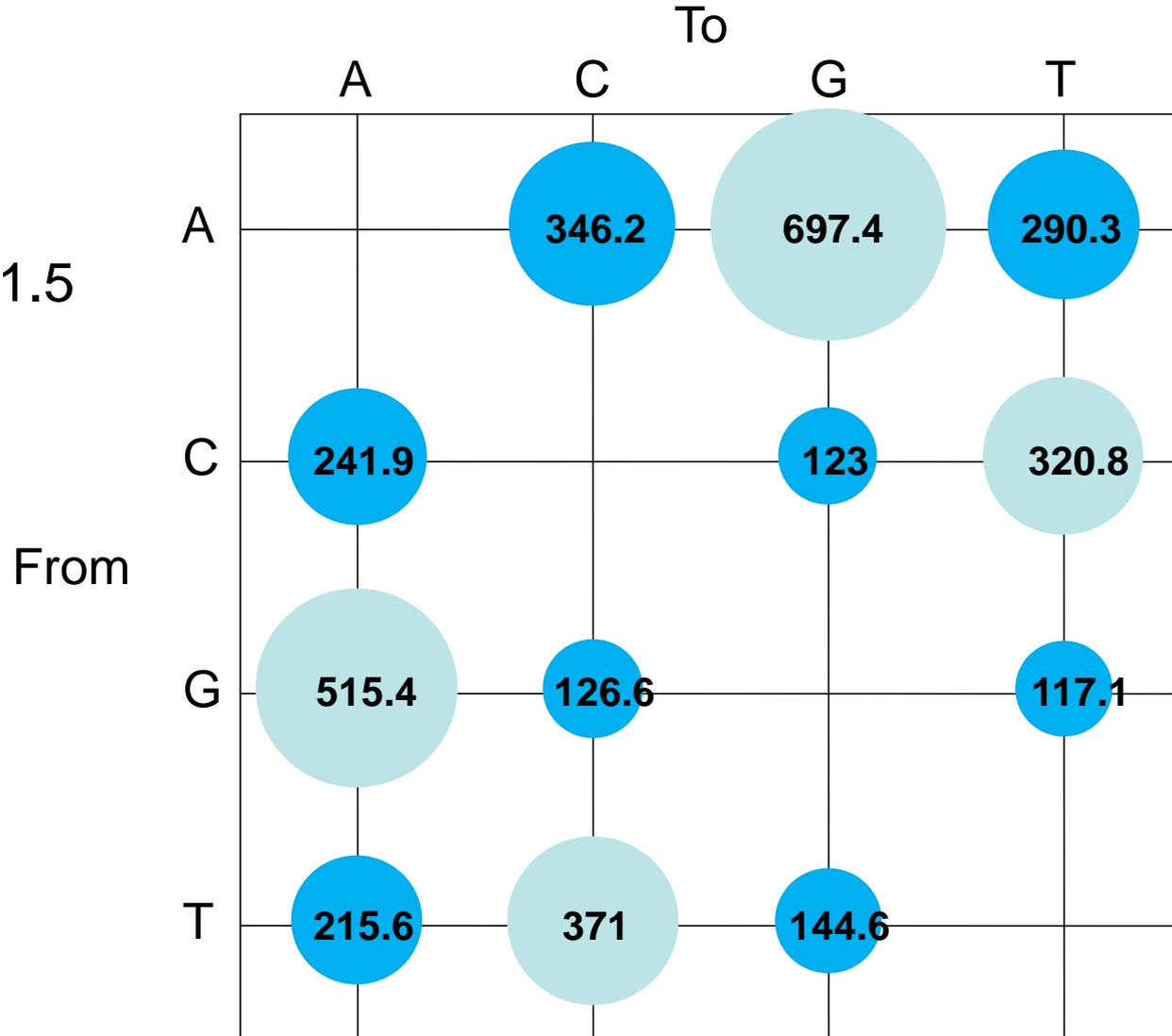
		Second Letter				
		T	C	A	G	
First Letter	T	TTT } Phe TTC } TTA } Leu TTG }	TCT } Ser TCC } TCA } TCG }	TAT } Tyr TAC } TAA } Stop TAG } Stop	TGT } Cys TGC } TGA } Stop TGG } Trp	T C A G
	C	CTT } Leu CTC } CTA } CTG }	CCT } Pro CCC } CCA } CCG }	CAT } His CAC } CAA } Gln CAG }	CGT } Arg CGC } CGA } CGG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

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

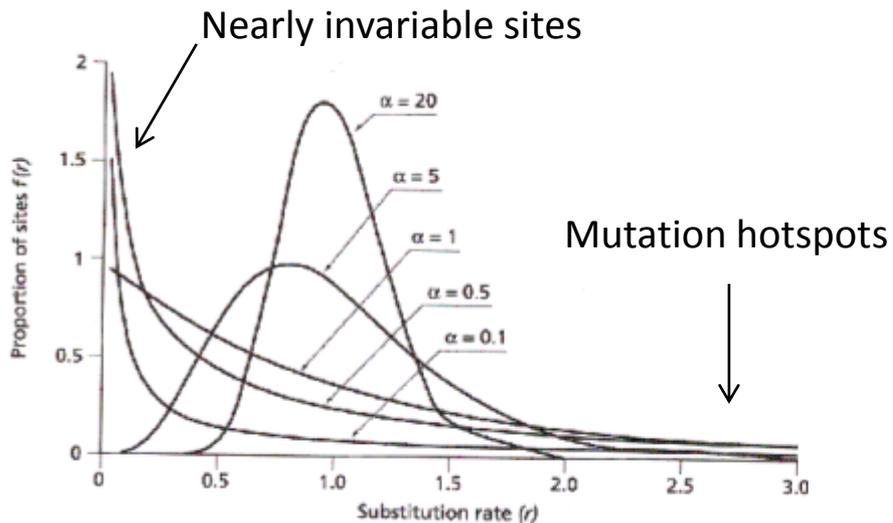
	T	C	A	G
T	f_N	a	b	b
C	a	f_N	b	b
A	b	b	f_N	a
G	b	b	a	f_N

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



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. The most complex model is therefore GTR+ Γ +I

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

Modeltest

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



jModelTest 0.1.1

File Edit Analysis Results Tools Help About

----- jModeltest 0.1.1
 (c) 2008 David Posada, Department of Biochemistry, Genetics and Immunology
 University of Vigo, 36310 Vigo, Spain. e-mail: dposada@uvigo.es

 Fri Nov 11 06:19:37 CET 2011 (Windows Vista 6.0, arch: x86)

***** NOTICE *****
 This program may contain errors. Please inspect the results carefully.

Reading data file "TZ_D_wRef_2253-3263.fasta"... OK.
 number of sequences: 27
 number of sites: 1011

Likelihood settings

Likelihood settings

Number of substitution schemes
 3 5 7 11 Number of models = 88

Base frequencies
 +F

Rate variation
 +I +G nCat: 4

Base tree for likelihood calculations
 Fixed BIONJ-JC Fixed user topology
 BIONJ ML optimized

Default Settings Cancel **Compute Likelihoods**

Likelihood scores not available TZ_D_wRef_2253-3263.fasta

Results

Models: AIC AICc BIC DT

ID	Name	Partition	-lnL	p	fA	fC	fG
88	GTR+I+G	012345	5 605,272	62	0.3912	0.1578	0.2043
72	TIM3+I+G	012032	5 611,779	60	0.3907	0.1586	0.2043
56	TIM1+I+G	012230	5 613,651	60	0.3932	0.1726	0.2043
80	TVM+I+G	012314	5 613,773	61	0.3794	0.1744	0.1862
48	TPM3uf+I...	012012	5 619,71	59	0.3789	0.175	0.1862
64	TIM2+I+G	010232	5 619,862	60	0.3876	0.184	0.2043
32	TPM1uf+I...	012210	5 620,383	59	0.383	0.186	0.1862
87	GTR+G	012345	5 624,254	61	0.3888	0.163	0.2043
71	TIM3+G	012032	5 629,285	59	0.3889	0.1634	0.2043
79	TVM+G	012314	5 633,502	60	0.3809	0.1744	0.1862
40	TPM2uf+I...	010212	5 633,734	59	0.3785	0.1948	0.1862
55	TIM1+G	012230	5 636,743	59	0.3924	0.1759	0.1862
31	TPM1uf+G	012210	5 636,874	58	0.3832	0.1862	0.1862
47	TPM3uf+G	012012	5 638,576	58	0.3802	0.1752	0.1862
39	TPM2uf+G	010212	5 638,792	58	0.3766	0.1961	0.1862
86	GTR+I	012345	5 650,849	61	0.3916	0.1613	0.2043
63	TIM2+G	010232	5 651,828	59	0.3883	0.1836	0.2043

Decimal numbers are rounded. Click on column headers to sort data in ascending or descending order (+Shift)
 11 november 2...

Maximum likelihood 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

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

Options

Use all 28 models

Construct initial tree using Neighbor PAUP* MrBayes

Always e-mail results

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.who.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:

Complete genomes	2832
Pol 2253-3263	59508
V3	133578

No of possible trees

6.4e7682
too big for the calculator...
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 time span for more accurate dating using the molecular clock.

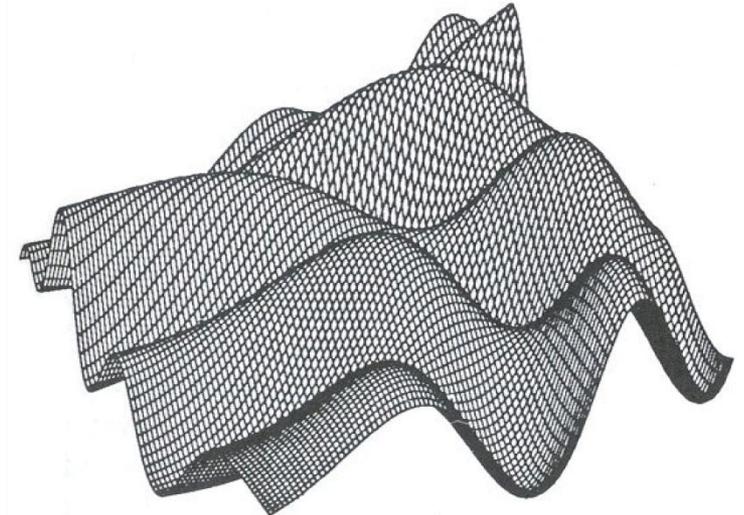
How do we find the best tree? (or at least a good one...)

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

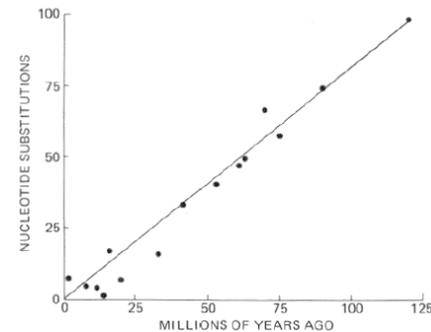
Heuristic – searches tree space by making small random changes to a starting tree, and accepting the new tree according to defined criteria. This method does not check all trees and cannot guarantee to find the best one.

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



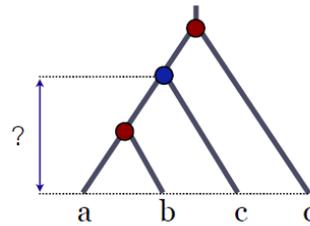
Molecular clocks

- 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)
- 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.



from AC Wilson, 1976

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.

$k = N \mu p =$ substitution rate (per generation)

$\mu =$ mutation rate (per individual per generation)

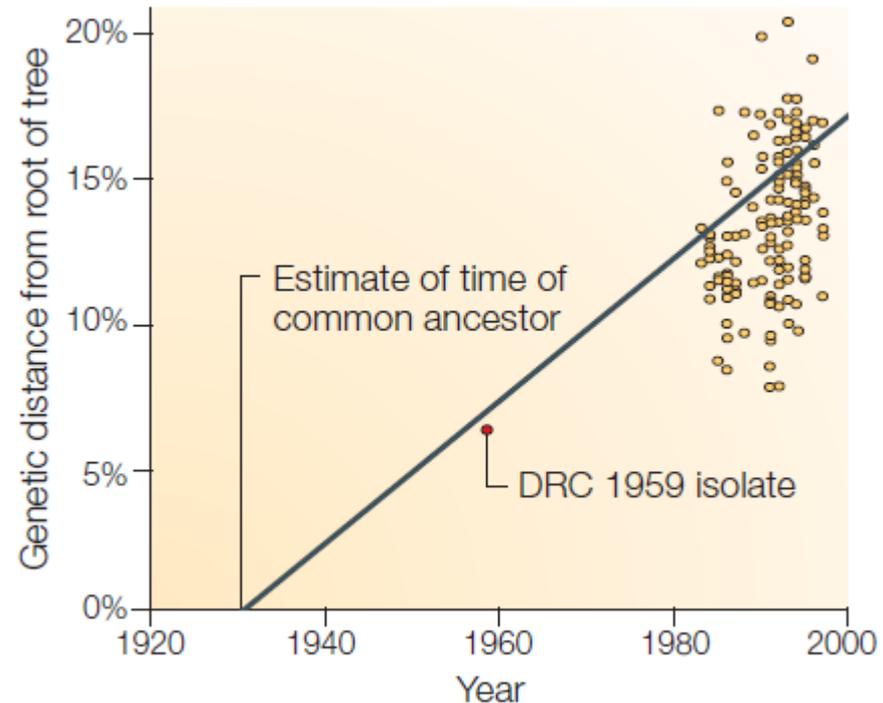
$p =$ probability of fixation

$N =$ population size ($2N$ for diploids)

Calibrating molecular clocks

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

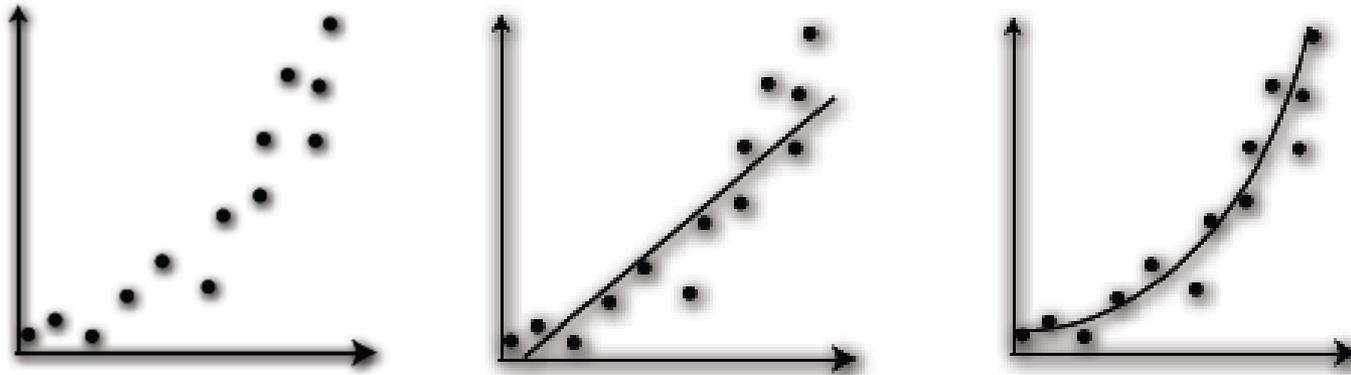
Therefore it is possible to use **tip dates** for calibration



Korber et al, 2000.Science

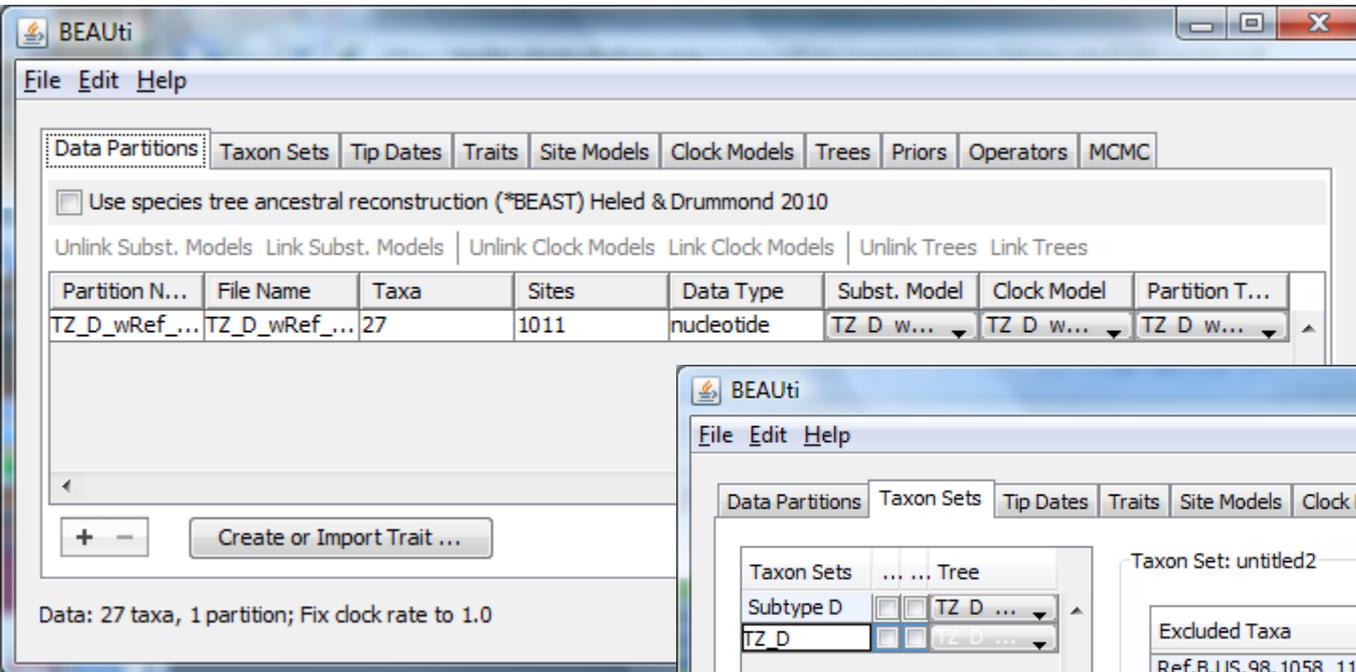
Relaxing the molecular clock

- 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)



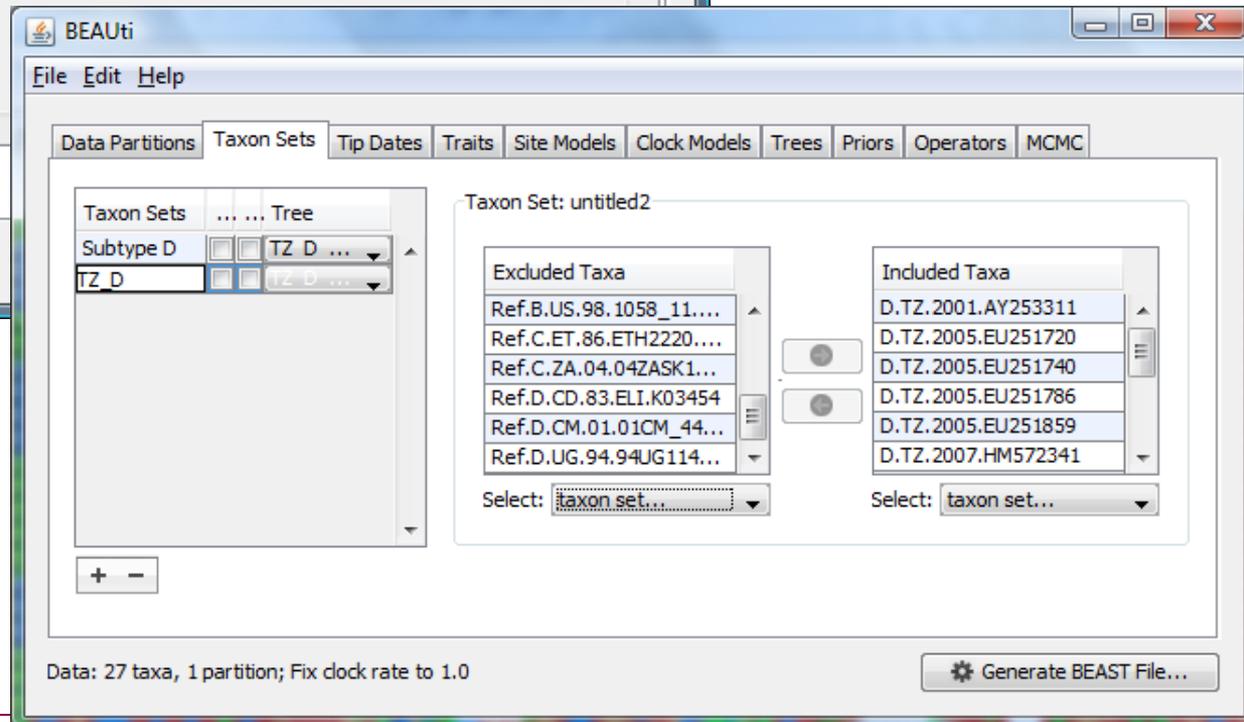
Pybus et al, 2006, Genome Biology

BEAUTi – a graphical user interface for BEAST

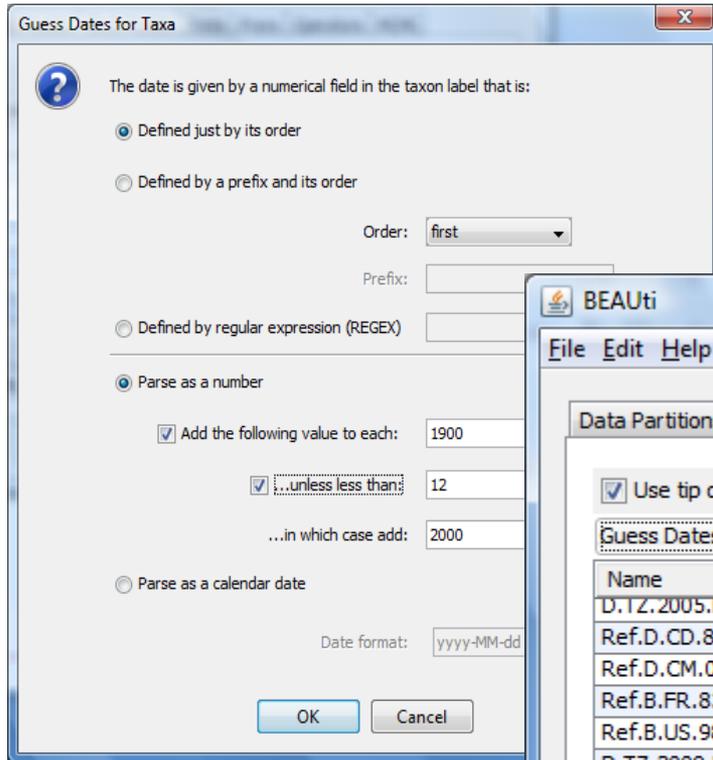


Import Nexus-file

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



Use of tip dates for calibration



Guess Dates for 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
- Defined by regular expression (REGEX)
- Parse as a number
- Parse as a calendar date

Order: first

Prefix:

Add the following value to each: 1900

...unless less than: 12

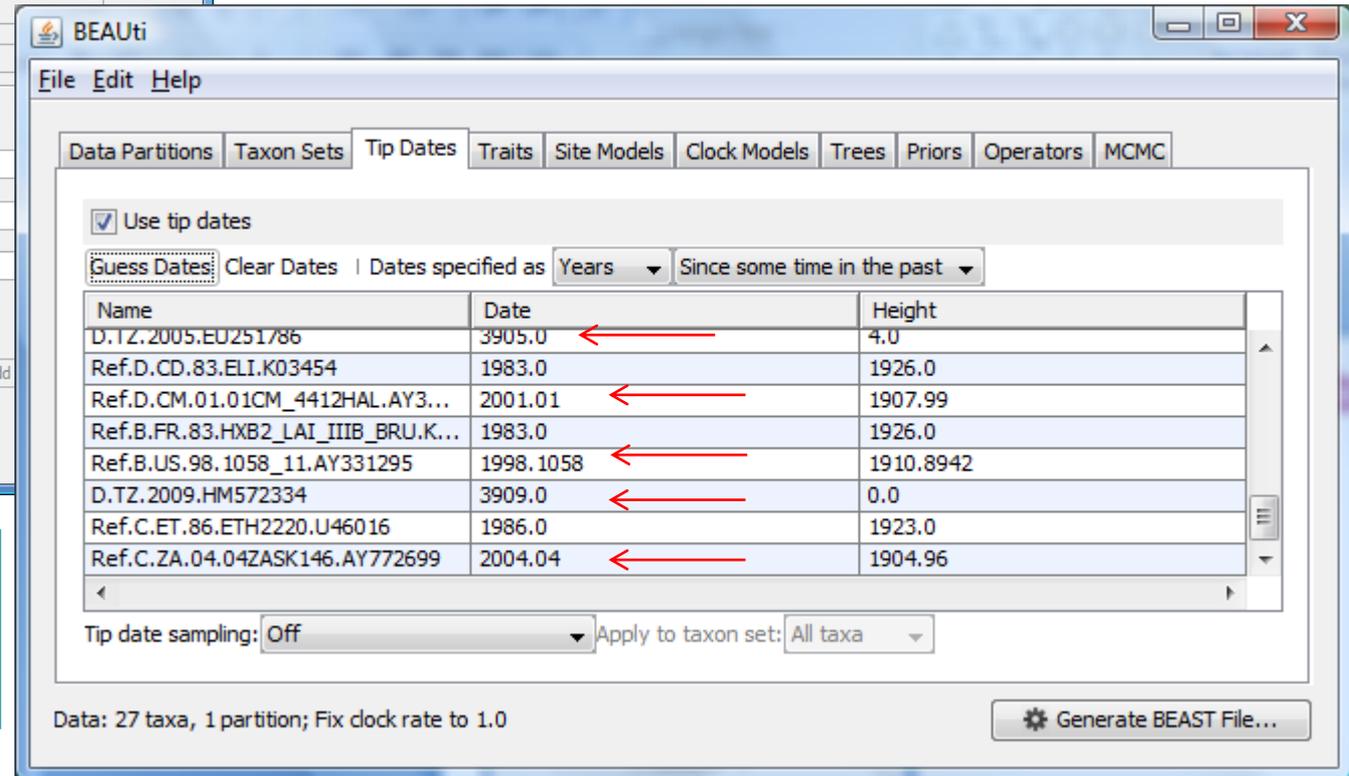
...in which case add: 2000

Date format: yyyy-MM-dd

OK Cancel

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

File Edit Help

Data Partitions Taxon Sets Tip Dates Traits Site Models Clock Models Trees Priors Operators MCMC

Use tip dates

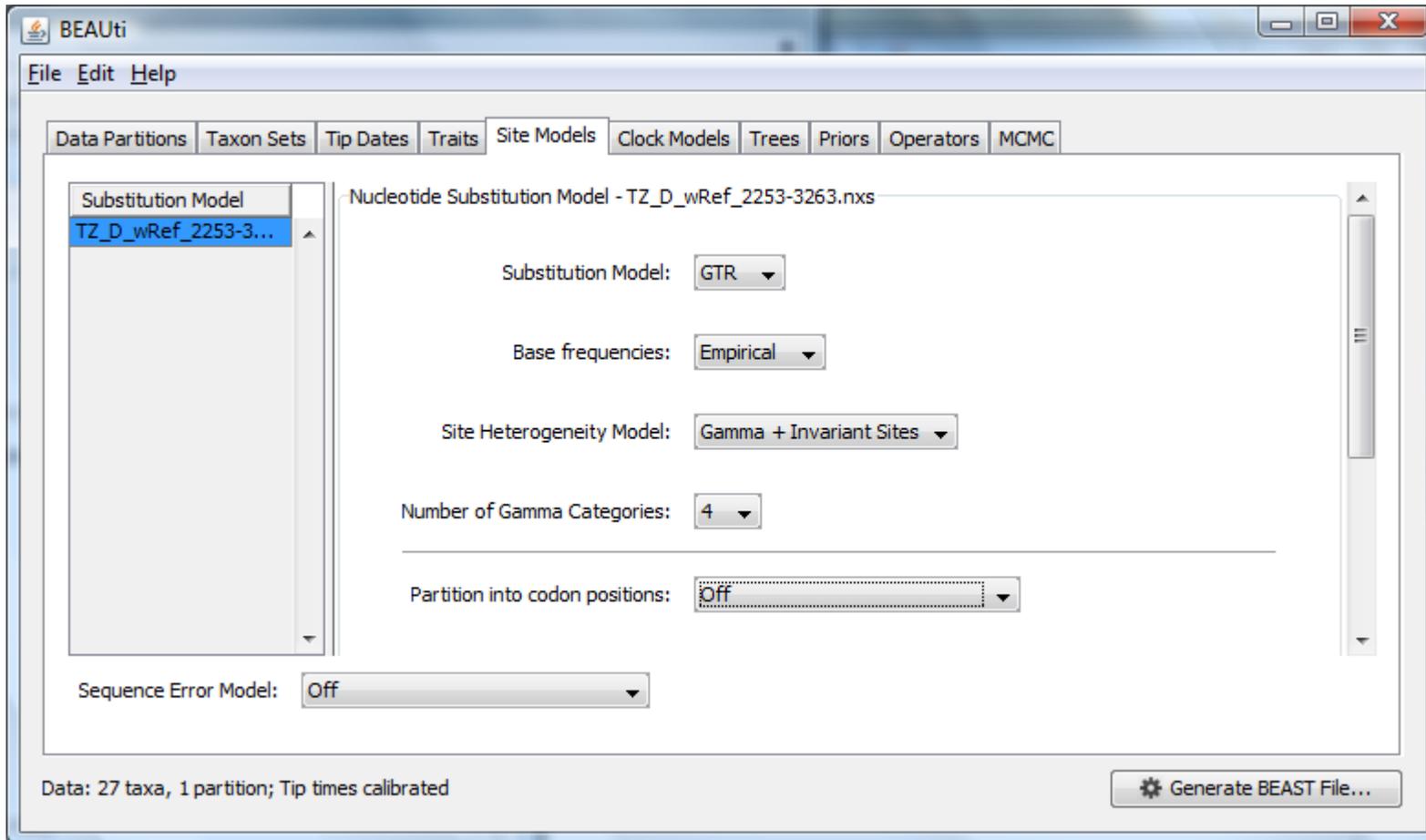
Guess Dates: Clear Dates | Dates specified as Years Since some time in the past

Name	Date	Height
D.T.Z.2005.EU251786	3905.0	4.0
Ref.D.CD.83.ELI.K03454	1983.0	1926.0
Ref.D.CM.01.01CM_4412HAL.AY3...	2001.01	1907.99
Ref.B.FR.83.HXB2_LAI_IIIB_BRU.K...	1983.0	1926.0
Ref.B.US.98.1058_11.AY331295	1998.1058	1910.8942
D.T.Z.2009.HM572334	3909.0	0.0
Ref.C.ET.86.ETH2220.U46016	1986.0	1923.0
Ref.C.ZA.04.04ZASK146.AY772699	2004.04	1904.96

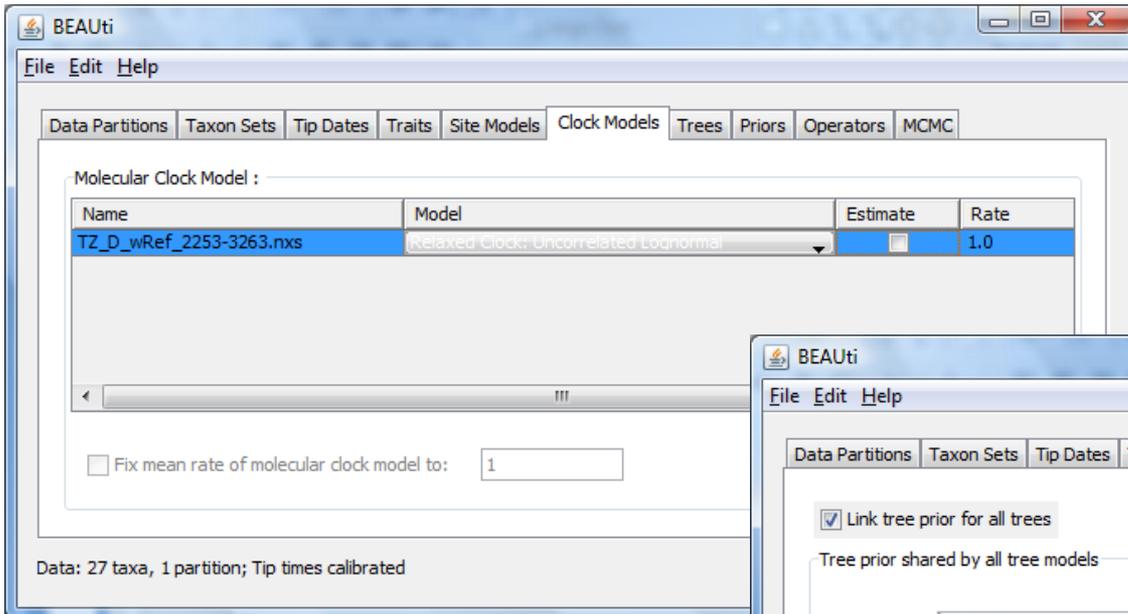
Tip date sampling: Off Apply to taxon set: All taxa

Data: 27 taxa, 1 partition; Fix clock rate to 1.0

Generate BEAST File...

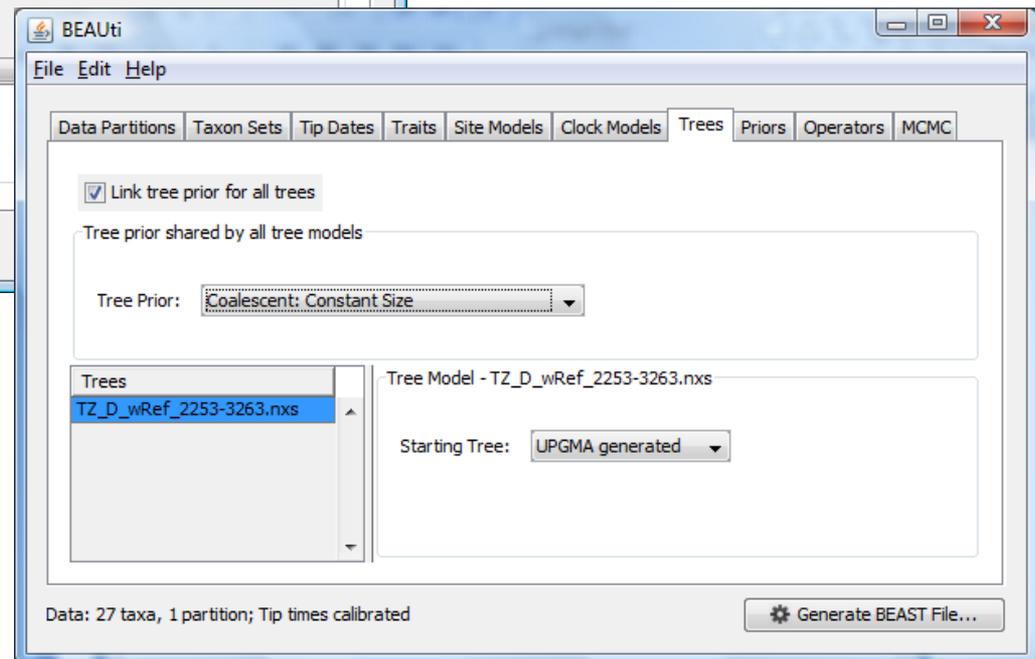


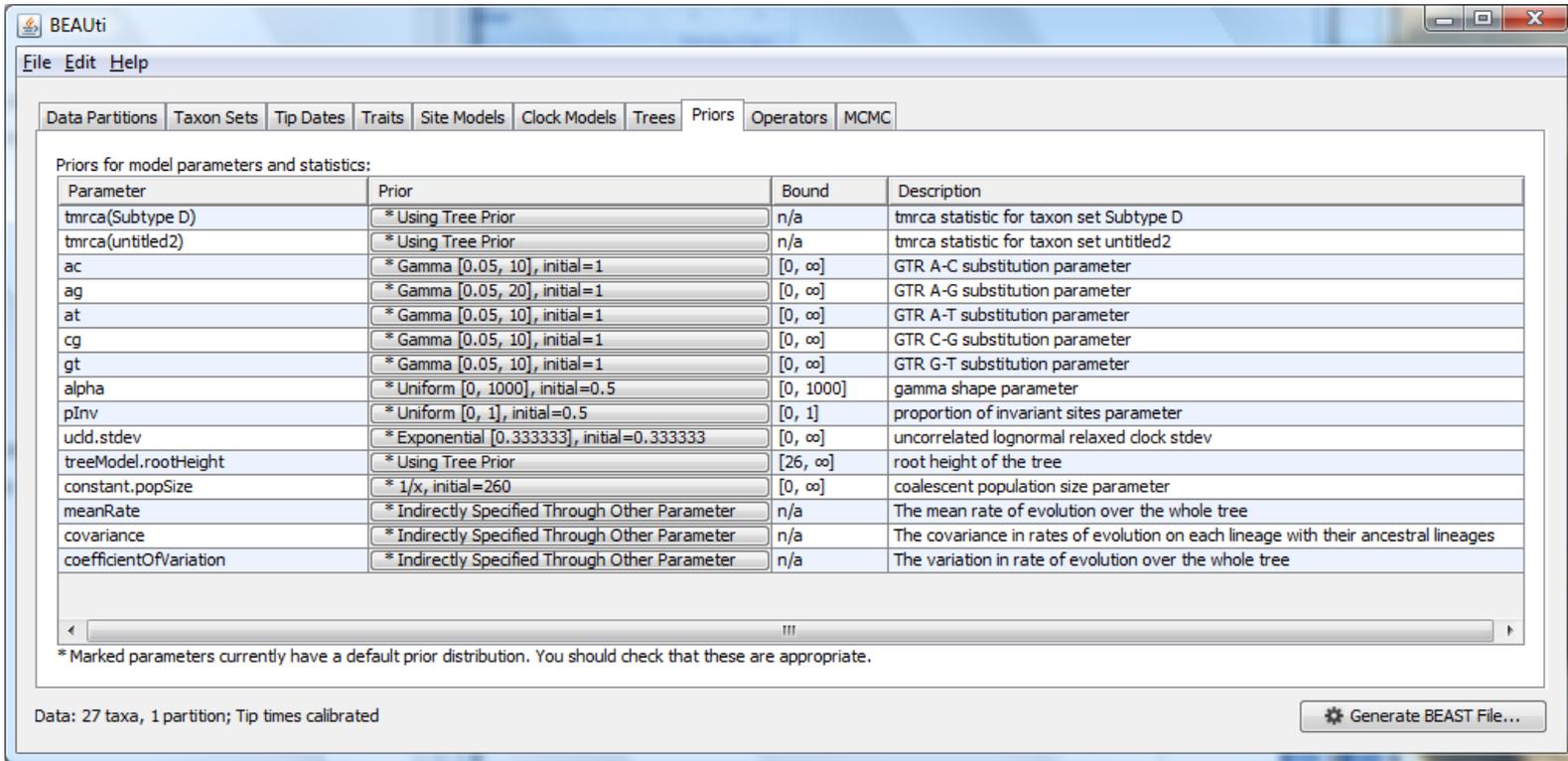
We know from the jModeltest result that GTR+ Γ +I fits this data best



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

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)	* Using Tree Prior	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=0.5	[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	* Using Tree Prior	[26, ∞]	root height of the tree
constant.popSize	* 1/x, initial=260	[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; Tip times calibrated

Generate BEAST File...

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 literature 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

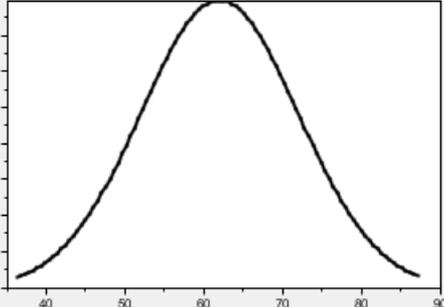
Prior for Parameter

Select prior distribution for tmrca(Subtype D)

Prior Distribution: **Normal**

Mean:

Stdev:



Quantiles: 2.5%: 42.4
5%: 45.55
Median: 62.0
95%: 78.45
97.5%: 81.6

Prior for Parameter

Select prior distribution for treeModel.rootHeight

Prior Distribution: **Uniform**

Initial Value:

Lower:

Upper:

BEAUTi

File Edit Help

Data Partitions Taxon Sets Tip Dates Traits Site Models Clock Models Trees Priors Operators MCMC

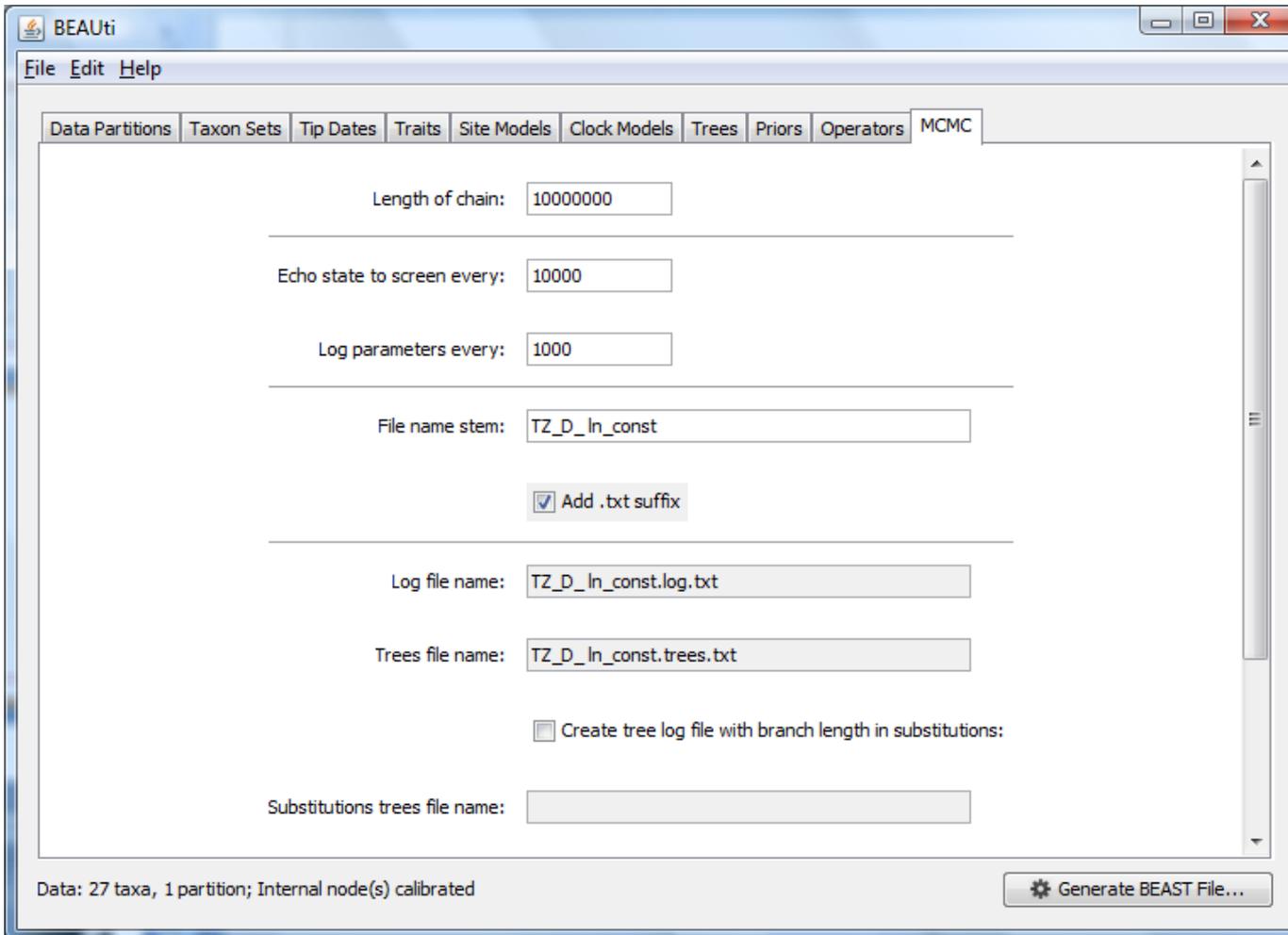
Priors for model parameters and statistics:

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
ucd.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

BEAUTi



The screenshot shows the BEAUTi software window with the MCMC tab selected. The interface includes a menu bar (File, Edit, Help) and a tabbed menu (Data Partitions, Taxon Sets, Tip Dates, Traits, Site Models, Clock Models, Trees, Priors, Operators, MCMC). The MCMC settings are as follows:

- Length of chain: 10000000
- Echo state to screen every: 10000
- Log parameters every: 1000
- File name stem: TZ_D_In_const
- Add .txt suffix
- Log file name: TZ_D_In_const.log.txt
- Trees file name: TZ_D_In_const.trees.txt
- Create tree log file with branch length in substitutions:
- Substitutions trees file name: (empty)

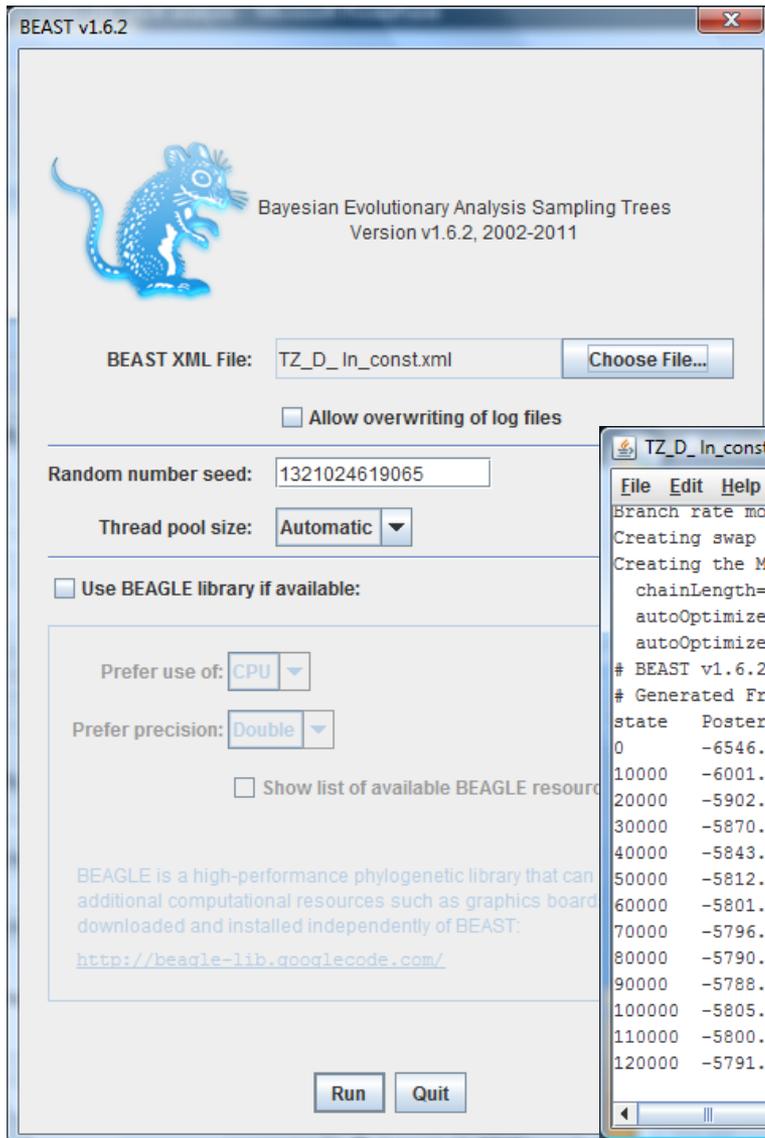
At the bottom left, it says "Data: 27 taxa, 1 partition; Internal node(s) calibrated". At the bottom right, there is a "Generate BEAST File..." button.

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_In_const
TZ_D_In_BSL
TZ_D_exp_const
TZ_D_exp_BSL

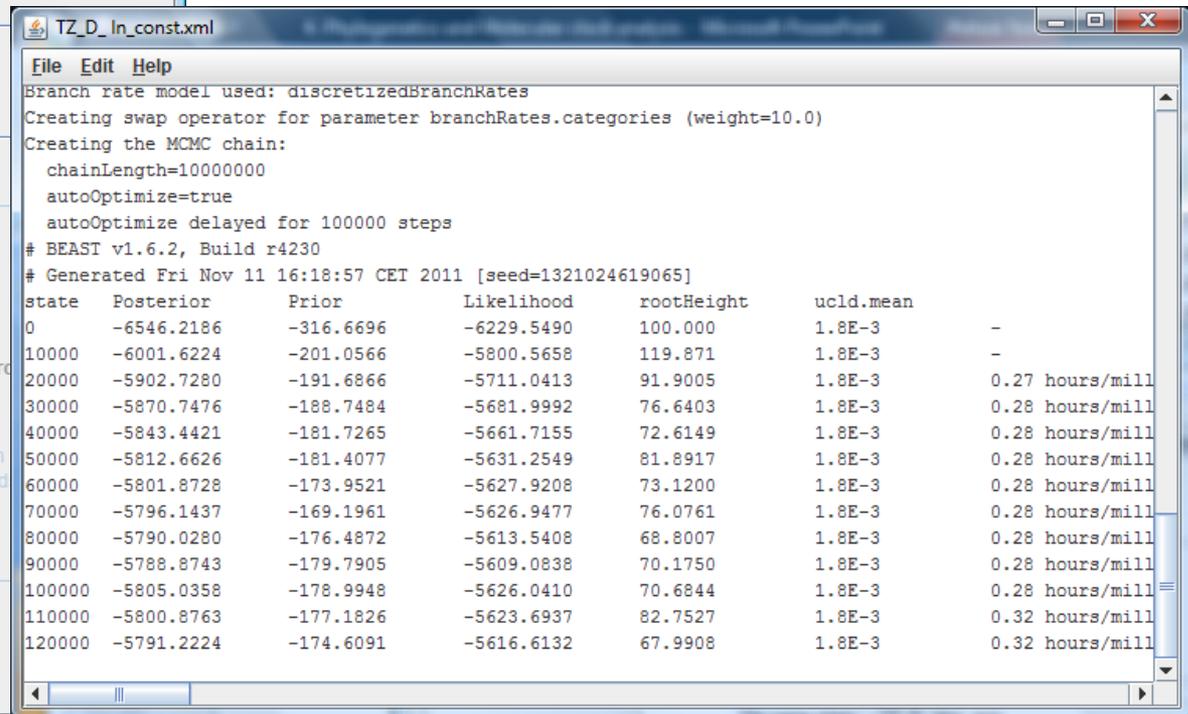
As these analyses are computationally demanding, start with "short" runs while testing the models (around 10 million). Then, after you have performed Bayes factor analysis to decide on the best model you can run longer chains if necessary (can be >100 million), use the Tracer file to check for convergence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

BEAST



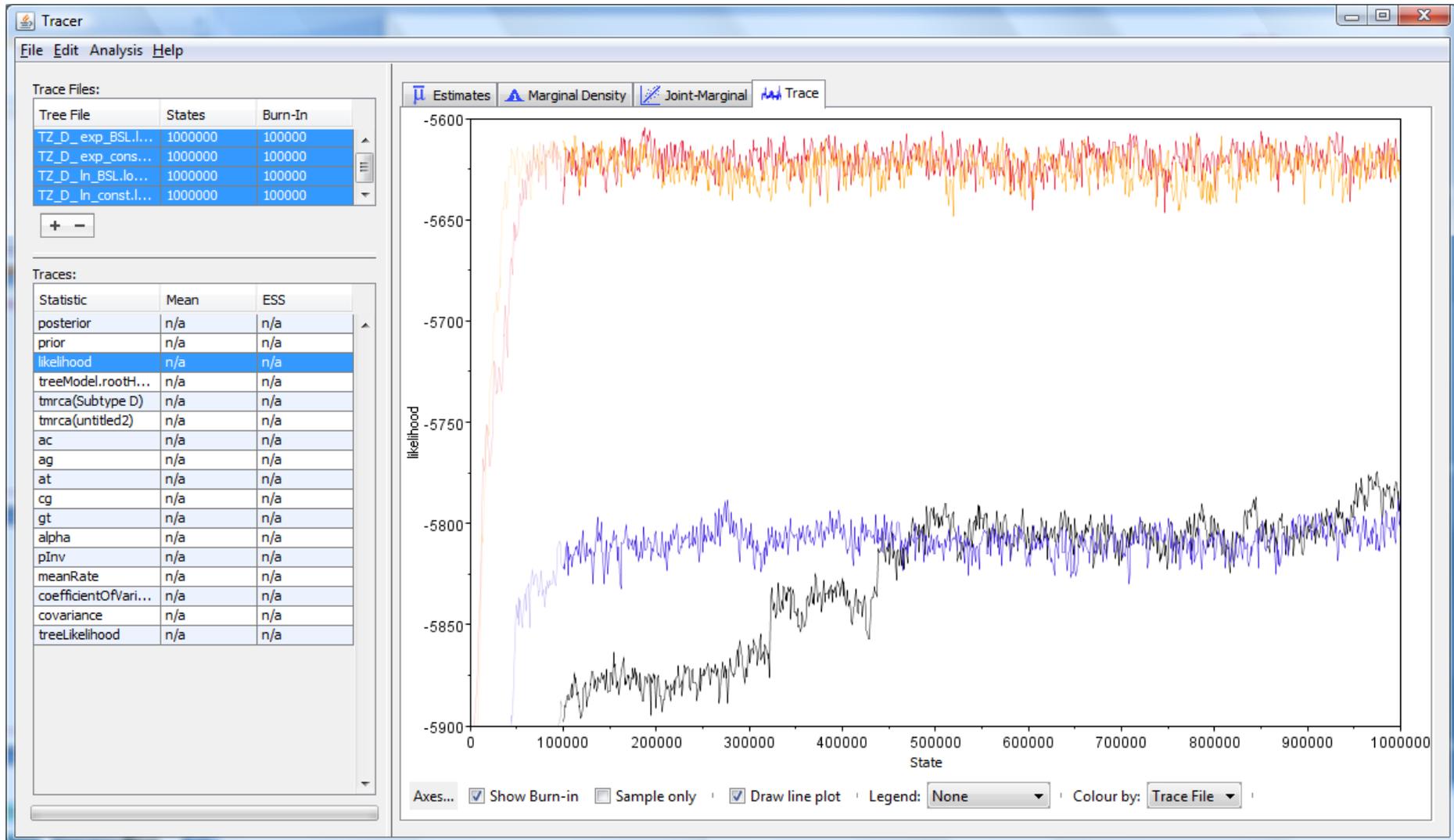
To make a BEAST run – just choose the .xml file created in BEAUTi and click Run.

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

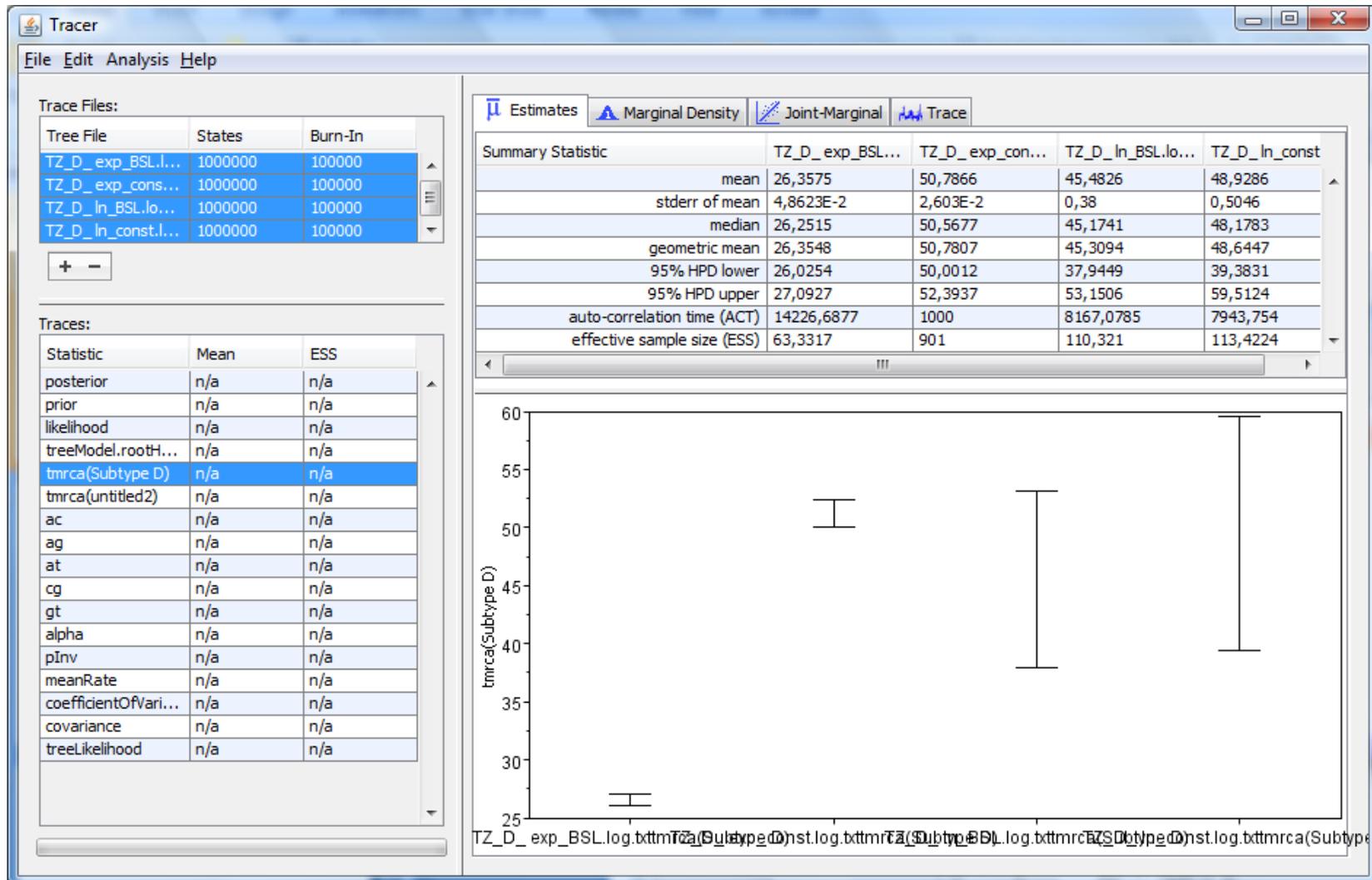


Tracer

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.

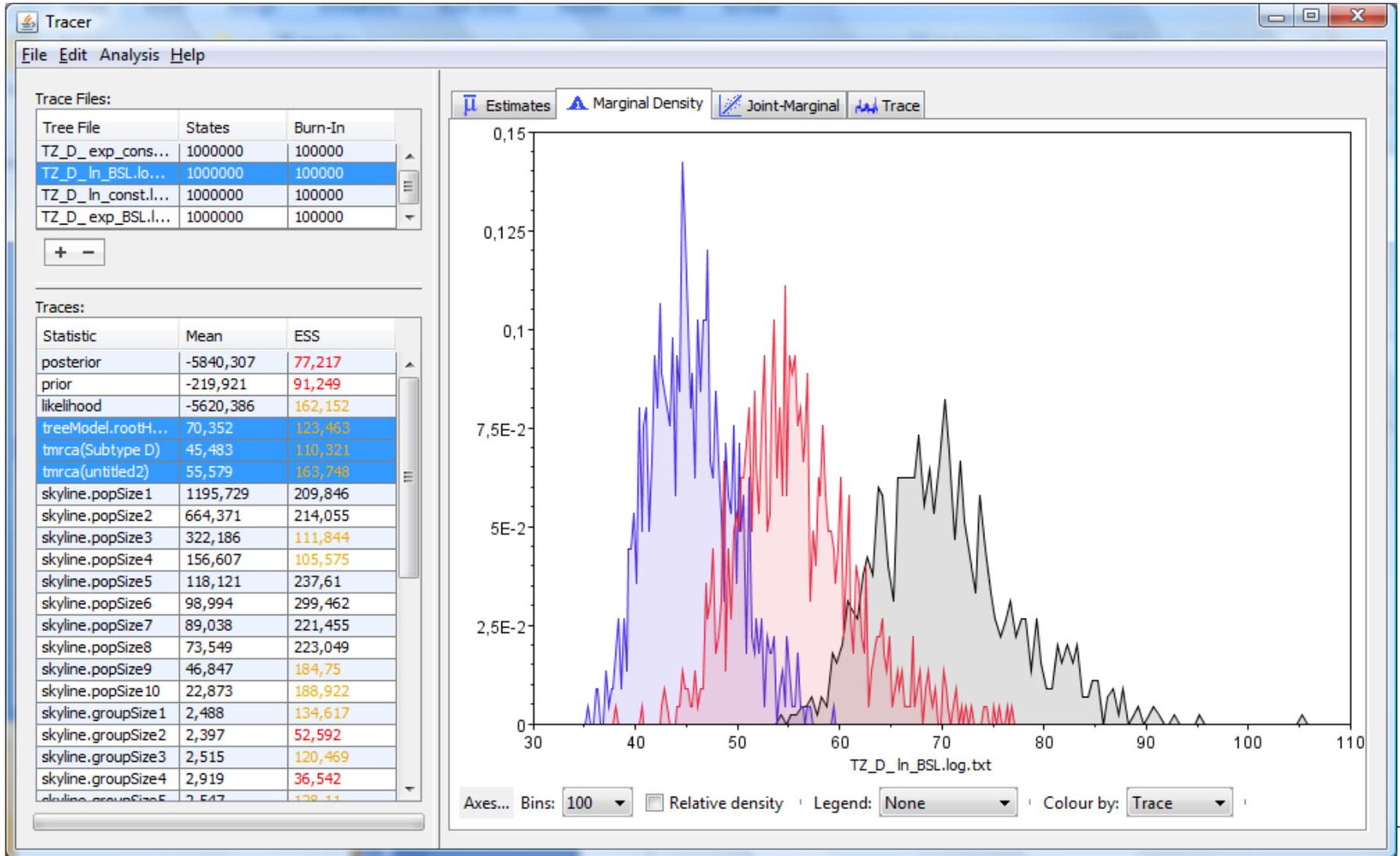


Tracer

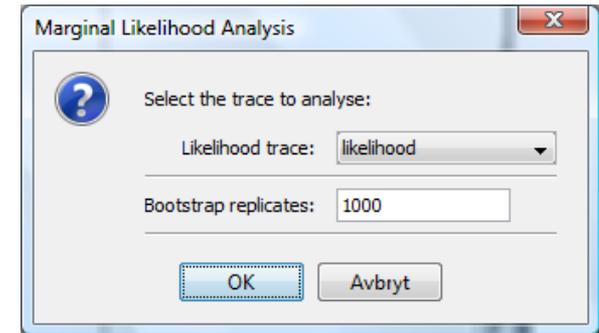
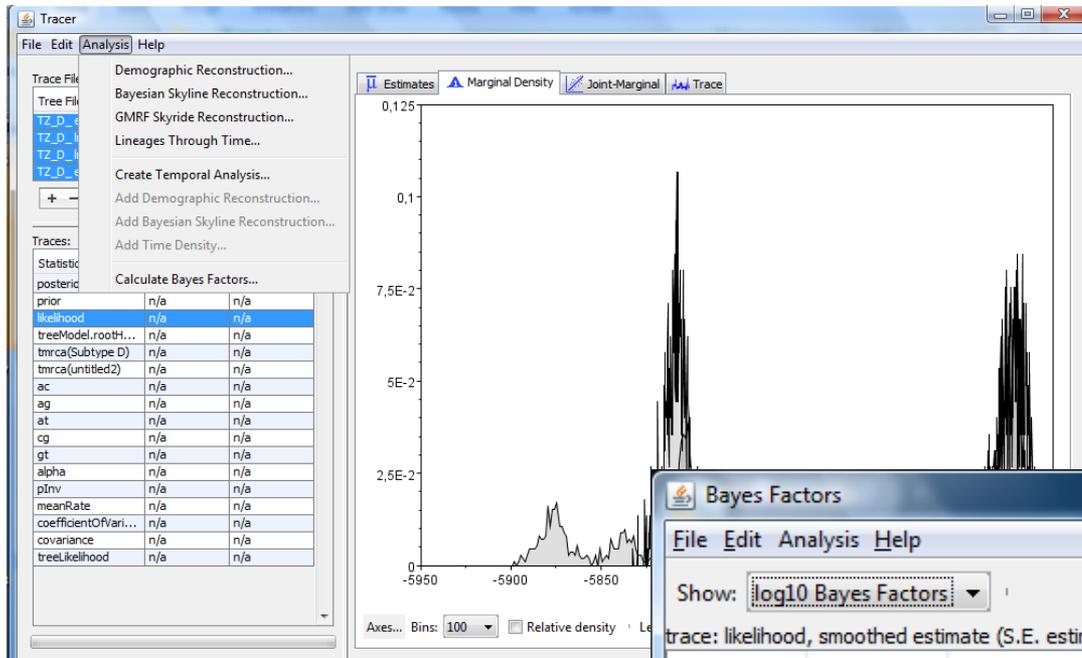
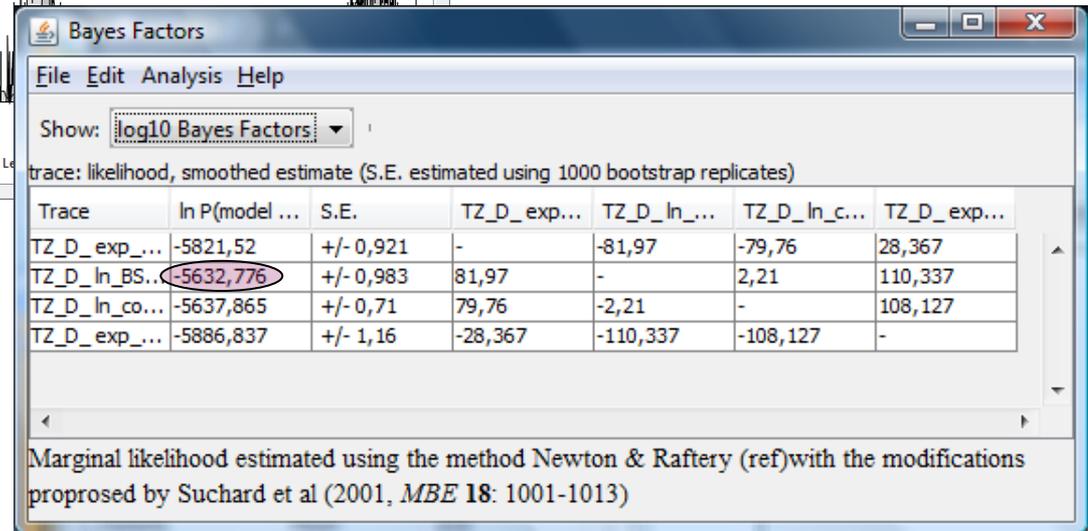


Tracer

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 Factors

Show: **log10 Bayes Factors**

Trace: likelihood, smoothed estimate (S.E. estimated using 1000 bootstrap replicates)

Trace	ln P(model ...)	S.E.	TZ_D_exp...	TZ_D_ln...	TZ_D_ln_c...	TZ_D_exp...
TZ_D_exp...	-5821,52	+/- 0,921	-	-81,97	-79,76	28,367
TZ_D_ln_BS...	-5632,776	+/- 0,983	81,97	-	2,21	110,337
TZ_D_ln_co...	-5637,865	+/- 0,71	79,76	-2,21	-	108,127
TZ_D_exp...	-5886,837	+/- 1,16	-28,367	-110,337	-108,127	-

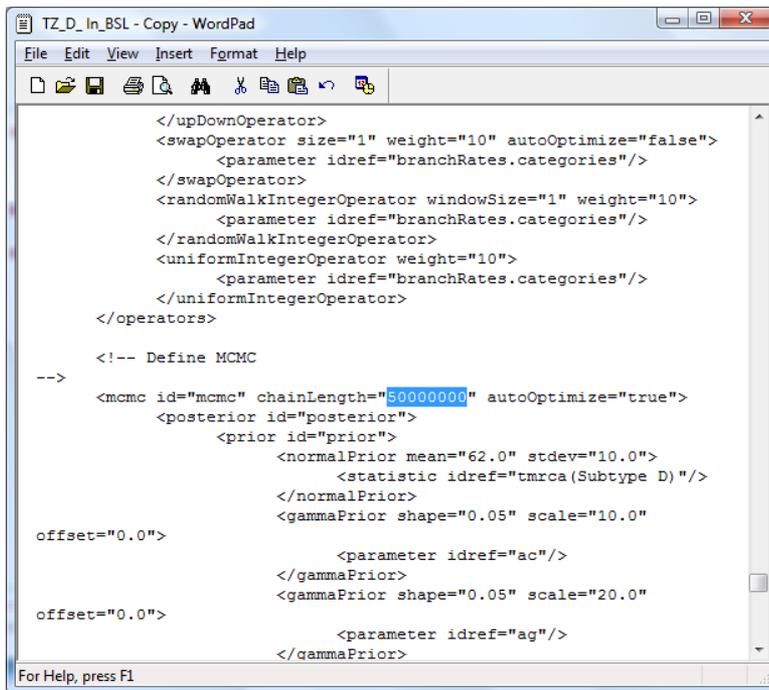
Marginal likelihood estimated using the method Newton & Raftery (ref) with the modifications proposed by Suchard et al (2001, *MBE* 18: 1001-1013)

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

TZ-D_In_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_In_const (2.21) -> can continue with longer runs for both models

Perform longer runs using the selected model(s)

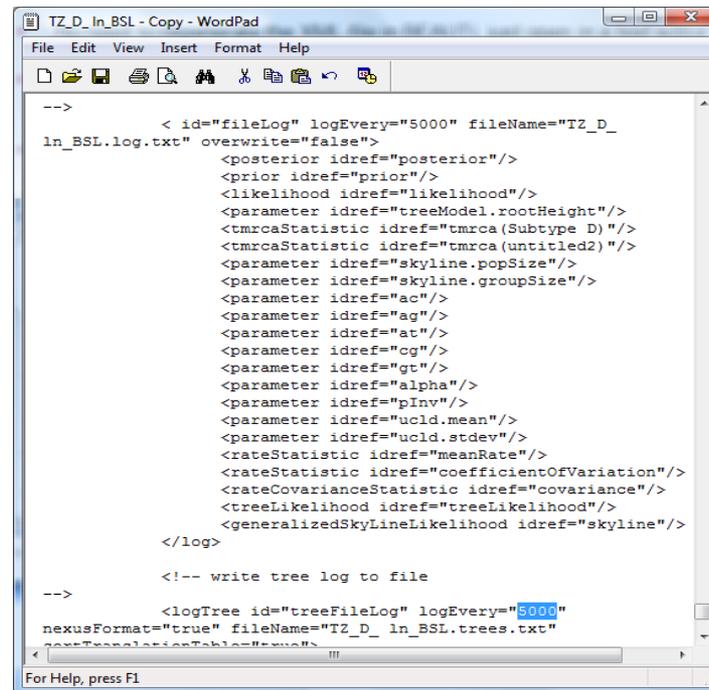
- No need to regenerate the XML-file in BEAUTi, just open in a text editor (for example WordPad), and change the chain length.
- It may also be useful to change the log frequencies (screenLog, fileLog, treeFileLog). Rule of thumb – obtain 10000 trees. Ex run of 50 million generations: fileLog and treeFileLog = 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!)



```
</upDownOperator>
<swapOperator size="1" weight="10" autoOptimize="false">
  <parameter idref="branchRates.categories"/>
</swapOperator>
<randomWalkIntegerOperator windowSize="1" weight="10">
  <parameter idref="branchRates.categories"/>
</randomWalkIntegerOperator>
<uniformIntegerOperator weight="10">
  <parameter idref="branchRates.categories"/>
</uniformIntegerOperator>
</operators>

<!-- Define MCMC
-->
<mcmc id="mcmc" chainLength="50000000" autoOptimize="true">
  <posterior id="posterior">
    <prior id="prior">
      <normalPrior mean="62.0" stdev="10.0">
        <statistic idref="tmrca(Subtype D)"/>
      </normalPrior>
      <gammaPrior shape="0.05" scale="10.0">
        <parameter idref="ac"/>
      </gammaPrior>
      <gammaPrior shape="0.05" scale="20.0">
        <parameter idref="ag"/>
      </gammaPrior>
    </prior>
  </posterior>
  <parameter idref="treeModel.rootHeight"/>
  <tmrcaStatistic idref="tmrca(Subtype D)"/>
  <tmrcaStatistic idref="tmrca(untitled2)"/>
  <parameter idref="skyline.popSize"/>
  <parameter idref="skyline.groupSize"/>
  <parameter idref="ac"/>
  <parameter idref="ag"/>
  <parameter idref="at"/>
  <parameter idref="cg"/>
  <parameter idref="gt"/>
  <parameter idref="alpha"/>
  <parameter idref="pInv"/>
  <parameter idref="uclid.mean"/>
  <parameter idref="uclid.stdev"/>
  <rateStatistic idref="meanRate"/>
  <rateStatistic idref="coefficientOfVariation"/>
  <rateCovarianceStatistic idref="covariance"/>
  <treeLikelihood idref="treeLikelihood"/>
  <generalizedSkyLineLikelihood idref="skyline"/>
</log>

-->
<!-- write tree log to file
-->
<logTree id="treeFileLog" logEvery="5000"
nexusFormat="true" fileName="TZ_D_in_BSL.trees.txt"
partTransalationTables="nexus">
  </logTree>
</mcmc>
```

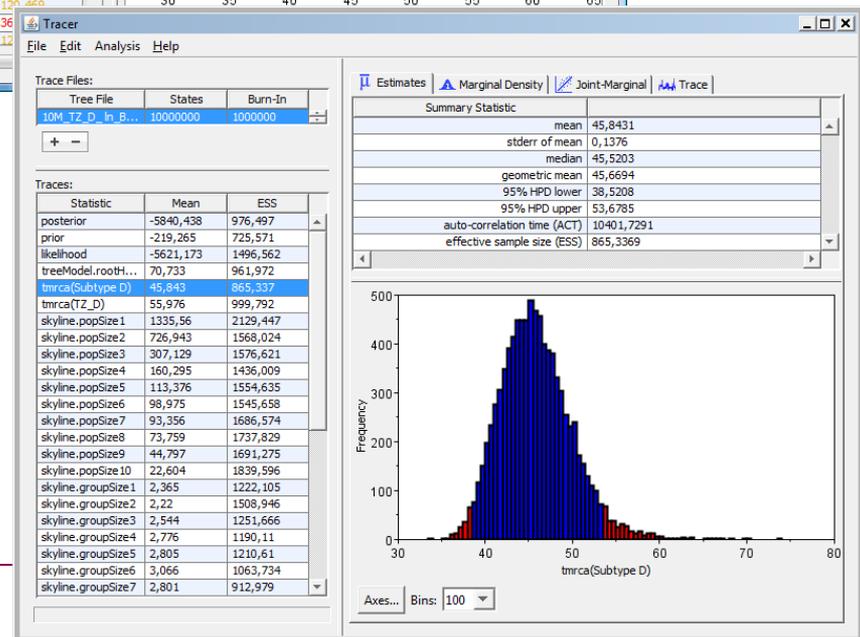
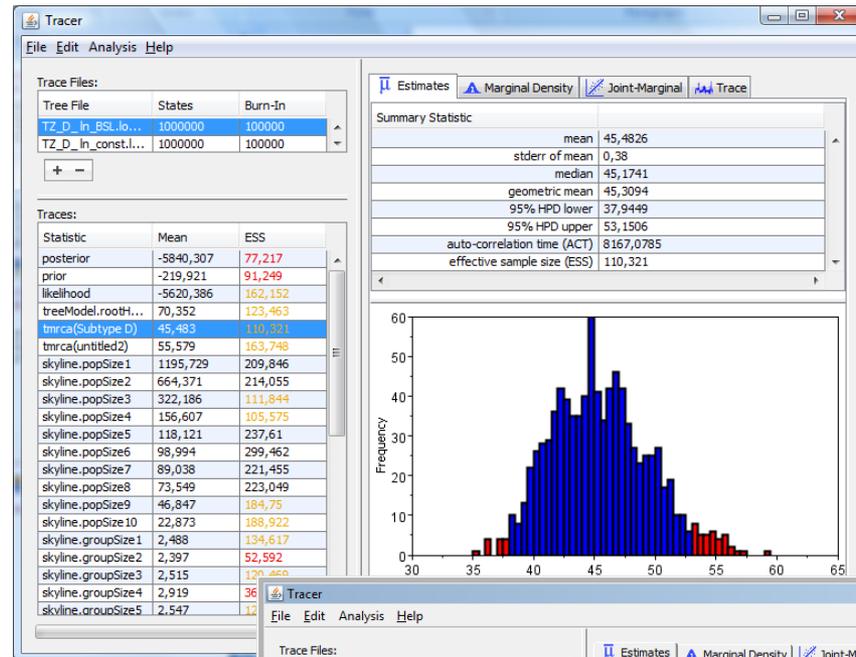


```
-->
  <id="fileLog" logEvery="5000" fileName="TZ_D_in_BSL.log.txt" overwrite="false">
    <posterior idref="posterior"/>
    <prior idref="prior"/>
    <likelihood idref="likelihood"/>
    <parameter idref="treeModel.rootHeight"/>
    <tmrcaStatistic idref="tmrca(Subtype D)"/>
    <tmrcaStatistic idref="tmrca(untitled2)"/>
    <parameter idref="skyline.popSize"/>
    <parameter idref="skyline.groupSize"/>
    <parameter idref="ac"/>
    <parameter idref="ag"/>
    <parameter idref="at"/>
    <parameter idref="cg"/>
    <parameter idref="gt"/>
    <parameter idref="alpha"/>
    <parameter idref="pInv"/>
    <parameter idref="uclid.mean"/>
    <parameter idref="uclid.stdev"/>
    <rateStatistic idref="meanRate"/>
    <rateStatistic idref="coefficientOfVariation"/>
    <rateCovarianceStatistic idref="covariance"/>
    <treeLikelihood idref="treeLikelihood"/>
    <generalizedSkyLineLikelihood idref="skyline"/>
  </log>

-->
  <!-- write tree log to file
-->
  <logTree id="treeFileLog" logEvery="5000"
nexusFormat="true" fileName="TZ_D_in_BSL.trees.txt"
partTransalationTables="nexus">
    </logTree>
  </log>
</mcmc>
```

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

Bayesian Skyline Analysis

Warning! This analysis should only be run on traces where the Bayesian Skyline plot was specified as the demographic in BEAST. Any other model will produce meaningless results.

Trees Log File: TZ_D_In_BSL.log.txt

Bayesian skyline variant: Stepwise (Constant) ▾

Plot growth rate

Select the traces to use for the arguments:

Population Size: skyline.popSize ▾

Group Size: skyline.groupSize ▾

Maximum time is the root height's: Lower 95% HPD ▾

Select the trace of the root height: treeModel.rootHeight ▾

Number of bins: 100

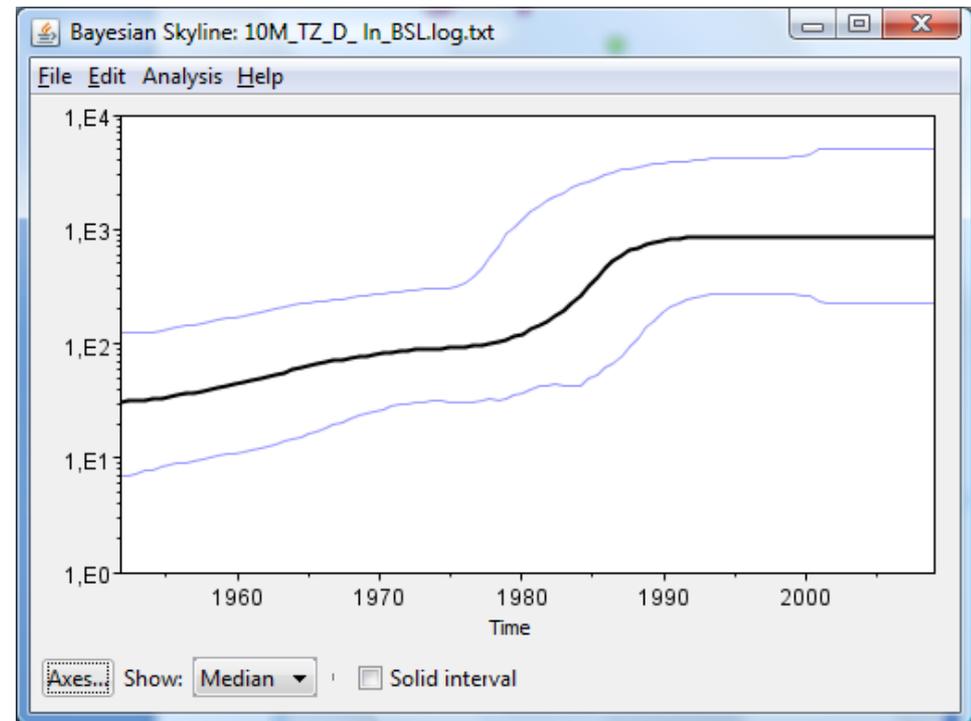
Use manual range for bins:

Minimum time:

Maximum time:

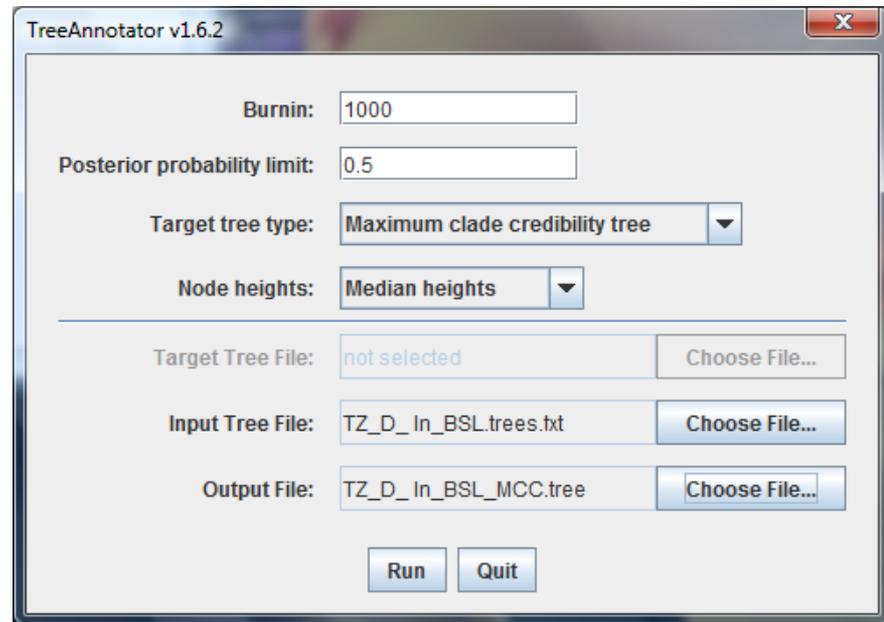
Age of youngest tip: 2009

You can set the age of sampling of the most recent tip in the tree. If this is set to zero then the plot is shown going backwards in time, otherwise forwards in time.



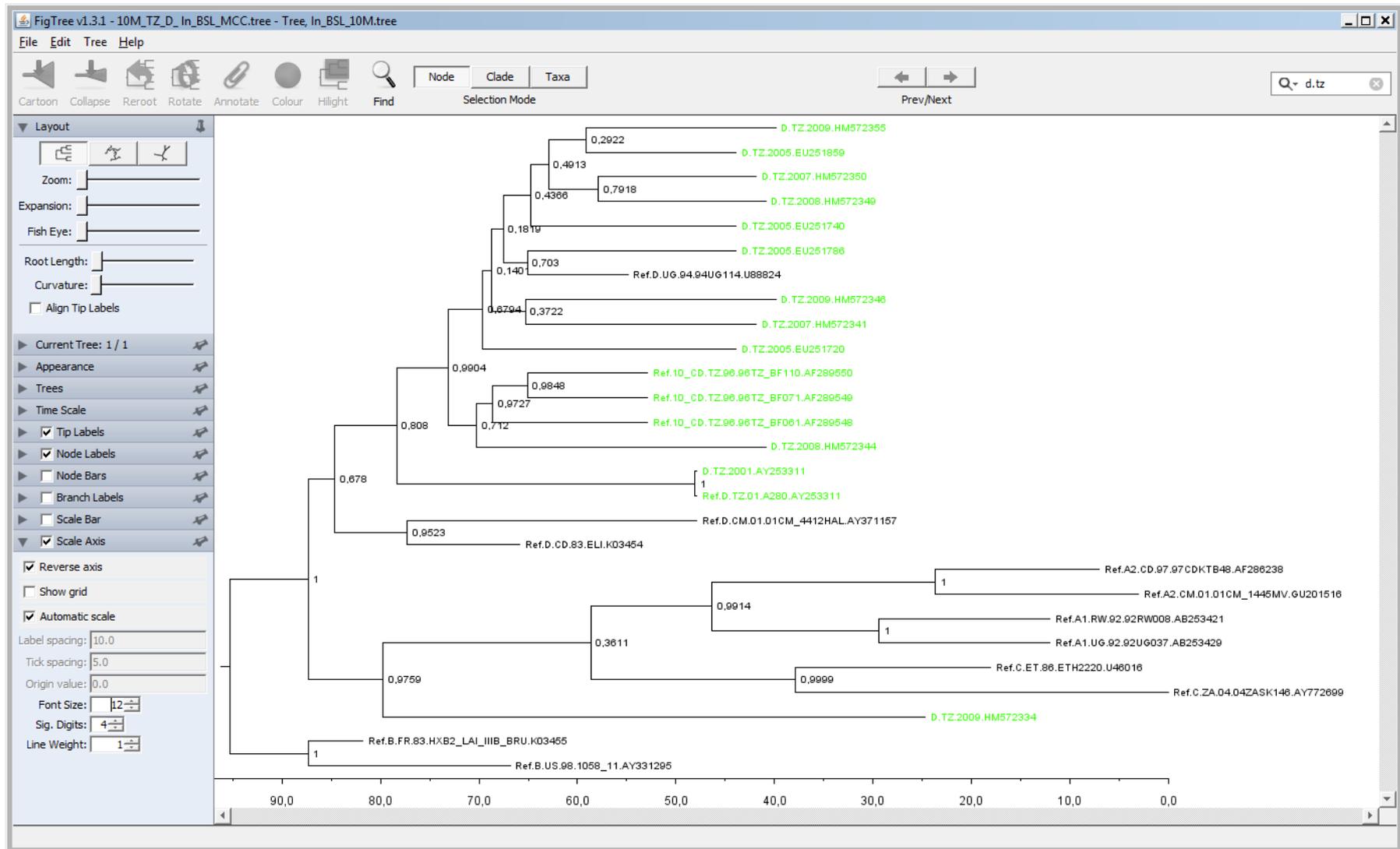
TreeAnnotator

- 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>

KATHOLIEKE UNIVERSITEIT
LEUVEN



About K.U.Leuven | Education | Research | Admissions | Living in Leuven | Alumni | Libraries | Faculties, Departments & Schools | International cooperation



Laboratory for Clinical and Evolutionary Virology > Meetings

International bioinformatics workshops

The seventeenth international bioinformatics workshop on virus evolution and molecular epidemiology will take place in 2012. Please stay tuned for updated information concerning place and time.

There will be no international bioinformatics workshop on virus evolution and molecular epidemiology in 2011. However we endorse the **Brazilian bioinformatics workshop on virus evolution and molecular epidemiology**, which will take place early September in Salvador de Bahia. Please contact Luiz Alcantara (lalcana@bahia.fiocruz.br) for further information.

- **Sixteenth international bioinformatics workshop on virus evolution and molecular epidemiology**
Johns Hopkins University, Baltimore, USA
Sunday, August 29, 2010 - Friday, September 3, 2010
- **Fifteenth international bioinformatics workshop on virus evolution and molecular epidemiology**
Erasmus Postgraduate School of Molecular Medicine, Erasmus Medical Center, Rotterdam, The Netherlands
Monday, September 7, 2009 - Friday, September 11, 2009
- **Fourteenth international bioinformatics workshop on virus evolution and molecular epidemiology**
South African National Bioinformatics Institute, Cape Town, South Africa
Monday, September 8, 2008 - Sunday, September 14, 2008
- **Thirteenth international bioinformatics workshop on virus evolution and molecular epidemiology**
Instituto Nacional de Saude Dr. Ricardo Jorge (INSA), Lisbon, Portugal
Sunday, September 9, 2007 - Friday, September 14, 2007
- **Twelfth international bioinformatics workshop on virus evolution and molecular epidemiology**
National Retrovirus Reference Center, Department of Hygiene and Epidemiology, Medical School, National and Kapodistrian University of Athens, Greece
Sunday, September 10, 2006 - Friday, September 15, 2006

Laboratory for Clinical and Evolutionary Virology

People

Publications

Research

Meetings

International bioinformatics workshops

European HIV & Hepatitis workshops

Software

Courses

Contact Us

How to get here

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

