Measurably evolving populations

Alexei J Drummond, Remco Bouckaert and Walter Xie

January 23, 2015

1 Time-stamped data

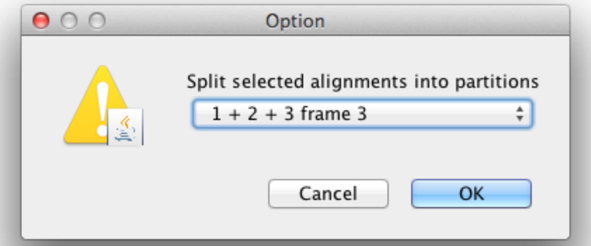
This tutorial estimates the rate of evolution from a set of virus sequences which have been isolated at different points in time (heterochronous or time-stamped data). The data are 129 sequences from the G (attachment protein) gene of human respiratory syncytial virus subgroup A (RSVA) from various parts of the world with isolation dates ranging from 1956-2002 [2, 1]. RSVA causes infections of the lower respiratory tract causing symptoms that are often indistinguishable from the common cold. By age 3, nearly all children will be infected and a small percentage ($< 3\%$) will develop more serious inflammation of the bronchioles requiring hospitalisation.

The aim of this tutorial is to obtain estimates for :

- the rate of molecular evolution
- the date of the most recent common ancestor
- the phylogenetic relationships with measures of statistical support.

The following software will be used in this tutorial:

- **BEAST** - this package contains the BEAST program, BEAUti, DensiTree, TreeAnnotator and other utility programs. This tutorial is written for BEAST v2.2.x, which has support for multiple partitions. It is available for download from `http://www.beast2.org`.
- **Tracer** - this program is used to explore the output of BEAST (and other Bayesian MCMC programs). It graphically and quantitively summarizes the distributions of continuous parameters and provides diagnostic information. At the time of writing, the current version is v1.6. It is available for download from `http://beast.bio.ed.ac.uk/`.
- **FigTree** - this is an application for displaying and printing molecular phylogenies, in particular those obtained using BEAST. At the time of writing, the current version is v1.4.2. It is available for download from `http://tree.bio.ed.ac.uk/`.



Split selected alignments into partitions

1 + 2 + 3 frame 3

Figure 1: Split the alignment

The NEXUS alignment

The data is in a file called `RSV2.nex` and you can find it in the `examples/nexus` directory in the directory where BEAST was installed. This file contains an alignment of 129 sequences from the G gene of RSVA virus, 629 nucleotides in length. Import this alignment into BEAUti. Because this is a protein-coding gene we are going to split the alignment into three partitions representing each of the three codon positions. To do this we will click the **Split** button at the bottom of the **Partitions** panel and then select the “1 + 2 + 3 frame 3” from the drop-down menu (Figure 1).

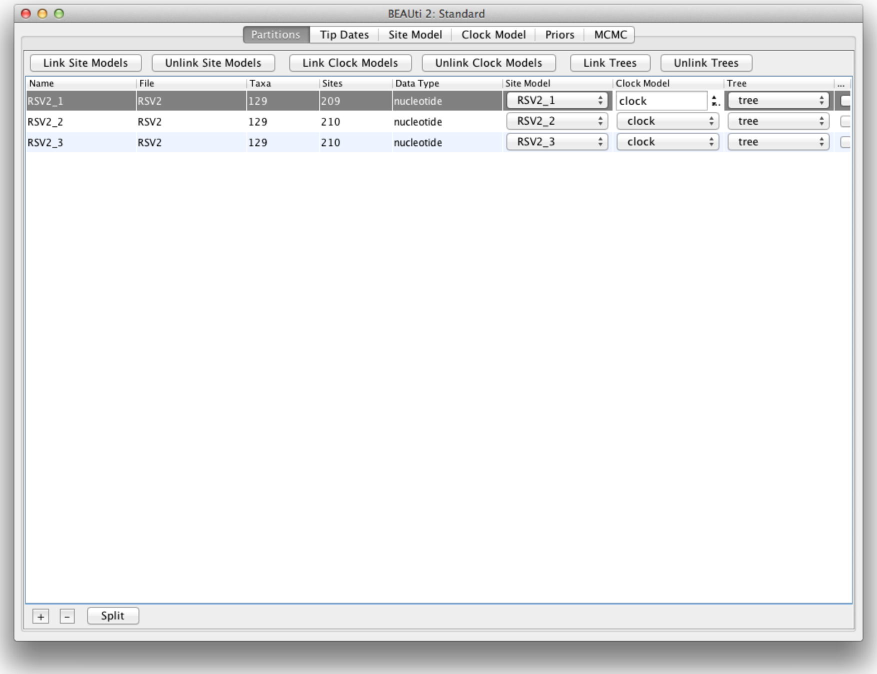
This signifies that the first full codon starts at the third nucleotide in the alignment. This will create three rows in the partitions panel. You will have to re-link the tree and clock models across the three partitions (and name them “tree” and “clock” respectively) before continuing to the next step. The partition panel should now look something like this:

Tip dates

By default all the taxa are assumed to have a date of zero (i.e. the sequences are assumed to be sampled at the same time). In this case, the RSVA sequences have been sampled at various dates going back to the 1950s. The actual year of sampling is given in the name of each taxon and we could simply edit the value in the Date column of the table to reflect these. However, if the taxa names contain the calibration information, then a convenient way to specify the dates of the sequences in BEAUti is to click the checkbox **Use tip dates** and then use the **Guess** button at the top of the **Tip Dates** panel. Clicking this will make a dialog box appear.

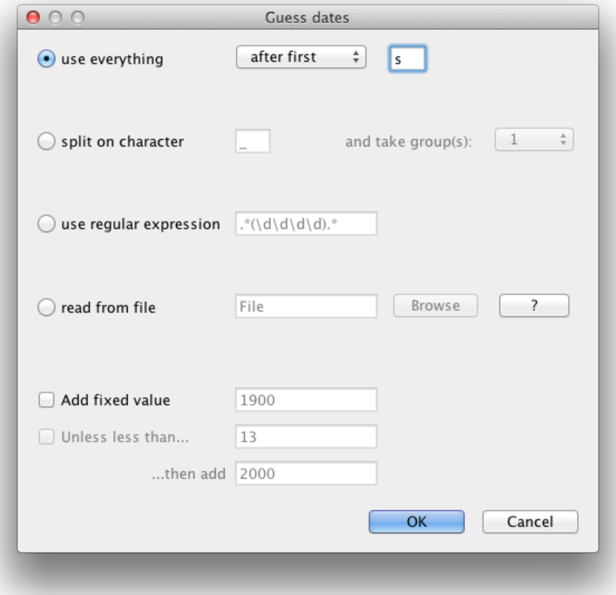
Select the option to *use everything*, choose “after last” from from drop-down box and type‘s’ into the corresponding text box. This will extract the trailing numbers from the taxon names after the last little ’s’, which are interpreted as the year (in this case since 1900) that the sample was isolated.

The dates panel should now look something like this:


BEAUti 2: Standard
Partitions
Tip Dates
Site Model
Clock Model
Priors
MCMC
Link Site Models
Unlink Site Models
Link Clock Models
Unlink Clock Models
Link Trees
Unlink Trees
+
-
Split

Name	File	Taxa	Sites	Data Type	Site Model	Clock Model	Tree	...
RSV2_1	RSV2	129	209	nucleotide	RSV2_1	clock	tree	☐
RSV2_2	RSV2	129	210	nucleotide	RSV2_2	clock	tree	☐
RSV2_3	RSV2	129	210	nucleotide	RSV2_3	clock	tree	☐

Figure 2: A screenshot of the Partitions tab in BEAUti


Guess dates
Browse
?
OK
Cancel

● use everything after first s

○ split on character _ and take group(s): 1

○ use regular expression .*(\d\d\d\d).*

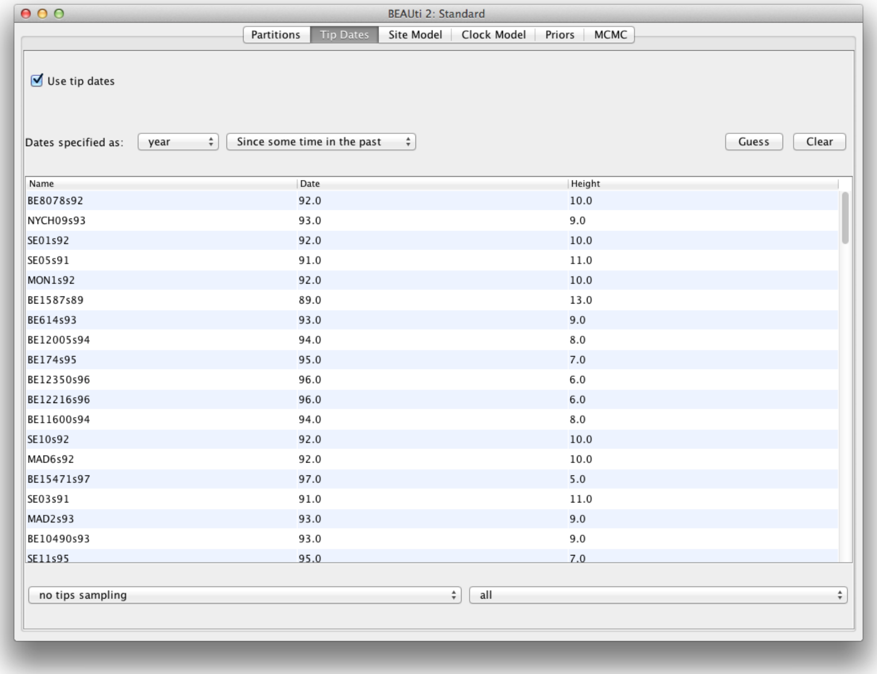
○ read from file File

☐ Add fixed value 1900

☐ Unless less than... 13

...then add 2000

Figure 3: Guess dates dialog



☑ Use tip dates

Dates specified as: year, Since some time in the past

Name	Date	Height
BE8078s92	92.0	10.0
NYCH09s93	93.0	9.0
SE01s92	92.0	10.0
SE05s91	91.0	11.0
MON1s92	92.0	10.0
BE1587s89	89.0	13.0
BE614s93	93.0	9.0
BE12005s94	94.0	8.0
BE174s95	95.0	7.0
BE12350s96	96.0	6.0
BE12216s96	96.0	6.0
BE11600s94	94.0	8.0
SE10s92	92.0	10.0
MAD6s92	92.0	10.0
BE15471s97	97.0	5.0
SE03s91	91.0	11.0
MAD2s93	93.0	9.0
BE10490s93	93.0	9.0
SE11s95	95.0	7.0

no tips sampling, all

Figure 4: Dates panel

Setting the substitution model

We will use the HKY model with empirical base frequencies for all three partitions. To do this first link the site partitions and then choose HKY and Empirical from the Subst Model and Frequencies drop-boxes. Also check the estimate box for the Mutation Rate,which will finally trigger to check the "Fix mean mutation rate" box.

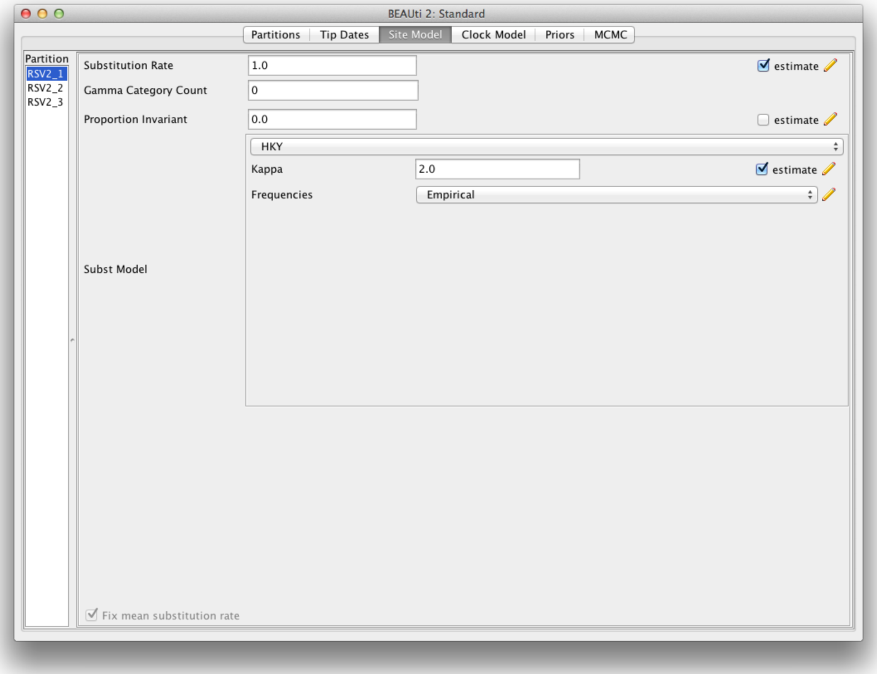
Then, hold 'shift' key to select all site models on the left side, and click **OK** to clone the setting from defined site model (Figure 6). Go through each site model, as you can see, their configurations are same now.

1.0.1 Priors

To set up the priors, select the "Priors" tab. Choose "Coalescent Constant Population" for the tree prior. Set the prior on the clockRate parameter to a log-normal with $M = -5$ and $S = 1.25$.

1.1 Setting the MCMC options

For this dataset let's initially set the chain length to 2,000,000 as this will run reasonably quickly on most modern computers. Set the sampling frequencies for the screen to to 10,000, the trace log file to 400 and the trees file to 400.


BEAUti 2: Standard
Partitions
Tip Dates
Site Model
Clock Model
Priors
MCMC

Partition

RSV2_1
RSV2_2
RSV2_3

Substitution Rate 1.0 ☑ estimate

Gamma Category Count 0

Proportion Invariant 0.0 ☐ estimate

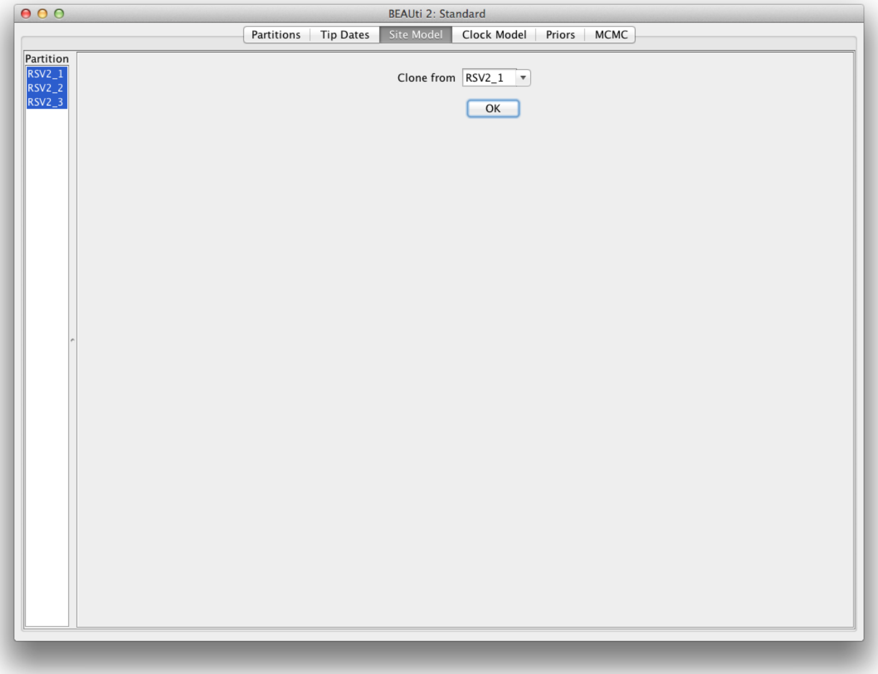
Subst Model HKY

Kappa 2.0 ☑ estimate

Frequencies Empirical

☑ Fix mean substitution rate

Figure 5: Site model

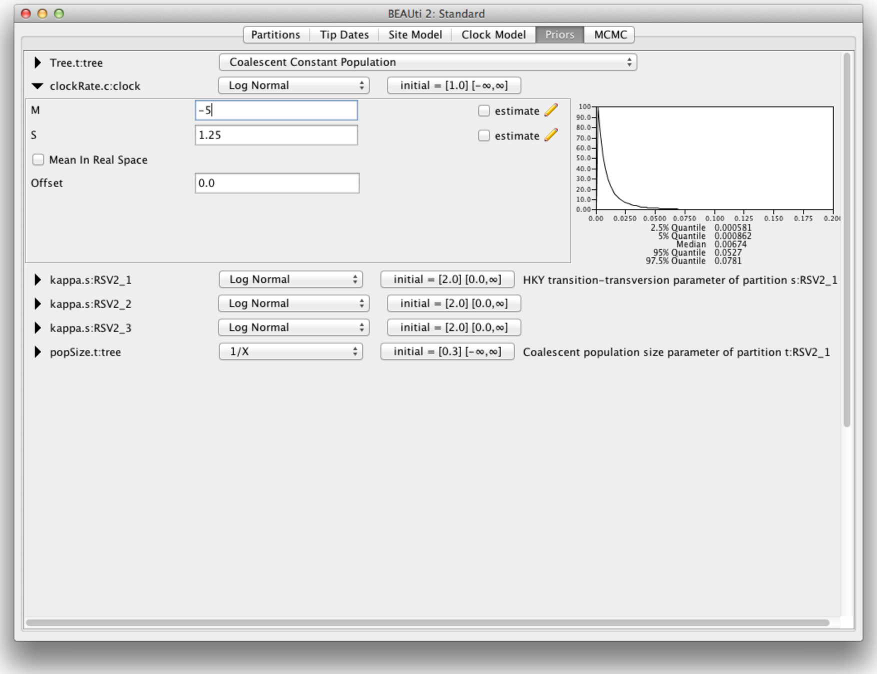

BEAUti 2: Standard
Partitions
Tip Dates
Site Model
Clock Model
Priors
MCMC
OK

Partition

RSV2_1
RSV2_2
RSV2_3

Clone from RSV2_1

Figure 6: Clone configuration from one site model to others.


BEAUti 2: Standard
Partitions
Tip Dates
Site Model
Clock Model
Priors
MCMC
0.00
0.0250
0.0500
0.0750
0.100
0.125
0.150
0.175
0.200
100
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.00

Tree.t:tree: Coalescent Constant Population

clockRate.c:clock: Log Normal — initial = [1.0] [−∞,∞]

M: −5 ☐ estimate

S: 1.25 ☐ estimate

☐ Mean In Real Space

Offset: 0.0

2.5% Quantile	0.000581
5% Quantile	0.000862
Median	0.00674
95% Quantile	0.0527
97.5% Quantile	0.0781

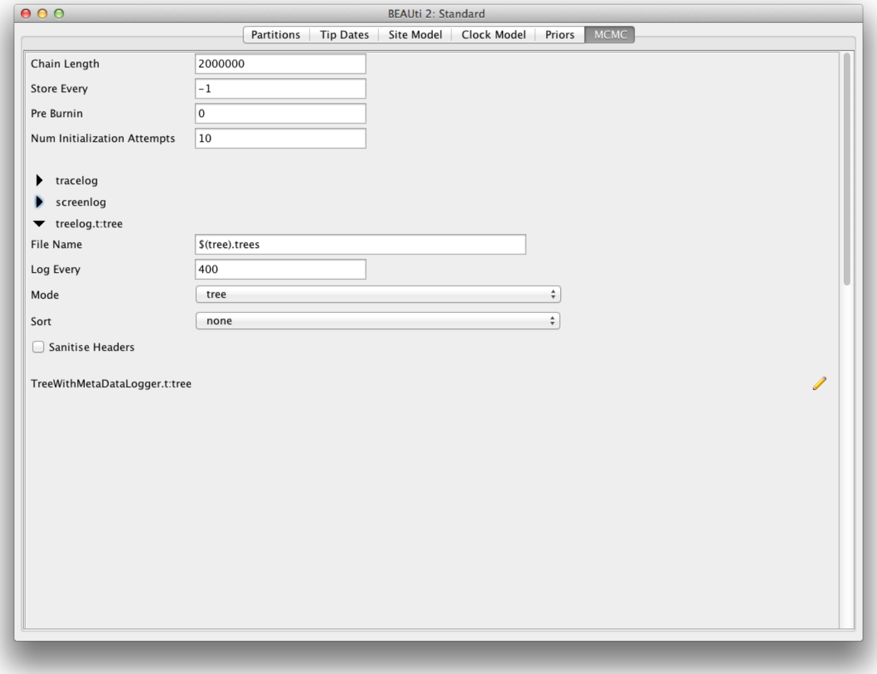
kappa.s:RSV2_1: Log Normal — initial = [2.0] [0.0,∞] — HKY transition−transversion parameter of partition s:RSV2_1

kappa.s:RSV2_2: Log Normal — initial = [2.0] [0.0,∞]

kappa.s:RSV2_3: Log Normal — initial = [2.0] [0.0,∞]

popSize.t:tree: 1/X — initial = [0.3] [−∞,∞] — Coalescent population size parameter of partition t:RSV2_1

Figure 7: Priors


BEAUti 2: Standard
Partitions
Tip Dates
Site Model
Clock Model
Priors
MCMC
tracelog
screenlog
treelog.t:tree

Chain Length: 2000000

Store Every: -1

Pre Burnin: 0

Num Initialization Attempts: 10

File Name: $(tree).trees

Log Every: 400

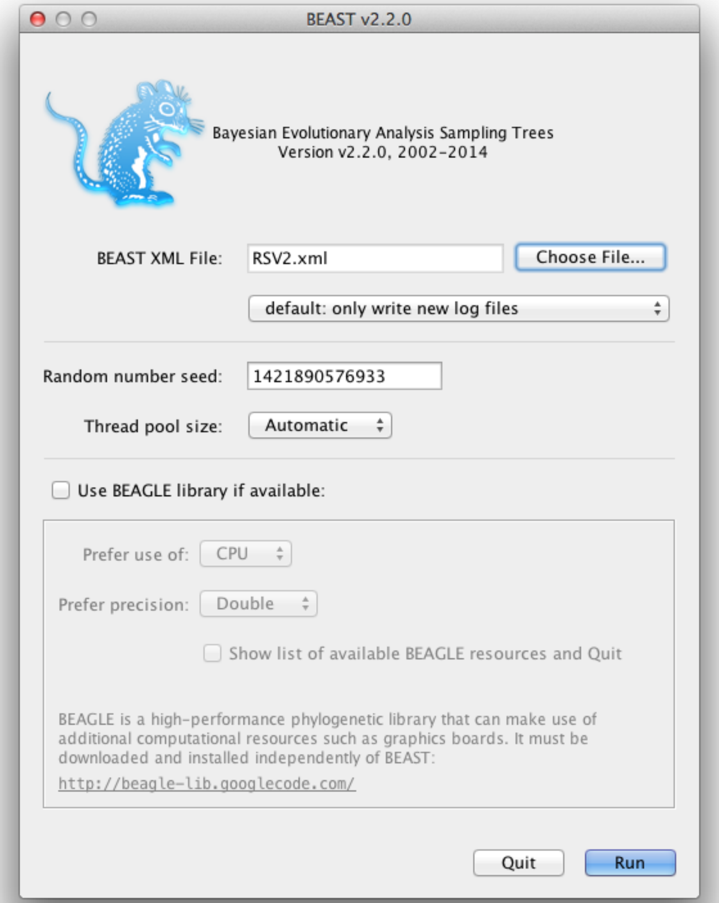
Mode: tree

Sort: none

☐ Sanitise Headers

TreeWithMetaDataLogger.t:tree

Figure 8: MCMC options



Bayesian Evolutionary Analysis Sampling Trees
Version v2.2.0, 2002-2014

BEAST XML File: RSV2.xml

default: only write new log files

Random number seed: 1421890576933

Thread pool size: Automatic

☐ Use BEAGLE library if available:

Prefer use of: CPU

Prefer precision: Double

☐ Show list of available BEAGLE resources and Quit

BEAGLE is a high-performance phylogenetic library that can make use of additional computational resources such as graphics boards. It must be downloaded and installed independently of BEAST:

http://beagle-lib.googlecode.com/

Figure 9: A screenshot of BEAST.

Running BEAST

Save the BEAST file (e.g. `RSV2.xml`) and run it in BEAST.

Now run BEAST and when it asks for an input file, provide your newly created XML file as input. BEAST will then run until it has finished reporting information to the screen. The actual results files are save to the disk in the same location as your input file. The output to the screen will look something like this:

```
             BEAST v2.2.0, 2002-2014
       Bayesian Evolutionary Analysis Sampling Trees
                 Designed and developed by
Remco Bouckaert, Alexei J. Drummond, Andrew Rambaut and Marc A. Suchard

               Department of Computer Science
                   University of Auckland
                  remco@cs.auckland.ac.nz
                  alexei@cs.auckland.ac.nz

             Institute of Evolutionary Biology
```

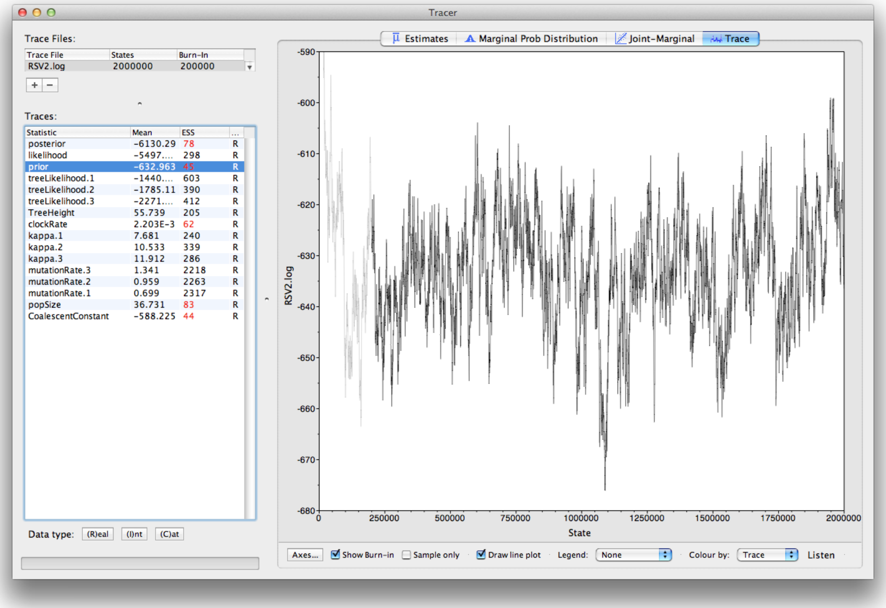
```
                  University of Edinburgh
                     a.rambaut@ed.ac.uk

              David Geffen School of Medicine
           University of California, Los Angeles
                     msuchard@ucla.edu

                Downloads, Help & Resources:
                     http://beast2.org/

Source code distributed under the GNU Lesser General Public License:
              http://github.com/CompEvol/beast2

                     BEAST developers:
Alex Alekseyenko, Trevor Bedford, Erik Bloomquist, Joseph Heled,
Sebastian Hoehna, Denise Kuehnert, Philippe Lemey, Wai Lok Sibon Li,
Gerton Lunter, Sidney Markowitz, Vladimir Minin, Michael Defoin Platel,
           Oliver Pybus, Chieh-Hsi Wu, Walter Xie

                         Thanks to:
     Roald Forsberg, Beth Shapiro and Korbinian Strimmer

... ...

        1999000     -6131.9540        80.2        -5496.4810      -635.4730 2m3s/Msamples
        2000000     -6142.7389        80.7        -5506.2818      -636.4570 2m3s/Msamples

Operator                                                Tuning    #accept    #reject      Pr(m)  Pr(acc|m)
ScaleOperator(treeScaler.t:tree)                        0.7224        546      71568     0.0360     0.0076 Try setting scaleFactor to about 0.85
ScaleOperator(treeRootScaler.t:tree)                    0.6019       5076      66818     0.0360     0.0706 Try setting scaleFactor to about 0.776
Uniform(UniformOperator.t:tree)                              -     387274     334094     0.3601     0.5369
SubtreeSlide(SubtreeSlide.t:tree)                       3.9556      57963     302626     0.1801     0.1607
Exchange(narrow.t:tree)                                      -      88179     271422     0.1801     0.2452
Exchange(wide.t:tree)                                        -        153      71541     0.0360     0.0021
WilsonBalding(WilsonBalding.t:tree)                          -        365      71560     0.0360     0.0051
ScaleOperator(StrictClockRateScaler.c:clock)            0.7869      17571      54573     0.0360     0.2436
UpDownOperator(strictClockUpDownOperator.c:clock)       0.7594        673      70876     0.0360     0.0094 Try setting scaleFactor to about 0.871
ScaleOperator(KappaScaler.s:RSV2_1)                     0.4069        581       1743     0.0012     0.2500
DeltaExchangeOperator(FixMeanMutationRatesOperator)     0.3833       8247      39773     0.0240     0.1717
ScaleOperator(KappaScaler.s:RSV2_2)                     0.4276        583       1790     0.0012     0.2457
ScaleOperator(KappaScaler.s:RSV2_3)                     0.4421        589       1774     0.0012     0.2493
ScaleOperator(PopSizeScaler.t:tree)                     0.5907      18802      53241     0.0360     0.2610

     Tuning: The value of the operator's tuning parameter, or '-' if the operator can't be optimized.
    #accept: The total number of times a proposal by this operator has been accepted.
    #reject: The total number of timEnd likelihood: -6142.738945357851
es a proposal by this operator has been rejected.
      Pr(m): The probability this operator is chosen in a step of the MCMC (i.e. the normalized weight).
  Pr(acc|m): The acceptance probability (#accept as a fraction of the total proposals for this operator).


Total calculation time: 249.968 seconds
```

Analysing the BEAST output

Note that the effective sample sizes (ESSs) for many of the logged quantities are small (ESSs less than 100 will be highlighted in red by Tracer). This is not good. A low ESS means that the trace contains a lot of correlated samples and thus may not represent the posterior distribution well. In the bottom right of the window is a frequency plot of the samples which is expected given the low ESSs is extremely rough.

If we select the tab on the right-hand-side labelled 'Trace' we can view the raw trace, that is, the sampled values against the step in the MCMC chain.

Here you can see how the samples are correlated. There are 5000 samples in the trace (we ran the MCMC for 2,000,000 steps sampling every 400) but adjacent samples often tend to have similar values. The ESS for the absolute rate of evolution (clockRate) is about 62 so we are only getting 1 independent sample to every 81 = 5000/62 actual samples). With a short run such as this one, it may also be the case that the default burn-in of 10% of the chain length is inadequate. Not excluding enough of the start of the chain as burn-in will render estimates of ESS unreliable.


Tracer
Estimates
Marginal Prob Distribution
Joint-Marginal
Trace
RSV2.log
-590
-600
-610
-620
-630
-640
-650
-660
-670
-680
0
250000
500000
750000
1000000
1250000
1500000
1750000
2000000
State
Axes...
Show Burn-in
Sample only
Draw line plot
Legend:
None
Colour by:
Trace
Listen
Data type:
(R)eal
(I)nt
(C)at

Trace Files:

Trace File	States	Burn-In
RSV2.log	2000000	200000

Traces:

Statistic	Mean	ESS	...
posterior	-6130.29	78	R
likelihood	-5497....	298	R
prior	-632.963	45	R
treeLikelihood.1	-1440....	603	R
treeLikelihood.2	-1785.11	390	R
treeLikelihood.3	-2271....	412	R
TreeHeight	55.739	205	R
clockRate	2.203E-3	62	R
kappa.1	7.681	240	R
kappa.2	10.533	339	R
kappa.3	11.912	286	R
mutationRate.3	1.341	2218	R
mutationRate.2	0.959	2263	R
mutationRate.1	0.699	2317	R
popSize	36.731	83	R
CoalescentConstant	-588.225	44	R

Figure 10: A screenshot of Tracer.

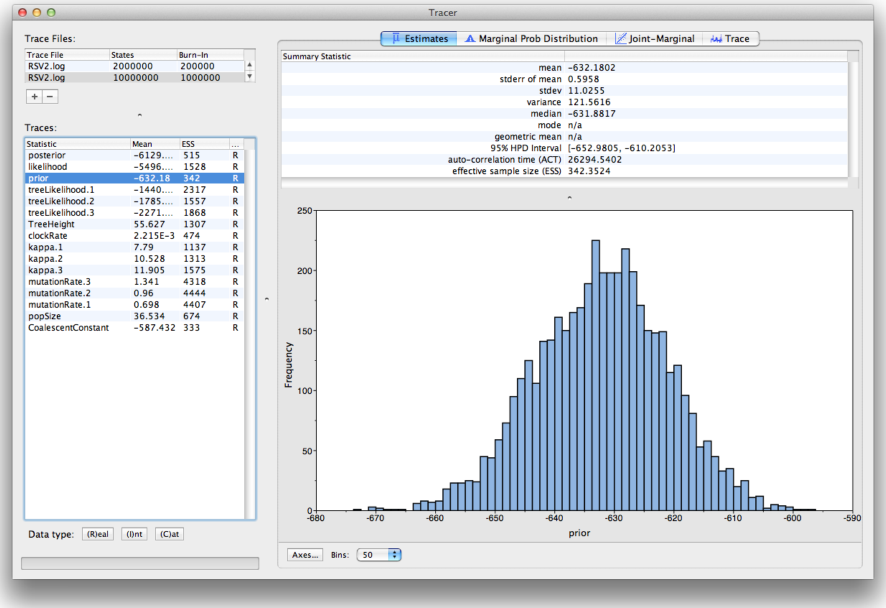
The simple response to this situation is that we need to run the chain for longer. Given the lowest ESS (for the constant coalescent) is 44, it would suggest that we have to run the chain for at least 5 times the length to get reasonable ESSs that are >200. So let's go for a chain length of 10,000,000 and log every 2,000. Go back to the **MCMC** options section in BEAUti, and create a new BEAST XML file with a longer chain length. Now run BEAST and load the new log file into Tracer (you can leave the old one loaded for comparison).

Click on the Trace tab and look at the raw trace plot.

Again we have chosen options that produce 5000 samples and with an ESS of about 342 there is still auto-correlation between the samples but >340 effectively independent samples will now provide a very good estimate of the posterior distribution. There are no obvious trends in the plot which would suggest that the MCMC has not yet converged, and there are no significant long range fluctuations in the trace which would suggest poor mixing.

As we are satisfied with the mixing we can now move on to one of the parameters of interest: substitution rate. Select `clockRate` in the left-hand table. This is the average substitution rate across all sites in the alignment. Now choose the density plot by selecting the tab labeled *Marginal Density*. This shows a plot of the marginal posterior probability density of this parameter. You should see a plot similar to this:

As you can see the posterior probability density is roughly bell-shaped. There is some sampling noise which would be reduced if we ran the chain for longer or sampled more often but we already have a good estimate of the mean and HPD interval. You

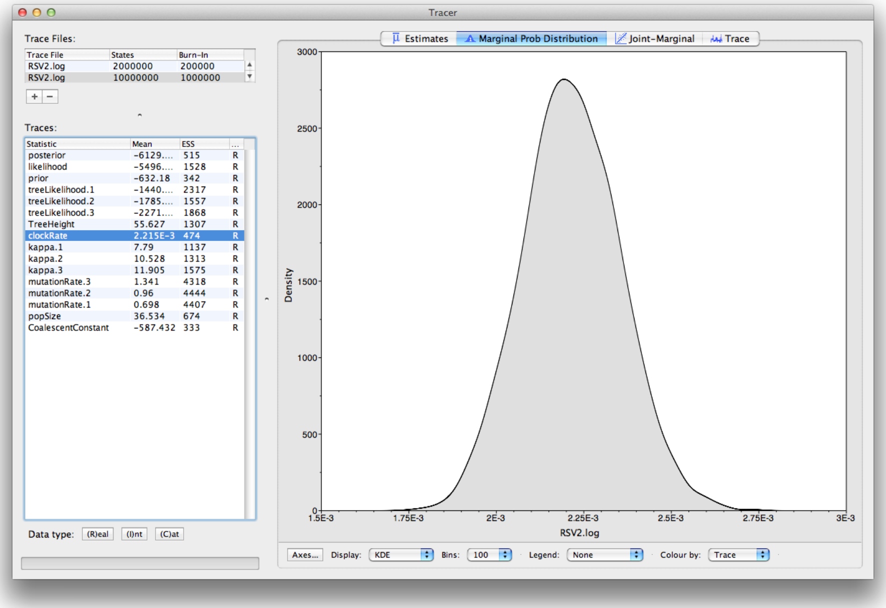

Tracer
Trace Files:
+
−
Traces:
Estimates
Marginal Prob Distribution
Joint-Marginal
Trace
Frequency
250
200
150
100
50
0
-680
-670
-660
-650
-640
-630
-620
-610
-600
-590
prior
Data type:
(R)eal
(I)nt
(C)at
Axes...

Trace File	States	Burn-In
RSV2.log	2000000	200000
RSV2.log	10000000	1000000

Statistic	Mean	ESS	...
posterior	-6129....	515	R
likelihood	-5496....	1528	R
prior	-632.18	342	R
treeLikelihood.1	-1440....	2317	R
treeLikelihood.2	-1785....	1557	R
treeLikelihood.3	-2271....	1868	R
TreeHeight	55.627	1307	R
clockRate	2.215E-3	474	R
kappa.1	7.79	1137	R
kappa.2	10.528	1313	R
kappa.3	11.905	1575	R
mutationRate.3	1.341	4318	R
mutationRate.2	0.96	4444	R
mutationRate.1	0.698	4407	R
popSize	36.534	674	R
CoalescentConstant	-587.432	333	R

Summary Statistic	
mean	-632.1802
stderr of mean	0.5958
stdev	11.0255
variance	121.5616
median	-631.8817
mode	n/a
geometric mean	n/a
95% HPD Interval	[-652.9805, -610.2053]
auto-correlation time (ACT)	26294.5402
effective sample size (ESS)	342.3524

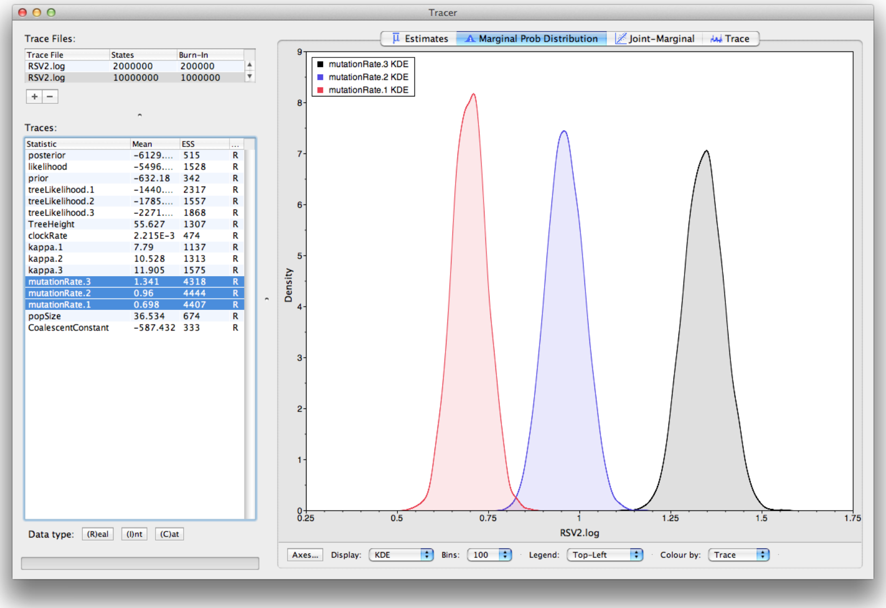
Bins: 50


Tracer
Trace Files:
+
−
Traces:
Estimates
Marginal Prob Distribution
Joint-Marginal
Trace
Density
3000
2500
2000
1500
1000
500
0
1.5E-3
1.75E-3
2E-3
2.25E-3
2.5E-3
2.75E-3
3E-3
RSV2.log
Data type:
(R)eal
(I)nt
(C)at
Axes...

Trace File	States	Burn-In
RSV2.log	2000000	200000
RSV2.log	10000000	1000000

Statistic	Mean	ESS	...
posterior	-6129....	515	R
likelihood	-5496....	1528	R
prior	-632.18	342	R
treeLikelihood.1	-1440....	2317	R
treeLikelihood.2	-1785....	1557	R
treeLikelihood.3	-2271....	1868	R
TreeHeight	55.627	1307	R
clockRate	2.215E-3	474	R
kappa.1	7.79	1137	R
kappa.2	10.528	1313	R
kappa.3	11.905	1575	R
mutationRate.3	1.341	4318	R
mutationRate.2	0.96	4444	R
mutationRate.1	0.698	4407	R
popSize	36.534	674	R
CoalescentConstant	-587.432	333	R

Display: KDE
Bins: 100
Legend: None
Colour by: Trace


Tracer
Trace Files:
+
−
Traces:
Estimates
Marginal Prob Distribution
Joint-Marginal
Trace
mutationRate.3 KDE
mutationRate.2 KDE
mutationRate.1 KDE
Density
RSV2.log
0
1
2
3
4
5
6
7
8
9
0.25
0.5
0.75
1
1.25
1.5
1.75
Data type:
(R)eal
(I)nt
(C)at
Axes...
Display: KDE
Bins: 100
Legend: Top-Left
Colour by: Trace

Trace File	States	Burn-In
RSV2.log	2000000	200000
RSV2.log	10000000	1000000

Statistic	Mean	ESS	...
posterior	-6129....	515	R
likelihood	-5496....	1528	R
prior	-632.18	342	R
treeLikelihood.1	-1440....	2317	R
treeLikelihood.2	-1785....	1557	R
treeLikelihood.3	-2271....	1868	R
TreeHeight	55.627	1307	R
clockRate	2.215E-3	474	R
kappa.1	7.79	1137	R
kappa.2	10.528	1313	R
kappa.3	11.905	1575	R
mutationRate.3	1.341	4318	R
mutationRate.2	0.96	4444	R
mutationRate.1	0.698	4407	R
popSize	36.534	674	R
CoalescentConstant	-587.432	333	R

Figure 11: The posterior probability densities for the relative substitution rates

can overlay the density plots of multiple traces in order to compare them (it is up to the user to determine whether they are comparable on the the same axis or not). Select the relative substitution rates for all three codon positions in the table to the left (labelled `mutationRate.1`, `mutationRate.2` and `mutationRate.3`). You will now see the posterior probability densities for the relative substitution rate at all three codon positions overlaid:

Summarizing the trees

Use the program TreeAnnotator to summarize the tree and view the results in Figtree (Figure 12).

DensiTree with clade height bars for clades with over 50% support. Root canal tree represents maximum clade credibility tree.

Questions

In what year did the common ancestor of all RSVA viruses sampled live? What is the 95% HPD?

Bonus section: Bayesian Skyline plot

We can reconstruct the population history using the Bayesian Skyline plot. In order to do so, load the XML file into BEAUti, select the priors-tab and change the tree prior

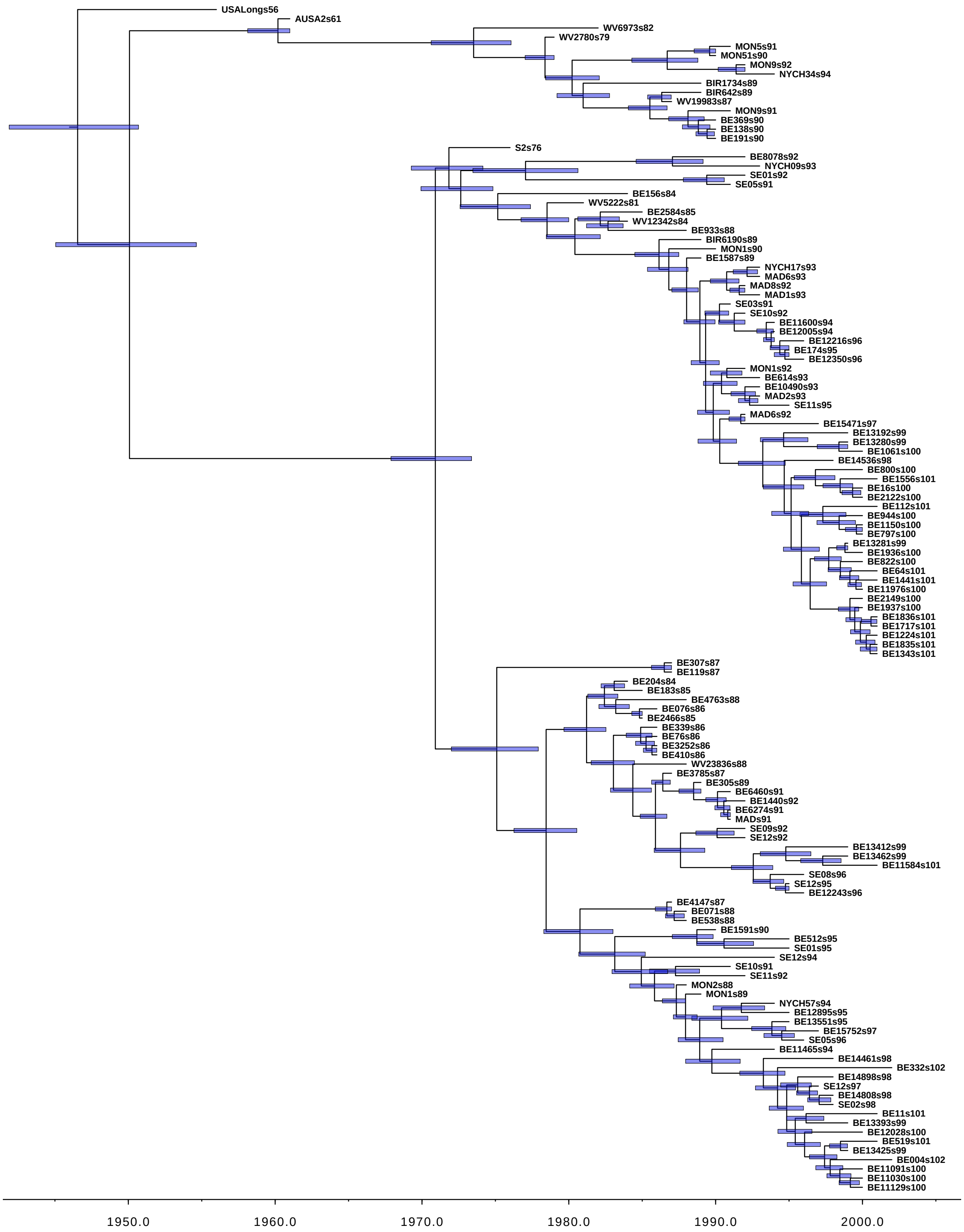
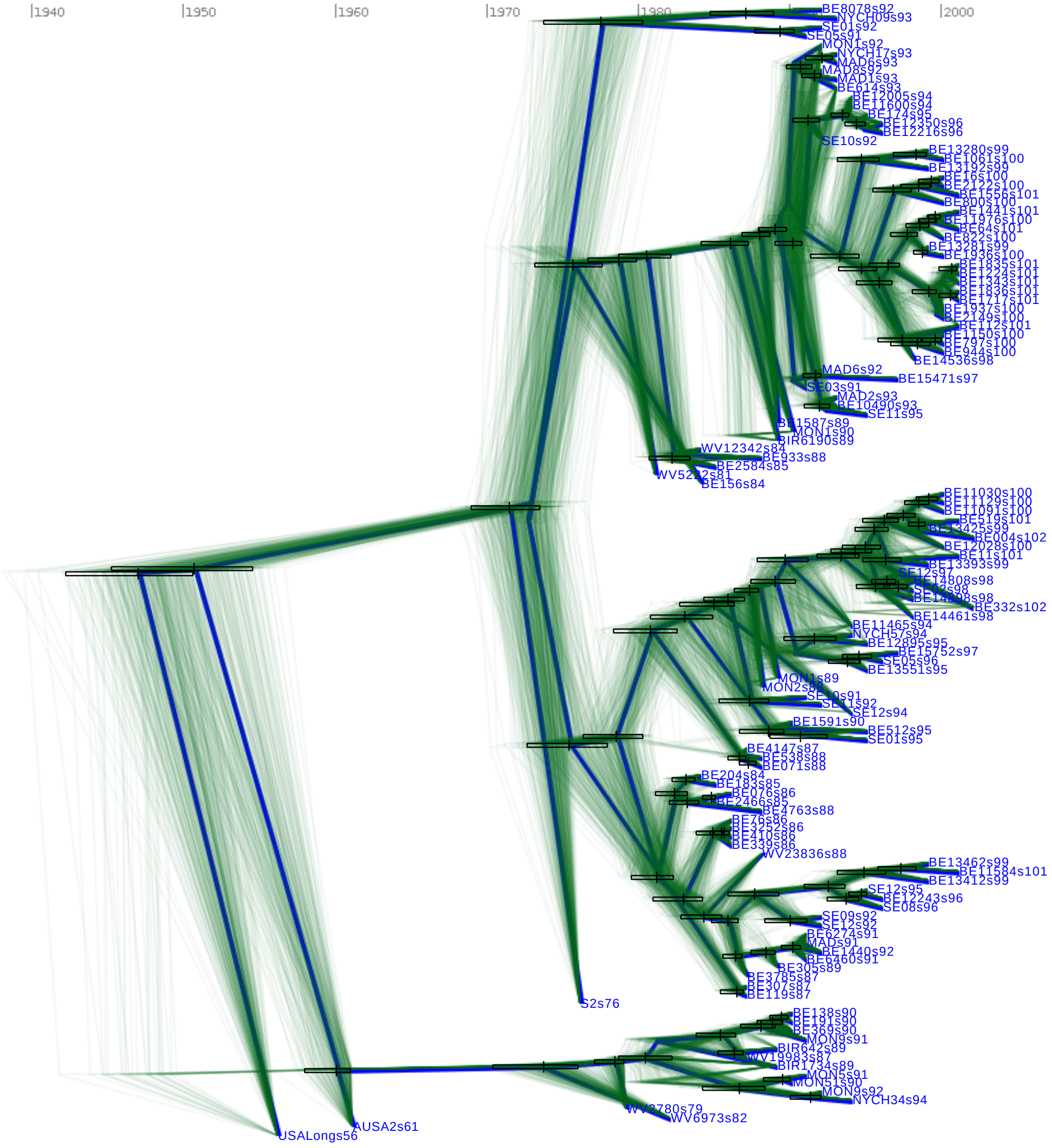
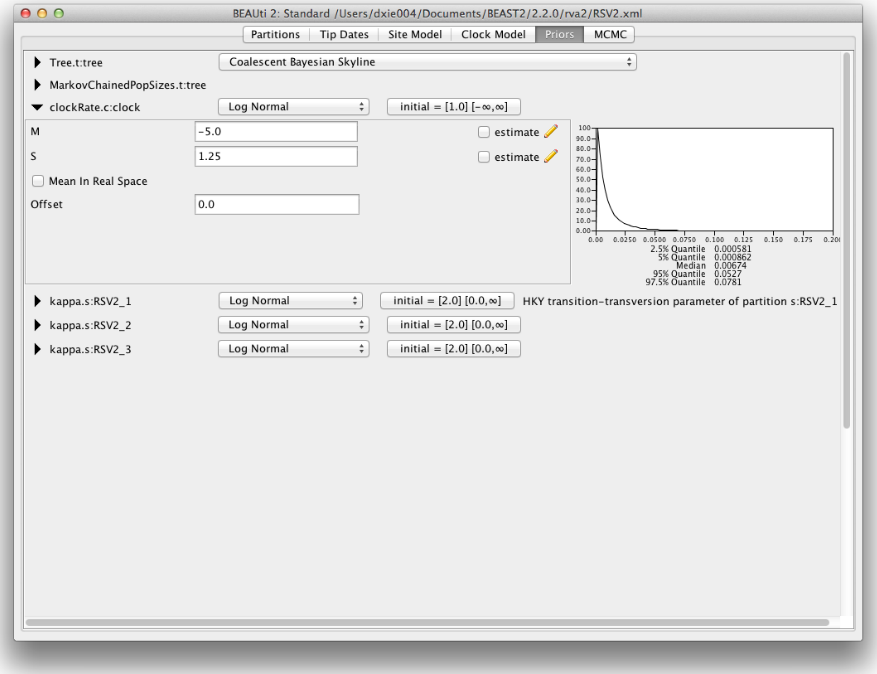

USALongs56
AUSA2s61
WV6973s82
WV2780s79
MON5s91
MON51s90
MON9s92
NYCH34s94
BIR1734s89
BIR642s89
WV19983s87
MON9s91
BE369s90
BE138s90
BE191s90
S2s76
BE8078s92
NYCH09s93
SE01s92
SE05s91
BE156s84
WV5222s81
BE2584s85
WV12342s84
BE933s88
BIR6190s89
MON1s90
BE1587s89
NYCH17s93
MAD6s93
MAD8s92
MAD1s93
SE03s91
SE10s92
BE11600s94
BE12005s94
BE12216s96
BE174s95
BE12350s96
MON1s92
BE614s93
BE10490s93
MAD2s93
SE11s95
MAD6s92
BE15471s97
BE13192s99
BE13280s99
BE1061s100
BE14536s98
BE800s100
BE1556s101
BE16s100
BE2122s100
BE112s101
BE944s100
BE1150s100
BE797s100
BE13281s99
BE1936s100
BE822s100
BE64s101
BE1441s101
BE11976s100
BE2149s100
BE1937s100
BE1836s101
BE1717s101
BE1224s101
BE1835s101
BE1343s101
BE307s87
BE119s87
BE204s84
BE183s85
BE4763s88
BE076s86
BE2466s85
BE339s86
BE76s86
BE3252s86
BE410s86
WV23836s88
BE3785s87
BE305s89
BE6460s91
BE1440s92
BE6274s91
MADs91
SE09s92
SE12s92
BE13412s99
BE13462s99
BE11584s101
SE08s96
SE12s95
BE12243s96
BE4147s87
BE071s88
BE538s88
BE1591s90
BE512s95
SE01s95
SE12s94
SE10s91
SE11s92
MON2s88
MON1s89
NYCH57s94
BE12895s95
BE13551s95
BE15752s97
SE05s96
BE11465s94
BE14461s98
BE332s102
BE14898s98
SE12s97
BE14808s98
SE02s98
BE11s101
BE13393s99
BE12028s100
BE519s101
BE13425s99
BE004s102
BE11091s100
BE11030s100
BE11129s100
1950.0
1960.0
1970.0
1980.0
1990.0
2000.0

Figure 12: The Maximum clade credibility tree for the G gene of 129 RSVA-2 viral samples.


1940
1950
1960
1970
1980
2000
BE8078s92
NYCH09s93
SE01s92
SE05s91
MON1s92
NYCH17s93
MAD6s93
MAD8s92
MAD1s93
BE614s93
BE12005s94
BE11600s94
BE174s95
BE12350s96
BE12216s96
SE10s92
BE13280s99
BE1061s100
BE13192s99
BE16s100
BE2122s100
BE1556s101
BE800s100
BE1441s101
BE11976s100
BE64s101
BE822s100
BE13281s99
BE1936s100
BE1835s101
BE1224s101
BE1343s101
BE1836s101
BE1717s101
BE1937s100
BE2149s100
BE112s101
BE1150s100
BE797s100
BE944s100
BE14536s98
MAD6s92
BE15471s97
SE03s91
MAD2s93
BE10490s93
SE11s95
BE1587s89
MON1s90
BIR6190s89
WV12342s84
BE933s88
BE2584s85
WV5222s81
BE156s84
BE11030s100
BE11129s100
BE11091s100
BE519s101
BE13425s99
BE004s102
BE12028s100
BE11s101
BE13393s99
SE12s97
BE14808s98
SE02s98
BE14898s98
BE332s102
BE14461s98
BE11465s94
NYCH57s94
BE12895s95
BE15752s97
SE05s96
BE13551s95
MON1s89
MON2s88
SE10s91
SE11s92
SE12s94
BE1591s90
BE512s95
SE01s95
BE4147s87
BE538s88
BE071s88
BE204s84
BE183s85
BE076s86
BE2466s85
BE4763s88
BE76s86
BE3252s86
BE410s86
BE339s86
WV23836s88
BE13462s99
BE11584s101
BE13412s99
SE12s95
BE12243s96
SE08s96
SE09s92
SE12s92
BE6274s91
MADs91
BE1440s92
BE6460s91
BE305s89
BE3785s87
BE307s87
BE119s87
S2s76
BE138s90
BE191s90
BE369s90
MON9s91
BIR642s89
WV19983s87
BIR1734s89
MON5s91
MON51s90
MON9s92
NYCH34s94
WV2780s79
WV6973s82
USALongs56
AUSA2s61



Tree.t:tree Coalescent Bayesian Skyline

MarkovChainedPopSizes.t:tree

clockRate.c:clock Log Normal initial = [1.0] [−∞,∞]

M −5.0 ☐ estimate

S 1.25 ☐ estimate

☐ Mean In Real Space

Offset 0.0

2.5% Quantile 0.000581
5% Quantile 0.000862
Median 0.00674
95% Quantile 0.0527
97.5% Quantile 0.0781

kappa.s:RSV2_1 Log Normal initial = [2.0] [0.0,∞] HKY transition−transversion parameter of partition s:RSV2_1

kappa.s:RSV2_2 Log Normal initial = [2.0] [0.0,∞]

kappa.s:RSV2_3 Log Normal initial = [2.0] [0.0,∞]

Figure 13: Priors

from coalescent with constant population size to coalescent with Bayesian skyline. Note that an extra item is added to the priors called 'Markov chained population sizes' which is a prior that ensures dependence between population sizes.

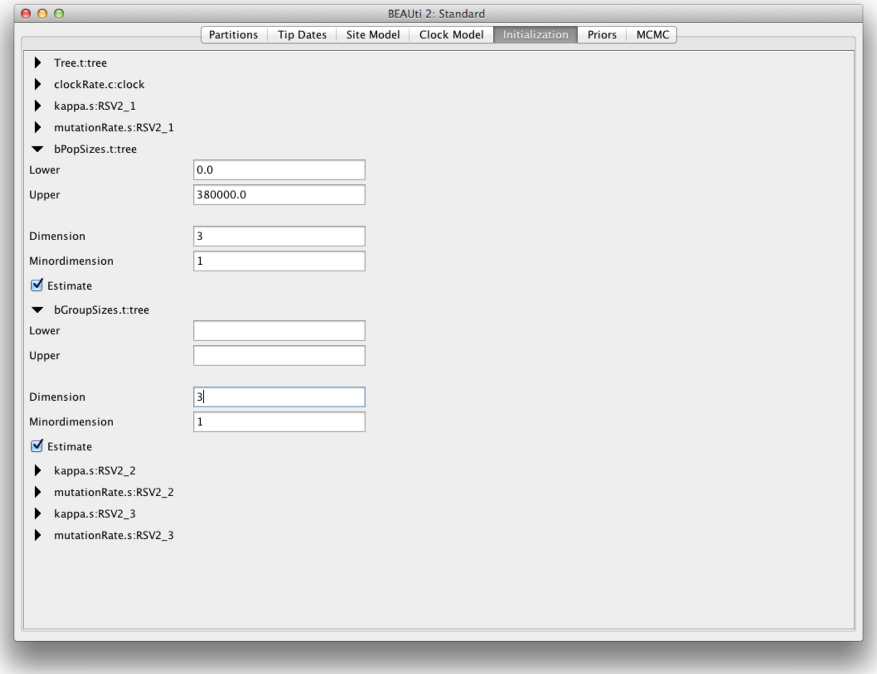
By default the number of groups used in the skyline analysis is set to 5, To change this, select menu View/Show Initialization panel and a list of parameters is shown. Select `bPopSizes.t:tree` and change the dimension to 3. Likewise, selection `bGroupSizes.t:tree` and change its dimension to 3. The dimensions of the two parameters should be the same. More groups mean more population changes can be detected, but it also means more parameters need to be estimated and the chain runs longer. The extended Bayesian skyline plot automatically detects the number of changes, so it could be used as an alternative tree prior.

This analysis requires a bit longer to converge, so change the MCMC chain length to 10 million, and the log intervals for the trace-log and tree-log to 10 thousand. Then, save the file and run BEAST.

To plot the population history, load the log file in tracer and select the menu Analysis/Bayesian Skyline Reconstruction.

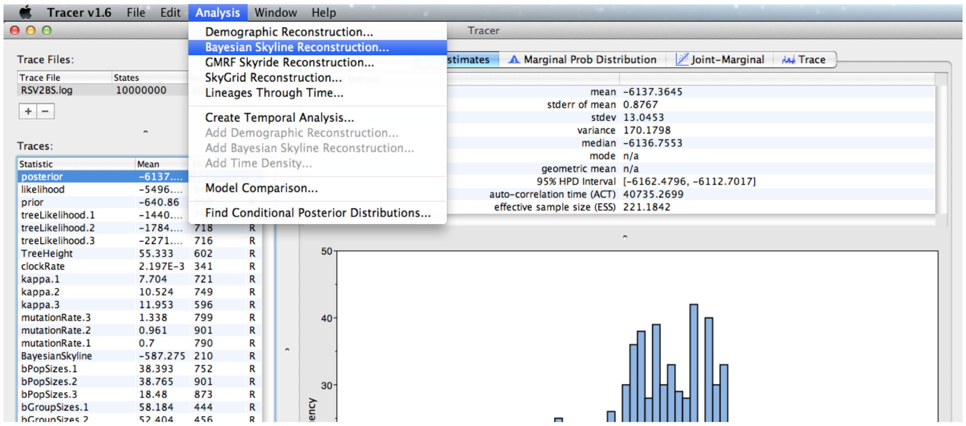
A dialog is shown where you can specify the tree file associated with the log file. Also, since the youngest sample is from 2002, change the entry for age of youngest tip to 2002.

After some calculation, a graph appears showing population history where the median and 95% HPD intervals are plotted. After selecting the 'solid interval' checkbox,


BEAUti 2: Standard
Partitions
Tip Dates
Site Model
Clock Model
Initialization
Priors
MCMC

Tree.t:tree
clockRate.c:clock
kappa.s:RSV2_1
mutationRate.s:RSV2_1
bPopSizes.t:tree
Lower 0.0
Upper 380000.0
Dimension 3
Minordimension 1
☑ Estimate
bGroupSizes.t:tree
Lower
Upper
Dimension 3
Minordimension 1
☑ Estimate
kappa.s:RSV2_2
mutationRate.s:RSV2_2
kappa.s:RSV2_3
mutationRate.s:RSV2_3

Figure 14: Initialization panel

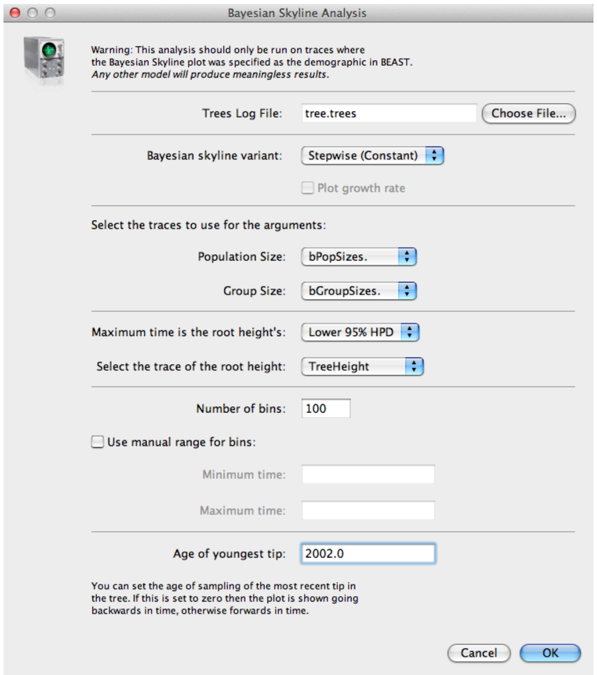

Tracer v1.6
File
Edit
Analysis
Window
Help
Demographic Reconstruction...
Bayesian Skyline Reconstruction...
GMRF Skyride Reconstruction...
SkyGrid Reconstruction...
Lineages Through Time...
Create Temporal Analysis...
Add Demographic Reconstruction...
Add Bayesian Skyline Reconstruction...
Add Time Density...
Model Comparison...
Find Conditional Posterior Distributions...
Tracer
stimates
Marginal Prob Distribution
Joint-Marginal
Trace
Trace Files:
+
−
Traces:
50
40
30
ency

Trace File	States
RSV2BS.log	10000000

Statistic	Mean		
posterior	-6137....		
likelihood	-5496....		
prior	-640.86		
treeLikelihood.1	-1440....		
treeLikelihood.2	-1784....	718	R
treeLikelihood.3	-2271....	716	R
TreeHeight	55.333	602	R
clockRate	2.197E-3	341	R
kappa.1	7.704	721	R
kappa.2	10.524	749	R
kappa.3	11.953	596	R
mutationRate.3	1.338	799	R
mutationRate.2	0.961	901	R
mutationRate.1	0.7	790	R
BayesianSkyline	-587.275	210	R
bPopSizes.1	38.393	752	R
bPopSizes.2	38.765	901	R
bPopSizes.3	18.48	873	R
bGroupSizes.1	58.184	444	R
bGroupSizes.2	52.404	456	R

mean	-6137.3645
stderr of mean	0.8767
stdev	13.0453
variance	170.1798
median	-6136.7553
mode	n/a
geometric mean	n/a
95% HPD Interval	[-6162.4796, -6112.7017]
auto-correlation time (ACT)	40735.2699
effective sample size (ESS)	221.1842

Figure 15: Bayesian Skyline Reconstruction



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: tree.trees

Bayesian skyline variant: Stepwise (Constant)

☐ Plot growth rate

Select the traces to use for the arguments:

Population Size: bPopSizes.

Group Size: bGroupSizes.

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

Select the trace of the root height: TreeHeight

Number of bins: 100

☐ Use manual range for bins:

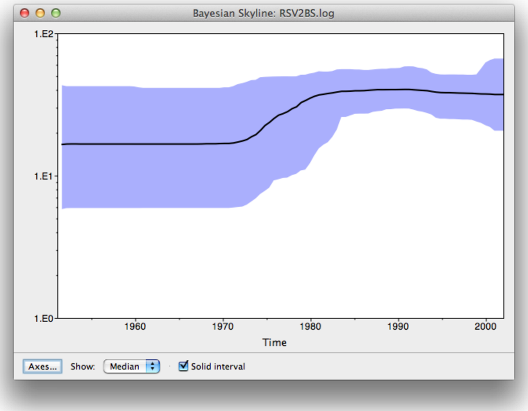
Minimum time:

Maximum time:

Age of youngest tip: 2002.0

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.

the graph should look something like this.



Questions

1. *By what amount did the effective population size of RSVA grow from 1970 to 2002 according to the BSP?*
2. *What are the underlying assumptions of the BSP? Are the violated by this data set?*

1.2 Exercise

Change the Bayesian skyline prior to extended Bayesian skyline plot (EBSP) prior and run till convergence. EBSP produces an extra log file, called EBSP.$(seed).log where $(seed) is replaced by the seed you used to run BEAST. A plot can be created by running the EBSPAnalyser utility, and loading the output file in a spreadsheet.

How many groups are indicated by the EBSP analysis? This is much lower than for BSP. How does this affect the population history plots?

References

[1] Kalina T Zlateva, Philippe Lemey, Elien Moës, Anne-Mieke Vandamme, and Marc Van Ranst, *Genetic variability and molecular evolution of the human respiratory syncytial virus subgroup b attachment g protein*, J Virol **79** (2005), no. 14, 9157–67.

[2] Kalina T Zlateva, Philippe Lemey, Anne-Mieke Vandamme, and Marc Van Ranst, *Molecular evolution and circulation patterns of human respiratory syncytial virus subgroup a: positively selected sites in the attachment g glycoprotein*, J Virol **78** (2004), no. 9, 4675–83.