BEAST: Bayesian Evolutionary Analysis by Sampling Trees

A STEP-BY-STEP TUTORIAL FOR PHYLOGEOGRAPHIC INFERENCE IN CONTINUOUS SPACE

This step-by-step tutorial guides you through a continuous space phylogeographic analysis for reconstructing the spatial dynamics of a large-scale rabies virus outbreak among North American raccoons. The data set consists of 47 N gene sequences of raccoon rabies virus (RABV), isolated at different time points (1982-2004) throughout the USA [Biek *et al*., 2007]. Here we describe a full Bayesian framework for phylogeographic reconstruction developed in [Lemey *et al*., 2009].

To undertake this tutorial, you will need to download the following software packages in a format compatible with your operating system (available for Linux/UNIX, Mac OS X and Windows):

- **BEAST** - this package contains the BEAST program, BEAUti and a couple of utility programs. At the time of writing, the current version is v1.8.0. It is available for download from http://beast.bio.ed.ac.uk
- **Tracer** - this program is used to explore the output of **BEAST** (and other Bayesian MCMC programs). It graphically and quantitatively summarizes the empirical 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://tree.bio.ed.ac.uk/software/tracer/
- **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/software/figtree/
- **SPREAD** - this is an application for the visualizing and quantifying support for phylogeographic analyses performed with **BEAST**. At the time of writing, the current version is v1.0.6. It is available for download from http://www.kuleuven.be/aidslab/phylogeography/home.html
- **Google Earth** - virtual globe software that can be used to visualize the KML output from SPREAD in an interactive fashion. It is available for download from http://www.google.com/earth/
- **BEAGLE library** - high-performance *library* that can perform the core calculations at the heart of most Bayesian and Maximum Likelihood phylogenetics packages. It is available for download from https://code.google.com/p/beagle-lib/

Suggested text editors for editing xml files:

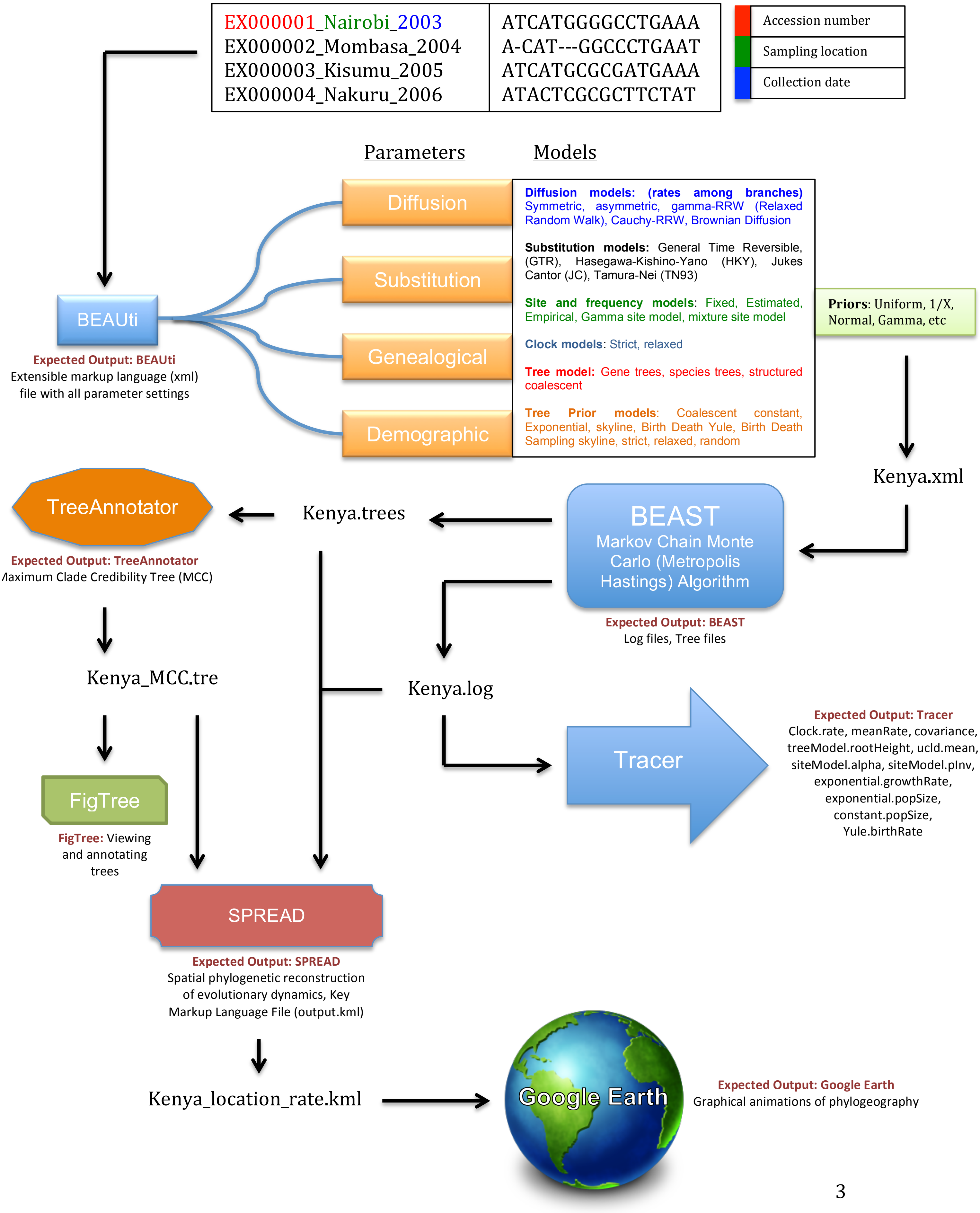
- **Geany** (Linux/UNIX):
  http://www.geany.org/Download/Releases

- **TextWrangler** (Mac OS X):
  http://www.barebones.com/products/textwrangler/download.html

- **Notepad++** (Windows):
  http://notepad-plus-plus.org/download/v6.6.8.html

ANALYSIS PIPELINE

Input: FASTA/ NEXUS file (e.g. Kenya.fasta)



EX000001_Nairobi_2003	ATCATGGGGCCTGAAA
EX000002_Mombasa_2004	A-CAT---GGCCCTGAAT
EX000003_Kisumu_2005	ATCATGCGCGATGAAA
EX000004_Nakuru_2006	ATACTCGCGCTTCTAT

🟥	Accession number
🟩	Sampling location
🟦	Collection date

Parameters	Models
Diffusion	**Diffusion models: (rates among branches)** Symmetric, asymmetric, gamma-RRW (Relaxed Random Walk), Cauchy-RRW, Brownian Diffusion
Substitution	**Substitution models:** General Time Reversible, (GTR), Hasegawa-Kishino-Yano (HKY), Jukes Cantor (JC), Tamura-Nei (TN93)
Genealogical	**Site and frequency models**: Fixed, Estimated, Empirical, Gamma site model, mixture site model **Clock models**: Strict, relaxed **Tree model:** Gene trees, species trees, structured coalescent
Demographic	**Tree Prior models**: Coalescent constant, Exponential, skyline, Birth Death Yule, Birth Death Sampling skyline, strict, relaxed, random

Priors: Uniform, 1/X, Normal, Gamma, etc

Expected Output: BEAUti
Extensible markup language (xml) file with all parameter settings

Expected Output: BEAST
Log files, Tree files

Expected Output: TreeAnnotator
Maximum Clade Credibility Tree (MCC)

FigTree: Viewing and annotating trees

Expected Output: Tracer
Clock.rate, meanRate, covariance, treeModel.rootHeight, ucld.mean, siteModel.alpha, siteModel.pInv, exponential.growthRate, exponential.popSize, constant.popSize, Yule.birthRate

Expected Output: SPREAD
Spatial phylogenetic reconstruction of evolutionary dynamics, Key Markup Language File (output.kml)

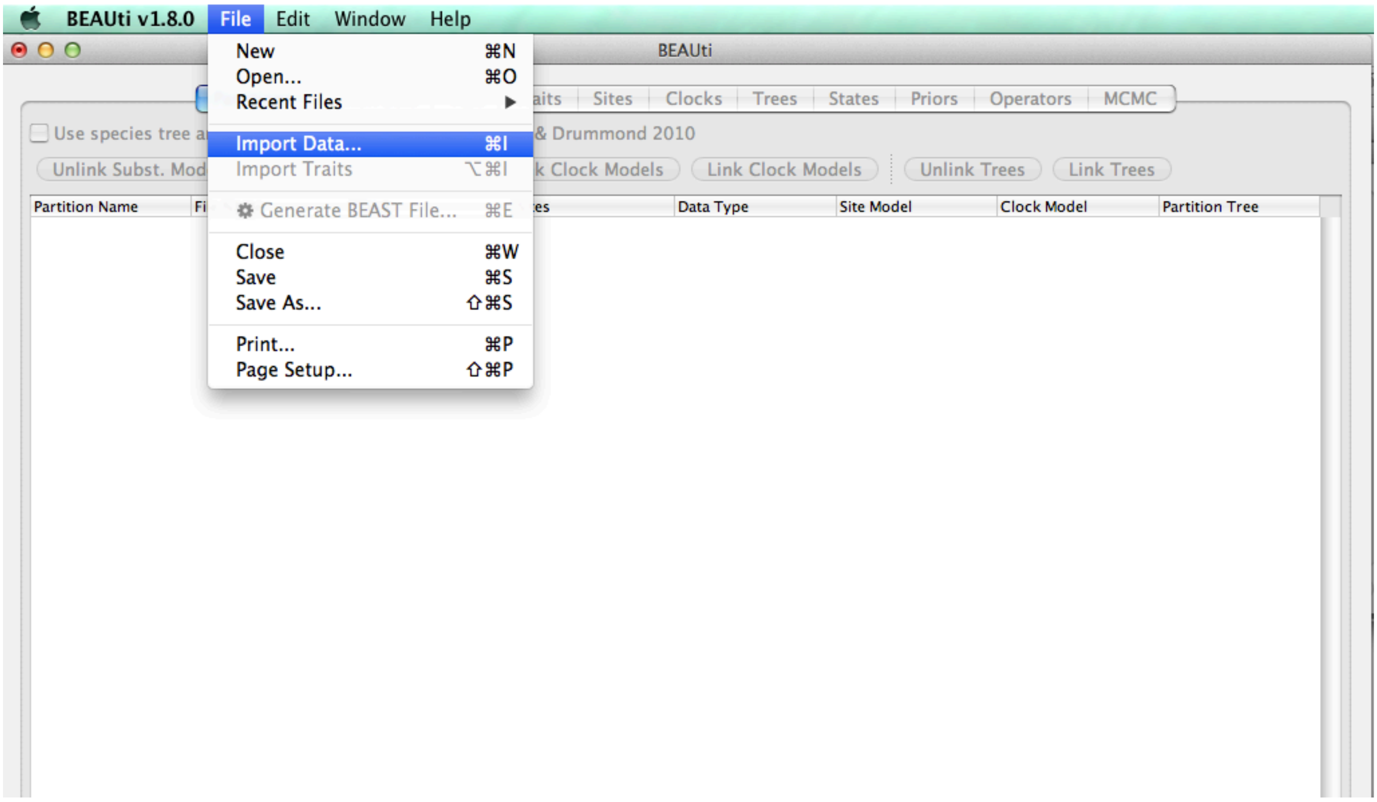
Expected Output: Google Earth
Graphical animations of phylogeography

RUNNING BEAUti

BEAUti is a user-friendly utility for setting model parameters for Bayesian analysis in BEAST.

1. **Partitions Tab**:

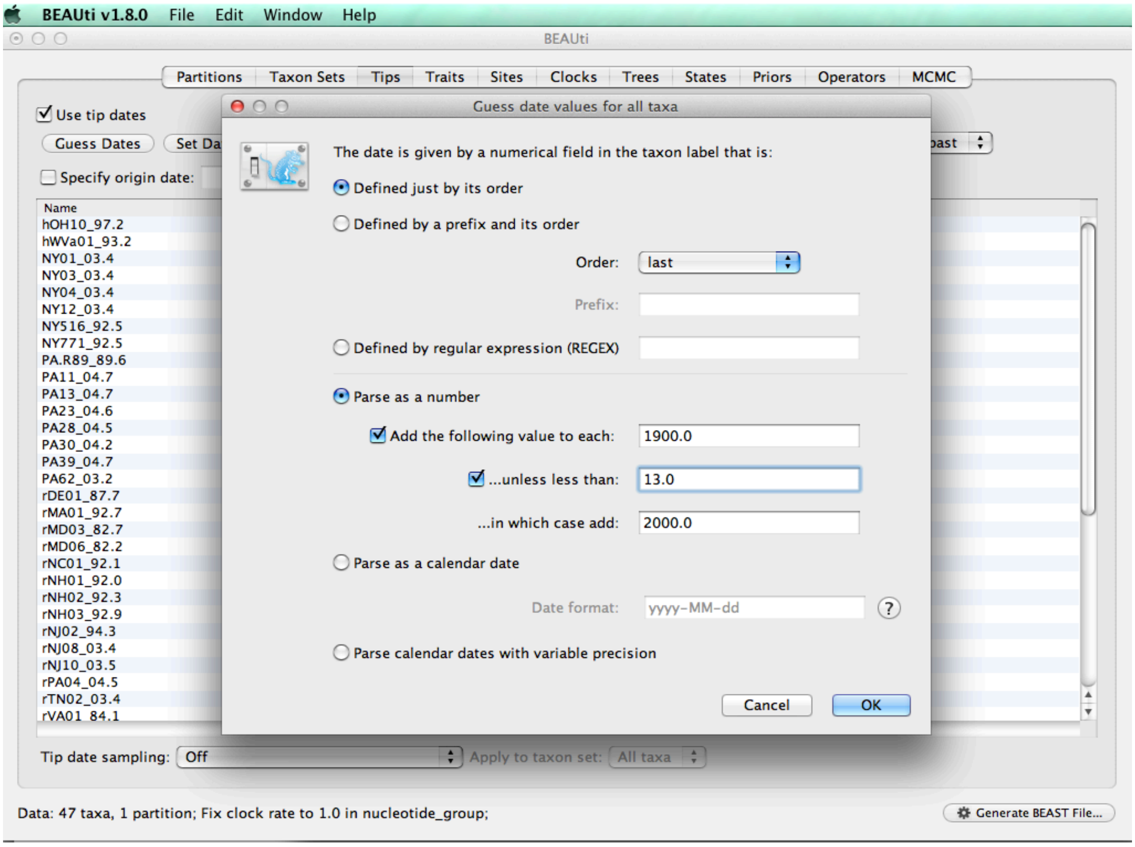
   - Load the **RacRABV.fasta** file by selecting **Import Data** from the **File** menu. Once loaded, the sequence data will be listed under **Partition Name**.



2. **Tips Tab:**

   - Select ‘**Use tip dates’** option. Click on ‘**Guess Dates**’**. I**n the dialogue box that appears keep the default ‘**Defined just by its order**’ and set the ‘**Order’**: to **last**.

   - Go to ‘**Parse as a number’** and select ‘**Add the following value to each:’.** Keep the default **1900.0.** Select ‘**unless less than:’.** Keep the default **15.0** and ‘…**in which case add:’ 2000.0**.


BEAUti v1.8.0 File Edit Window Help
BEAUti
Partitions Taxon Sets Tips Traits Sites Clocks Trees States Priors Operators MCMC
Guess Dates
Set Da
Cancel
OK
Generate BEAST File...

☑ Use tip dates

☐ Specify origin date:

Name
hOH10_97.2
hWVa01_93.2
NY01_03.4
NY03_03.4
NY04_03.4
NY12_03.4
NY516_92.5
NY771_92.5
PA.R89_89.6
PA11_04.7
PA13_04.7
PA23_04.6
PA28_04.5
PA30_04.2
PA39_04.7
PA62_03.2
rDE01_87.7
rMA01_92.7
rMD03_82.7
rMD06_82.2
rNC01_92.1
rNH01_92.0
rNH02_92.3
rNH03_92.9
rNJ02_94.3
rNJ08_03.4
rNJ10_03.5
rPA04_04.5
rTN02_03.4
rVA01_84.1

Guess date values for all taxa

The date is given by a numerical field in the taxon label that is:

● Defined just by its order

○ Defined by a prefix and its order

Order: last

Prefix:

○ Defined by regular expression (REGEX)

● Parse as a number

☑ Add the following value to each: 1900.0

☑ ...unless less than: 13.0

...in which case add: 2000.0

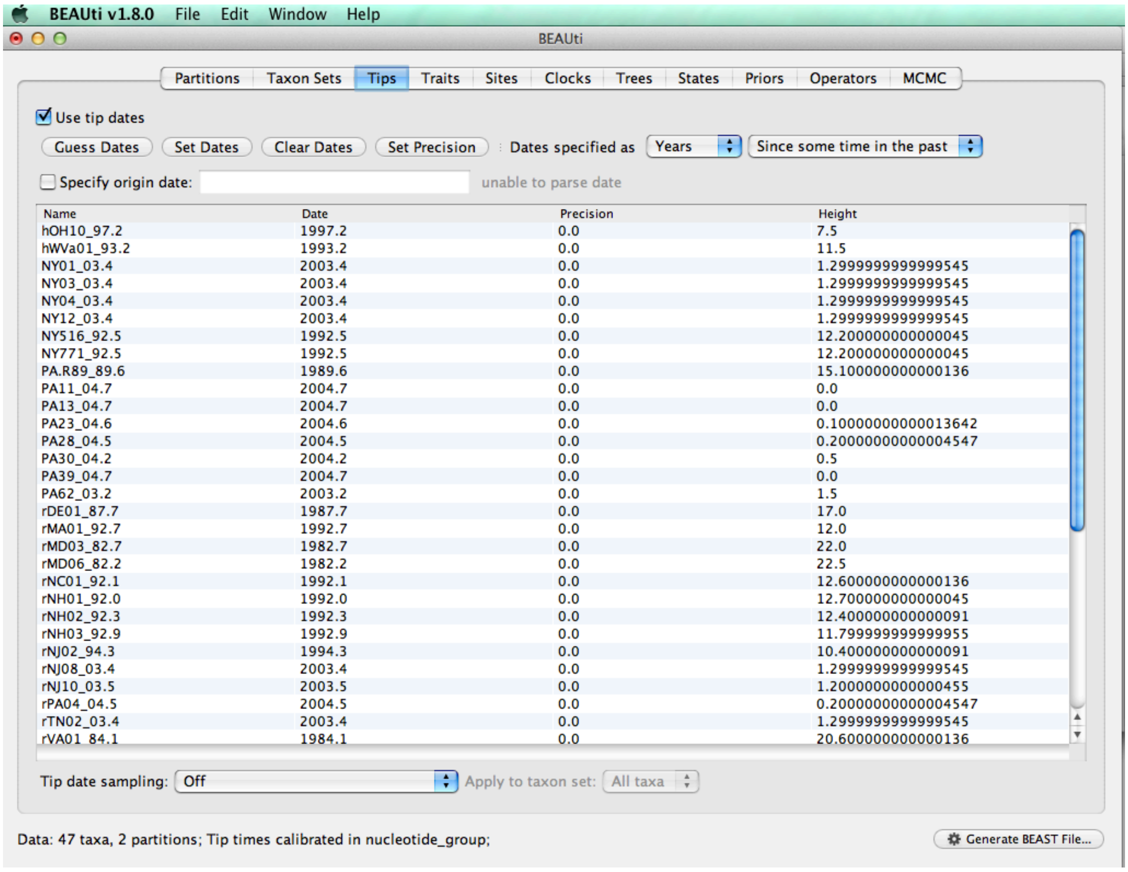
○ Parse as a calendar date

Date format: yyyy-MM-dd

○ Parse calendar dates with variable precision

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

Data: 47 taxa, 1 partition; Fix clock rate to 1.0 in nucleotide_group;


BEAUti v1.8.0 File Edit Window Help
BEAUti
Partitions Taxon Sets Tips Traits Sites Clocks Trees States Priors Operators MCMC
Guess Dates
Set Dates
Clear Dates
Set Precision
Generate BEAST File...

☑ Use tip dates

Dates specified as Years   Since some time in the past

☐ Specify origin date: unable to parse date

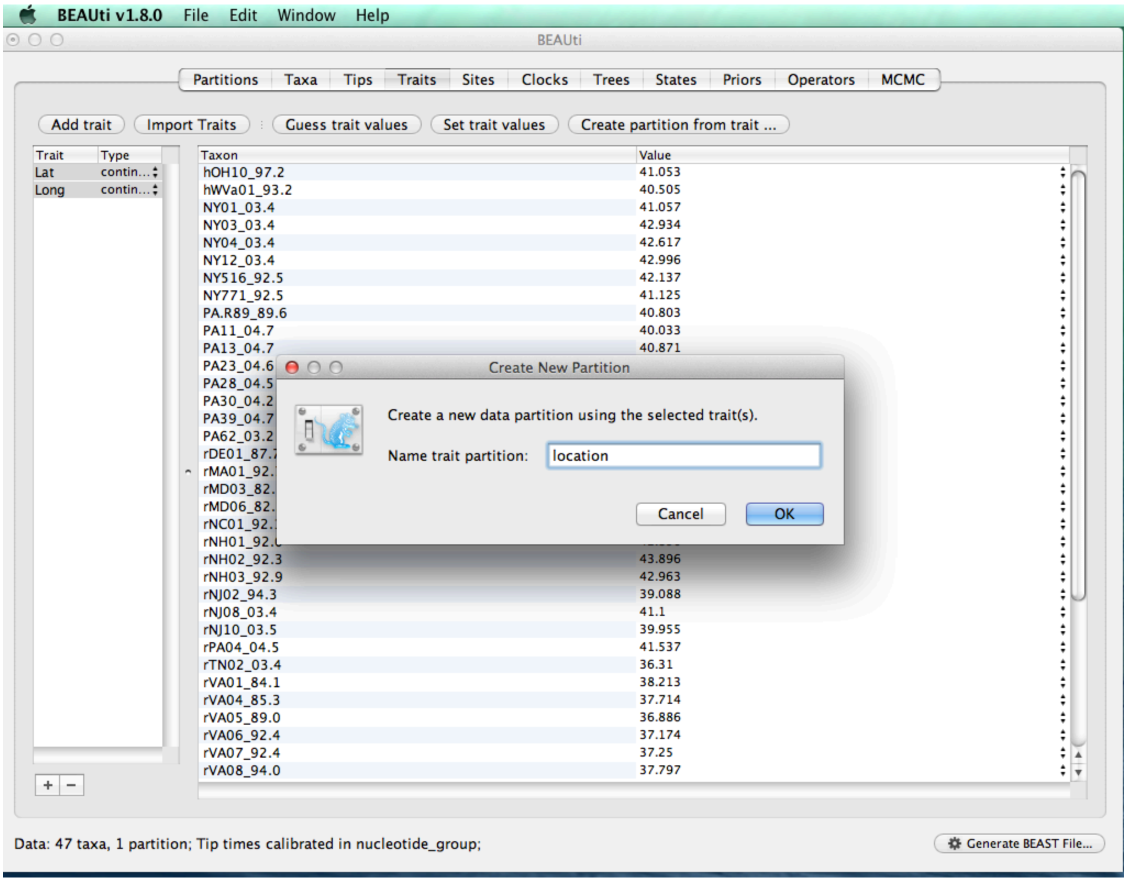
Name	Date	Precision	Height
hOH10_97.2	1997.2	0.0	7.5
hWVa01_93.2	1993.2	0.0	11.5
NY01_03.4	2003.4	0.0	1.2999999999999545
NY03_03.4	2003.4	0.0	1.2999999999999545
NY04_03.4	2003.4	0.0	1.2999999999999545
NY12_03.4	2003.4	0.0	1.2999999999999545
NY516_92.5	1992.5	0.0	12.200000000000045
NY771_92.5	1992.5	0.0	12.200000000000045
PA.R89_89.6	1989.6	0.0	15.100000000000136
PA11_04.7	2004.7	0.0	0.0
PA13_04.7	2004.7	0.0	0.0
PA23_04.6	2004.6	0.0	0.10000000000013642
PA28_04.5	2004.5	0.0	0.20000000000004547
PA30_04.2	2004.2	0.0	0.5
PA39_04.7	2004.7	0.0	0.0
PA62_03.2	2003.2	0.0	1.5
rDE01_87.7	1987.7	0.0	17.0
rMA01_92.7	1992.7	0.0	12.0
rMD03_82.7	1982.7	0.0	22.0
rMD06_82.2	1982.2	0.0	22.5
rNC01_92.1	1992.1	0.0	12.600000000000136
rNH01_92.0	1992.0	0.0	12.700000000000045
rNH02_92.3	1992.3	0.0	12.400000000000091
rNH03_92.9	1992.9	0.0	11.799999999999955
rNJ02_94.3	1994.3	0.0	10.400000000000091
rNJ08_03.4	2003.4	0.0	1.2999999999999545
rNJ10_03.5	2003.5	0.0	1.200000000000455
rPA04_04.5	2004.5	0.0	0.20000000000004547
rTN02_03.4	2003.4	0.0	1.2999999999999545
rVA01_84.1	1984.1	0.0	20.600000000000136

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

Data: 47 taxa, 2 partitions; Tip times calibrated in nucleotide_group;

3. Traits Tab:

- Click ‘**Add trait’.** In the dialogue box that appears, select ‘**Import trait(s) from a mapping file format’** and click **OK**.
- Load the **LatLong.txt** tab-delimited file, which contains latitudes and longitudes associated with each sequence. Click **OK.**
- *Select both* the **Lat** and **Long** traits and click the ‘**Create partition from trait…’** button. Enter the name **location** for this partition and click **OK**. This new partition will now appear in the Partitions tab as a continuous Data Type.



Trait	Type
Lat	contin...
Long	contin...

Taxon	Value
hOH10_97.2	41.053
hWVa01_93.2	40.505
NY01_03.4	41.057
NY03_03.4	42.934
NY04_03.4	42.617
NY12_03.4	42.996
NY516_92.5	42.137
NY771_92.5	41.125
PA.R89_89.6	40.803
PA11_04.7	40.033
PA13_04.7	40.871
PA23_04.6	
PA28_04.5	
PA30_04.2	
PA39_04.7	
PA62_03.2	
rDE01_87.	
rMA01_92.	
rMD03_82.	
rMD06_82.	
rNC01_92.	
rNH01_92.	
rNH02_92.3	43.896
rNH03_92.9	42.963
rNJ02_94.3	39.088
rNJ08_03.4	41.1
rNJ10_03.5	39.955
rPA04_04.5	41.537
rTN02_03.4	36.31
rVA01_84.1	38.213
rVA04_85.3	37.714
rVA05_89.0	36.886
rVA06_92.4	37.174
rVA07_92.4	37.25
rVA08_94.0	37.797

Create New Partition

Create a new data partition using the selected trait(s).

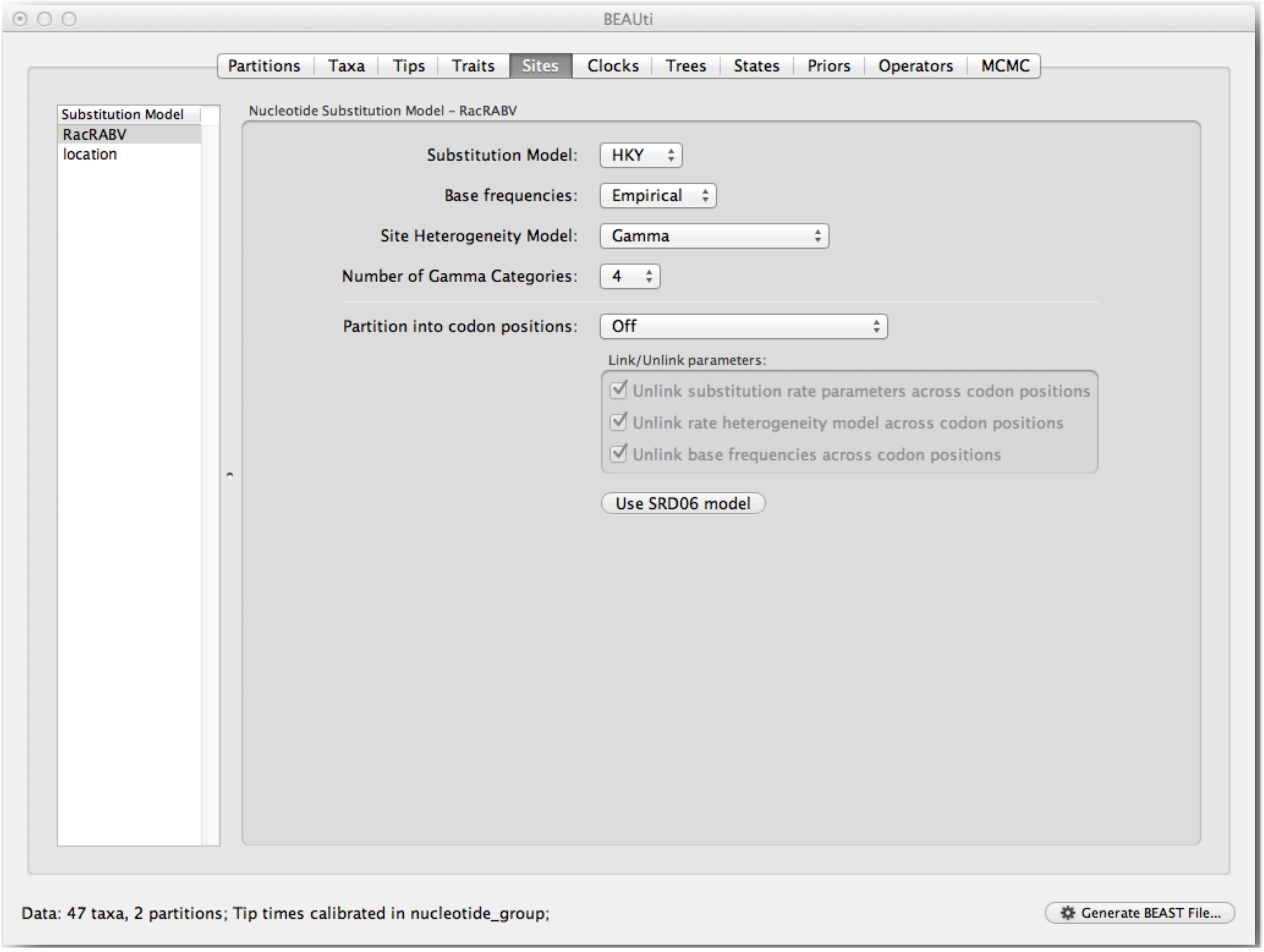
Name trait partition: location

Data: 47 taxa, 1 partition; Tip times calibrated in nucleotide_group;

4. Sites Tab:

- Keep **Substitution Model** as the default **HKY**
- Set **Base frequencies** to **Empirical**, and
- Set **Site Heterogeneity Model** to **Gamma** distributed rate variation among sites.
- Keep the **Number of Gamma Categories** at **4.**

- Click on **location** in the **Substitution Model** window on the left and keep the default **Homogenous Brownian model.**



Nucleotide Substitution Model – RacRABV

Substitution Model: HKY

Base frequencies: Empirical

Site Heterogeneity Model: Gamma

Number of Gamma Categories: 4

Partition into codon positions: Off

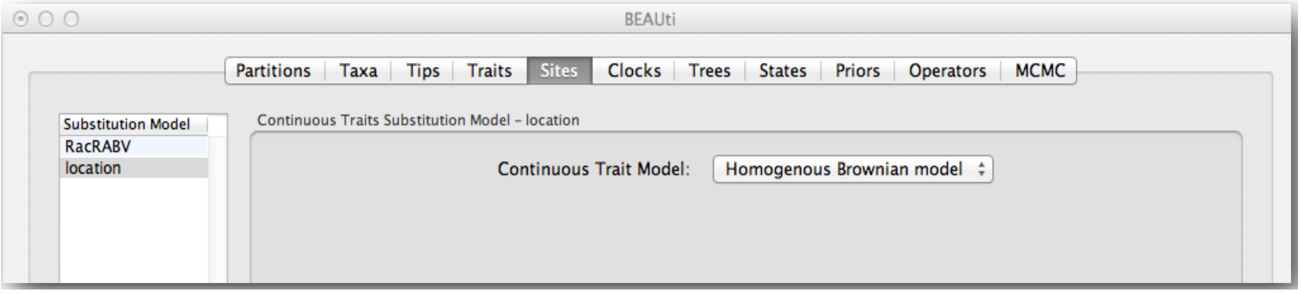
Link/Unlink parameters:

☑ Unlink substitution rate parameters across codon positions

☑ Unlink rate heterogeneity model across codon positions

☑ Unlink base frequencies across codon positions

Data: 47 taxa, 2 partitions; Tip times calibrated in nucleotide_group;

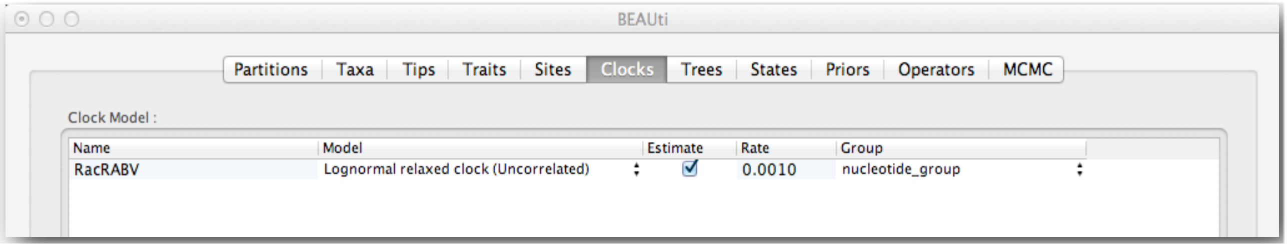


Continuous Traits Substitution Model – location

Continuous Trait Model: Homogenous Brownian model

5. Clocks Tab:

- Select the model **Lognormal relaxed molecular clock (Uncorrelated)** and set the initial **Rate** value to **0.0010**.

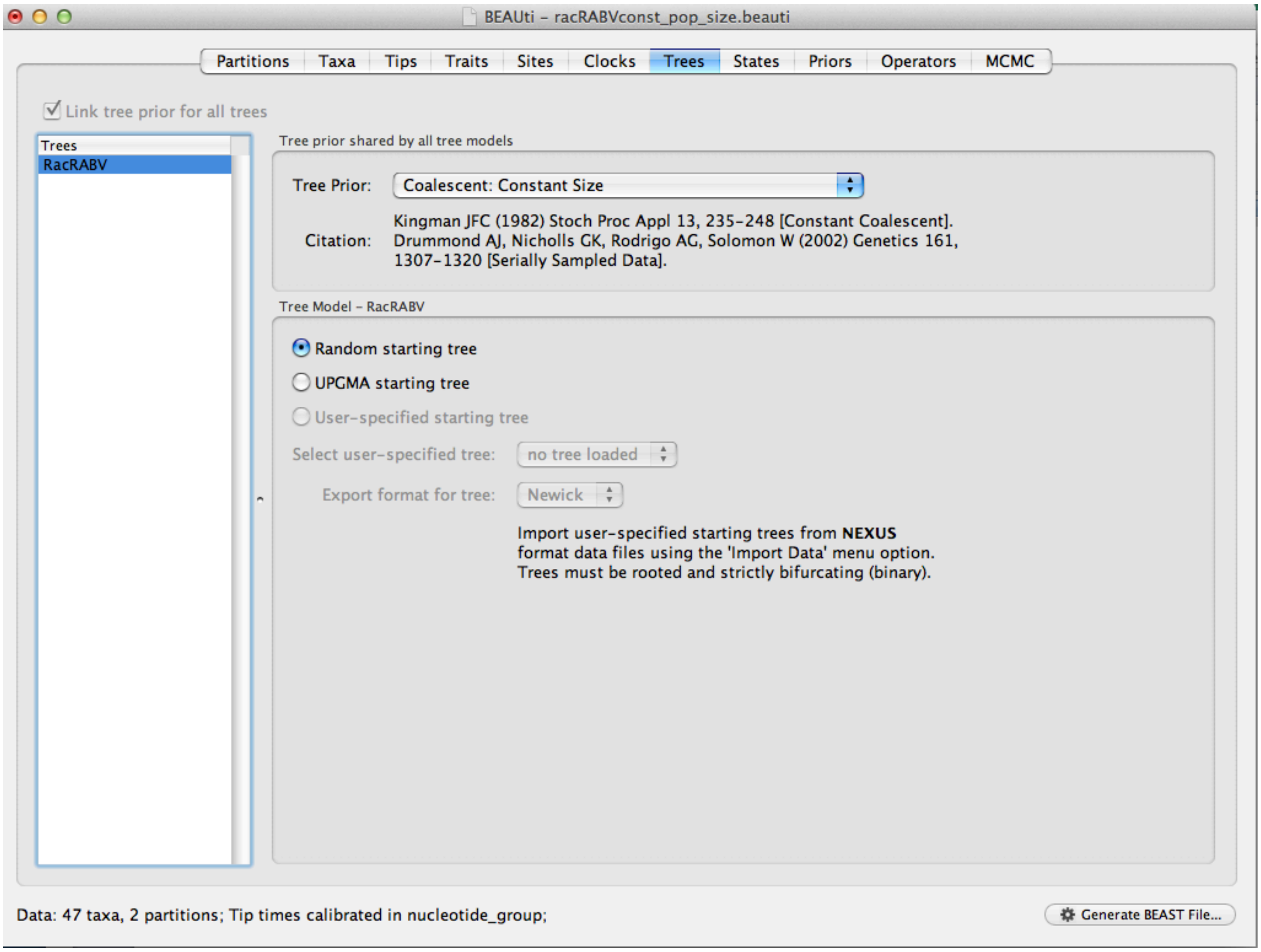


Clock Model :

Name	Model	Estimate	Rate	Group
RacRABV	Lognormal relaxed clock (Uncorrelated)	☑	0.0010	nucleotide_group

6. Trees Tab:

- On the **Tree Prior,** select the **Coalescent: Constant Size** model and keep the other default settings.



☑ Link tree prior for all trees

Tree prior shared by all tree models

Tree Prior: Coalescent: Constant Size

Citation: Kingman JFC (1982) Stoch Proc Appl 13, 235-248 [Constant Coalescent]. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Genetics 161, 1307-1320 [Serially Sampled Data].

Tree Model - RacRABV

● Random starting tree

○ UPGMA starting tree

○ User-specified starting tree

Select user-specified tree: no tree loaded

Export format for tree: Newick

Import user-specified starting trees from **NEXUS** format data files using the 'Import Data' menu option. Trees must be rooted and strictly bifurcating (binary).

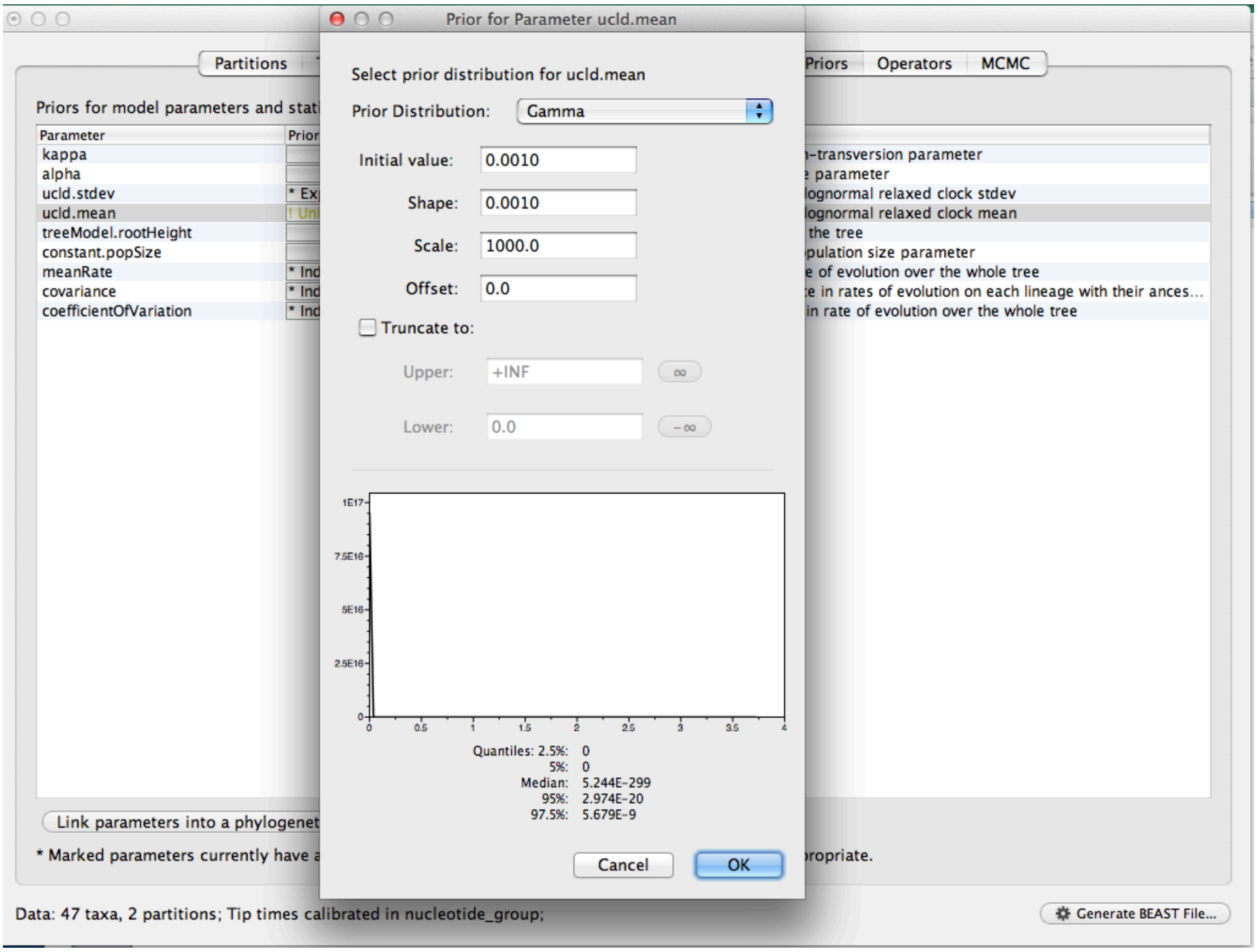
Data: 47 taxa, 2 partitions; Tip times calibrated in nucleotide_group;

7. States Tab:

- Click on the **location** partition on the left, and select the option ‘**Reconstruct states at all ancestors**’.

8. Priors Tab:

- Priors that have not been set yet appear in red (e.g. **ucld.mean**).
- Click on the prior for the **ucld.mean** parameter.
- In the dialog window that appears, set the ‘**Prior Distribution**’ to **Gamma** with **initial value = 0.001, shape = 0.001** and **scale = 1000.** Click **OK.**



Prior for Parameter ucld.mean

Select prior distribution for ucld.mean

Prior Distribution: Gamma

Initial value: 0.0010

Shape: 0.0010

Scale: 1000.0

Offset: 0.0

☐ Truncate to:

Upper: +INF

Lower: 0.0

Quantiles: 2.5%: 0
5%: 0
Median: 5.244E-299
95%: 2.974E-20
97.5%: 5.679E-9

Priors for model parameters and stati

Parameter	Prior	
kappa		-transversion parameter
alpha		parameter
ucld.stdev	* Ex	lognormal relaxed clock stdev
ucld.mean	! Uni	lognormal relaxed clock mean
treeModel.rootHeight		the tree
constant.popSize		pulation size parameter
meanRate	* Ind	e of evolution over the whole tree
covariance	* Ind	e in rates of evolution on each lineage with their ances...
coefficientOfVariation	* Ind	in rate of evolution over the whole tree

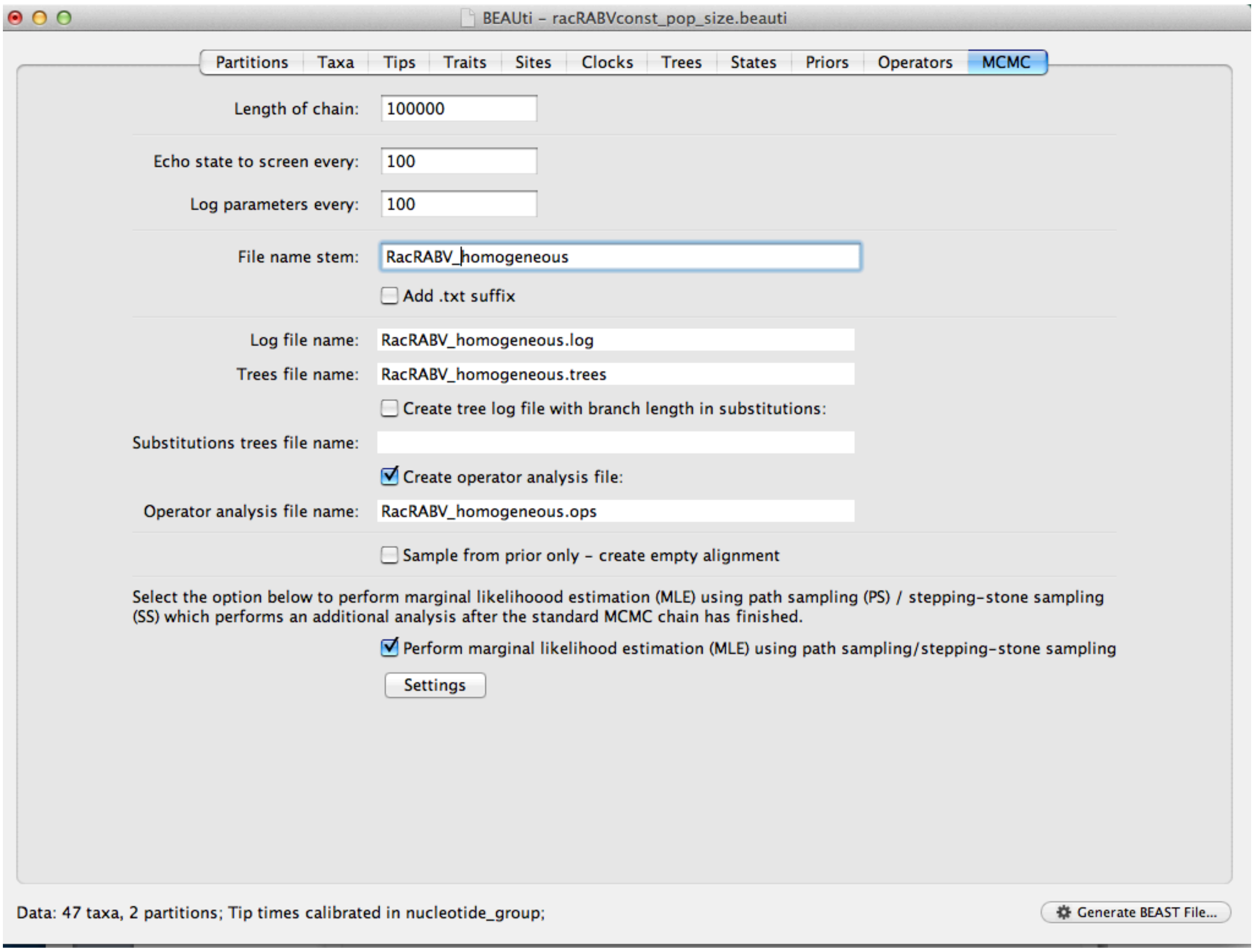
Link parameters into a phylogenet

* Marked parameters currently have a ropriate.

Data: 47 taxa, 2 partitions; Tip times calibrated in nucleotide_group;

9. MCMC Tab:

- Set the **Length of Chain** to **1 000 000**
- Set both sampling frequencies (**Echo state to screen every:** and **Log parameters every:**) to 10 000.
- Set **File name stem** to **RacRABV_homogeneous**.
- Select **Create operator analysis file.**
- At the bottom, select **Perform marginal likelihood estimation (MLE) using path sampling/ stepping-stone sampling**. Keep default settings.
- Click the **Generate BEAST File** at the bottom of the page.
- Click **Continue**. Choose the name of the file and the directory where it will be saved. Click **Save**.



Length of chain: 100000

Echo state to screen every: 100

Log parameters every: 100

File name stem: RacRABV_homogeneous

☐ Add .txt suffix

Log file name: RacRABV_homogeneous.log

Trees file name: RacRABV_homogeneous.trees

☐ Create tree log file with branch length in substitutions:

Substitutions trees file name:

☑ Create operator analysis file:

Operator analysis file name: RacRABV_homogeneous.ops

☐ Sample from prior only – create empty alignment

Select the option below to perform marginal likelihoood estimation (MLE) using path sampling (PS) / stepping-stone sampling (SS) which performs an additional analysis after the standard MCMC chain has finished.

☑ Perform marginal likelihood estimation (MLE) using path sampling/stepping-stone sampling

Data: 47 taxa, 2 partitions; Tip times calibrated in nucleotide_group;

Leave the BEAUti window open, so that you can change the values and re-generate the BEAST file if necessary.

Open the **RacRABV_homogeneous.xml** file in an xml text editor to manually add an additional diffusion model parameter (**treeDispersionStatistic)**, which *cannot* be specified by this version of **BEAUti.** This parameter records the *rate of diffusion* along each branch, and provides a diffusion estimate in km/yr.

Add the following block (e.g., by copy- pasting) just before the operators block in the xml:

```
<treeDispersionStatistic id="dispersionRate" greatCircleDistance="true">

        <treeModel idref="treeModel"/>

        <multivariateTraitLikelihood idref="location.traitLikelihood"/>

</treeDispersionStatistic>

<!-- END Multivariate diffusion model                    -->

<!-- Define operators                                    -->
```

And also add a reference to it in the log file:

```
<log id=”fileLog” logEvery=”100” fileName=”RacRABV_homogeneous.log”
overwrite=”false”>

<posterior idref=”posterior”/>

<prior idref=”prior”/>

....

<treeDispersionStatistic idref=”dispersionRate”/>

 ....
```

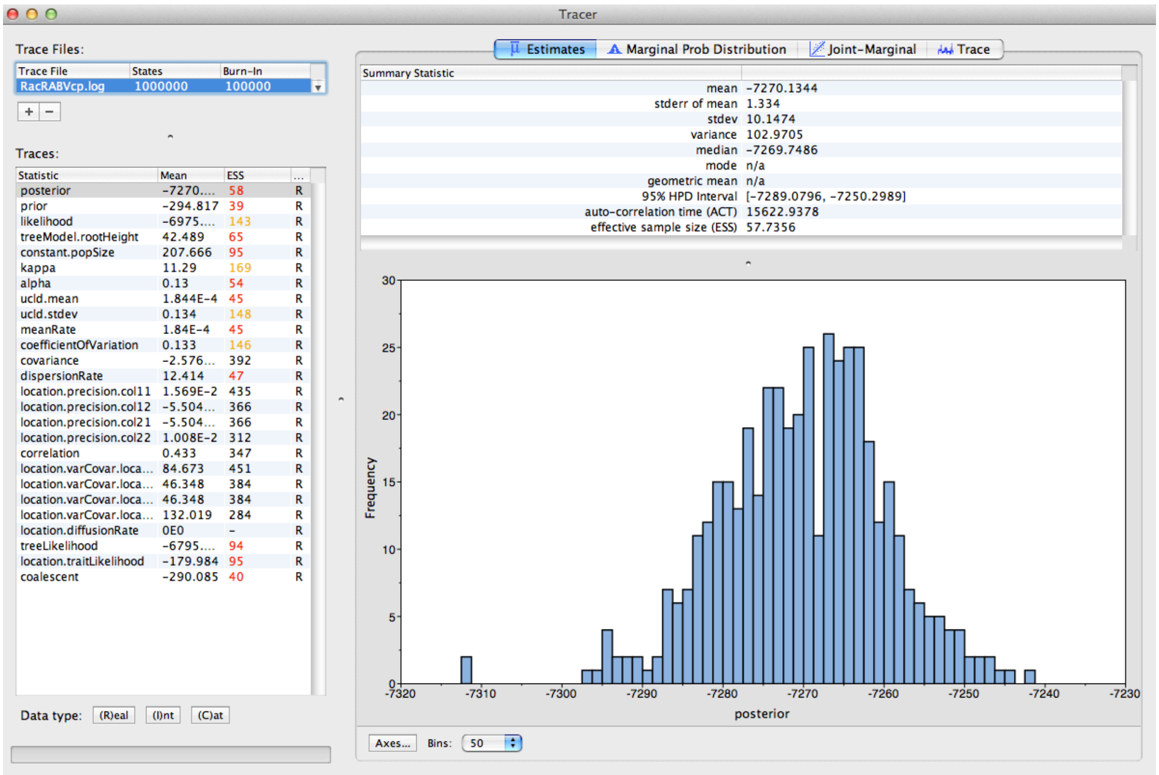
Save the file.

RUNNING BEAST

- Click on **Choose File…** and open **RacRABV_homogeneous.xml**.
- Select the ‘**Use BEAGLE library if available**’ option.
- Keep default settings and select **Run.**

ANALYSING THE BEAST OUTPUT

- Execute **Tracer**
- Select **Import Trace File…** from the **File** menu and load the log file created by BEAST (**RacRABV_homogeneous.log).**



Trace File	States	Burn-In
RacRABVcp.log	1000000	100000

Statistic	Mean	ESS	...
posterior	-7270....	58	R
prior	-294.817	39	R
likelihood	-6975....	143	R
treeModel.rootHeight	42.489	65	R
constant.popSize	207.666	95	R
kappa	11.29	169	R
alpha	0.13	54	R
ucld.mean	1.844E-4	45	R
ucld.stdev	0.134	148	R
meanRate	1.84E-4	45	R
coefficientOfVariation	0.133	146	R
covariance	-2.576...	392	R
dispersionRate	12.414	47	R
location.precision.col11	1.569E-2	435	R
location.precision.col12	-5.504...	366	R
location.precision.col21	-5.504...	366	R
location.precision.col22	1.008E-2	312	R
correlation	0.433	347	R
location.varCovar.loca...	84.673	451	R
location.varCovar.loca...	46.348	384	R
location.varCovar.loca...	46.348	384	R
location.varCovar.loca...	132.019	284	R
location.diffusionRate	0E0	-	R
treeLikelihood	-6795....	94	R
location.traitLikelihood	-179.984	95	R
coalescent	-290.085	40	R

Summary Statistic	
mean	-7270.1344
stderr of mean	1.334
stdev	10.1474
variance	102.9705
median	-7269.7486
mode	n/a
geometric mean	n/a
95% HPD Interval	[-7289.0796, -7250.2989]
auto-correlation time (ACT)	15622.9378
effective sample size (ESS)	57.7356

Answer the following questions

a) At what rate did RABV spread in the raccoon population in North America? (Note that the dispersionRate statistic is measured in km/year).
b) What is the age of most recent common ancestor (TMRCA)?
c) What does it mean when the EES values are displayed, in red, yellow or black? How can we improve ESS estimates?

SUMMARISING THE TREES

- Open **TreeAnnotator.** Click on **Choose File…** next to **Input Tree File** and open **RacRABV_homogeneous.trees**

- **Burn-in** - This is the number of trees in the input file that should be excluded from the summarization. This value is given as the number of trees rather than the number of steps in the MCMC chain. Thus for a chain length of 10 000 000 steps, sampling every 10 000 steps will result in 1000 trees. To obtain a 10% **Burnin**, set the value to **100**.

- **Posterior probability limit** - This is the minimum acceptable posterior probability for a node. Keep the **default** is **0.0** so all nodes will be annotated.

- **Target tree type** – Keep the default **Maximum clade credibility tree** option. TreeAnnotator will examine every tree in the Input Tree File and select the tree that has the highest sum of the posterior probabilities of all its nodes.

- **Node heights** - Keep the default **Median heights**.

- **Output File** – Click on **Choose File…** next to **Output File**. Enter a name for the output file (e.g. **RacRABV_homogeneous_mcc.tre**).

- Click **Run**

VIEWING THE ANNOTATED TREE

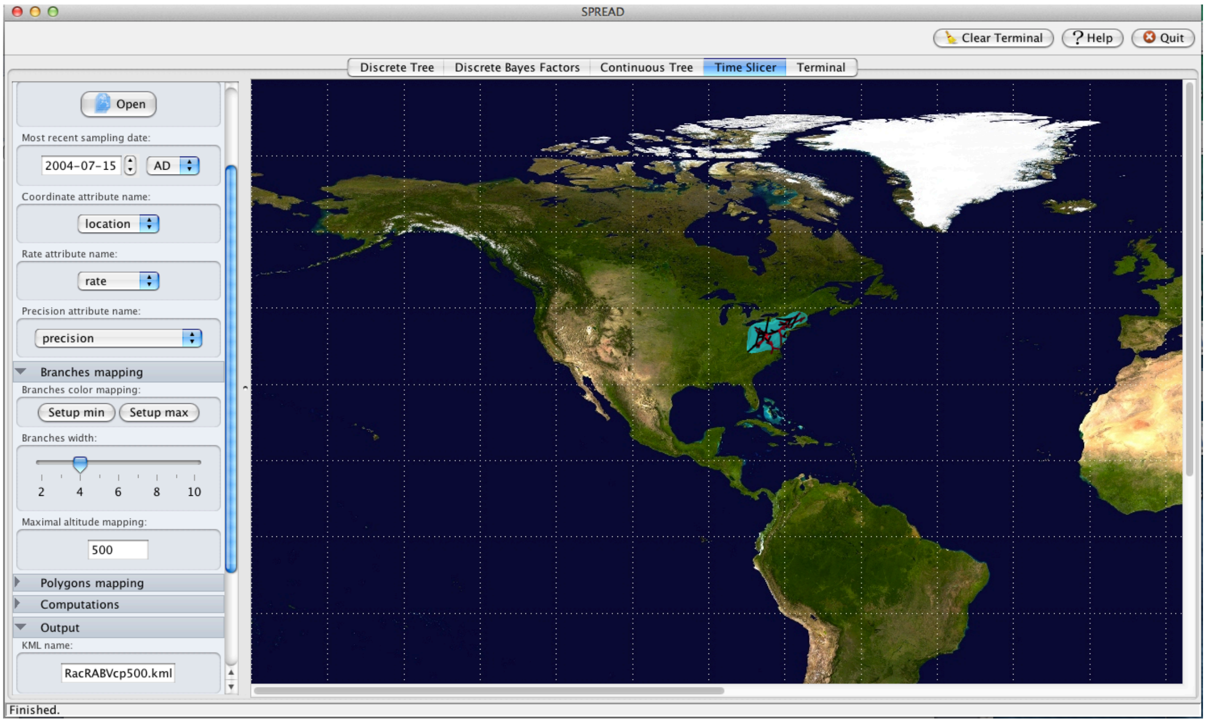
- Open **Figtree**. Under the **File** menu, select **Open.**
- Select and open the **RacRABV_homogeneous_mcc.tre** file created by TreeAnnotator
- Check next to and open the **Branch Labels tab** on the left.
- Next to **Display,** select **Posterior** from the dropdown menu.
- We now want to display bars on the tree to represent the estimated uncertainty in the inferred date for each node. Check and open the **Node Bars tab** and choose **Display: height_95%_HPD** from the dropdown menu.

- To display a time scale for this phylogeny, deselect the **Scale bar tab.**
- For appropriate scaling, open the **Time Scale tab**, and set the **Offset by:** to **2004.7**, the **Scale Factor:** to **-1.0**.
- Tick and open the **Scale Axis tab.** Check the **Reverse axis** option.
- In the **Layout tab,** tick **Align Tip Labels**.
- Finally, open the **Appearance tab** and set the **Line Weight** to **3**.

None of the options actually alter the tree's topology or branch lengths in anyway, so feel free to explore the options and settings. You can also save the tree with all your settings so that you can reload it into FigTree.

EVALUATING DIFFUSION RATE VARIATION

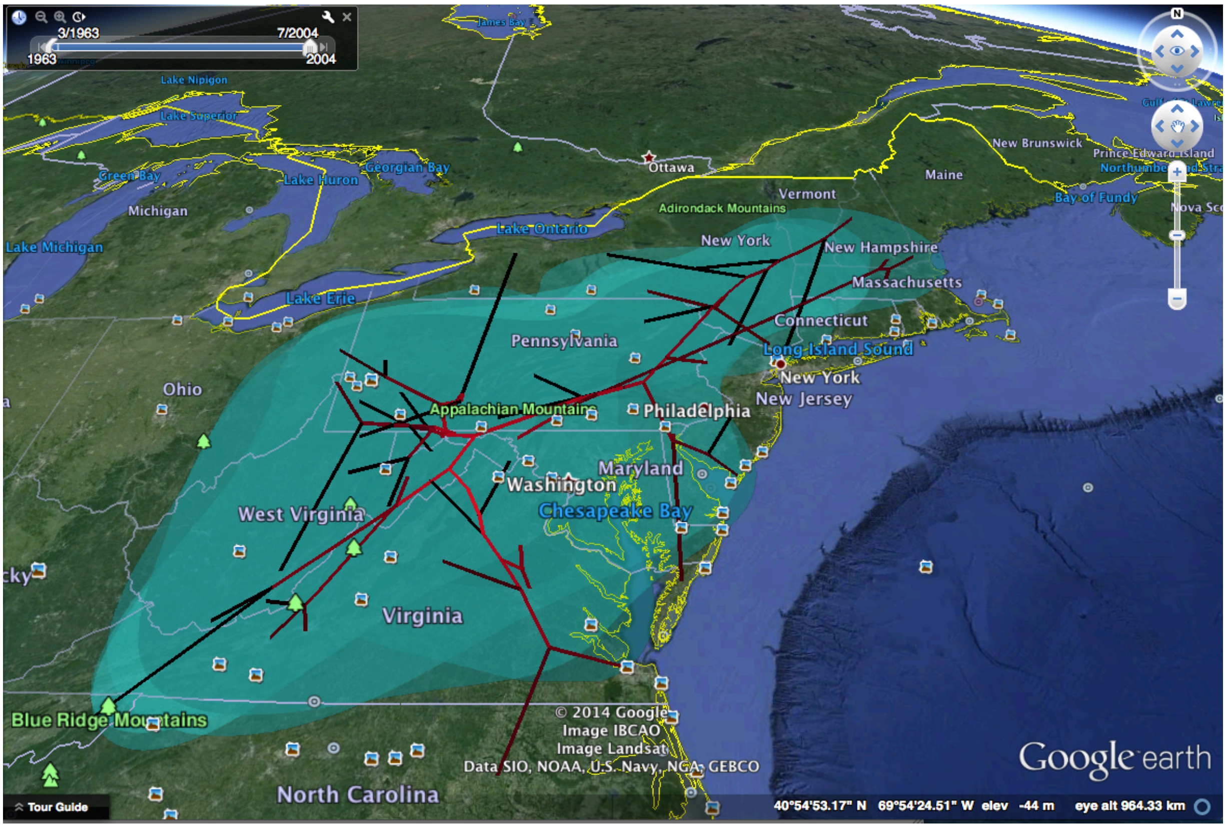
- Open **SPREAD**
- Click on the **Time Slicer tab.**
- Under **Load slice heights/ tree file,** open **RacRABV_homogeneous_mcc.tre.**
- Next to **Load trees file,** open **RacRABV_homogeneous.trees**
- Set the **Most recent sampling date** to **2004-07-15.**
- Under **Coordinate attribute name**, select **location.**
- Under **Rate attribute name,** keep the default **rate.**
- Under the **Branches mapping tab**, set the **Maximal altitude mapping** to **500.**
- Under the **Output tab,** give an appropriate **KML name** (e.g. **RacRABVcp500.kml**).
- Click **Plot** and **Generate.**



Most recent sampling date: 2004-07-15 AD
Coordinate attribute name: location
Rate attribute name: rate
Precision attribute name: precision
Branches color mapping:
Branches width:
Maximal altitude mapping: 500
KML name: RacRABVcp500.kml
Finished.

- The **RacRABVcp500.kml** file will automatically be saved in the same directory from which the **tree** file was loaded.
- Open **RacRABVcp500.kml** in **Google Earth**.

- Move the slider in the top-left corner to the start.
- Click the **Spanner icon** in the same slider panel.
- Set the **Animation speed** to **Slower** and click **OK.**
- In the slider panel, click the **Clock and triangle** button to start the animation.



REFERENCES

Biek R, Caroline Hendorson J, Waller LA, Rupprecht CE, Real LA (2007). **A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus.** Proceedings of the National Academy of Sciences, USA 104(19): 7993–7998.

Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009). **Bayesian Phylogeography Finds Its Roots.** PLoS Computational Biology 5(9): e1000520. doi:10.1371/journal.pcbi.1000520

Recommended Literature

BEAST2 manual
http://beast2.org/

Selecting proper priors
http://code.google.com/p/beast-mcmc/wiki/ParameterPriors

Lemey P *et al.,* 2010. Phylogeography Takes a Relaxed Random Walk in Continuous Space and Time.
http://mbe.oxfordjournals.org/content/27/8/1877.full

BEAST: Bayesian evolutionary analysis by sampling trees. Alexei J Drummond and Andrew Rambaut. 2007.
http://www.biomedcentral.com/1471-2148/7/214

SPREAD Tutorial.
http://www.kuleuven.be/aidslab/phylogeography/tutorial/spread_tutorial.html

Phylogeographic Inference in continuous space: A hands-on Practical. Suchard M and Lemey P. July 2013.
https://perswww.kuleuven.be/~u0036765/SISMID/handouts_files/2013_SISMID_12_16.pdf