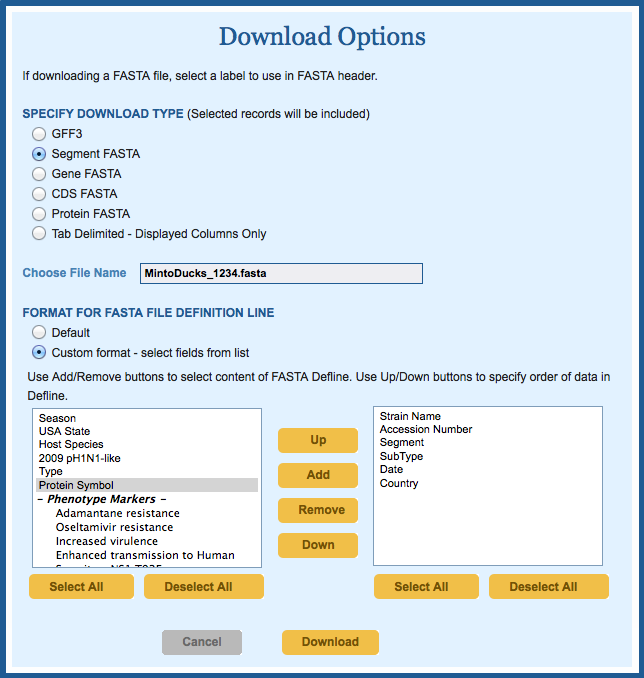
**Most Recent Common Ancestor method**

This is a step-by-step explanation of how to perform Most Recent Common Ancestor (MRCA) method using phylogenetic software. The end result is a gene tree that estimates divergence times in years and is useful for pinpointing when a particular influenza isolate split from a lineage of interest. This ‘cheat sheet’ was prepared by Nichola Hill and Islam Hussein to summarize training received from Justin Bahl. Please email with updates/corrections (nhill@mit.edu).

Creating a well-balanced tree using PAUP…

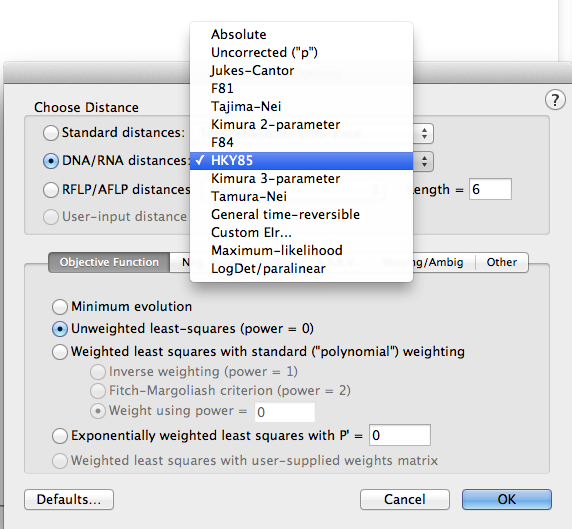


**Influenza Research Database (IRD)…**

1. Run a broad search for sequences of interest within IRD
2. Select sequences from 1976 onwards. Before this time, -80 freezers did not exist so viruses were periodically passaged in eggs to keep them viable (risking accumulation of mutations due to adaptation)
3. Include partial sequences (these can be informative)
4. Download sequences as ‘Segment FASTA’ (Fig. 1)
5. Choose ‘Custom Format’ and select metadata of interest: strain name, accession, subtype and date (important!).

**Geneious…**

1. Click Import > From File. Browse to fasta file downloaded from IRD
2. Ensure all nucleotides are DNA (PAUP does not like ‘U’ ie. uracil). Select any RNA sequences and click ‘Convert between RNA and DNA’
3. Remove any sequences less than 1000bp
4. Run an alignment using MUSCLE with settings ‘Terminal gaps: half’ (everything else is default settings).
5. On the Display side panel click Translation (Frame 1). Trim off any sequence before the start codon. Translation usually begins with an ATG which specifies the amino acid methionine; the next triplet, GAT, specifies aspartic acid, and so on.
6. Trouble segments: NS and M because they have overlapping reading frames. For M: translate according to Frame 1 to find start codon. To find stop codon, translate to Frame 2.
7. Check for indels in the consensus sequence. In coding regions of the genome, unless the length of an indel is a multiple of 3, a frameshift mutation will occur. Knowing that frameshift mutations are deleterious for the virus, we can assume that indels 1-2 nucleotides or > 3 nts in length are a result of i) sequencing artifacts or ii) poor alignments by alignment program.
8. Remove entire sequences that contain indels (ie. gaps of 1-2 in consensus sequence)
9. Export alignment in nexus format with interleaved format



**PAUP\*…**

1. File > Log Output to Disk. Important to keep a record of all steps you took to build tree. Tick box for ‘Include contents of display buffer in log file’. Open the log file.
2. File > Open and browse to nexus file. Ensure ‘Initial Mode: Execute’ option button is on.
3. PAUP is fussy with syntax. You may need to clean up your nexus file in TextWrangler. Escape any commas (‘) ie. Bewick’s swan was a troublemaker. So put (‘’). Save.
4. Back to PAUP. Analysis > Distance.
5. Analysis > Distance Settings. Choose HKY85 substitution model (Fig. 2). This model distinguishes between the rate of [transitions](http://en.wikipedia.org/wiki/Transition_%28genetics%29) & [transversions](http://en.wikipedia.org/wiki/Transversion) (using the κ parameter) & allows unequal base frequencies.
6. Options > Tree Drawing Options. Select: “Ladderize right” and “Never truncate taxon names”
7. Options > Rooting. Select Midpoint.
8. Analysis>Neighbor-Joining/UPGMA…this will spit out your tree!
9. Tree > Show Tree. This will produce a tree with aligned tips so you can select them.
10. \*If working from the command line your PAUP block will look like this:

Begin paup;

Log File='/Users/Nic/PAUP/ENTER FILE NAME HERE’;

DSet distance=HKY85;

Set tOrder=right taxLabels=full;

Set rootMethod=midpoint;

NJ;

Showtree;

End;

1. Prune your tree. The idea is to keep all taxa that cluster together with your strain of interest (ie. preserve your ingroup) and remove extraneous sequences that clutter the tree.
2. Make a duplicate of your .log file and save the copy with the file extension .txt (important to save as Plain Text otherwise it may truncates names!).
3. Open this file in TextEdit. Highlight removable taxa with Alt + Command. Paste into new TextEdit file called ‘DeleteX’. Find (.+) and replace with ‘\1’ to wrap taxa names in apostraphies (to make PAUP happy).
4. Import text file into Excel (taxa to delete now appears in column A). In column B write formula ‘=RAND()’. This will assign a random number to all taxa. Order the column either by ascending or descending. Select and copy half of those taxa (or any fraction depending on the size of your desired tree).
5. Back in your nexus file, under your taxa names type:

Begin paup;

Delete

YOUR LIST OF NAMES

/only;

export file=YOURNEW-FILENAME.nex fomat=nex interleave=NO;

End;

…then Execute in the GUI. This step will prune the tree of extraneous taxa and make it more manageable for downstream analysis in BEAST.

1. Trees >Save Trees To File. Save as nexus file.

**Fig Tree…**

1. Open the tree in FigTree to visualise results of pruning (& decide if more is needed)

**Text Wrangler…**

1. Open nexus file in TextWrangler. Clean up dates in taxa name using regular expressions (make sure ‘GREP’ box is ticked). This is critical because all isolates must have a date to assess MRCA. There are two options for the dates; non-Americanised (DD/MM/YYYY) or American (MM/DD/YYYY) depending on your target journal or outlet.

DD/MM/YYYY…

1) for aesthetics: Find: \r\s+ Replace:

Find: ; Replace: ;\r

2) for year with no month or date, add prefix ‘15/06’:

Find: \|(\d\d\d\d’) Replace: |15/06/\1

3) for year and month but no day, add prefix 15:

Find: \|(\d\d/\d\d\d\d’) Replace: |15/\1

MM/DD/YYYY…

1) for aesthetics: Find: \r\s+ Replace:

Find: ; Replace: ;\r

2) for year with no month or date, add prefix ‘06/15’:

Find: \|(\d\d\d\d’) Replace: |06/15/\1

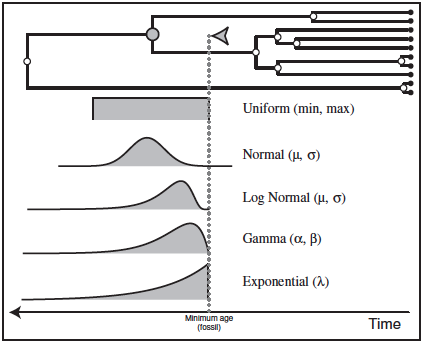
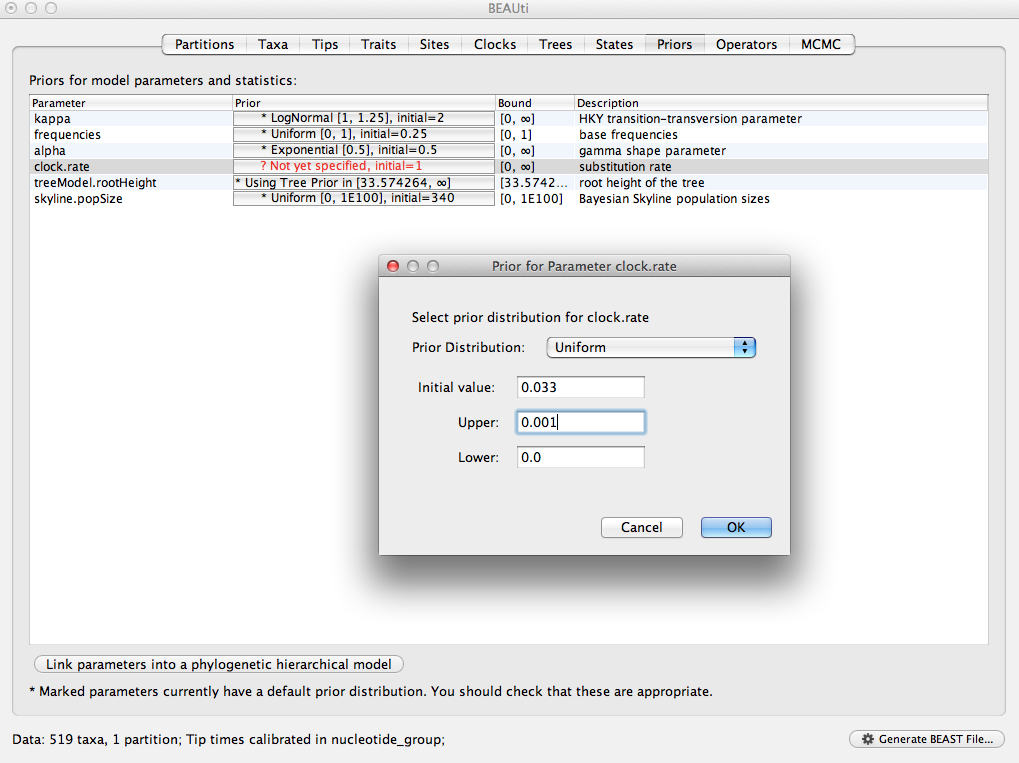
3) for year and month but no day, add 15 between month and year:

Find: \|(\d\d)(/\d\d\d\d’) Replace: |\1/15\2

Estimating most recent common ancestor using BEAST…

**BEAUti…**

1. File > ‘Import Data’ to open nexus file from PAUP
2. Select Tips tab. If isolates have been collected by year or month, it is possible to add this information into the analysis. Click ‘Use tip dates’, then ‘Guess dates’ to extract the date from the taxa name. Select Define By a Prefix and Its Order’, ‘Order = Last’, Prefix = |. Select ‘Parse as a calendar date’ and enter format: MM/DD/YYYY or DD/MM/YYYY
3. Select Traits tab. Add the traits that you want to guess. If geographic location or host species is available in strain name, use ‘Guess traits’ to extract this information. This is handy for colour coding the tree later on in FigTree. This will also generate likelihood values for your trait of interest.
4. Select Sites tab: Substitution model: HKY or GTR. Base frequencies: Estimated. Site heterogeneity model: G+I. Partition into codons: Off (unless sequence was trimmed to the ORF). What if we did trim the sequence to the ORF? Which option do we choose? 2 or 3 partitions? I’m comfortable with keeping this off because we trimmed the sequences, so thre should be only one working reading frame. We shouldn’t need to partition into multiple reading frames. Codon models offer slightly more power than nucleotide models.
5. Select Clocks tab: Lognormal relaxed clock (Uncorrelated). Tick ‘Estimate’
6. Select Trees tab: Choose a coalescent model either: \* Coalescent: Bayesian Skyline: Skyline models require that the number of coalescent events be specified before analysis. The time between coalescent events depends on the population size. Therefore the number of coalescent events can be approximated by dividing the total number of taxa (example: 300 isolates) by the number of years that your data encompass (example: 2013-1976 = ~30 years) resulting in ~ 10 coalescent events. \* Coalescent: GMRF Bayesian Skyride: Skyride is an extension of the Bayesian skyline plot and assumes that populations change gradually. Therefore population size changes between intervals are smoothed. In contrast, Skyride models do not require you to specify the number of coalescent events but this can be demanding of the program and does not necessarily improve the results. However, there can be substantial coalescent error if you estimate the wrong number of coalescent events.
7. Select Priors tab and click ‘ucld.mean’. Change Prior Distribution to Uniform, Initial value=0.033 (~33=2010-1977, determined by span of dates), Upper=0.1(now), Lower=0.0 (start of time) (Fig. 3). Can keep the default values if the time span is small. You must have both max. and minimum age bounds when applying a uniform calibration prior. This distribution places equal probability across all ages spanning the interval between the lower and upper bounds (Fig. 4)

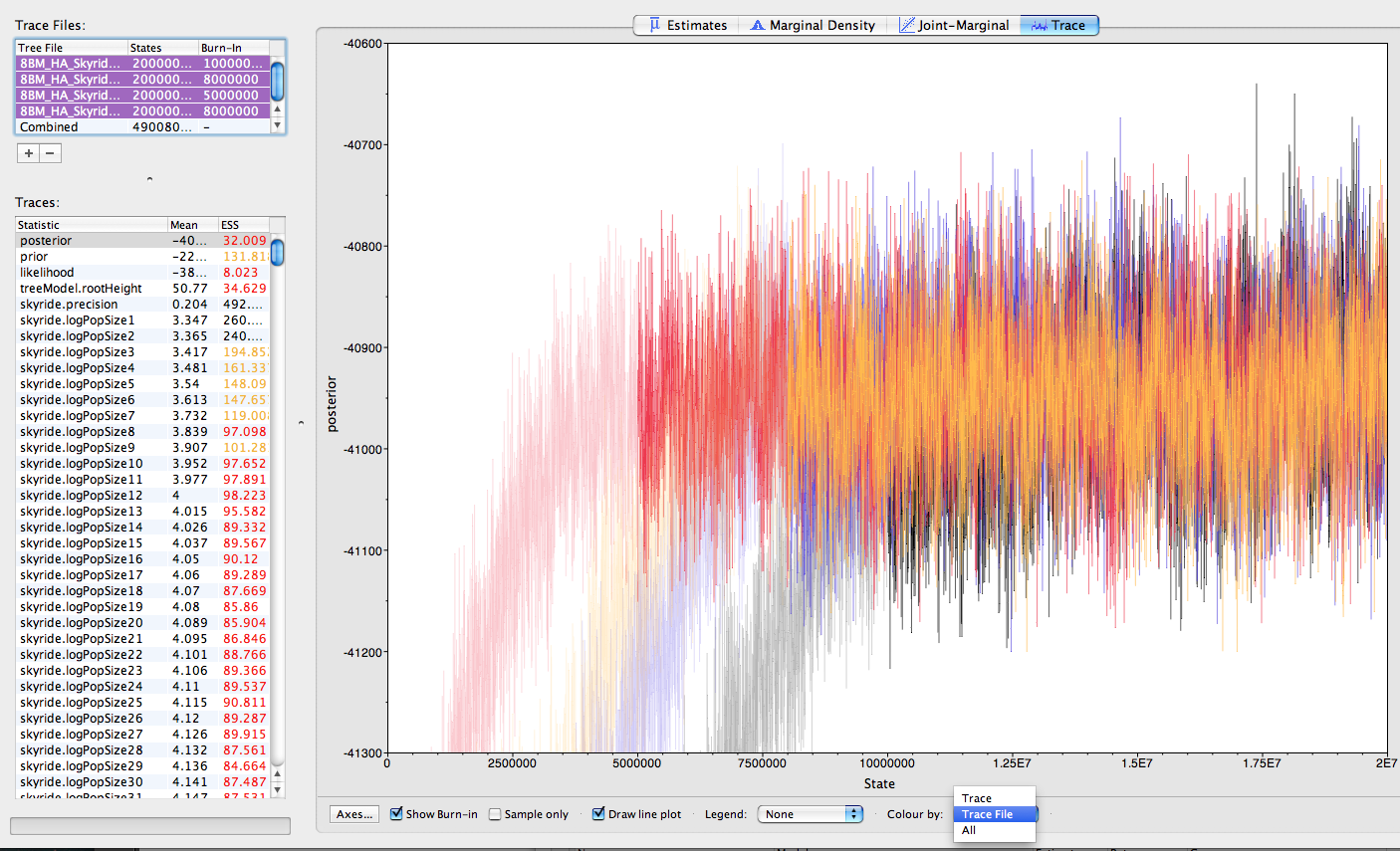
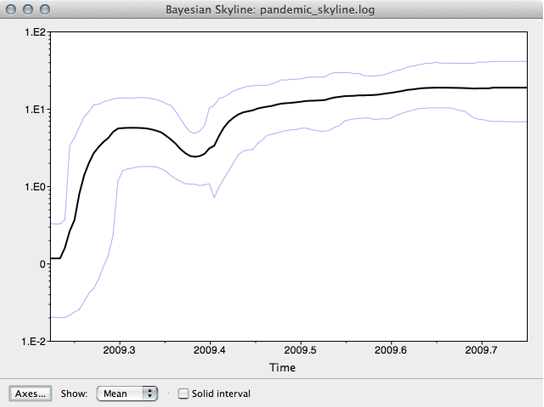
****

1. Select MCMC tab: Length of chain: 20,000,000. Try to log parameters (sample frequency) to obtain 10,000 trees, otherwise TreeAnnotator will run out of memory attempting to analyse too many trees. i.e. for chain length of 20 million, log parameters every 2000.
2. Make sure your output includes 4 files: log, (time)trees, (substitution)trees and operators file. You also don’t need to tick in the Add .txt suffix box.
3. Name your output files ‘Job\_Gene \_Skyride/Skyline\_runX\_Date’

**BEAST…**

1. Import .xml file…easy peasy.
2. Select BEAGLE library to increase efficiency of MCMC simulation. This works by utilizing GPU processor to help run the analysis.
3. *Highly recommended*: run 4 independent BEAST runs using the same .xml file (then use LogCombiner to merge the MCMC chains (later on)
4. Click on Run. This will start the BEAST run and log MCMC setup and progress to a window. The TMRCA ('rootHeight') and the per-year substitution rate ('clock.rate') will be logged to this window. These two parameters give a good idea of whether the MCMC appears to be reaching convergence and you can estimate the time of the run.
5. *Highly recommended*: get the XML up and running and tested locally and then run BEAST on a cluster node to perform the full analysis. To do this…ssh into rous.mit.edu (Koch Institute Cluster), change directory to locate your .xml file on the server and use script file (.sh) to submit job to SGE.

**Tracer…**

1. Import .log file into Tracer to diagnose the quality of your runs
2. Check the effective sample size (ESS) for each of the log files. If they are <100 and red, this is not good and means the sampling is auto-correlated (not independent). Ideally the ESS>200 and the trace should show a ‘hairy caterpillar’ indicating good mixing.
3. To improve ESS Option 1 (easy): Burn-in. Click on ‘Posterior’ and then the ‘Trace’ window and select a burn-in value that removes any samples prior to convergence (plateau). The ESS values should be >100 for the following statistics: posterior, prior, likelihood, treeModel.rootHeight, ucld.mean, treeLikelihood and skyride. The other statistics are noise and only relevant when analysing demographic statistics for population genetics. Repeat for all your runs and assess simultaneously by selecting each log file and clicking ‘Colour by: TraceFile’ (bottom right of window: see figure below).
4. To improve ESS Option 2 (easy): Use the operators to tune the parameters. Open the Operators file (.ops) in Text Wrangler. Enter the suggested operators in the XML file.
5. To improve ESS Option 3 (time consuming): Increase chain length. Find the lowest ESS for logged quantities ie. 20, then run chain 10 times longer to get reliable ESS >200.
6. Click the Estimates tab. Check the prior for ‘ucld.mean.rate’ ie. clock rate. Mean should be centered ~10-3 and be roughly bell shaped ie. unimodal.
7. Analysis>GMRF Skyride Analysis. Input the log file and specify the Age of youngest tip (ie. 2013.75)
8. Click OK to run analyses and generate a skyline plot (Fig. 6) to see a distribution of coalescent events
9. If you have multiple MCMC runs, just drop them into Tracer, select ‘combined’ log file and it will automatically concatenate the runs in the display. However use LogCombiner to implement concatenated file (Tracer will not do this for you).

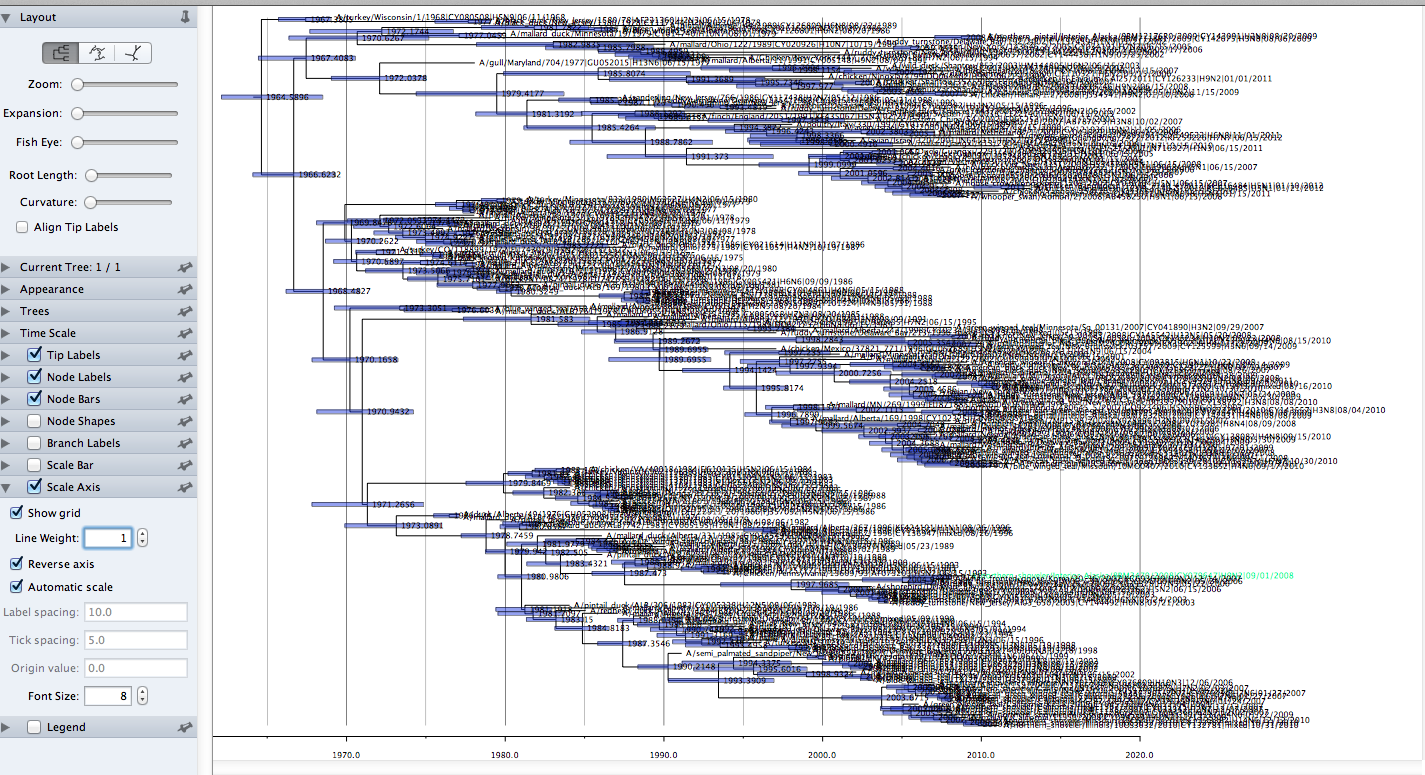
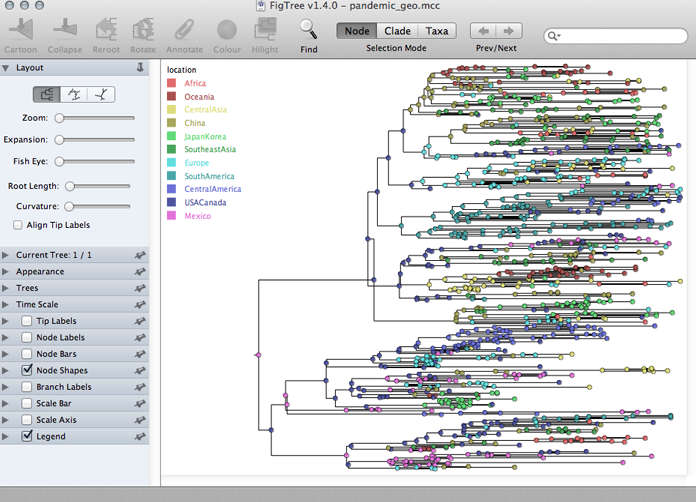
**Log Combiner…**

1. Select File Type: Log Files
2. \*If using log files from an interrupted MCMC run (ie. only reached 10 million and not 20 million chains) then use TextWrangler to remove the last incomplete line in the log file
3. Add your log files to Log Combiner panel by dragging. This step is much more efficient if your log files are on your hard drive (not cluster or server).
4. Burnin = the number of chains to remove from the start of the run as the model explores the probability space. Use values directly from Tracer.
5. Down sampling is required to thin the tree. Check ‘Resample states at lower frequency’ and enter a value that is a multiple of your original sampling frequency. If your original sampling frequency is 2000 then enter 4000 to reduce the sampling frequency by half.
6. Save with ‘.comb.log’ in the file name.
7. Inspect the combined log file in Tracer again. The ESS values should be high and all runs should have converged on the same optimal tree.
8. Select File Type: Tree Files
9. \*If using tree files from an interrupted MCMC run, then use TextWrangler to remove last tree and append with ‘;End’ on new line (in keeping with nexus syntax).
10. Add your tree files to Log Combiner panel using ‘+’.
11. Burnin is specified in trees ie. for chain length 20,000,000 with sampling every 2,000 steps, the tree file will contain 10,000 trees. Burnin ~ 10% will be 1,000 (i.e. 1/10 \* 10,000) for each run. Also you may divide the burn-in by [chain length/10,000].
12. Down sampling is required to thin the tree. Check ‘Resample states at lower frequency’ and enter a value that is a multiple of your original sampling frequency. If your original sampling frequency is 2000 then enter 4000 to reduce the sampling frequency by half.
13. Save with ‘.comb.trees’ in the file name

**Tree Annotator…**

1. \*If performing burnin now, specify the actual number of trees (not the actual number of samples). To specify a 10% burnin use the value of 1000. *Highly recommended*: Use a higher burnin (ie. 15-20%) if you have multiple BEAST runs
2. Add your tree files to Log Combiner panel by dragging. This step is much more efficient if your tree files are on your hard drive (not cluster or server).
3. Posterior probability limit: value of 0.5 means that only nodes seen in the majority of trees will be annotated with a ‘confidence value’. Or select 0.0 to show all nodes
4. Target tree type: choose ‘Maximum clade credibility tree’, this will find the tree with the highest product of the posterior probabilities
5. Node heights: choose ‘Mean heights’
6. Open the ‘.comb.trees’ file
7. Save with ‘.annot.mcctree’ in the file name

**Fig Tree…**

1. Open the combined tree file in Fig Tree
2. Appearance > Line weight: 1.5
3. Trees > Order nodes > increasing (youngest taxa at bottom)
4. Time scale > Offset by: age of youngest tip (i.e. 2013.67, check XML file in Beauti)
5. Time scale > Scale by factor: -1 (so date goes backwards from tips to root)
6. Tip labels > Display: names
7. Tip labels > Font size: 9, Font: Helvetica
8. Node labels > Display: node ages
9. Node labels > Font size: 8, Font: Arial
10. Node bars > Display: height\_95%\_HPD (this helps to asses the variance in age estimates)
11. Scale axis: reverse
12. Scale axis: label spacing: 10.0
13. Scale axis: tick spacing: 10.0
14. Tip labels=arial, 12
15. Scale bar weight =2 & scale bar range = 3.0
16. If you have specified a trait during Beauti (ie. species or country), turn on node shapes and ‘Colour by’: your trait. Sometimes FigTree will keep with the default colour scheme. So refresh by clicking ‘Colour by’: User Selection and then reselect your trait. Add legend and choose trait for ‘Attribute’ (see figure).
17. Troubleshooting: if you have negative branch lengths in FigTree this is not good. Return to step 17-19 above to correct this.
18. Press ‘Node; button and colour Eurasian (aqua), American (maroon), 8BM (tangerine)