

Phylodynamics and Phylogeography Practical

Dr Samantha Lycett, Roslin Institute, University of Edinburgh

Workshop information

Pathogen sequences (especially viruses) mutate rapidly over time, and this information can be used to infer how the disease is spreading. In an outbreak, virus sequence data can be used to infer possible source dates, locations and possibly species. In this workshop you will learn how to run a basic analysis in BEAST on virus sequences and create and interpret a time scaled tree.

Software:

MEGA

<http://www.megasoftware.net/>

For sequence alignment and simple tree building (note it runs on Windows and Macs, but for Macs there is sometimes a 'Wineskin' issue and I had to have a couple of attempts to get it to run properly)

Tempest

<http://tree.bio.ed.ac.uk/software/tempest/>

For checking clock-like-ness using the trees (non-time scaled) created in MEGA.

BEAST 1.10

<http://beast.community/>

This is the main software for time scaled trees, BEAUTi and TreeAnnotator are also in the download zip / dmg file and these are also required.

Note version 1.8.4 is mostly OK for this too – but BEAST 2 is something different.

[Also note – you need Java if you don't have this already – http://beast.community/install_on_windows]

BEAGLE

<http://beast.community/beagle>

High performance library, speeds up certain calculations within BEAST. Also required to be installed for spatial analysis (and you definitely need this for version 1.10).

Tracer

<http://beast.community/tracer>

To analyse and summarise the BEAST MCMC trace output files

FigTree

<https://github.com/rambaut/figtree/releases> and <http://beast.community/figtree>

To visualise the time scaled tree (and any nexus or newick format tree)

Other software:

Useful for post-run analysis but not covered in detail in this practical:

Spread <https://rega.kuleuven.be/cev/ecv/software/spread> (original)

Spread3 <http://beast.community/spread3> (improved)

Google Earth https://www.google.co.uk/intl/en_uk/earth/

R (3.3.2+) + packages: ape, maps, mapdata, mapproj, OutbreakTools, RgoogleMaps, shiny

Data

Sequences: [H5N1_HA_sel5regions.fas](#)

92 sequences of Highly pathogenic Avian Influenza H5N1, in fasta format. These sequences are the coding region of Hemagglutinin (HA), and the isolates are from Domestic-galliformes, Domestic-anseriformes, Wild-anseriformes in Asia, Africa and Europe from 2003 – 2014.

Traits table: [H5N1_HA_sel5regions_traitsTbl_with_lat_lon.txt](#)

This tab separated data file contains the corresponding 'trait' data for each sequence, contains the discrete Host (bird species type), geographic Region, and Latitude and Longitude values.

Neighbour joining tree: [H5N1_HA_sel5regions_tn93_nj.nwk](#)

A simple neighbour joining tree from the sequences has already been made (see Step 2 for details), and is included here for reference.

BEAST parameter Log files: [H5N1_HA_logfiles.zip](#)

These are the log files from Step 3 – download and unzip before use

BEAST Tree files: [H5N1_HA_1000_trees.zip](#)

These are the tree files from Step 3 – download and unzip before use

MCC tree files: [*_mcc.tre](#)

The Maximum clade credibility tree files with Host, Region or Latitude-Longitude annotations (Step 5)

KML files: [*.kml](#)

For use with Google Earth (Step 6)

Step 1 – Sequence Alignment / Check Sequence Alignment

Data familiarisation: Open the *.fas file in MEGA, there you will see the nucleotide alignment. If you click on the translate tab, you will see the corresponding protein sequence.

These sequences are actually aligned to each other, however to finish the job / check, click on Alignment and then Align by MUSCLE (when in the protein view). In general, if you are looking at influenza, bear in mind that there are sometimes insertions/deletions in Hemagglutinin for highly pathogenic H5 or H7, and in the Neuraminidase stalk region. However, for a within-year within-country single outbreak, or the internal segments protein coding segments of influenza there should not be many (or even any) insertions / deletions.

Click back on the nucleotide tab and (supposing something has changed) save the file by “exporting” as fasta format with the name of *_al.fas (or similar).

<http://www.megasoftware.net/> (link for MEGA software for reference)

Step 2 – Simple trees and molecular clocks

A neighbour joining tree has already been made and included - H5N1_HA_sel5regions_tn93_nj.nwk, but below are the details of how this was done (and incase you want to repeat this yourself). When you have a tree, go to Step 2b.

Step 2a: Neighbour Joining Trees

A neighbour joining tree (the most similar sequences are joined together first) can be created very quickly in MEGA. Re-open the file you made earlier *_al.fas, and chose Phylogenetic analysis (left hand side menu). As with many of these things there are a lot of options, but the ones which are good for this Influenza data are the following – these settings are also applicable for SARS-CoV-2 and most other viruses. If you have bacterial SNP data, or not much diversity in your set, then try the HKY substitution model.

Test of Phylogeny	= none
Substitution model	= Nucleotide
Model / Method	= Tamura-Nei model
Substitutions to include	= d: Transitions and Transversions
Rates among sites	= Gamma distributed
Gamma parameter	= 1
Pattern among lineages	= Different (heterogenous)
Gaps / Missing data treatment	= Pairwise deletion

Note that when you click Phylogenetic Analysis, you will have to move the alignment window out of the way in order to see the other one behind. On this click Phylogenetic analysis again, choose Neighbour joining and then enter the model settings as above (these are Tamura-Nei Model, with transitions and transversions, gamma distributed rates among sites with gamma parameter = 1, different pattern among lineages and pairwise deletion for gaps and missing data treatment. No need to use a test of phylogeny here).

(you can formally test which are the best settings using Likelihood scores, and there are separate programs to do this, e.g. ModelTest. This is different to “test of phylogeny” which is about changing the shape of the tree to find the best one.)

When the tree is built – save it as the *.mts native format and also export as “newick” (use file extension *.nwk).

You can now use FigTree to open the exported newick file (*.nwk) – FigTree will let you change the order that the branches are displayed in on the vertical axis (allowed because it is not changing the underlying tree topology), make the tip labels (sequence names) bigger and also colour them.

<http://tree.bio.ed.ac.uk/software/figtree/> (link for FigTree software for reference)

Step 2b: Adding a time-scale (quick)

Open the supplied or previously generated newick tree file (*.nwk) in TempEst. This program tries a simple method to fit a molecular clock to the tree data, specifically it will try to perform a root-to-tip regression meaning it will fit a straight line to “genetic distance from root” (y-axis) vs time to sampling (x-axis). The genetic distance from root is given by the tree (add branch lengths from sampled sequence to root) and the sampling time is obtained from each sequence name.

The decimal dates have been appended to the end of the sequence names, so you can click “guess dates” to add dates of the sequences into the program.

However, when the tree was made, no particular root was specified; tick the find best root box to make the program search over different possible root positions.

- What is the estimated mutation rate (slope of the root-to-tip fitted line, units are substitutions per site per year) ? and approximately how many mutations is that per year per sequence (HA sequences are 1707 bases long).
- What happens with the different root-to-tip estimating methods – do the trees look different ?

If you notice that one sequence seems out of place in the root-to-tip plot and the residuals plot, highlight it and click to the main tree plot pane to see which one it is. It is likely that this sample contains sequencing errors.

The re-rooted trees can also be exported for further use, click File -> Export tree. Note that this will actually export as a nexus format tree, so give it a *.nex filename when saving.

<http://tree.bio.ed.ac.uk/software/tempest/> (link to TempEst for reference)

Step 3 – Timescaled trees using BEAST

Step 3a: Make the BEAUTi xml

Now the data is prepared, you are ready to make your first BEAST tree. To do this you need to make the *.xml configuration file in **BEAUTi**, and there are a lot of steps and settings, but essentially you work your way along the horizontal tabs at the top.

The ** are important parameters and you would normally try a few of these in an initial analysis

Tab 1 (Partitions) – do File -> Import data : import the fasta *.fas file you made (change the file selector to all files)

Tab 2 (Taxa) – no action

Tab 3 (Tips) – click use tip dates; then Guess dates. The sequence name have been appended with the decimal date, so select 'Defined just by its order' and Order: last

Tab 4 (Traits) – will be used in the next section (no action just now)

Tab 5 (Sites) – Set the nucleotide substitution model, I recommend using the SRD06 model button for this data (this has 2 x HKY models with Gamma site heterogeneity, one model for codon positions 1+2 and a different one for codon position 3) **

Tab 6 (Clock) – Choose between Strict Clock or Uncorrelated relaxed clock with Lognormal distribution **

Tab 7 (Trees) – Choose between Coalescent: Constant Size or Coalescent: Exponential Growth **

Tab 8 (States) – No action

Tab 9 (Priors) – some can be set here (although the defaults in 1.10.4 are often OK)

- clock.rate (if Strict clock) or ucl.d.mean (if relaxed clock): from the TempEst analysis you have an estimated clock rate of around $5e-3$ per site per year, you can use this to set a normal distribution prior for the clock rate if you like (a strong prior). Normal with mean = $5e-3$, std = $3e-3$, range 0-0.1 is OK on this data **

Tab 10 (Operators) – no action

Tab 11 (MCMC) – for a small dataset, the default MCMC parameters are OK (that is the chain length = 10,000,000 and 'log parameters every' = 1000, which gives an output of 10,000 trees and the corresponding log parameters file). For larger data sets increase these values.

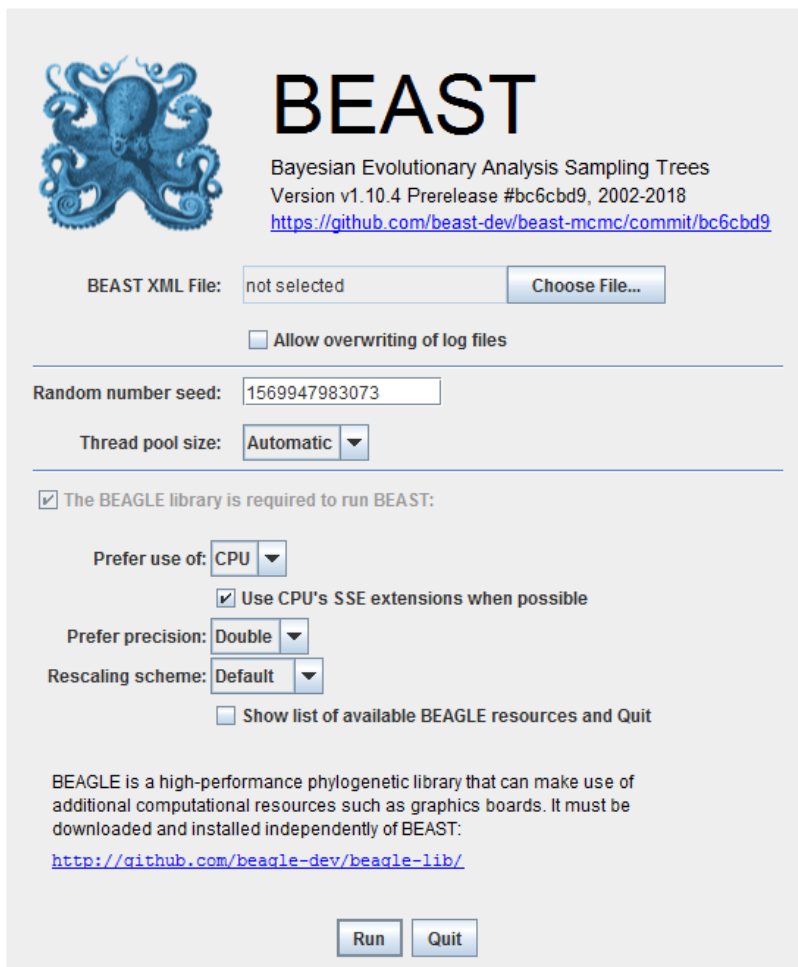
Change the filename stem to reflect the model choices – I use things like this which seem to work well: (fasta file name)_(subst model)_(clock model)_(trees prior model)_(replicate) for example:

- H5N1_HA_sel5regions_SRD06_strict_constPop_1
- H5N1_HA_sel5regions_SRD06_relaxLn_expGrowth_2

Finally you are done, now click the generate BEAST file to make the (filename stem).xml

Step 3b: Run BEAST

Open **BEAST** (double click), you will get an xml selector window - Choose your xml and click RUN !



Actually you can do it from the command line too – there will be a `beast.jar` in the downloaded part, something like this would be OK:

```
java -jar beast.jar your.xml
```

(do `java -jar beast.jar` with no xml to get the BEAGLE & GPU etc run time options)

The 92 sequences will run for a long time if allow to go to completion

A parameters log file (*.log.txt) and trees file (*.trees.txt) will be generated as well as a diagnostic *.ops file (at completion).

I suggest that you stop the run once it has gone for a short while and use the runs I prepared earlier ! – these are H5N1_HA_logfiles.zip (download and unzip before use).

Step 3c: Analyse Log File Output using Tracer

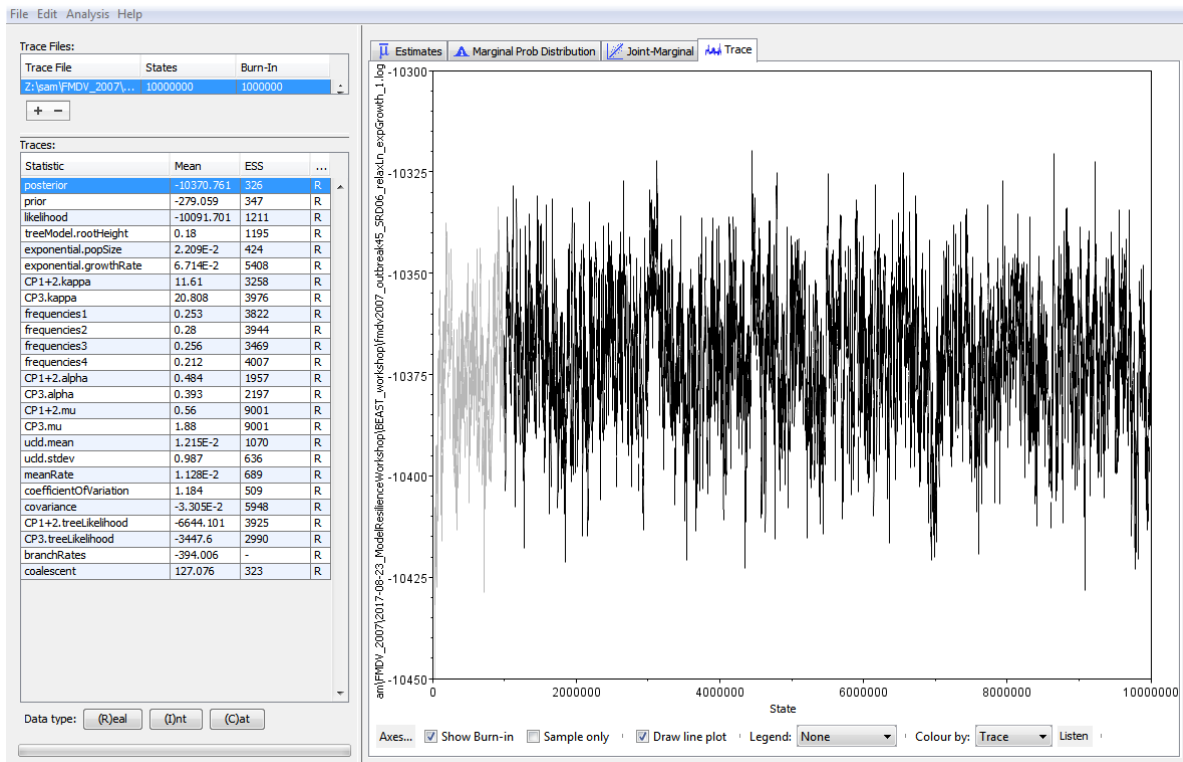
In the unzipped H5N1_HA_logfiles.zip, there are 4 log files – chose one to look at:

Open a *.log.txt file in **Tracer**. For each parameter in the model (and MCMC output) you can see its statistics and trace.

To quickly see if the MCMC is OK, look to see whether the Effective Sample Size of all parameters is ≥ 200 .

Tracer automatically assumes a burn-in of 10% (the greyed out part), but you can change this if you like.

A nice trace looks like this:



From a log file:

- What is the estimated root height ?
- What is the estimated clock rate ?

Note - the estimated start of the epidemic i.e. root-age, or TMRCA is = youngest tip – root height

There the youngest tip of the 92 HA sequences is 2014.342

Step 3d: Make the MCC Tree

The posterior set of trees are in the *.trees.txt file. DO NOT OPEN THIS FILE IT WILL BE TOO BIG, instead use **TreeAnnotator** to summarise the trees for human viewing.

From a default run, there would be 10,000 trees (about 90Mbytes in this case). I have already sampled 1000 trees out of tree numbers 1000-10,000 (i.e. the good 90% after the 10% of burnin has been removed).

These are in file H5N1_HA_1000.trees.zip. Download and unzip it – the resulting file is called H5N1_HA_selected_SRD06_relaxLn_constPop_1.combined.trees.txt

Tree Annotator settings:

- Because I have already removed the burnin, no further trees need to be removed from the file. (If this was not already done then since you would have 10,000 posterior trees, and a burn-in of 10%, then you would specify a burn-in of 1000 trees).
- Also, the default settings of Posterior probability limit (=0), Target Tree Type (Maximum Clade Credibility tree, i.e. the MCC tree), and Node heights (median heights) are good for this data.
- Choose your *.trees.txt file as input, and give it a sensible name for output – I recommend: *.tre
- Click run and create the file (it will take a few moments)

The screenshot shows the TreeAnnotator application window. At the top, there are two radio buttons for specifying burnin: "Specify the burnin as the number of states" (unselected) and "Specify the burnin as the number of trees" (selected). Below the selected option, there is a text input field for "Burnin (as trees)" with the value "0". Further down, there is a text input field for "Posterior probability limit" with the value "0.0". Below that, there are two dropdown menus: "Target tree type" set to "Maximum clade credibility tree" and "Node heights" set to "Median heights". A horizontal line separates these settings from the file selection section. In this section, there are three rows, each with a label, a text input field, and a "Choose File..." button. The first row is "Target Tree File:" with the text "not selected". The second row is "Input Tree File:" with the text "_constPop_1.combined.trees.txt". The third row is "Output File:" with the text "axLn_constPop_1.combined.tre". At the bottom of the window, there are two buttons: "Run" and "Quit".

Step 3e: Viewing the MCC in FigTree

Using the MCC tree file: *.tre (DO NOT USE THE *.trees.txt FILE IT WILL BE TOO BIG), you may now see the results in Fig Tree.

There are a lot of options in FigTree, these are down the left hand side and also in the bar at the top. When you first open the *.tre it will look abit like this:

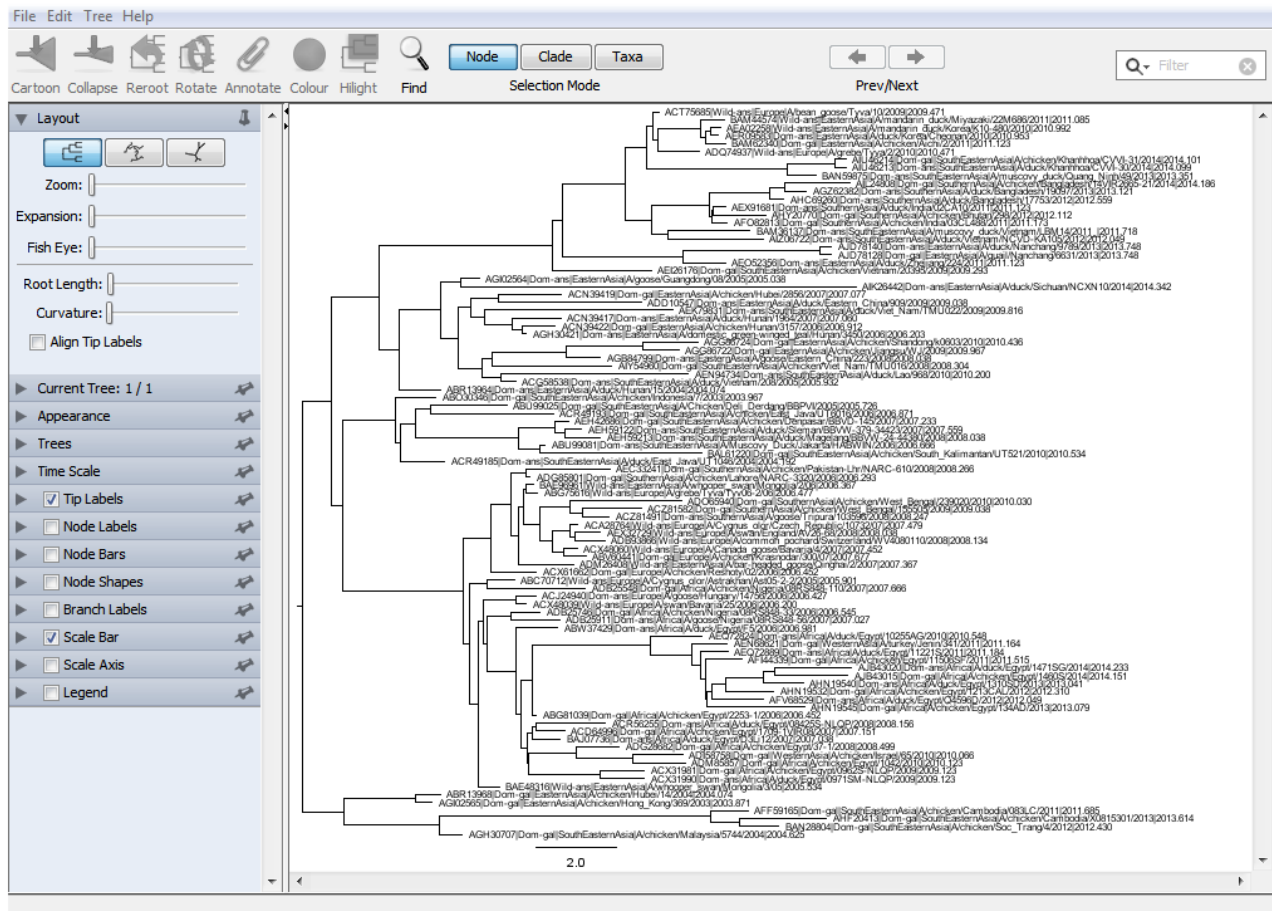
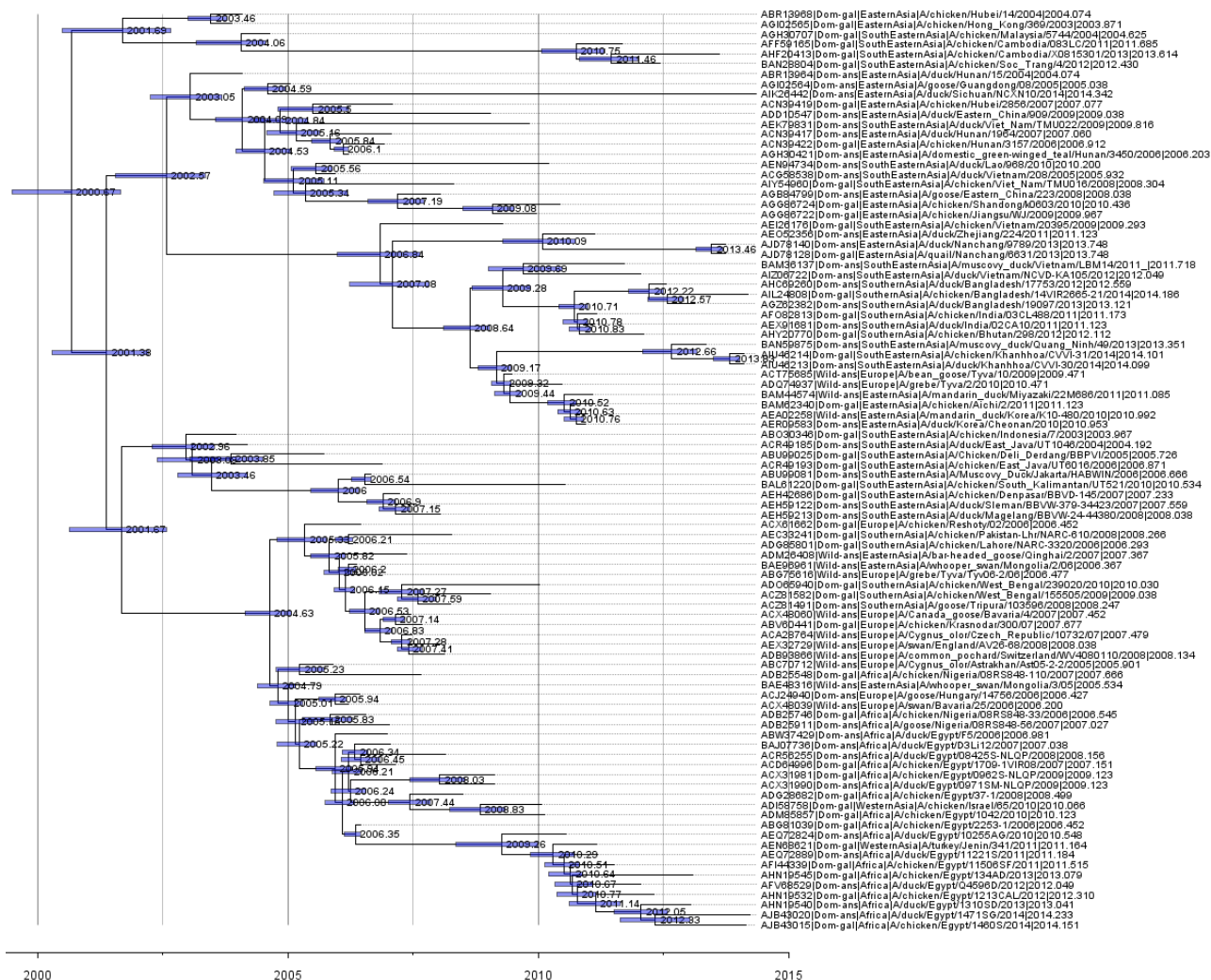


Fig Tree settings:

- Set the time scale – offset by = 2014.342 (the youngest tip)
- Click scale axis – then click reverse axis (increase font size too) - remember coalescent theory, everything goes backwards from the present
- Un-click scale bar
- (Top bars) – Tree: click decreasing node order
- Tip labels – increase the font size, and then click align tip labels (left hand bar)
- Node bars: display 95% HPD
- Node labels: display Node ages, but turn the sig digits down to 2 and make the font arial bold and increase the font size if wanted
- Save this as a *.figTree file
- Export as an image – e.g. *.png



You now have a time-scaled BEAST Tree and image !

Step 4 – Adding traits to the BEAST analysis

Step 3 was about getting the basic time-scaled tree, but now we will add some traits, an example discrete trait (Host) and also a continuous spatial trait (Latitude and Longitude).

I have already done these, and run the analysis to completion, but here are the instructions incase you want to make the xml (otherwise just go to Step 5).

To Configure an XML for Discrete Traits (e.g. Host)

Use BEAUTi, if you have left it open from step 3 then you can just add the new parts (otherwise you will need to repeat step 3 with the appropriate model settings).

Tab 4 (traits): Import the traits file H5N1_HA_sel5regions_traitsTbl_with_lat_lon.txt

- This is a tab separated file with the first column of sequence names, and the other columns of traits. The first row should say: traits Host Region Lat Lon, and when you have imported the traits file you will see the possible traits of Host, Region Lat and Lon.
- Click on Host, then create partition from Host (this will take you to the Partitions tab)

Tab 5 (Sites):

- Select the Host partition and choose Asymmetric model

Tab 8 (States):

- check that the reconstruct states at all ancestors is selected for Host

Now to steps 3b – 3e as before to run BEAST, examine the log parameters file and make the MCC tree.

To Configure an XML for Continuous Traits – Latitude and Longitude

Use BEAUTi, if you have left it open from step 3 then you can just add the new parts (otherwise you will need to repeat step 3 with the appropriate model settings).

Tab 4 (traits): Import the traits file H5N1_HA_sel5regions_traitsTbl_with_lat_lon.txt

- Go to the traits tab, and select Lat and Lon together, then click create partition and re-name it latlon (in my own subsequent R code, it is very important that the exact name latlon is used)

Tab 5 (Sites):

- Select the latlon partition and choose Homogeneous Brownian Motion model, and click 'Bivariate trait represents latitude and longitude', also add a jitter of 0.01

Tab 8 (States):

- check that the reconstruct states at all ancestors is selected latlon partitions

Now to steps 3b – 3e as before to run BEAST, examine the log parameters file and make the MCC tree. As before, if allowed to run to completion this will take a long time; so use my pre-built files for the final parts.

Step 5 – Displaying Trees with Discrete Traits

I have made and run a BEAST analysis with Host-type as discrete trait, and the MCC tree is H5N1_HA_sel5regions_Host_mcc.tre

This tree is like the 'plain' tree, except it also has Host annotations and these can be displayed in FigTree.

In addition to the basic steps (3e), add the Host annotation as colour.

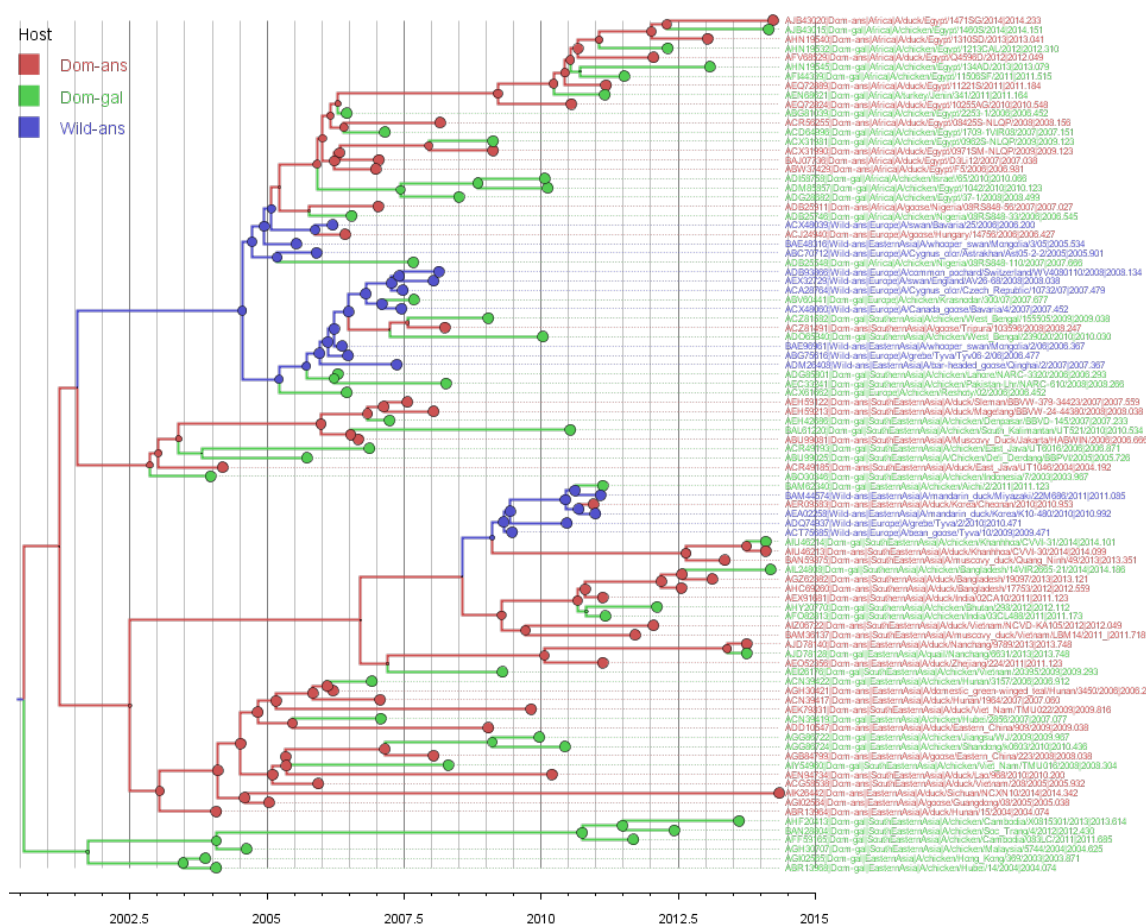
Appearance (left hand side): Colour by Host, and increase the line weight to 2

Align the tip labels

Tip labels (left hand side): Colour by Host

Node Shapes (left hand side): Max size 9, Size by = Host.prob, Min size 0, Colour by Host

Click Legend (left hand side): Attribute Host, increase the font size to 14



Some questions to answer from this tree:

- What does the tree tell you about the origin Host in this case ?
- Which Domestic species transmits to Wild birds ?
- Do the epidemics in Domestic galliformes form a continuous pattern ? or are there multiple introductions ? and which species does these introductions ?

Step 6 - Visualising trees in space using SPREAD/Google Earth

Similarly to the Host tree, BEAST analyses were made and run with Region as a discrete trait, and also Latitude-Longitude as a continuous trait. The MCC trees are:

H5N1_HA_sel5regions_Region_mcc.tre

H5N1_HA_sel5regions_latlon_mcc.tre

The Region tree can be displayed with coloured nodes and branches similarly to the Host tree if wanted.

However, since both the discrete and continuous traits relate to geographic coordinates, it is also possible to convert them into something that can be plotted on a map.

In particular, the programme SPREAD and its successor SpreaD3 is designed to work with BEAST output files. SPREAD takes an MCC tree as input, and outputs a KML format file – which can be viewed using Google Earth. I have already used Spread and made the KML files (see the links if you want to do the tutorials at a later time)

- <https://rega.kuleuven.be/cev/ecv/software/Software>
- https://rega.kuleuven.be/cev/ecv/software/SpreaD3_tutorial

Open the KML files in Google Earth: H5N1_HA_Region.kml & H5N1_HA_latlon.kml

Some questions to answer:

- Do both analyses show the same thing ?
- what about the apparent transmission from Europe back into Asia (discrete) ?
- Does this happen with the continuous ?
- Would increasing the number of discrete states help ?

