Assignment Evolutionary and Quantitative Genetics (I0D53A)

Paper: Worobey *et al*. (2008). Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**: 661-664.

Objective: estimate a time-scaled evolutionary history using BEAST to date the origin of the virus.

Specific tasks and questions:

1. Perform a BEAST analysis on this data set. For this analysis, specify a GTR substitution model with empirical base frequencies and discrete gamma-distributed rate variation, a relaxed molecular clock with an underlying lognormal distribution, and an exponential growth coalescent prior.
    a. Based on the BEAST results, how does your evolutionary rate estimate compare to (i) to the one approximated by TempEst and (ii) to the one reported in Worobey *et al*. (2008)?
    b. What is the estimate for the time of the most recent common ancestor of the HIV-1 sequences used in this study? Rank the subtypes according to their mean age estimates (MRCAs).
    c. Is the clustering between the subtypes consistent between the MCC tree and the ML tree? Are the answers to question 1d for the ML tree the same for this MCC tree?

2. Perform a BEAST analysis with the same settings on this data set, but now estimate the age of both the 1960 sequences and the 1959 sequences (cfr. last tip below). In addition, because only short stretches are obtained for the 1960 sequences, constrain these sequences to be monophyletic with the subtype A sequences. Is this a reasonable assumption?
3. Does this affect the time estimate for the MRCA of the tree? And the evolutionary rate?
4. How accurately are the ages for the 1959 and 1960 sequences estimated?

Tips:

- Use the last version of BEAST for phylogenetic inference. The input file can be constructed using BEAUti, a Java interface provided with BEAST, and the output of BEAST runs can be diagnosed (and continuous parameters be summarized) using Tracer). A BEAST analysis on a relatively large data set can take a considerable amount of time to converge for all parameters. Ensure that your BEAST runs have converged (stationarity of trace plots, ESS values for continuous parameters, multiple runs if computationally feasible). A maximum clade credibility (MCC) tree can be summarized from the posterior trees file using the program TreeAnnotator, and visualized using FigTree, two other Java interfaces provided with BEAST.
- To estimate tip ages, a taxon set needs to be defined first in the Taxa panel in BEAuti for these taxa. At the bottom of the Tips panel, the option ‘Tip date sampling’ can then be set to ‘Sampling with individual priors’, and applied to the specified taxon set. The prior on this taxon age needs to be set in the priors panel; specify a uniform distribution between 0 and 100 years.

Files to upload:

- An answer sheet that provides sufficiently detailed but to-the-point answers to the specific questions. Including illustrations (e.g. trees, Tracer plots) to support your answers could be particularly useful.
- BEAST XML input files for both analyses.
- BEAST log file for both analyses.
- BEAST MCC trees for both analyses.