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.

> Human immunodeficiency virus type 1 (HIV-1) sequences that pre-date the recognition of AIDS are critical for accurately timing the origin of the AIDS pandemic because they provide the required 'anchor points' for evolutionary reconstructions. Worobey *et al*. (2008) report the amplification and characterization of viral sequences from a Bouin's-fixed paraffin-embedded lymph node biopsy specimen obtained in 1960 from Kinshasa in the Democratic Republic of the Congo (DRC), which is only the second HIV sequence obtained from a pre-1976 isolate. Through phylogenetic analyses, the authors use the newly obtained DRC60 sequences (there are two independent replications for this sample) and a previously obtained ZR59 sequence together with other sequences sampled through time to date the most recent common ancestor of the M group near the beginning of the twentieth century.

Data: We provide a Nexus file - Worobey.1960.all.cc3.nex - with the sequence alignment of the phylogenetic analyses performed by Worobey *et al*. (2008). No further aligment needs to be done. In addition to the DRC60 sequences and the ZR59 sequence, the data set contains representatives of subtype A, B, C, D, F, H, J and K, as well as a few untypable strains ('U'). Subtypes and sampling years are included in the sequence names.

Objective: Examine the phylogenetic relationships, explore the temporal structure using TempEst, and estimate a time-scaled evolutionary history using BEAST to date the origin of the virus.

Specific tasks and questions:

1. Reconstruct a maximum-likelihood tree (PhyML) tree using a GTR model of evolution and gamma-distributed rate variation among sites. Except for cluster support, most questions can be addressed using this tree, including the TempEst analysis below. To assess cluster support, repeat the same maximum likelihood inference with 100 bootstrap replicates, which will require a long run time. Root the tree using midpoint rooting in FigTree and answer the following specific questions:
    a. Do the subtypes form monophyletic clades? If so, is there good support for this?
    b. Without doing detailed calculations, which subtype do you expect to have the highest diversity: subtype A or subtype C?
    c. Which sequence is most similar to the 1959 strain? What is the divergence between these two sequences?
    d. Do the 1959 and the two 1960 sequences fall in a subtype cluster. If so, which one? If not, to which subtype is/are the sequence(s) most closely related?
2. Use the reconstructed maximum-likelihood tree for a TempEst analysis and answer the following questions:
    a. Is there a temporal signal in this HIV-1 data set?
    b. Does the best-fit root under the R-squared function correspond to the midpoint rooting used above?
    c. What would be the point estimate of the evolutionary rate based on this regression analysis?
    d. Do the 1960 sequences and the 1959 sequence show more or less divergence as expected for their sampling dates based on the fitted regression line? Should they be considered as outliers?
3. 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?

4. 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?).
5. Does this affect the time estimate for the MRCA of the tree? And the evolutionary rate?
6. How accurately are the ages for the 1959 and 1960 sequences estimated?

Tips:

- Simple distance-based tree reconstruction methods such a NJ and/or BioNJ sometimes have difficulties handling datasets that contain (very) divergent sequences or many gaps. Using maximum-likelihood inference in SeaView should not be a problem.
- Bootstrapping maximum likelihood trees can be time consuming. Therefore, it could be interesting to first infer the tree without bootstrapping (this tree can then already be used for the TempEst analysis).
- Temporal signal can be evaluated with TempEst (tree.bio.ed.ac.uk/software/tempest). This software requires phylogenetic trees with branch lengths measured in genetic distance (and not in time units). Such trees can be inferred using the phylogenetic methods implemented in SeaView. Do not restrict yourself to a single function to find the best root, but explore different ones.
- 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.

Files to upload:

- An answer sheet that provides sufficiently detailed but to-the-point answers to the specific questions. Including illustrations (e.g. trees, root-to-tip divergence plots) to support your answers could be particularly useful.
- PhyML tree with bootstrap values (in Newick or nexus format).
- BEAST XML input files for both analyses.
- BEAST log file for both analyses.
- BEAST MCC trees for both analyses.