

Viral Bioinformatics 2022

Topic: Wrap up of the day!

Instructors: Marta Giovanetti and Carolina Torres

Hands-on exercises

1. Scientific question:
 - 1.1 Reconstruct the phylogenetic/dynamic history of our query sequence
Lab_Salvador|2016
 - 1.2 Genotyping assignment
 - 1.3 Temporal Signal
 - 1.4 tMRCA estimation
 - 1.5 Data visualization
2. Perform multiple genome alignment of ZIKV isolates (ZIKV.fasta) using MAFFT.
Input dataset:
LAC-2022/Typing-tool_Data-Visualization/1.Viral_typing_tool/2.Hands-on/ZIKV.fasta
3. Use the genome alignment of ZIKV isolates and generate whole genome phylogenetic tree with the help of IQTREE. Select the best nucleotide substitution model that fits the data using ModelSelector. Reconstruct maximum likelihood-based phylogeny with 1000 bootstrap replicates using UltraFast method available in IQTREE. Which nucleotide model best fits the data provided? Using the consensus tree, find out the number of clusters in which the isolates are observed.
4. Using the consensus tree generated with IQTREE, check the presence or absence of temporal signal in the ZIKV data with the help of TempEst. What can be the initial value for nucleotide substitution rate? What about the R squared and the Coefficient Correlation?
5. **Optional Step (Might be time-consuming).**
Estimate the genome-wide nucleotide substitution rate of ZIKV dataset using BEAST package. With the help of 'BeauTi' tool, choose GTR+I+G as the nucleotide substitution model. Set molecular clock to 'uncorrelated lognormal distribution' with demographic model as 'skyline. MCMC to be set to 10 million steps with log at every 10,000 steps. Generate an xml file with all the parameters set and use this as input to run 'beast'(Takes 10-11 minutes on 8GB laptop or desktop). After the beast run, two files are obtained namely, log file and tree file. Check convergence of the log file using Tracer (Hint: ESS values to be greater than 200 for every parameter). (If convergence is obtained, run the same in triplicate and combine the log files using the tool 'logcombiner'). If convergence is not obtained, then increase the MCMC steps to 50 million and repeat the same in triplicate. Generate the maximum clade credibility tree using the tool 'treeannotator' with trees file as input and visualise the same using FigTree.

6. Perform lineage assignment of the ZIKV isolate collected in the state of Bahia strain label: **Lab_Salvador2016**
7. Play a bit with your imagination, present your trees and your Root-to-tip regression (Tempest) in a Fancy way... just be creative!
8. Try to discuss your results!

Have fun!

Marta & Carolina