Phylogenetics

Introduction to the hands-on exercises

DATASET 1: SARS-CoV-2

SARS-CoV-2 diversity and evolution are reflected by both variants and lineages. The Pango nomenclature is a system for identifying SARS-CoV-2 genetic lineages of epidemiological relevance and is being used by researchers and public health agencies worldwide to track the transmission and spread of SARS-CoV-2, including variants of concern (VOCs) (https://www.pango.network/).

The **objective** of the practical activity is to assign the Pango lineage to a group of SARS-CoV-2 sequences obtained in Latin America.

For that purpose, we have built a dataset of complete genome sequences consisting of 40 reference sequences of several Pango lineages (indicated at the beginning of the sequence names) and 10 sequences from some Latin American countries (indicated as "query" at the beginning of the sequence names).

DATASET 2: Dengue virus (DENV)

Dengue infections are caused by four closely related viruses (DENV-1 to DENV-4). DENV-4 is classified into four genotypes (I-III and Sylvatic).

The **objective** of the practical activity is to genotype three DENV-4 isolates that circulated in Brazil in 2012-2013 ("query") and determine if they belong to a single transmission chain.

For that purpose, we have built a dataset consisting of 37 complete genome sequences of this virus including **reference sequences of the DENV-4 genotypes** (indicated at the end of the sequence names as I-III and Sylvatic), **the sequences of interest** (indicated as "query" at the beginning of the sequence names) and **sequences of DENV-1 to DENV-3 as outgroup**.

Activity I: Multiple Sequence Alignment (MSA)

Specific objectives of this practice:

- To become familiar with the MAFFT and Aliview programs.

- To obtain a sequence alignment for later use in phylogeny.

To carry out an alignment, the sequences need to be available in a file that can be read by the programs. In general, the FASTA format is accepted by most sequence alignment and edition programs. The time required for the analysis will depend on the power of the computer and the number of sequences to be analyzed. In general, it can be estimated that the computation time will increase linearly with the length of the sequences, and exponentially with the number of sequences to be aligned.

The datasets (unaligned set of sequences: `LAC_SARSCoV2.fasta / LAC_DENV.fasta`) are located in:

```
/home/manager/course_data/Phylogenetics_methods_and_tree_building/LAC/1_SARSCoV2/
LAC_SARSCoV2.fasta
```

```
/home/manager/course_data/Phylogenetics_methods_and_tree_building/LAC/2_DENV.fasta
```

Exercise

A. Alignment with MAFFT

MAFFT is an advanced tool that can align using different alignment algorithms for different applications such as L-INS-i (accurate; recommended for <200 sequences), FFT-NS-2 (fast; recommended for >2,000 sequences), etc. It can be run locally or on online servers. To understand the algorithms and their use cases, please refer to https://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html
To use it on the VM, type `mafft` on the command-line, `mafft --help` will give you information about the proper syntax.

Please note that the procedure below is for the SARS-CoV-2 dataset, so if you will be analyzing the Dengue dataset as well, you need to go to the directory where that dataset is located and replace the file names in the instructions below.

1. Open a Terminal and go into the directory that contains the dataset to align: `LAC_SARSCoV2.fasta`

```
/home/manager/course_data/Phylogenetics_methods_and_tree_building/LAC/1_SAR
SCoV2/LAC_SARSCoV2.fasta
```

2. Type: `mafft --auto LAC_SARSCoV2.fasta > LAC_SARSCoV2_aln.fasta`

Usage:
```
mafft [options] input > output
```

`--auto` automatically switches algorithm according to data size.

[OPTIONAL] Alignment with Muscle (in Aliview).

1. Execute Aliview and open the alignment: File → Open File: **LAC_SARSCoV2.fasta**

2. Explore the Aliview window and locate the following elements: Sequence names, Sequences, Ruler.

3. Perform an alignment with the default program (Muscle): Align → Realign everything → OK.

[The program will start, and different steps will be shown. Once the alignment is completed, the output file will be automatically shown.]

4. Check the alignment and realign regions if needed:

   Select the region to realign → Align → Realign selected block.

5. Save this alignment: File → Save as fasta → **LAC_SARSCoV2_muscle_aln.fasta**

B. Editing the sequence alignment

After aligning, it is advisable to visually review the sequence alignments obtained before proceeding to phylogenetic analysis. Sometimes, a manual edition is needed, especially at the ends of the alignment, where only some sequences have reliable information. For this exercise, you can use any of the alignments generated from MAFFT or Aliview (Muscle).

The following instructions show the use of the alignment of SARS-CoV-2 generated with MAFFT.

Edition in Aliview:

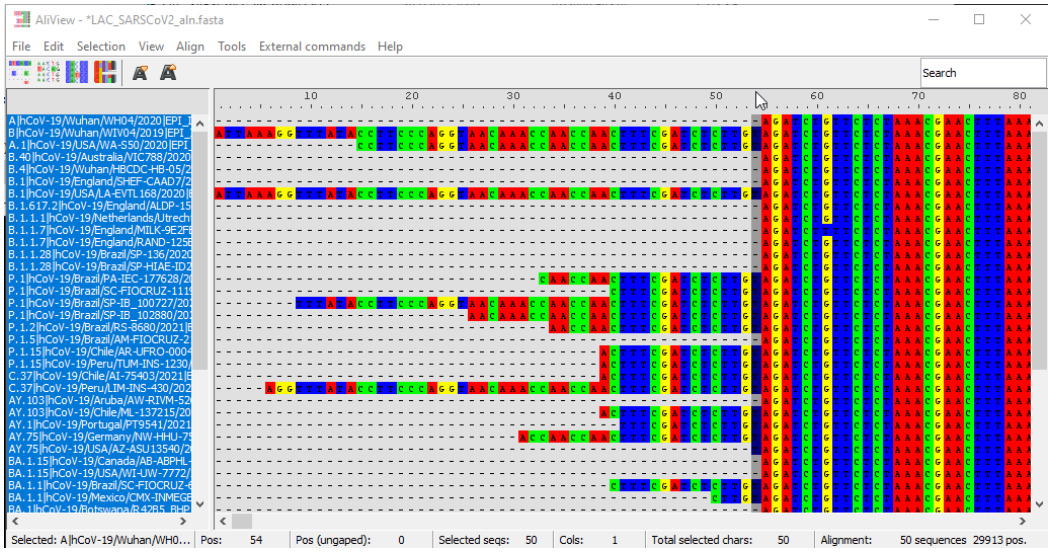
1. Execute Aliview and open the alignment: File → Open File: **LAC_SARSCoV2_aln.fasta**

2. Check the alignment and realign regions if needed:

   Select the region to realign → Align → Realign selected block.

3. Select the region to be deleted:

   a) For the left end of the alignment, select the last nucleotide of the region to be deleted (as in the figure below):



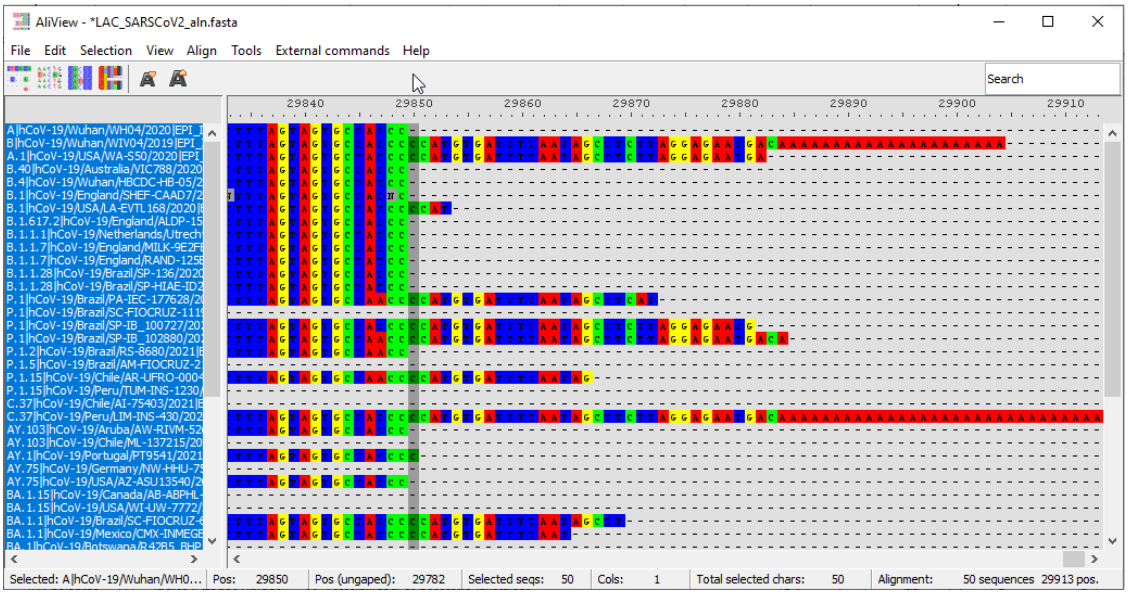
   - Select → Expand Selection Left

   - Edit → Delete selected

b) Repeat the procedure for the right end, select the first nucleotide of the region to be deleted (as in the figure below):

- *Select → Expand Selection Right*

- *Edit → Delete selected*



c) Repeat the procedure for internal regions if needed:

- *Select the region → Edit → Delete selected.*

4. Save the edited alignment:

- File → Save as fasta → **LAC_SARSCoV2_aln_edit.fasta**

This file will be used to estimate the substitution model and infer the phylogenetic tree.

Activity 2: Phylogenetic analysis

Specific objectives of this practice:

- To become familiar with the IQ-TREE and Figtree programs.

- To build a Maximum Likelihood tree to assign the Pango lineage to a group of SARS-CoV-2 sequences obtained in Latin America.

- To build a Maximum Likelihood tree to genotype three DENV-4 isolates that circulated in Brazil in 2012-2013 and determine if they belong to a unique transmission chain.

Datasets to use: **`LAC_SARSCoV2_aln_cut.fasta`** (or `DENV_aln_cut.fasta`), from the practical activity 1.

Introduction to the IQ-TREE program:

This program allows you to perform phylogenetic analysis by Maximum Likelihood. It uses efficient algorithms to explore the tree space, allowing very large matrices to be analyzed with reliable results (hundreds or thousands of sequences). It allows estimating the evolutionary model (ModelFinder module) followed by the phylogenetic inference, and implements support measures to evaluate the reliability of the groupings or branches (Bootstrap, Ultrafast Bootstrap Approximation and probabilistic contrasts). The program can be downloaded and run locally (http://www.iqtree.org/), or on online servers such as http://iqtree.cibiv.univie.ac.at/ | https://www.phylo.org/ | https://www.hiv.lanl.gov/content/sequence/IQTREE/iqtree.html

You can find many basic and advanced tutorials at http://www.iqtree.org/doc/

Please note that the procedure below is for the SARS-CoV-2 dataset, so if you will be analyzing the Dengue dataset as well, you need to go to the directory where that dataset is located and replace the file names in the instructions below.

Exercise

A. Phylogenetic Inference by Maximum Likelihood with IQ-TREE

Phylogenetic inference + support (Ultrafast Bootstrap Approximation + SH-aLRT)

1. Open a Terminal and go into the folder that contains the edited alignment to analyze: `(~/course_data/Phylogenetics_methods_and_tree_building/LAC/1_SARSCoV2/LAC_SARSCoV2_aln_cut.fasta)`
2. Type: `iqtree2 -h` (this command allows you to see all available options, check those that you will use in the next step).
3. Type: `iqtree2 -s LAC_SARSCoV2_aln_cut.fasta -m MFP -B 10000 -alrt 1000`

`-s` to specify the name of the alignment file, always required by IQ-TREE to work.

`-m` to specify a model selection strategy (if no option is specified, `-m MFP` is used by default).

`-B` to specify the number of replicates for Ultrafast Bootstrap Approximation in IQ-TREE v2.

`-alrt` to specify the number of replicates for SH-aLRT.

Once the process is finished, the output files will be found in the folder, including:

.treefile: the ML tree in NEWICK format, which can be visualized by any supported tree viewer programs like FigTree.

.iqtree: the main report file that is self-readable. You should look at this file to see the computational results. It also contains a textual representation of the final tree.

.log: log file of the entire run (also printed on the screen).

Questions

1. Which is the best-fit evolutionary model for this dataset according to the Bayesian information criterion (BIC)? (open the ".iqtree" or the ".log" files with TextEditor)
2. What parameters does the best-fit model have?

B. Tree visualization

1. Open the FigTree program → In the terminal, type: `figtree`

   File → Open → select the file ***LAC_SARSCoV2_aln_cut.treefile***

3. Select a name for annotated values: "SH/UFB"
4. In this exercise, it is recommended to root the tree in the branch that connects lineage A with lineage B:

   Select the branch → Reroot.

5. It is recommended to order the tree to improve its visualization:

   Tree → Increasing Node Order

6. Display support values: Branch labels → *Display "SH/UFB".*
7. Annotate and color the tree according to the locations:

   -Menu File → Import annotations → load the file "LAC_SARSCoV2_location.txt" from its folder.

   -Menu Tree → Annotate Nodes from tips → Annotation: Region.

   -Appearance → Colour by → Region

   -Tip Labels → Colour by: Region

   -Legend → Click to show the legend → Attribute: Region.

   In addition, you can modify the size of the fonts (in Tip Labels, Legend, etc).

Activity:

1. Assign a Pango lineage to the "query" sequences of SARS-CoV-2.
2. To what genotype of DENV-4 do the "query" sequences belong? Are they part of a single transmission chain?
3. DENV-4 sequences from Brazil are monophyletic?

C. [Optional] Additional activity: Phylogenetic signal

IQ-TREE implements the likelihood mapping approach (Strimmer and von Haeseler, 1997; https://doi.org/10.1073/pnas.94.13.6815) to assess the phylogenetic information of an input alignment. The detailed results will be printed to .iqtree report file. The likelihood mapping plots will be printed to .lmap.svg and .lmap.eps files.

To perform a likelihood mapping analysis (ignoring tree search) with 2000 quartets for the alignment `SARSCoV2_aln_cut.fasta` with a model being automatically selected, create a new directory "PhyloSignal" and paste within the file `SARSCoV2_aln_cut.fasta`.

Open a terminal in that directory and type:

```
iqtree2 -s SARSCoV2_aln_cut.fasta -lmap 2000 -n 0 -m MF
```

`-lmap` Specify the number of quartets to be randomly drawn. If you specify -lmap ALL, all unique quartets will be drawn, instead.

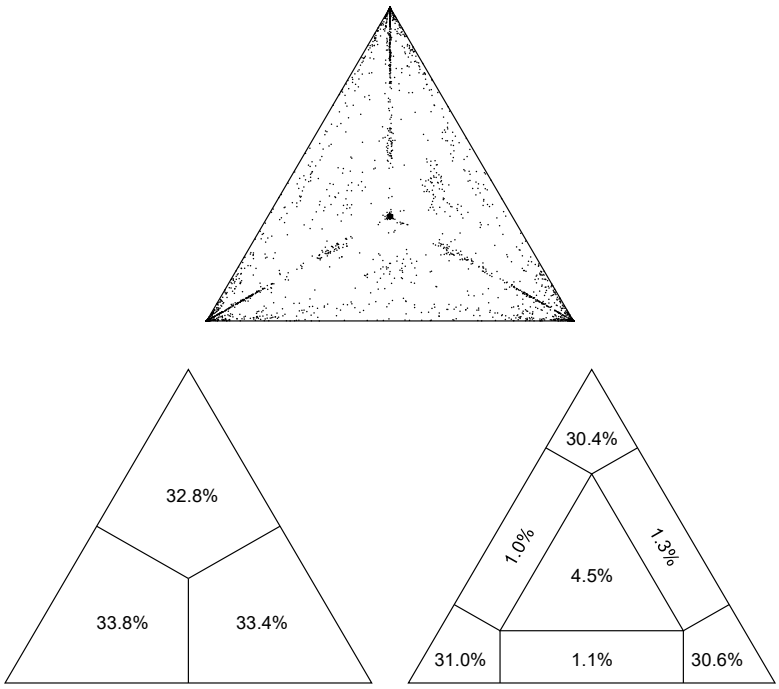
[TIP: The number of quartets specified via -lmap is recommended to be at least 25 times the number of sequences in the alignment, such that each sequence is covered ~100 times in the set of quartets drawn.]

`-n 0` Skip subsequent tree search, useful when you only want to assess the phylogenetic information of the alignment.

[Note that if you already have selected an evolutionary model from a previous analysis with this dataset, you can specify it in the command option -m, for example: `-m TIM2+F+I+G4]`

You can now view the likelihood mapping plot file **SARSCoV2_aln_cut.lmap.eps (or .svg file)**, which shows phylogenetic information of the alignment **SARSCoV2_aln_cut.fasta**.

The figure will look like this:



On the top: distribution of quartets depicted by dots on the likelihood mapping plot.

On the left: the three areas show support for one of the different groupings like (a,b)-(c,d).

On the right: quartets falling into the three corners are informative.

Quartets in the three rectangles are partly informative and those in the center are uninformative. **A good data set should have a high number of informative quartets and a low number of uninformative quartets.** The meanings can also be found in the LIKELIHOOD MAPPING STATISTICS section of the report file .iqtree.