RSV Consensus Genome Analysis

10th of December, 2024

Practical Exercise of

RSV Insights: Genomic Evolution, Treatments, and Public Health Strategies

This exercise will ensure that participants gain hands-on experience of real-world RSV genome analysis workflows using widely available online tools. For both **RSVA** and **RSVB**, participants will focus on **quality control**, **clade assignment**, **diversity/evolutionary history analysis** and **resistance mutation identification**. Participants will use online versions of **Nextclade**, **RSVSurver** and **NGPhylogeny.fr**.

Participants will be given two sets of datasets. The first ones are comprising of "contextual" sequences that gives a good representation of RSVA and RSVB diversity:

- **RSVA_contextual.fasta**
- **RSVB_contextual.fasta**

The second ones are comprising of "query" sequences that must be processed with sequences from the first dataset:

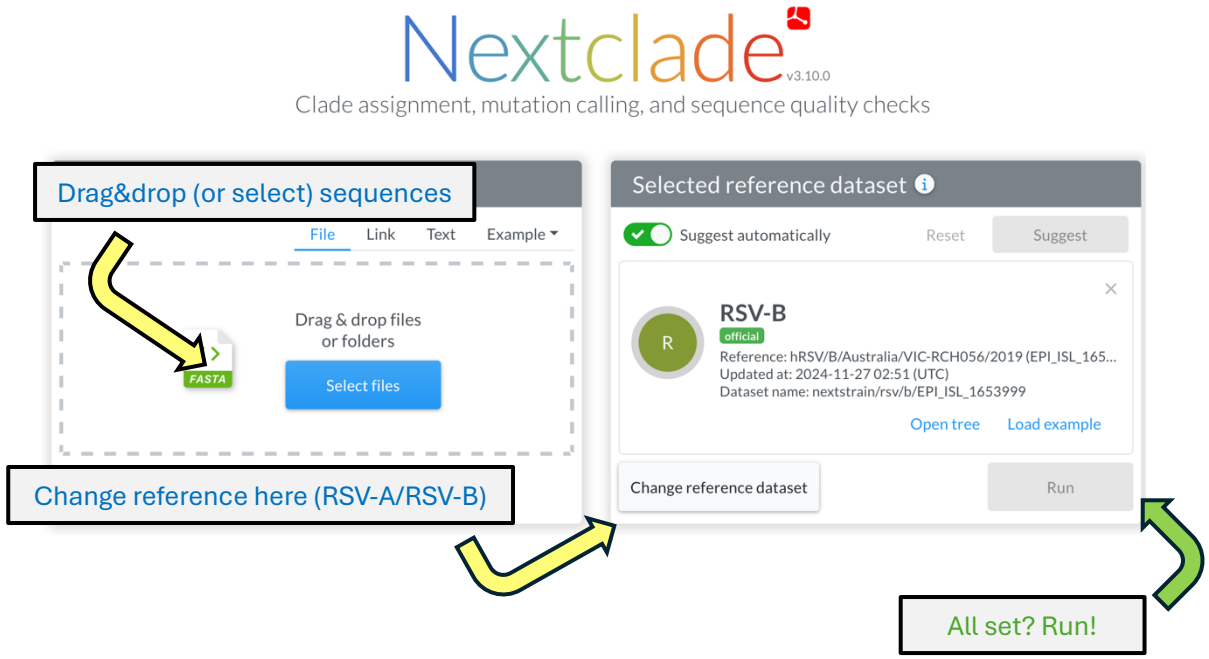
- **RSVA_queries.fasta**
- **RSVB_queries.fasta**

Part 1: Contextual datasets analysis

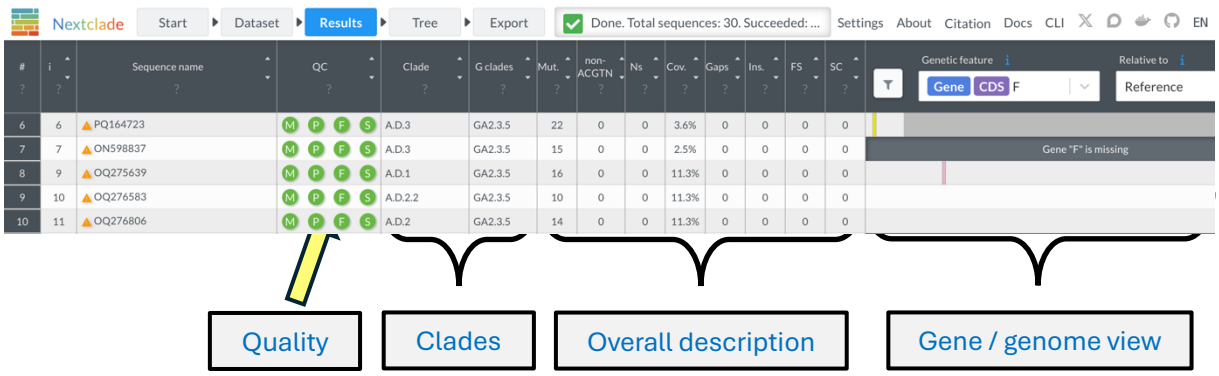
The first datasets correspond to fasta files ***RSVA/B_contextual.fasta*** containing RSVA/B sequences. Your task here is to select sequences that are suitable for further analysis. You should use the online tool to Nextclade (https://www.nextclade.org) to do this:

1) Overall quality such as coverage of the whole genome and of the G / F genes, frameshits, stop codons, accumulation of private mutations, high levels of ambiguous bases, ...
2) Determine clades and verify that your dataset is a good representation of RSV diversity.
3) Prepare your dataset for the next step.

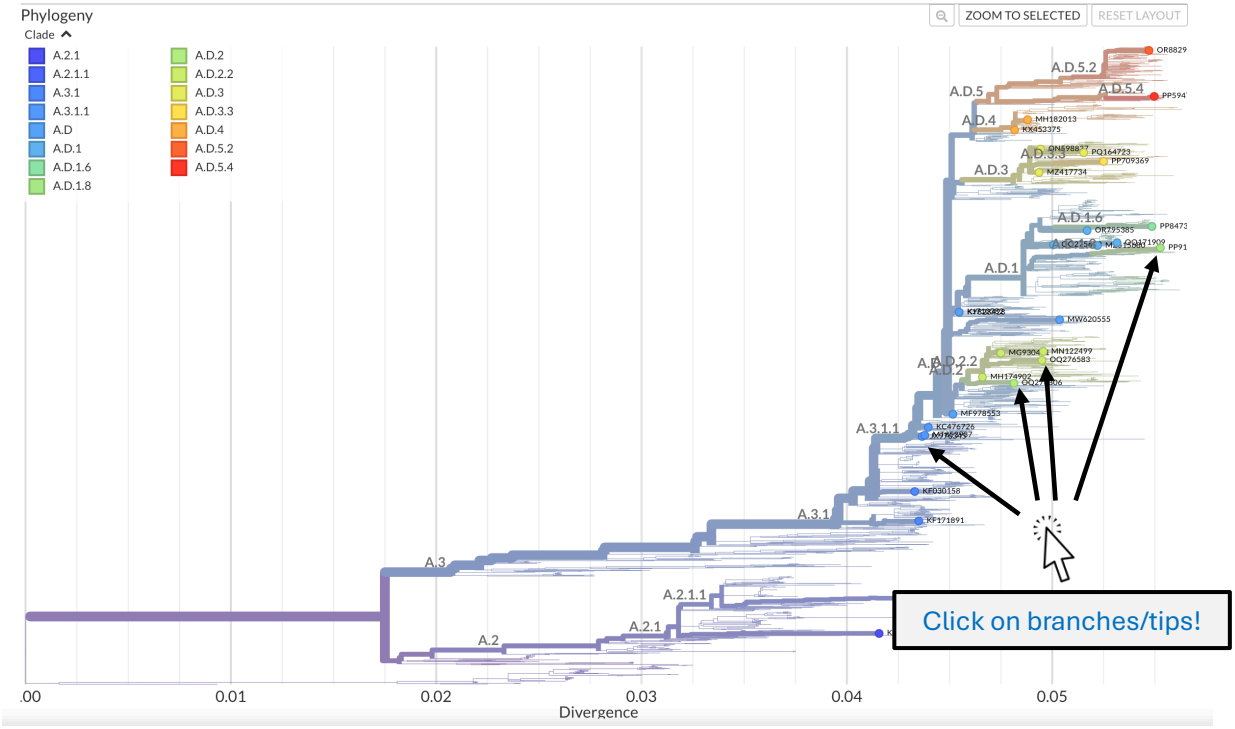
1. **Upload the Dataset**:
    - Go to Nextclade (https://www.nextclade.org).
    - Upload the provided ***RSVA/B_contextual.fasta***.
    - Since Nextclade proposes different references for each RSV, you must launch the analysis multiple times (one time for each reference).



2. **Sequence Quality Control**:
    - Review the sequence quality scores. Identify any sequences with significant issues (e.g., high number of missing bases, ambiguous bases). Sequences with an overall score "bad" (red ones) should be rejected.
    - Check for non-covered regions, focusing on the **G** and **F genes**. Document sequences with significant loss of information on these regions. A sequence without the F gene should not be kept.

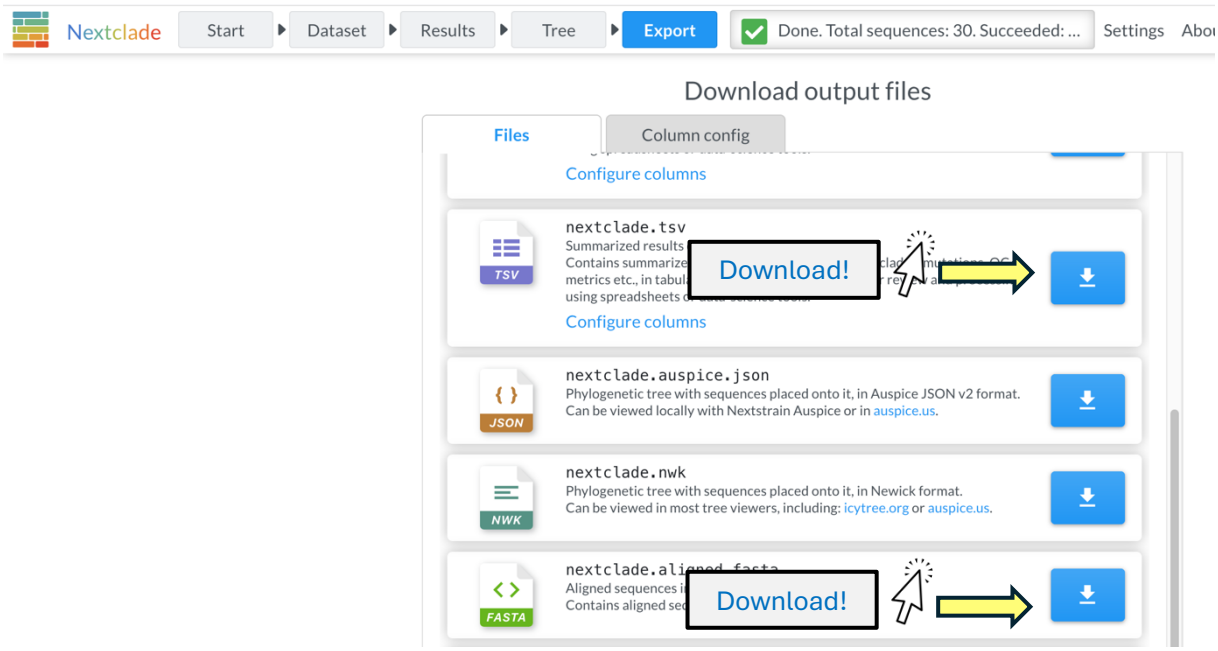


3. **Determine Clades**:
    - Analyse the clades assigned to the sequences and check that all major clades are well represented. You can do this by looking at the assigned clades in the Results tab, but you can also visualise the phylogenetic placement in the Tree tab to check for diversity. You can click on tips / branches to get more information.
    - If possible, document missing clades.



4. **Save Curated Contextual Datasets**:

   For each dataset, save both ***nextclade.tsv*** and ***nextclade.aligned.fasta*** files.
   - ***nextclade.tsv*** contains all results that you can see in the “results” tab.
   - ***nextclade.aligned.fasta*** contains the sequences that have been aligned to the reference (nextalign).
   - Save these files for both RSVA and RSVB analyses.



Questions

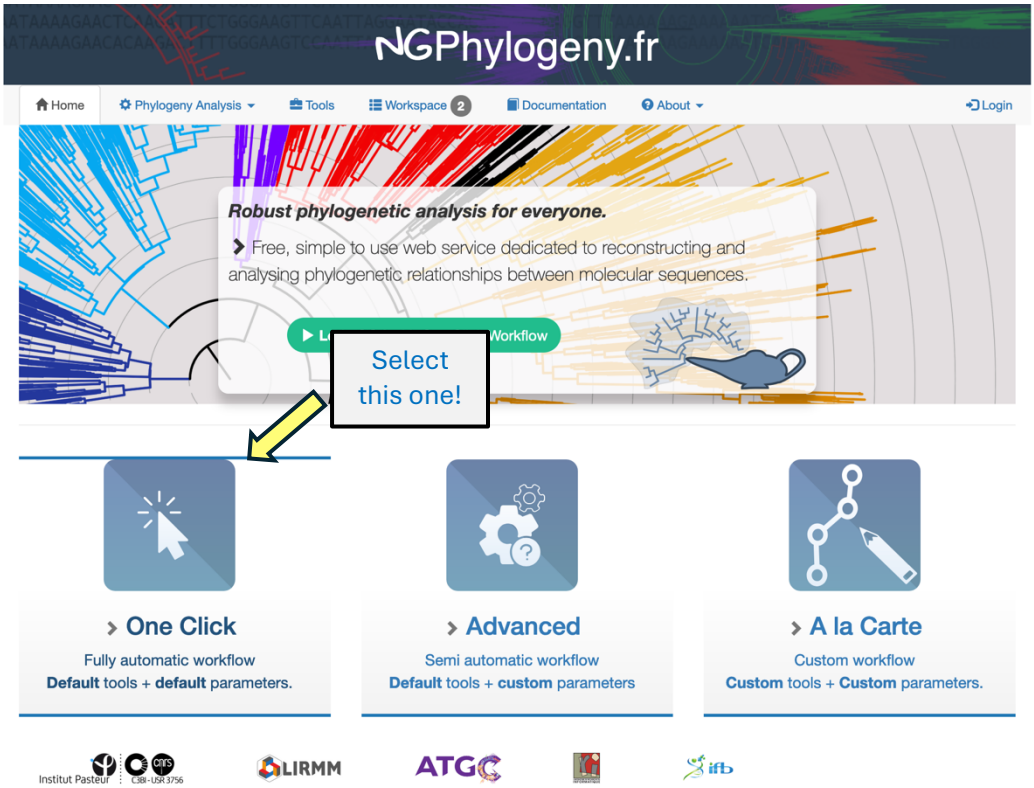
- Did both context files have any "bad" sequences? If so, which ones? What were the reasons for classifying them as such?
- Do you think the data sets represent the diversity of RSVs well?
- Are there any pre-duplication individuals in the data set? How do you recognise them?

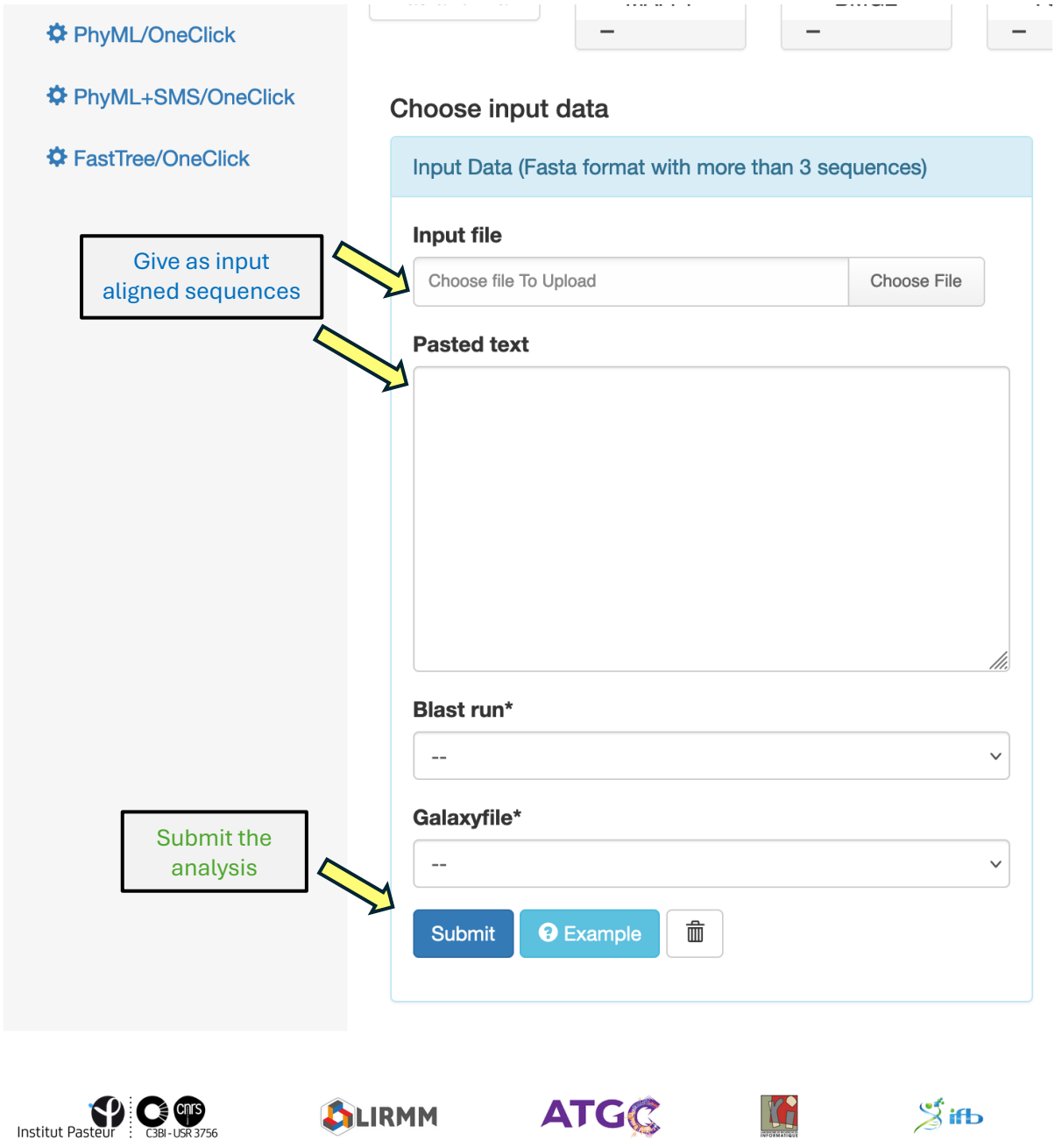
Part 2: Clade placement and DRMs for New Sequences

In this exercise, you will determine the clades of few sequences contained in the files ***RSVA/B_queries.fasta*.** But this time, instead of using Nextclade, you will use NGPhylogeny.fr (https://ngphylogeny.fr/) webserver.

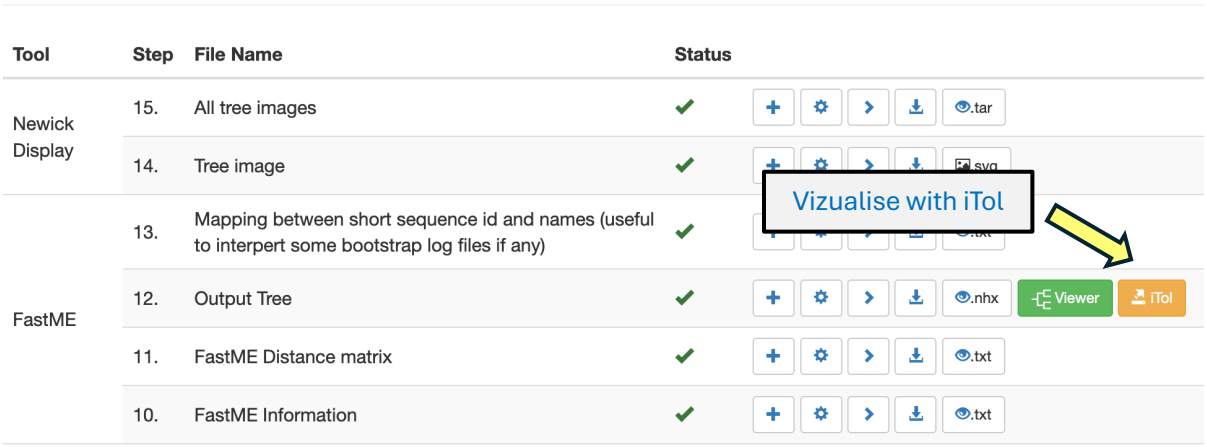
In addition, you will determine the presence of drug-resistance mutations in these individuals by using the tool RSVSurver (https://rsvsurver.bii.a-star.edu.sg/) as well as the Virus French Resistance database https://virusfrenchresistance.org/virus-french-resistance-rsv/.

1. **Construct Phylogenetic Trees**:
    - Upload the merged files (contextual + queries) to NGPhylogeny.fr (https://ngphylogeny.fr/) to align these sequences and build phylogenetic trees.
        - Perform a “One-click” analysis.
        - Use the default settings for tree construction. In this step, sequences are aligned and then used to infer a phylogeny.
    - Analyse the phylogenetic trees to determine the clades for the new sequences. To do this, check where your query sequences sit in relation to contextual sequences (for which you already know the clade).

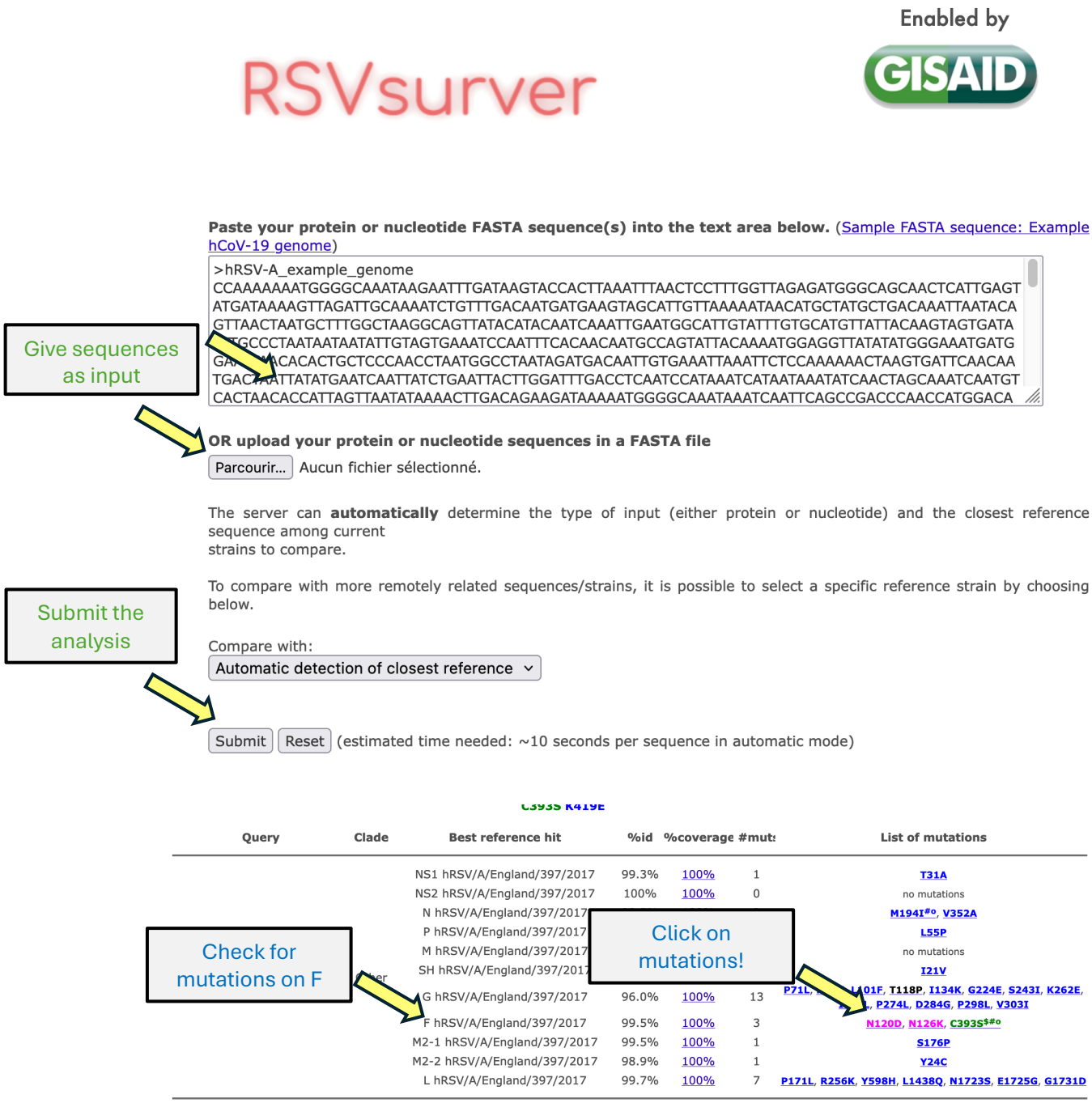



PhyML/OneClick
PhyML+SMS/OneClick
FastTree/OneClick
Choose input data
Input Data (Fasta format with more than 3 sequences)
Give as input aligned sequences
Input file
Choose file To Upload
Choose File
Pasted text
Blast run*
--
Galaxyfile*
--
Submit the analysis
Submit
Example

Institut Pasteur
C3BI - USR 3756
LIRMM
ATGC
ifb


Tool
Step
File Name
Status
Newick Display
15.
All tree images
.tar
14.
Tree image
.svg
Vizualise with iTol
FastME
13.
Mapping between short sequence id and names (useful to interpert some bootstrap log files if any)
12.
Output Tree
.nhx
Viewer
iTol
11.
FastME Distance matrix
.txt
10.
FastME Information
.txt

2. **Check for Resistance Mutations**:
   - Focus on key resistance-related mutations in the F gene. Look at the clade assignment and specific mutation annotations (if available in the tree or sequence metadata). Use the RSVSurver software (https://rsvsurver.bii.a-star.edu.sg/).
   - You'll find a list of mutations of Resistance in the following webpage: https://virusfrenchresistance.org/virus-french-resistance-rsv/



Questions

- To which clades these query RSVs belong to?
- Do you see any drug resistance mutations?