# Mini project 3:

## Bayesian and likelihood-based phylogenetics.
## SARS-CoV-2: selection and clock models

In this mini project you will continue your investigation of the evolution of SARS-CoV-2. Specifically, you will use likelihood and Bayesian methods to investigate the selective forces acting on the virus, and refine the previous analysis where you tried to date the original zoonotic transfer. As in the other mini projects the idea is to expose you to a realistic workflow for a research project, meaning you will not initially know how to do everything. Use google when you run into problems, and also use the discussion forum to discuss with other students, and ask questions from me.

## 1: Construction of data sets

You will need two SARS-CoV-2 data sets for this mini project:

- A full-genome data set for clock-analysis (derived from the data set you made for mini project 2)
- A new data set consisting only of coding sequences for the spike protein, to be used for analysing selection (with the PAML software)

**Construction of derived full-genome data set:**

Based on the SARS-CoV-2 data you collected for mini project 2, create an updated data set:

- Remove the original SARS (2003) sequence you previously used as outgroup
- Remove the sequences that do not have a human host
- Edit the remaining SARS-CoV-2 sequence names such that they specify the sample date: Make sure each name in the fasta file ends with the date in this format:
    - "_yyyy/MM/dd"
    - i.e., underscore, 4-digit year, 2-digit month number, and 2-digit day. (You should not include the quotes). For instance: "seq1_2009/05/20" for May 5th 2009.

**Construction of spike-protein data set:**

Use the NCBI Virus ressource (Search by sequence) to collect a set of spike protein sequences: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/

- Virus name: SARS-CoV-2
- Nucleotide completeness: complete genomes
- Host: Human
- Proteins: surface glycoprotein

- Select about 30-50 sequences in all. Try to cover a few different Pango lineages (e.g. B.1.1.7 is the “British variant”, B.1.429 is the “California variant”, and B.1.617 is the “India variant”)
- Download the sequences: Under “Sequence data” make sure to select “Coding region”. This will give you just the coding sequence corresponding to the spike protein.
- You do not need sampling date for the sequences.
- You should not include an outgroup sequence

Briefly describe the steps you used to construct the data set: what sequences did you select? What filters did you use? Include two tables listing the names for all sequences.

## 2: Alignment of sequences

Align the SARS-CoV-2 sequences. Prepare versions of the alignments for use in:

- PAUP (spike gene sequences)
- PAML (spike gene sequences)
- BEAST2 (full-genome alignment)

Briefly describe how you performed these steps.

## 3: Max likelihood phylogeny of spike gene sequences

Use PAUP to create a maximum likelihood tree for the spike gene sequences. Use the GTR+G model.

What commands did you use to construct the max likelihood tree? Include a plot of the tree.

## 4: Investigate selective pressure on spike gene

Use the tree you just constructed and the alignment of spike gene sequences as input to codeml (from the PAML package), and perform an analysis where you test for the presence of positive selection and estimate dN/dS for all codons.

Inspect output files and answer these questions:

- What are the estimated dN/dS rate ratios for the 2 or 3 different codon classes in the spike genes? What fraction of codons belong to each codon class?
- Are there any sites in your selected spike gene sequences that show signs of positive selection?
- The spike protein is important for binding cellular receptors and is also a target of many vaccine projects. Try to find a few residues that are important for binding cellular receptors or interacting with antibodies (from recent literature) and find the estimated dN/dS for these sites.
- What selective pressure would we prefer for sites used as targets of vaccines?

## 5: Do clock model analysis

Fit a strict clock model to the full-genome alignment using BEAST2. Estimate the date when the first human got infected (give 95% credible interval).

## 6: Conclusion

Briefly summarise what you have learned about SARS-CoV-2 evolution from these analyses.