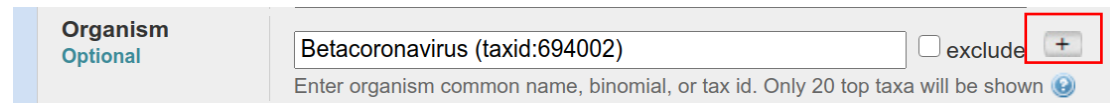


1. Download nucleotide entry NC_045512 from NCBI and save as fasta. If interested - look at available coronavirus sequences in RefSeq with search term betacoronavirus[orgn].

2. Lets collect related genomes.

- a) Go to https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch
- b) Set search using the COVID-19 sequence you downloaded before.
- c) Restrict search to Betacoronavirus



- d) Add additional organism search term and this time check the box next to the “exclude” entry.
- e) Set to exclude sequences matching taxid 2697049. Why we exclude this? Try a search excluding and not this term.
- f) Set “Entrez query” term to *complete genome*[title]
- g) Set maximum number of return sequences to 1000.
- h) Download complete sequences that has coverage $\geq 50\%$ as fasta file and add the NC_045512 entry on the top.
- i) Also add camel virus (MN514967.1) sequence

3. Remove redundant sequences:

- a) Download and compile <https://github.com/niu-lab/gclust>
- b) Sort the input genomes in decreasing order of length (look at gclust github page)
- c) Cluster with gclust at 97 identity cut-off.
- d) Play with grep/linux utilities and get ids of the representatives.
- e) Use seqkit grep to extract representatives from the initial set.

4. Protein based analysis

- a) Search this protein <https://www.uniprot.org/uniprot/D3W8N4> against the collected viral genomes using tblastn (word size 2, e=10).
- b) Download the aligned parts.
- c) Translate with seqkit translate command.
- d) By using `seqkit seq -m` discard all protein sequences that are shorter than 800.
- e) Align with mafft (`$ mafft --maxiterate 1000 --localpair`)
- f) For easier interpretation and annotation you could remove “:” and spaces from the alignment files.
- g) Generate tree with fasttree (use option “-gamma”). Google about this program.

5. Analysis

- a) Use ETE3 python package to add root on the camel virus (<http://etetoolkit.org/docs/latest/tutorial/index.html>). Command “set_outgroup”

6. Interpretation.....how did the Covid-19 evolve, what path through hosts was taken? Would it be different interpretation if out-group is not used? What about Urbani SARS origin? Is the Palm Civet origin evident?