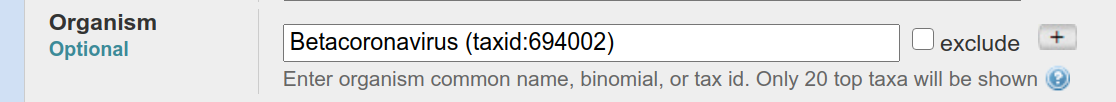
1. **Download nucleotide entry NC\_045512 from NCBI and save as fasta. If interested - look at available coronavirus sequences in RefSeq with search term betacoronavirus[orgn].**

**https://www.ncbi.nlm.nih.gov/nuccore/NC\_045512.2?report=fasta**

1. Lets collect related genomes.
   1. Go to <https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch>
   2. Set search using the COVID-19 sequence you downoaded before.
   3. Restrict search to Betacoronavirus



* 1. Add additional organism search term and this time check the box next to the “exclude” entry.
  2. Set to exclude sequences matching taxid 2697049. Why we exclude this? Try a search excluding and not this term.
  3. Set “Entrez query” term to \*complete genome\*[title]
  4. Set maximum number of return sequences to 1000.
  5. Download complete sequences that has coverage >=50% as fasta file and add the NC\_045512 entry on the top.
  6. Also add camel virus (MN514967.1) sequence

https://www.ncbi.nlm.nih.gov/nuccore/MN514967.1?report=fasta

* 1. Also do the search changing the database to ”RefSeq Genome Database” - add the collected sequences to the analysis. For this step remove the settings for “exclude”

“collected sequences” – cia kaip suprantu ta camel virus ir NC\_045512? Nes dedant visus is h) zingsnio meta error’a kad per daug base’u (12 mil > 1 mil max). Ir cia filtruot tik tuos kur >50% irgi ar ne?

1. Remove redundant sequences:
   1. Download and compile <https://github.com/niu-lab/gclust>
   2. Sort the input genomes in decreasing order of length (look at gclust github page)

perl script/sortgenome.pl --genomes-file seqdump.txt --sortedgenomes-file seqdump\_sorted.txt

* 1. Cluster with gclust at 97 identity cut-off.

./gclust -threads 8 -memiden 97 seqdump\_sorted.txt > seqdump\_sorted\_output.txt

* 1. Play with grep/linux utilities and get ids of the representatives.

cat seqdump\_sorted\_output.txt | grep -Eo ", .\*\.\.\. \\*" | cut -b 4- | rev | cut -b 6- | rev > seqdump\_sorted\_output\_ids.txt

* 1. Use seqkit grep to extract representatives from the initial set.

**seqkit grep -f seqdump\_sorted\_output\_ids.txt seqdump.txt -o res**

**(same stats, niekas nepasikeite)**

1. Protein based analysis
   1. Search this protein <https://www.uniprot.org/uniprot/D3W8N4.> against the collected viral genomes using tblastn (word size 2, e=10).

Paduodant E visur 0?

* 1. Download the aligned parts.
  2. Translate with seqkit translate command.

seqkit translate aligned\_tblastn.txt -o aligned\_tblastn\_translated.txt

* 1. By using seqkit seq -m discard all protein sequences that are shorter than 800.

seqkit seq -m 800 aligned\_tblastn\_translated.txt -o aligned\_tblastn\_translated\_filtered.txt

* 1. Align with mafft ($ mafft --maxiterate 1000 —localpair)

mafft --maxiterate 1000 --localpair aligned\_tblastn\_translated\_filtered.txt > aligned\_mafft.txt

* 1. For easier interpretation and annotation you could remove “:” and spaces from the alignment files.

cp aligned\_mafft.txt aligned\_mafft\_mod.txt && sed "-i" "" -E "-e" "s/^>[^:]\*:[^[:space:]]\* />/g" "-e" "s/ /\_/g" aligned\_mafft\_mod.txt

* 1. Generate tree with fasttree (use option “-gamma”). Google about this program.

./FastTree -gamma aligned\_mafft\_mod.txt > tree.txt

1. Analysis
   1. Use ETE3 python package to add root on the camel virus (http://etetoolkit.org/docs/latest/tutorial/index.html). Command “set\_outgroup”
2. 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?

**1) How did Covid-19 evolve, what path through hosts was taken?**

Covid-19 evolved/transmitted to humans through horseshoe bats. Data suggests that maybe Pangolins served as an intermediary since Covid-19 contains genetic similarity with them as well.

**2) Would it be a different interpretation if out-group is not used?**

If out-group is not used, camel virus clusters quite closely with the Covid-19 - it could misleadingly imply that the camel virus is an off-shoot of the Bat & Pangolin host variants, rather than a common ancestor.

**3) What about Urbani SARS origin?**

Urbani SARS branches out from Covid-19 (SARS-CoV-2), indicating that, although they are both SARS-related coronaviruses and share a common bat ancestor - the evolutionary / transmission path they took to jump to humans was different.

**4) Is the Palm Civet origin evident?**

From my collected files, there were no Palm Civet data points. However, after looking on the internet - they may have served as intermediates between bats and humans.