


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

2. Lets collect related genomes.

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

Organism Optional	Betacoronavirus (taxid:694002)	<input type="checkbox"/> exclude	
	Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown		

- Add additional organism search term and this time check the box next to the “exclude” entry.
- Set to exclude sequences matching taxid 2697049. Why we exclude this? Try a search excluding and not this term.
- Set “Entrez query” term to *complete genome*[title]
- Set maximum number of return sequences to 1000.
- Download complete sequences that has coverage >=50% as fasta file and add the NC_045512 entry on the top.
- Also add camel virus (MN514967.1) sequence
<https://www.ncbi.nlm.nih.gov/nuccore/MN514967.1?report=fasta>
- 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?

3. Remove redundant sequences:

- Download and compile <https://github.com/niu-lab/gclust>
- 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`

- Cluster with gclust at 97 identity cut-off.
`./gclust -threads 8 -memiden 97 seqdump_sorted.txt > seqdump_sorted_output.txt`
- 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`

- 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)

4. Protein based analysis

- 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?
- Download the aligned parts.
- Translate with seqkit translate command.
`seqkit translate aligned_tblastn.txt -o aligned_tblastn_translated.txt`
- 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`
- Align with mafft (\$ mafft --maxiterate 1000 --localpair)
`mafft --maxiterate 1000 --localpair aligned_tblastn_translated_filtered.txt >`

`aligned_mafft.txt`

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

`./FastTree -gamma aligned_mafft_mod.txt > tree.txt`

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?

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.