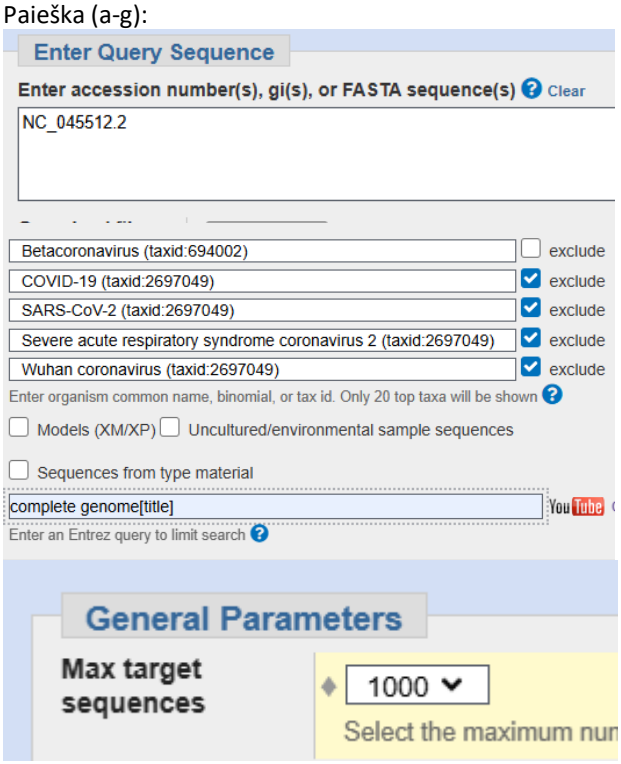
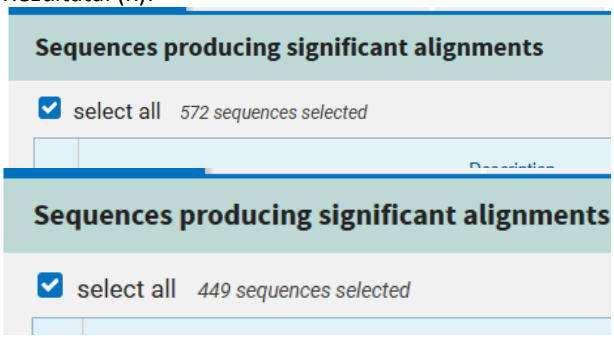
Linas Janavičius, VU, PSI, MIF, 4k., 4gr., 2110560

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.

Paieška (a-g):

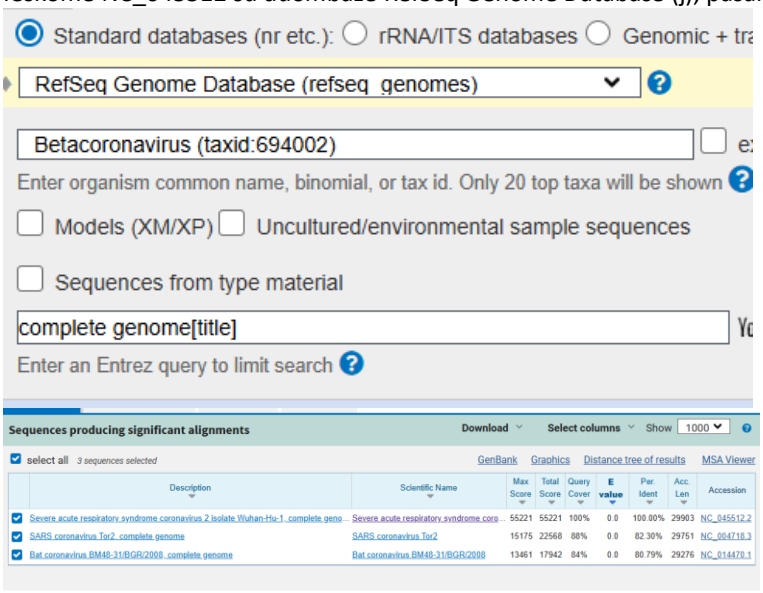


Rezultatai (h):



Rastos 572 sekos. Pritaikius coverage >=50% gauname 449 sekas. Id 2697049 pašalinimas iš paieškos leidžia susitelkti į SARS-CoV-2 kilmės ir evoliucijos analizę, išvengiant duomenų, susijusių su NC_045512 sekų kartotinėmis kopijomis, kurios gali užgožti giminingų virusų informaciją.

Ieškome NC_045512 su duombaze RefSeq Genome Database (j), pašalinus taxid 2697049:



Gauname 449 + 3 + NC_045512 + virus (MN514967.1) = 454 sekos.

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)

```
#Sort
!perl script/sortgenome.pl --genomes-file
/content/gclust/outputs/all_sequences.fasta --sortedgenomes-file
/content/gclust/outputs/sorted_all.fasta
```

   c) Cluster with gclust at 97 identity cut-off.

```
#Cluster
!./gclust -minlen 20 -both -nuc -threads 8 -ext 1 -sparse 2 -memiden 97
/content/gclust/outputs/sorted_all.fasta >
/content/gclust/outputs/clustering.out

Total clusters: 147
```

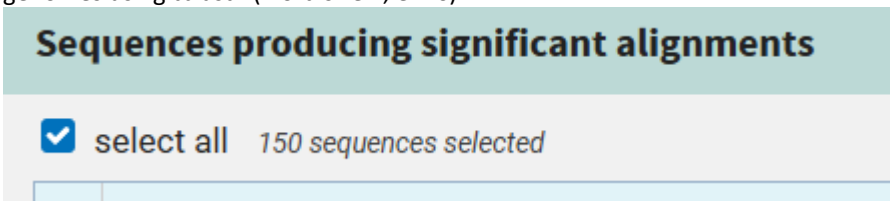
   d) Play with grep/linux utilities and get ids of the representatives.

```
#Get ids of the representatives
!grep ".*\*" /content/gclust/outputs/clustering.out | cut -d'>' -f2 | awk
'{print $1}' | sed 's/\.\.\.$//' >
/content/gclust/outputs/representative_ids.txt
```

   e) Use seqkit grep to extract representatives from the initial set.

```
#Use seqkit to extract sequences
!./seqkit grep -i -f /content/gclust/outputs/representative_ids.txt
/content/gclust/outputs/all_sequences.fasta -o
/content/gclust/outputs/result.fasta
```

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).



   e visur 0.
   b) Download the aligned parts.
   Išsaugota `aligned.fasta dokumente.`
   c) Translate with seqkit translate command.

```
#Translate
!./seqkit translate /content/gclust/outputs/aligned.fasta -o
/content/gclust/outputs/translated_proteins.fasta
```

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

```
#Filter shorter than 800
!./seqkit seq -g -m 800
/content/gclust/outputs/translated_proteins.fasta -o
/content/gclust/outputs/filtered_proteins.fasta
```

   e) Align with mafft ($ mafft --maxiterate 1000 --localpair)

```
#mafft alignment
!mafft --maxiterate 1000 --localpair
/content/gclust/outputs/filtered_proteins.fasta >
/content/gclust/outputs/output_aligned.fasta
```

   f) For easier interpretation and annotation you could remove ":" and spaces from the alignment files.

```
input_file = "/content/gclust/outputs/output_aligned.fasta"
output_file = "/content/gclust/outputs/output_cleaned_aligned.fasta"
# Open the input file, clean it, and write to the output file
with open(input_file, "r") as infile, open(output_file, "w") as outfile:
    for line in infile:
        # Remove spaces and ':' characters
        cleaned_line = line.replace(" ", "_").replace(":", "_")
        outfile.write(cleaned_line)
```

   g) Generate tree with fasttree (use option "-gamma"). Google about this program.

```
!FastTree -gamma /content/gclust/outputs/output_cleaned_aligned.fasta
> /content/gclust/outputs/phylogenetic_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"

```
tree = Tree("/content/gclust/outputs/phylogenetic_tree.txt")
camel_virus_label =
"lcl|Query_4358037_4901-8458_MN514967.1_Dromedary_camel_coronavirus_H
KU23_isolate_DcCoV-HKU23/camel/Nigeria/NV1385/2016"
```

```
tree.set_outgroup(camel_virus_label)
```

6. Interpretation.
    a) How did the Covid-19 evolve, what path through hosts was taken?
        OVID-19, caused by SARS-CoV-2, likely evolved from a bat coronavirus (e.g., RaTG13) and may have passed through an intermediate host, such as a pangolin, before spilling over to humans.
    b) Would it be different interpretation if out-group is not used?
        Without an out-group, the phylogenetic tree would be unrooted, making it difficult to determine the direction of evolution and which lineage is ancestral. This would complicate the interpretation of SARS-CoV-2's origin and could lead to inaccuracies in identifying the primary host (e.g., bats or pangolins).

        Without an out-group, the phylogenetic tree might misleadingly suggest that the Dromedary camel virus is another type of COVID-19 or closely related to SARS-CoV-2. However, this is not accurate—camel coronavirus is not a type of COVID-19 but rather part of a common ancestor of various betacoronaviruses.
    c) What about Urbani SARS origin?
        Urbani SARS originated from bat coronaviruses.
    d) Is the Palm Civet origin evident?
        While the tree does not explicitly include a palm civet-derived coronavirus, the evolutionary position of SARS-CoV (Urbani) and historical evidence strongly indicate the involvement of palm civets as the intermediate host.