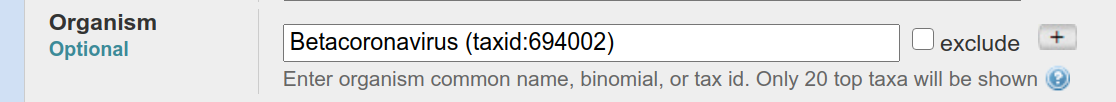
1. Download nucleotide entry NC\_045512 from NCBI and save as fasta. If interested - look at available coronavirus sequences in RefSeq with search term betacoronavirus[orgn].
2. Lets collect related genomes.
   1. Go to <https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch>
   2. Set search 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
  7. 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”

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)
   3. Cluster with gclust at 97 identity cut-off.
   4. Play with grep/linux utilities and get ids of the representatives.
   5. Use seqkit grep to extract representatives from the initial set.
2. 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).
   2. Download the aligned parts.
   3. Translate with seqkit translate command.
   4. By using seqkit seq -m discard all protein sequences that are shorter than 800.
   5. Align with mafft ($ mafft --maxiterate 1000 --localpair)
   6. For easier interpretation and annotation you could remove “:” and spaces from the alignment files.
   7. Generate tree with fasttree (use option “-gamma”). Google about this program.
3. Analysis
   1. Use ETE3 python package to add root on the camel virus (http://etetoolkit.org/docs/latest/tutorial/index.html). Command “set\_outgroup”
4. 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?