Pipeline

Note: The pdf version of this file explains the pipeline while the Rmd version contains the code.

Instructions:

1. Filter genomes on NCBI using the following criteria and download the csv results table:
   - Taxon ID: 2697049 (CoV-2)
   - Ambiguous Characters < 30
   - Nucleotide Completeness: Complete
   - Host: 9696 (human)
   - Collection date: 30th December 2003 - current date
2. Extract accession IDs from the results tables.
3. Download genomes using NCBI datasets and remove duplicates using SeqKit.
4. Extract ORFs for S gene using EMBOSS and remove duplicates using SeqKit.
5. Cluster the sequences using h-clustering.
6. Align clusters roughly using MAFFT and remove duplicates using SeqKit.
7. Improve genetic diversity for each lineage.