Pipeline

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

Instructions:

1. Filter genomes on NCBI and download the csv results table:
2. Extract accession IDs from the results tables.
3. Download genomes using NCBI datasets and remove duplicates using SeqKit.
4. Extract ORFs using EMBOSS and remove duplicates using SeqKit.
5. Cluster the sequences using h-clustering (or some other method).
6. Align clusters roughly using MAFFT and remove duplicates using SeqKit.
7. Improve genetic diversity for each lineage.