BigSmallDNA Example Pipeline

For studies with DNA datasets, big data is good but computationally expensive. This method aims to reduce it without losing the benefits. It does this by improving the nucleotide diversity of the sample.

This example uses SARS-CoV-2 virus (responsible for 2019 pandemic and has >9M seqs on GenBank) to show how to implement this method and how good it is.

From Obtaining Original Dataset to Reduced Dataset

NCBI Virus web server hosts >9M sequences for SARS-CoV-2 but many of these are poorly read and can negatively impact results. Hence a strict inclusion/exclusion criteria was used to filter the sequences. However, the web server can not download large datasets. And, the command line tool "datasets" by NCBI is developed for this purpose, it lacks the filtration capabilities of the web server. Hence after filtration, the results table was downloaded from which the Accession IDs were be extracted and fed to datasets. This process was done in batches due to poor internet stability. After downloading the dataset, duplicates were removed using "SeqKit".

Since distinct genomes need not have distinct genes, the S gene was extracted from all genomes using "EMBOSS getORF" and the duplicates were removed using SeqKit. Since SARS-CoV-2 is classified using the Pangolin lineage, we used it to cluster the sequences and large clusters (>6000 seqs) were sub-clustered using kmeans and h-clustering. If your set of sequences does not have an existing classification, you can use clustering softwares such as MeShClust and MMSeqs2 as they do not require aligned sequences.

Each cluster was then roughly aligned using "MAFFT" with loose parameters and the genetic diversity was improved using the provided method. The corresponding unaligned sequences were combined to form the reduced dataset.

Technical Specification

Software Dependencies (versions)

- R (4.4.1) with packages: tidyverse, stringr, ape, pegas, kmer, and dendextend
- GNU bash (5.2.26)
- datasets (16.22.1)
- SeqKit (2.8.2)
- EMBOSS (6.6.0.0)
- MAFFT (7.526)

This pipeline was run on a laptop with the following hardware:

- Model: Dell Inc. Precision 7730
- CPU: Intel Core i7-8750H × 12
- RAM: 32 GB
- VRAM: 8 GB
- OS: Fedora Linux 40 (Workstation) with Linux 6.11.4-201.fc40.x86_64

Instructions

1. Filter genomes on NCBI web server using the following criteria and download the csv results table with collection date and Pangolin lineage,
    - Taxon ID: 2697049
    - Ambiguous Characters < 30
    - Nucleotide Completeness: Complete
    - Host: 9696 (human)
    - Collection date: 30th December 2003 - current date
2. Extract accession IDs (with version) from the results table.
3. Download genomes using NCBI datasets and remove duplicates using SeqKit.
4. Extract ORFs using EMBOSS and remove duplicates using SeqKit.
5. Cluster sequences using Pangolin lineage
6. Sub-cluster by kmeans and hierarchical clustering.
7. Roughly align the sequences using MAFFT.
8. Improve genetic diversity of all clusters using provided method.
9. Finely align all roughly using MAFFT.

Code

See Rmd version of this document.

Results

NCBI hosts 9,066,813 (on 5th Feb 2024) from which 1,365,731 were downloaded using a strict filtration criteria and after extracting the S gene and removing duplicates, the original dataset comprised of 183,780 sequences. While the data had 2658 lineages, 880 had less than 3 sequences and were grouped together. The remaining were clustered and 6 were sub-clustered, forming 1849 clusters. The general improvements are shown below in the Table 1 and Figure 1 while Figure 2 shows the changes in the distributions of nucleotide diversity and number of sequences.

Table 1: Summary of changes in dataset

Property	Original Dataset	Reduced Dataset	Change (%)
mean π per cluster	0.0009177	0.0074139	708
median π per cluster	0.0006168	0.0013001	111
min π per cluster	0.0000000	0.0000000	0
max π per cluster	0.1621207	0.6161369	280
mean seqs per cluster	99	3	3200
median seqs per cluster	12	2	500
min seqs per cluster	3	2	50
max seqs per cluster	4,749	34	13,868
Total number of seqs	183,780	6,673	2,654

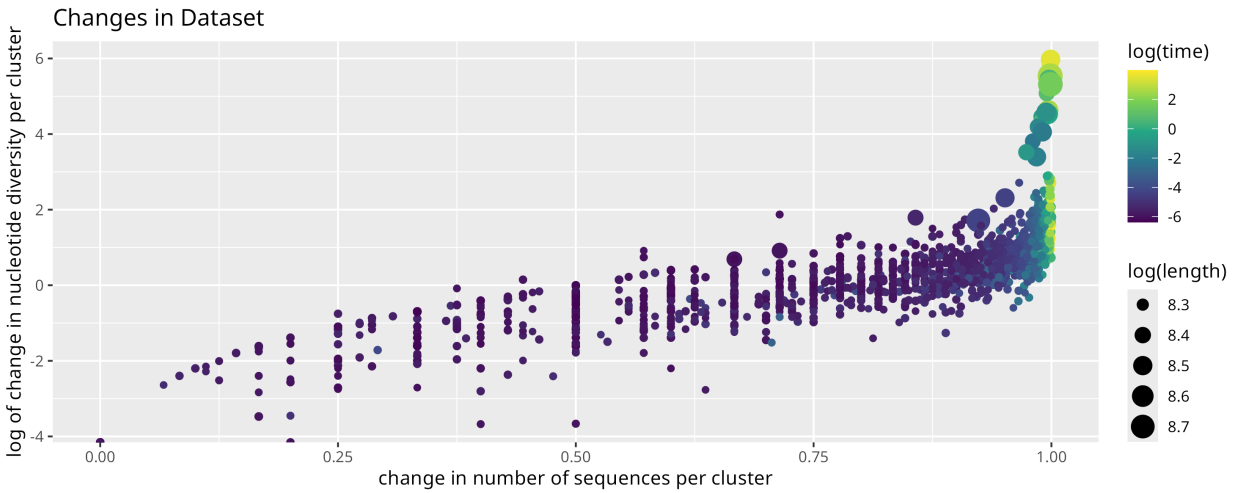

Figure 1: Overall Changes in the Dataset

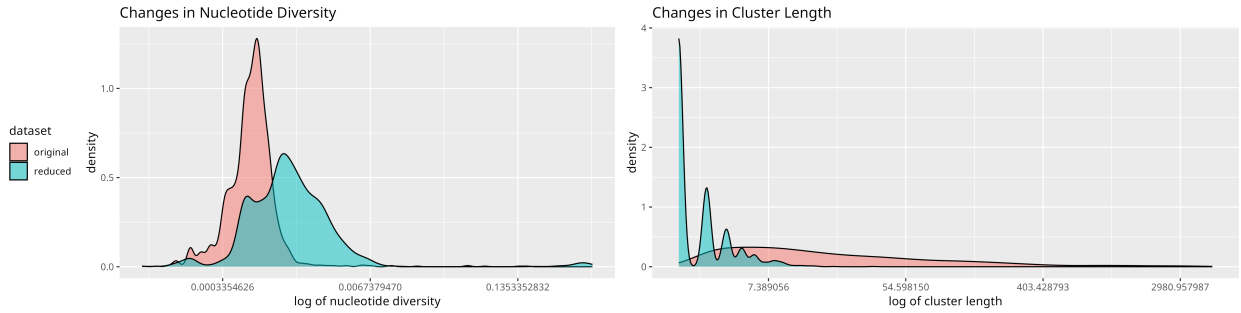

Figure 2: Changes in distribution of length and nucleotide diversity.