MAR 21, 2024

DOI:
dx.doi.org/10.17504/protocols.io.e6nvw1p9wlmk/v1

Protocol Citation: Spoorthi R Kulkarni, Lavanya C, Baishali Garai, Vidya Niranjan 2024. Unveiling Prolonged COVID Variants: A Protocol for Clustering and Phylogenetic Analysis. **protocols.io** https://dx.doi.org/10.17504/protocols.io.e6nvw1p9wlmk/v1

Protocol status: Working
We use this protocol and it's working

Unveiling Prolonged COVID Variants: A Protocol for Clustering and Phylogenetic Analysis

Spoorthi R Kulkarni[1], Lavanya C[1], Baishali Garai[2], Vidya Niranjan[1]

[1]R V College of Engineering;
[2]R V University, School of Computer Science and Engineering

Vidya Niranjan
R V College of Engineering

ABSTRACT

Prolonged COVID-19 has emerged as a significant concern globally, with a subset of individuals experiencing persistent symptoms long after the acute phase of the infection. Understanding the genomic basis of prolonged COVID-19 can provide crucial insights into its pathophysiology and aid in the development of targeted therapeutic strategies. In this study, we retrieved genome sequences of COVID-19 from the National Center for Biotechnology Information (NCBI) database and employed a comprehensive pipeline to identify single nucleotide polymorphisms (SNPs). The primary objective was to investigate the clustering similarity of different variants of COVID-19 prevalent in the Indian population. By focusing solely on Indian population genome sequences, we aimed to capture the unique genetic landscape of COVID-19 variants circulating in this demographic. Our analysis revealed distinct SNP patterns across the Indian population, indicative of genetic diversity within the viral strains. Furthermore, we plan to map these identified SNPs to relevant pathways to elucidate their potential functional significance in the context of prolonged COVID-19.

Created: Mar 18, 2024

Last Modified: Mar 21, 2024

PROTOCOL integer ID: 96838

Keywords: Prolonged Covid-19, Single Nucleotide Polymorphism, Multiple Sequence Alignment, Phylogenetic Tree

SAMPLE COLLECTION AND REFERENCE GENOME

1 The genomic sequences utilized in this protocol were sourced from the NCBI Virus Database, which encompasses region-specific genome sequences. Focused on the Indian population, the protocol primarily analyzes sample sequences provided in FASTA format, alongside reference genomes obtained from the NCBI Virus Database.

Dataset

NAME	LINK
Sample and reference genome retrieved from NCBI virus	https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Severe%20acute%20respiratory%20syndrome%20coronavirus%202,%20taxid:2697049&Country_s=India

REFERENCE GENOME INDEXING

2 Index the reference genome of COVID-19 by executing the 'bwa index' command followed by the name of the FASTA file containing the reference genome. This step facilitates efficient alignment of sequencing reads to the reference genome during subsequent analyses

Command

bwa index "Reference_genome.fasta" initiates the process of indexing a reference genome file using the Burrows-Wheeler Aligner (BWA).

```
bwa index Reference_genome.fasta
```

ALIGNMENT AND MAPPING

3 The bwa command is used to align the sequencing reads to the indexed reference genome. This command efficiently aligns the reads, taking into account possible mismatches, insertions, deletions, and sequencing errors. SAMtools is a suite of programs for interacting with high-throughput sequencing data in SAM/BAM format.

Command

The provided command executes the alignment procedure for COVID-19 samples against a reference genome using the Burrows-Wheeler Aligner (BWA) tool. It takes as input the reference genome file ("Reference_genome.fasta") and the FASTA file containing the sequence of the sample ("sample_ID.fasta"). Upon execution, the output, formatted as a Sequence Alignment/Map (SAM) file, contains the alignment information detailing how the sample sequences align with the reference genome.

```
bwa mem Reference_genome.fasta sample_ID.fasta > sample_ID.sam
```

4

Command

SAM to BAM conversion using SAM tools

```
samtools view -@ 4 -Sb -o sample_ID.bam sample_ID.sam
```

VARIANT CALLING AND SNP GENERATION USING SAMtools MPILEUP AN…

5 This protocol step involves variant calling and single nucleotide polymorphism (SNP) generation using SAMtools mpileup and BCFtools call. The SAMtools mpileup command is used to generate a pileup format file from multiple BAM files aligned to a reference genome. This file contains information about the alignment of sequencing reads to the reference genome at each genomic position. The BCFtools call command then analyzes the pileup data to identify variants, including SNPs, insertions, deletions, and complex variants. The output is generated in variant call format (VCF), providing detailed information about the detected variants, including their genomic coordinates, allele frequencies, and quality scores. This protocol step is crucial for identifying genetic variations and understanding the genomic landscape of the samples under investigation.

Command

VCF generation

```
samtools mpileup -uf Reference_genome.fasta sample_ID.bam | bcftools call -O v -
mv -o 1_output.vcf
```

VARIANT GENERATION USING MULTIPLE SEQUENCE ALIGNMENT

6 The Multiple Sequence Alignment (MSA) was conducted using the MAFFT version 7 tool, leveraging the alignment parameters specifically chosen for the analysis. The FASTA files, obtained from earlier

downloads, served as input for the MSA procedure. The primary objective of this step was to align the genomic sequences to visualize shared single nucleotide polymorphisms (SNPs) and subsequently generate a phylogenetic tree and similarity score clusters.

Note

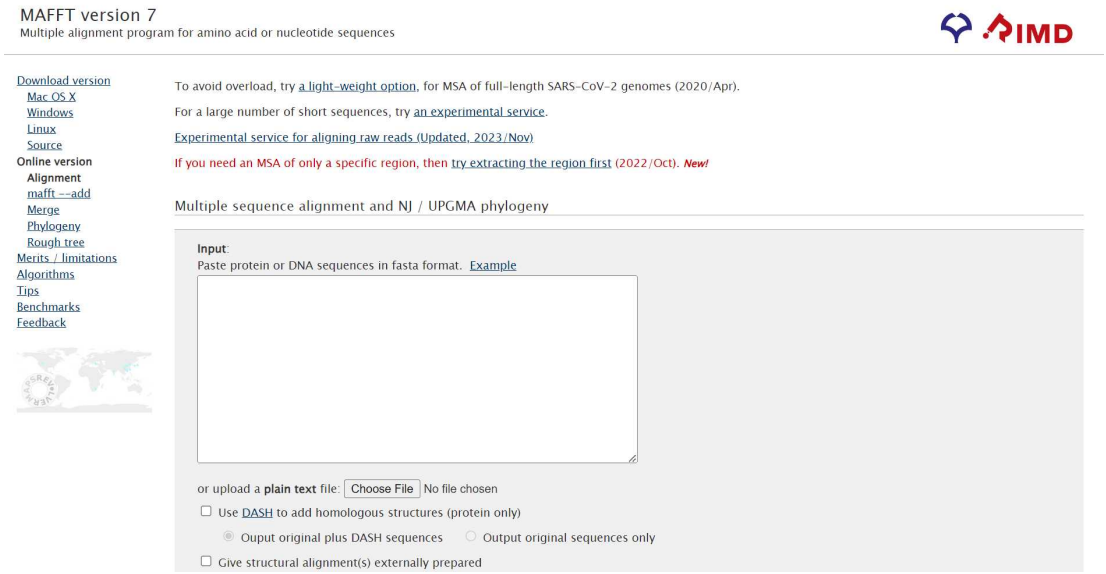
SNPs were additionally identified using the pipeline established with mpileup, enabling the exploration of genetic variations within the COVID-19 genomic sequences. These SNPs were further analyzed using the Multiple Sequence Alignment (MSA) web server to assess their similarity. Consequently, the generated phylogenetic tree and clusters were leveraged to elucidate the evolutionary relationships and genetic similarities among the identified SNPs.

Command

MAFFT version 7 web server

```
https://mafft.cbrc.jp/alignment/server/
```

Expected result

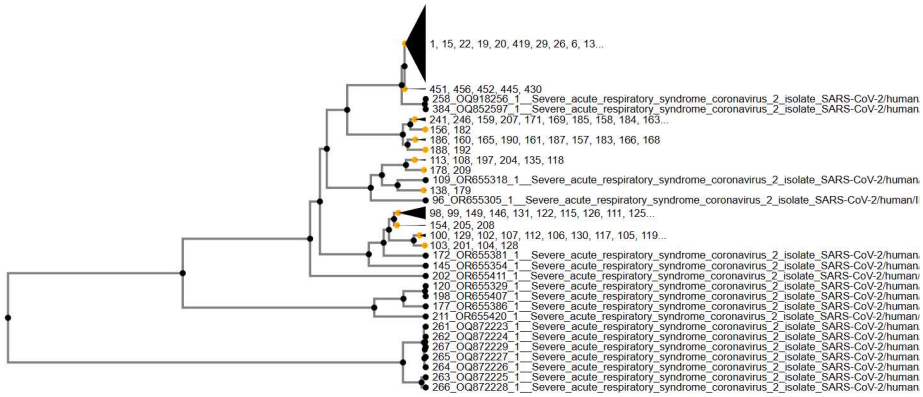


The MAFFT version 7 web server

As depicted in the image, the acquired FASTA files can be uploaded by opting for the designated "choose File" button, and then selecting the alignment parameter located in the upper-left corner of the web server interface.

7 Clustering and Phylogenetic Tree Analysis

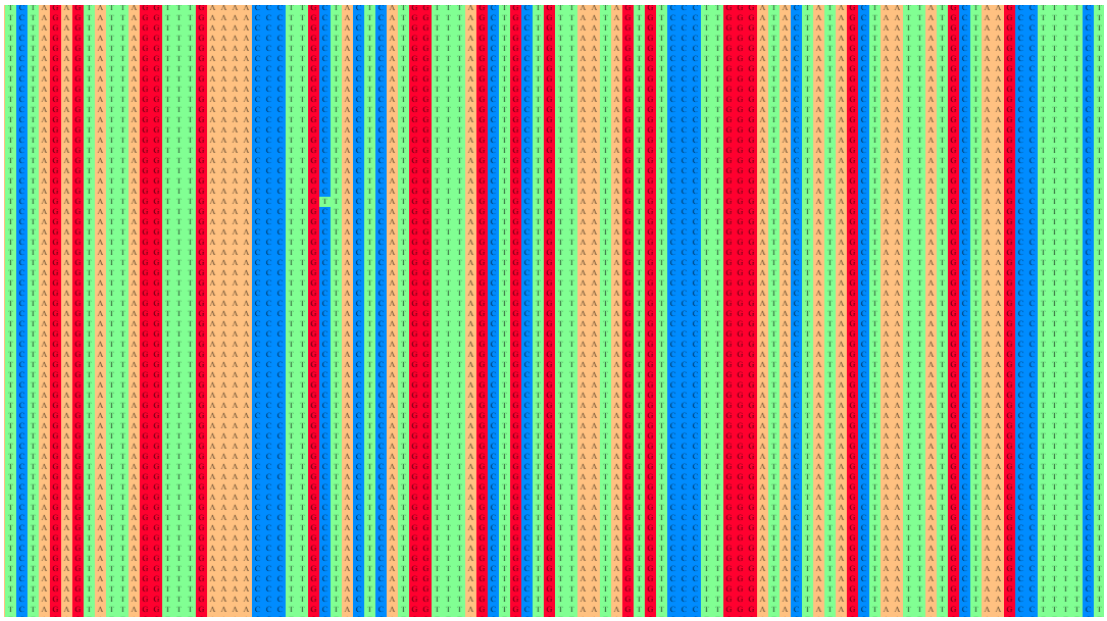
Expected result



The phylogenetic tree depicted here is constructed based on the alignment of COVID-19 genomic sequences obtained from FASTA files. The tree illustrates the evolutionary relationships among the samples, with branches representing genetic divergence and nodes indicating common ancestors. Clusters within the tree highlight groups of sequences sharing similar genetic features.

8

Expected result



The picture depicts the similarity observed between the SNPs generated from the pipeline and the Multiple Sequence Alignment (MSA) tool suggesting a congruence in the identified genetic variations. This congruence underscores the reliability and accuracy of both the pipeline and the MSA tool in detecting single nucleotide polymorphisms (SNPs). Such consistency reinforces confidence in the analytical methodologies employed and enhances our understanding of genetic diversity within the studied population or species.