Phylogenetic analysis of SARS-COV-2 genomes using MAFFT and iqtree

About

This task was collaboratively carried out by team-venter, named after famous Craig Venter

Team members:
@WGatua - *Team leader* | @Pawan - *Team assistant*
@Harinath, @Josiah24, @Ahmad, @Maruf, @Caroline, @Sefunmi, @Tracyallen

Background

The objective of the Stage 1 task on HackBio internship was to perform an analysis that is simple, but informative. Due to the current advancement in high throughput sequencing, there are numerous samples of the viral genome which are being sequenced on a daily basis and deposited on online repositories such as NCBI, GISAID etc. We sought to render a phylogenetic tree of complete human severe acute respiratory syndrome coronavirus 2 (SARS-COV-2). To make things a little interesting, we used the sequences from the respective countries of our team members (Figure 1). With this analysis, we will show how closely the different viral genomes in the different home countries of our team members are related, by showing their evolutionary paths and their relations with the ancestral isolated in Wuhan, China.

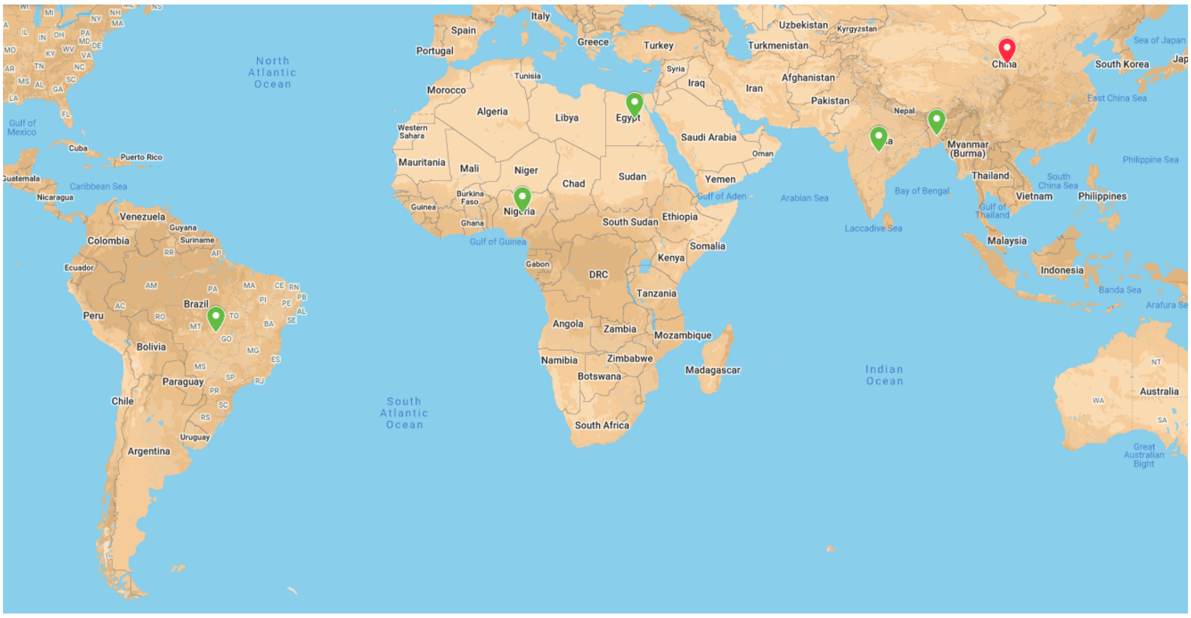


Figure 1. Countries included in the phylogenetic analysis of SARS-COV-2. Green pin map: home countries of Team-venter members; Brazil, Nigeria, Egypt, India and Bangladesh. Red pin map: reference genome of SARS-COV-2 in China.

Methodology

Data collection

Human SARS-CoV-2 data was collected from The National Center for Biotechnology Information (NCBI), which is a major resource for bioinformatics tools and services. The downloaded data set contains viral genome, protein sequences, annotation and detailed data report. Genomic sequence from SARS-2, Wuhan, China (NCBI Reference Sequence: NC_045512.2) was used as reference. Complete genome sequences were retrieved across the 6 countries (Ten sequences each from Bangladesh, Brazil, India, Kenya, Egypt and one from Nigeria) were also collected (Table 1). Although one of our team members is from Kenya, unfortunately no sequences were found in the NCBI database from Kenya.

Software packages

i. MAFFT version 7.471/Ubuntu version

ii. Jalview version 2.11.1.0

iii. IQ-TREE Multicore version 2.0.6

iv. FigTree version 1.4.4

v. Biopython version 1.77s

Analysis

For dataset collection, curl command was used to retrieve the data from NCBI in a terminal in a command-line as in:

```
curl -o datasets
'https://ftp.ncbi.nlm.nih.gov/pub/datasets/command-line/LATEST/linux-amd64/datasets'
```

Next, to give permission to access the datasets module we used chmod +x datasets and enabled the execution of the tasks.

Because we were interested in the genomes of coronavirus that cause severe acute respiratory syndrome we did not include experimental inputs. For that, we used the following command-line. The datasets command was used to download the complete genomic data for the human coronavirus.

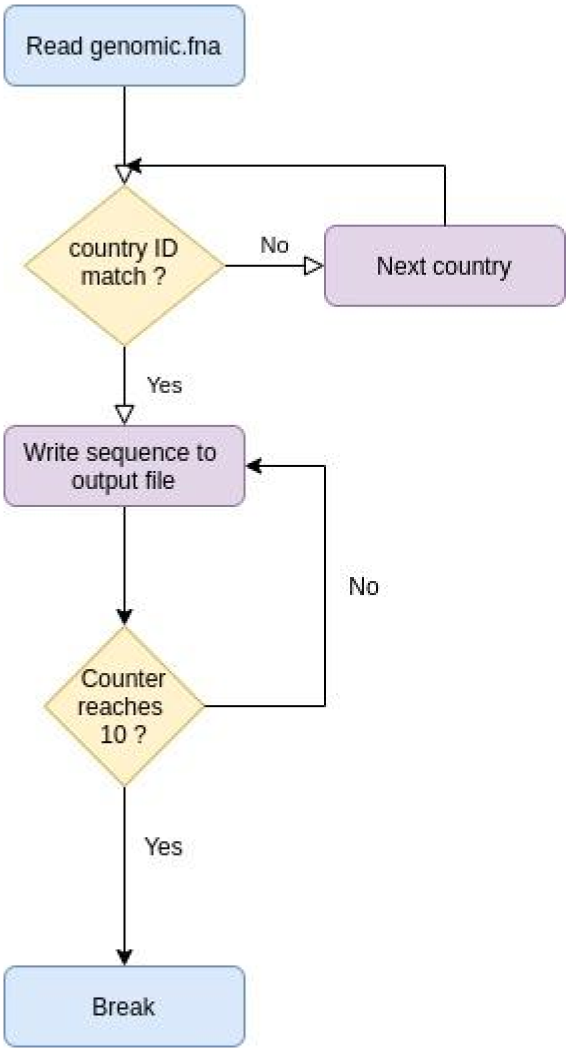
```
./datasets download virus genome tax-name SARS2 --host human --filename
SARS2-all.zip
```

The dataset is compressed in a zip file. We used unzip command to next extract batch data from the dataset:

```
echo "Extracting batch sequences from the SARS-CoV2 dataset"

python getseq.py
```

To extract the sequences of interest from the selected countries, we used the getseq.py script below on Biopython using the genomic.fna as input.
Below is a flowchart to describe the algorithm for sampling sequences of interest.



A sequence.fasta file will be generated which will contain at most 10 sequences for each country.

The alignment was performed with the MAFFT program in reference to the genome from Wuhan, China. The command-line below directs the alignment with reorder and sets maximum number of **iterative refinement** (an improvement of the accuracy) in 1000, with maxiterate.

```
echo "aligning sequences using MAFFT"

mafft --auto --reorder --maxiterate 1000 sequence.fasta  > seqalign.fasta
```

The file seqalign.fasta was viewed using jalview

```
echo "viewing the alignment"

jalview seqalign.fasta
```

Using iqtree, a maximum likelihood phylogenetic tree was generated with seqalign.fasta as input

```
echo "constructing a phylogenetic tree using iqtree"

iqtree -s seqalign.fasta -nt AUTO -m TEST -bb 1000
```

A phylogenetic tree was viewed and annotated with the seqalign.fasta.treefile as input using FigTree

```
echo "opening the tree file"

figtree *.treefile
```

Results & Discussion

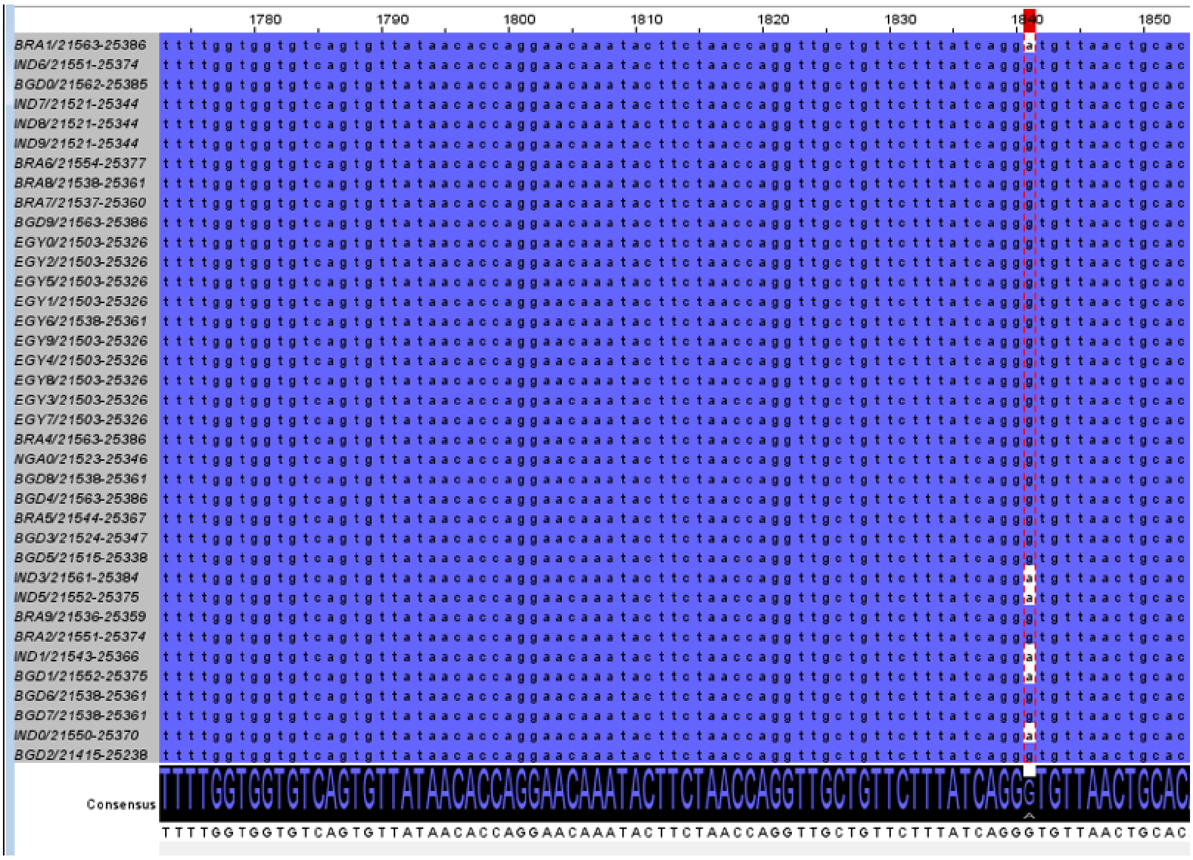
The genome retrieval from NCBI of 10 sequences from each country was successful for all the countries analysed, with exception of Nigeria that had one sequence reposited (Table 1).

Table 1. Countries and accession number of genomic sequences of SARS-2 collected from NCBI.

Country	Sequences no.	Accession number
Bangladesh	BGD0 -BGD9	MT476385.1, MT502774.1, MT509958.1, MT539158.1,

		MT539159.1, MT539160.1, MT566435.1, MT566436.1, MT566437.1, MT576639.1
Brazil	BRA0 - BRA9	MT126808.1, MT350282.1, MT710714.1, MT738101.1, MT738173.1, MT807936.1, MT827074.1, MT827075.1, MT827190.1, MT827202.1
India	IND0 - IND9	MT012098.1, MT050493.1, MT358637.1, MT415320.1, MT415321.1, MT415322.1, MT415323.1, MT435079.1, MT435080.1, MT435081.1
Egypt	EGY0 - EGY9	MT510690.1, MT510691.1, MT510692.1, MT510693.1, MT510694.1, MT510695.1, MT510696.1, MT510697.1, MT510698.1, MT510699.1
Nigeria	NGA0	MT576584.1

When aligned to the reference genome, high similarity was observed in the highly conserved regions within the spike of SARS-COV-2 from the different countries. However, mismatches could be spotted (Figure 2), evidencing mutations possibly resulting from different evolutionary paths in SARS-COV-2 across the world. Such alterations in the genome sequences can result in distinct protein translation and different strains of the same species.



```
BRA1/21563-25386   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcaggatgttaactgcac
IND6/21551-25374   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
BGD0/21562-25385   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
IND7/21521-25344   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
IND8/21521-25344   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
IND9/21521-25344   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
BRA6/21554-25377   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
BRA8/21538-25361   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
BRA7/21537-25360   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
BGD9/21563-25386   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
EGY0/21503-25326   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
EGY2/21503-25326   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
EGY5/21503-25326   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
EGY1/21503-25326   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
EGY6/21538-25361   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
EGY9/21503-25326   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
EGY4/21503-25326   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
EGY8/21503-25326   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
EGY3/21503-25326   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
EGY7/21503-25326   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
BRA4/21563-25386   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
NGA0/21523-25346   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
BGD8/21538-25361   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
BGD4/21563-25386   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
BRA5/21544-25367   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
BGD3/21524-25347   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
BGD5/21515-25338   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
IND3/21561-25384   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcaggatgttaactgcac
IND5/21552-25375   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcaggatgttaactgcac
BRA9/21536-25359   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
BRA2/21551-25374   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
IND1/21543-25366   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcaggatgttaactgcac
BGD1/21552-25375   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcaggatgttaactgcac
BGD6/21538-25361   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
BGD7/21538-25361   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
IND0/21550-25370   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcaggatgttaactgcac
BGD2/21415-25238   ttttggtggtgtcagtgttataacaccaggaacaaatacttctaaccaggttgctgttctttatcagggtgttaactgcac
         Consensus TTTTGGTGGTGTCAGTGTTATAACACCAGGAACAAATACTTCTAACCAGGTTGCTGTTCTTTATCAGGGTGTTAACTGCAC
                   TTTTGGTGGTGTCAGTGTTATAACACCAGGAACAAATACTTCTAACCAGGTTGCTGTTCTTTATCAGGGTGTTAACTGCAC
```

Figure 2. Visualization of the alignment using Jalview. Diagram evidencing in red dashed line an example of mismatch G-A (in white) in sequences of SARS-COV-2 from Brazil, Bangladesh and India.
]

The pattern of branching in the analysis has revealed that the virus went through mutations that had an effect on the distance from the original root in China, indicating a variability (Figure 3). The closest strains from China were found in Brazil, India and Bangladesh. Five out of ten sequences of Egypt SARS-COV-2 appeared to cluster and form a distinct branch, distant from the root. Six out of ten sequences from Bangladesh are closely related from the most recent common ancestor and originated a new branch with two different strains. Strains from India and Brazil share the most recent common ancestor in different branches of the tree, along with the only Nigerian sequence.

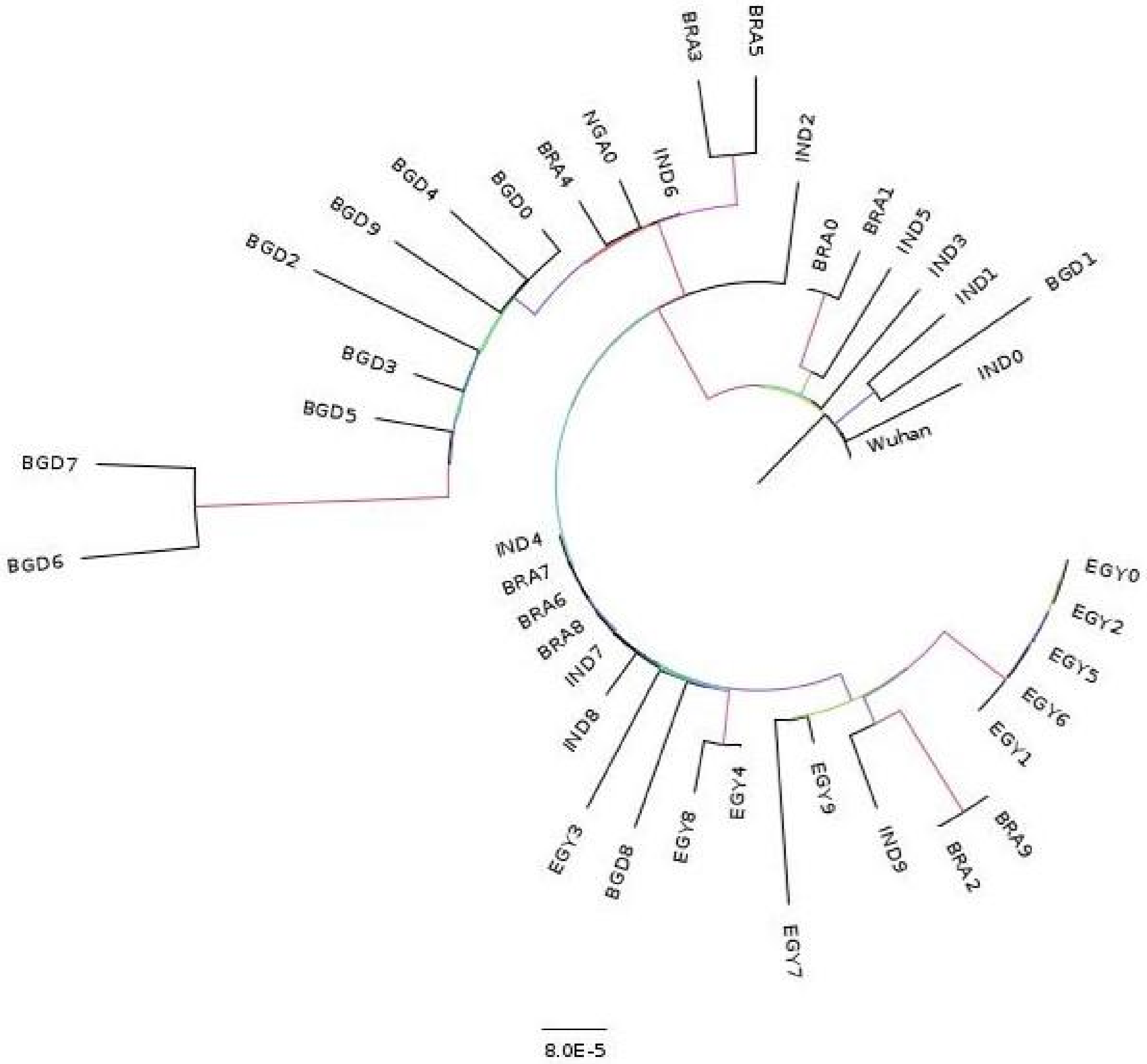


Figure 3. Phylogenetic tree of 41 SARS-COV-2 sequences. Diagram representing evolutionary phylogenetic relationships between strains of SARS-COV-2 in Bangladesh (BGD), Brazil (BRA), India (IND), Egypt (EGY) and Nigeria (NGA) related to China (Wuhan).

Conclusion

The analysis was done with contribution from most of the team members, with exception to those that had specific problems that would require longer time to assess. The analysis was not computationally complex, and all the team members followed all the steps made to complete it. We have had daily zoom calls for discussing methods, troubleshooting and evaluating the results. With a simple phylogenetic analysis with pre-existing data, we were able to show the known variability on the genome of SARS-COV-2 in a small sampling from our teammates' home countries. Since we were on a tight schedule, we used only a small sample which limits the scope of the study. Given a more liberal time frame, a much larger ensemble could have been studied which would have provided greater insight into the dynamic nature of mutation and transmission of the virus.

Contributions

Conceptualization	@WGatua, @Pawan
Creating script	@Pawan, @WGatua
Data collection	@WGatua, @Pawan
Software install and/or running analysis	@WGatua, @Pawan, @Harinath, @Josiah24, @Maruf, @Caroline, @Sefunmi, @Tracyallen, @Ahmad
Report	@Maruf, @Caroline, @Josiah24
Video Presentation	@Sefunmi, @Tracyallen, @Ahmad, @Harinath

Code

The scripts for the above pipeline are available at https://github.com/MountainMan12/SARS-Cov2-phylo