HB HackBio

PHYLOGENETIC ANALYSIS OF MAJOR COVID-19 LINEAGE GENOMES

HACKBIO FINAL PROJECT REPORT

PREPARED BY
ANNABEL SHARON ANYANG
2022 HACKBIO GENOMICS FELLOW

Phylogenetics Analysis of Major COVID-19 Lineage Genomes

About

This project was selected from the HackBio genomics project list by the participant (named below) for the final stage of the 2022 July/August HackBio Genomics cohort

Name of Participant

Annabel Sharon Anyang

Overview

The aim of this stage was to provide participants with the opportunity to solve real problems encountered in academic and industrial settings.

Task Background

Coronavirus (Covid-19) is an infectious disease caused by the SARS-COV-2 virus (WHO, 2022). It is a highly contagious, zoonotic viral respiratory pathogen responsible for the 2020 global health crisis (Wikipedia, 2022)

At the beginning of the COVID-19 pandemic in Nigeria, three key institutions collaborated and secured an Illumina Hi-Seq device for rapid sequencing and genotyping of samples in 4 Nigeria states (Kwara, Oyo, Osun and Ogun). Upon sequencing, they generated full-length sequences from the first 10 positive samples.

As a public health bioinformatician, the task was to explore the epidemiology of COVID-19 in Nigeria and answer the following questions:

- What is the genetic relatedness of COVID-19 strains in Nigeria with previously reported isolates globally?
- What is/are the most likely geographical sources of the samples in each state?

These questions will explore the phylogenetic relationship between the circulating strains in the country.

- Considering the mortality rate data from John Hopkins COVID-19 Dashboard, Does the circulating strain have anything to do with the mortality rate?

Methodology

Data Collection

There is a glut of viral sequencing data being deposited daily on online repositories such as NCBI, GISAID, ENSEMBL, amongst others. The National Center for Biotechnology Information (NCBI) is a major platform for bioinformatics tools and services, this was the chosen data base source for this project. Human SARS-COV2 data was retrieved from NCBI. Genomics sequences was collected from across 10 different countries globally. The selected countries include Brazil, China, Egypt, France, Great Britain, India, Italy, South Africa, Russia, and the USA. The Nigeria dataset was retrieved from a GitHub repository specified by HackBio in the project description. The COVID-19 reference genome from Wuhan, China was also retrieved.

Country		**Sequence ID**	**Year**
Brazil		MT126808.1, MT827074.1	2020
China		ON965801.1, ON965799.1, ON965480.1, ON965458.1	2022
Egypt		MT510690.1, MT510698.1	2020
France		OP160217.1, OP160215.1	2022
		OP160091.1, OP160071.1	2021
GBR		OL989059.1, OL989057.1, OK041130.1, OK041131.1	2020
India		OP132964.1, OP132963.1, MT012098.1, MT415322.1	2020
Italy		OP002140.1, ON974847.1,	2022
		MT509660.1, MT509655.1	2020
Nigeria	**Nig States**		2020
	OYO	OY045B-CV35B	2020
	OYO	OY045A-CV35	2020
	OYO	OY035-CV34	2020
	OYO	OY008-CV29	2020
	OSUN	OS129-CV21	2020
	OSUN	OS125-CV20	2020
	OSUN	OS122-CV18	2020
	OSUN	OS116-CV17	2020
	OSUN	OS085-CV14	2020
	OSUN	OS075-CV12	2020
	OSUN	OS070-CV11	2020
	OSUN	OS068-CV10	2020

	OGUN	OG007-CV22	2020
	KWARA	KW017-CV24	2020
Russia		MW305251.1, MW305250.1, MZ047312.1, MZ044345.1	2020
S/Africa		ON287357.1, ON287306.1	2021
USA		OP236311.1, OP235619.1, OP235618.1, OP236310.1	2022
REF-WUHAN		NC_045512.2	

Table 1. Metadata Table. Table of the Genomic Sequences of Human SARS-Cov-2 Data used in Project.

States	Cases	Deaths
Oyo	10,219	202
Ogun	2,103	31
Osun	3,311	92
Kwara	1,088	27

Table 2. Mortality Rate Data from the John Hopkins Covid-19 dashboard.

Software Packages Used

- MEGA X v11.0
- SDT v1.2

Analysis

Genomic sequences were downloaded directly from the NCBI database in a FASTA format. The sequences were compiled into a list in a text editor and outputted as a FASTA file. This was then sent to MEGA X (version 11.0) which was first used to build the alignment of the sequences. The multiple sequence alignment was carried out using the MUSCLE algorithm.

Phylogenetic analysis was conducted in MEGA X using the aligned COVID-19 sequences of the Nigerian sequences and the selected global sequences sourced from NCBI database as input. The

tree was built using the Neighbor-Joining tree method based on the Jones-Taylor-Thornton model to compute the evolutionary distances. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches.

Pairwise sequence comparison analysis was done in SDT using the Muscle alignment program to cluster sequences using a neighbor joining tree.

Results & Discussion

The multiple sequence alignment revealed that there are 4342/30018 conserved regions and 25570/30018 variable regions (Figure 1.). Considering the high percentage of mismatches that indicate a high degree of mutations that possibly have arisen from the different worldwide evolutionary paths of the COVID-19. Mismatches in genomic sequence ultimately affects the protein translation outcome and may even result in different variants of the virus, as is the case with COVID-19.

However, the poor quality of some of the Nigerian sequences in the analysis have to be considered. This could simply be the reason for the high percentage of mismatches in the alignment.

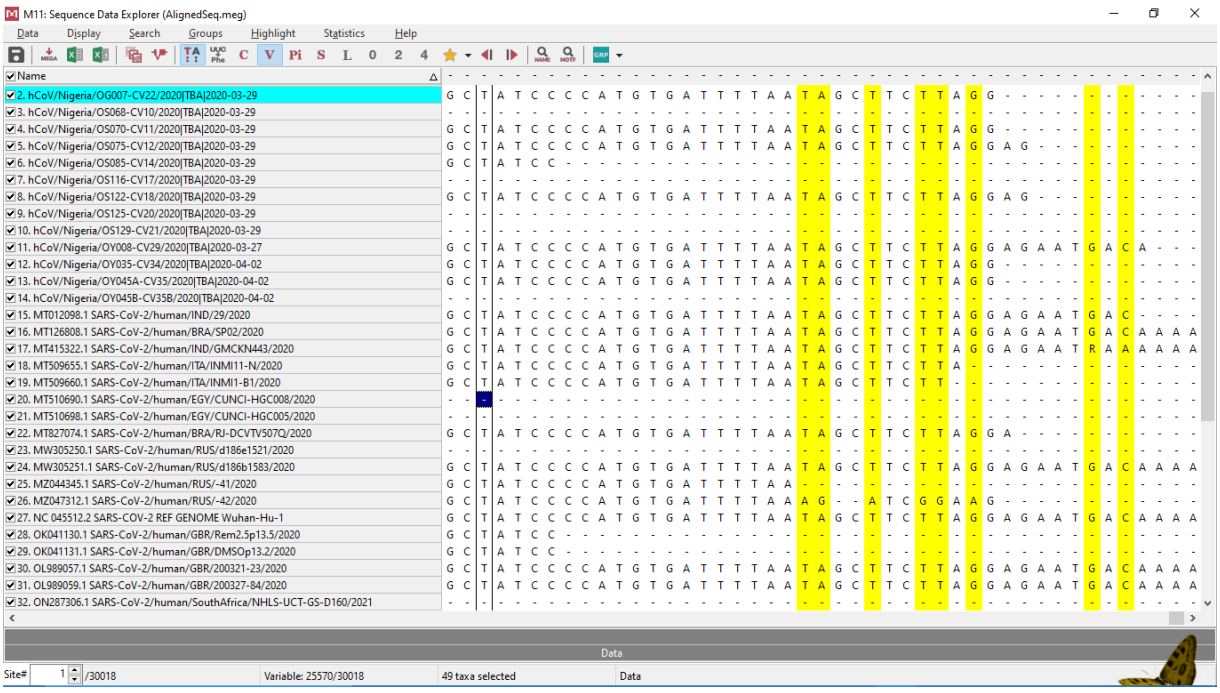

Figure 1. Visualization of Multiple Sequence Alignment. Diagram showing the sequence data explorer of the aligned sequences. The regions highlighted in yellow shows some of the variable regions from the alignment.

Phylogenetic analysis of COVID-19 genomes revealed that all the Nigerian isolates included in this study have a complex, diverse evolutionary relationship both with each other and also with the international isolates (Figure 2.).

The genetic relatedness of some of the Nigerian isolates cluster together, while some other Nigerian isolates cluster with a group of viruses from diverse geographical sources. This clade of Nigerian SARS-COV-2 isolates are related to strains that have been found in a range of countries including strains detected in Brazil, Egypt, Italy and the UK.

This suggests the introduction of exotic SARS-COV-2 strains into the country which most likely occurred due to poor COVID-19 screening of international travelers into the country. This highlights the need for the implementation of stricter COVID-19 screening and quarantine protocols for international travelers. In addition, more COVID-19 surveillance should be encouraged among researchers to further understand the true status, epidemiological extent and impact of COVID-19 in Nigeria.

When considering the mortality rate data from the John Hopkins dashboard, Oyo state recorded the highest number of deaths. One can use the phylogeny tree to postulate on the probable links in the dynamics between the circulating strain and the mortality rate. Oyo isolates can be observed to be in clades with foreign isolates from countries such as Brazil and Great Britain. Countries that have had massive numbers in Covid-19 related mortality. Furthermore, the length of branching of certain Oyo state isolates signifies that those isolates have undergone a lot of evolution over time in comparison to the other isolates in the same clade.

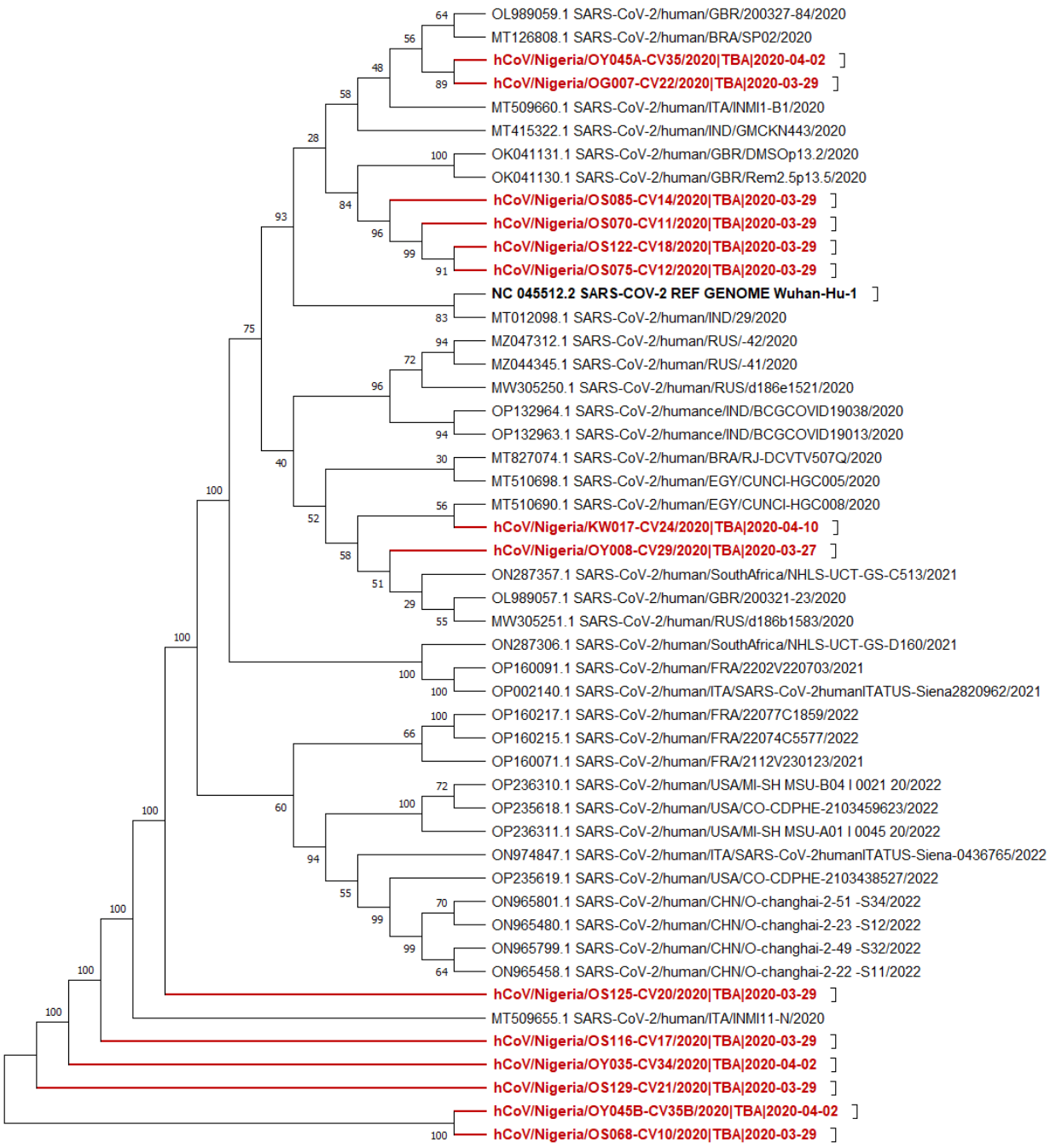

Figure 2. Phylogenetic Tree of 49 SARS-COV-2 Sequences. Diagram depicting the evolutionary phylogenetic relationship between human SARS-COV-2 Nigerian sequences (Highlighted in red) with the reference Wuhan sequence (Sequence in Bold) and other global SARS-COV-2 strains.

The Pairwise sequences comparisons of the Nigerian isolates produced 95-100% identity with a large proportion of the sequences from countries such as China, Great Britain, Italy, USA amongst others. (Figure 1). However, there were a couple of outliers as certain Nigerian sequences showed poor similarity values ranging from 74% to even as 12%.

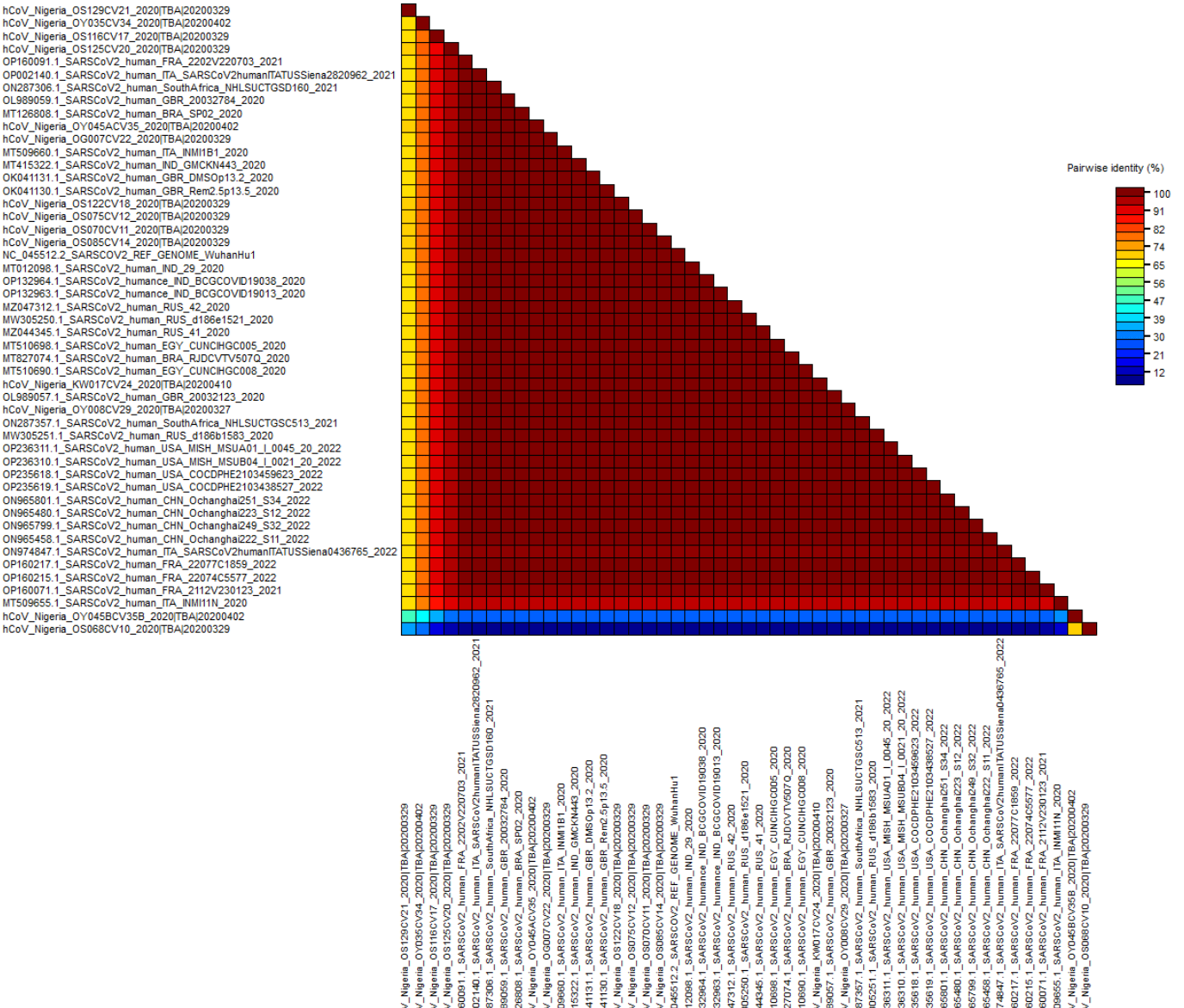

Figure 3. Sequence Demarcation Tool Based Pairwise sequence comparisons. Diagram showing the pairwise sequence comparisons of Nigerian SARS-COV-2 sequences with global SARS-COV-2 sequences.

Conclusion

This project elucidates that the Nigerian SARS-COV-2 strains are genetically related to global SARS-COV-2 strains. It also tries to explore a link between the circulating strain and mortality rates in the Nigerian states of origin for the dataset used in this project.

Future directions for studying SARS-COV-2 in Nigeria include further sequence analysis, ideally more full-length genome sequencing to better elucidate the evolutionary relationship between the Nigerian SARS-COV-2 strains and other global strains, as well as to explore the transmission of the virus and track mutational events in the viral genome.

Resources

The Nigerian viral genome dataset used in this project is available at https://github.com/josoga2/sc_data/tree/main/covid%20data

Mortality rate data from John Hopkins Covid-19 Dashboard is available at https://www.google.com/search?q=mortality+rate+of+covid+19+in+Nigeria&oq=mortality+rate+of+covid+19+in+Nigeria&aqs=chrome..69i57.13247j0j7&sourceid=chrome&ie=UTF-8

References

Coronavirus disease (COVID-19). (2022). Retrieved 21 August 2022, from https://www.who.int/health-topics/coronavirus#tab=tab_1

COVID-19. (2022). Retrieved 21 August 2022, from https://en.wikipedia.org/wiki/COVID-19

HackBio team Venter technical report. Retrieved August 2020, from https://github.com/HackBio-Internship/wale-home-tasks/raw/main/Report .pdf

This project and the resources are available at the GitHub repo:

asanyang/HackBio-Genomics: Projects and tasks from the 2022 HackBio Genomics Workshop (github.com)