UGWU PASCHAL
+2348166207095 | ugwupaschal@gmail.com
https://www.linkedin.com/in/paschal-ugwu-52abb6229/
HackBio Genomics Final Project
Topic: Phylogenetic analysis of major covid-19 lineage genomes.
August, 2020

Abstract

The recent COVID-19 pandemic has greatly affected the world negatively, both medically and financially. In Nigeria, Johns Hopkins corona resource center has confirms 262,664 cases with 3,147 death. Next generation sequencing has been useful in understanding this disease. In this project, I applied Next Generation Sequencing techniques: Multiple Sequence Alignment and phylogeny analysis to compare samples from Nigeria with those from Italy, Russia, France, China, UK, and USA. I used NCBI virus sequence repository to search for sample and reference genomes, and MEGA 11 software to perform Multiple Sequence Alignment and construct a phylogeny tree. From this tree, I observed the relationship between collected samples and their mutation rates. This study confirms the claim by Johns Hopkins and will be useful to other scientists in future research on the nature, structure, and mutability of coronavirus.

Introduction

The remarkable capacity of some viruses to adopt to new hosts and environment is highly dependent on their ability to generate de novo diversity in a short period of time. Rates of spontaneous mutation vary amply among viruses. RNA viruses mutate faster than DNA viruses, single stranded viruses mutate faster than double strand virus, and genome size appears to correlate negatively with mutation rate. Viral mutation rates are modulated at different levels, including polymerase fidelity, sequence context, template secondary structure, cellular microenvironment, replication mechanisms, proofreading, and access to post-replicative repair (Sanjuan and Domingo-Calap, 2016).

Members of the family of coronavirus have the largest genome of all RNA viruses, and express up to 29 protein establishing complex interactions with the host proteome (Perrin-Cocoon *et al.,* 2020). The coronavirus spike protein is a multi-functional molecular machine that mediates coronavirus entry into host cells (Li, 2016).

Coronavirus causes respiratory and digestive diseases in vertebrates. The recent pandemic caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-COV-2), is taking a heavy toll on society and planetary health, and illustrates the threat emerging coronavirus can pose to the well-being of humans and other animals (Nova, 2021).

The coronavirus disease 2019 (COVID-19) spread throughout China and received worldwide attention. On 30 January 2020 World Health Organization (WHO) officially declared the COVID-19 pandemic as a public health emergency of international concern (Guo *et al*., 2020). The COVID-19 pandemic shocked the world, overwhelming the health systems of even high-income countries. In Nigeria, the situation elicited social and medical responses from the public and government respectively (Amzat *et al*., 2020). According to the Johns Hopkins University and Medicine, coronavirus resource center cases and deaths data in Nigeria on 19th August 2022, the total number of confirmed cases of infection by the virus is 262,664 with deaths amounting 3,147.

Various next generation sequencing (NGS) based strategies have been successfully used in recent past for tracing origins and understanding the evolution of infectious agents, investigating the spread and transmission chains

of outbreaks, as well as facilitating the development of effective and rapid molecular diagnostic tests, and contributing to the hunt for treatments and vaccines (Chiara *et al*., 2021).

In this project, I aim to use Multiple Sequence Alignment to compare multiple genomes of coronavirus from Nigerian states. I will also perform phylogenic analysis for the samples in Nigeria, in comparison to the samples from Italy, Russia, France, China, UK, and USA. Then, I will check the most likely geographical sources of the samples from Nigerian states. Finally, I will determine if the circulating strain has anything to do with mortality rates of data collected from John Hopkins.

Methods and Results

I used NCBI virus as sequence repository to search for sample genomes and reference genome by geolocation. For each of the locations: Nigeria, Italy, Russia, France, China, UK, and USA, 2 genomes including the reference genome were collected as shown in the table below.

Table 1: Reference and sample genomes used for multiple sequence alignment and phylogenetic analysis.

Accession number	Species	Molecule type	Length	Geolocation	Collection date
NC_045512 (Ref Sequence)	SARS-COV-2	ssRNA(+)	29903	China	2019-12
OP247736	SARS-COV-2	ssRNA(+)	29720	USA: Texas	2022-07-29
OP247737	SARS-COV-2	ssRNA(+)	29376	USA: Illinois	2022-07-30
OX274044	SARS-COV-2	ssRNA(+)	29844	UK:England	2022-07-31
OX274045	SARS-COV-2	ssRNA(+)	29844	UK:England	2022-08-02
OP160034	SARS-COV-2	ssRNA(+)	29818	France	2021-12-21
OP160035	SARS-COV-2	ssRNA(+)	29823	France	2022-01-19
OP002141	SARS-COV-2	ssRNA(+)	29733	Italy	2021-05-20
ON974845	SARS-COV-2	ssRNA(+)	29632	Italy	2022-06-06
ON965361	SARS-COV-2	ssRNA(+)	29784	China	2022-04-07
ON692745	SARS-COV-2	ssRNA(+)	3822	Russia	2021-11-12
ON692746	SARS-COV-2	ssRNA(+)	3822	Russia	2020-11-06
ON564648	SARS-COV-2	ssRNA(+)	29489	Nigeria	2021-07-29
ON564649	SARS-COV-2	ssRNA(+)	29798	Nigeria	2021-06-02

I downloaded the selected nucleotide sequences in FASTA format and made sure the genome sequence title was in GeneBank format (GB). In the next step of analysis, I used MEGA 11 software to perform Multiple Sequence Alignment of the selected samples against the consensus reference genome. The resulting aligned sequence was downloaded in MEGA format. The aligned sequences were then used to construct a phylogeny tree.

The evolutionary history was inferred using the UPGMA method (Sneath and Sokal, 1973). The bootstrap consensus tree inferred from 1000 replicates (Felsenstein, 1985) is taken to represent the evolutionary history of the taxa analyzed (Felsenstein, 1985). Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura and Kumar, 2004) and are in the units of the

number of base substitutions per site. This analysis involved 14 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated (complete deletion option). There were a total of 2029 positions in the final dataset. Evolutionary analyses were conducted in MEGA11 (Tamura *et al*., 2022).

Figure 1: The phylogeny tree constructed from the the aligned sample and reference genomes.

Discussion

Multiple sequence alignment is a prerequisite to virtually all comparative genomic analyses, including the identification of conserved sequence motifs, estimation of evolutionary divergence between sequences, and inference of historical relationships among genes and species (Kumar and Filipski, 2007).

A dendrogram, or phylogeny tree, is a branching diagram or "tree" showing the evolutionary history between biological species or other entities based on their genetic characteristics. Species or entities joined together by nodes represent descendants from a common ancestor and are more similar genetically. With the advancement of DNA sequencing technologies, phylogeny trees have been used widely in infectious disease control to depict the genetic similarities and differences between strains and variants of a certain disease pathogen. In this case, our disease pathogen is SARS COV-2. This enables us to know whether coronavirus disease occurring in different areas are from the same strain, and provides key information on the sources of infection and how the disease may be transmitted (Carroll *et al*., 2014).

From Figure1 above, we draw the following inference: firstly, the genomes from the Nigerian states differ but belong to the same ancestry as depicted by the bootstrap value of 100 that connects them. Secondly, the strain I collected from Nigeria, closely relates to the deadly strain from Wuhan-Hu-1 China which caused massive death rates

in the beginning of the pandemic. This must have attributed to the massive death rates Johns Hopkins data shows for Nigeria.

Conclusion

The origin and identity of coronavirus will be useful to study the outbreak of the virus and plan practical measures to study how it can be treated. The phylogenic analysis will also help to determine the relationship between existing and previously sequenced coronaviruses. This study confirms the data collected from Johns Hopkins as seen in the mutability that exists in our samples.

Consequently, scientists can use these results to further their research on the nature, structure, and mutability of coronavirus.

Acknowledgement

I wish to thank HackBio and all its reviewers for providing me the opportunity to learn so much within a short period of time. I specially wish to thank Mr. Wale and all my team members for their support throughout this workshop.

References

Amzat, J., Aminu, K., Kolo, V. I., Akinyele, A. A., Ogundairo, J. A., & Danjibo, M. C. (2020). Coronavirus outbreak in Nigeria: Burden and socio-medical response during the first 100 days. *International Journal of Infectious Diseases*, *98*, 218-224.

Carroll, L. N., Au, A. P., Detwiler, L. T., Fu, T. C., Painter, I. S., & Abernethy, N. F. (2014). Visualization and analytics tools for infectious disease epidemiology: a systematic review. *Journal of biomedical informatics*, *51*, 287-298.

Chiara, M., D'Erchia, A. M., Gissi, C., Manzari, C., Parisi, A., Resta, N., ... & Pesole, G. (2021). Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. *Briefings in Bioinformatics*, *22*(2), 616-630.

Felsenstein J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783-791.

Guo, Y. R., Cao, Q. D., Hong, Z. S., Tan, Y. Y., Chen, S. D., Jin, H. J., ... & Yan, Y. (2020). The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status. *Military medical research*, *7*(1), 1-10.

Kumar, S., and Filipski, A. (2007). Multiple sequence alignment: in pursuit of homologous DNA positions. *Genome research*, *17*(2), 127-135.

Li, F. (2016). Structure, function, and evolution of coronavirus spike proteins. *Annual review of virology*, *3*(1), 237.

Nova, N. (2021). Cross-species transmission of coronaviruses in humans and domestic mammals, what are the ecological mechanisms driving transmission, spillover, and disease emergence?. *Frontiers in Public Health*, *9*.

Perrin-Cocon, L., Diaz, O., Jacquemin, C., Barthel, V., Ogire, E., Ramière, C., ... & Vidalain, P. O. (2020). The current landscape of coronavirus-host protein–protein interactions. *Journal of translational medicine*, *18*(1), 1-15.

Sanjuán, R., and Domingo-Calap, P. (2016). Mechanisms of viral mutation. *Cellular and molecular life sciences*, *73*(23), 4433-4448.

Sneath P.H.A. and Sokal R.R. (1973). *Numerical Taxonomy*. Freeman, San Francisco.

Tamura K., Nei M., and Kumar S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences* (*USA*) 101:11030-11035.

Tamura K., Stecher G., and Kumar S. (2021). MEGA 11: Molecular Evolutionary Genetics Analysis Version *Molecular Biology and Evolution* https://doi.org/10.1093/molbev/msab120.