BIOINFORMATIC ANALYSIS OF SOME GLOBAL STRAINS OF SARS-COV-2 GENOMES REVEAL DISTRIBUTION OF SIX MAJOR PHYLOGENETIC GROUPS

[1]Thomas, Benjamin Thoha, [1]Efuntoye, Moses Olusola, [2]Folorunsho, Jamiu Bello, [1]Popoola, Omolara Dorcas and [1]Tajudeen, Ahmed Olanrewaju

[1].Department of Microbiology, Olabisi Onabanjo University, Ago Iwoye, Ogun State, Nigeria
[2].Directorate of Health Sciences, Olabisi Onabanjo University Health Centre, Ago Iwoye

ABSTRACT

The global strains of SARS-CoV-2 were evaluated using bioinformatic approach in order to infer their levels of variation. Phylogeny, toggling and secondary structure were carried out using maximum likelihood, percentages and SOPMA tools respectively while tajima's neutrality test and maximum likelihood estimate for transition-transversion bias were estimated using standard recommended procedures. Results obtained delineate the evaluated global SARS-Cov-2 into six distinct haplotypes with several conserved sites at different levels of percentages ranging from 50-100%. The tajima's neutrality test and maximum likelihood for transition-transversion bias were found to be approximately 5.85 and 0.59 respectively while the secondary structure of these six haplotypes shows similar representation of protein content with approximately 67% (cluster 3-6) showing higher content of alpha helix than other protein structures. The remaining 33% (cluster 1 and 2) however also resembles themselves as their random coil was found to be higher than other protein structures. Results of this study have shown that, even though, there is a little variation occurring in the SARS-Cov-2 genome, the rate at which is happening is a cause for concern and there might be SARS-Cov-2 variants escaping vaccines in the nearest future

Keywords; Bioinformatics, SARS-CoV-2, phylogenetics, secondary structures.

Accepted Date: 10 November, 2021

Introduction

It is no more news that the world was stormed by the sudden outbreak of coronavirus disease, which begins in November 2019 in Wuhan, China as a pneumonia-like infection until it was confirmed as a lower respiratory tract disease affecting the lower respiratory tract of patients (Zhou *et al.,* 2020). Consequent to its confirmation to the World Health Organization (WHO) on the 29th of December 2019, it was soon observed as a major catastrophic threat sickening nearly two million people worldwide and causing approximately 125,000 death globally as at 14th of April, 2020 (WHO, 2020). Currently, the global infection and death rates stood at 21, 412, 643 and 764470 respectively while the total recovering rate from the disease was estimated to be 14,198,089 as at 15th of August, 2020 (Covid-19 Worldometer, 2020). In Nigeria, the number of confirmed infection so far according to the NCDC was estimated to be 48,445 with not less than 973 death (NCDC, 2020) accruing from it thereby resulting in fear and panic.

The panic resulting from this global pandemic further exacerbated the impact of this public health crisis on both the global economy as well as our social lives and irresistibly, the whole world was forced to go on a compulsory holiday (Ozili and Arun, 2020). This obligate intracellular parasite is a family of viruses that causes respiratory illnesses ranging from asymptomatic to mildly symptomatic and severe cases with many more mild cases that might have resolved without any serious diagnosis (Lu *et al.*, 2020; Zhou *et al.,* 2020).

What is most compounding about this disease is that despite the level of spread on the global scale and the economic hardship resulting from the outbreak and/or the aggressive spread of this disease, there is not yet a definite therapy for this public health problem and treatment is only been administered symptomatically (Cascella *et al., 2*020). It is however true that the world governments are engaging in developments of countermeasures to abrogate the annihilating impacts of this disease and it has been evaluated that strict shutdowns would save several lives (BBC News, 2020). Currently, remedial schemes to combat the infection are only supportive and prophylaxis targeted at minimizing the spread of the virus is still the best weapon so far (Cascella *et al.,* 2020).

Despite the significant threat pose by coronavirus, there are yet to be any known vaccines proven to protect the body against covid-19 and this apparently appears the whole world is still vulnerable at the moments. However, the presence of potent vaccines would train the body immune systems on how to fight the virus so they should not become sick (UNICEF, 2020) and subsequently ease lockdown while social distancing can also be relaxed. The development of successful vaccines could hitherto be contingent upon the genetic variability of the global coronavirus strains as effort to develop a good vaccine may be undercut if the virus changes in a way that lets it evade the vaccine (Lemaire *et al.,* 2009). This study was therefore aimed at comparing the different strains of SARS-CoV-2 from different part of the world in order to infer their phylogenetic groups using bioinformatic analysis.

Materials and Methods

Twenty SARS-CoV-2 strains were retrieved from the gene bank at NCBI by using nucleotide BLAST program (http: ncbi.gov.blast.cgi). These sequences were subsequently aligned using the CLUSTAL W in MEGA 17.0 (Tamura *et al.,* 2007). The evolutionary history was deduced by using the Maximum Likelihood method based on the Jukes and Cantor model (1969). The tree with the highest log likelihood (-15230.57) is shown. The initial trees which were heuristically searched were axiomatically computed through Neighbour-Join and BioNJ algorithms matched with a matrix of pairwise distance that were approximated by the maximum composite likelihood approach (MCL).The sequences were subsequently toggled at percentages ranging between 50-100% to determine the level of conserved sites. The secondary structure of the RNA sequences was predicted following the recommended SOPMA bioinformatic tool (Geourjon and Deleage, 1995). The tajima's neutrality test were determined out by removing all missing data and gaps prior to subjecting the data to a statistical test as described by Kumar *et al.,* (2018). The maximum likelihood estimate of Transition/Transversion bias was determined under the Kimura (1980) parameter model.

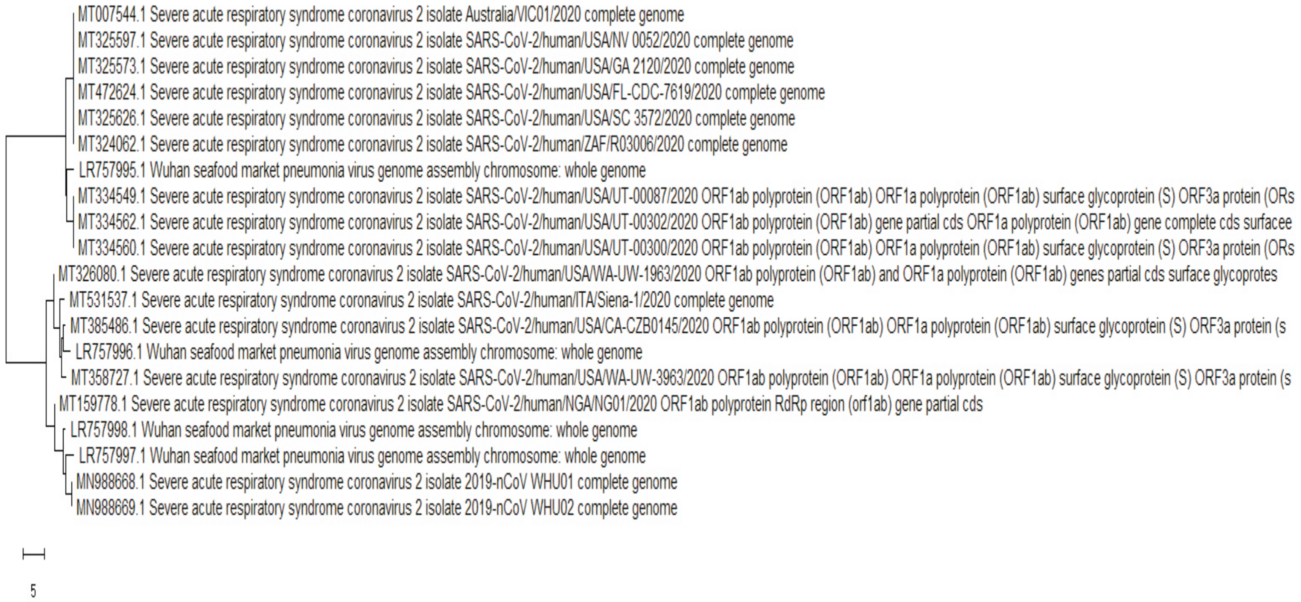
MT007544.1 Severe acute respiratory syndrome coronavirus 2 isolate Australia/VIC01/2020 complete genome
MT325597.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/NV 0052/2020 complete genome
MT325573.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/GA 2120/2020 complete genome
MT472624.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/FL-CDC-7619/2020 complete genome
MT325626.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/SC 3572/2020 complete genome
MT324062.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/ZAF/R03006/2020 complete genome
LR757995.1 Wuhan seafood market pneumonia virus genome assembly chromosome: whole genome
MT334549.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/UT-00087/2020 ORF1ab polyprotein (ORF1ab) ORF1a polyprotein (ORF1ab) surface glycoprotein (S) ORF3a protein (ORs
MT334562.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/UT-00302/2020 ORF1ab polyprotein (ORF1ab) gene partial cds ORF1a polyprotein (ORF1ab) gene complete cds surfacee
MT334560.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/UT-00300/2020 ORF1ab polyprotein (ORF1ab) ORF1a polyprotein (ORF1ab) surface glycoprotein (S) ORF3a protein (ORs
MT326080.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/WA-UW-1963/2020 ORF1ab polyprotein (ORF1ab) and ORF1a polyprotein (ORF1ab) genes partial cds surface glycoprotes
MT531537.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/ITA/Siena-1/2020 complete genome
MT385486.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/CA-CZB0145/2020 ORF1ab polyprotein (ORF1ab) ORF1a polyprotein (ORF1ab) surface glycoprotein (S) ORF3a protein (s
LR757996.1 Wuhan seafood market pneumonia virus genome assembly chromosome: whole genome
MT358727.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/WA-UW-3963/2020 ORF1ab polyprotein (ORF1ab) ORF1a polyprotein (ORF1ab) surface glycoprotein (S) ORF3a protein (s
MT159778.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/NGA/NG01/2020 ORF1ab polyprotein RdRp region (orf1ab) gene partial cds
LR757998.1 Wuhan seafood market pneumonia virus genome assembly chromosome: whole genome
LR757997.1 Wuhan seafood market pneumonia virus genome assembly chromosome: whole genome
MN988668.1 Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV WHU01 complete genome
MN988669.1 Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV WHU02 complete genome
5

Figure 1 : Phylogenetic analysis of the global strains of SARS-CoV-2

Key: MT007544.1 (Australian strain), MT325597.1, MT325573.1, MT472624.1, MT325626.1 (USA), MT324062.1(South Africa), LR757995.1 (Wuhan, China), MT334549.1 , MT334562.1, MT334560.1, MT326080.1 (USA), , MT531537.1(Siena, Italy), MT385486.1 (USA), LR757996.1 (Wuhan, China), MT358727.1 (USA), MT159778.1 (Nigeria), LR757998.1, LR757997.1,MN988668.1, MN988669.1 (China

Table 1: Results from Tajima's Neutrality Test

m	S	p_s	Θ	π	D
20	915	1.000000	0.281870	0.675703	5.851387

Abbreviations: m = number of sequences, n = total number of sites, S = Number of segregating sites, $p_s = S/n$, $\Theta = p_s/a_1$, π = nucleotide diversity, and D is the Tajima test statistic

The phylogenetic analysis of the evaluated strains of SARS-CoV-2 categorized the analyzed strains into six different clusters (cluster 1-6). Members of cluster 1 include strains from Australia, United States and South Africa. Cluster 2 has strains from Wuhan sea market in china and the United State of America. The members of both cluster 1 and 2 are however found to be closely related proteomically as both show higher content of random coil than other protein structures. Alpha helix was consequently found to be second most abundant protein structure in these strains of organisms. Member of the third cluster (cluster 3) are only found in China and they include strains with accession number (LR 75799.7, MN 988668.1 and MN 988669.1). cluster 4 contain only one strain with accession number LR 757997.1 that also domicile in China (Wuhan sea food market). Cluster 5 include members circulating in Nigeria, United State of America and Siena in Italy while the members of cluster 6 are mainly found in United States and China. All the members of SARS-CoV-2 found in cluster 3 to cluster 6 are higher in alpha helix than other protein structures with random coil being the second most abundant. The toggling of these sequences revealed several conserved sites with the tajima's neutrality test as well as maximum likelihood estimate for transition-transversion bias approximately found to be 5.85 and 0.59 respectively.

DISCUSSION AND CONCLUSION

The use of bioinformatic tools for deducing important gene function and structure has been long documented (Zuckerkandl and Pauling, 1965;Thomas *et al.,* 2016). In this study, phylogenetic analysis of the selected global strains of SARS-CoV-2 categorized them into six distinct clusters to describe their patterns of relatedness as well as their evolutionary relationships (Thomas *et al.,* 2016). According to Tamura *et al.,* (2007), phylogenetic appraise the moment different organisms deviate from their common ancestor. In this study, phylogenetic analysis revealed six distinct categories of organisms as an indication that mutation is slowly accumulating in these newly discovered strains of coronavirus. The aforesaid mutations have appeared verificatory for the natural selection of organisms (Pascarella and Argoi, 1992; Benner *et al.,* 1993*;* Wolf *et al*., 2007; Thomas *et al*., 2019). The fact that the strains of SARS-CoV-2 from Australia, United States and South Africa clustered together further reveal that these organism share a common ancestor (Tamura et al., 2007). Cluster 2 isolates which include strain from Wuhan sea market and United States though clustered separately from those in cluster 1, evidence from SOPMA tool analysis reveal both strains sharing significant similarity as both have higher content of random coil than alpha helix to demonstrate higher proteomic similarities. This observation pointed to the fact that even though mutation is accumulating genomically, the level of proteomic disparity is still relatively low. Members of cluster 3-6 also demonstrated similar proteomic activity as alpha helix was found to be the most abundant as against what was observed in cluster 1 and 2. This observation further reinforced relatively lower proteomic mutation in these strains of SARS-CoV-2 as against what was observed genomically. However, sharp variation was found proteomically between strains in cluster 1 and 2 and those found in cluster 3-6. Such variation hitherto may be emphasizing several types of transition/transversion mutation occurring and may consequently be surmising that the degree of microbial mutation divergence are not limited to among species but even to different strains of the same organisms as well as inside equipotent gene at varied moments (Moxon and Thaler, 1997; Drake *et al*., 1998; Thomas *et al*.,2019). Variations observed both genomically and proteomically may be an indication of gradual accumulation of mutation and so providing significant insight into how these proteins would function (Romero *et al*., 2006; Zhang *et al.,* 2011) and evolve (Wolf *et al.,* 2001) in future. The observation of several conserved sites noticed during toggling of the genomic sequence is a critical pointer that even though mutation is occurring gradually, the level of such mutation is not significant enough to halt the development of vaccine against them. Although, assemblage of mutation as well as several defective multiplication of viruses in their host coupled with the possibility of carrying over such levels of defection through a population influences their natural selection . However, coronavirus has proofreading machinery that curtails the margin of error and the rapidity of mutation. The positive valve of the tajima's test of neutrality is simpatico with the global trademark of positive selection (Zeng *et al*., 2017). This positive selection may be ascribed to significant alteration in the amino acid sequence of proteins despite the rare presence of such non synonymous substitution (Sobrinho and de Brito, 2010).

In conclusion, the results obtained from this study have shown that the strains of SARS-CoV-2 circulating globally is accumulating mutation at a very low level that cannot halt development of a potent vaccine against them. Our study however shows that the spike gene in SARS-CoV-2 is relatively conserved and the little variation observed are mostly maintained by positive selection. Consequently, the rate at which the mutation is accumulating is a cause for concern and might be suggesting the possibility of the evolution of some SARS-Cov-2 variants escaping vaccines in the nearest future

REFERENCES

Benner, S.A., Cohen, M.A. & Gonnet, G.H. (1993). Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol*. 229, 1065–1082.

British Brocasting Corporation. (2020). Coronavirus update retrieved @ www.worldometers.info on the 15th of August, 2020.

Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S.C. & Napoli, R.D. (2020). Features, Evaluation and treatment coronavirus. In :StatPearls (Internet). Treasure Island (FL): StatPearls Publishing.

Drake, J.W., Charlesworth, D., Charlesworth, D. & Crow, J.F. (1998). Rates of spontaneous mutation. *Genetics* 148, 1667–1686.

Geourjon, C. & Deleage, G. (1995). SOPMA significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Appl. Biosci*. 11(6), 681-684.

Gojobori, T., Li, W.H. & Graur, D. (1982). Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol*. 18, 360–369.

Hartwell, L.H., Goldberg, M.L., Fischer, J.A. & Hood, L. (2011). Genetics: from genes to genomes, 4th ed. McGraw-Hill Education, New York, NY, pp. 1–730.

Implications for virus origins and receptor binding. *Lancet*, 395: 565–574.

Jukes, T.H. & Cantor, C.R. (1969). Evolution of protein molecules. In: Munro H.N., editor. *Mammalian Protein Metabolism*. Academic Press; New York. pp. 21–132.

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci*. Unit. States Am. 78:454–458.

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci*. Unit. States Am. 78, 454–458.

Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura K. (**2018**). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* 35:1547-1549.

Lemaine, D., Barbosa, T. & Rihet, P. (2012). Coping with genetic diversity: contribution of pathogen and human genomics to modern vaccinology. *Brazilian Journal of medical and biological research*, 45: 376-385.

Lu, R. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus:

Moxon, E.R. & Thaler, D.S. (1997). The tinkerer's evolving tool-box. *Nature* 387, 659–662.

NCDC coronavirus update. (2020) Coronavirus Update retrieved @ www.covid19.ncdc.gov.ng on the 15th of August, 2020.

Ozili, P. & Arun, T. (2020). Spill over of COVID-19: impact on the Global Economy. *SSRN Electronic Journal,* pp 1-24.

Pascarella, S. & Argos, P. (1992). Analysis of insertions/deletions in protein structures. *J. Mol. Biol*. 224, 461–471.

Petrov, D.A. & Hartl, D.L. (1999). Patterns of nucleotide substitution in Drosophila and mammalian genomes. *Proc. Natl. Acad. Sci*. Unit. States Am. 96, 1475–1479.

Podlaha, O. & Zhang, J. (2003). Positive selection on protein-length in the evolution of a primate sperm ion channel. *Proc. Natl. Acad. Sci*. Unit. States Am. 100, 12241–12246.

Qi, J., Wiljeratne, A.J., Tomsho, L.P., Hu, L.I., Schuster, S.C. & Ma, H. (2009). Characterization of meiotic crossovers and gene conversion by whole-genome sequencing in Saccharomyces cerevisiae. *BMC Genomics* 10, 475.

Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T., Obradovic, Z. & Dunker, A.K. (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc. Natl. Acad. Sci*. Unit. States Am. 103, 8390–8395.

Sobrinho, I.S.& de Brito, R.A. (2012). Positive and purifying selection influence the evolution of doublesex in the Anastrepha fraterculus species Group. *PLoS One* 7 (3), 1–10.

Spatz, S.J. & Rue, C.A. (2008). Sequence determination of a mildly virulent strain (CU-2) of Gallid herpesvirus type 2 using 454 pyrosequencing. *Virus Genes* 36, 479–489.

Tamura, K., Dudley, J. & Nei, M. (2007). MEGA 4 Molecular Evolutionary Genetics Analysis (MEGA software version 4.0. *Molecular Biology and Evolution*; 24, 1596-1599.

Thomas, B.T., Agu, G.C., Oso, O.A., James, E.S., Davies, A. & Dele-Osinbanjo, T.A. (2016). Evolutionary and Secondary Structure of Multi Drug Resistance Class 1 Integron from Gram Negative Bacteria, *International Journal of Microbiological Research* 7 (1): 21-29, 2016.

Thomas, B.T., Ogunkanmi, L.A., Iwalokun, B.A. &

Popoola, O.D. (2019). Transition-Transversion mutation in the polyketide synthase gene of *Aspergillus* Section *Nigri*. *Heliyon*, 5(6) e01881-1-6.

UNICEF. (2020). Coronavirus Update retrieved @ www.unicef.org on the 15th of August, 2020

Wolff, J., Bresch, H., Cholmakov-Bodechtel, C., Engel, G., Garais, M., Majerus, P., Rosner, H. & Scheuer, R. (2000). Contamination of foods and consumer exposure. *Arch. Lebensm. Hyg*. 51, 81–128.

World Health Organization (W.H.O). (2020). Coronavirus disease 2019 (COVID-19) Situation Report 84.retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/ (13 April 2020, date accessed).

Worldometer. (2020). Coronavirus Update retrieved @ www.worldometers.info on the 15th of August, 2020.

Zeng, K., Shi, S. & Wu, C.I. (2007). Compound tests for the detection of hitchhiking under positive selection. *Mol. Biol. Evol*. 24, 1898–1908.

Zhang, Z. & Gerstein, M. (2003). Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res*. 31, 5338–5348.

Zhang, Z., Huang, J., Wang, Z., Wang, L. & Gao, P. (2011). Impact of indels on the flanking regions in structural domains. *Mol. Biol. Evol*., **28**:291–301.

Zhang, Z., Huang, J., Wang, Z., Wang, L. & Gao, P. (2011a). Impact of indels on the flanking regions in structural domains. *Mol. Biol. Evol*. 28, 291–301.

Zhang, Z., Xing, C., Wang, L., Gong, B. & Liu, H. (2011b). Indel FR: a database of indels in protein structures and their flanking ***regions***. *Nucleic Acids Res*. 40, 512–518.

Zhao, Z. & Boerwinkle, E. (2002). Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res*. 12, 1679–1686.

Zhou, X.L., Yang, X.G. & Wang T.U. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature***,** 579 (7798): 270–273.

Zuckerkandl, E. & Pauling, L. (1965). Molecules as documents of evolutionary history. *J. Theor. Biol.. Appl Environ Microbiol*., 8(2): 357-366.