Analysis of SARS-CoV-2 genomes from across Africa reveals potentially clinically relevant mutations.

Benson C. Iweriebor PhD[1], Olivia S. Egbule PhD[2], Samuel O. Danso PhD[3], Eugene Akujuru[4], Victor T. Ibubeleye[5], Christabel I. Oweredaba[6], Theodora Ogharanduku[7,8], Adama Traore MD[9], Alexander Manu MD PhD[10] and Modeline N. Longjohn[11].

Affiliations

1. Safako Makgatho Health Sciences University, Pretoria, South Africa
2. Department of Microbiology, Delta State University, Abraka, Delta State, Nigeria
3. University of Edinburgh Medical School, Scotland, United Kingdom
4. Department of Haematology, Blood Transfusion and Immunology, University of Port Harcourt, Rivers State, Nigeria.
5. African Centre for Excellence in Public Health and Toxicological Research University of Port Harcourt, Rivers State, Nigeria.
6. School of Science, Engineering and Environment, University of Salford, United Kingdom.
7. Department of Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom
8. Department of Epidemiology, Delta State University, Abraka, Delta State, Nigeria
9. World Bank, Cote d'Ivoire.
10. Centre for Maternal & Newborn Health, Liverpool School of Tropical Medicine, Liverpool, United Kingdom
11. Department of Biochemistry, Memorial University of Newfoundland, Newfoundland and Labrador, Canada.

Address correspondence to:

mnlongjohn@mun.ca

Department of Biochemistry

Memorial University of Newfoundland

232 Elizabeth Avenue,

St. John's, NL CANADA A1B 3X9

Funding details

MNL is supported by an NSERC held by Dr. Sherri Christian at Memorial University, and a school of graduate studies funding.

Disclosure Statement

No potential competing interest was reported by the authors.

Acknowledgement

MNL conceived and designed the project, while BCI carried out the bioinformatics analysis with MNL and OSE. The manuscript was written by MNL, BCI and OSE, with proofreading and additional inputs by SOD, EA, VTI, CIO, TO, AT, AM.

ORCID

MNL: 0000-0001-8278-8556

BCI: 0000-0002-2621-6462

OSE: 0000-0003-2364-3327

VTI: 0000-0003-3079-0413

AM: 0000-0001-5230-6413

Abstract

SARS-CoV-2 is a betacoronavirus, the etiologic agent of the novel Coronavirus disease 2019 (COVID-19). In December 2019, an outbreak of COVID-19 began in Wuhan province of the Hubei district in China and rapidly spread across the globe. On March 11th, 2020, the World Health Organization officially designated COVID-19 as a pandemic. Across the continents and specifically in Africa, all index cases were travel related. Thus, it is crucial to compare COVID-19 genome sequences from the African continent with sequences from COVID-19 hotspots (including China, Brazil, Italy, United State of America and the United Kingdom). To identify if there are distinguishing mutations in the African SARS-CoV-2 genomes compared to genomes from other countries, including disease hotspots, we conducted *in silico* analyses and comparisons. Complete African SARS-CoV-2 genomes deposited in GISAID and NCBI databases as of June 2020 were downloaded and aligned with genomes from Wuhan, China and other SARS-CoV-2 hotspots. Using phylogenetic analysis and amino acid sequence alignments of the spike and replicase (NSP12) proteins, we searched for possible targets for vaccine coverage or potential therapeutic agents. Our results showed a similarity between the African SARS-CoV-2 genomes and genomes in countries including China, Brazil, France, the United Kingdom, Italy, France and the United States of America. This study shows for the first time, an in-depth analysis of the SARS-CoV-2 landscape across Africa and will potentially provide insights into specific mutations to relevant proteins in the SARS-CoV-2 genomes in African populations.

Keywords

SARS-CoV-2, COVID-19, Africa, genome, phylogeny

1. Introduction

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is the aetiologic agent for the so-called Coronavirus disease 2019 (COVID-19). COVID-19 first emerged in Hubei Province, China in December 2019; the second coronavirus to have arisen from China in the last two decades [1] [2]. Within months, SARS-CoV-2 spread rapidly both nationally and internationally, resulting in global outbreaks of pneumonia-like symptom clusters and death in some cases. Consequently, the World Health Organization publicly declared COVID-19 a pandemic on 11th March 2020. Besides SARS-CoV-2, four strains of coronavirus pathogenic to humans are known to be associated with clinical symptoms, including common cold and pneumonia [4]. Two other coronaviruses, SARS-CoV (responsible for the Severe acute respiratory syndrome – SARS) and MERS (which causes the Middle Eastern respiratory syndrome) were notable for causing infections with higher fatalities and deaths [5] [6]. However, outbreaks caused by SARS-CoV and MERS (though still circulating) were relatively limited compared to those caused by SARS-CoV-2 [7].

The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is a novel betacoronavirus (family *Coronaviridae*, genus *Betacoronavirus*, species *Severe acute respiratory syndrome-related coronavirus*) [1] [2]. SARS-CoV-2 is an enveloped virus which has a positive sense single-stranded viral RNA genome of approximately 30kb (13). The SARS-CoV-2 genome consists of 14 open reading frames (ORFs) that are preceded by transcriptional regulatory sequences (TRSs), which form transcriptional units [8]. The largest transcriptional unit includes the ORF1a and b regions, which translates into two precursor polyproteins (PP1a and PP1b), that are subsequently cleaved by viral proteases into functional non-structural proteins (nsps) [8]. Other ORFs encodes four main structural proteins including surface spike (S) glycoprotein, envelope (E) protein, membrane (M) protein and nucleocapsid (N) protein *(See Fig. 1)*, each of which has specific functions that dictate SARS-CoV-2 interactions and biology [9] [10] [11].

The pathophysiology and virulence of SARS-CoV-2 is heavily reliant on the function of nsps and other structural proteins. Specifically, nsps are able to prevent host immune response [1]. For structural proteins, the E protein is crucial in viral pathogenicity and promotes viral assembly and release [2]. In addition, the S protein facilitates viral entry into target cells, having an affinity for the angiotensin converting enzyme 2 (ACE2) receptor expressed on the surface of cells of

the lungs, heart, renal system [3] [2]. Since the onset of the pandemic, scientists have contributed over 34,300 genome sequence to the Global Initiative on Sharing Avian Influenza Data (GISAID) and NCBI GenBank databases as at May 2020. Of this, Africa has only contributed 224 genome sequences. This is mostly due to Africa's reduced scientific research capacity, despite an abundance of expertise. As of 12th May 2020, Africa accounts for about 2% of the global COVID-19 caseload [4]. This has a huge implication for the African continent, especially as the continent currently lacks the health care system capacity to effectively manage a full-blown outbreak.

Currently, there is no known cure for COVID-19, with most pandemic response systems across the globe not being adequate in curbing the spread of the virus. Hence, there is a need for the development of vaccine(s) and drug(s), which research teams across the globe are in a race against time to develop. However, this process is a multi-step process which rely heavily on specific molecular and pharmacological data. For insights into the molecular basis of developing vaccines and drugs, the pool of genome sequences of SARS-CoV-2 from all over the world are currently being mined and analyzed by researchers. Conversely, many vaccines being development are currently at phase III of trials, implying that the vaccine development may have commenced before index case was reported in some African countries. This in turn raises an important question; would the treatments and/or vaccines when available be effective in Africa? A previous study showed that residues near some amino acid (Lysine 31, Tyrosine 41, 82-84, and 353-357) in human ACE2 were important for SARS-CoV-2 binding [5]. To investigate mutations in the amino acid residues of SARS-CoV-2 circulating in Africa, we analyzed whole genome sequences of SARS-CoV-2 from Africa deposited in the GISAID and NCBI GenBank databases. Because all cases in Africa arose from travel, it became compelling to compare COVID-19 sequences generated from the African continent with strains from COVID-19 hotspots Wuhan, China and other countries across the different regions of the world. We conducted complete genome analyses of viruses from Africa and compared them with reference sequences from Wuhan and other parts of the world using phylogenetic analysis. We also investigated the amino acid sequences of the spike protein and the replicase (NSP12) to identify similarities and differences. This study is showing for the first time, an in-depth analysis of the SARS-CoV-2 molecular epidemiology in Africa and will potentially provide insights into the mutations/variations in relevant proteins of viruses circulating in Africa.

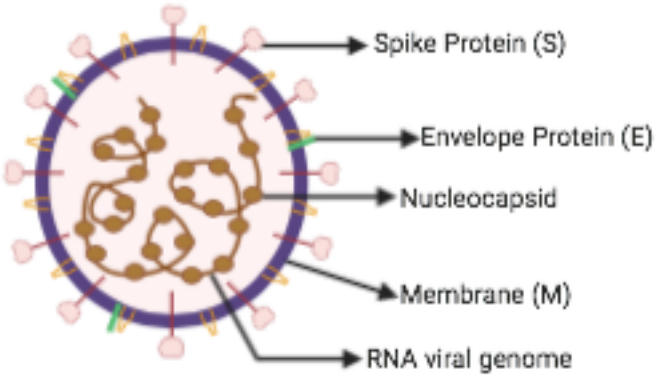

Figure 1: SARS-CoV-2 viral structure

2.0 Materials and Methods

2.1 Search Strategy for Identification of COVID-19 SEQUENCES

We searched the SARS-CoV-2 NCBI (https://www.ncbi.nlm.nih.gov/sars-cov-2/) and GISAID (https://www.gisaid.org/) databases for SARS-CoV-2 genomic sequences from Africa, specific countries within Asia, Oceania, North America and Europe. For our analyses, nucleotide and amino acid sequences of the SARS-CoV-2 genome were used. Also included were spike, NSP12 and ORF1ab nucleotide and protein sequences respectively.

2.2 Inclusion Criteria

Only complete genomes with at least 29kb of length were selected for the phylogenetic analysis while ORF1ab and surface glycoproteins with 1273 amino acids were downloaded from the Coronavirus genomes NCBI website. Table 1 shows the details of curated proteins of novel coronavirus data downloaded. Efforts were made to download proteins for spike and NSP12 from different geographical regions of the world, including Africa, Asia, Europe, North America, South America, and Oceania. All curated surface glycoproteins from Africa with omitted or ambiguous amino acids residues were eliminated from the analysis.

	African Country	**Country of Origin of first imported case**	**City of Index case**	**Date of index case recording**
1	Egypt	China	Cairo	14th February, 2020
2	Nigeria	Milan, Italy	Lagos	27th February, 2020
3	Ghana	Norway/Turkey/France	Accra	12th March, 2020
4	Senegal	France	Dakar	2nd March, 2020
5	Gambia	United Kingdom	Banjul	17th March, 2020
6	Democratic Republic of Congo (DRC)	France	Kinshasa	10th March, 2020
7	Algeria	Italy	Blida	17th February, 2020
8	South Africa	Milan, Italy via Dubai	Hilton, KwaZulu-Natal, Johannesburg	1st March, 2020
9	Uganda	Dubai	Entebbe	18th March,2020

Table 1. African Countries and the country of Origin of the index case: *All initial cases of SARS-CoV-2 in African countries were travel related. The source of the index cases for the sequences used in this study are outlined in the table above.*

2.3 Amino Acid alignment of the surface glycoprotein (spike protein)

All African surface glycoproteins of the SARS-CoV-2 obtained from the NCBI GenBank were aligned using ClustalW and compared to a reference strain from Wuhan, China. Thereafter, a consensus sequence was created for the African sequences as well as a global consensus sequence for all spike proteins from different geographical regions of the world respectively. A

total of 44 African surface glycoproteins sequences along with a reference strain that were analyzed are as follow: YP_009724390_Wuhan, QJY78020.1,QJY78032.1, QJY78044.1, QJY78056.1, QJY78068.1, QJY78092.1, QJY78104.1, QJY78116.1, QJY78128.1, QJY78140.1, QJY78152.1, QJY78164.1, QJY78176.1, QJZ28114.1, QJZ28126.1, QJZ28347.1, QJZ28359.1, QKE11078.1, QJX45321.1, QJX45333.1, QJX45344.1, QJX45356.1, QJX45380.1QKK12863.1, QKI36913.1, QKD20860.1,QKG82981.1,QKG82993.1,QKF95995.1, QIZ15537.1, EPI_ISL_418241_Algeria, EPI_ISL_418217_Senegal,EPI_ISL_418207_Senegal, EPI_ISL_418217_Senegal, EPI_ISL_421573_SouthAfrica, EPI_ISL_418241_Algeria, EPI_ISL_421574_SouthAfrica, EPI_ISL_428857_Gambia, EPI_ISL_428855_Gambia, EPI_ISL_430819_Egypt, EPI_ISL_437339_DRC, EPI_ISL_447234_DRC, EPI_ISL_413550_Nigeria, EPI_ISL_430820_Egypt.

2.4 Amino acid alignment of the NSP12- Replicase protein

ORF1ab amino acid sequences of SARS-CoV-2 from Africa were downloaded and 1050 NSP12 sequences were identified. Consensus sequence of the African NSP12 was created and aligned with the Wuhan reference strain and identity plot was created with BioEdit. Analysis of the NSP12 also was performed with global sequences of the replicase protein and efforts were made to include representatives of the protein from the six geographical regions of the world. Global consensus was created from these sequences as was done with the African sequences and the consensuses were aligned with the Wuhan strain and identity was plotted with the reference strain as was performed earlier with the African consensus sequence. The 28 replicase proteins from Africa and the reference strain that were analyzed are as follow: YP_009724389.1_Wuhan, QJY78018.1, QJY78030.1,QJY78042.1, QJY78054.1, QJY78066.1, QJY78078.1, QJY78090.1, QJY78102.1, QJY78114.1, QJY78126.1, QJY78162.1, QJY78174.1, QJZ28112.1, QJZ28124.1, QJX45319.1, QJX45331.1, QJX45342.1, QJX45354.1, QJX45366.1, QJX45378.1, QKK12861.1,QKI36911.1,QKD20858.1, QKG82979.1, QKG82991.1,QKF95993.1, QKF96005.1, QIZ15535.1

2.5 Genomes phylogenetic analysis

For phylogenetic analyses, complete SARS-CoV-2 genomes (a total of 56 sequences from Africa and other regions of the world) were obtained from GenBank. Genomes with at least 29kb were used in the analysis, with bat coronavirus genome used as an outgroup. Phylogenetic tree construction by the neighbor joining method was performed using MEGA X software, with bootstrap values of 1000 [6]. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) was indicated next to the branches [7]. The tree was drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Poisson correction method was used to compute the evolutionary distances considering the units of the number of amino acid substitutions per site [8]. All ambiguous positions were removed for each sequence pair (pairwise deletion option). Evolutionary analyses were conducted in MEGA X software while multiple alignment was performed using MUSCLE alignment as implemented in MEGA [9] [10].

3.0 Results

3.1 Amino acid alignment of the surface glycoprotein

Results of the amino acid sequence alignment of African surface glycoproteins against the Wuhan reference strain showed that sequences were almost identical, except for few positions where non-synonymous substitutions occurred. Thirteen mutations were identified, to be different between African and non-African SARS-CoV-2 spike proteins (See Table 2). 90% of the African spike proteins had D614G substitution while 7% had a C677H mutation. L5F and R408I mutations occurred in 2% of all cases. The remaining mutations nine mutations occurred at a rate 1%.

Of the identified mutations, the D614G substitution (shown in the alignment of the amino acid in Figure 2) is the most famous mutation among global SARS-CoV-2 sequences. Analysis of the spike proteins from the different geographical regions of the globe revealed that sequences arising from Asia have predominance of D614G substitution. Also, European sequences have mostly the D614G substitution as well as the Oceania sequences.

There is currently no record of the C677H, V70F, L242F, N354B, A288T, C314R, A653V and S604A mutations in the literature. Therefore, this study is showing for the first time the existence

of these mutations though at a low percentage. However, the L5F, R408I and H49Ymutations have previously been reported in the literature.

S/N	Substitution	Percentage
1	L5F	2%
2	H49Y	1%
3	V70F	1%
4	L242F	1%
5	Y279N	1%
6	N354B	1%
7	A288T	1%
8	C314R	1%
9	R408I	2%
10	D614G	90%
11	C677H	7%
12	A653V	1%
13	S604A	1%

Table 2: *Amino acid substitutions observed in the spike proteins of Coronavirus sequences circulating in Africa.*

Fig 2. Amino acids alignment of the Spike protein of the novel Coronavirus: *Consensus sequences were created from Wuhan, Global and African sequences of spike protein sequences obtained from NCBI GenBank. The consensuses of African and global spike proteins were compared with the reference strain from Wuhan. Dots represent similarity in amino acid while there was no deletion in the protein. There was an amino acid substitution in the African and global sequences at position D614G as shown by an arrow.*

3.2 Amino Acid alignment of the NSP12- Replicase protein

Analysis of the NSP12 of African, global sequences in comparison with Wuhan consensus sequence showed that sequences were almost identical except for a deletion at position P314- in both the African and global sequences as shown by an arrow in Fig 3.

3.3 Phylogenetic analysis of genomes of SARS-CoV-2 from Africa

Phylogenetic analysis of trees based on amino acid sequences of SARS-CoV-2 from different parts of Africa along with global reference strains with bat coronavirus sequence showed that sequences originating from African countries clustered with each other, but overall African sequences did not form a monophyletic cluster as shown in Fig 4. However, there seemed to be multiple strains across South Africa, which did not all align together. This may likely mean that there were multiple introductions of SARS-CoV-2 strains from different regions. Furthermore, the DRC and Algeria also have two different strains while all other African countries have only one strain. Overall, there seems to be multiple African clades which demonstrates local transmission within the African countries.

The number of complete genomes of African SARS-CoV-2 that were found were twenty-eight (check to confirm) in total. These sequences were selected to be included in the tree based on the completeness of the genomes. A tree with showing all sequences are shown in the supplementary figures (supplementary fig. 1).

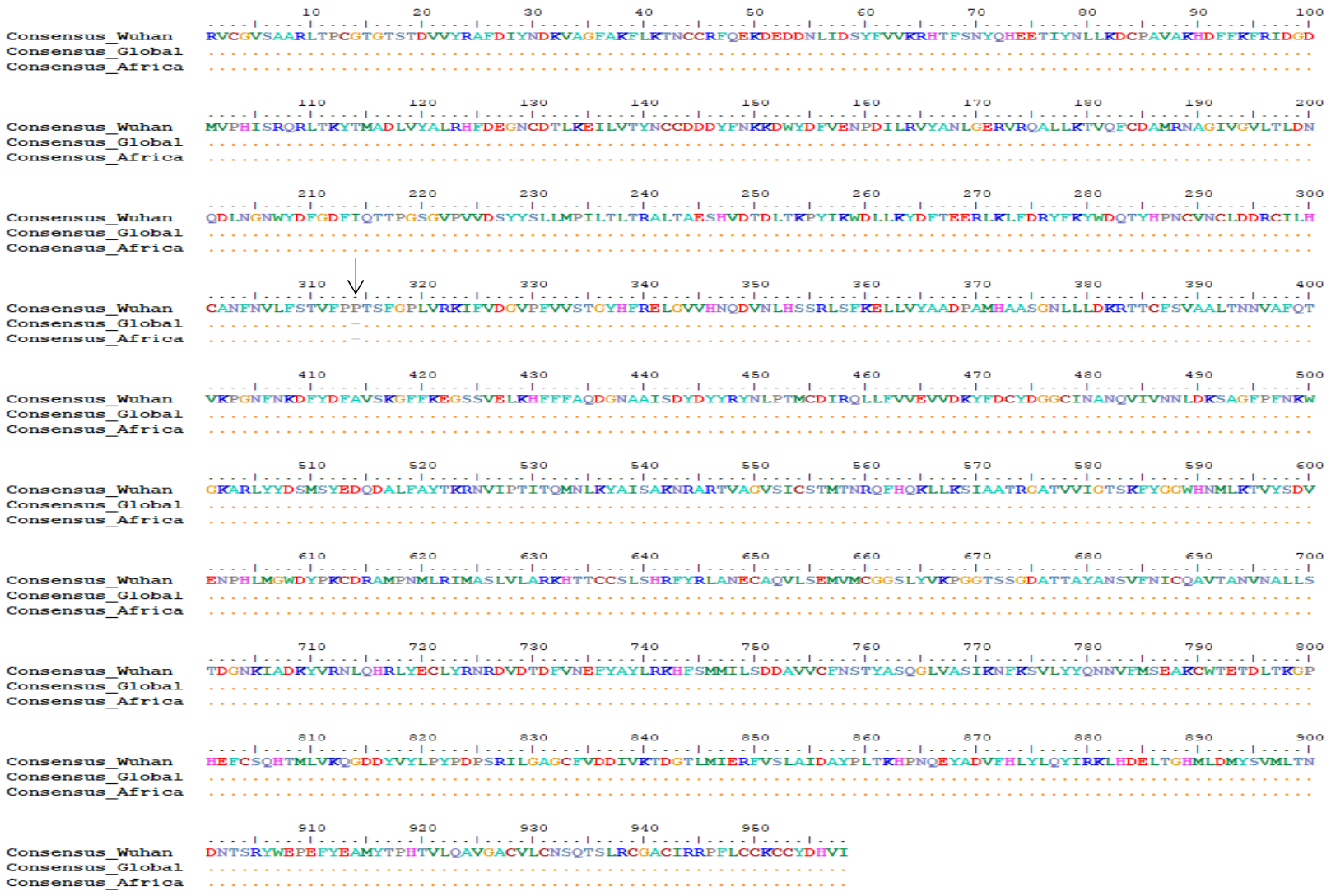

Fig.3. Amino acids alignment of the NSP12 of the novel Coronavirus: *Consensus sequences were created from Wuhan, Global and African sequences of NSP12 within the OR1ab sequences obtained from NCBI GenBank. The consensuses of African and global were compared with the consensus generated from Wuhan ORF1ab sequences. Dots represent similarity in amino acid while the dash denotes a deletion. There was a deletion in the African and global sequences at position P314- as indicated by an arrow.*

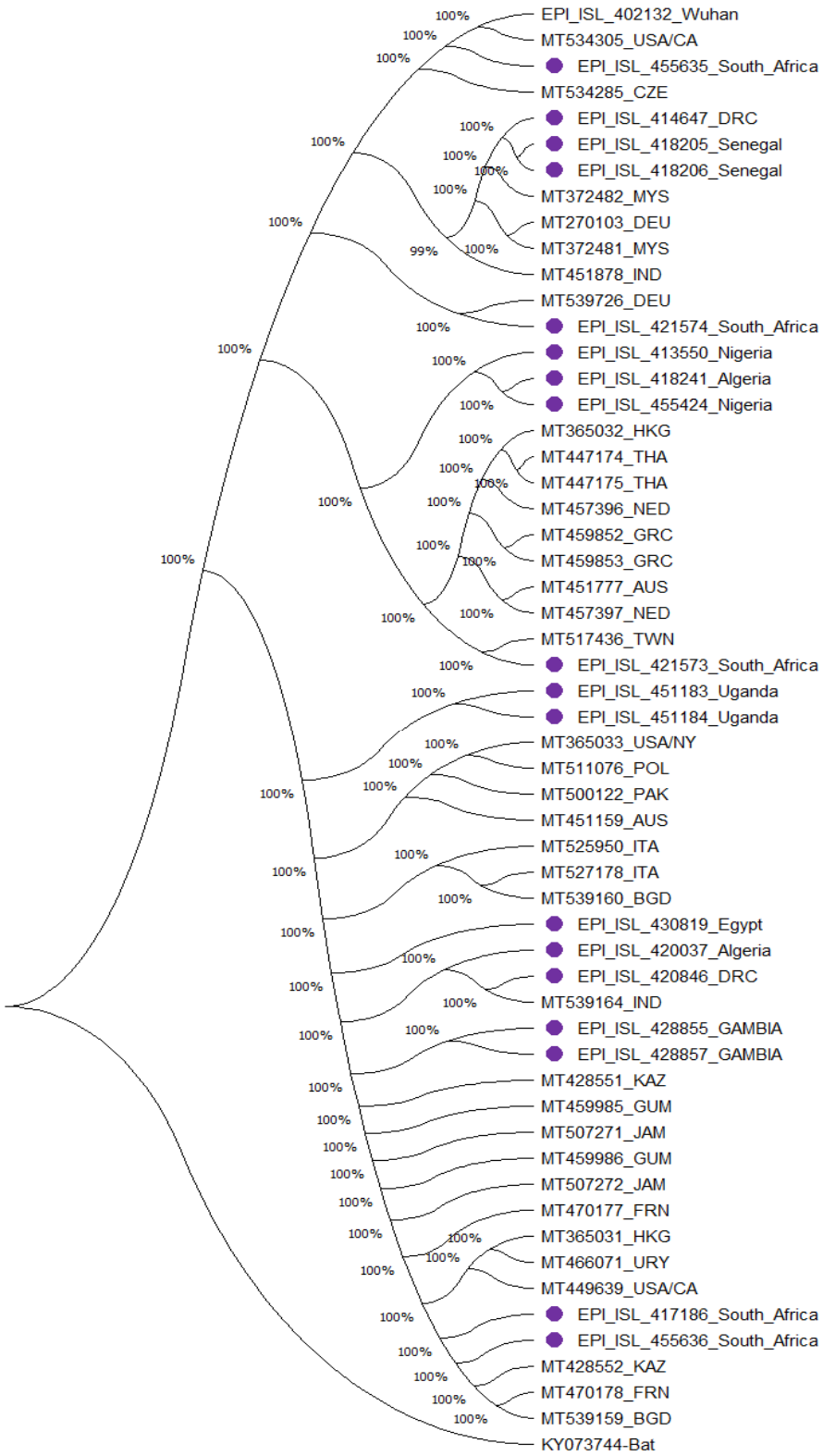

Fig 4. Evolutionary relationships of SARS-CoV-2 genome taxa. *The evolutionary relationship between African SARS-CoV-2 sequence in comparison to reference sequences*

from Wuhan, China, Asia, Australia, Europe and the Americas were inferred using the Neighbor-Joining method. The analysis involved 56 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All ambiguous positions were removed for each sequence pair. There was a total of 27873 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [4]. African sequences are denoted in purple colored balls and abbreviation of the name of country while sequences from Non-African countries were denoted using abbreviation of the country name. This figure shows that the SARS-CoV-2 strain in Africa are evolutionarily related to strain in different countries.

4.0 Discussion

The entry of coronaviruses including SARS-CoV and SARS-CoV-2 into host cells is important for viral infectivity and pathogenicity [11] [12]. SARS-CoV-2 host cell entry is mediated by surface spike proteins recognizing and binding its receptor human angiotensin converting enzyme – hACE2 via its Receptor binding domain (RBD), in a manner that is dependent on host cell pre-activation by host cell proprotein convertase furin [13]. Therefore, SARS-CoV-2 surface spike protein may affect the efficiency of viral spread. Thus, to compare African SARS-CoV-2 spike proteins from Africa to other geographical regions, we aligned the amino acid sequences of African surface spike glycoproteins against the Wuhan reference strain. Our results showed that African SARS-CoV-2 spike proteins were highly identical to those of viruses from other geographical regions. In addition, a majority of African spike proteins have a D614G substitution as shown in the alignment of the amino acid in Figure 2. The D614G substitution was also identified to be predominant in Asian and European SARS-CoV-2 sequences.

D614G substitution is a missense mutation (Aspartic acid (D) to Glycine (G) substitution), which predominantly emerged as a clade in Europe that eventually spread globally [14] [15]. D614G substitution was one of eight missense mutations occurring in the S protein, and suggested to be relatively conserved, with potential effects on S protein interactions and functions [16]. The biological and clinical relevance of this mutation is linked to the type of amino acid and subsequent protein changes that arise. Studies using (*in silico)* structural bioinformatics to assess the effect of D614G mutation on SARS-CoV-2 virulence and epidemiology suggests that this mutation may be neutral to protein function [15]. Based on the structure of SARS-CoV-2 Spike protein (S) determined by Wrapp *et al* using cryo-EM and 3-D modeling of the S-protein structure by Isabel *et al*, it was uncovered that this mutation leads to loss of four inter-chain destabilizing (hydrophobic-hydrophilic) contacts at

a position relatively distal from the RBD with overall null effects on ACE2 and S protein interaction dynamics [16] [15]. However, Yurkovetsky and colleagues showed that the D614G substitution, which is present in >97% of SARS-CoV-2 genomes globally, disrupts contact between S1 and S2 domains of the spike protein and significantly causes conformation shift [17]. Even though this mutation may have increased infectivity of SARS-CoV-2, it does not alter spike protein interactions including with monoclonal antibodies targeting the S protein [17].

The increase in prevalence of D614G mutation across different geographical regions as shown by Korber *et al* may be because this clade has a fitness advantage [18]. This variant of the SARS-CoV-2 virus was shown to grow to higher titers as pseudotyped virions, infect a broad range of human cell types such as lungs, liver and kidney and potentially have higher upper respiratory tract viral loads [19] [18] [20]. The presence of multiple mutations, with the D614G mutation is the most predominant may be important for future studies. Specifically, the D614G substitution could potentially act as an anchor point for divergent S protein mutations especially within the African continent. Currently, efforts are geared toward the formulation of vaccines to combat the global scourge of the coronavirus pandemic. The target of the vaccines is the spike protein, which plays a central role in the viral entry into the host. The implications of the observed D614G mutation on the coverage of any vaccine that will be produced against the virus is not known. However, several studies [29] [21] [22] have reported that antibodies generated due to natural infections of the two variants D614 or G614 showed cross neutralization re-activities thus indicating that the locus might not be within the epitope recognized by the host immune system and therefore not critical for humoral mediated immunity. As the D614G mutation does not prevent cross re-activities of antibodies generated against the variants, it is therefore very unlikely to interfere with the efficacy of any vaccine that will be eventually formulated. Therefore, we are of the courteous optimism that whichever vaccine that is finally approved will be beneficial to the African continent as the observed amino acid substitution is common to majority of the virus circulating in the different regions of the world apart from those from Wuhan, China.

The polymerization of SARS-CoV-2 depends on the main polymerase, the RNA-dependent RNA polymerase (RdRp, also named nsp12). Nsp12 is a 932-residue long enzyme which consists of two conserved domains, the nidovirus RdRp-associated nucleotidyltransferase (NiRAN) and the polymerase domains [23]. Using cryo-EM, Gao and colleagues determined the structure of SARS-CoV-2 Nsp12 bound to Nsp7 and Nsp8 which act as co-factors [24]. Upon comparison of sequences, SARS-CoV and SARS-CoV-2 sequences were found to be

highly identical, including in their interactions with Nsps 7 and 8 [23] [25]. To identify if there are differences between the African SARS-CoV-2 and global NSP12 sequences, we performed a comparative bioinformatic analysis of the NSP12 of African and global sequences in comparison with Wuhan consensus sequence. We were interested in identifying any substitution, deletion or insertion that might interfere with the function of this viral protein. We observed an amino acid deletion at position P314 where there was a deletion in both African and global consensuses in comparison with Wuhan consensus sequence as seen in Fig 3. The deletion occurred in the interface region of the enzyme, which may have significant effect on the interaction between nsp12 and nsps 8 and 9. Though the implications of this deletion on viral fitness and its replicative capacity is not yet understood, more work needs to be done especially to validate this deletion and how it affects viral fitness. Apart from the deletion in the NSP12, no substitution was observed, this is not strange as an enzymatic protein cannot afford much substitutions as it will interfere with viral fitness and its replicative capacity.

Some nsps have been the subject of clinical research into SARS-CoV-2 therapeutics. For instance, Remdesivir (GS-5734) is a nucleotide prodrug, a 1' cyano-substituted adenosine analogue which has been shown to have broad antiviral activity including against coronaviruses MERS-CoV and SARS-CoV, by inhibiting CoV genome replication [26, 27] [28]. Specifically, nsp12 is a major target for remdesivir plays a critical role in the replication of coronavirus genome, hence a mutation in the nsp12 of African SARS-CoV-2 genome may have critical ramifications for the ability of Africans to utilize Remdesivir and any nsp12 targeting drugs.

Genomes of viruses undergo evolution over time. [29]. To analyze the evolutionary similarities between the African versus global SARS-CoV-2 sequences, we performed phylogenetic analysis of full genome of SARS-CoV-2 from different parts of Africa along with global reference strains using bat coronavirus sequence as outgroup. As shown in Fig. 4, the evolutionary history was inferred using the Neighbor-Joining method [6]. Results showed that sequences originating from African countries clustered with each other, but overall African sequences did not form a monophyletic cluster (Fig 4). This lack of a monophyletic cluster may be because the analyzed sequences from the different African country were introduced from other countries. The fact that the genomes formed clades shows and suggests that at this time, there has been divergence of the SARS-CoV-2 from the continent to form new clades. This suggests that there may be multiple strains of SARS-CoV-2 virus

circulating in the Africa continent. Further prospective studies are needed to monitor the evolution of the virus overtime both within the continent and globally.

5.0 Conclusion

The fact that this study showed multiple clades of SARS-CoV-2 in the African continent shows that there are multiple SARS-CoV-2 strains in Africa. However, because these clades/strains are shown to predominantly cluster with strains from other continents, therapeutics (vaccines and drugs) designed based on SARS-CoV-2 strains in these continents could potentially be usable in Africa. This study is showing for the first time, an in-depth analysis of the SARS-CoV-2 landscape in Africa and potentially provide insights into the mutations/variations to relevant proteins of the virus driving the pandemic in African populations.

References

[1] Cascella, M.; Rajnik, M.; Cuomo, A.; Dulebohn, S.C.; Di Napoli, R., "Features, evaluation and treatment coronavirus (COVID-19).," 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK554776/. [Accessed 23rd August 2020].

[2] Francesco Di Gennaro, Damiano Pizzol, Claudia Marotta, Mario Antunes, Vincenzo Racalbuto, Nicola Veronese and Lee Smith , "Coronavirus Diseases (COVID-19) Current Status and Future Perspectives: A Narrative Review," *International Journal of Environmental Research and Public Health,* vol. 17, no. 2690, pp. 1-11, 2020.

[3] Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S. Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, Marcel A.Müller, Christian Drosten, Stefan Pöhlmann, "SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor," *Cell,* vol. 181, no. 2, pp. 271-280, 2020.

[4] A. Union, "Outbreak Brief #17: Coronavirus Disease 2019 (COVID-19) Pandemic Date of Issue: 12 May 2020," 2020. [Online]. Available: https://africacdc.org/download/outbreak-brief-17-covid-19-pandemic-12-may-2020/. [Accessed May 2020].

[5] Wenhui Li, Chengsheng Zhang, Jianhua Sui, Jens H Kuhn, Michael J Moore, Shiwen Luo, Swee-Kee Wong, I-Chueh Huang, Keming Xu, Natalya Vasilieva, Akikazu Murakami, Yaqing He, Wayne A Marasco, Yi Guan, Hyeryun Choe, and Michael Farzan, "Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2," *The EMBO Journal ,* vol. 24, no. 8, pp. 1634-1643, 2005.

[6] Saitou N. and Nei M., "The neighbor-joining method: A new method for reconstructing phylogenetic trees.," *Molecular Biology and Evolution,* vol. 4, pp. 406-425, 1987.

[7] Felsenstein J., "Confidence limits on phylogenies: An approach using the bootstrap.," *Evolution,* vol. 39, pp. 783-791., 1985.

[8] Zuckerkandl, E. and Pauling, L., "Evolutionary Divergence and Convergence in Proteins.," in

Evolving Genes and Proteins, New York, Academic Press, 1965, pp. 97-165.

[9] Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. , "MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms.," *Molecular Biology and Evolution,* vol. 35, pp. 1547-1549., 2018.

[10] Robert C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research,* vol. 32, no. 5, pp. 1792-1797, 2004.

[11] F. Li, "Structure, function, and evolution of coronavirus spike proteins," *Annu. Rev. Virol.,* vol. 3, pp. 237-261, 2016.

[12] S. Perlman, J. Netland, "Coronaviruses post-SARS: Update on replication and pathogenesis.," *Nat. Rev. Microbiol.,* vol. 7, pp. 439-450, 2009.

[13] Jian Shang, Yushun Wan, Chuming Luo, Gang Ye, Qibin Geng, Ashley Auerbach, and Fang Li, "Cell entry mechanisms of SARS-CoV-2," *PNAS,* vol. 117, no. 21, pp. 11727-11734, 2020.

[14] Muthukrishnan Eaaswarkhanth, Ashraf Al Madhoun, Fahd Al-Mulla, "Could the D614G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality?," *Elsevier,* vol. 96, pp. 459-460, 2020.

[15] Sandra Isabel, Lucía Graña-Miraglia, Jahir M. Gutierrez, Cedoljub Bundalovic-Torma, Helen E. Groves, Marc R. Isabel, AliReza Eshaghi, Samir N. Patel, Jonathan B. Gubbay, Tomi Poutanen, David S. Guttman, Susan M. Poutanen, "Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide," *BioRxiv,* 2020.

[16] Daniel Wrapp, Nianshuang Wang, Kizzmekia S. Corbett, Jory A. Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S. Graham, and Jason S. McLellan, "Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation," *Science,* vol. 367, no. 6483, pp. 1260-1263, 2020.

[17] K. E. P. C. T.-T. T. N. Y. W. A. B. W. E. D. A. D. C. C. K. V. S. B. E. S. F. S. J. E. L. J. M. P. C. S. Leonid Yurkovetskiy, "SARS-CoV-2 Spike protein variant D614G increases infectivity and retains sensitivity to antibodies that target the receptor binding domain," *bioRxiv,* 2020.

[18] B. Korber, W.M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E.E. Giorgi, T. Bhattacharya, B. Foley, K.M. Hastie, M.D. Parker, D.G. Partridge, C.M. Evans, T.M. Freeman, T.I. de Silva, C. McDanal, L.G. Perez, H. Tang, A. Moon, "Tracking changes in SARS-CoV-

2 Spike: evidence that D614G increases infectivity of the COVID-19 virus," *CELL,* vol. 20, pp. 30820-5, 2020.

[19] Lizhou Zhang, Cody B Jackson, Huihui Mou, Amrita Ojha, Erumbi S Rangarajan, Tina Izard, Michael Farzan, Hyeryun Choe, " The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity".

[20] Zharko Daniloski, Xinyi Guo, Neville E. Sanjana, "The D614G mutation in SARS-CoV-2 Spike increases transduction of multiple human cell types," *bioRxiv,* 2020.

[21] Ozono, S., Zhang, Y., Ode, H., Seng, T.T., Imai, K.,Miyoshi, K., Kishigami, S., Ueno, T., Iwatani, Y.,Suzuki, T., et al. , "Naturally mutated spike proteins of SARS-CoV-2 variants show differential levels of cell entry.," *bioRxiv,* 2020.

[22] Hu, J., He, C.-L., Gao, Q.-Z., Zhang, G.-J., Cao,X.-X., Long, Q.-X., Deng, H.-J., Huang, L.-Y.,Chen, J., Wang, K., et al., "The D614G mutation of SARS-CoV-2 spike protein enhances viral infectivity and decreases neutralization sensitivity to individual convalescent sera," *bio-Rxiv,* 2020.

[23] Maria Romano, Alessia Ruggiero, Flavia Squeglia, Giovanni Maga and Rita Berisio, "A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping," *Cells,* vol. 9, no. 1267, pp. 1-22, 2020.

[24] Gao, Y.; Yan, L.; Huang, Y.; Liu, F.; Zhao, Y.; Cao, L.; Wang, T.; Sun, Q.; Ming, Z.; Zhang, L.; et al., "Structure of the RNA-dependent RNA polymerase from COVID-19 virus.," *Science,* 2020.

[25] Subissi, L.; Posthuma, C.C.; Collet, A.; Zevenhoven-Dobbe, J.C.; Gorbalenya, A.E.; Decroly, E.; Snijder, E.J.; Canard, B.; Imbert, I., "One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities.," *Proc. Natl. Acad. Sci. USA ,* vol. 111, pp. E3900-E3909, 2014.

[26] M. K. Lo et al., "GS-5734 and its parent nucleoside analog inhibit Filo-, Pneumo-, and Paramyxoviruses.," *Sci. Rep,* vol. 7, no. 43395, 2017.

[27] Ka-Tim Choy, Alvina Yin-Lam Wong, Prathanporn Kaewpreedee, Sin Fun Sia, Dongdong Chen, Kenrie Pui Yan Hui, Daniel Ka Wing Chu, Michael Chi Wai Chan, Peter Pak-Hang Cheung, Xuhui Huang, Malik Peiris, Hui-Ling Yen, "Remdesivir, lopinavir, emetine, and homoharringtonine inhibit SARS-CoV-2 replication in vitro," *Antiviral Research,* vol. 178, no. 104786, 2020.

[28] T. P. Sheahan et al, "Broad-spectrum antiviral GS-5734 inhibits both epidemic and zoonotic coronaviruses.," *Transl. Med.,* vol. 9, 2017.

[29] Nikita Alexeev and Max A. Alekseyev, "Estimation of the true evolutionary distance under the fragile breakage model," *BMC Genomics,* vol. 18, no. 4, p. 356, 2017.

[30] J S M Peiris 1 , S T Lai, L L M Poon, Y Guan, L Y C Yam, W Lim, J Nicholls, W K S Yee, W W Yan, M T Cheung, V C C Cheng, K H Chan, D N C Tsang, R W H Yung, T K Ng, K Y Yuen, "Coronavirus as a Possible Cause of Severe Acute Respiratory Syndrome," *Lancet,* pp. 1319-25, 2003.

[31] Fenggang Yu, Rufu Ji, Yongyong Tang, Jin Liu and Benjie Wei, "SARS-CoV-2 infection and stem cells: Interaction and intervention," *Stem Cell Research,* vol. 46, no. 101859, 2020.

[32] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, Zhenshun Cheng, Ting Yu, Jiaan Xia, Yuan Wei, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *The Lancet,* vol. 395, no. 10223, pp. 497-506, 2020.

[33] Shuo Su, Gary Wong, Weifeng Shi, Jun Liu, Alexander C K Lai, Jiyong Zhou, Wenjun Liu, Yuhai Bi, George F Gao, "Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses," *Trends Microbiology,* vol. 24, no. 6, pp. 490-502, 2016.

[34] Moira Chan-Yeung, "Severe acute respiratory syndrome Patients were epidemiologically linked," *The British Medical Journal,* vol. 326, no. 7403, p. 1393, 2003.

[35] Ji Yeon Lee, You-Jin Kim, Eun Hee Chung, Dae-Won Kim, Ina Jeong, Yeonjae Kim, Mi-ran Yun, Sung Soon Kim, Gayeon Kim & Joon-Sung Joh , "The clinical and virological features of the first imported case causing MERS-CoV outbreak in South Korea, 2015," *BMC Infectious Diseases,* vol. 17, no. 498, 2017.

[36] Akhtar Hussain, Bishwajit Bhowmik, and Nayla Cristina do Vale Moreira, "COVID-19 and diabetes: Knowledge in progress," *Diabetes research and clinical practice ,* vol. 162, no. 108142, 2020.

[37] Sisi Kang, Mei Yang, Zhongsi Hong, Liping Zhang, Zhaoxia Huang, Xiaoxue Chen, Suhua He, Ziliang Zhou, Zhechong Zhou, Qiuyue Chen, Yan Yan, Changsheng Zhang, Hong Shan, Shoudeng Chen , "Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals

potential unique drug targeting sites," *Acta Pharm Sin B.,* 2020.

[38] Ahmed, S.F., Quadeer, A.A., McKay, M.R, "Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies.," *Viruses,* vol. 12, no. 254, 2020.

[39] Rahman, M.S., Hoque, M.N., Islam, M.R., Akter, S., Rubayet-Ul-Alam, A., Siddique, M.A., Saha, O., Rahaman, M.M., Sultana, M., Hossain, M.A.,, "Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2 etiologic agent of global pandemic COVID-19: an in silico approach.," *bioRxix,* 2020.

[40] Wu, C., Liu, Y., Yang, Y., Zhang, P., Zhong, W., Wang, Y., Wang, Q., Xu, Y., Li, M., Li, X, "Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods," *Acta Pharmaceutica Sinica B,* vol. 10, no. 5, pp. 766-788, 2020.

[41] Walls, AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D., "Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein.," *Cell,* vol. 181, pp. 281-292, 2020.

[42] Martina Bianchi,Domenico Benvenuto, Marta Giovanetti, Silvia Angeletti, Massimo Ciccozzi , and Stefano Pascarella, "Sars-CoV-2 Envelope and Membrane Proteins: Structural Differences Linked to Virus Characteristics?," *BioMed Research International ,* 2020.

[43] Taimá N. Furuyama, Fernando Antoneli, Isabel M. V. G. Carvalho, Marcelo R. S. Briones, Luiz M. R. Janini, "Temporal data series of COVID-19 epidemics in the USA, Asia and Europe suggests a selective sweep of SARS-CoV-2 Spike D614G variant," *arXiV,* 2020.

[44] Nei M. and Kumar S., Molecular Evolution and Phylogenetics., New York: Oxford University Press, 2000.

[45] Lanying Du, Yuxian He, Yusen Zhou, Shuwen Liu, Bo-Jian Zheng, Shibo Jiang, "The spike protein of SARS-CoV--a target for vaccine and therapeutic development," *Review Nat Rev Microbiol .,* vol. 7, no. 3, pp. 226-236, 2009.

[46] Lanying Du, Yang Yang, Yusen Zhou, Lu Lu, Fang Li, Shibo Jiang , "MERS-CoV spike protein: a key target for antivirals," *Expert Opin Ther Targets .,* vol. 21, no. 2, pp. 131-143, 2017.

[47] Schoeman, D. & Fielding, B. C. , "Coronavirus envelope protein: current knowledge.," *Virol. J.,*

vol. 16, no. 69, 2019.

[48] Weilong Shang, Yi Yang, Yifan Rao & Xiancai Rao , "The outbreak of SARS-CoV-2 pneumonia calls for viral vaccines," *npj Vaccines,* vol. 5, no. 18, 2020.

[49] Indwiani Astuti and Ysrafil, "Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response," *Diabetes Metab Syndr.,* vol. 14, no. 4, pp. 407-412, 2020.

[50] T.K. Tang, M.P.J. Wu, S.T. Chen, M.H. Hou, M.H. Hong, F.M. Pan, et al., "Biochemical and immunological studies of nucleocapsid proteins of severe acute respiratory syndrome and 229E human coronaviruses," *Proteomics,* vol. 5, pp. 925-937, 2005.

[51] L. Du, G. Zhao, Y. Lin, C. Chan, Y. He, S. Jiang, et al., "Priming with rAAV encoding RBD of SARS-CoV S protein and boosting with RBD-specific peptides for T cell epitopes elevated humoral and cellular immune responses against SARS-CoV infection," *Vaccine,* vol. 26, pp. 1644-1651, 2008.

[52] M. Surjit, B. Liu, V.T. Chow, S.K. Lal, "The nucleocapsid protein of severe acute respiratory syndrome-coronavirus inhibits the activity of cyclin-cyclin-dependent kinase complex and blocks S phase progression in mammalian cells," *J Biol Chem,* vol. 281, pp. 10669-10681, 2006.

[53] P.K. Hsieh, S.C. Chang, C.C. Huang, T.T. Lee, C.W. Hsiao, Y.H. Kou, et al., "Assembly of severe acute respiratory syndrome coronavirus RNA packaging signal into virus-like particles is nucleocapsid dependent," *J Virol,* vol. 79, pp. 13848-13855, 2005.

[54] P.S. Masters, L.S. Sturman, "Background paper: functions of the coronavirus nucleocapsid protein," *Adv Exp Med Biol,* vol. 276, pp. 235-238, 1990.

[55] R. McBride, M. van Zyl, B.C. Fielding, "The coronavirus nucleocapsid is a multifunctional protein," *Viruses,* vol. 6, pp. 2991-3018, 2014.

[56] S.F. Ahmed, A.A. Quadeer, M.R. McKay, "Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies," *Viruses,* vol. 12, p. E254, 2020.

[57] S.J. Liu, C.H. Leng, S.P. Lien, H.Y. Chi, C.Y. Huang, C.L. Lin, et al., "Immunological characterizations of the nucleocapsid protein based SARS vaccine candidates," *Vaccine,* vol. 24,

pp. 3100-3108, 2006.

[58] B. Shang, X.Y. Wang, J.W. Yuan, A. Vabret, X.D. Wu, R.F. Yang, et al., "Characterization and application of monoclonal antibodies against N protein of SARS-coronavirus," *Biochem Biophys Res Commun,* vol. 336, pp. 110-117, 2005.

[59] Y. Lin, X. Shen, R.F. Yang, Y.X. Li, Y.Y. Ji, Y.Y. He, et al., "Identification of an epitope of SARS-coronavirus nucleocapsid protein," *Cell Res,* vol. 13, pp. 141-145, 2003.

[60] T. K. Warren et al.,, "Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys," *Nature,* vol. 531, pp. 381-385, 2016.

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