Taiwo *et al Pan African Journal of Life Sciences* (2020): 4(3): 154-165
DOI: 10.36108/pajols/0202/40(0370)

Publication of Faculty of Basic Medical Sciences,
Ladoke Akintola University of Technology, Ogbomoso

www.pajols.org
Online ISSN:2672-5924

RESEARCH **Open Access**

Genomic Diversity and Phylogenetic Analysis of SARS-CoV-2 Circulating in Africa and Other Continents: Implications for Diagnosis, Transmission, and Prevention

Idowu A. Taiwo[1,2], Bamidele A. Iwalokun[1,3] ,Titilola A. Samuel[1,4], Adesola Olalekan[1,5], Khalid O. Adekoya[1,2], Oluwabukola M. Akinloye[6], Gloria Amegatcher[7], Eyitayo Adenipekun[1,5], Daniel Adewuni[1,5], Fatimah O. Anwoju[1,2], Olayiwola A. Popoola[1,5], Oluwatoyin P. Popoola[1,8], Bolanle Iranloye[1,9], Olubunmi A. Magbagbeola[1,4] and Oluyemi Akinloye[1,3, 5*]

[1]Centre for Genomics of Non-communicable Diseases and Personalized Healthcare (CGNPH). [2]Department of Cell Biology and Genetics, University of Lagos. [3]Department of Molecular Biology and Biotechnology, Nigerian Institute of Medical Research (NIMR), Yaba, Lagos. [4]Department of Biochemistry and University of Lagos. [5]Department of Medical Laboratory Sciences, University of Lagos. [6]Department of Medical Laboratory Sciences, Oulton College, Moncton, New Brunswick, Canada. [7]Department of Medical Laboratory Sciences and West African Centre for Cell Biology of Infectious Pathogen (WACCBIP), University of Ghana. [8]Department of Biomedical Engineering, University of Lagos. [9]Department of Physiology, University of Lagos

*Correspondence should be addressed to Oluyemi Akinloye : oakinloye@unilag.edu.ng, Director, CGNPH

Received 18th November 2020; Revised 22nd December 2020; Accepted 23rd December 2020

Abstract

Background: COVID-19 pandemic caused by SARS-CoV-2 remains a global health threat. Assessment of the genetic relatedness of the genome sequence is a prerequisite to understanding the dynamics, which is important to improve diagnosis and preventive measures. This study determined genomic diversity and SNP characteristic of genomes of SARS-CoV-2 from Africa and the rest of the world. The study involved molecular and phylogenetic analyses to understand the phylogeny and transmission dynamics of the virus.

Methods: The SARS-CoV-2 genome sequence data were mined and retrieved from major databases for one year in two phases: Phase 1; December 2019 to May 2020 and Phase 2; June 2020 to December 2020. A maximum of the four sequences that fulfilled the following predetermined criteria from each country were randomly selected for inclusion in the study: (i) sequence length >29,700 nt, (ii) number of Ns in the sequence not >5%, (iii) inclusion of Poly-A tail in the sequence record to ensure completeness.

Results: The similarity of SARS-Cov-2 genomes within and between countries was generally high with an average of 99.9%. Thus, SARS-CoV-2 vary between countries and continents by 0.1% as a result of SNPs in its genome. Phylogenetic data revealed multiple origin of SARS-CoV-2 in Africa and also suggested that the virus spreads by 'founder's effect'; whereby few viruses newly introduced into a population multiply rapidly and accumulate mutations as they spread quickly by community transfer to create population-based identity. Tree of continental consensus sequences retrieved in Phase 1 suggested that SARS-CoV-2 virus is of two major clusters: African cluster consisting of Africa, Europe, and North America and Asian cluster made up of Asia, South America, and Oceania. However, this clustering pattern vanished in phase 2. Thus, upholding the view that SARS-CoV-2 is constantly evolving.

Conclusion: This dynamism and genetic diversity of SARS-CoV-2 have important implications in diagnosis, transmission, and prevention strategy.

Keywords: SARS-CoV-2, SNPs, Genomics, Phylogenetic, Data Mining

1.0 INTRODUCTION

COVID-19 pandemic caused by a novel severe acute respiratory syndrome coronavirus -2 continues to wreak havoc in human populations living in many countries of the world. As of 6th of December 2020, COVID-19 pandemic has caused 68,161,156 cases and 1,555,898 deaths globally (https://coronavirus.jhu.edu/map.html). Out of these, 2,274,651 cases and 53,921 deaths were from Africa (africacdc.org/covid-19). After its detection in late December 2019, it spread rapidly across different parts of the world causing lockdown and paralysis of social and economic activities [1]. SARS-CoV-2 isolates have been detected in many parts of the body including tracheal aspirates, bronchial alveolar lavage, sputum, nasopharyngeal swab, oropharyngeal swab, saliva and blood samples of infected patients [2, 3].

In the absence of specific anti-SARS-CoV-2 drugs and vaccines, experimental and supportive therapies under clinical trials are currently being used for case management. Presently, different combinations of preventive measures such as lockdown, social distancing, face mask wearing, handwashing and social gathering restrictions are being employed as strategies to halt SARS-CoV-2 transmission and put a stop to the ongoing pandemic in several countries across the globe. Unfortunately, non-pharmaceutical measures have not been able to stop SARS-CoV-2 transmission and have necessitated the need to develop novel specific drug and vaccines for prophylaxis and treatment of COVID-19 as a matter of urgency [4,5,6]. Improvement in COVID-19 diagnosis in terms of higher sensitivity and specificity of diagnostic kits capable of ruling out other diseases that are clinically like COVID-19 is also needed urgently.

SARS-CoV-2 is a positive single stranded enveloped RNA virus belonging to Betacoronavirus genus of the Orthocoronaviridae sub-family. The virus has an average genome size of 29.9 kb, which comprises 11 coding regions for structural, non-structural and accessory proteins [7]. Like other related coronaviruses, the order of arrangement of the encoded proteins from 5' to 3' ends of the genome comprises the ORF1ab polyproteins (which account for more than two-thirds of the virus genome), followed by structural proteins named as spike (S) glycoprotein, envelope (E) protein, membrane (M) protein and nucleocapsid (N) protein as well as accessory proteins defined as ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF10 [8, 9]. Further structural analysis of SARS-CoV-2 genome has revealed the ORF1ab polyprotein to be divided into 15 non-structural proteins (NSP1 – NSP16) following cleavage by three viral proteases [10,11].

The first genome sequence of the novel SARs-CoV-2, using the next generation sequencing (NGS) technology was reported in January, 2020 [12]. Since then, tremendous NGS sequencing efforts have been put up by researchers and public health institutions from several countries of the world to generate over 10,000 high quality genome assemblies of SARS-CoV-2 strains recovered from patients from different continents of the world, including Africa [13,14].These sequences have been deposited in public databases such as the Global Initiative on Sharing of all Influenza Data (GISAID) at https://www.gisaid.org. Others include GenBank (https://www.ncbi.nlm.nih.gov/genbank), European nucleotide archive (ENA) at https://www.ebi/ac/uk/ena, DNA Databank of Japan (DDBJ) at http://www.ddbj.nig.ac.jp and Virus Pathogen and Analysis Resource (ViPR) at http://www.ViPRbrc.org. Visualization and analysis tools such as Nextstrain and COV-GLUE have also been provided to enable real-time analysis and visualization of genomic signatures in the genome of SARS-CoV-2 as they emerge and spread from country to country and region to region via travel associated or community human-human transmission [15,16].

Generally, RNA viruses are highly susceptible to mutations due to the low fidelity of RNA dependent RNA polymerase [17]. However, Coronaviruses are exceptions because their genomes encode a proofreading exonuclease (ExoN) in nonstructural protein 14 (nsp 14) of ORF 1b [18]. Considering the SARS-CoV-2 genome size of 29.9 kb, its estimated evolutionary rate is 0.9 x 10 -3 nucleotide substitutions/site/year [19]. This is relatively low when compared with other RNA viruses such as influenza virus. The fact that coronaviruses do not evolve only by mutation, but also by recombination [19] coupled with the recent outbreak of COVID-19 pandemic, there has been a resurgence of interest in the evolution of coronaviruses especially SARS-CoV-2 Earlier SARS-CoV-2 genome diversity studies conducted in China, USA and India have confirmed the plasticity of SARS-CoV-2 genome with genetic mutations and single nucleotide polymorphisms (SNPs) playing a major role in the microevolution of SARS-CoV-2 strains into lineages and sub-strains [12,20,21].

The three forms of mutations that have been reported among circulating SARS-CoV-2 strains were

substitutions, (synonymous or non-synonymous), insertion and deletion collectively referred to as indels [21]. SNPs refer to substitutions of equal or greater than 1% occurrence in the genomes of SARS-CoV-2 strains. Since genetic mutations in SARS-CoV-2 genome accumulate with time and vary between settings, it has become very critical to monitor genomic variations in SARS-CoV-2 in real-time on country and regional basis. This will aid a better understanding of local, regional and global epidemiology of SARS-CoV-2 via improved knowledge about drivers of genetic mutations, evolution and lineage emergence in SARS-CoV-2 strains circulating in Africa and the rest of the world.

Although few studies have attempted to unravel the phylogenetic relationship, evolutionary path, and transmission dynamics of SARS-CoV-2, none of these studies included African particularly Nigerian strains in its analysis [22]. There is paucity of information regarding the level of mutations and SNPs as well as the propensity of these drivers to cause evolution among African strains of SARS-CoV-2 with their origins and reservoirs known. However, with increasing number of genome sequences of SARS-CoV-2 strains circulating in Africa including Nigeria, it is now possible to carry out a phylogenetic study of SARS-CoV-2 that would involve African samples.

The first African genome sequences of the new coronavirus was obtained from an index case who travelled from Europe to Nigeria. SARS-CoV-2 genomes from Nigeria and other African countries have been sequenced and deposited in databases most especially in GISAID. Information on the genomic variation of SARS-CoV-2 strains from Africa is crucial for the identification of markers that can be targeted for personalized drug and vaccine design as well as the development of regionally effective molecular diagnostics. Findings from this study will improve understanding of transmission dynamics and microevolution of SARS-CoV-2 in the African region.

This study was therefore carried out to determine the genomic diversity and SNP characteristics of genomes of SARS-CoV-2 and assess the phylogenetic characteristics of the circulating strains from Africa and the rest of the world.

2.0 METHODOLOGY

2.1 Comparative SARS-CoV-2 Genome Sequence Data in Five Major Databases

The dataset for this study were retrieved from five major databases namely National Center for Biotechnology Information (NCBI: www.ncbi.nlm.nih.gov/genbank), European Nucleotide Archive (ENA: www.https://www.ebi.ac.uk/ena. DNA Databank of Japan (DDBJ: www.ddbj.nig.ac.jp) Global Initiative on Sharing All Influenza Data (GISAID: www.gisaid.org), and Virus Pathogen Database and Analysis Resource (ViPR: www.gisaid.org). They were mined for SARS-CoV-2 genome data records that had accumulated for one year since the outbreak of COVID-19 from December, 2019 to December, 2020. The retrieved sequences were given Accession Number|Country|Collection Date annotation before analyses.

In order to compare evolutionary changes and genetic divergence in SARS-CoV-2 during early and late period of its emergence, sampling was divided into two phases: Phase 1 (December 05, 2019 to May 31, 2020) and Phase 2 (June 01 to December 06, 2020). The data were accessed after issuing appropriate commands, which captured nucleotide completeness (complete or partial), continent/country of origin of data, and sequence quality in terms of proportion of percentage of unidentified bases designated as 'Ns'. Derivation of continental and country percentages respectively was computed in MS Excel version 10 by writing appropriate formulae.

2.2 Data Retrieval for Molecular and Phylogenetic Analysis

In addition to their popularity and ease of navigation, NCBI and GISAID databases contained the largest volume of SARS-CoV-2 genome dataset. Thus, out of the 5 databases compared for their viral genome information, only NCBI and GISAID were used for further analysis that included molecular characterization and phylogenetic reconstruction throughout the entire study (December 2019 – December 2020). For inclusion in the study, the retrieved sequence data had to satisfy predetermined selection criteria that included nucleotide completeness (sequence length > 29,700 nt), adequate genome coverage, and sequence quality in terms of the proportion of unidentified nucleotides designated as 'N', inclusion of poly-A tail in the whole genome sequence. To remove technical variations and standardize

comparison, the retrieved sequences were pruned at the 5' end such that their Site 1 nt mapped to Site 41 nt of NCBI reference sequence (Accession No: NC 045512.2). Since the sequences included poly-A tails, pruning at the 3' end was merely to remove the tail.

2.3 Multiple Sequence Alignment and Variation Analysis

Multiple sequence alignment (MSA) was by MAFFT Version 7 while phylogenetic reconstruction by p-distance (in the units of number of base differences per site) was implemented in MEGA X. The genomes were initially aligned with MAUVE to check for large scale genomic changes including large deletions, gene inversion, and genome rearrangements. Then, the sequences were re-aligned in MAFFT for outputs that were used in DnaSP (Version 6.12.03) for SNP detection. Visualization and determination of allelic frequency were done by Jalview 2.10.5. Genomic location of SNPs was determined relative to the NCBI reference sequence form the GenBank (NC 045512.2). Sites where the frequency of the second most frequent allele is greater than 1% were regarded as polymorphic; otherwise, they were referred to as monomorphic. Monomorphic sites with minor allele frequency less than 1% were rare alleles and the viruses containing them were regarded as rare variants. Consensus sequences were generated for each continent in Jalview. These sequences were realigned in MAFFT to generate aligned sequences in FASTA format. Using MEGA X, a continental consensus tree of the aligned sequences was produced to reveal continental affiliations and the overall pattern of global spread of the virus.

2.4 Linkage Disequilibrium and Haplotype Analysis for L and S Novel Coronavirus Strains

Linkage disequilibrium and haplotype analysis for L and S strains of SARS-CoV were based on allelic frequencies at C8782 and T28144 sites as previously described by Tang *et al.,* [23]. Detection of haplotypes and determination of frequencies of CT (L strain), TC (S strain) haplotypes and their recombinants were analyzed by DnaSP and Jalview, respectively.

2.5 Phylogenetic Tree Reconstruction

Construction of the maximum likelihood tree was implemented in MEGA X using MAFFT-aligned FASTA sequence output based on Tamura-Nei evolutionary model. After obtaining the initial trees, the tree that had topology with superior log likelihood value was selected by applying neighbor-join (NJ) and bio-NJ algorithm in a heuristic search. Two global trees were constructed in this study: one consists of all the genome sequences considered for the study while the second is a tree constructed from continental consensus sequences. Validation of the continental global tree was by 500 bootstrap replicates.

3.0 RESULTS

3.1 Sequence Alignment and Large-Scale Genomic Changes in SARS-CoV-2

Multiple submission into the same and different databases were observed. However, similarity of SARS-CoV-2 genome within and between countries was generally very high (99.9 – 100%). Multiple Sequence alignment (MSA) revealed that the aligned SARS-CoV-2 genomes are made up of a major collinear block of conserved sequences that cover up an average size of 99.9% of the viral genome. The blocks are interspersed with SNPs that accounted for remaining 0.1% dissimilarity in SARS-C0V-2 genome. Large mutations such as gene loss, duplication, inversion, and rearrangements were not detected using MAUVE. Majority of detected SNPs were transitions (C/T & G/A) with bias towards C/T transitions.

3.2 Distribution of SNPs and Variation Analysis

The genomic distribution of SARS-CoV-2 SNPs is shown in Table 1 with ORF1a representing the region with the greatest number of SNPs in all the continents. However, when the number of SNPs were normalized by frequency of SNPs per 1000 nts in a genomic region, the N (nucleocapsid) gene contained the highest frequency of SNPs. The greatest number of SNPs (a total of 113) were detected in Asian SARS-CoV-2, which also have the greatest frequency of SNPs (102.5 SNPs/1000 nts). In terms of number and frequency, Asian SARS-CoV-2 had the greatest diversity as compared to other continents (Table 2). Further analyses revealed the occurrence of 1 – 15 SNPs, per loci with an average frequency of 2.5 SNPs per 1,000 nucleotides in the genomes of SARS-COV-2 from Africa. For strains circulating from the rest of the world 3 -27 SNPs per locus (Asia), 1 -6 SNPs per locus (North America, 2 – 27 SNPs per locus (Europe) and 1 -4 SNPs per locus (Oceania) were found. No SNPs were found in the M gene and ORF 7a, 7b and 10 among the African strains, while the SNP frequency in other genes

Table 1. Genomic Distribution of SNPs in the Studied SARS-CoV-2 Genomes from Different Continents

Region	Size	Africa		Asia		N/America		S/America		Europe		Oceania	
		SNPs	Frequency (SNPs / 1000 nts)	SNPs	Frequency (SNPs / 1000 nts)	SNPs	Frequency (SNPs / 1000 nts)	SNPs	Frequency (SNPs / 1000 nts)	SNPs	Frequency (SNPs / 1000 nts)	SNPs	Frequency (SNPs / 1000 nts)
5'UTR	265	1	3.8	3	11.3	1	3.8	1	3.8	2	7.5	1	3.8
1a	13218	15	1.1	32	2.4	14	1.1	6	0.45	27	2.0	4	0.3
1b	8088	8	1.0	27	3.3	11	1.4	4	0.49	14	1.7	4	0.5
S	3822	6	1.6	21	5.5	15	3.9	4	1.1	6	1.6	2	0.5
3a	828	3	3.6	5	6.0	2	2.4	3	3.6	3	3.6	3	3.6
M	227	-	-	7	30.8	-	-	-	-	2	8.8	1	4.4
E	186	1	5.4	-	-	-	-	-	-	-	-	-	-
6	366	1	2.7	-	-	-	-	-	-	1	2.7	-	-
7a	132	-	-	-	-	-	-	-	-	-	-	-	-
7b	366	-	-	-	-	-	-	-	-	-	-	-	-
8	1260	1	0.8	2	1.6	2	1.6	1	0.8	-	-	-	-
N	117	7	59.8	12	102.5	5	42.7	3	25.6	8	68.4	1	8.5
10	232	-	-	3	12.9	-	-	-	-	1	4.3	-	-
3'UTR	764	3	3.9	1	1	1	1	-	-	-	1.3	-	-
Total	29871	46	2.5	113	113	41	41	22	22	1.2	3.4	16	21.6

Table 2. Haplotype frequencies of S and L SARS-CoV-2 Strains in Different Continents

Haplotype	Strain	Africa		Asia		N/America		S/America		Europe		Oceania		Total	
		No	%	No	%	No	%	No	%	No	%	No	%	No	%
CT	L	14	82.4	26	78.8	17	84.5	5	71.4	23	100.0	4	100.0	89	86.4
TC	S	3	17.6	7	21.2	2	10.5	2	28.6	0	0.0	0	0.0	14	13.6
Total		17	100	33	100	19	100	7	100	23	100	4	100	103	100

Table 3. Common SNPs across all Continents

	Location	Genomic Region	Substitutions
1	241	5'UTR	C/T
2	3071	Orf1a	CT
3	11083	Orf1a	G/T
4	25563	ORF3a	G/T
5	28881	N	G/A
6	28882	N	G/A
7	28883	N	G/C

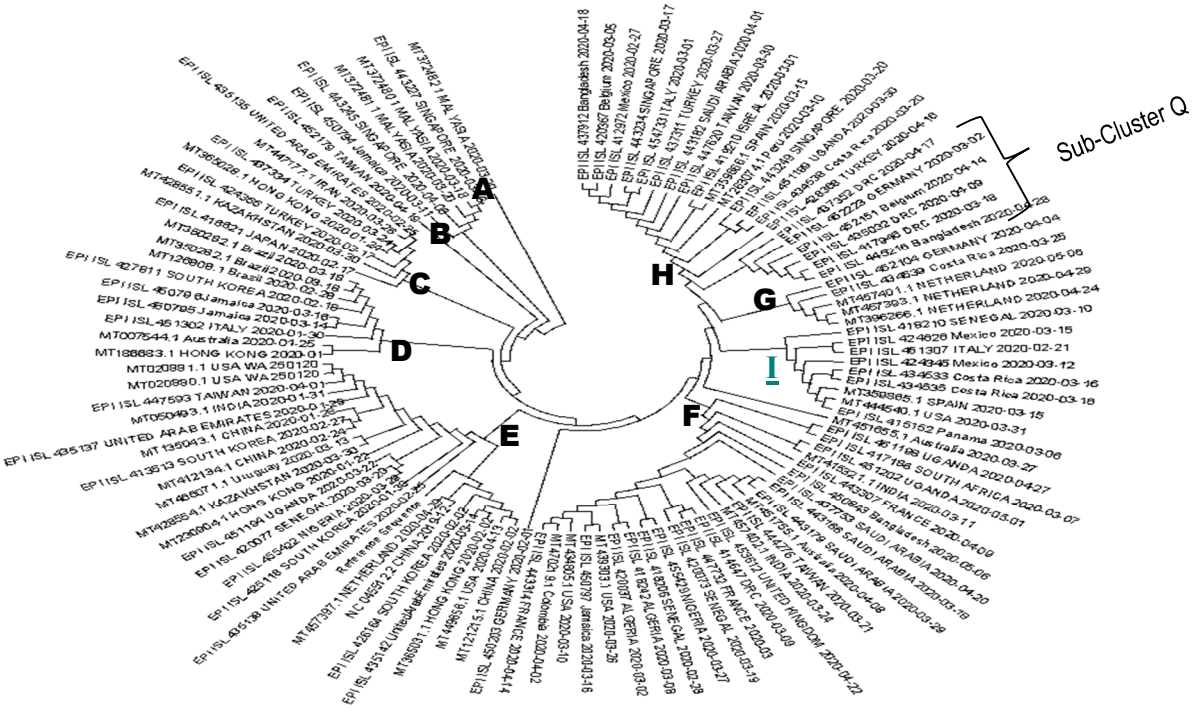

Figure 1. A Maximum Likelihood Phylogenetic Tree of All the Countries with Nine Clusters (A-I) from December 2019 to May 2020 (Phase 1).

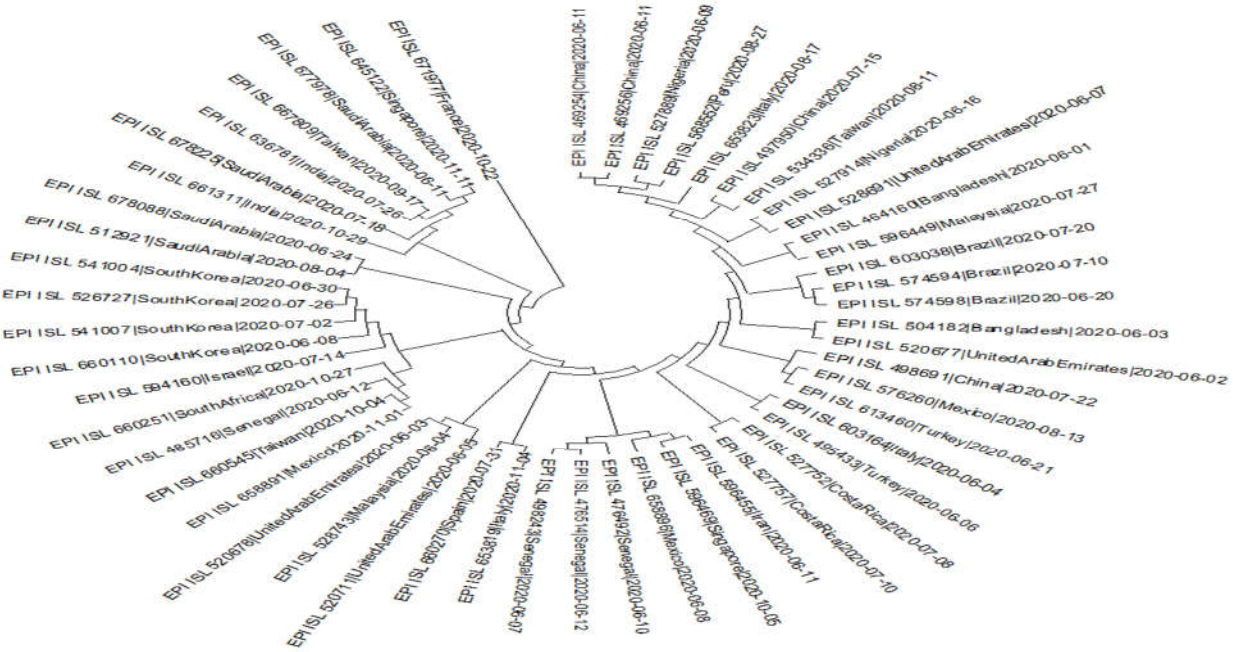

Figure 2. A Maximum Likelihood Phylogenetic Tree of All the Countries from June 2020 to December 2020 (Phase 2).

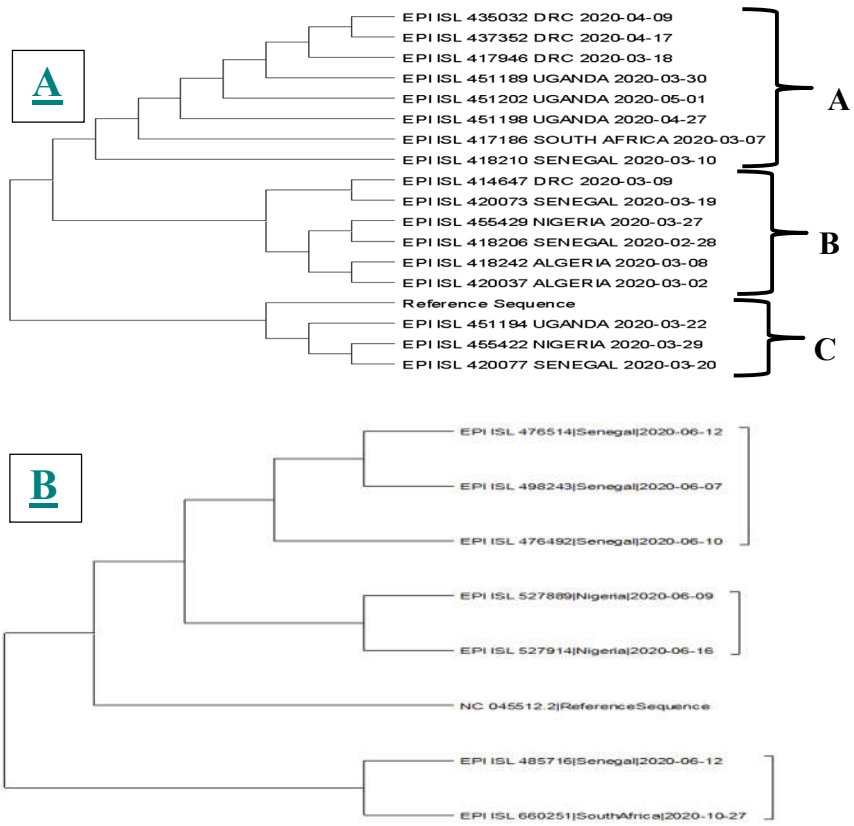

Figure 3. A Maximum Likelihood Phylogenetic Tree of Selected Africa Countries in (A) Phase 1 (December 2019 to May 2020) and (B) Phase 2 (June 2020 to December 2020).

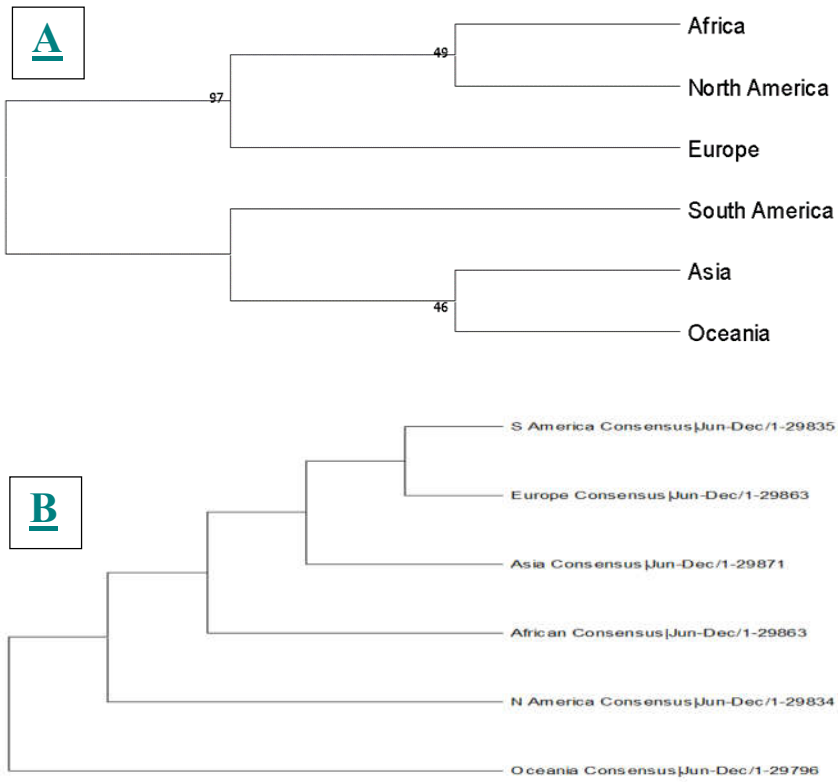

Figure 4. Continental Phylogenetic Tree for (A) Phase 1 (December 2019 to May 2020) and (B) Phase 2 (June 2020 to December 2020).

in decreasing order was orf1a (15)>orf1b (8)>N (7)>S (6) (Table 2). Some of the SNPs detected in this study were unique for each continent, however 7 SNPs located at 241, 3071, 11083, 25563, and 28881-28883 were common to all continents (Table 3).

3.3 Haplotype Analysis for S and L SARS-CoV-2 Strains

Two haplotypes namely CT and TC, which define L and S strains of SARS-CoV-2 [23] had different frequencies (Table 2). The total number of CT haplotypes which identifies the L strain was 89 (86.4%) while that of TC haplotypes (S strain) was 14 (13.6%) indicating that the L strain is the more common SARS-CoV-2 strain in the world (Table 2). It is noteworthy that loci C8782 and T28144 were in absolute linkage disequilibrium because no single recombinants for the two loci was recovered. Viruses of S lineage were not found in European and Asian samples used for this study.

3.4 Phylogenetic Analysis

The tree that had the highest log likelihood topology (-43546.61) is shown in Figure 1. Eight main clusters (A-H) and many sub-clusters were recognized on the tree. An example is Sub-Cluster Q which forms a sub-cluster under the main cluster I. Some clades are made up mostly from countries in the same continents while some clusters were not well defined because of mis-clustering involving countries located in different continents. For instance, Cluster B could be considered as an Asian cluster consisting of mostly Asian countries while Cluster G is a European cluster consisting of Netherlands and Germany with Costa Rica mis-clustering with these European countries. Comparatively, Cluster E is not well defined. Two countries that formed an out-cluster were Australia and Germany. Figure 1 also revealed a tree with different topology when sequences retrieved between December 2019 – May 2020 and those retrieved between June 2020 to December 2020 (Figure 2) were used for phylogenetic analysis. Figure 2 has many more but smaller clades than Figure 2. However, there was no significant difference in Africa's SARS-CoV-2 phylogeny when Figures 3a and 3b was compared.

The continental consensus tree showing bootstrap support for some of its nodes is depicted in Figure 4a. According to the tree, all continents in the world could be grouped into two major clusters regarding SARS-CoV-2 transmission: the first cluster, which could be called African Cluster consists of Africa, North America, and Europe while the second cluster, which could be referred to as Asian Cluster contains Asia, South America and Oceania. This clustering pattern vanished later between June 2020 to December 2020 and was replaced by a tree with a single large cluster (Figure 4b).

4.0 DISCUSSION

The voluminous genome sequence data that accumulated in databases within a short time of discovery of the novel coronavirus was fundamental to understanding the molecular characteristics and phylogeny of the novel virus. Using data mining, we have carried out a genomics and phylogenetic study with a view to understanding the mutational, evolutionary, and transmission dynamics of the novel virus. At the commencement of the present study, we examined eighteen databases for genome information on the novel coronavirus and discovered that many of them are specialized databases that accept data from special or mandate organisms or items with little or no data on coronaviruses especially SARS-CoV-2. By empirical estimation, NCBI GenBank is regarded as the most popular database; however, it appeared that GISAID was the most preferred for submitting SARS-CoV-2 genome sequence data. Furthermore, our experience during sequence retrieval shows that these two databases were the most interactive and the easiest to navigate. Thus, after database SARS-CoV-2 sequence record comparison in the Phase 1 period of this study (December 2019 – May 2020), we focused on sequences in NCBI and GISAID for sequence retrieval for molecular characterization and phylogenetic analysis throughout the period of the study (December 2019 – December 2020).

A scrutiny of data in the 5 databases revealed multiple submissions such that a sequence data might be found in more than one database. Precaution was taken to avoid data duplication and also to avoid inadvertent error that might affect accuracy and interpretation of results, especially since the study involves phylogenetic reconstruction. However, Comparative analysis of whole genome sequences of the novel coronavirus revealed high similarity of 99.9-100% within and between different countries and continents in agreement with previous reports [14]. In attestation to this, we did not detect large genome changes such as gene duplication, loss and rearrangement among the studied SARS-CoV-2 sequences, which should have created considerable differences between the novel coronavirus genomes. Previous report has also ob-

served that SARS-CoV-2 is an evolving virus with no recent significant gene rearrangement including recombination [24]. Based on phylogenetic studies and molecular clock theory, there is general agreement that the novel coronavirus originated in Wuhan, China, around mid-November, 2019 i.e. seven to twelve months ago [24] as at the time of this report (December, 2020). The detection of many SNPs in this study suggests that point mutations, majority of which are C/T transitions and a few indels are the major processes currently driving SARS-CoV-2 evolution. Thus, despite the high genome similarity among the novel coronaviruses, scanning the genome with appropriate software revealed several SNPs with N gene serving as a SNP forest in view of the high frequency of SNPs on the genome of African SARS-CoV-2.

Based on our observation in the present study, we hereby propose 'founder's effect' strategy to account for the method usually adopted by SARS-CoV-2 as it spreads globally. The results of this study strongly suggests that after few multiple introductions of the viruses into a population it usually multiplies rapidly and accumulate mutations as they spread quickly by community transfer to create a population-based identity. This opinion is supported by our observation that many of the detected SNPs, and their genomic distribution varied from one region to the other especially, due to migration of the virus via human-to-human transmission.

Several reports have indicated the basis of the unprecedented survival advantage of SARS-CoV-2 [23,24]. The most acclaimed properties are the possession of spike protein that is optimized for binding to angiotensin converting enzyme 2 (ACE2) of human cells [24]. Like many other RNA viruses, replication of the viral genetic material is rapid and expected to be error prone [23,24]. However, coronaviruses have relatively low mutation rate when compared to other RNA viruses, because they (coronaviruses) have genomes that encode a proofreading exonuclease (ExoN). Nonetheless, evolution of coronaviruses especially SARS-CoV-2 has been the subject of investigation, especially in view of the current COVID-19 pandemic.

In support of the proposed 'Founder's Effect' theory for SARS-CoV-2 transmission dynamics, the novel virus has such ability that when a few viruses enter a population, they rapidly multiply and acquire fresh mutations to establish and give themselves a unique identity in the population. Going by statistical sampling theory and principle of genetic drift [25], the small viral sample introduced into a new human population may not be representative of the original population from where they came from. As the viruses multiply and mutate, they quickly acquire a new identity which might be unique for that population. Natural selection can then act to select SARS-CoV-2 variants with features that confer the best survival advantage in a particular environment. It is therefore understandable that only 7 out of the hundred SNPs in this study were common to all the sampled SARS-CoV-2 genomes in all continents. The phylogenetic trees are to a large extent, reflection of the SNP profile obtained. Thus, in the global phylogenetic tree constructed in this study, there were clades that could be easily defined based on the continents where the countries are located. Some clades were not well defined because countries that constitute such clusters are from different continents. Using continental consensus sequences, SARS-CoV-2 in the world fell into two major clusters. Tang *et al.,* [23] had similarly divided the novel coronavirus into two major groups based on linkage disequilibrium and haplotype analysis of cytosine residue found at C8782 and thymine residue located at T28144. Majority of viral strains in this study fell into the L group by possessing CT haplotype. It is not yet clear, however, whether the two major clusters in this study are connected to the two major lineages identified by Tang *et al.,* [23]. Many SARS-CoV-2 strains are now well documented; however, L and S strains are of special interest because to the best of our knowledge, they are of fundamental importance being the earliest (if not the first) SARS-CoV-2 strains identified, characterized and reported in literature [23]. Correlation between S and L strains (and other strains) of SARS-CoV-2 and severity/death rate of COVID-19 is the subject of future study . In contrast to our study in which we obtained many clades, Giorgi and Mercatelli [26] identified 3 main clades because they considered only a few well-defined mutations of SARS-CoV-2 genome in their study.

The number of samples used in this study was limited to 4 per country to accommodate computational limitation and prevent frequent abortion of some complex analytical processes experienced during the study. This may be considered as likely limitation of the study because of few numbers of genome samples included per country. This constraint was compounded by the need to have adequate global sampling spread of SARS-CoV-2 genomes in the databases for a year (December 2019 – December 2020). Despite this limitation, this study attests to the high mutational and adaptive evolutionary capability to different populations in the world.

This evolutionary attributes could account for the possession of unique receptor binding domain on SARS-CoV-2 glycoprotein spike, a structure that has evolutionarily enhanced the virus for efficient attachment to human angiotensin converting enzyme 2 (ACE2) receptors to enhance viral entry into human cells [27]. In this study, we observed several SNPs on the S gene in agreement with Woo *et al.,* [27], who indicated that spike proteins of the novel virus have the most variable sequences in the coronavirus genomes. It will be of interest to see if the mutations identified in this study are adaptive or non-adaptive in subsequent studies. In contrast to Woo *et al.,* [27], data from this study suggests that the nucleocapsid protein of the novel coronavirus is the most variable in view of the high density of SNPs detected on the gene. Functional genomics and proteomics studies on SARS-CoV-2 are needed to see the impact of SNPs on the N gene on nucleocapsid function.

The advantages conferred on the novel coronavirus in terms of its infectivity and transmissibility implies that it might be challenging to start fighting COVID-19 once it has entered a community. This justifies the option of closing borders immediately the news of the disease breaks out in any part of the world. Furthermore, the finding that coronaviruses might have evolved different SNP signatures in different populations as it is transmitted from one country to the other raises some concern. To effectively control a disease, proper testing and diagnosis is crucial. Since many COVID-19 testing kits are nucleic acid-based, developing a universal testing kit with equally high efficiency in different parts of the world may be challenging or difficult. Thus, many of the kits available today should be validated locally to determine its sensitivity and specificity, because a test kit that works well in a particular population may have reduced efficiency in the other. Several reasons including mode, site and time of sample collection and transportation have been given as the basis of the recent increasing reported cases of false negatives [14,28]. The limited sensitivity of the currently available nucleic acid detection test system, and implicated viral genetic variation have been implicated [14,29,30].

Since COVID-19 testing kits are usually based on amplification of nucleic acid by PCR, which involves annealing of short oligonucleotide primers and probes to SARS-CoV-2 genome, it might be difficult or impossible to have a generalized testing kit with equally high efficiency in all populations considering the high genetic diversity of SARS-CoV-2 indicated in this study. Efforts by African countries like Nigeria to produce their testing kits, drugs and vacinesvaccines based on the strains of SARS-CoV-2 in their peculiar populations is desirable.

Artesi *et al.,* [31] have emphasized that accurate testing is an important strategy to prevent spread of the disease as this would facilitate identification and isolation of infected individuals. Puty *et al.,* [32] also pointed out that mutations or SNPs have great implications for SARS-CoV-2 testing, target drug binding and antibody binding. In producing antiviral drugs and vaccines for the treatment and prevention of COVID-19 therefore, cognizance should be taken of SARS-CoV-2 SNP profiles, which vary from region to region. Thus, in view of the region-specific genetic diversity of SARS-CoV-2, production of testing kits, drugs and vaccines should be region specific for proper diagnosis, treatment and prevention of COVID-19. Therefore, attempt to reduce or stop the spread/ transmission of SARS-CoV-2 should consider region specific intervention as suggested above. Furthermore, an effective intervention in a particular region should be reviewed from time to time in view of the continuing evolution of the virus.

The present study suggests that the high mutability of the novel coronavirus gives the novel coronavirus the opportunity to evolve different variants in different communities. Hence, the concern for variability of the novel Corona Virus is not only for the development of testing kit, but also for effective drug designing and vaccine development. The latter is in respect of the S gene on which many unique SNPs were found. The effect of these SNPs on human T cell and B- cell responses need further investigation as this is critical for assessing effects on neutralization antibody production and induced TH2 associated antibody dependent enhancement [33,34]. Findings from this study uphold the view of previous recommendations of Heng et al, and Vankadari, [35,36] that every population should identify key SARS-CoV-2 SNPs peculiar to their population for effective and appropriate intervention.

Conflict of Interest

The authors declare that there is no conflict of interest.

Funding

The study is part of the bioinformatics and molecular biology capacity development planning program supported by the NIH-Fogarty Grant (1D71TW011487-01).

Authors Contribution

IAT, BAI contributed to data analysis tools, performed data analysis and interpretation and manuscript writing; **TAS, AO** performed data collection, contributed to data analysis tools and interpretation of data; **KOA** contributed to data collection, data analysis tools and analysis of data**; OMA** contributed to data analysis tools and data interpretation; G**A, EA, DA, FOA, OAP, OPP** contributed to data collection and data analysis tools; **BI, OAM** contributed to data interpretation and manuscript writing; **OA** conceived and designed the study, contributed to data analysis tools, analysis of data, interpretation and writing of manuscript.

References

1. WHO. Coronavirus disease (COVID-19) situation report-132. 31 May 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200531-covid-19-sitrep-132.pdf?sfvrsn=d9c2eaef_2
2. Sullivan PS, Sailey C, Guest JL, Guarner J, Kelley C, Siegler AJ, et al., Detection of SARS-CoV-2 RNA and Antibodies in Diverse Samples: Protocol to Validate the Sufficiency of Provider-Observed, Home-Collected Blood, Saliva, and Oropharyngeal Samples. JMIR Public Health Surveill. 2020; 6(2): e19054. doi: 10.2196/19054.
3. Thabet L, Mhalla S, Naija H, Jaoua MA, Hannachi N, Fki-Berrajah L, Toumi A, Karray-Hakim H. SARS-CoV-2 infection virological diagnosis. Tunis Med. 2020; 98 (4):304-308.
4. Ye ZW, Jin DY. [Diagnosis, treatment, control and prevention of SARS-CoV-2 and coronavirus disease 2019: back to the future]. Sheng Wu Gong Cheng Xue Bao. 2020;36(4):571-592. Chinese. doi: 10.13345/j.cjb.200115.
5. Amanat F, Krammer F. SARS-CoV-2 Vaccines: Status Report. Immunity. 2020; 52(4):583-589. doi: 10.1016/j.immuni.2020.03.007.
6. McKee DL, Sternberg A, Stange U, Laufer S, Naujokat C. Candidate drugs against SARS-CoV-2 and COVID-19. Pharmacol Res. 2020; 157:104859. doi: 10.1016/j.phrs.2020.104859.
7. Naqvi AAT, Fatima K, Mohammad T, Fatima U, Singh IK, Singh A, et al. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. Biochim Biophys Acta Mol Basis Dis. 2020 1866(10):165878. doi: 10.1016/j.bbadis.2020.165878.
8. Srinivasan S, Cui H, Gao Z, Liu M, Lu S, Mkandawire W, Narykov O, Sun M, Korkin D. Structural Genomics of SARS-CoV-2 Indicates Evolutionary Conserved Functional Regions of Viral Proteins. Viruses. 2020; 12(4):360. doi: 10.3390/12040360.
9. Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. Gene Rep. 2020; 19:100682. doi: 10.1016/j.genrep.2020.100682
10. Romano M, Ruggiero A, Squeglia F, Maga G, and Berisio R. A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping Cells. 2020; 9(5): 1267. doi: 10.3390/cells9051267
11. Yoshimoto, F.K. The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19. Protein J 2020; 39, 198–216 https://doi.org/10.1007/s10930-020-09901-4.
12. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet. 2020; DOI:https://doi.org/10.1016/S0140-6736(20)30251-8
13. Pillay S, Giandhari J, Tegally H, Wilkinson E, Chimukangara B, Lessells R, Moosa et al. Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation during a Pandemic. Genes (Basel). 2020; 11(8):949. doi: 10.3390/genes11080949.
14. Wang C, Liu Z, Chen Z, Huang X, Xu M, He T, Zhang Z. The establishment of reference sequence for SARS-CoV-2 and variation analysis. J Med Virol. 2020;92(6):667-674. doi: 10.1002/jmv.25762.
15. Lai CC, Shih TP, Ko WC, Tang HJ, Hsueh PR. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. Int J Antimicrob Agents. 2020;55 (3):105924. doi: 10.1016/j.ijantimicag.2020.105924.
16. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, Ortiz AT, Balloux F. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol. 2020 83:104351. doi: 10.1016/j.meegid.2020.104351.
17. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med. 2020;18(1):179. doi: 10.1186/s12967-020-02344-6.
18. Smith EC, Blanc H, Surdel MC, Vignuzzi M, Denison MR. Coronaviruses lacking xoribonuclease activity are susceptible to lethal mutagenesis:evidencemutagenesis: evidence for proofreading and potential therapeutics. PLoS Pathog. 2013; 9(8):e1003565.
19. Nakagawa S and Miyazawa T. Genome evolution of SARS-CoV-2 and its virological characteristics. Inflammation and Regeneration 2020; 40:17. https://doi.org/10.1186/s41232-020-00126-7.

20. Deng X, Gu W, Federman S, du Plessis L, Pybus OG, Faria N, et al. Genomic Survey of SARS-CoV-2 Reveals Multiple Introductions into Northern California without a Predominant Lineage. medRxiv [Preprint]. 2020:2020.03.27.20044925. doi: 10.1101/2020.03.27.20044925. Update in: Science. 2020 Jul 31;369(6503):582-587.

21. Singh H, Singh J, Khubaib M, Jamal S, Sheikh JA, Kohli S et al. Mapping the genomic landscape & diversity of COVID-19 based on >3950 clinical isolates of SARS-CoV -2: Likely origin & transmission dynamics of isolates sequenced in India. Indian J Med Res. 2020 May;151(5):474 -478. doi: 10.4103/ijmr.IJMR_1253_20.

22. Giovanetti M, Benvenuto D, Angeletti S, Ciccozzi M. The first two cases of 2019nCoV in Italy: Where they come from? Journal of Medical Microbiology. 2020; 92(5):518-521.

23. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui, J. and Lu J. On the origin and continuing evolution of SARS-CoV-2. National Science Review. 2020; 7(6):1012-1023. https://doi.org/10/1093/nsr/nwaa/036

24. Andersen KG, Rambaut A. Lipkin, W.I. Holmes EC and Garry RF. The proximal origin of SARS-CoV-2. Nat Med 2020; 26, 450–452. https://doi.org/10.1038/s41591-020-0820-9

25. Templeton AR modes of speciation and inferences based on genetic distances. Evolution. 1980; 34(4): 719-729. https://doi.org/10.1111/j.1558-5646.1980.tb04011.

26. Mercatelli D, Giorgi FM. Geographic and genomic distribution of SARS-CoV-2 mutations. Front Microbiol. 2020; https://doi.org/103389/fmicb.2020.01800.

27. Woo PCY, Huang Y, Lau SKP, and Yuen K. Coronavirus Genomics and Bioinformatics Analysis Viruses. 2010 Aug; 2(8): 1804–1820.doi: 10.3390/v2081803

28. Olalekan A., Iwalokun BA., Akinloye OM., Popoola O., Samuel TA and Akinloye O Covid-19 rapid diagnostic test could contain transmission in low- and middle-income countries. African Journal of Laboratory Medicine 2020a; 9 (1), a1255. https://doi.org/10.4102/ajlm.v9i1.1255.

29. Iwaloku BA, Olalekan A, Adenipekun E, Ojo O, Senapon O, , Orija O, Adegbola R, Salako B, Akinloye O (2020). Improving the Understanding of Immunopathogenesis of Lymphopenia as a Correlate of SARS-COV-2 Infection Risk and Disease Progression in African Patients: UGLY SARS-COV-2 Study Protocol. JMIR Preprints 2020 (09/06/2020:21242) DOI: https://doi.org/10.2196/preprints.21242.

30. Olalekan AO, Iwalokun BA, Adekunle OC, Makun HA, Mirabeau T, Akinloye O. Understanding the use of real time reverse trascriptase polymerase chain reaction (Rrt-PCR) for COVID 19 Diagnosis. Pan African Journal of Life Sciences 2020b; 4 (2): 86 – 97 DOI: 10.36108/pajols/0202/40(0270).

31. Artesi M, Bontems S, Gobbels P, Frankh M, Mau P, Boreux R, Meex C, et al. A recurrent mutation at position 26340 of SARS-CoV-2 is associated with failure of the E gene qRT-PCR utilized in a commercial dual-target diagnostic assay. Doi:101128/JCM.01598-20.

32. Puty TC, Saraf JS, Almeida TC, Filho VC, de Carvalho LE, Fonseca FL, Adami F. Evaluation of the impact of single nucleotide polymorphism on treatment response, survival and toxicity with cytarabines and anthracyclines in patients with acute myeloid leukemia: a systematic review.protocol. Syst Rev 2019; 8(1): 109.

33. Zost SJ, Gilchuk P, Case JB, Binshtein E, Chen RE, Nkolola JP, Schäfer A et al. Potently neutralizing and protective human antibodies against SARS-CoV-2. Nature. 2020 Aug;584(7821):443-449. doi: 10.1038/s41586-020-2548-6.

34. Ulrich H, Pillat MM, Tárnok A. Dengue Fever, COVID-19 (SARS-CoV-2), and Antibody-Dependent Enhancement (ADE): A Perspective. Cytometry A. 2020 Jul;97 (7):662-667. doi: 10.1002/cyto.a.24047.

35. Heng Li, Shang-Ming Liu, Xiao-Hua Yu Shi-Lin Tang, and Chao-Ke Tang Coronavirus disease 2019 (COVID-19): current status and future perspectives. Int J Antimicrob Agents. 2020 May; 55(5): 105951. doi: 10.1016/j.ijantimicag.2020.105951.

36. Vankadari N. Overwhelming mutations or SNPs of SARS-CoV-2: A point of caution Gene. 2020 Aug 20; 752: 144792. doi: 10.1016/j.gene.2020.144792