ORIGINAL ARTICLE

Sequence analysis for SNP detection and phylogenetic reconstruction of SARS-cov-2 isolated from Nigerian COVID-19 cases

I. A. Taiwo[2], N. Adeleye[2], F. O. Anwoju[2], A. Adeyinka[2], I. C. Uzoma[1,2] and T. T. Bankole[2]

1) *Molecular-Haematology Unit, Department of Medical Laboratory Science, Faculty of Health Sciences and Technology, College of Medicine, University of Nigeria, Nsukka, Enugu Campus, Nigeria and* 2) *Genetics and Bioinformatics Units, Department of Cell Biology and Genetics, Faculty of Science, University of Lagos, Akoka, Nigeria*

Abstract

Background: Coronaviruses are a group of viruses that belong to the Family Coronaviridae, genus *Betacoronavirus*. In December 2019, a new coronavirus disease (COVID-19) characterized by severe respiratory symptoms was discovered. The causative pathogen was a novel coronavirus known as 2019-nCoV and later as SARS-CoV-2. Within two months of its discovery, COVID-19 became a pandemic causing widespread morbidity and mortality.

Methodology: Whole genome sequence data of SARS-CoV-2 isolated from Nigerian COVID-19 cases were retrieved by downloading from GISAID database. A total of 18 sequences that satisfied quality assurance (length $\geq$29,700 nts and number of unknown bases denoted as "N" $\leq$ 5%) were used for the study. In addition, genome sequence of SARS-CoV-2 obtained from Nigeria's COVID-19 index case (Accession ID: EPI_ISL_413550) and the reference genome (Accession NC_ 045512.2) were obtained from GISAID and the GenBank databases, respectively. Multiple sequence alignment (MSA) was done in MAFFT (Version 7.471) while SNP calling was implemented in DnaSP (Version 6.12.03), respectively and then visualized in Jalview (Version 2.11.1.0). Phylogenetic analysis was with MEGA X software.

Results: Nigerian SARS-CoV-2 had 99.9% genomic similarity with four large conserved genomic regions. A total of 66 SNPs were identified out of which 31 were informative. Nucleotide diversity assessment gave Pi = 0.00048 and average SNP frequency of 2.22 SNPs per 1000 nts. Non-coding genomic regions particularly 5′UTR and 3′UTR had a SNP density of 3.77 and 35.4, respectively. The region with the highest SNP density was ORF10 with a frequency of 8.55 SNPs/1000 nts). This value was significantly higher ($P < 0.01$) than that of the spike gene, the region of greatest interest in SARS-CoV-2 genomics. Majority (72.2%) of viruses in Nigeria are of L lineage with preponderance of D614G mutation which accounted for 11 (61.1%) out of the 18 viral sequences. Nigeria SARS-CoV-2 revealed 3 major clades namely Oyo, Ekiti and Osun on a maximum likelihood phylogenetic tree.

Conclusion and Recommendation: There was a preponderance of L lineage (to include the new lineage scheme) and D614G mutants. Nigerian SARS-CoV-2 genome revealed ORF1ab as the region containing the highest SNP density as compared to the spike gene. The implication of this distribution of SNPs for the empirical lower infectivity of SARS-CoV-2 in Nigeria is discussed. This also underscores the need for more aggressive testing and treatment of COVID-19 in Nigeria. Additionally, attempt to produce testing kits for COVID-19 in Nigeria should consider the conserved regions identified in this study. Strict adherence to COVID-19 preventive measure is recommended in view of Nigerian SARS-CoV-2 phylogenetic clustering pattern, which suggests intensive community transmission possibly rooted in communal culture characteristic of many ethnicities in Nigeria.

Keywords: COVID-19, Nigeria, phylogeny, SARS-CoV-2, SNPs
Article published online: 18 January 2022

Corresponding author: I.C. Uzoma
E-mails: itaiwo@unilag.edu.ng **(I.A. Taiwo),** nikeadeleye@yahoo.com**,** tai_dex@yahoo.com **(N. Adeleye),** anwojufatimah@gmail.com **(F.O. Anwoju),** adeyinkaadeyemi495@gmail.com **(A. Adeyinka),** ijeoma.uzoma@unn.edu.ng **(I.C. - Uzoma),** ttbankole@unilag.edu.ng **(T.T. Bankole)**

Introduction

Coronaviruses are a group of viruses that belong to the family Coronaviridae, genus *Betacoronavirus* [1,2]. These viruses are of special interest because they possess the largest genome among RNA viruses and also have the capability to infect a wide host range causing intestinal and respiratory infections in animals and humans [3,4]. In recent times, coronaviruses have attracted renewed interest in view of a novel coronavirus disease outbreak of 2019 (COVID-19) that originated in Wuhan, China

https://doi.org/10.1016/j.nmni.2022.100955

[5]. The causative agent was found to be a novel coronavirus (2019-nCoV) that was identified in December, 2019. Barely two months after its discovery, had the disease become a pandemic of global concern causing widespread morbidity and mortality [6,7]. Because COVID-19 is a serious respiratory disease, the causative pathogen was later known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).

Before the origin of SARS-CoV-2, six coronaviruses were known to infect man [5,8]. Out of these, severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV) caused severe respiratory illness in humans [9,10]. In 2002, SARS-CoV, originating from Guangdong Province, Southern China, and MERS-CoV emerging from Saudi Arabia caused serious epidemics in 2002/2003 and 2012 respectively [11,12]. The severity of the respiratory diseases occasioned by these viruses and the resulting wide spread morbidity and mortality had been reported elsewhere [13].

Immediately after its identification, scientists in several laboratories swiftly sequenced SARS-CoV-2 genome making the complete and partial sequences available in public databases. Within few months, the number of viral genome entries in databanks grew exponentially, and this afforded scientists unprecedented opportunity for undertaking molecular, phylogenetic and epidemiological studies on SARS-CoV-2. It was re-established that coronaviruses including SARS-CoV-2 possess the largest genome ranging from 26.4 to 31.7 kb among all RNA viruses [14,15]. Like other coronaviruses, SARS-CoV-2 contains a positive-sense single-stranded RNA (ssRNA). The first two-third portion of the genome beginning from the 5′ end codes for a polyprotein pp1ab, whose cleavage gave several non-structural proteins (nsps) that are mainly associated with genome replication and transcription [15,16]. On the remaining portion of coronavirus genome (i.e. towards the 3′ end), there are four genes that encode viral structural proteins namely spike (S), envelope (E), membrane (M) and nucleocapsid (N) necessary for viral infectivity and transmissibility [17]. Out of these four structural genes, the S gene has attracted the greatest attention because it encodes the spike protein which binds human angiotensin-converting enzyme (ACE2) for viral attachment and entry into human cells. SARS-CoV-2 appears to be highly optimized for binding to human ACE2 receptor using the receptor binding domain (RBD) on the spike glycoprotein [18].

The first coronavirus genome sequence report in Africa was from a Nigerian index case that traveled to the country from Italy in February, 2020 [19]. In early March, 2020, complete (29,798 nt) and partial (486 nt) genome sequences of SARS-CoV-2 collected from the index case were produced by Nigerian scientists at African Centre of Excellence for the Genomics of Infectious Disease (ACEGID) at the Redeemers University of Nigeria (RUN), and the Centre for Virology and Genomics at the Nigerian Institute of Medical Research (NIMR), Nigeria, respectively. The partial and the complete sequences were deposited in the National Center for Biotechnology Information (NCBI) GenBank (Accession No: MT159778) and the Global Initiative on Sharing All Influenza Data (GISAID) databases (EPI_ISL_413550), respectively. At the time of writing this article on the 5th August, 2020, additional 41 complete SARS-CoV-2 Nigerian SARS-CoV-2 genome sequences have been added to GISAID by ACEGID to give a total of 42 complete sequence records. Preliminary molecular and phylogenetic analyses gave a phylogenetic tree where the Nigerian index case clustered with a European clade [19]. Subsequent analysis of 18 complete sequences suggested multiple introduction and community transfer in Nigeria [20].

These preliminary reports focused on SARS-CoV-2 spike protein mutations without a comprehensive molecular characterization in terms of number and distribution of SNPs, haplotype, and conserved genomic regions of Nigerian SARS-CoV-2. In order to understand SARS-CoV-2 evolutionary dynamics in any current or future studies in Nigeria, it is important to evaluate the diversity and the phylogenetic relationship of the SARS-CoV-2 circulating in Nigeria during the early period of COVID-19. We hereby present a report of a detailed computational analysis and phylogenetic reconstruction of the Nigerian SARS-CoV-2 using the viral genome data deposited in GISAID database between March 2019, when the viral genome of Nigeria's index case was sequenced and July 18, 2020. The results of this study, which concentrated on genomes obtained from early COVID-19 cases in Nigeria, will aid the understanding of variability and evolution of SARS-CoV-2 genome in this region. Such knowledge is crucial for understanding the epidemiological and evolutionary dynamics of SARS-CoV-2 in any present and future phylogenetic studies. Furthermore, the results will aid diagnosis, prognosis, and treatment of COVID-19 in Nigeria.

Materials and methods

Data retrieval

Whole genome sequence data of SARS-CoV-2 isolated from Nigerian COVID-19 cases were retrieved by downloading from GISAID (Global Initiative on Sharing All Influenza Data) database. A total of 43 sequences were deposited as at 18th July, 2020. Out of these, 42 sequences were complete genome sequences when filtered out by selecting the "complete" filter button in GISAID database page implying that one sequence was partial. To ensure acceptable size and quality of SARS-CoV-2 genome sequence for the study, "high coverage" filtering button was also selected. This and quality check by visualization

left us with 18 sequences that satisfied all the selection criteria (Table 1). According to GISAID database, "complete" genome sequences are sequences with > 29,000 nucleotides (nts) while "high coverage" sequences are those with < 1% "Ns" and no insertions and deletions unless verified by the submitter. The sequences were retrieved by downloading in FASTA file format from GISAID database for further analysis. In addition Genome sequence of SARS-CoV-2 of Nigeria's index COVID-19 case (EPI_ISL_413550) and the reference genome (NC_045512.2) were obtained from GISAID and NCBI GenBank databases, respectively. In view of the limited sample obtained in Nigeria during the early period of COVID-19 outbreak and the constrain imposed by the requirement for high quality genome sequence in Nigeria, variants identified in selected countries in the Northern (Sweden) and Southern Europe (Italy and Croatia) were included in the study.

Sequence homology and mapping

To ensure true homology, non-biological differences (i.e. differences due to technical variations) were removed from the retrieved sequences. The sequences were aligned in MAFFT and trimmed at the 5′ and 3′ ends in MEGA X to obtain homologous sequences of 29,787 nts each. Alignment revealed that 49 nucleotides should be removed at the 5′ end while 67 nucleotides should be pruned away at the 3′ end to obtain the 29,787 nts for each of the homologous sequences used for the study. The trimmed sequences were mapped onto a reference genome sequence of SARS-CoV-2 retrieved by downloading from NCBI GenBank (Accession No: NC_045512.2) in order to determine the location of sites on the sampled genome sequences [15]. The 5′ and the 3′ ends of the aligned sequences are shown in Fig. 1.

Multiple sequence alignment

Multiple sequence alignment (MSA) was by MAFFT Version 7.471 while phylogenetic reconstruction was 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 (Fig.1) to produce aligned sequences that were fed into DnaSP for SNP and haplotype analysis and subsequently into Jalview 2.11.1.0 for visualization and automatic determination of allelic frequency of SNPs.

TABLE 1. Summary statistics of the retrieved sequences

Statistical parameter	Value
Mean (nts)	29,863
SD	35.11
Maximum Size (nts)	29,903
Minimum Size (nts)	29,787
Coeff of Variation (%)	0.12

Variation and SNP analysis

Genomic locations of SNPs were determined relative to the SARS-CoV-2 reference genome sequence obtained from NCBI GenBank (Accession No: NC_045512.2). A polymorphic site is considered as a genomic location with the frequency of the second most frequent allele greater than 1%; otherwise, they were referred to as monomorphic. Rare alleles are those with minor allele frequency (MAF) less than 1%, and the viruses carrying them were rare variants. Detection of SNPs, linkage disequilibrium and haplotype analysis was done in DnaSPv6.12.03 while SNP distribution and density was obtained by a series of MS Excel formulae packaged by us.

Lineage analysis

Lineage analysis using the retrieved SARS-CoV-2 genome sequences were by the pangolin v2.4.2 package [30].

Phylogenetic analysis

Maximum likelihood phylogenetic tree construction was implemented in MEGA X using sequences that had been aligned by MAFFT employing Tamura-Nei evolutionary model under assumption of uniform nucleotide substitution [21,22]. The tree that had topology with superior log likelihood value was selected from the initial trees by applying neighbor-joining (NJ) and bio-NJ algorithm in a heuristic search. Analysis of the tree was based on topology and clustering pattern analysis. Tree validation was by 1000 bootstrap replicates [23].

Results

Summary statistics of the retrieved sequences

The summary statistics of 18 SARS-CoV-2 complete sequences that satisfied our selection criteria are presented in Table 1. The complete genome sequences of mean size 29,863 ± 35.11 (mean + SD) were obtained from the retrieved sequences from SARS-CoV-2 isolated from 18 subjects (8M: 6F: 4 Unknown Sex) from six states in Nigeria namely, Kwara, Ogun, Osun, Oyo, Ekiti, and Ondo.

Results of SNP analysis

Nigerian SARS-CoV-2 had 99.9% genomic similarity with four large conserved genomic regions (Region 1: 13,020-14407, Region 2: 16,742-18876, Region 3: 20,375-21,897, and Region 4: 23,407-25,470). A total of 66 SNPs were identified in the SARS-CoV-2 genomes used for this study out of which 31 were

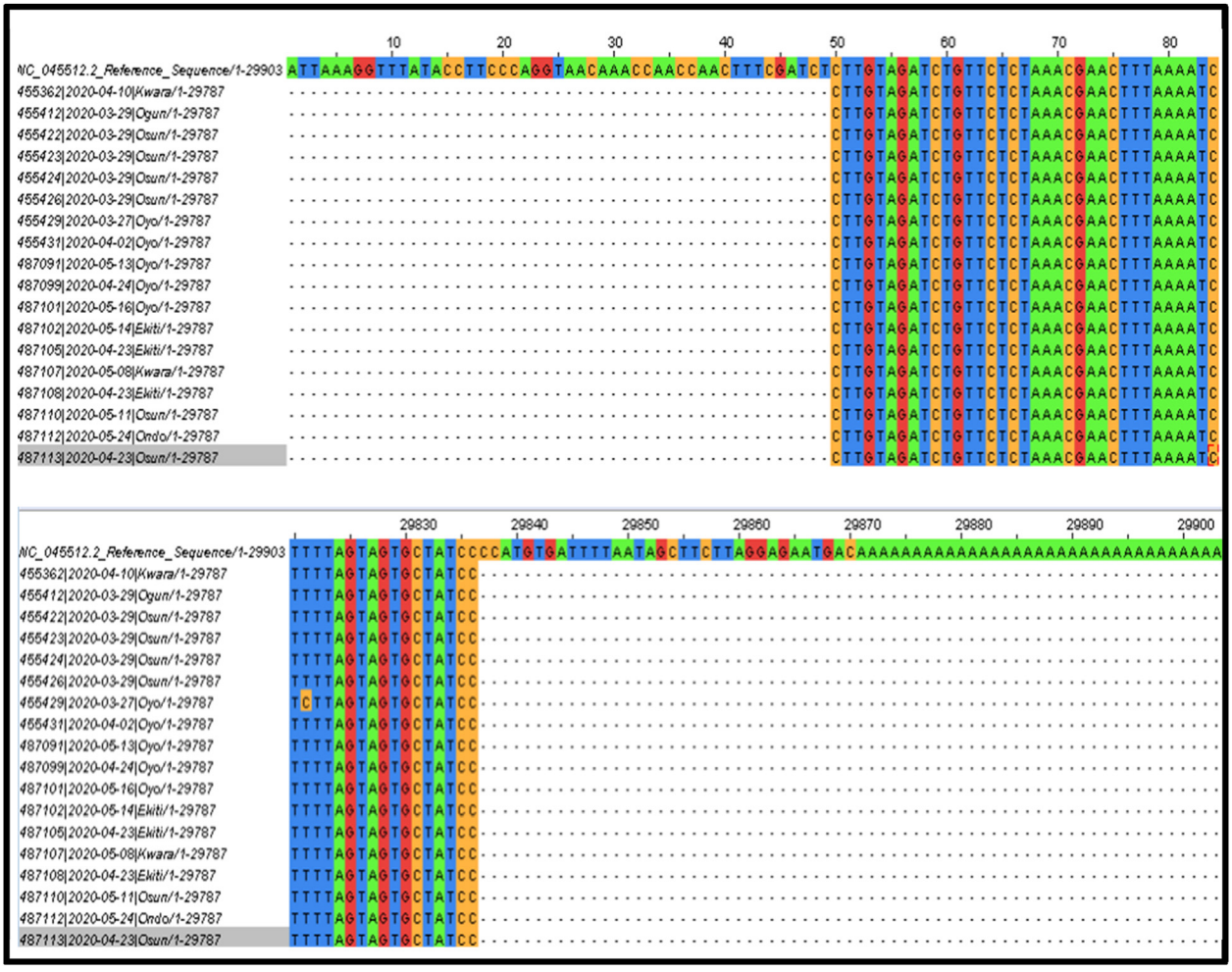

FIG. 1. Aligned Sequences Trimmed at 5′ and 3′ ends to Obtain True Homology.

informative. All the SNPs were diallelic except a SNP located at 29,791 nt which was triallelic. The overall nucleotide diversity among the SARS-CoV-2 genomes analyzed was Pi = 0.00048. The average frequency of SNPs on SARS-CoV-2 genomes in Nigeria was 2.22 SNPs per 1000 nts (Table 2). The SNPs were not, however, evenly distributed in the entire genome: highest SNP densities were observed in the 5′UTR and the 3′UTR regions with SNP frequency of 3.77 and 35.4, respectively. The coding region with the highest SNP density was ORF10 (8.55 SNPs/1000 nts). No SNP was detected in ORFs 6, 7a, 7b, and E gene.

Haplotype analysis for delineation of L and S lineages

Linkage analysis involving pairwise comparison of SNPs showed that several SNPs are in 2-locus linkage disequilibrium (LD) in SARS-CoV-2 genome with majority of the SNPs above the significant ($P < 0.05$) horizontal threshold line indicated in Fig. 2. The total number of significant pairwise haplotypes detected by Fisher's exact test was 108 out of which 12 were highly significant ($P < 0.001$) by Bonferoni procedure (Table 3). A linkage of two SNPs: a C/T SNP at location 8782 nt in the ORF1ab region and a T/C SNP at 28,144 nt in ORF8 region (Fig. 3) originally identified by Cui *et al.*, [1], were also detected in this study (Haplotype 7). The frequency of CT haplotype (L lineage) and TC (S lineage) were 13 (72.2%) and 5 (27.8%), respectively (Table 4).

Results of D614G mutation analysis

Multiple sequence alignment of translated SARS-CoV-2 spike (S) gene revealed a D/G amino acid substitution at position 614 of the spike protein (Fig. 4). There was a preponderance of D614G mutation because majority 11 (61.1%) of the analyzed genomic sequences of SARS-CoV-2 had G at 614 position of spike protein as compared to 7 (39.9%) that had D at the site.

TABLE 2. Distribution and density of SNPs in SARS-CoV-2 isolated in Nigeria

	5′ UTR	ORF 1a	ORF 1b	S	ORF 3a	E	M	ORF 6	ORF 7a	ORF 7b	ORF 8	N	ORF 10	3′ UTR	Spacer	Whole Genome
Length of Region	265	13,218	8088	3822	828	228	669	186	366	132	366	1260	117	113	138	29,787
SNP Distribution	1	31	10	5	5	0	2	0	0	0	2	5	1	4	0	66
SNP Density (SNPs/1000 nts)	3.77	2.35	1.24	1.31	6.04	0	2.99	0	0	0	5.46	3.97	8.55	35.4	0	2.22

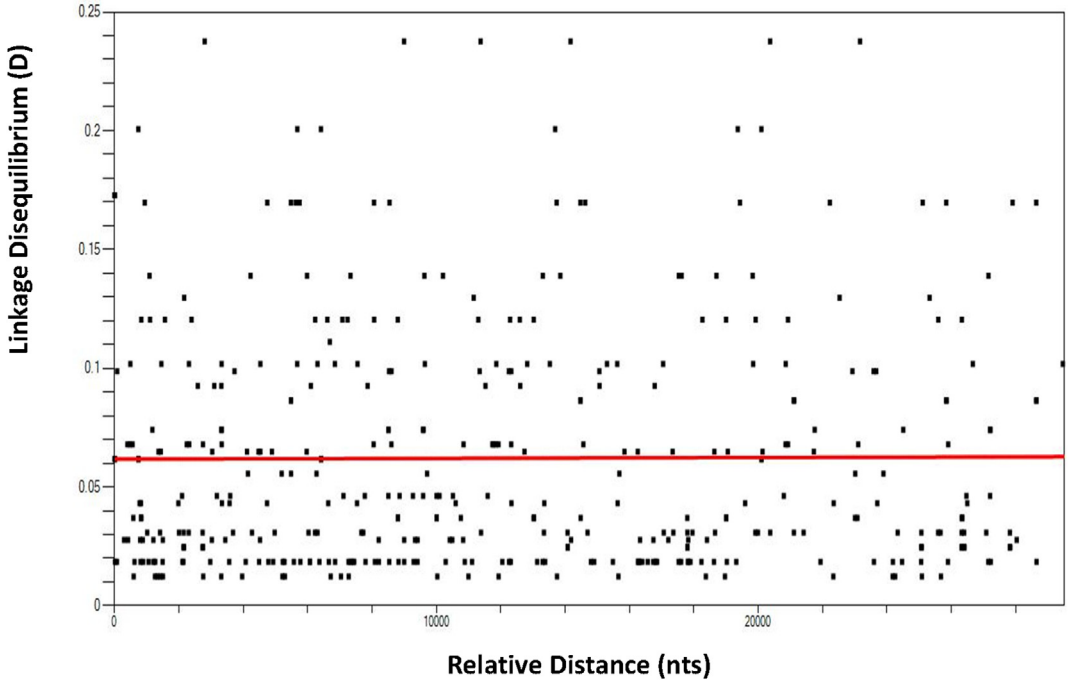

FIG. 2. SNPs in linkage disequilibrium in Nigerian SARS-CoV-2 genome. **Note:** SNPs above the line are in significant linkage disequilibrium.

Nigeria SARS-CoV-2 lineages

Out of the 32 sequences used for this study, majority (27 = 84.4%) were of B lineage while only 5 (15.6%) were of A lineage. All European lineages were B. Nineteen sequences (including the index case) were from Nigeria, retrieved for analysis, out of these, 5 (55.6 %) were of A lineage. Interestingly, 4 out of the 5 A lineages were from the same state (i.e. Osun State).

TABLE 3. Haplotypes obtained from sites in highly significant linkage disequilibrium

Haplotype No	Site 1	Site 2	Distance (nts)	Linkage Diseq. (D)
1	241	3037	2796	0.238[a]
2	241	14,408	14,167	0.238[a]
3	241	23,403	23,162	0.238[a]
4	3037	14,408	11,371	0.238[a]
5	3037	23,403	20,366	0.238[a]
6	8782	22,468	13,686	0.201[a]
7	8782	28,144	19,362	0.201[a]
8	8782	28,878	20,096	0.201[a]
9	14,408	23,403	8995	0.238[a]
10	22,468	28,144	5676	0.201[a]
11	22,468	28,878	6410	0.201[a]
12	28,144	28,878	734	0.201[a]

[a]P < 0.001 (Bonferoni Correction).

Phylogenetic analysis

Clustering analysis of the maximum likelihood phylogenetic tree gave 3 major clades. On the basis of the preponderance of taxa in each cluster, the clusters were identified as Oyo, Ekiti and Osun clades respectively. Reference sequence clustered with Osun clade while the index case clustered with Ekiti clade (Fig. 5). Majority of the nodes have high bootstrap support (60%–100%).

Four major clusters were identified as depicted in Fig. 6. The index case clustered mostly with European countries as expected. However, it was closer to Sweden in Northern Europe than Italy and Croatia in Southern Europe.

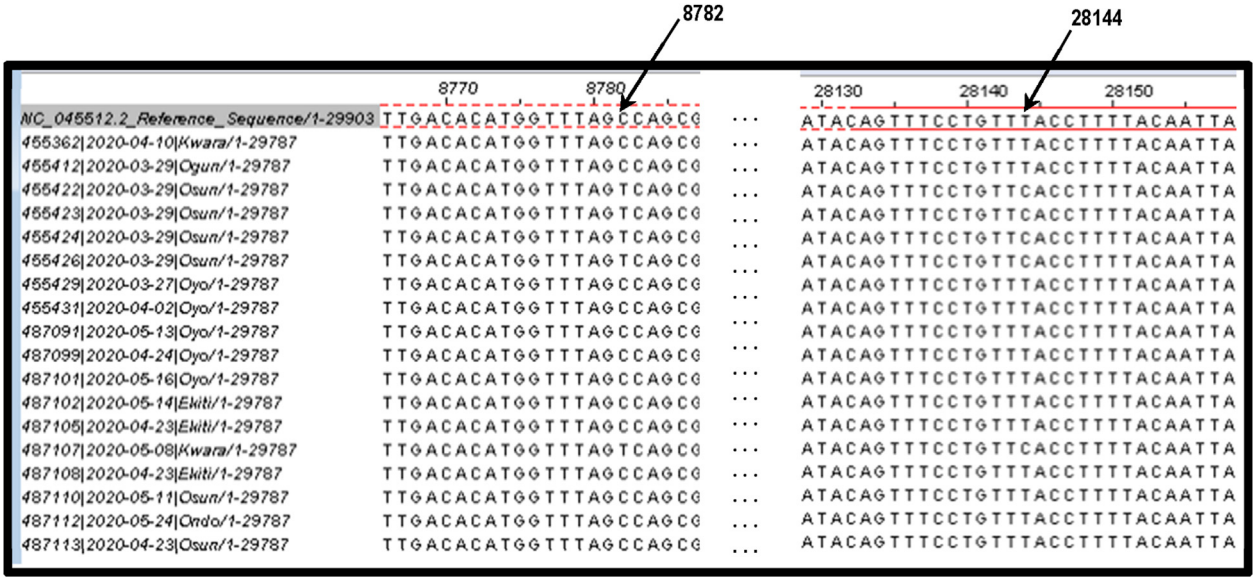

FIG. 3. Complete linkage disequilibrium of C8782 and T28144 locations in the genome of Nigerian SARS-CoV-2.

Discussion

In the present study, whole-genome sequence characterization and phylogenetic reconstruction of SARS-CoV-2 isolated from Nigerian COVID-19 subjects was carried out. Out of the 42 Nigerian complete genome sequences in GISAID database, only 18 fulfilled our selection criteria for inclusion in the study. In view of the implications of data quality for result validity, it was better, in our opinion, to use 18 sequences of acceptable quality rather than using 42 sequences where more than half of the sequences are of low or questionable quality. Although the genome sequence from the Nigerian index case did not fulfil our selection criteria, we had to include it in the analysis being the first SARS-CoV-2 genome sequenced in Africa [19]. The index case clustered with a European clade on the phylogenetic tree. This is consistent with the travel history of the index case. It is not clear why the index case clustered more with Sweden in Northern Europe than with Italy or Croatia in Southern Europe, since the index case travelled from Italy to Nigeria.

Nigerian SARS-CoV-2 had 99.9% similarity, implying 0.01% dissimilarity to the reference genome sequence which is in accordance with the global trend [15,16]. It was therefore not surprising that four large conserved genomic regions were identified in this study. This observation supports the well-established view that the novel virus has a recent origin estimated to be between 6th October and 11th December 2019 [24]. Furthermore, Nigeria has extended family and communal structure, which could favour community transmission of the novel virus. Thus, Nigeria's socio-cultural attributes may contribute (at least in part) to the high degree of similarity and the ethnic based clustering pattern on the phylogenetic tree obtained in this study. It will be of interest to carry out comparative analysis of conserved regions involving genomes of SARS-CoV-2 in other parts of the world to compare their tree topologies with that of Nigeria.

Despite the high degree of similarity, some degree of diversity resulting from SNPs found in SARS-CoV-2 genome, was detected. The number of SNPs detected in this study was low when compared with other studies [25]. A limitation of this study is the small sample size of nineteen genome sequences used in this study, which we consider too small for detection of all SARS-CoV-2 SNPs in Nigeria. In subsequent studies, it will be of interest to see if distribution of SNPs and conserved regions identified in this study are peculiar to Nigerian SARS-CoV-2 or have complete overlap with genomes of SARS-CoV-2 found elsewhere. If there are regions of non-overlap, any attempt to produce testing kits with high sensitivity for Nigeria's COVID-19 cases should take note of the conserved regions and SNP distribution in the genomes of SARS-CoV-2 detected in this study.

TABLE 4. Haplotype analysis for the frequency of L and S lineages

Site 8782	Site 28,144	Haplotype	Lineages	Number	Frequency (%)
C	T	CT	L	13	72.2
T	C	TC	S	5	29.8

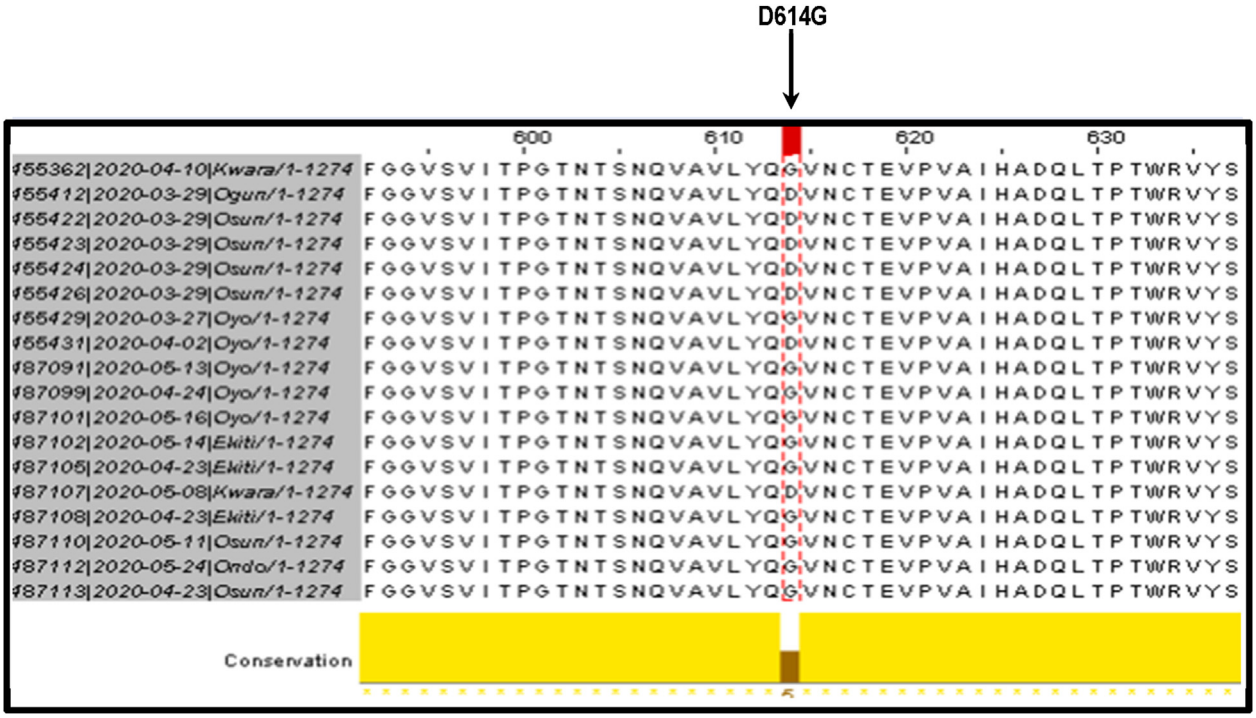

FIG. 4. Multiple sequence alignment of translated spike protein revealing preponderance of D614G mutation.

The frequency of SNPs observed in the Nigerian SARS-CoV-2 genome is higher than twice the frequency of SNPs in human genome which is generally taken to be 1 SNP per 1000 bps. The UTRs, especially 3′UTR, are mutation hot spots in Nigerian SARS-CoV-2 genome. This is expected because UTRs, unlike coding sequences, are generally under more relaxed or neutral

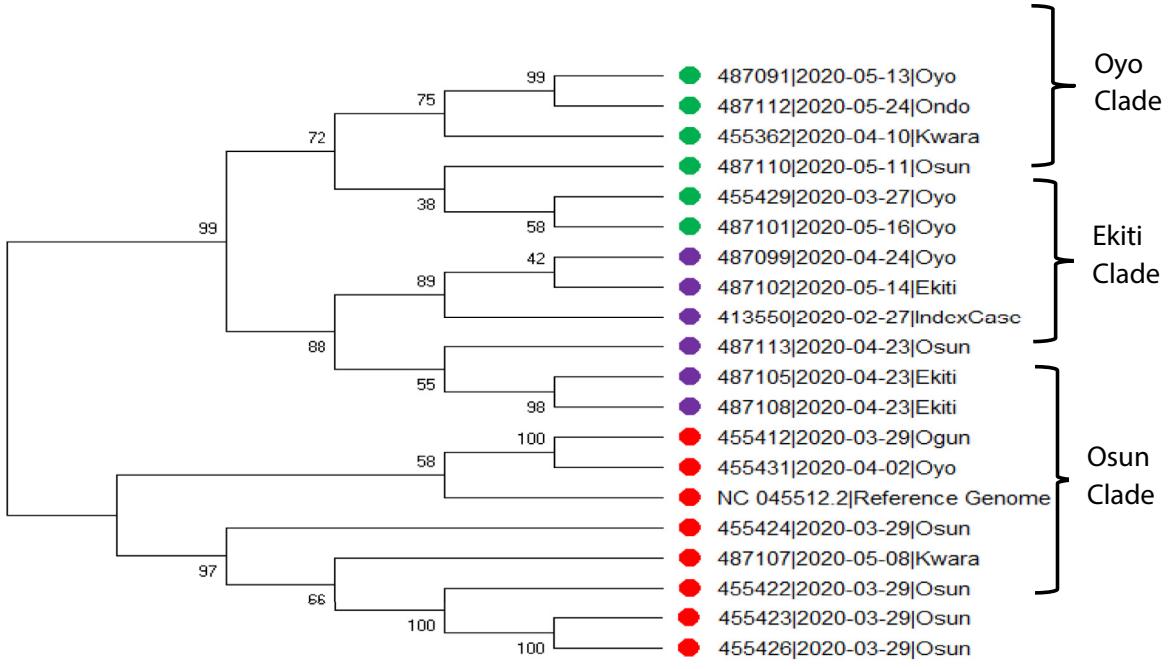

FIG. 5. Maximum likelihood phylogenetic tree of Nigerian SARS-CoV-2 using whole genome sequences.

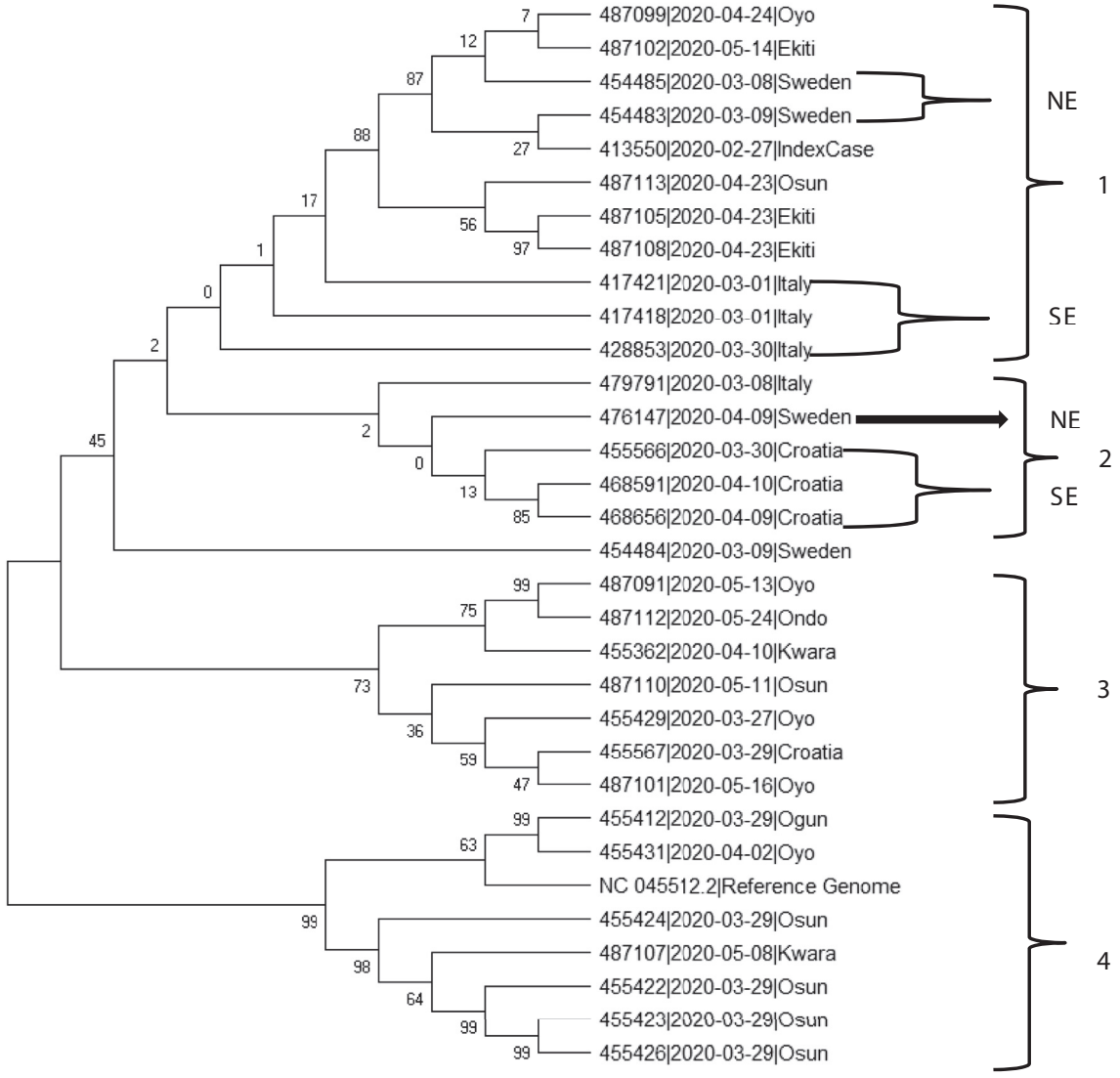

FIG. 6. Maximum likelihood tree of Nigerian and selected European SARS-CoV-2 genome sequences. Note: NE: Northern Europe, SE: Southern Europe.

selection pressure which allow mutations to accumulate at a higher rate in that region [25]. Relatively high SNP densities were also recorded in some of the coding regions of Nigerian SARS-CoV-2. An example is ORF 10 region with a SNP density of 8.55 SNPs/1000 nts. This coding region had the highest SNP density, implying that this is the most polymorphic region in Nigeria SARS-CoV-2 genome. It is of note that the spike gene, the region of greatest interest has lower SNP density when compared to ORF1a and some other regions in this study. Although coronaviridae have proofreading capability due to their exonuclease activity during nucleotide replication, mutation rate of SARS-CoV-2 was estimated at ~6 × 10^{-4} nucleotides/genome/year with the capacity to mutate during human-to-human transmission [24].

We share the view that characterizing organisms and viruses by haplotypes is more informative than using individual SNPs. Such haplotype analysis gave results which agreed with an early report by Cui *et al.*, [1] who proposed two major lineages namely the L lineage carrying CT haplotype at the closely linked positions C8782 and T28144 respectively and the S lineage with nucleotides T and C (TC haplotype) at the corresponding genome locations. The higher frequency of L lineage viruses

which was higher than those of S lineage in this study is consistent with the original finding by Cui *et al.*, [1]. Few recombinants were observed in their study unlike the present study where both loci were absolutely linked such that no recombinants were recovered. According to Cui *et al*, [1], identification as L and S lineages was based on the amino acids resulting from T28144C substitution in which leucine (L lineage) is replaced with serine (S lineage).

Similar increased frequency of viruses in the L lineage as compared to those in the S lineage was observed at 23,403 nt position where A23403G nucleotide substitution caused D614G mutation on SaRS-CoV-2 spike protein. Majority of SARS-CoV-2 in Nigeria were G614 (G type) mutants as opposed to the D614 (D type) ancestral form in agreement with the global trend [26]. Preponderance of L and G variants in many populations was accounted for by their increased Darwinian fitness whereby these newly evolved strains have competitive advantage over their alternative and more ancestral forms in terms of infectivity, transmissibility and, possibly, virulence. Others felt that the preponderance of the L and the G mutants might result from sampling bias and founder effect [27].

Available data suggest that Lineage A shared the same genome sequence with the most recent common ancestor (MRCA) of SARS-CoV-2, although Lineage B sequence was published first [30]. Existence of Lineage A in Ekiti State during the early period of COVID-19 pandemic suggests that SARS-CoV-2 was already circulating in southern Nigeria before the index case brought SARS-CoV-2 to the country.

Results of studies from our group suggested that SNPs and mutation pattern of SARS-CoV-2 vary between continents and, in some cases, between different countries in the same continent [31]. In view of the fact that testing kits are generally nucleic acid based and the potential of SARS-CoV-2 to evade detection due to its mutational dynamics, it is difficult or impossible to have universal testing kits with equal efficiency for detecting the novel coronavirus in all parts of the world [28]. Therefore, the identified conserved regions and the distribution of SNPs in the genome of SARS-CoV-2 circulating in Nigeria has implications for Nigeria-based intervention programmes in the area of viral detection, drug designing and treatment options as had been generally noted by Vankadari [29]. According to Puty *et al.* [28], it is important for every population to identify key SARS-CoV-2 mutations and SNPs for appropriate population-based intervention specific for its population. The view of Wang *et al.* [15], that limited sensitivity of SARS-CoV-2 detection kits is possibly associated with genetic variation of the virus is pertinent. Thus, any meaningful intervention activity should consider SARS-CoV-2 strains in the concerned population.

Compliance with ethical standards

Complied with ethical standards. No ethical clearance required.

Consent for publication

The author(s) read and approved the final manuscript.

Transparency declaration

The authors declare no competing interest.

Funding

No funding.

Authors' contribution

IT designed the study; IU, NA & FA contributed to the design of the study IT, IU, NA, FA & AA were involved in data retrieval; IT, NA, IU & FA wrote the draft of the manuscript, IT, TB, IU & AA sourced the software used. IT provided critical revision of the manuscript. All authors have read and agreed to the published version of the manuscript.

Credit authors statement

IT: Conceptualization; IU, NA & FA: contributed to Conceptualization; IT, IU, NA, FA & AA: Data curation; Formal analysis; Resources; IT, NA, IU & FA: Roles/Writing – original draft; IT, TB, IU & AA: Project administration; Resources; Software; IU: Project administration and; IT: Supervision and review & editing. All authors have read and agreed to the published version of the manuscript.

Acknowledgement

We thank scientists at African Centre of Excellence for Genomics of Infectious Disease (ACEGID) at the Redeemers University of Nigeria (RUN), and Centre for Human Virology and Genomics, Nigerian Institute of Medical Research (NIMR), Nigeria, respectively for depositing complete and partial SARS-

CoV-2 sequences at GISAID and NCBI GenBank respectively. We also thank their partners and collaborators namely Nigerian Centre for Disease Control (NCDC), Centre for Human and Zoonotic Virology (CHAZVY), College of Medicine, University of Lagos, Nigeria, and Infectious Disease Hospital, Yaba, Nigeria.

References

[1] Cui J, Li F, Shi Z-L. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 2019;17(3):181–92. https://doi.org/10.1038/s41579-018-0118-9.

[2] Sternberg A, Naujokat C. Structural features of coronavirus SARS-CoV-2 spike protein: Targets for vaccination. Life Sci 2020:118056. https://doi.org/10.1016/j.lfs.2020.118056. PubMed PMID: 32645344; PubMed Central PMCID: PMCPMC7336130.

[3] Li G, Fan Y, Lai Y, Han T, Li Z, Zhou P, et al. Coronavirus infections and immune responses. J Med Virol 2020;92(4):424–32. https://doi.org/10.1002/jmv.25685.

[4] Abdel-Moneim AS, Abdelwhab EM. Evidence for SARS-CoV-2 infection of animal hosts. Pathogens 2020;(7):9. https://doi.org/10.3390/pathogens9070529. PubMed PMID: 32629960.

[5] Li D, Zhang J, Li J. Primer design for quantitative real-time PCR for the emerging Coronavirus SARS-CoV-2. Theranostics 2020;10(16):7150–62. https://doi.org/10.7150/thno.47649. PubMed PMID: 32641984; PubMed Central PMCID: PMCPMC7330846.

[6] Chen J, Bai H, Liu J, Chen G, Liao Q, Yang J, et al. Distinct clinical characteristics and risk factors for mortality in female COVID-19 inpatients: a sex-stratified large-scale cohort study in Wuhan, China. Clin Infect Dis 2020. https://doi.org/10.1093/cid/ciaa920. PubMed PMID: 32634830.

[7] Du RH, Liang LR, Yang CQ, Wang W, Cao TZ, Li M, et al. Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study. Eur Respir J 2020. https://doi.org/10.1183/13993003.00524-2020. PubMed PMID: 32269088; PubMed Central PMCID: PMCPMC7144257.

[8] Cheng VCC, Lau SKP, Woo PCY, Yuen KY. Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection. Clin Microbiol Rev 2007;20(4):660–94. https://doi.org/10.1128/cmr.00023-07.

[9] Baddal B, Cakir N. Co-infection of MERS-CoV and SARS-CoV-2 in the same host: a silent threat. J Infect Publ Health 2020. https://doi.org/10.1016/j.jiph.2020.06.017. PubMed PMID: 32622797; PubMed Central PMCID: PMCPMC7306724.

[10] Fouchier RAM, Kuiken T, Schutten M, van Amerongen G, van Doornum GJJ, van den Hoogen BG, et al. Koch's postulates fulfilled for SARS virus. Nature 2003;423(6937):240. https://doi.org/10.1038/423240a.

[11] Peiris JSM, Guan Y, Yuen KY. Severe acute respiratory syndrome. Nat Med 2004;10(12):S88–97. https://doi.org/10.1038/nm1143.

[12] Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. New Engl J Med 2012;367(19):1814–20. https://doi.org/10.1056/NEJMoa1211721. PubMed PMID: 23075143.

[13] Arabi YM, Balkhy HH, Hayden FG, Bouchama A, Luke T, Baillie JK, et al. Middle East respiratory syndrome. New Engl J Med 2017;376(6):584–94. https://doi.org/10.1056/NEJMsr1408795. PubMed PMID: 28177862.

[14] Woo PCY, Huang Y, Lau SKP, Yuen K-Y. Coronavirus genomics and bioinformatics analysis. Viruses 2010;2(8):1804–20. https://doi.org/10.3390/v2081803. PubMed PMID.

[15] Wang C, Liu Z, Chen Z, Huang X, Xu M, He T, et al. The establishment of reference sequence for SARS-CoV-2 and variation analysis. J Med Virol 2020;92(6):667–74. https://doi.org/10.1002/jmv.25762.

[16] Sah R, Rodriguez-Morales AJ, Jha R, Chu DKW, Gu H, Peiris M, et al. Complete genome sequence of a 2019 novel coronavirus (SARS-CoV-2) strain isolated in Nepal. Microbiol Res Announc 2020;9(11). https://doi.org/10.1128/mra.00169-20. e00169–20.

[17] Paraskevis D, Kostaki EG, Magiorkinis G, Panayiotakopoulos G, Sourvinos G, Tsiodras S. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. Infect Genet Evol 2020;79:104212. https://doi.org/10.1016/j.meegid.2020.104212.

[18] Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med 2020;26(4):450–2. https://doi.org/10.1038/s41591-020-0820-9.

[19] Ihekweazu C, Happi C, Omilabu S, Salako BL, Abayomi A, Oluniyi PE. First African SARS-CoV-2 genome sequence from Nigerian COVID-19 case. Genome Rep 6th March 2020. Available from: https://virological.org/t/first-african-sars-cov-2-genome-sequence-from-nigerian-covid-19-case/421.

[20] Happi C, Ihekweazu C, Oluniyi PE, Olawoye I. SARS-CoV-2 Genomes from Nigeria Reveal Community Transmission, Multiple Virus Lineages and Spike Protein Mutation Associated with Higher Transmission and Pathogenicity. Genome Rep 6th March 2020. Available from: https://virological.org/t/sars-cov-2-genomes-from-nigeria-reveal-community-transmission-multiple-virus-lineages-and-spike-protein-mutation-associated-with-higher-transmission-and-pathogenicity/494.

[21] Kumar S, Stecher G, Li M, Knyaz C, Tamura K, Mega X. Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 2018;35(6):1547–9. https://doi.org/10.1093/molbev/msy096.

[22] Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 1993;10(3):512–26. https://doi.org/10.1093/oxfordjournals.molbev.a040023.

[23] Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 1985;39:783–91.

[24] van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection, genetics and evolution. Epub 05/05 J Mol Epidemiol Evol Genet Infect Dise 2020;83:104351. https://doi.org/10.1016/j.meegid.2020.104351. PubMed PMID: 32387564.

[25] Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. Nat Sci Rev 2020;7(6):1012–23. https://doi.org/10.1093/nsr/nwaa036.

[26] Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 2020. https://doi.org/10.1016/j.cell.2020.06.043.

[27] Templeton AR. THE THEORY OF SPECIATION VIA THE FOUNDER PRINCIPLE. Genetics 1980;94(4):1011.

[28] Puty TC, Sarraf JS, Do Carmo Almeida TC, Filho VCB, de Carvalho LEW, Fonseca FLA, et al. Evaluation of the impact of single-nucleotide polymorphisms on treatment response, survival and toxicity with cytarabine and anthracyclines in patients with acute myeloid leukaemia: a systematic review protocol. Syst Rev 2019;8(1):109. https://doi.org/10.1186/s13643-019-1011-y.

[29] Vankadari N. Overwhelming mutations or SNPs of SARS-CoV-2: a point of caution. Epub 05/20 Gene 2020;752:144792. https://doi.org/10.1016/j.gene.2020.144792. PubMed PMID: 32445924.

[30] Rambaut A, Holmes EC, O'Toole Á, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020;5:1403–7. https://doi.org/10.1038/s41564-020-0770-5.

[31] Taiwo IA, Iwalokun BA, Samuel TA, et al. Genomic diversity and phylogenetic analysis of SARS-CoV-2 circulating in Africa and other continents: implications for diagnosis, transmission, and prevention. Pan Afr J Life Sci 2020;4(3):154–65. https://doi.org/10.36108/pajols/0202/40(0370).