microorganisms

MDPI

Article

The Spike Protein of SARS-coV2 19B (S) Clade Mirrors Critical Features of Viral Adaptation and Coevolution

Bidour K. Hussein, Omnia M. Ibrahium, Marwa F. Alamin, Lamees A. M. Ahmed, Safa A. E. Abuswar, Mohammed H. Abdelraheem and Muntaser E. Ibrahim *

Unit of Disease and Diversity, Department of Molecular Biology, Institute of Endemic Diseases, Khartoum University, Khartoum P.O Box 102, Sudan
* Correspondence: mibrahim@iend.org

Abstract: Pathogens including viruses evolve in tandem with diversity in their animal and human hosts. For *SARS-coV2*, the focus is generally for understanding such coevolution on the virus spike protein, since it demonstrates high mutation rates compared to other genome regions, particularly in the receptor-binding domain (RBD). Viral sequences of the *SARS-coV2* 19B (S) clade and variants of concern from different continents were investigated, with a focus on the A.29 lineage, which presented with different mutational patterns within the 19B (S) lineages in order to learn more about how *SARS-coV2* may have evolved and adapted to widely diverse populations globally. Results indicated that *SARS-coV2* went through evolutionary constrains and intense selective pressure, particularly in Africa. This was manifested in a departure from neutrality with excess nonsynonymous mutations and a negative Tajima D consistent with rapid expansion and directional selection as well as deletion and deletion–frameshifts in the N-terminal domain (NTD region) of the spike protein. In conclusion, we hypothesize that viral transmission during epidemics through populations of diverse genomic structures and marked complexity may be a significant factor for the virus to acquire distinct patterns of mutations within these populations in order to ensure its survival and fitness, explaining the emergence of novel variants and strains.

Keywords: *SARS-coV2*; Spike protein; 19B (S) clade; N-terminal domain (NTD region); synonymous and nonsynonymous mutations; deletion; deletion–frameshift; adaptation; neutrality

check for updates

Citation: Hussein, B.K.; Ibrahium, O.M.; Alamin, M.F.; Ahmed, L.A.M.; Abuswar, S.A.E.; Abdelraheem, M.H.; Ibrahim, M.E. The Spike Protein of SARS-coV2 19B (S) Clade Mirrors Critical Features of Viral Adaptation and Coevolution. *Microorganisms* **2022**, *10*, 2017. https://doi.org/10.3390/microorganisms10102017

Academic Editor: Deepak Shukla

Received: 15 September 2022
Accepted: 2 October 2022
Published: 12 October 2022

1. Introduction

SARS-coV2 is a member of the *Coronaviridae* family, with a wide range of viruses that affect humans, animals, or both [1]. Pathogens are being shown to coevolve in tandem with diversity in animals and humans [2–6]. The exact pathway and timeline of the virus's emergence and appearance of human cases is still unknown. Many theories exist to track the possible evolution of *SARS-coV2* from animals, including recombination events that began with bat corona viruses (*RmYN02*, *RpYN06*, *and PrC31*); it was found to be the closest ancestor for the virus in the whole genome apart from the spike protein in which *RaTG13* bat-derived virus is the closest. No intermediate host has been determined so far [2]. In the past two years, the virus evolved to give 12 variants, five of them are dominating and each with specific unique set of mutations. These variants include Alpha (B.1.1.7 and Q lineages), Beta (B.1.351 and descendent lineages), Gamma (P.1 and descendent lineages), Epsilon (B.1.427 and B.1.429), Eta (B.1.525), Iota (B.1.526), Kappa (B.1.617.1, B.1.617.3), Mu (B.1.621, B.1.621.1), Zeta (P.2), Delta (B.1.617.2 and AY lineages) and lastly Omicron (B.1.1.529 and BA lineages), which was reported in early November 2021. These variants have been divided into three categories either being variants of concern, interest or high consequences, the last which does not include any variant to date [7]. Although they circulate the globe, some dominate specific countries. *SARS-coV2* mutations are mainly concentrated on the spike protein and open reading frame 1 (ORF1), but as time passed, mutations expanded to include other open reading frames (ORFs) and structural proteins including the membrane, envelope and nucleocapsid proteins.

From the host side, a Spike protein's main function is coupling angiotensin-converting enzyme 2 receptor (ACE2), recognizing and fusing to facilitate viral entry to the host cell [8]. With the emergence of more transmissible and mutable variants, understanding the evolutionary characteristic of *SARS-coV2*'s spike genomic region is critical for predicting the path that reinfection, vaccination, and therapeutics will take [9]. The spike protein contains several conserved areas, but the region of RBD in the S1 subunit is the highest mutable region of the spike. S1 is where the initiation of the attachment of the virus to the ACE2 starts [10,11]. The *SARS-coV2* N-terminal domain (NTD region) of the spike protein also brought attention because its evolution is related to alteration of the viral antigenicity and promoting immune escaping. The more mutations and/or deletions in this region, the faster *SARS-coV2* will adapt, evolve and evade the immune system [12]. The mutational robustness found in this virus demonstrates its strength to tolerate host range expansion and adaptation, phenotypic plasticity or environmental stressors such as temperature, virulence or attenuation, antigenicity and immune escape [13]. Furthermore, the uncontrolled community transmission of *SARS-coV2* increases the possibility of the emergence of more transmissible variants, which is determined by host diversity in specific countries [14,15].

Here, we investigate *SARS-coV2* sequences: specifically, the 19B (S) clade, the first dominating variant after the ancestral virus of Wuhan [16], obtained from various countries and continents in order to understand the sequel of evolutionary processes and viral adaptation in disparate environments, particularly the putative effect of the population variation on its distribution and mutational variations.

2. Materials and Methods

2.1. Study Design

This is a retrospective cross-sectional study to address the global variation in the *SARS-coV2* evolution and how it may have adapted to the host selective pressure. This analysis is for the 19B (S) clade, the first dominating variant after the ancestral virus of Wuhan, covering the period from the start of the pandemic until February 2022.

2.2. Viral Genome Sequences

A total of 15,537 viral sequences of the 19B (S) clade were procured from the GISAID EpiCoV database [17] for the different continents at the period from the start of the pandemic until February 2022. Sequences of the lineages from A to A30, Bat, Pangolin as well as the variants of concern including Alpha, Beta, Gamma, Delta, Lambda and Omicron were also downloaded from the same database. They were chosen according to their first appearance in the collection dates. More focus was on the 19B (S) lineage A.29 that has been downloaded also.

We filtered sequences for the study based on the length of the sequence not to be less than 29,000 nucleotides, percentage of gaps, N stretches (unidentified nucleotides) of less than 5%, lack of clusters of mutations, and overlapping of reading frames. It was evaluated using the GISAID EpiCoV database and NEXTSTRAIN web tool [16,17].

2.3. Sequences Analysis

Sequences were analyzed using the Next-Generation Sequencing (NGS) analysis packages targeting areas of variations in all sequences included in the analysis, the COVID-19 genome annotator online tool for annotation and the NEXTSTRAIN online tool for defining clades for each sequence [16,18]. Sequences were aligned to *SARS-coV2* isolate Wuhan-Hu-1 with the accession number NC 045512.2S, and mutations were examined. The variations of nonsynonymous and synonymous mutations were estimated for the whole genome and the spike protein for each sample.

2.4. Phylogenetic Analysis

Phylogenetic analysis was carried out for those variants using the MEGAX software [19]. Maximum likelihood analysis was performed with bootstrapping (1000 iterations).

2.5. Evolutionary Distance and Neutrality Testing

The evolutionary distance and neutrality were tested by the Tajima neutrality test using MEGAX software. p values were estimated for the proportion of synonymous to nonsynonymous as an indicator of neutrality according to Kimura [20].

2.6. 19B (S) A.29 Lineage Analysis

The lineage of 19B (S) clade A.29 was taken as an example showing different patterns of mutations. Using a Python script, the frequency of mutations was calculated for each country involving the whole genome. Countries that reported A.29 were from Africa (Gambia and Sudan), Asia (India and Jordan), Europe (United Kingdom, Germany, and Belgium) as well as Oceania (Australia) and North America (the United States of America and Canada).

2.7. Data Visualization

The frequency of mutations for the lineage A.29 was displayed on bar charts and tables. It was displayed also in a timeline chart. Secondary RNA structures for the spike protein were then constructed by first extracting the Spike protein region using the seqkit command [21] and then uploading the sequences to the RNAFOLD online tool [22]. For the secondary structures created, the ensemble diversity and Minimal Free Energy (MFE) were estimated. To show the variation in the structure of the spike protein and the NTD region, a 3D model was created using the EXPASY translation [23] and the SWISSMODEL online tools [24].

3. Results

3.1. The Variation of the 19 (S) Clade across Continents

The distribution of the 19B (S) clade samples varied, expectedly, between and within countries due to the wide differences in the sequencing efforts. Findings revealed the existence of different patterns of distribution of the 19B (S) clade lineages across continents, in which for example, the A.23, A and A.27 were dominating Africa with approximate percentages of 39.8%, 16.8% and 14.9%, respectively, while the A lineage dominated Asia with 61.1%. The A.2 lineage was found to be common in Europe, North America, South America and Oceania with approximate percentages of 37.2%, 40.6%, 76.2% and 56.9%, respectively. In Africa, the lineages A18 to A30 were found to be more common than in other continents (Supplementary Table S1).

3.2. Nonsynonymous Mutations Found to Be Exceeding in Africa

In Africa, the total number of the 19B(S) clade samples was 1147, the total number of nonsynonymous to synonymous mutations was 1468\1013, and there were 28 deletions and 18 deletion–frameshifts across the whole genome. Whereas in North America, where $n = 8313$, there were 3822 nonsynonymous mutations, 2225 synonymous, and 55 deletions and deletions-frameshifts (each) (Table 1).

The same pattern of variation applies to other lineages and variants of concern. Based on the dates of collection reported in the databases combined with phylogenetic analysis, 11 of the lineages were found to have originated most likely in Africa including Senegal (A.11), Serra Leone (A.12), Burkina Faso (A.18), Côte d'Ivoire (A.19), Mali (A.21), Uganda (A.23), Kenya (A.25), Niger (A.27), Egypt (A.28), Gambia (A.29), and finally Angola (A.30). Nine of the other sub variants originated in Asia, and the rest were scattered in Europe and North America (Figure 1).

Table 1. The distribution of the 19B (S) clade in the different continents and the variation in the nonsynonymous, synonymous, deletion and deletion–frameshift. Both the whole genome and the spike alone were presented. Abbreviations: NS-WG: nonsynonymous mutations in the whole genome, S-WG: synonymous mutations in the whole genome, NS-Spike: nonsynonymous mutations in the spike protein and S-Spike: synonymous mutations in the spike protein. Note: Not all countries have updated sample uploading in the GISAID platform.

Continent	Total Number of 19B (S)	NS-WG	S-WG	Total Substitutions	NS-Spike	S-Spike	Total Substitution in the Spike	Deletion in WG	Deletion–Frameshift In WG	Percentages of Deletion–Frameshift
Africa	1147	1468	1013	2481	205	132	337	28	18	~1.6%
Asia	1349	1078	718	1796	150	152	302	18	27	~2%
Europe	3546	1985	1205	3190	277	154	431	39	45	~1.3%
North America	8313	3822	2225	6047	546	299	845	55	55	~0.66%
South America	244	342	241	583	47	29	76	3	5	~2.1%
Oceania	938	323	198	521	55	30	85	11	11	~1.2%

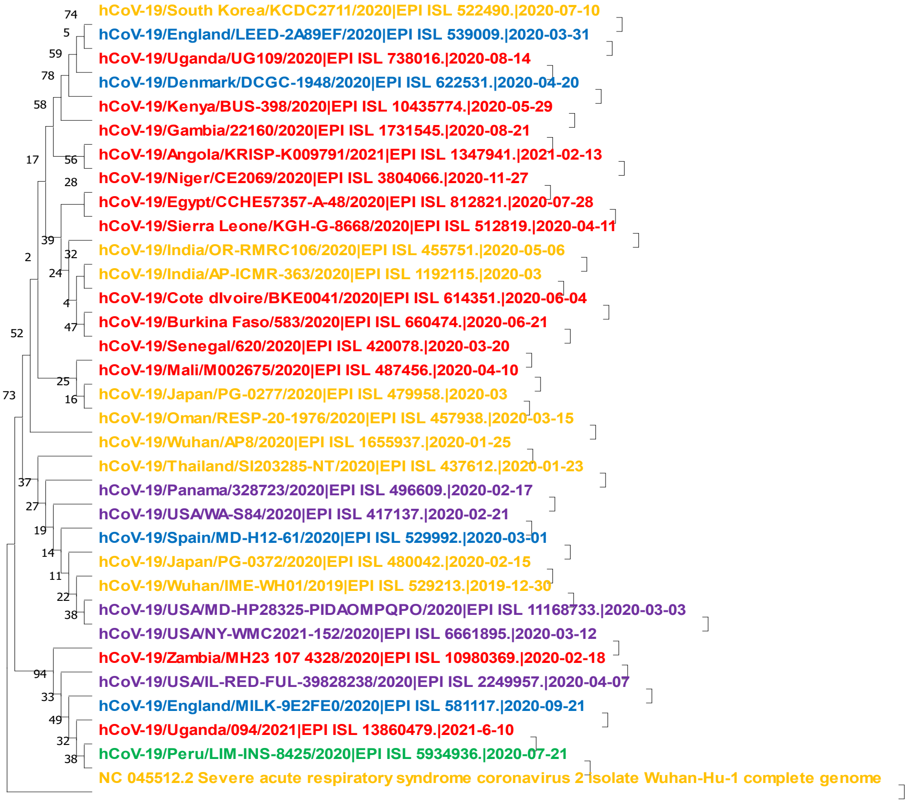

hCoV-19/South Korea/KCDC2711/2020|EPI ISL 522490.|2020-07-10
hCoV-19/England/LEED-2A89EF/2020|EPI ISL 539009.|2020-03-31
hCoV-19/Uganda/UG109/2020|EPI ISL 738016.|2020-08-14
hCoV-19/Denmark/DCGC-1948/2020|EPI ISL 622531.|2020-04-20
hCoV-19/Kenya/BUS-398/2020|EPI ISL 10435774.|2020-05-29
hCoV-19/Gambia/22160/2020|EPI ISL 1731545.|2020-08-21
hCoV-19/Angola/KRISP-K009791/2021|EPI ISL 1347941.|2021-02-13
hCoV-19/Niger/CE2069/2020|EPI ISL 3804066.|2020-11-27
hCoV-19/Egypt/CCHE57357-A-48/2020|EPI ISL 812821.|2020-07-28
hCoV-19/Sierra Leone/KGH-G-8668/2020|EPI ISL 512819.|2020-04-11
hCoV-19/India/OR-RMRC106/2020|EPI ISL 455751.|2020-05-06
hCoV-19/India/AP-ICMR-363/2020|EPI ISL 1192115.|2020-03
hCoV-19/Cote dIvoire/BKE0041/2020|EPI ISL 614351.|2020-06-04
hCoV-19/Burkina Faso/583/2020|EPI ISL 660474.|2020-06-21
hCoV-19/Senegal/620/2020|EPI ISL 420078.|2020-03-20
hCoV-19/Mali/M002675/2020|EPI ISL 487456.|2020-04-10
hCoV-19/Japan/PG-0277/2020|EPI ISL 479958.|2020-03
hCoV-19/Oman/RESP-20-1976/2020|EPI ISL 457938.|2020-03-15
hCoV-19/Wuhan/AP8/2020|EPI ISL 1655937.|2020-01-25
hCoV-19/Thailand/SI203285-NT/2020|EPI ISL 437612.|2020-01-23
hCoV-19/Panama/328723/2020|EPI ISL 496609.|2020-02-17
hCoV-19/USA/WA-S84/2020|EPI ISL 417137.|2020-02-21
hCoV-19/Spain/MD-H12-61/2020|EPI ISL 529992.|2020-03-01
hCoV-19/Japan/PG-0372/2020|EPI ISL 480042.|2020-02-15
hCoV-19/Wuhan/IME-WH01/2019|EPI ISL 529213.|2019-12-30
hCoV-19/USA/MD-HP28325-PIDAOMPQPO/2020|EPI ISL 11168733.|2020-03-03
hCoV-19/USA/NY-WMC2021-152/2020|EPI ISL 6661895.|2020-03-12
hCoV-19/Zambia/MH23 107 4328/2020|EPI ISL 10980369.|2020-02-18
hCoV-19/USA/IL-RED-FUL-39828238/2020|EPI ISL 2249957.|2020-04-07
hCoV-19/England/MILK-9E2FE0/2020|EPI ISL 581117.|2020-09-21
hCoV-19/Uganda/094/2021|EPI ISL 13860479.|2021-6-10
hCoV-19/Peru/LIM-INS-8425/2020|EPI ISL 5934936.|2020-07-21
NC 045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1 complete genome
74
5
59
78
58
17
56
28
39
2
32
24
4
47
52
25
73
16
37
27
19
14
11
22
38
94
33
49
32
38

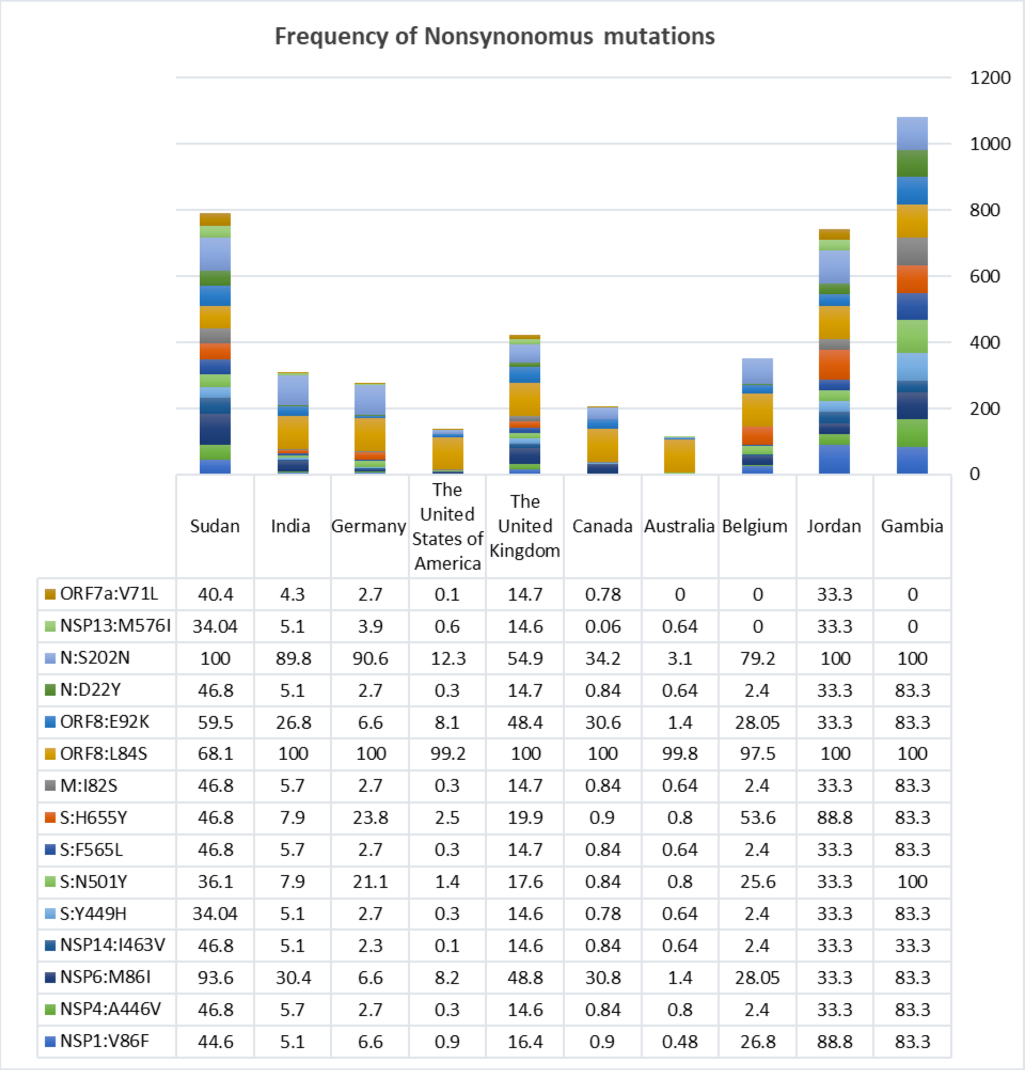
Figure 1. Evolutionary analysis by Maximum Likelihood method. The evolutionary history was inferred by using the Maximum Likelihood method and Kimura 2-parameter model. The tree with the highest log likelihood (−45061.28) is shown. The percentage of trees in which the associated taxa clustered together in 1000 bootstraps is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach and then selecting the topology with superior log likelihood value. This analysis involved the first reported lineages sequences of the 19B (S) reported in the databases based on their dates of collection. They were 33 sequences in number. Note: Red (Africa), Yellow (Asia), Blue (Europe), Purple (North America) and Green (South America).

3.3. Tajima's Neutrality Test Returned a Negative D Value Hence Directional Selection

Signals of selection were manifested in the excess of nonsynonymous mutations in the global sample x235.7 (p = 0.0001), particularly in the African continent in comparison to other continents (Z = 3.91 p = 0.0001) despite the small sample size. Tajima's neutrality test returned a negative D value of –2.646764, which is consistent with expansion and directional selection.

3.4. Mutational Variation in the A.29 Lineage of the 19B (S) Clade and the Spike Protein Structural Variation

The A.29 lineage of the 19B (S) clade presented with a specific mutational pattern and hence was selected to address questions of adaptation and coevolution. These samples had shared mutations that were different from other sub variants in the 19B (S) clade, which include: 15 nonsynonymous mutations, 9 synonymous mutations, 2 deletions, 1 frameshift and 1 extra-genic mutation. Frequencies of those mutations are shown in Figure 2 as well as Supplementary Figures S1 and S2.



	Sudan	India	Germany	The United States of America	The United Kingdom	Canada	Australia	Belgium	Jordan	Gambia
ORF7a:V71L	40.4	4.3	2.7	0.1	14.7	0.78	0	0	33.3	0
NSP13:M576I	34.04	5.1	3.9	0.6	14.6	0.06	0.64	0	33.3	0
N:S202N	100	89.8	90.6	12.3	54.9	34.2	3.1	79.2	100	100
N:D22Y	46.8	5.1	2.7	0.3	14.7	0.84	0.64	2.4	33.3	83.3
ORF8:E92K	59.5	26.8	6.6	8.1	48.4	30.6	1.4	28.05	33.3	83.3
ORF8:L84S	68.1	100	100	99.2	100	100	99.8	97.5	100	100
M:I82S	46.8	5.7	2.7	0.3	14.7	0.84	0.64	2.4	33.3	83.3
S:H655Y	46.8	7.9	23.8	2.5	19.9	0.9	0.8	53.6	88.8	83.3
S:F565L	46.8	5.7	2.7	0.3	14.7	0.84	0.64	2.4	33.3	83.3
S:N501Y	36.1	7.9	21.1	1.4	17.6	0.84	0.8	25.6	33.3	100
S:Y449H	34.04	5.1	2.7	0.3	14.6	0.78	0.64	2.4	33.3	83.3
NSP14:I463V	46.8	5.1	2.3	0.1	14.6	0.84	0.64	2.4	33.3	33.3
NSP6:M86I	93.6	30.4	6.6	8.2	48.8	30.8	1.4	28.05	33.3	83.3
NSP4:A446V	46.8	5.7	2.7	0.3	14.6	0.84	0.8	2.4	33.3	83.3
NSP1:V86F	44.6	5.1	6.6	0.9	16.4	0.9	0.48	26.8	88.8	83.3

Figure 2. The frequency of the shared nonsynonymous mutations in samples of the 19B (S) clade A.29 lineage with different patterns among countries over the whole genome include: N:S202N, N:D22Y, ORF8:E92K, ORF8:L84S, M:I82S, S:H655Y, S:F565L, S:N501Y, S:Y449H, NSP14:I463V, NSP6:M86I, NSP4:A446V, and NSP1:V86F.

Deletions and deletion–frameshifts were concentrated in the NTD region of the S1 subunit of the spike protein, whereas nonsynonymous mutations were scattered in different ORFs of the viral genome including the spike protein itself. In the 3′ UTR of the genome, there were extra-genic mutations at coronavirus 3′ stem-loop II-like motif (s2m; located in the region from 29,728 to 29,768) at position 29,742 and 29,739 (Supplementary Figure S3).

Based on the CoVariants website [25], which gives an overview of *SARS-coV2* variants and their mutations, which is supported by data from the GISAID platform, mutations from the above-mentioned samples were found mainly in those variants of concern as nonsynonymous mutations including: S:N501Y dominating 20I (Alpha, V1), 20H (Beta, V2) and 20J (Gamma, V3), S:H655Y dominating 20J (Gamma, V3), 21K (Omicron) and 21L (Omicron), M:I82S dominating 21B (Kappa), S:V143 and S:N211 dominating 21K (Omicron), and finally the extragenic mutation in the 3′ UTR:29,742 in 21A, I and J (Delta) and 21B (Kappa) as the G nucleotide substituted to T, which is a synonymous mutation, but in those samples, it was substituted with an A nucleotide. None of the known variants showed substitution at 3′ UTR: 29,739, which changed from C to T.

Visualization of structural variations in the spike protein were presented in secondary and 3D structures, all in comparison to the reference genome of *SARS-coV2*, Bat-derived virus of Yunnan, and Pangolin. There were small variations from the reference genome in the minimal free energy and ensemble diversity in secondary structures, and multiple areas of deletions were demonstrated in the 3D structures (Supplementary Figures S4–S10).

Those samples were scattered across the pandemic years. The majority appeared in 2021 with a peak in March, with the exception of samples from Kombo city in Gambia and Kassala city in Sudan, whose samples were reported in August and December 2020, respectively (Figure 3).

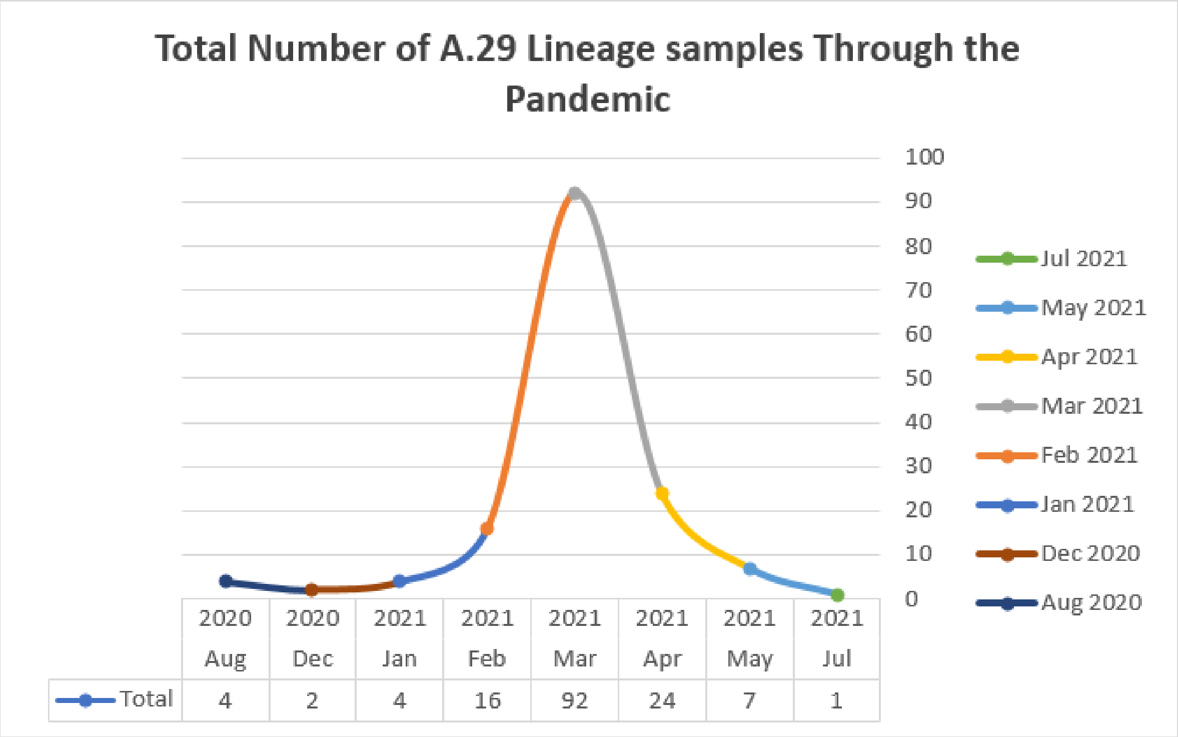


Figure 3. Illustrates the increase in the number of A.29 lineage samples across countries over two years of the pandemic to show a higher peak in March 2021. The first sample was reported in Kombo Gambia on August 2020 *n* = four, followed by Sudan in December 2020 in Kassala *n* = two.

4. Discussion

For a virus such as *SARS-coV2*, to jump the species barriers, it needs to pass through stages of natural selection and incur more genomic changes to increase its affinity to the human genomic counterparts including the ACE2. This was applied to the structure of the spike protein of the Bat-derived virus of Yunnan and Pangolin samples, which showed multiple deletions and deletion–frameshifts not represented in the reference genome of *SARS-coV2* [26]. Viruses are able to evolve within the host itself, resulting in variants and strains that could infect and transmit better [27]. Once this happens in a certain population, a variant will form clusters and allow for more transmission and escaping of the immune surveillances of the host. Still, in continents such as in Africa, where highly diverse populations exist [28–30], the dynamics and timeline for the virus evolution and adaptation is not well understood due to the lack of adequate sampling. Taking Sudan as an example in which the 19B (S) clade was dominating and represented with different patterns of mutations in high frequency (49%) could be a clue for the virus track to adapt with population variability in order to increase its survival [26,31]. An example of some mutations found is the H655Y substitution, which enhances spike cleavage and viral growth [32]. N501Y and Y449H substitutions are believed to increase the virus transmissibility and fitness, although it has been shown that the Y449H substitution alone leads to a decrease in the affinity of binding to the ACE2 receptor, but when it came along with the N501Y, it was speculated that they will enhance binding affinity and escaping, based on epistatic shifting and modulation [33]. In addition, there is the deletion and deletion–frameshift in the NTD region, which plays an important role in conformation of the spike protein structure, binding to ACE2 and immune escaping, although it shows a higher rate of conservation [34]. These mutational patterns are scattered in different variants and strains of concern and interest, rather than being grouped in one variant including the 19B (S) strain. Patterns of similar mutations were shown in a few other countries including Gambia, India, Jordan, the United Kingdom, Germany, Belgium, the United States of America, Canada, and Australia with low frequencies, possibly due to migration, since the first samples were detected in Gambia in August followed by Sudan in December in the year 2020 based on collection dates reported in the GISAID platform.

Deletions could be either deleterious, in which the virus starts reverting to its ancestral state, or compensatory with changes that could mask the effect of the previous mutations. Such a mechanism will guarantee recovery of fitness, all depending on the effective population size of the organism. The mechanism is common in the RNA viruses [35]. For the virus to have a deletion–frameshift followed by deletion means that the virus is going through an intense selective pressure and through a survival strategy to overcome variation, which has been shown in the Omicron and Delta variants that manifest such a breadth of deletions in the S1 subunit [36]. By comparing *SARS-coV2* to the *SARS-coV2-related coronavirus* (*SC2r-CoV*) lineages from bats and pangolins, researchers found out that the NTD area's gained indels during viral transmission across animals are the same as those reported during human transmissions [8]. According to one theory, increasing global human population immunity to natural *SARS-coV2* infection is driving such a large selection pressure on the virus genome, which may select for convergent deletions at NTD to avoid being neutralized by neutralizing antibodies against NTD [37]. Another study demonstrated that rather than working in an antibody avoidance mechanism, NTD deletion in certain locations boosts viral infectivity by boosting the incorporation of cleaved S into virion [14].

As an RNA virus, *SARS-coV2* could affect the host as a quasispecies population; hence, for indels mutation, it can be maintained, which will not be the case if the infected host came as one clonal population with the same genome sequences [38–40]. In addition, these quasispecies are maintained after selection and bottleneck events [41]. Having an adaptive landscape will allow overcoming the high functional constrains of the virus for more transmissibility and infectivity [35]. However, this landscape should not be from the pathogen side alone. A hypothesis called the Red Queen mentioned that the host and pathogen reciprocally struggle to maintain constant levels of fitness [42], and the virus copy number and mutational rate are highly determined by the variability within the host [36,43,44]. Many studies investigated the coadaptation of different ancient viruses; taking Adenovirus lineages as an example, it appeared to have speciated and coevolved with different vertebrate families, and host shifts to new taxa were accompanied by changes in genome content and those non-coevolved showed more severe pathology [43]. Others studied how this relationship can affect different part of the human genome: for example, the *hepatitis C virus* is able to affect reversibly the diversity of mitochondrial DNA, in which it can be used as a biomarker to investigate the stage of infection [44]. Other studies mentioned that having some genetic markers represented in certain population has a large effect on the disease progression: for example, the representation of the Y chromosome ancestry marker *R1b1b2* in certain continents including in Africa with low frequencies is associated with a lower mortality rate from COVID-19 infection [45]. Interestingly, African viral sequences appeared to be ancestral to most of the lineages of the 19B (S) clade and for two variants of concern including Beta and Omicron, making the continent together with Asia the site for the emergence of these variants and in tally with the hypothesis that places of high effective population size are probably the main sources of novel variants. Tajima's Neutrality test D value is consistent with rapid expansion and directional selection, which are major features of the pandemic. In numerous studies, it was discovered that Africans have higher nucleotide and haplotype diversity than non-Africans [5,28–30]. In another study, ACE2 receptors are presented with some different nonsynonymous mutations in African populations which were identified as signatures of selection affecting variation at regulatory regions related to ACE2 expression [46] along with multiple risk factors such as HIV coinfection, tuberculosis, anemia and population age [47–49].

The dearth of viral sequences from the African continent in this manuscript might be related to the paucity of sampling due to poorly resourced and ailing health systems or perhaps to the relatively low disease burden. Circumventing such hurdles is essential for a lucid understanding of the epidemiology of the pandemic and in verifying concepts such as the above [50,51].

5. Conclusions

Finally, for the virus to evolve within complex populations characterized by a higher effective population size, it may be pivotal for the virus to acquire distinct patterns of mutations that ensure its survival and fitness, thus becoming a source of novel variants of concern, as was observed early on in the 19 (S) clade of *SARS-coV2*, specifically in the A.29 lineage, which was first dominating the African continent and revealed distinct mutational variations that were scattered in other variants of concern. These variations raise many questions related to the function of the mutations and structural motif within the host that may lead to better viral fitness or transmissibility.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/microorganisms10102017/s1, Table S1: The distribution in the 19B (S) lineages in the different contents, in which the A.23 and A.27 were dominating Africa with approximate percentages of 39.67% and 14.996% respectively, while the A Lineage dominated Asia with 61.1%. The A.2 found to be common in Europe, North America, South America and Oceania with approximate percentages of 37.23%, 40.623%, 76.23% and 56.93% respectively Figure S1: The frequency of the shared synonymous mutations in samples of 19B (S) clade A.29 lineage with different pattern among countries over the whole genome include NSP16:A199A, NSP16:K137K, NSP9:N95N, NSP4:S76S, NSP4 - Figure S2: The frequency of the shared Deletion and Deletion-Frameshifts in samples of 19B (S) clade A.29 lineage with different pattern among countries over the whole genome include S: F140, S: V143, S: N211. Figure S3: The frequency of shared Extragenic mutations in the 3′UTR region in samples of 19B (S) clade A.29 lineage with different pattern among countries over the whole genome. Figure S4: A. Secondary structure, for the sample hCoV-19_Wuhan_HB-WH5-234_2020_EPI_ISL_455393_2020-03-18 (Reference: accession number NC 045512.2S), Minimum free energy of −1072.20 kcal/mol, the ensemble diversity is 898.62. B. 3D structure of the spike protein using the template 7cn8.1.A, seq identity: 91.68% with the description of Glycoprotein Cryo-EM structure of PCoV_GX spike glycoprotein. C. show no deletion in the spike protein. Figure S5: A.Secondary structure for the sample hCoV-19_bat_Yunnan_RsYN03_2019_EPI_ISL_1699443_2019-10-22, Minimum free energy of −1079.70 kcal/mol, and the ensemble diversity is 826.31. B. 3D structure of the spike protein using the template 7sbo.1.A, seq identity: 78.48% with the description of Glycoprotein 1 RBD-up 2 of pre-fusion SARS-CoV-2 Delta variant spike protein. C. shows multiple deletions in the spike protein marked in blue. Figure S6: A.Secondary structure for the sample hCoV-19_pangolin_Guangdong_A22-2_2019_EPI_ISL_471467_2019, Minimum free energy of −1055.30 kcal/mol, and the ensemble diversity is 705.75. B. 3D structure of the spike protein using the template 7cn8.1.A, seq identity: 90.73% with the description of Glycoprotein Cryo-EM structure of PCoV_GX spike glycoprotein. C. shows multiple deletions in the spike protein marked in blue. Figure S7: A.Secondary structure for the sample hCoV-19_England_NORT-1BC5CAB_2021_EPI_ISL_2434626_2021-05-09, Minimal free energy of −1071.87 kcal/mol, and the ensemble diversity is 909.34. B. 3D structure of the spike protein using the template 7cn8.1.A, seq identity: 92.18% with the description of Glycoprotein Cryo-EM structure of PCoV_GX spike glycoprotein. C. shows multiple deletions in the spike protein marked in blue. Figure S8: A.Secondary structure for the sample hCoV-19_Gambia_22178_2020_EPI_ISL_1731548_2020-08-29, Minimum free energy of −1064.20 kcal/mol, and the ensemble diversity is 813.74. B. 3D structure of the spike protein using the template 7cn8.1.A, seq identity: 92.20% with the description of Glycoprotein Cryo-EM structure of PCoV_GX spike glycoprotein. C. shows multiple deletions in the spike protein marked in blue. Figure S9: A.Secondary structure for the sample hCoV-19_Sudan_N6392_2021_EPI_ISL_4739539_2021-07-03, Minimum free energy of −1075.40 kcal/mol, and the ensemble diversity is 848.83. B. 3D structure of the spike protein using the template 7cn8.1.A, seq identity: 92.08% with the description of Glycoprotein Cryo-EM structure of PCoV_GX spike glycoprotein. C. shows multiple deletions in the spike protein marked in blue. Figure S10: A.Secondary structure for the sample hCoV-19_Sudan_N6418_2021_EPI_ISL_4739626_2021-03-07, Minimum free energy of −1023.60 kcal/mol, and the ensemble diversity is 681.89. B. 3D structure of the spike protein using the template 7krs.1.A, seq identity: 100.00% with the description of Spike glycoprotein Structural impact on SARS-CoV-2 spike protein by D614G substitution. C. shows multiple deletions in the spike protein marked in blue.

Author Contributions: Conceptualization, B.K.H.; Data curation, B.K.H. and O.M.I.; Formal analysis, B.K.H., O.M.I. and L.A.M.A.; Funding acquisition, M.E.I.; Methodology, B.K.H. and O.M.I.; Project administration, B.K.H., M.H.A. and M.E.I.; Software, B.K.H.; Supervision, M.E.I.; Validation, B.K.H. and M.H.A.; Visualization, B.K.H., M.F.A. and S.A.E.A.; Writing—original draft, B.K.H.; Writing—review and editing, B.K.H., M.H.A. and M.E.I. All authors have read and agreed to the published version of the manuscript.

Funding: UNESCO office Khartoum, Khartoum, Sudan funded this research. Contract N°: 4500429831.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data used in this study are available at the Global Initiative for Sharing All Influenza Data (GISAID EpiCoV database) portal [17].

Acknowledgments: The completion of this work would not have been possible without the participation and assistance of a large number of people. Their contributions are gratefully acknowledged and appreciated. However, we would like to express our gratitude to: Shahd Osman, Mudathir Salih, Muiz M. Hamza, Ahmed Berir, and Mahmoud Koko for their kind assistance.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gorbalenya, A.E.; Baker, S.C.; Baric, R.S.; de Groot, R.J.; Drosten, C.; Gulyaeva, A.A.; Haagmans, B.L.; Lauber, C.; Leontovich, A.M.; Neuman, B.W.; et al. The Species Severe Acute Respiratory Syndrome-Related Coronavirus: Classifying 2019-NCoV and Naming It SARS-CoV-2. *Nat. Microbiol.* **2020**, *5*, 536–544. [CrossRef]
2. Li, L.L.; Wang, J.L.; Ma, X.H.; Sun, X.M.; Li, J.S.; Yang, X.F.; Shi, W.F.; Duan, Z.J. A Novel SARS-CoV-2 Related Coronavirus with Complex Recombination Isolated from Bats in Yunnan Province, China. *Emerg. Microbes Infect.* **2021**, *10*, 1683–1690. [CrossRef] [PubMed]

3. Gutierrez, M.C.; Brisse, S.; Brosch, R.; Fabre, M.; Omaïs, B.; Marmiesse, M.; Supply, P.; Vincent, V. Ancient Origin and Gene Mosaicism of the Progenitor of Mycobacterium Tuberculosis. *PLoS Pathog.* **2005**, *1*, e5. [CrossRef]
4. Wiens, K.E.; Woyczynski, L.P.; Ledesma, J.R.; Ross, J.M.; Zenteno-Cuevas, R.; Goodridge, A.; Ullah, I.; Mathema, B.; Djoba Siawaya, J.F.; Biehl, M.H.; et al. Global Variation in Bacterial Strains That Cause Tuberculosis Disease: A Systematic Review and Meta-Analysis. *BMC Med.* **2018**, *16*, 196. [CrossRef]
5. Ibrahim, M.E.; Barker, D.C. The Origin and Evolution of the Leishmania Donovani Complex as Inferred from a Mitochondrial Cytochrome Oxidase II Gene Sequence. *Infect. Genet. Evol.* **2001**, *1*, 61–68. [CrossRef]
6. Falush, D.; Wirth, T.; Linz, B.; Pritchard, J.K.; Stephens, M.; Kidd, M.; Blaser, M.J.; Graham, D.Y.; Vacher, S.; Perez-Perez, G.I.; et al. Traces of Human Migrations in Helicobacter Pylori Populations. *Science* **2003**, *299*, 1582–1585. [CrossRef]
7. CDC. SARS-CoV-2 Variant Classifications and Definitions. *CDC*. 13 July 2021; pp. 1–12. Available online: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html#Interest (accessed on 15 August 2022).
8. Letko, M.; Marzi, A.; Munster, V. Functional Assessment of Cell Entry and Receptor Usage for SARS-CoV-2 and Other Lineage B Betacoronaviruses. *Nat. Microbiol.* **2020**, *5*, 562–569. [CrossRef] [PubMed]
9. Pereson, M.J.; Flichman, D.M.; Martínez, A.P.; Baré, P.; Garcia, G.H.; Di Lello, F.A. Evolutionary Analysis of SARS-CoV-2 Spike Protein for Its Different Clades. *J. Med. Virol.* **2021**, *93*, 3000–3006. [CrossRef] [PubMed]
10. Hussman, J.P. Cellular and Molecular Pathways of COVID-19 and Potential Points of Therapeutic Intervention. *Front. Pharmacol.* **2020**, *11*, 1169. [CrossRef]
11. Alejandra Tortorici, M.; Walls, A.C.; Lang, Y.; Wang, C.; Li, Z.; Koerhuis, D.; Boons, G.J.; Bosch, B.J.; Rey, F.A.; de Groot, R.J.; et al. Structural Basis for Human Coronavirus Attachment to Sialic Acid Receptors. *Nat. Struct. Mol. Biol.* **2019**, *26*, 481–489. [CrossRef]
12. Tegally, H.; Wilkinson, E.; Giovanetti, M.; Iranzadeh, A.; Fonseca, V.; Giandhari, J.; Doolabh, D.; Pillay, S.; San, E.J.; Msomi, N.; et al. Detection of a SARS-CoV-2 Variant of Concern in South Africa. *Nature* **2021**, *592*, 438–443. [CrossRef] [PubMed]
13. Smith, E.C.; Denison, M.R. Implications of Altered Replication Fidelity on the Evolution and Pathogenesis of Coronaviruses. *Curr. Opin. Virol.* **2012**, *2*, 519–524. [CrossRef] [PubMed]
14. Meng, B.; Kemp, S.A.; Papa, G.; Datir, R.; Ferreira, I.A.T.M.; Marelli, S.; Harvey, W.T.; Lytras, S.; Mohamed, A.; Gallo, G.; et al. Recurrent Emergence of SARS-CoV-2 Spike Deletion H69/V70 and Its Role in the Alpha Variant B.1.1.7. *Cell Rep.* **2021**, *35*, 109292. [CrossRef] [PubMed]
15. Wertheim, J.O.; Leigh Brown, A.J.; Hepler, N.L.; Mehta, S.R.; Richman, D.D.; Smith, D.M.; Kosakovsky Pond, S.L. The Global Transmission Network of HIV-1. *J. Infect. Dis.* **2014**, *209*, 304–313. [CrossRef] [PubMed]
16. Aksamentov, I.; Roemer, C.; Hodcroft, E.; Neher, R. Nextclade: Clade Assignment, Mutation Calling and Quality Control for Viral Genomes. *J. Open Source Softw.* **2021**, *6*, 3773. [CrossRef]
17. GISAID—Gisaid. 2021. Available online: https://gisaid.org/ (accessed on 15 August 2022).
18. Mercatelli, D.; Triboli, L.; Fornasari, E.; Ray, F.; Giorgi, F.M. Coronapp: A Web Application to Annotate and Monitor SARS-CoV-2 Mutations. *J. Med. Virol.* **2021**, *93*, 3238–3245. [CrossRef]
19. Tamura, K.; Stecher, G.; Kumar, S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.* **2021**, *38*, 3022–3027. [CrossRef]
20. Kimura, M. Diffusion Models in Population Genetics. *J. Appl. Probab.* **1964**, *1*, 177–232. [CrossRef]
21. Chudalayandi, S. Bioinformatics_Workbook. 2021. Available online: https://bioinformaticsworkbook.org/about.html#gsc.tab=0 (accessed on 15 August 2022).
22. Format, V.; Format, C.; Converter, I.; Explorer, I.; Svg, A. RNAfold Web Server. *Monatshefte*. 2011, pp. 8–10. Available online: http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi (accessed on 15 August 2022).
23. SIB Swiss Institute of Bioinformatics ExPASy-Translate Tool 2011. Available online: https://www.expasy.org/ (accessed on 15 August 2022).
24. Swiss Institue of Bioinformatics SWISS-MODEL Interactive Workspace. Database 2021. Available online: https://swissmodel.expasy.org/ (accessed on 15 August 2022).
25. CoVariants. Available online: https://covariants.org/ (accessed on 15 August 2022).
26. Andersen, K.G.; Rambaut, A.; Lipkin, W.I.; Holmes, E.C.; Garry, R.F. The Proximal Origin of SARS-CoV-2. *Nat. Med.* **2020**, *26*, 450–452. [CrossRef]
27. Tonkin-Hill, G.; Martincorena, I.; Amato, R.; Lawson, A.R.; Gerstung, M.; Johnston, I.; Jackson, D.K.; Park, N.; Lensing, S.V.; Quail, M.A.; et al. Patterns of Within-Host Genetic Diversity in SARS-CoV-2. *Elife* **2021**, *10*, e66857. [CrossRef]
28. Elhassan, N.; Gebremeskel, E.I.; Elnour, M.A.; Isabirye, D.; Okello, J.; Hussien, A.; Kwiatksowski, D.; Hirbo, J.; Tishkoff, S.; Ibrahim, M.E. The Episode of Genetic Drift Defining the Migration of Humans out of Africa Is Derived from a Large East African Population Size. *PLoS ONE* **2014**, *9*, e97674. [CrossRef] [PubMed]
29. Campbell, M.C.; Tishkoff, S.A. African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annu. Rev. Genomics Hum. Genet.* **2008**, *9*, 403–433. [CrossRef] [PubMed]
30. Sirugo, G.; Hennig, B.J.; Adeyemo, A.A.; Matimba, A.; Newport, M.J.; Ibrahim, M.E.; Ryckman, K.K.; Tacconelli, A.; Mariani-Costantini, R.; Novelli, G.; et al. Genetic Studies of African Populations: An Overview on Disease Susceptibility and Response to Vaccines and Therapeutics. *Hum. Genet.* **2008**, *123*, 557–598. [CrossRef] [PubMed]
31. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.M.; Wang, W.; Song, Z.G.; Hu, Y.; Tao, Z.W.; Tian, J.H.; Pei, Y.Y.; et al. A New Coronavirus Associated with Human Respiratory Disease in China. *Nature* **2020**, *579*, 265–269. [CrossRef] [PubMed]

32. Escalera, A.; Gonzalez-Reiche, A.S.; Aslam, S.; Mena, I.; Laporte, M.; Pearl, R.L.; Fossati, A.; Rathnasinghe, R.; Alshammary, H.; van de Guchte, A.; et al. Mutations in SARS-CoV-2 Variants of Concern Link to Increased Spike Cleavage and Virus Transmission. *Cell Host Microbe* **2022**, *30*, 373–387.e7. [CrossRef]
33. Starr, T.N.; Greaney, A.J.; Hannon, W.W.; Loes, A.N.; Hauser, K.; Dillen, J.R.; Ferri, E.; Farrell, A.G.; Dadonaite, B.; McCallum, M.; et al. Shifting Mutational Constraints in the SARS-CoV-2 Receptor-Binding Domain during Viral Evolution. *Science* **2022**, *424*, eabo7896. [CrossRef]
34. Klinakis, A.; Cournia, Z.; Rampias, T. N-Terminal Domain Mutations of the Spike Protein Are Structurally Implicated in Epitope Recognition in Emerging SARS-CoV-2 Strains. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 5556–5567. [CrossRef]
35. Burch, C.L.; Chao, L. Evolution by Small Steps and Rugged Landscapes in the RNA Virus Φ6. *Genetics* **1999**, *151*, 921–927. [CrossRef]
36. Xue, S.-A.; Jones, M.D.; Lu, Q.-L.; Middeldorp, J.M.; Griffin, B.E. Genetic Diversity: Frameshift Mechanisms Alter Coding of a Gene (Epstein-Barr Virus LF3 Gene) That Contains Multiple 102-Base-Pair Direct Sequence Repeats. *Mol. Cell. Biol.* **2003**, *23*, 2192–2201. [CrossRef]
37. Mathur, P.; Goyal, P.; Verma, G.; Yadav, P. Entropy Based Analysis of SARS-CoV-2 Spread in India Using Informative Subtype Markers. *Sci. Rep.* **2021**, *11*, 15972. [CrossRef]
38. Sanjuán, R.; Thoulouze, M.I. Why Viruses Sometimes Disperse in Groups. *Virus Evol.* **2019**, *5*, vez014. [CrossRef] [PubMed]
39. Domingo, E.; Sheldon, J.; Perales, C. Viral Quasispecies Evolution. *Microbiol. Mol. Biol. Rev.* **2012**, *76*, 159–216. [CrossRef] [PubMed]
40. Aaskov, J.; Buzacott, K.; Thu, H.M.; Lowry, K.; Holmes, E.C. Long-Term Transmission of Defective RNA Viruses in Humans and Aedes Mosquitoes. *Science* **2006**, *311*, 236–238. [CrossRef]
41. Park, D.; Hahn, Y. Rapid Protein Sequence Evolution via Compensatory Frameshift Is Widespread in RNA Virus Genomes. *BMC Bioinform.* **2021**, *22*, 251. [CrossRef] [PubMed]
42. McLaughlin, R.N.; Malik, H.S. Genetic Conflicts: The Usual Suspects and Beyond. *J. Exp. Biol.* **2017**, *220*, 6–17. [CrossRef]
43. Kaján, G.L.; Doszpoly, A.; Tarján, Z.L.; Vidovszky, M.Z.; Papp, T. Virus–Host Coevolution with a Focus on Animal and Human DNA Viruses. *J. Mol. Evol.* **2020**, *88*, 41–56. [CrossRef]
44. Campo, D.S.; Roh, H.J.; Pearlman, B.L.; Fierer, D.S.; Ramachandran, S.; Vaughan, G.; Hinds, A.; Dimitrova, Z.; Skums, P.; Khudyakov, Y. Increased Mitochondrial Genetic Diversity in Persons Infected With Hepatitis C Virus. *Cmgh* **2016**, *2*, 676–684. [CrossRef]
45. Ibrahim, M.; Salih, A. The Y Chromosome Ancestry Marker R1b1b2: A Surrogate of the SARS-CoV-2 Population Affinity. *Hum. Genome Var.* **2021**, *8*, 11. [CrossRef]
46. Zhang, C.; Verma, A.; Feng, Y.; Melo, M.C.R.; McQuillan, M.; Hansen, M.; Lucas, A.; Park, J.; Ranciaro, A.; Thompson, S.; et al. Impact of Natural Selection on Global Patterns of Genetic Variation and Association with Clinical Phenotypes at Genes Involved in SARS-CoV-2 Infection. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2123000119. [CrossRef]
47. Cele, S.; Karim, F.; Lustig, G.; San, J.E.; Hermanus, T.; Tegally, H.; Snyman, J.; Moyo-Gwete, T.; Wilkinson, E.; Bernstein, M.; et al. SARS-CoV-2 Prolonged Infection during Advanced HIV Disease Evolves Extensive Immune Escape. *Cell Host Microbe* **2022**, *30*, 154–162.e5. [CrossRef]
48. Maponga, T.G.; Jeffries, M.; Tegally, H.; Sutherland, A.; Wilkinson, E.; Lessells, R.J.; Msomi, N.; van Zyl, G.; de Oliveira, T.; Preiser, W. Persistent Severe Acute Respiratory Syndrome Coronavirus 2 Infection With Accumulation of Mutations in a Patient with Poorly Controlled Human Immunodeficiency Virus Infection. *Clin. Infect. Dis.* **2022**, ciac548. [CrossRef] [PubMed]
49. Gesesew, H.A.; Koye, D.N.; Fetene, D.M.; Woldegiorgis, M.; Kinfu, Y.; Geleto, A.B.; Melaku, Y.A.; Mohammed, H.; Alene, K.A.; Awoke, M.A.; et al. Risk Factors for COVID-19 Infection, Disease Severity and Related Deaths in Africa: A Systematic Review. *BMJ Open* **2021**, *11*, e044618. [CrossRef] [PubMed]
50. Tegally, H.; San, J.E.; Cotten, M.; Tegomoh, B.; Mboowa, G.; Martin, D.P.; Baxter, C.; Moir, M.; Lambisia, A.; Diallo, A.; et al. The Evolving SARS-CoV-2 Epidemic in Africa: Insights from Rapidly Expanding Genomic Surveillance. *medRxiv* **2022**. [CrossRef] [PubMed]
51. Aziz, R.K.; Giri, B.; Guzmán, M.G.; Jiang, H.; Abbès, S. Establishment of Regional Genomic Surveillance Networks in Lower and Lower-Middle Income Countries. 2022; *preprints*. [CrossRef]