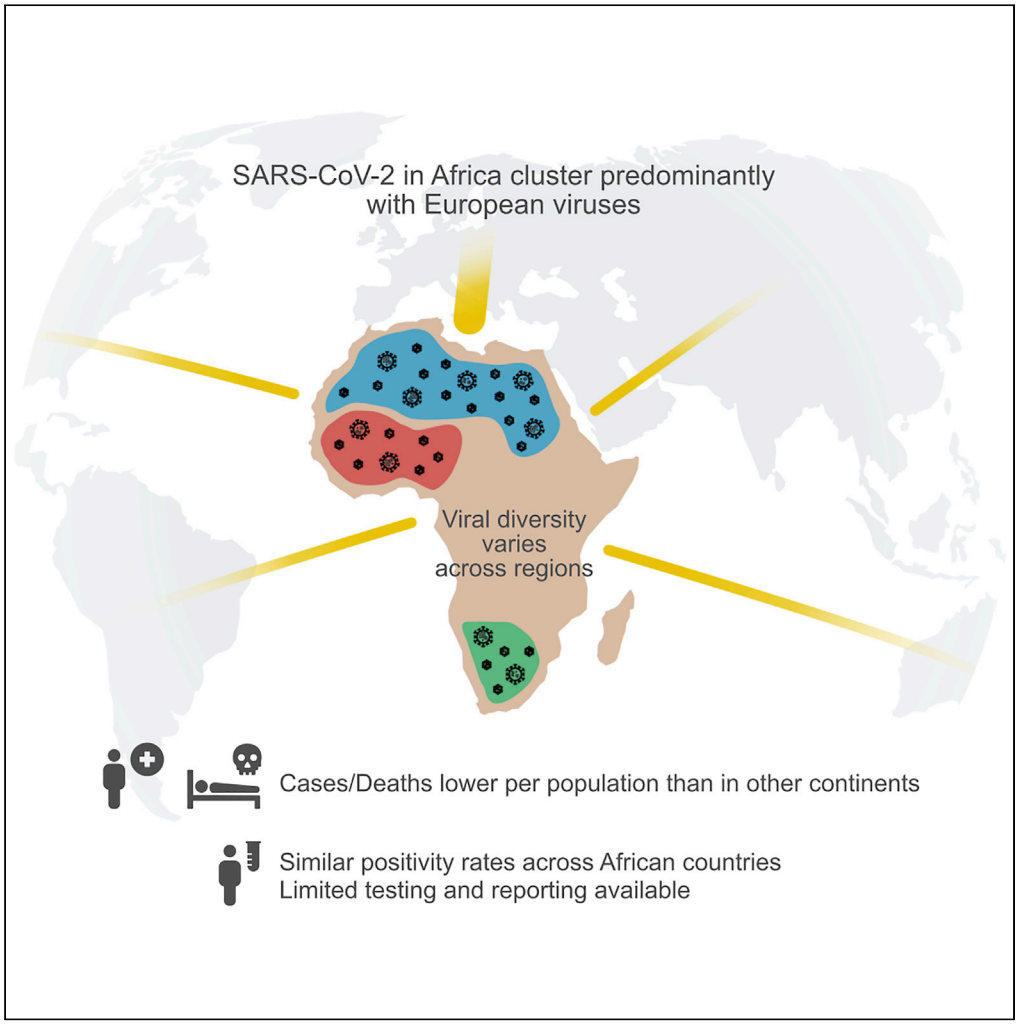
iScience

CellPress
OPEN ACCESS

Article

Epidemiology and genetic diversity of SARS-CoV-2 lineages circulating in Africa



Olayinka Sunday Okoh, Nicholas Israel Nii-Trebi, Abdulrokeeb Jakkari, ..., David J. Spiro, Elijah Kolawole Oladipo, Nídia S. Trovão

nidiastrovao@gmail.com

Highlights

SARS-CoV-2 viruses from Africa cluster predominantly with European strains

Lower viral diversity observed in Africa is likely due to genomic under-surveillance

Number of cases, deaths, and testing show substantial heterogeneity across Africa

Two motifs could function as integrin-binding sites and N-glycosylation domains

Okoh et al., iScience 25, 103880
March 18, 2022
https://doi.org/10.1016/j.isci.2022.103880

iScience

CellPress
OPEN ACCESS

Article

Epidemiology and genetic diversity of SARS-CoV-2 lineages circulating in Africa

Olayinka Sunday Okoh,[1,15] Nicholas Israel Nii-Trebi,[2,15] Abdulrokeeb Jakkari,[3] Tosin Titus Olaniran,[4,5] Tosin Yetunde Senbadejo,[6] Anna Aba Kafintu-kwashie,[7] Emmanuel Oluwatobi Dairo,[5,8] Tajudeen Oladunni Ganiyu,[6] Ifiokakaninyene Ekpo Akaninyene,[4,5] Louis Odinakaose Ezediuno,[9] Idowu Jesulayomi Adeosun,[10,11] Michael Asebake Ockiya,[12] Esther Moradeyo Jimah,[5,13] David J. Spiro,[14] Elijah Kolawole Oladipo,[5,10] and Nídia S. Trovão[14,16,*]

SUMMARY

There is a dearth of information on COVID-19 disease dynamics in Africa. To fill this gap, we investigated the epidemiology and genetic diversity of SARS-CoV-2 lineages circulating in the continent. We retrieved 5229 complete genomes collected in 33 African countries from the GISAID database. We investigated the circulating diversity, reconstructed the viral evolutionary divergence and history, and studied the case and death trends in the continent. Almost a fifth (144/782, 18.4%) of Pango lineages found worldwide circulated in Africa, with five different lineages dominating over time. Phylogenetic analysis revealed that African viruses cluster more closely with those from Europe. We also identified two motifs that could function as integrin-binding sites and N-glycosylation domains. These results shed light on the epidemiological and evolutionary dynamics of the circulating viral diversity in Africa. They also emphasize the need to expand surveillance efforts in Africa to help inform and implement better public health measures.

INTRODUCTION

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that was first reported in Wuhan, China in December 2019, emerged as a novel virus causing a cluster of unusual pneumonia cases. Its outbreak soon became a worldwide pandemic that resulted in a global public health emergency (Guo et al., 2020). Coronavirus disease 2019 (COVID-19) caused by SARS-CoV-2 has affected all seven continents, with Africa currently being the least stricken by the pandemic (Lone and Ahmad, 2020). The first case confirmed in Africa was in Egypt on February 14, 2020, followed by Algeria on February 25, 2020. The first case reported in sub-Saharan Africa was confirmed in Nigeria on February 27, 2020 (NCDC, 2020). The first cases in other African countries were recorded in March 2020 (African Centres for Disease Control, 2020), including in Ghana on March 12, 2020. By September 6, 2021, about 7.9 million confirmed cases and more than 199 thousand deaths were reported by Africa CDC, as part of the more than 219 million confirmed cases and more than 4.5 million deaths reported globally by the World Health Organization (WHO).

SARS-CoV-2 is the seventh coronavirus known to infect humans (Corman et al., 2020), and it is the third novel coronavirus known to have caused large-scale outbreaks in the 21st century. The first was the SARS-CoV in 2003 that also emerged in China (Rosling and Rosling, 2003; Alanagreh et al., 2020) followed by the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) (Alanagreh et al., 2020; Cui et al., 2019; Zaki et al., 2012) that emerged in Saudi Arabia in 2012. Unlike SARS-CoV and MERS-CoV that cause severe disease in humans (Kumar et al., 2020), SARS-CoV-2 is more infectious but appears to have a lower case fatality rate (CFR) (Zhang and Holmes, 2020).

The genome of SARS-CoV-2 is a positive-sense single-stranded RNA (+ssRNA) of approximately 29.9 kilobases (29,891 nucleotides) encoding 9,860 amino acids (Sapkota, 2020). The virus is classified among the Betacoronaviruses (β-CoVs) group under the *Coronavirinae* subfamily of the *Coronaviridae* family. The β-CoVs genus is known to infect humans, bats, and other wild animals (Chen et al., 2020). Genome replication produces two large ORFs that are translated into polyproteins processed post-translationally to

[1]Department of Chemical Sciences, Anchor University, Lagos, Nigeria

[2]Department of Medical Laboratory Sciences, School of Biomedical and Allied Health Sciences, University of Ghana, Accra, Ghana

[3]Department of Microbiology, Faculty of Science, Lagos State University, Ojo, Lagos, Nigeria

[4]Department of Pure and Applied Biology (Microbiology Unit), Ladoke Akintola University of Technology, Ogbomoso, Nigeria

[5]Helix Biogen Institute, Ogbomoso, Nigeria

[6]Department of Biological Sciences, College of Natural and Applied Sciences, Fountain University, Osogbo, Nigeria

[7]Department of Medical Microbiology, Clinical Virology Unit, University of Ghana Medical School, Accra, Ghana

[8]Department of Virology, College of Medicine, University of Ibadan, Ibadan, Nigeria

[9]Department of Microbiology, Faculty of Life Sciences, University of Ilorin,1515 P.M.B, Ilorin, Nigeria

[10]Department of Microbiology, Laboratory of Molecular Biology, Immunology and Bioinformatics, Adeleke University, Ede, Osun, Nigeria

[11]Division of Microbiology, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Private

Continued

produce 16 proteins comprising four structural proteins, namely envelope (E), spike (S), membrane (M), nucleocapsid (N), and at least nine accessory proteins, some of which are unique to SARS-CoV-2 and others conserved among coronaviruses (Wu et al., 2020). These 16 proteins play a crucial role in the viral RNA synthesis and immune evasion (Snijder et al., 2020; Shereen et al., 2020). Of the four structural proteins, the S-protein, which is made up of S1 and S2 domains, is known to play a unique role in SARS-CoV-2 replication. The protein functions during host cell attachment and entry by primarily mediating binding to the extracellular domains of its receptor, the angiotensin-converting enzyme 2 (ACE2), a transmembrane protein that is also used by SARS-CoV for cell entry (Wan et al., 2020). SARS-CoV-2 binding to ACE2 and fusion with cellular membrane are facilitated by the S1 receptor-binding domain (RBD) and the S2 subunit, respectively. This unique function makes the spike glycoprotein a target for the development of antibodies, therapeutics, and vaccines. Therefore, the mutational patterns in the S-protein and their circulation trends warrant surveillance for effective interventions against the ongoing COVID-19 pandemic.

Protein motifs are small regions of amino acid sequences that facilitate the function of the protein and protein-protein interactions (PPIs). They mediate interactions with cellular proteins and molecular processes within the host cells (Sobhy, 2016). Infection by the virus involves a large number of PPIs between the infective virus and the target host cell (Alguwaizani et al., 2018). Motifs are nucleotide or amino acid sequences that are significant in the genome structure formation, function, and conserved regions in protein molecules. The conserved amino acid sequence may be responsible for protein-substrate binding, determining the active domain for enzymatic cleavage, binding to transcription factors, and the plasticity of protein, and thus repeated motifs are an essential evolutionary mechanism. Hence, identifying motifs and their repeated patterns is important in determining the binding domain of SARS-CoV-2 and to elucidate the evolutionary relationships among sequences (Luo and Nijveen, 2014). Hence, studying and understanding the functional motifs and repeat patterns of SARS-CoV-2 may aid in the prediction of viral protein characteristics, virus-host protein interactions or other putative roles.

RNA viruses, including SARS-CoV-2, commonly generate and accumulate mutations in their genomes during viral replication. In humans, immunological pressure facilitates the accumulation and fixation of mutations as the epidemic persists. Mutations in the S protein constitute a major cause of public health concern, as they have the potential to alter the viral tropism and thereby potentially confer adaptation to new tissues and hosts, influence transmissibility and clinical outcomes, and/or confer resistance to neutralizing antibodies and therapeutics (Sui et al., 2008; Wibmer et al., 2021). In February 2020, a non-synonymous mutation was detected in the S-protein of SARS-CoV-2 viruses sampled from individuals in China and Europe. The mutation caused an amino acid change from aspartic acid to glycine at position 614 (D614G). Experimental and clinical findings associated the G614 variant with a selective advantage over the D614 virus, resulting in higher viral loads and increased infectivity (Korber et al., 2020), but not necessarily increasing the mortality rate (Plante et al., 2021). Between late summer and early autumn of 2020, several variants of the SARS-CoV-2 virus emerged. Variant 20B/501Y.V1 202012/01 classified as Pango lineage B.1.1.7, was identified in the United Kingdom (UK) and designated the Alpha Variant of Concern (VOC) (Volz et al., 2021). This variant emerged with an unusually large number of mutations and has since been detected in numerous countries around the globe, including several in Africa. Almost simultaneously, the independent emergence of the VOC 20C/501Y.V2 belonging to Pango lineage B.1.351, or Beta VOC was detected in South Africa (Tegally et al., 2021). Cases attributed to this variant have since been detected outside of South Africa (Volz et al., 2021), disseminating northwards in Africa. Early in 2021, the VOC 20J/501Y.V3 classified as Pango lineage P.1, or Gamma VOC was first identified in Brazil (Faria et al., 2021) and rapidly spread throughout the Americas, Europe, and Oceania. Despite the independent emergence of the 20B/501Y.V1, 20C/501Y.V2 and 20J/501Y.V3, they share a few common mutations. Because most of the current SARS-CoV-2 immunotherapeutic strategies target the RBD of the S-protein to prevent the binding of SARS-CoV-2 with ACE2 (Chan et al., 2020), alterations in the S-protein sequence could potentially affect the efficacy of immune-based therapeutic agents (Wibmer et al., 2021), making surveillance of spike mutations imperative to aid in the development of effective pharmaceutical interventions.

As of January 7, 2021, nearly one year since the first case was reported in Africa, a total of 5229 SARS-CoV-2 complete genome sequences from 33 African countries had been deposited in public sequence databases including the Global Initiative on Sharing All Influenza Data (GISAID) (Elbe and Buckland-Merrett, 2017) (Shu and Mccauley, 2017), which can be studied to better understand the ongoing molecular epidemiology of SARS-CoV-2 in the continent (Oladipo et al., 2020). Initial genome sequence analysis suggested the

Bag X20, Hatfield Pretoria 0028, South Africa

[12]Department of Animal Science, Niger Delta University, Wilberforce Island, Bayelsa, Nigeria

[13]Department of Medical Microbiology and Parasitology, University of Ilorin 1515, P.M.B, Ilorin, Nigeria

[14]Fogarty International Center, National Institutes of Health, Bethesda, MD, USA

[15]These authors are contributed equally

[16]Lead contact

*Correspondence: nidiastrovao@gmail.com

https://doi.org/10.1016/j.isci.2022.103880

importation of multiple SARS-CoV-2 strains, mainly of European origin and partly from China (Tessema et al., 2020).

Knowledge of the evolutionary dynamics underlying the viral genome variation allows tracing the ongoing outbreak and informs the development and deployment of diagnostic tests (Wang et al., 2020a), as well as effective antiviral and vaccination strategies (Avise, 2000). For example, a recent genome-wide association study on SARS-CoV-2 genomes found variations at the genomic position 11,083 within the coding region of non-structural protein six to be associated with COVID-19 severity. The study showed that the G11083 variant was more commonly found in symptomatic cases, while the T11083 variant appeared to be more frequently associated with asymptomatic infections (Aiewsakun et al., 2020). Toyoshima and colleagues (2020) also performed a comprehensive investigation of 12,343 SARS-CoV-2 genome sequences isolated from individuals in six geographic areas and found that ORF1ab L4715 and S protein G614 variants showed significant positive correlations with fatality rates, which supports the finding suggesting that SARS-CoV-2 mutations might affect the susceptibility to SARS-CoV-2 infection or severity of COVID-19 (Toyoshima et al., 2020). It is to be noted that most sequences from Africa included in the genome variation analysis described above were mostly from North African countries including Egypt, but not those of sub-Saharan Africa.

The epidemiology of SARS-CoV-2 in the African continent calls for a comprehensive study of the genomic and evolutionary patterns of this virus. Comparative analysis of viral genome sequences represents a very useful approach to provide insight into pathogen emergence and evolution. This study, therefore, pursues an in-depth investigation into the epidemiology, evolution, and molecular motifs of SARS-CoV-2 in Africa to shed light on the pandemic dynamics, and aid in informing the development and implementation of control measures in the African continent.

RESULTS

Epidemiological trends of SARS-CoV-2 in Africa

Focusing only on sequence data will not give a true representation of the disease dynamics of SARS-CoV-2 in Africa, as only a fraction of the cases in Africa are sequenced. Therefore, we analyzed reported COVID-19 cases using data from OurWorldInData.org downloaded on January 8, 2021. As presented in Figure S1, North America has the highest number of COVID-19 cases (n = 25 million), followed by Europe (n = 19 million) and Asia (n = 18 million). Oceania has the least number of cases reported of the six continents (n = 20,575). The absolute number of cases in Africa (n = 2 million) and South America (n = 6 million) are a very small fraction of the cases in Asia, North America, and Europe. The global average of COVID-19 cases per 100,000 population (hereafter referred to as cases/pop) is 895, represented by the red line in Figure 1. We observed that the average number of COVID-19 cases per 100,000 persons in Oceania, Africa, and Asia are all below the global average. Considering the absolute cases of COVID-19, Asia is more affected than South America. However, taking population into consideration reveals that South America (1,287 cases/pop) has a higher burden of COVID-19 cases than Asia (390 cases/pop).

The number of deaths per 100,000 population (hereafter referred to as deaths/pop) followed the same trend as the number of cases per 100,000. Deaths per 100,000 in Oceania (2 deaths/pop), Africa (4 deaths/pop), Asia (6 deaths/pop) are far below the global value (19 deaths/pop), while it is above in South America (39 deaths/pop), Europe (55 deaths/pop), and North America (91 deaths/pop). This shows a positive correlation between the number of COVID-19 tests and the number of COVID-19 cases and deaths reported.

In Africa, an analysis of COVID-19 cases per 100,000 population (Figure 2) showed that, while the continent had comparatively low case numbers, individual nations had high COVID-19 burdens. South Africa has been the most seriously affected with 1,260 cases/pop; however, Tunisia (678 cases/pop) and Morocco (791 cases/pop) in North Africa also appear to be greatly impacted. Similar patterns were observed for Libya in North Africa (1,076 cases/pop), Namibia (528 cases/pop) and Botswana (347 cases/pop) in Southern Africa, and Gabon in Central Africa (404 cases/pop). Other African countries were mildly impacted as well. Overall, the number of cases/pop demonstrates that northern and southern countries, and Gabon in Central Africa, were the most affected regions/country in Africa.

To get a deeper insight into the impact of SARS-CoV-2 on African countries, we analyzed the reported deaths as represented in Figure 3. Considering the absolute number of deaths, South Africa remained

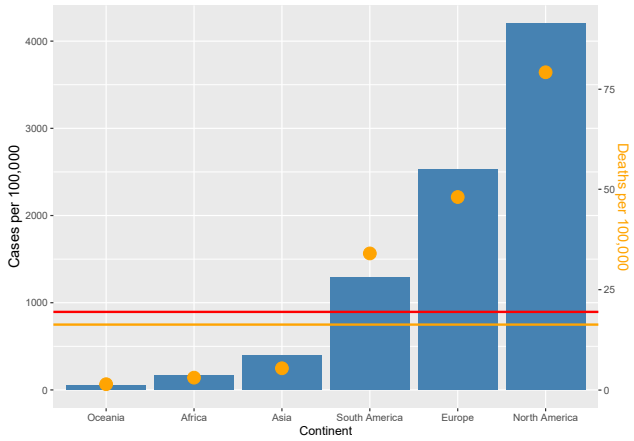


Figure 1. Number of COVID-19 reported cases and deaths per 100,000 population in the different continents

The red line represents the average absolute number of COVID-19 cases worldwide per 100,000 population in the world (i.e., (global COVID-19 cases/world population) x 100,000). The orange points represent the average number of deaths per 100,000 population in the different continents with its scale on the left y axis. The orange line represents the number of deaths globally per 100,000 world population.

the worst hit on the continent with 20,241 deaths, distantly followed by Egypt (n = 6,453), and other North African countries. While Ethiopia, Kenya, Nigeria, Sudan, and Libya have reported more than 1,000 deaths each, about 20 African countries have recorded less than a total of 500 deaths. Considering the number of deaths recorded per 1,000 reported cases (deaths/cases) (Figure 3 - right), the analysis showed Western Sahara as the worst affected with 100 death/cases, followed by Sudan (76 death/cases), Egypt in the North (58 death/cases), and Chad in the West (63 death/cases).

South Africa and Morocco recorded the highest absolute numbers of COVID-19 cases in Africa (Figure 2 - left), though this might reflect the large number of tests (5,110,384 and 3,646,330 for South Africa and Morocco, respectively) conducted in these two countries (Figure 4). The third highest number of tests was conducted in Ethiopia (n = 1,562,008) albeit approximately 60% less than the number of tests conducted in Morocco. Taking population into consideration (number of tests conducted per 100,000 population; tests/pop), the islands of Mauritius and Cape Verde performed the highest number of tests (22,389 tests/pop and 18,250 tests/pop, respectively) followed by Botswana (13,955 tests/pop) and Gabon (11,860 tests/pop) who carried out more tests, relative to population, than South Africa (8,576 tests/pop) and Morocco (9,835 tests/pop).

We analyzed the number of positive tests per 1,000 COVID-19 tests (pos/test) conducted (Figure 5). Mayotte was estimated to have the highest positivity rate (271 pos/test), followed by Guinea (250 pos/test), South Sudan (249 pos/test), and Tunisia (200 pos/test). Other African countries with more than 100 positive cases per 1,000 COVID-19 tests include Libya (198 pos/test), Madagascar (186 pos/test), Gambia

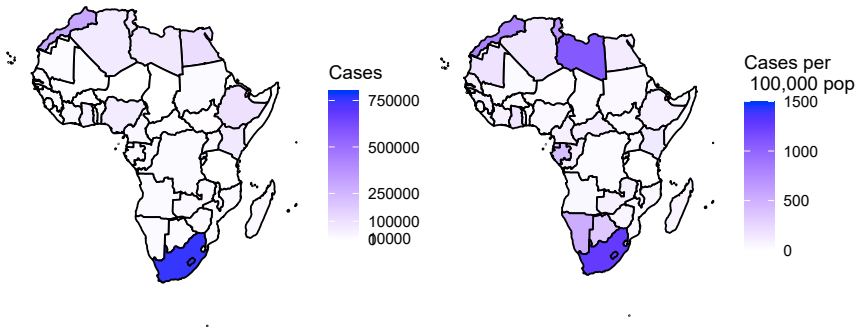


Figure 2. COVID-19 cases reported in African countries

Absolute number of cases (left). Number of cases per 100,000 population (right).

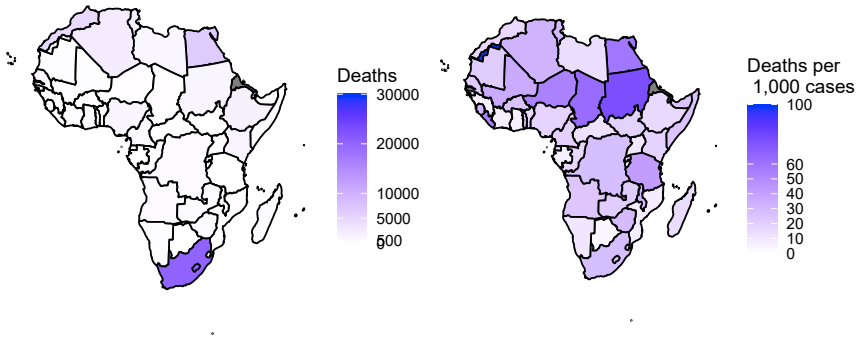


Figure 3. Reported deaths from COVID-19 in Africa
Absolute number of total deaths (left) per country. Number of deaths per 1,000 reported cases (right).

(166 pos/test), Cameroon (152 pos/test), Central African Republic (150 pos/test), South Africa (147 pos/test), Reunion (141 pos/test), Sao Tome and Principe (136 pos/test), Swaziland (110 pos/test), Egypt (110), and Ivory Coast (102 pos/test). The lowest positivity rates were observed in countries such as Benin (10 pos/test), Rwanda (9 pos/test), and Mauritius (2 pos/test).

Evolutionary history

Despite the lower number of genetic sequences available compared to epidemiological data, the former allow us to gain insight into the viral diversity circulating in the continent as well as the evolutionary relationships and transmission dynamics of viruses in different regions.

Worldwide, we observed 782 Pango lineages, nine GISAID clades and 10 Nextstrain clades. Europe submitted about 65% (n = 208,538) of the SARS-CoV-2 sequences in GISAID, the majority being from the United Kingdom (46%, n = 147,137). Africa submitted 2% of the SARS-CoV-2 sequences (n = 5,229) with most of the sequences (55%, n = 2,882) coming from South Africa. Democratic Republic of the Congo and Gambia followed with 7% (n = 360) each, then Kenya with 290 sequences (6%) and Nigeria with 4% (n = 223) (Figure S2).

Even though in absolute terms, South Africa submitted most of the sequences (Figure 6 - left), a closer look at the number of sequences submitted (Figure 6 - right) revealed that the Democratic Republic of the Congo was the largest contributor of genomic data per 1,000 reported COVID-19 cases in Africa.

Using Nextstrain clade nomenclature (Figure S3), we observed that 20A.EU1 was the dominant circulating clade in Europe followed by 20B and 20A. Clades 20C and 20A are predominant in North America, while 20B was predominant in Oceania. In Africa (Table S3), 20B and 20A were the dominant circulating clades, accounting for about 82% of all sequences available for the continent.

For simplicity, we present the most prevalent (top 1%) of the 782 Pango lineages identified worldwide in Figure S2. Europe had the most diverse lineages with B.1.177 being the most prevalent, followed by

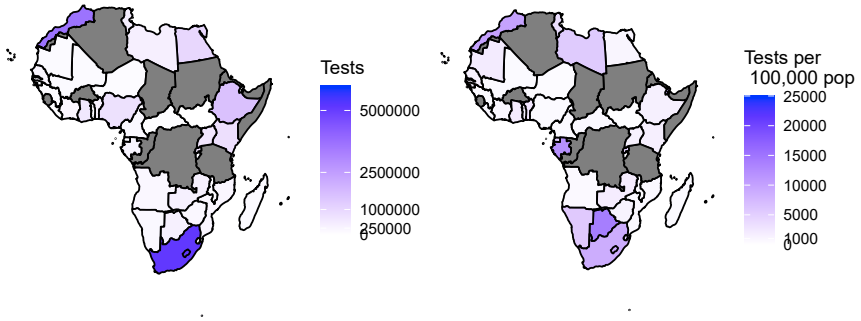


Figure 4. Number of SARS-CoV-2 tests conducted in African countries
Absolute number of tests (left). Number of tests per 100,000 population (right).

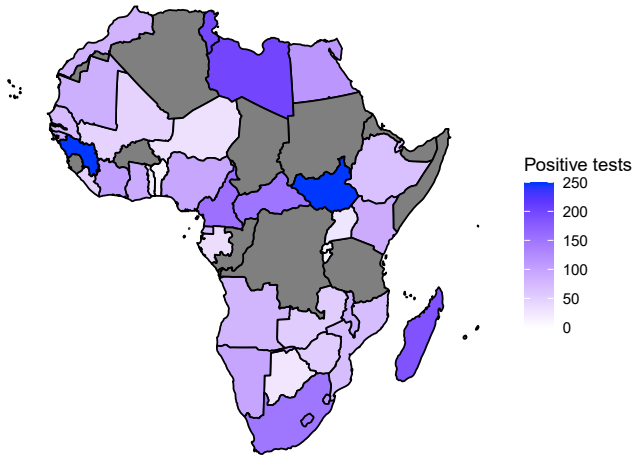


Figure 5. COVID-19 positivity rate in Africa

Number of positive patients out of every 1000 COVID-19 tests. Gray shades represent countries for which data is not available.

B.1.1 and B.1. In Asia, B.1.1 and B.1.1.284 were the dominant lineages, while D.2 was the most prevalent in Oceania. In North America, B.1, B.1.2, and B.1.1 were the most prevalent. In contrast, the top 1% Pango lineages (Figure S4) circulating elsewhere in the world were not observed in sequences from Africa and South America. The top 10% of the lineages circulating in Africa is represented in Figure S5. Pango lineage B.1.5 was the dominant lineage circulating in Africa, representing 11.3% (n = 591) of all diversity. Other prevalent lineages in Africa are B.1 (n = 546, 10.4%), B.1.1 (n = 518, 9.91%), B.1.1.206 (n = 481, 9.20%), B.1.351 (n = 349, 7%), and C.1 (n = 271, 5%).

The first SARS-CoV-2 sequence collected from Africa on March 1, 2020 was found to belong to the Pango lineage B.1.5 (Figure 7), while the second reported on March 8, 2020 belongs to the B.1 lineage. Pango lineages B.1, B.1.1, and B.1.5 circulated in Africa from March 1, 2020 through June 7, 2020; and were replaced by Pango lineage B.1.1.206 that was first reported in the continent on June 8, 2020 and dominated between June 8 and November 2, 2020. The Pango lineage B.1.351 was first reported in South Africa on October 10, 2020 and it became the most prevalent lineage on November 3, 2020, likely as a consequence of the increased sequencing efforts in the country.

Further, we analyzed the phylogenetic relationships among African sequences and those from other parts of the world (Figure 8). The phylogenetic tree demonstrates that African viruses cluster closely with viruses from all continents, but mostly with those from Europe, a source that generated some of the large

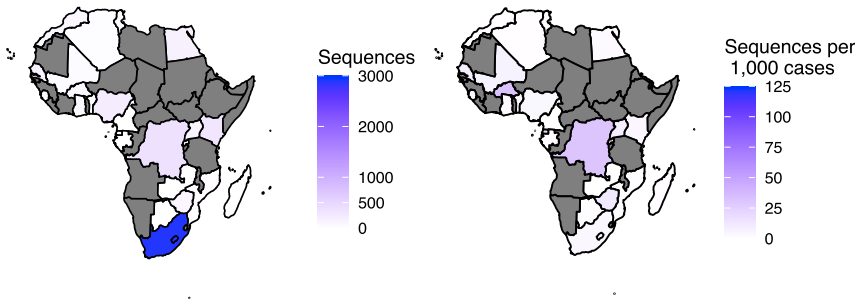


Figure 6. Sequences from African countries submitted to GISAID

Absolute number of sequences from Africa submitted to GISAID (left). Sequences available in GISAID per 1,000 SARS-CoV-2 cases in Africa (right). Gray shade represents countries for which no sequences were available (n = 27/57 (47.4%) countries).

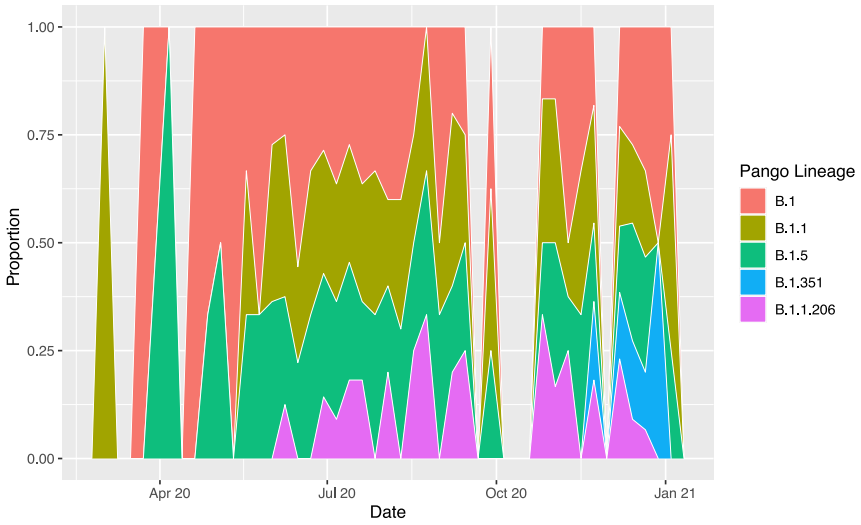


Figure 7. The incidence of the top five Pango lineages circulating in Africa between March 1, 2020 and January 7, 2021.

outbreaks detected in the phylogenetic tree. Of note are the viruses sampled from Uganda in March 2020 which appear in large clusters of European viruses. Interestingly, we also observed a cluster of a Ugandan virus with another from Oceania, both with identical collection dates. Although Europe appeared to seed many African outbreaks, African viruses also frequently clustered with those from Asia, Oceania, and especially South America, which can be seen associated with viruses from South African, Congolese, and Gambian clusters. A similar phenomenon was observed with North America which had viruses clustering with those from Morocco, South Africa, Egypt, Kenya, and Nigeria. It was also observed that African clusters mostly contained sequences from the same or adjacent countries, which is evident in South African clusters (Data S1 and Data S2). A number of studies have focused on the South African outbreaks in depth (Giandhari et al., 2021; Tegally et al., 2021, Tegally et al., 2021). The current study identified outbreaks (clusters

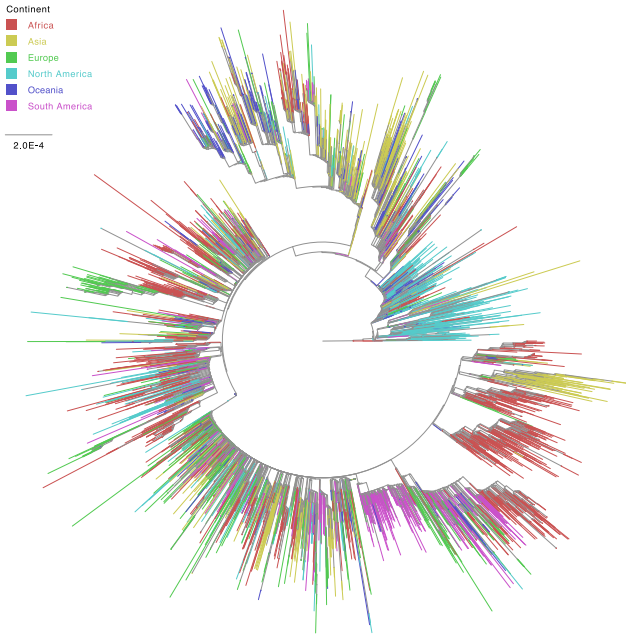


Figure 8. Maximum likelihood tree colored by continent
Phylogenetic tree inferred for a dataset with genetic sequences from all continents.

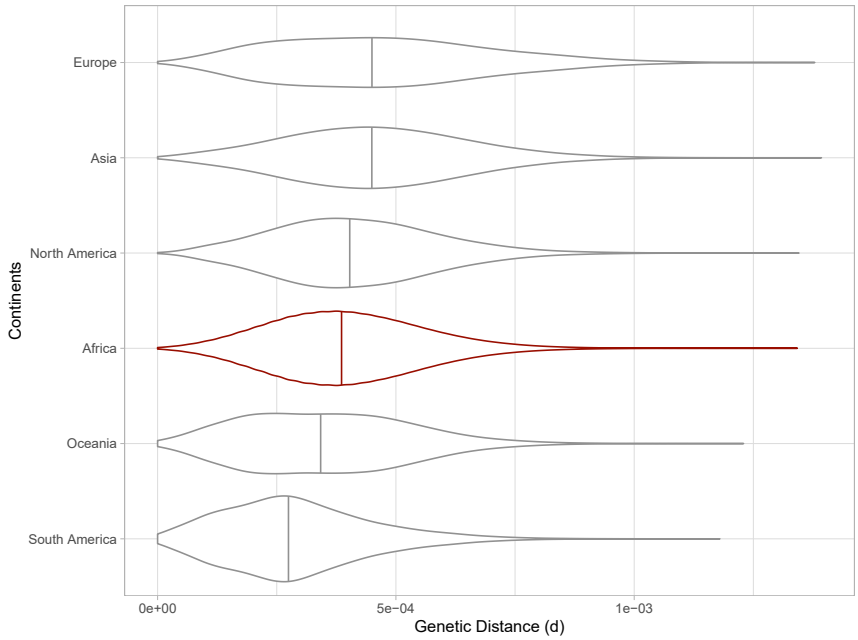


Figure 9. Evolutionary divergence of SARS-CoV-2 across continents

Violin plots represent the distribution of pairwise genetic distances between all sequences for isolates in each continent. Vertical lines depict the mean pairwise genetic distance between all samples in each continent.

with more than 15 sequences) in Egypt, Democratic Republic of the Congo, and Gambia. Some of these clustered closely with viruses from other African countries as observed for instance with some of the Gambian outbreaks, which were related to West African neighbors, such as Senegal. (Data S2). This phenomenon was also observed for Kenya and neighboring Uganda viruses.

We looked specifically at the viral genetic diversity within Africa, as compared to the genetic diversity observed in other continents (Figure 9 and Data S3). Inspection of the continent-specific genetic distance distributions with a Wilcoxon signed-rank test revealed that the viral diversity circulating in Africa is significantly higher (p value < 2.2e-16) than that estimated in Oceania and South America, but significantly lower than that in Asia, Europe, and North America. These findings indicate that the African epidemic is closest to that of South America and farthest from those of Asia and North America (Figure S6).

We also investigated the genetic diversity across countries in the African continent (Figure S7). The lowest viral diversities were observed in Madagascar, Zambia, and Algeria, while the highest viral diversity circulated in Nigeria, Tunisia, and Sierra Leone. Of note, distinct viral populations were estimated among countries in North Africa (namely in Tunisia), West African countries (including Gambia, Ghana, Mali, Nigeria, Senegal, and Sierra Leone), and South Africa. Viruses collected in West Africa were also genetically distant from those in South Africa. Generally, with the exception of countries with overall lower within-country diversities, substantial diversity was observed among African countries.

Identification of repeat patterns and motifs

GLAM2 determines recurring motifs with deletions and insertions, and their presence in functional proteins. Using the Tomtom motif tool, we identified the motifs represented in the table below, as well as their functional class when compared with the motif database (Eukaryotic Linear Motif (ELM) resource - http://elm.eu.org/elms/LIG_IBS_1.html) (Table 1). The p-value denotes the likelihood of a random motif with equal width to function as a target and align better to motifs in the database, thus producing a match score as either good or better than another target. The e-value shows the expected number of false positives in the matches, with a threshold of ten or less.

The motifs were matched with the study dataset but were not found in all isolates (data not shown). We observed a position shift in the motifs across the isolates, but it was not clear if the shifts altered the

Table 1. Repeat patterns and motifs in SARS-CoV-2 genomes from Africa

ID	Motif	Accession number	Length	Functional site class	p-value	e-value
1	ILRKGGR TIAFGGC VFSYVG CHNKC AYWVPR ASANIG CNHTGV VGEGSEG	ELM: ELME000129 (LIG_IBS_1)	50	Integrin binding sites N-glycosylation site	3.43E-04 4.70e-03	5.63E-02 7.71E-01
2	ILRKGGR TIAFGGC VFSYVG CHNKC AYWVPR ASANIG CNHTGV VGEGSEG	ELM: ELME000316 (LIG_Integrin_isoDGR_1)	50	Integrin binding sites	2.68E-02	2.68E-02

functionality of the motifs. Therefore, it indicates that the motifs are present but at different positions in the genome, probably due to deletions and substitutions along the genome of African SARS-CoV-2 isolates.

Two of the motifs revealed by the analysis were found in the ORF1ab gene (locus Gu280-gp01) and were identified as integrin binding sites occurring mainly at positions 396–445 (ID 1) and 3,361–3,409 (ID 2). Our analysis also uncovered a motif that could function as an N-glycosylation site, mostly in positions 396–445 (ID 1) (Table 1).

DISCUSSION

Ongoing research on SARS-CoV-2 classical and genomic epidemiology in Africa is crucial for monitoring the circulating genetic diversity of the virus, its clinical presentation, and epidemiological profiles, as well as estimating the magnitude of a pandemic, and informing the development and implementation of effective control measures in the African continent. Mutations in the viral genome may also raise concerns for reliable and effective diagnostic surveillance and for monitoring of SARS-COV-2 transmission dynamics (Galloway et al., 2021).

We assessed the impact of the COVID-19 pandemic in Africa by evaluating the trends of absolute number of cases and deaths, but given the limited testing on the continent, we also relied on the proportion of the population that was screened for COVID-19 and the proportion that tested positive (positivity rate), which were relatively elevated for Mayotte, Guinea, South Sudan, and Tunisia. Compared to other continents, Africa appears to be relatively spared in terms of case fatality rate. Nonetheless, Egypt, Sudan, Chad, and Niger, all of which share borders, were found to have the highest numbers of COVID-19-related deaths, and thus further investigation is necessary to uncover the factors that led to this public health burden. We estimated that the impact of SARS-CoV-2 in Africa has been below the global average, both in terms of cases and mortality. However, this is based on the available information associated with the reported metrics on cases, deaths, and number of tests conducted, which appear to be underestimated in Africa. The younger African population might also contribute to keeping the number of severe cases low compared to older populations elsewhere (Lee et al., 2020). The under-reporting calls for policy directions in Africa to be tailored toward expanding screening and improving implementation of measures that curb community spread. It has also been hypothesized that the apparently lower impact of the disease in Africa might be partly due to the heavy use of chloroquine, and its derivative hydroxychloroquine, to prevent or treat malaria, as well as autoimmune conditions and other diseases (Tönnesmann et al., 2013; Ben-Zvi et al., 2012). These antimalarials have been shown, *in vitro* and *in vivo*, to inhibit the pH-dependent steps in the replication of several viruses, including SARS-CoV-2 (Yao et al., 2020; Colson et al., 2020; Gao et al., 2020; Liu et al., 2020, Wang et al., 2020b). Although both hydroxychloroquine and chloroquine, either alone or in combination with azithromycin, are commonly used in several African countries for the treatment of COVID-19 (Abena et al., 2020), the

use of these drugs has not been recommended by international health organizations, and findings suggest further studies are warranted to arrive at a conclusive basis for their use (Pastick et al., 2020).

Furthermore, it is not certain that control measures that have proved effective in the global north will be equally effective in Africa. For instance, as shown in our findings, lockdown was not only counterproductive in different socio-economic areas, but also ineffective in curbing COVID-19 transmission in Africa. Some reasons that could be assigned include illiteracy, poverty, and cultural norms. Effective pursuance of grass-root education on good public health practices, mass distribution of disposable masks, free access to running water and soap, and availability of sanitizers in various public places, represent important avenues to tackling the current and future outbreaks.

Although SARS-CoV-2 sequence data constitute an integral part of the decision making in other continents (Lu et al., 2020; Zhu et al., 2020), Africa has yet to fully employ sequence information to manage its COVID-19 outbreaks. This could be attributed to the limited technical competencies and infrastructural deficiencies (Jerving, 2020), such as low sequencing capacity, bioinformatics, computational skills and pipelines, and limited funding for and access to sequencing reagents. To date, African countries have contributed few SARS-CoV-2 genomic sequences in the global pool of open access repositories, such as NCBI GenBank and GISAID databases.

Molecular epidemiology studies remain imperative as African countries reopen their borders with some level of COVID-19 testing but without real-time genomic surveillance to monitor the emergence of viral variants. To this end, this study pursued detailed phylogenetic inference, comparison of the evolutionary divergence, detection of repeat patterns and motifs, and analysis of the geographical distribution of SARS-CoV-2 trends in Africa. Here, we employed, among other methods, the Pango nomenclature system designed to implement a dynamic classification of SARS-CoV-2 lineages that incorporates both genetic and geographical components (O'Toole et al., 2021. in prep). The Pango nomenclature contains molecular signatures that are helpful for tracking SARS-CoV-2 introduction, emergence and spread (Andersen et al., 2020). We observed differences in the lineages circulating in Africa from those in most parts of the world. Lineage B.1.5 was identified as the dominant genetic lineage circulating in Africa, but most of the top 1% lineages in circulation worldwide were not found. This observation might be a consequence of under-surveillance as there is a relatively low number of African sequences available in the genetic databases, founder effects, or the inefficient implementation of control measures, such as testing of travelers that may contribute to viral introductions and subsequent spread in the community. Consequently, the biological significance of the B.1.5 lineage, its epidemiologic features and spatial patterns deserve monitoring and further exploration, as there have been no reports of changes in transmissibility, fatality rates, or vaccine efficacy, in contrast with lineage B.1.351 which dominated in early 2021 (Chen et al., 2021; Luo et al., 2021; Irfan and Chagla, 2021; Abu-Raddad et al., 2021; Shinde et al., 2021; Davies et al., 2021; Jassat et al., 2021).

The phylogenetic topological relationships revealed that African genomes tended to cluster with those from Europe, which is in line with the high cultural and business connectivity between these continents. We also observed several noteworthy outbreaks in Egypt, Democratic Republic of the Congo, Gambia, and South Africa, which reflects the higher number of sequences available for these countries. The higher viral diversity in Africa, compared to that in Oceania and South America, can initially be thought to be a reflection of the within-continent and inter-continent connectivity, as well as the travel patterns of individuals in Africa, which includes several major metropolitan areas, such as those in South Africa, Egypt, Ethiopia, and Morocco. However, it might also be a consequence of diversity bottlenecks or the limited genomic surveillance in the other continents. In addition, we also observed that viruses circulating in different African regions are relatively genetically diverse. This might be a consequence of varied sources of introduction of a variety of lineages into the different regions of the continent. It can also be due to easier viral spread among neighboring countries or those that share language or economic ties. This observation highlights the need for more in-depth phylodynamic studies to gain insight into the transmission routes leading to viral introductions and outbreaks throughout the continent, though these have been partially addressed in other studies (Wilkinson et al., 2021).

We also identified *de novo* protein motifs that may have functional significance for SARS-CoV-2. Integrins are essential eukaryotic cells' collagen receptors formed by a noncovalent interaction of two transmembrane glycoproteins subunits developing into about 24 varieties of heterodimers that facilitate

the binding of cells to extracellular matrix and junctions. Hence, the integrin-binding domain facilitates cell-attachment and cell-adhesion (Sigrist et al., 2020). Integrins may be used in place of the ACE-2 receptor because there is an integrin binding motif (arginine-glycine-aspartate [RGD]) on the spike protein (Beddingfield et al., 2021), that could potentially mediate viral entry into host cells and influence SARS-CoV-2 tissue tropism, viral transmission, and pathogenicity. Therefore, integrins should be further studied in order to gain insight into their roles in the viral pathogenicity and transmission. Despite several studies focused on ACE2 (Makowski et al., 2021) (Lan et al., 2020), the hypothesis that SARS-CoV-2 integrins could serve as alternative viral receptors needs to be validated experimentally. In addition, several integrins are believed to be co-receptors of SARS-CoV-2 infections, and thus primary infection assays focusing on integrins should be carried out (Beddingfield et al., 2021).

An N-glycosylation site employs a biosynthetic process of high complexity that is responsible for protein maturation along its secretory pathway (Yang et al., 2019; Galbán and Duckett, 2010). Glycosylation is a posttranslational modification in viral proteins that determines protein conformation, function, and host adaptation. It can also act as a defense mechanism for SARS-CoV-2 against the immune cells and antibodies of the host, making it difficult to distinguish, identify, and target the virus for elimination (Watanabe et al., 2019; Grant et al., 2020; Ramírez Hernández et al., 2021). This may in turn contribute to the cell infection rate and therefore to disease severity (Reily et al., 2019). We hypothesize that the presence and potential mutation of N-glycosylation domains on the SARS-CoV-2 genome may have implications for the binding affinity as previously described by Zhao et al. (2020). This may likely account for the immune evasion for SARS-CoV-2 by camouflaging immunogenic viral protein epitopes (Watanabe et al., 2020) as previously described by Watanabe et al. (2020). Further investigation using animal models would add to the understanding of the impact of mutations in the N-glycosylation domains on the efficacy of ongoing vaccination against SARS-CoV-2 around the world.

This work was produced not without certain challenges. Our search for data brought to the fore a seeming lack of transparency in data disclosure and availability, both at the genetic and epidemiological fronts. This is concerning as it may deprive the continent and the global scientific community of useful information for consideration in the fight against the COVID-19 pandemic. Secondly, very few sequences from Africa were available in public and semi-public sequence databases as compared to other continents, and certain factors might have contributed to this outcome; however, key factors among these are the lack of resources and technical proficiencies. Sequencing, bioinformatics, and computational expertise can be greatly improved with capacity-building trainings organized by African entities and other international partners, such as the initiatives by the Fogarty International Center, National Institutes of Health and Johns Hopkins University Applied Physics Laboratory that have been regularly training scientists from low and middle-income countries, particularly during the COVID-19 pandemic. These trainings have greatly improved the technical skills of participants toward analyzing the epidemiological and evolutionary trends of SARS-CoV-2 in Africa, as presented in this study. Furthermore, funding from African countries to support African scientists to carry out in-depth research on various aspects of SARS-CoV-2 in Africa is scarce (Oladipo et al., 2020). Local governmental commitment to funding research would allow scientists to be more independent in their research pursuits.

In conclusion, this work describes the molecular epidemiology, analyzes the genetic variability of SARS-CoV-2 in Africa, and highlights the need for continuous genomic and epidemiological surveillance, which is imperative for tracing the emergence of genetic variants that can have significant effects on antigenicity, immunity, transmissibility, and potential vaccine escape. This information will also allow investigation of the transmission dynamics and resurgence of waves of infection, as well as optimize public health measures, such as the deployment of vaccine formulations across the continent.

LIMITATIONS OF THE STUDY

Despite being one of the few studies that comprehensively explored the viral genetic diversity, evolutionary history, and functional genome patterns of SARS-CoV-2 in Africa, we understand that our observations might have been conditioned by sampling bias. The limited testing capacity and/or under-reporting of cases and mortality might influence our estimation of the impact of SARS-CoV-2 in Africa, which was found to be below the global average. We note that where cases and deaths were reported, in some of the instances the actual number of tests conducted was not reported, which did not allow estimation of positivity rate. Therefore, it is conceivable that underestimation in the reported cases and mortality metrics could mask the actual incidence and impact of the pandemic in Africa. Another caveat lies in the full genome

sequencing technology being also limited in most public health, research, or academic institutions in Africa. The associated cost, especially in less endowed countries, makes genome sequencing an option not considered routinely. Consequently, the number of sequences generated from most of the countries was scarce, and the fact that some countries were not represented at all in our analysis due to lack of sequence data might suggest that our observations may not represent the true state of the situation in the African continent. However, our findings shed light on the state of affairs and may help inform public health policies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Compilation of genomic datasets
  - Phylogenetic inference
  - Comparison of evolutionary divergence
  - Geographical distribution of COVID-19 pandemic in Africa
  - Detection of repeat patterns and motifs
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.103880.

ACKNOWLEDGMENTS

We acknowledge the authors and laboratories that generated and submitted sequences into GISAID's EpiFlu Database. A full table of sequence authors is available in Table S2. We also thank Helix Biogen Institute, also known as Helix Biogen Consult for their support. We gratefully acknowledge the Fogarty International Center at the National Institutes of Health and the Johns Hopkins University Applied Physics Laboratory for developing in-country capacity for whole genome sequencing and phylogenetics of SARS-CoV-2 and providing technical guidance. The opinions expressed in this article are those of the authors and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

AUTHOR CONTRIBUTIONS

Conceptualization, EKO and NST; Software, OSO and NST; Validation, OSO, NINT, and NST; Analysis, OSO, NINT, AJ, TTO, TYS, TOG, IEA, LOE, MAO, EMJ, and NST; Visualization, OSO and NST; Writing - original draft, OSO, NINT, AJ, TTO, TYS, AAKK, EOD, TOG, IEA, LOE, IJA, MAO, EMJ, EKO, and NST; Writing – review and editing, OSO, NINT, AJ, TTO, TYS, AAKK, EOD, TOG, IJA, DJS, EKO, and NST; Supervision and project administration, NST; Funding acquisition, DJS and EKO.

DECLARATIONS OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in science. The author list of this paper includes contributors from the location where the research was conducted who participated in the data collection, design, analysis, and/or interpretation of the work.

Received: August 3, 2021
Revised: November 29, 2021
Accepted: February 3, 2022
Published: March 18, 2022

REFERENCES

Abena, P.M., Decloedt, E.H., Bottieau, E., Suleman, F., Adejumo, P., Sam-Agudu, N.A., Tamfum, J.J.M., Seydi, M., Eholie, S.P., and Mills, E.J. (2020). Chloroquine and hydroxychloroquine for the prevention or treatment of COVID-19 in Africa: caution for inappropriate off-label use in healthcare settings. Am. J. Trop. Med. Hyg. *102*, 1184–1188.

Abu-Raddad, L.J., Chemaitelly, H., and Butt, A.A. (2021). Effectiveness of the BNT162b2 Covid-19 vaccine against the B. 1.1. 7 and B. 1.351 variants. New Engl. J. Med. *385*, 187–189.

Adler, D., and Kelly, S.T. (2020). Vioplot: violin plot. R package version 0.3.7. https://github.com/TomKellyGenetics/vioplot.

African Centres for Disease Control (2020). Novel Coronavirus (2019-nCoV) Global Epidemic – 31 March 2020. https://africacdc.org/disease-outbreak/novel-coronavirus-2019-ncov-global-epidemic-31-march-2020/.

Aiewsakun, P., Wongtrakoongate, P., Thawornwattana, Y., Hongeng, S., and Thitithanyanont, A. (2020). SARS-CoV-2 genetic variations associated with COVID-19 severity. Preprint at MedRxiv. https://doi.org/10.1101/2020.05.27.20114546.

Alanagreh, L.A., Alzoughool, F., and Atoum, M. (2020). The human coronavirus disease COVID-19: its origin, characteristics, and insights into potential drugs and its mechanisms. Pathogens *9*, 331.

Alguwaizani, S., Park, B., Zhou, X., Huang, D.S., and Han, K. (2018). Predicting interactions between virus and host proteins using repeat patterns and composition of amino acids. J. Healthc. Eng. *2018*, 1391265.

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., and Garry, R.F. (2020). The proximal origin of SARS-CoV-2. Nat. Med. *26*, 450–452.

Auguie, B. (2017). gridExtra: Miscellaneous Functions for" Grid" Graphics (R Package version 2.3) [Computer software].

Avise, J.C. (2000). Phylogeography: The History and Formation of Species (Harvard university press).

Becker, R., Wilks, A.R., Brownrigg, R., Minka, T.P., and Deckmyn, A. (2018). maps: Draw Geographical Maps. R package version 3.3. 0. https://CRAN.R-project.org/package=maps.

Beddingfield, B.J., Iwanaga, N., Chapagain, P.P., Zheng, W., Roy, C.J., Hu, T.Y., Kolls, J.K., and Bix, G.J. (2021). The integrin binding peptide, ATN-161, as a novel therapy for SARS-CoV-2 infection. Basic Transl. Sci. *6*, 1–8.

Ben-Zvi, I., Kivity, S., Langevitz, P., and Shoenfeld, Y. (2012). Hydroxychloroquine: from malaria to autoimmunity. Clin. Rev. Allergy Immunol. *42*, 145–153.

Bivand, R., Lewin-Koh, N., Pebesma, E., Archer, E., Baddeley, A., Bearman, N., Bibiko, H.-J., Brey, S., Callahan, J., and Carrillo, G. (2021). Package 'maptools'.

Chan, J.F.W., Kok, K.-H., Zhu, Z., Chu, H., To, K.K.W., Yuan, S., and Yuen, K.-Y. (2020). Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerging Microbes Infections *9*, 221–236.

Bowman, A.W., and Azzalini, A. (2018). R Package 'sm': Nonparametric Smoothing Methods (version 2.2-5.6). http://www.stats.gla.ac.uk/~adrian/sm.

Chan, C., Chan, G.C., Leeper, T.J., and Becker, J. (2021). Rio: A Swiss-army Knife for Data File I/O. R Package version 0.5.29.

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., and Wei, Y. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. The lancet *395*, 507–513.

Chen, X., Chen, Z., Azman, A.S., Sun, R., Lu, W., Zheng, N., Zhou, J., Wu, Q., Deng, X., and Zhao, Z. (2021). Comprehensive mapping of neutralizing antibodies against SARS-CoV-2 variants induced by natural infection or vaccination. Preprint at medRxiv. https://doi.org/10.1101/10.1101/2021.05.03.21256506.

Colson, P., Rolain, J.-M., and Raoult, D. (2020). Chloroquine for the 2019 novel coronavirus SARS-CoV-2. Int.J.Antimicrob.Agents *55*, 105923.

Corman, V.M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D.K., Bleicker, T., Brünink, S., Schneider, J., and Schmidt, M.L. (2020). Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Eurosurveillance *25*, 2000045.

Cui, J., Li, F., and Shi, Z.-L. (2019). Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. *17*, 181–192.

Davies, N.G., Abbott, S., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J., Pearson, C.A., Russell, T.W., Tully, D.C., and Washburne, A.D. (2021). Estimated transmissibility and severity of novel SARS-CoV-2 variant of concern 202012/01 in England, Preprint at medRxiv. https://doi.org/10.1101/2020.12.24.20248822.

Elbe, S., and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob. Challenges *1*, 33–46. https://doi.org/10.1002/gch2.1018 PMCID: 31565258.

Faria, N.R., Mellan, T.A., Whittaker, C., Claro, I.M., Candido, D.D.S., Mishra, S., Crispim, M.A., Sales, F.C., Hawryluk, I., and Mccrone, J.T. (2021). Genomics and epidemiology of a novel SARS-CoV-2 lineage in Manaus Brazil. Preprint at medRxiv. https://doi.org/10.1101/2021.02.26.21252554.

Galbán, S., and Duckett, C.S. (2010). XIAP as a ubiquitin ligase in cellular signaling. Cell Death Differ. *17*, 54–60.

Galloway, S.E., Paul, P., Maccannell, D.R., Johansson, M.A., Brooks, J.T., Macneil, A., Slayton, R.B., Tong, S., Silk, B.J., and Armstrong, G.L. (2021). Emergence of SARS-CoV-2 b. 1.1. 7 lineage—United States, december 29, 2020–january 12, 2021. Morbidity Mortality Weekly Rep. *70*, 95.

Gao, J., Tian, Z., and Yang, X. (2020). Breakthrough: chloroquine phosphate has shown apparent efficacy in treatment of COVID-19 associated pneumonia in clinical studies. Biosci. Trends *14*, 72–73.

Grolemund, G., and Wickham, H. (2011). Dates and times made easy with lubridate. J. Stat. Softw. *40*, 1–25. https://www.jstatsoft.org/v40/i03/.

Giandhari, J., Pillay, S., Wilkinson, E., Tegally, H., Sinayskiy, I., Schuld, M., Lourenço, J., Chimukangara, B., Lessells, R., and Moosa, Y. (2021). Early transmission of SARS-CoV-2 in South Africa: an epidemiological and phylogenetic report. Int. J. Infect. Dis. *103*, 234–241.

Grant, O.C., Montgomery, D., Ito, K., and Woods, R.J. (2020). Analysis of the SARS-CoV-2 spike protein glycan shield reveals implications for immune recognition. Scientific Rep. *10*, 1–11.

Guo, Y.-R., Cao, Q.-D., Hong, Z.-S., Tan, Y.-Y., Chen, S.-D., Jin, H.-J., Tan, K.-S., Wang, D.-Y., and Yan, Y. (2020). The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status. Mil. Med. Res. *7*, 1–10.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol. *8*, 1–9.

Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. *35*, 518–522.

Irfan, N., and Chagla, Z. (2021). In South Africa, a 2-dose Oxford/AZ vaccine did not prevent mild to moderate COVID-19 (cases mainly B. 1.351 variant). Ann. Intern. Med. *174*, JC50.

Jassat, W., Mudara, C., Ozougwu, L., Tempia, S., Blumberg, L., Davies, M.-A., Pillay, Y., Carter, T., Morewane, R., and Wolmarans, M. (2021). Increased mortality among individuals hospitalised with COVID-19 during the second wave in South Africa. medRxiv.

Jerving, S. (2020). Strengthening Africa's Ability to 'decode' the Coronavirus. https://www.devex.com/news/strengthening-africa-s-ability-to-decode-the-coronavirus-97319.

Kahle, D., and Wickham, H. (2013). ggmap: spatial Visualization with ggplot2. R. Journal *5*, 144–161.

Kamvar, Z.N. (2021). aweek: Convert Dates to Arbitrary Week Definitions. R package version 1.0.2, https://CRAN.R-project.org/package=aweek.

Kassambara, A. (2019). Ggcorrplot: Visualization of a Correlation Matrix Using'ggplot2.

Kassambara, A. (2020). ggpubr:'ggplot2'based Publication Ready Plots (R Package Version 0.4. 0) [Computer software].

Katoh, K., Rozewicki, J., and Yamada, K.D. (2019). MAFFT online service: multiple sequence

alignment, interactive sequence choice and visualization. Brief. Bioinformatics *20*, 1160–1166.

Korber, B., Fischer, W., Gnanakaran, S.G., Yoon, H., Theiler, J., Abfalterer, W., Foley, B., Giorgi, E.E., Bhattacharya, T., and Parker, M.D. (2020). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. Preprint at bioRxiv. https://doi.org/10.1101/2020.04.29.069054.

Kumar, R., Verma, H., Singhvi, N., Sood, U., Gupta, V., Singh, M., Kumari, R., Hira, P., Nagar, S., and Talwar, C. (2020). Comparative genomic analysis of rapidly evolving sars-cov-2 reveals mosaic pattern of phylogeographical distribution. Msystems *5*, e00505–e00520.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). Mega X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. *35*, 1547–1549.

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics *30*, 3276–3278.

Lee, P.I., Hu, Y.L., Chen, P.Y., Huang, Y.C., and Hsueh, P.R. (2020). Are children less susceptible to COVID-19? J. Microbiol. Immunol. Infect. *53*, 371.

Liu, J., Cao, R., Xu, M., Wang, X., Zhang, H., Hu, H., Li, Y., Hu, Z., Zhong, W., and Wang, M. (2020). Hydroxychloroquine, a less toxic derivative of chloroquine, is effective in inhibiting SARS-CoV-2 infection *in vitro*. Cell Discov. *6*, 1–4.

Lone, S.A., and Ahmad, A. (2020). COVID-19 pandemic–an African perspective. Emerging microbes and infections *9*, 1300–1308.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., and Zhu, N. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The lancet *395*, 565–574.

Lan, J., Ge, J., YU, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., and Wang, X. (2020). Crystal Structure of the 2019-nCoV Spike Receptor-Binding Domain Bound with the ACE2 Receptor, Preprint at bioRxiv. https://doi.org/10.1101/2020.02.19.956235.

Luo, G., Hu, Z., and Letterio, J. (2021). Modeling and predicting antibody durability for mRNA-1273 vaccine for SARS-CoV-2 variants. medRxiv.

Luo, H., and Nijveen, H. (2014). Understanding and identifying amino acid repeats. Brief. Bioinformatics *15*, 582–591.

Makowski, L., Olson-Sidford, W., and WEISEL, J. (2021). Biological and clinical consequences of integrin binding via a rogue RGD motif in the SARS CoV-2 spike protein. Viruses *13.2*, 146. https://doi.org/10.3390/v13020146.

Frith, M.C., Saunders, N.F.W., Kobe, B., and Bailey, T.L. (2008). Discovering sequence motifs with arbitrary insertions and deletions. PLoS Comput. Biol. *4*, e1000071.

Mcilroy, D., Brownrigg, R., Minka, T.P., and Bivand, R. (2020). Mapproj: Map Projections. 2017, R package version, 1. https://CRAN.R-project.org/package=mapproj.

Nigeria Centres for Disease Control (2020). First Case of Coronavirus Disease Confirmed in Nigeria. https://ncdc.gov.ng/news/227/first-case-of-corona-virus-disease-confirmed-in-nigeria.

Neuwirth, E. (2014). RColorBrewer: ColorBrewer palettes. R. Package Version, pp. 1–2.

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.

Oladipo, E., Ajayi, A., Oladipo, A., Ariyo, O., Oladipo, B., Ajayi, L., and Oloke, J. (2020). A call: COVID-19 research funding in Africa. Afr. J. Clin. Exp. Microbiol. *21*, 256–257.

O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, Colquhoun R, Ruis C, Abu-Dahab K, Taylor B, Yeats C, du Plessis L, Maloney D, Medd N, Attwood SW, Aanensen DM, Holmes EC, Pybus OG, Rambaut A. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol. 2021 Jul 30;7(2):veab064. https://doi.org/10.1093/ve/veab064. PMID: 34527285; PMCID: PMC8344591.

Pastick, K., Okafor, E., Wang, F., Lofgren, S., Skipper, C., Nicol, M., Pullen, M., Rajasingham, R., Mcdonald, E., and Lee, T. (2020). Review: hydroxychloroquine and chloroquine for treatment of SARS-CoV-2 (COVID-19). Open Forum Infect. Dis. 1–9.

Pagès, H., Aboyoun, P., Gentleman, R., and Debroy, S. (2021). Biostrings: efficient manipulation of biological strings. R package version 2.62.0.

Plante, J.A., Liu, Y., Liu, J., Xia, H., Johnson, B.A., Lokugamage, K.G., Zhang, X., Muruato, A.E., Zou, J., and Fontes-Garfias, C.R. (2021). Spike mutation D614G alters SARS-CoV-2 fitness. Nature *592*, 116–121.

Shu, Y., and Mccauley, J. (2017). GISAID: global initiative on sharing all influenza data – from vision to reality. EuroSurveillance *22*, 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494 PMCID: PMC5388101.

Rambaut, A., Lam, T.T., Max Carvalho, L., and Pybus, O.G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. *2*, vew007.

Ramírez Hernández, E., Hernández-Zimbrón, L.F., Martínez Zúñiga, N., Leal-García, J.J., Ignacio Hernández, V., Ucharima-Corona, L.E., Pérez Campos, E., and Zenteno, E. (2021). The Role of the SARS-CoV-2 S-protein glycosylation in the interaction of SARS-CoV-2/ACE2 and immunological responses. Viral Immunol. *34.3*, 165–173.

R Core Team (2019). R: A Language and Environment for Statistical Computing, p. 201.

Reily, C., Stewart, T.J., Renfrow, M.B., and Novak, J. (2019). Glycosylation in health and disease. Nat. Rev. Nephrol. *15*, 346–366.

Roser, M., Ritchie, H., Ortiz-Ospina, E., and Hasell, J. (2020). Coronavirus pandemic (COVID-19) (Our World in Data).

Rosling, L., and Rosling, M. (2003). Pneumonia Causes Panic in Guangdong Province (British Medical Journal Publishing Group).

Sapkota, A. (2020). Structure and Genome of SARS-CoV-2 (COVID-19) with diagram. https://microbenotes.com/structure-and-genome-of-sars-cov-2/#genomic-organization-of-sars-cov-2.

Shereen, M.A., Khan, S., Kazmi, A., Bashir, N., and Siddique, R. (2020). COVID-19 infection: origin, transmission, and characteristics of human coronaviruses. J. Adv. Res. *24*, 91–98.

Shinde, V., Bhikha, S., Hoosain, Z., Archary, M., Bhorat, Q., Fairlie, L., Lalloo, U., Masilela, M.S., Moodley, D., and Hanley, S. (2021). Efficacy of NVX-CoV2373 Covid-19 vaccine against the B. 1.351 variant. New Engl. J. Med. *384*, 1899–1909.

Sigrist, C.J., Bridge, A., and Le Mercier, P. (2020). A potential role for integrins in host cell entry by SARS-CoV-2. Antivir. Res. *177*, 104759.

Gupta, S., Stamatoyannopolous, J.A., Bailey, T., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol. *8*, R24.

Snijder, E.J., Limpens, R.W., De Wilde, A.H., De Jong, A.W., Zevenhoven-Dobbe, J.C., Maier, H.J., Faas, F.F., Koster, A.J., and Bárcena, M. (2020). A unifying structural and functional model of the coronavirus replication organelle: tracking down RNA synthesis. PLoS Biol. *18*, e3000715.

Sobhy, H. (2016). A review of functional motifs utilized by viruses. Proteomes *4*, 3.

Stecher, G., Tamura, K., and Kumar, S. (2020). Molecular evolutionary genetics analysis (MEGA) for macOS. Mol. Biol. Evol. *37*, 1237–1239.

Sui, J., Aird, D.R., Tamin, A., Murakami, A., Yan, M., Yammanuru, A., Jing, H., Kan, B., Liu, X., Zhu, Q., et al. (2008). Broadening of neutralization activity to directly block a dominant antibody-driven SARS-coronavirus evolution pathway. PLoS Pathog. e1000197. https://doi.org/10.1371/journal.ppat.1000197.

Tamura, K., and Kumar, S. (2002). Evolutionary distance estimation under heterogeneous substitution pattern among lineages. Mol. Biol. Evol. *19*, 1727–1736.

Tamura, K., Nei, M., and Kumar, S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc. Natl. Acad. Sci. *101*, 11030–11035.

Tegally, H.W.E., Lessells, R.J., Giandhari, J., Pillay, S., Msomi, N., Mlisana, K., Bhiman, J.N., Von Gottberg, A., Walaza, S., Fonseca, V., et al. (2021). Sixteen novel lineages of SARS-CoV-2 in South Africa. Nat. Med. *27*, 440–446.

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E.J., and Msomi, N. (2021). Emergence of a SARS-CoV-2 variant of concern with mutations in spike glycoprotein. Nature, 1–8.

Tessema, S.K., Inzaule, S.C., Christoffels, A., Kebede, Y., De Oliveira, T., Ouma, A.E.O., Happi, C.T., and Nkengasong, J.N. (2020). Accelerating genomics-based surveillance for COVID-19 response in Africa. The Lancet Microbe *1*, e227–e228.

Tönnesmann, E., Kandolf, R., and Lewalter, T. (2013). Chloroquine cardiomyopathy–a review of the literature. Immunopharmacol. Immunotoxicol. *35*, 434–442.

Toyoshima, Y., Nemoto, K., Matsumoto, S., Nakamura, Y., and Kiyotani, K. (2020). SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. J. Hum. Genet. *65*, 1075–1082.

Volz, E., Mishra, S., Chand, M., Barrett, J.C., Johnson, R., Geidelberg, L., Hinsley, W.R., Laydon, D.J., Dabrera, G., and O'toole, Á. (2021). Transmission of SARS-CoV-2 Lineage B. 1.1. 7 in England: insights from linking epidemiological and genetic data. Preprint at medRxiv. https://doi.org/10.1101/2020.12.30.20249034.

Wan, Y., Shang, J., Graham, R., Baric, R.S., and Li, F. (2020). Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. J. Virol. *94*, e00127-20.

Wang, C., Horby, P.W., Hayden, F.G., and Gao, G.F. (2020a). A novel coronavirus outbreak of global health concern. The lancet *395*, 470–473.

Wang, M., Cao, R., Zhang, L., Yang, X., Liu, J., Xu, M., Shi, Z., Hu, Z., Zhong, W., and Xiao, G. (2020b). Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. Cell Res. *30*, 269–271. https://doi.org/10.1038/s41422-020-0282-0.

Watanabe, Y., Bowden, T.A., Wilson, I.A., and Crispin, M. (2019). Exploitation of glycosylation in enveloped virus pathobiology. Biochim. Biophys. Acta *1863*, 1480–1497.

Watanabe, Y., Allen, J.D., Wrapp, D., Mclellan, J.S., and Crispin, M. (2020). Site-specific glycan analysis of the SARS-CoV-2 spike. Science *369*, 330–333. https://doi.org/10.1126/science.abb9983.

Wibmer, C.K., Ayres, F., Hermanus, T., Madzivhandila, M., Kgagudi, P., Oosthuysen, B., Lambson, B.E., De Oliveira, T., Vermeulen, M., and Van Der Berg, K. (2021). SARS-CoV-2 501Y. V2 escapes neutralization by South African COVID-19 donor plasma. Nat. Med. 1–4.

Wickham, H., and Bryan, J. (2019). Readxl: Read Excel Files. R package version, 1.

Wickham, H., and Henry, L. (2020). Tidyr: Tidy Messy Data. R package version 1.0. 2.

Wickham, H., Averick, M., Bryan, J., Chang, W., Mcgowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. J. Open Source Softw. *4*, 1686. https://doi.org/10.21105/joss.01686.

Wickham, H., Hester, J., and Francois, R. (2016). Readr: Read Tabular Data. R Package Version 1.0.0. https://CRAN.R-project.org/package=readr.

Wickham, H., Romain, F., Henry, L., and Muller, K. (2018). Dplyr: A Grammar of Data Manipulation. R package version 0.7.6.

Wilkinson, E., Giovanetti, M., Tegally, H., San, J.E., Lessells, R., Cuadros, D., Martin, D.P., Rasmussen, D.A., Zekri, A.N., Sangare, A.K., et al. (2021). A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. Science *374*, 423–431. https://doi.org/10.1126/science abj4336.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., and Pei, Y.-Y. (2020). A new coronavirus associated with human respiratory disease in China. Nature *579*, 265–269.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q.V. (2019). Xlnet: generalized autoregressive pretraining for language understanding, preprint at. arXiv, 1906.08237.

Yao, X., Ye, F., Zhang, M., Cui, C., Huang, B., Niu, P., Liu, X., Zhao, L., Dong, E., and Song, C. (2020). In vitro antiviral activity and projection of optimized dosing design of hydroxychloroquine for the treatment of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Clin. Infect. Dis. *71*, 732–739.

Zaki, A.M., Van Boheemen, S., Bestebroer, T.M., Osterhaus, A.D., and Fouchier, R.A. (2012). Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. New Engl. J. Med. *367*, 1814–1820.

Zhang, Y.-Z., and Holmes, E.C. (2020). A genomic perspective on the origin and emergence of SARS-CoV-2. Cell *181*, 223–227.

Zhao, P., Praissman, J.L., Grant, O.C., Cai, Y., Xiao, T., Rosenbalm, K.E., Aoki, K., Kellman, B.P., Bridger, R., and Barouch, D.H. (2020). Virus-receptor interactions of glycosylated SARS-CoV-2 spike and human ACE2 receptor. Cell Host Microbe *28*, 586–601 e6.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., and Lu, R. (2020). A novel coronavirus from patients with pneumonia in China. New Engl. J. Med. *382*, 727–733.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
R	R Core Team 2019. R: A Language and Environment for Statistical Computing	https://cran.r-project.org/
Biostrings R package	Pagès et al., (2021)	https://bioconductor.org/packages/release/bioc/html/Biostrings.html; R package version 2.62.0
maptools R package	Bivand et al., (2021)	https://cran.r-project.org/web/packages/maptools/index.html; Package 'maptools'
RColorBrewer R package	Neuwirth, 2014	https://cran.r-project.org/web/packages/RColorBrewer/index.html; RColorBrewer: ColorBrewer palettes. R package version 1.1-2.
maps R package	Becker et al., (2018)	https://cran.r-project.org/web/packages/maps/index.html; maps: Draw Geographical Maps. R package version 3.3. 0.
mapdata R package	Becker et al., (2018)	https://cran.r-project.org/web/packages/mapdata/index.html; maps: Draw Geographical Maps. R package version 3.3. 0.
readxl R package	Wickham and Bryan, 2019	https://cran.r-project.org/web/packages/readxl/; readxl: Read excel files. *R package version*, 1.
ggplot2 R package	Kahle and Wickham, 2013	https://cran.r-project.org/web/packages/ggplot2/index.html
dplyr R package	Wickham and Henry, 2020; Wickham et al., 2018	https://cran.r-project.org/web/packages/dplyr/index.html; R package version 0.7.6
gridExtra R package	Auguie, 2017	https://cran.r-project.org/web/packages/gridExtra/index.html; (R package version 2.3)
ggcorrplot R package	Kassambara, 2019	https://cran.r-project.org/web/packages/ggcorrplot/
ggpubr R package	Kassambara, 2020	https://cran.r-project.org/web/packages/ggpubr/index.html
ggmap R package	Kahle and Wickham, 2013	https://cran.r-project.org/web/packages/ggmap/index.html
mapproj R package	Mcilroy et al., 2020	https://cran.r-project.org/web/packages/mapproj/index.html
rio R package	Chan et al., (2021).	https://cran.r-project.org/web/packages/rio/index.html
tidyverse R package	Wickham et al., (2019).	https://cran.r-project.org/web/packages/tidyverse/index.html
readr R package	Wickham et al., (2016).	https://CRAN.R-project.org/package=readr
graphics R package	R Core Team (2019).	https://www.R-project.org/
sm R package	Bowman and Azzalini, (2018).	https://cran.r-project.org/web/packages/sm/index.html
Lubridate	Grolemund and Wickham (2011).	https://cran.r-project.org/web/packages/lubridate/index.html
Aweek	Kamvar (2021)	https://cran.r-project.org/web/packages/aweek/index.html
vioplot R package	Adler and Kelly, 2020	https://cran.r-project.org/web/packages/vioplot/index.html; https://github.com/TomKellyGenetics/vioplot
MAFFT	Katoh et al., 2019	https://mafft.cbrc.jp/alignment/server/add_fragments.html?frommanualnov6
Aliview	Larsson, 2014	https://ormbunkar.se/aliview/
Figtree	Not applicable	https://github.com/rambaut/figtree/releases
TempEst	Rambaut. et al.,. 2016	http://tree.bio.ed.ac.uk/software/tempest/
IQTree	Nguyen et al., 2015; Hoang et al., 2018	http://www.iqtree.org/
GLAM2	Frith et al., 2008	http://meme-suite.org/tools/glam2

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Tomtom	Shobhit Gupta, JA Stamatoyannopolous, Timothy Bailey and William Stafford Noble, 2007	http://meme-suite.org/tools/tomtom
Data and code	This study	https://github.com/Yinkaokoh/updatedSARCoV2_project https://doi.org/10.17632/bczg8z7yg2.1
Other		
Sequence data from GISAID	Elbe and Buckland-Merrett, 2017; Shu and Mccauley, 2017	https://www.gisaid.org/
GISAID database authors and laboratories	This study	Data S5

RESOURCE AVAILABILITY

Lead contact

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request, Dr. Nídia S. Trovão (nidia.trovao@nih.gov; nidiastrovao@gmail.com).

Materials availability

This study did not generate new unique reagents or genetic sequences.

Data and code availability

- The accession numbers of sequences used in the study and respective GISAID acknowledgment table are provided in Data S5.
- All original code used in the manuscript has been deposited at: https://github.com/Yinkaokoh/updatedSARCoV2_project. Additional supplemental items are available from Mendeley Data at: https://doi.org/10.17632/bczg8z7yg2.1.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

METHOD DETAILS

Compilation of genomic datasets

Assembly of study dataset. SARS-CoV-2 genome sequences collected from Africa were obtained from the GISAID database (https://www.gisaid.org/) (Elbe and Buckland-Merrett, 2017) (Shu and Mccauley, 2017) on January 7, 2021, by only selecting complete genomes and excluding those with low coverage. As of January 7, 2021, nearly eleven months since the first case was reported in Africa, a total of 5229 SARS-CoV-2 complete genome sequences from 33 African countries were available in the GISAID database. The sequences were aligned using the online version of the MAFFT (Katoh et al., 2019) multiple sequence alignment tool hosted at https://mafft.cbrc.jp/alignment/software/closelyrelatedviralgenomes.html, with the Wuhan-Hu-1 (www.ncbi.nlm.nih.gov/nuccore/MN908947.3) as the reference sequence. The aligned sequences were manually edited and cleaned in AliView version 1.26 (Larsson, 2014), by excluding sequences with fewer than 75% unambiguous bases, and trimming the alignment at the 5′ and 3′ ends. We also removed duplicate sequences defined as those having identical nucleotide composition and having been collected on the same date and in the same country. This dataset was subjected to multiple iterations of phylogeny reconstruction using IQ-TREE multicore software version v1.6.12 (Nguyen et al., 2015) with parameters -m GTR+G -nt 50. TempEst (Rambaut et al., 2016) was used to exclude outlier sequences whose genetic divergence and sampling date were incongruent, resulting in a dataset with 2,414 sequences with 29,796 nucleotide base pairs.

Selecting a genomic background dataset. We used the Pango lineage classification available in the metadata associated with the sequences to identify the lineages circulating in Africa, as this nomenclature system is designed to integrate both genetic and geographical information about SARS-CoV-2 dynamics

(O'Toole et al., in prep). In total, there were 143 Pango lineages circulating in Africa, as listed in Data S4. On January 7, 2021, we obtained from GISAID all available sequences belonging to lineages that circulate in Africa. A similar approach to that described above (including alignment using MAFFT, manual inspection using AliView, phylogenetic tree reconstruction with IQ-TREE and exclusion of root-to-tip outliers using TempEst) was employed, resulting in a dataset with 5002 sequences with 29,796 nucleotide base pairs.

Phylogenetic inference

We merged the study and background datasets, resulting in a final dataset with 7,416 sequences (Data S5). We computed the phylogeny with ultrafast bootstraps using IQ-TREE v1.6.12 (Hoang et al., 2018; Nguyen et al., 2015) with parameters -m GTR+G -bb 1,000 -bnni -nt 50. These analyses were conducted using the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health (Bethesda, MD, USA) (http://biowulf.nih.gov). Trees were rooted in the Wuhan-Hu-1 (GenBank: MN908947.3) reference genome and visualized in FigTree version 1.4.4.

Comparison of evolutionary divergence

We estimated the evolutionary divergence of several sequence datasets from each continent. Each continent-specific dataset consisted of sequences from the final dataset described above. The final continent-specific datasets were as follows: Africa (n = 2,414); Asia (n = 1,008); Europe (n = 999); North America (n = 997); Oceania (n = 1,000); South America (n = 998).

To estimate the evolutionary divergence, we calculated the pairwise distance (in base substitutions per site) between all pairs of sequences within and between each continent. We conducted the analyses using the Molecular Evolutionary Genetics Analysis software version 10 (MEGA X) (Kumar et al., 2018; Stecher et al., 2020) and applied the maximum composite likelihood model (Tamura et al., 2004). The rate variation among sites was modeled with a gamma distribution (shape parameter = 4), and the differences in the composition bias among sequences were considered in evolutionary comparisons (Tamura and Kumar, 2002). We included 1st+2nd+3rd+non-coding codon positions, and all with less than 50% site coverage due to alignment gaps, missing data, and ambiguous bases, were eliminated (partial deletion option). R 4.0.3 software (R CoreTeam, 2019) was used for the visualization. The pairwise genetic distances were summarized and plotted using scripts designed in R 4.0.3 software (R CoreTeam, 2019). The R packages used were rio (Chan et al., 2021), tidyverse (Wickham et al., 2019), readr (Wickham et al., 2016), graphics (R Core Team, 2019), sm (Bowman and Azzalini, 2018), vioplot (Adler and Kelly, 2020), gridExtra (Auguie, 2017), and ggplot2 (Kahle and Wickham, 2013) [ref].

Geographical distribution of COVID-19 pandemic in Africa

SARS-CoV-2 genetic data was collected from GISAID as described above and epidemiological data was obtained from OurWorldInData.org (Roser et al., 2020) on January 8, 2021. One sequence did not have a clade assignment and was excluded from the statistics. We developed scripts for the statistical analysis using R 4.0.3 software (R CoreTeam, 2019). The R packages used were maptools (Bivand et al., 2021), RColorBrewer (Neuwirth, 2014), maps (Becker et al., 2018), mapdata (Becker et al., 2018), readxl (Wickham and Bryan, 2019), ggplot2 (Kahle and Wickham, 2013) (Kassambara, 2019), dplyr (Wickham et al., 2018), gridExtra (Auguie, 2017), ggcorrplot (Kassambara, 2019), ggpubr (Kassambara, 2020), ggmap (Kahle and Wickham, 2013), lubridate (Grolemund G et al., 2011), aweek (Kamvar, 2021), and mapproj (Mcilroy et al., 2020). The data and the R script for the analysis can be accessed at https://github.com/Yinkaokoh/updatedSARCoV2_project.

Detection of repeat patterns and motifs

The retrieved SARS-CoV-2 sequences from Africa were annotated using GLAM2 (http://meme-suite.org/tools/glam2). GLAM2 is a deletion and motif finding software for either nucleotide or amino acid sequences (Frith et al., 2008). The Wuhan isolate with accession number GenBank: NC_045512.2 was annotated for novel motifs, and the Biostrings R package from Bioconductor (Pagès et al., 2021) was used to find the motifs' appearance in the retrieved African SARS-CoV-2 sequences. The Tomtom tool (http://meme-suite.org/tools/tomtom) (Gupta et al., 2007) [, which equates one or more motifs against a database of known motifs, was employed to find overlapping positions across the motif database (Gupta et al., 2007).

QUANTIFICATION AND STATISTICAL ANALYSIS

To investigate how the viral genetic diversity circulating in Africa compares to that circulating in other continents we employed a Wilcoxon signed-rank test on the within-continent pairwise distances, using R 4.0.3 software (R CoreTeam, 2019). For each continent there are 2.912.491 estimates of within-continent pairwaise distances for the Africa set, 507.528 for the Asia set, 498.501 for the Europe set, 496.506 for the North America set, 499.500 for the Oceania set, and 497.503 for the South America set. We did not use any methods to determine whether the data met athe ssumptions of the statistical approach.