Epidemiological and phylogenetic analyses of COVID-19 in Africa using open-source sequence data.

SARS-CoV-2 genetic epidemiology in Africa.

Nwachukwu, Chigozie John[1*]; MacIntyre, C. Raina[1]; Stone, Haley [1].

[1]School of Population Health, University of New South Wales, Sydney, Australia

[*]Corresponding author

z5237971@zmail.unsw.edu.au

Summary

Between late December 2019 to early September 2020, over 10 million people globally were reportedly infected by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), responsible for the coronavirus disease-2019 (COVID-19). In Africa, more than 300,000 infection occurred within the period, from which several viral genetic sequences were generated. Phylogenetic reconstruction of genomic data can provide epidemiological inferences about time of pathogen introduction, epidemic growth rate and temporal-spatial spread of the infection during disease outbreak. In this work, we studied the genetic epidemiology of COVID-19 in Africa. Genetic sequence data of SARS-CoV-2 and metadata from African countries were obtained from open-source sequence database hosted by the GISAID initiative. Whole genome sequences were subjected to multiple sequence alignment, from which Maximum Likelihood phylogenetic tree was constructed based on the general time reversible model. Of the 227 genetic sequences obtained for 9 African countries (DRC=133, Senegal=23, South Africa=20, Ghana=15, Tunisia=6, Algeria=3, Gambia=3, Egypt=2 and Nigeria=2), 220 were whole genome sequences while 7 were partial genome sequences of the surface glycoprotein S. Phylogenetic analysis confirmed multiple introductions of the virus to the continent from multiple external sources prior to local adaptation and spread. The very close alignment of three viruses - *Ghana/1659_S14/2020|EPI_ISL_422405*, *DRC/KN0054/2020|EPI_ISL_417437*, *and South_Africa/R05475/2020|EPI_ISL_435059* – to the reference Wuhan strain on the time tree, suggests possible introduction and circulation of the virus into the continent much earlier than when the first case was announced on February 15 2020. In conclusion, this study provided evidence to support multiple introductions of SARS-CoV-2 into Africa, and further suggests that the virus may have already been circulating in the continent

prior to official reporting of the first case. Also, there is strong impression to infer likely genetic adaptation of the virus in the continent that may account for the close clustering of isolates from different countries.

Key words: Africa, Coronavirus, COVID-19, Humans, Molecular Epidemiology, Phylogenetics.

Introduction

In late December 2019, the World Health Organisation (WHO) was notified of cluster of cases of pneumonia of unknown aetiology in the Hubei province of China (WHO, 2020e). They were later confirmed to be caused by a novel coronavirus, severe acute respiratory syndrome coronaviruses-2 (SARS-CoV-2) (Gorbalenya et al., 2020) and the associated disease named coronavirus disease-2019 [COVID-19] (WHO, 2020d). Up till early September 2020, over 10 million infections have been reported globally (WHO, 2020c). The first case of COVID-19 in Africa was reported on 15 February 2020 in Egypt, and as of ending of June 2020, Africa has recorded over 300,000 cases (WHO, 2020c). As of early September 2020, a total of 6,155 COVID-19 associated deaths has been recorded in Africa with over 150,000 recoveries. With a global case-fatality ratio (CFR) of 5.0%, Africa's contribution to worldwide COVID-19 associated death of 503,862 stood at 1.2%.

As efforts are ongoing to contain the pandemic, researchers have been working assiduously to understand various characteristics of the virus including its pathogenicity, immunogenicity, pathobiology, and epidemiological characteristics. Early studies on the clinico-epidemiological characteristics of the virus reported mean incubation period of 5.2 days, ranging up to 12.5 days in 95th percentile of the distribution (Li et al., 2020). This finding formed the basis of the

current recommendation of 14 days isolation period for exposed persons. Cases in Africa however presented with some characteristics that are different from the global outlook. The disease has been reported in all age groups in the continent, with a male to female ratio of 1.7:1 and median age of 36 years among affected cases (WHO, 2020c).

In terms of geographical spread, COVID-19 has been reported from 55 African countries and territories (including Mayotte), with South Africa having the highest number of cases in sub-Sahara Africa with total confirmed cases of over 350,000 and 4,948 deaths as of end of June 2020 (WHO, 2020b). According to the Africa Centre for Disease Control and Prevention (Africa CDC), eight countries in Africa are reporting case fatality rate that is similar or higher than the 4.2% obtained globally. These include: Chad (8.4%), Liberia (6.3%), Sudan (6.3%), Niger (6.2%), Burkina Faso (5.0%), Egypt (4.9%), Mali (4.9%) and Algeria (4.7%) (Africa CDC, 2020). Contrary to situation seen in past epidemics, the continent demonstrated an improved laboratory diagnostic capacity in response to COVID-19, with ability for local viral sequencing in several countries. This is an improvement from situations in years past whereby external collaboration is usually needed for definitive diagnosis of a disease outbreak (Makoni, 2020). One outcome of this situation is the array of SARS-CoV-2 partial and whole-genome sequence data that has been submitted to the public database of the GISAID Initiative (formerly Global Initiative on Sharing All Influenza Data) (Shu & McCauley, 2017) by scientists from various laboratories in Africa. Analysing these sequence data is expected to furnish information on the genetic epidemiology of the SARS-CoV-2 infection in Africa.

Molecular or genetic epidemiology is concerned with how genomic, genetic, and other molecular attributes contribute to the aetiology of a disease, its distribution and approach to prevention (Llanes et al., 2020). Through the analysis of molecular data, inferences could be

made on the rate of evolution of a viral pathogen, the evolutionary changes influencing host and tissue specificity, and the pattern of transmission within human population. An important tool in achieving this is phylogenetic studies. Phylogenetic analysis is a critical tool for understanding the historical evolution of viruses and serves as the basic building block for their classification (Gorbalenya et al., 2020; Llanes et al., 2020). They are usually carried out on conserved genes or genomic segment of an organism with adequate sequence divergence to enable their clear delineation on a phylogenetic tree. In the case of coronaviruses, whole or partial sequences of the Orf1ab, S and N genes are often deemed adequate for this analysis.

To inform public health intervention, genome sequence data requires rapid analyses and wider dissemination of the results obtained. Since the identification of the aetiological agent of COVID-19 and curation of sequence data on GISAID, large scale phylogenetic analyses of the sequences have been undertaken by NextStrain (Hadfield et al., 2018), an open-source project aimed at harnessing "the scientific and public health potential of pathogen genome data" (Bedford et al., 2020). In their last situation report, NextStrain noted some patterns displayed by the COVID-19 pandemic based on analyses conducted. Inferences from the analysis showed that, firstly, outbreaks in far-flung geographical locations are intertwined. Secondly, there has been multiple introductions of the virus to communities through migration and human travel. Thirdly, not all introduced variants result in local spread while few others spawn local transmission. And lastly, once local transmission is established, these send off their own sparks into other communities (Nextstrain, 2020). Thus, the interest of the current work is to understand how outbreaks in Africa which were seeded from other continents, especially Europe (Makoni, 2020 ; Oluniyi, 2020), have interacted to produce the pattern currently predominant in Africa. This will help in mounting robust surveillance system that could alert to genetic changes or trans-continental introduction of new variant.

Relying on two main sources of publicly available data, our objective was to compare the genetic epidemiology of COVID-19 infection in African countries through phylogenetic studies. Our result showed that the data curated by the COVID-19 Data Working Group (Xu et al., 2020) was inadequate for making inferences on the general descriptive epidemiology of COVID-19 in Africa on one hand. On the other hand, with exception to wide disparities in the numbers of SARS-CoV-2 sequences submitted from few African countries to the GISAID database (GISAID, 2020), the quality of the data enabled inferential deductions that may reflect the epidemiological pattern of the COVID-19 pandemic in Africa.

Materials and methods

Source of data and data retrieval

Data for this study were obtained from two main sources – GISAID database (GISAID, 2020) and Open COVID-19 Data Working Group Repository (Xu et al., 2020) (detailed acknowledgment available in table S1), supplemented by COVID-19 infection data from WHO-OCHA (WHO-OCHA, 2020) and population data from the World Bank's Health Nutrition and Population Statistics database (World Bank, 2020). The GISAID database is a publicly available open-access platform for submission of genome sequence data. The COVID-19 epidemiological line list is a centralised repository of individual-level information on patients with laboratory-confirmed COVID-19. The data are openly available (Xu et al., 2020), and a live version of the data record, which is continually updated, can be downloaded from the group's github repository.

On 27 May 2020, the GISAID database was accessed and the EpiCoV tab selected. On the browser pane, Africa was typed in to filter the data to display only entries from Africa, excluding contents from other parts of the world. The 227 sequences from Africa as of that

125 date were downloaded manually as individual files in *Fasta* format. The downloaded files were
126 then imported into Geneious prime® (*Geneious version 2020.2 created by Biomatters. Available*
127 *from* https://www.geneious.com) for editing. Also, the reference strain,
128 *China/WHU01/2020/EPI_ISL_406716* was downloaded separately and imported into the
129 Geneious® software.

130 Similarly, latest data up to 17 July 2020 on global COVID-19 epidemiological line list was
131 downloaded from the github repository of the COVID-19 Data Working Group. This contained
132 data up till end of May 2020. These were filtered to select only for African countries and data
133 for African countries were extracted into a separate Excel sheet for analysis.

134 Epidemiological analysis on line-list data

135 The data on the line list was analysed for count of cases and number of cases per country.
136 However, age and gender distribution could not be meaningfully estimated because these data
137 points are not available in several of the line list entries. Data from this source were used to
138 plot the number of daily cases in Africa up to the end of May 2020 and the number of cases per
139 country. The results were compared to WHO-OCHA daily cases report from African countries
140 for the period up to April 30, 2020.

141 Updating and analysing sequence metadata

142 The manually downloaded sequence data contained only the nucleotide bases without
143 metadata. For each sequence, metadata including age, sex, location, and clades were updated
144 in Geneious prime® (*Geneious version 2020.2 created by Biomatters. Available from*
145 https://www.geneious.com) and on Excel file by triangulating back to the GISAID database. The

age and sex distribution of the sequence data were analysed and the count of lineages and clades from each country annotated.

Sequence alignment

Sequences were manually filtered and partial gene sequences with sequence size of 1kb were excluded. Whole genome sequences with average sequence length of 29,000 to 30,000 were selected and aligned with the MUSCLE algorithm (Edgar, 2004) in MEGA version X software (Kumar, Stecher, Li, Knyaz, & Tamura, 2018) using default settings.

Building of phylogenetic tree

In MEGA version X software (Kumar et al., 2018), the aligned sequences were used to estimate the tree by the maximum likelihood (ML) model in which gaps and missing data were set for complete deletion and Neighbour-joining (NJ) tree specified as preferred tree to use. The General Time Reversible (GTR) model (Nei & Kumar, 2000) was used to fit the tree. By applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances obtained from the Maximum Composite Likelihood (MCL) approach, initial tree(s) for heuristic search were obtained automatically and the topology with superior log likelihood value was selected. Evolutionary rate differences among sites [5 categories (+G parameter=200.0)] was modelled by discrete Gamma distribution. The tree was drawn to scale, with branch lengths measured in the number of substitutions per site and the tree with the highest log likehood (-586456.26) was displayed.

Tree visualisation and editing

The phylogenetic tree obtained from MEGA version X (Kumar et al., 2018) was exported in Newick tree format into FigTree version 1.4.4 (Rambaut, 2018) where it was edited and time scale added to infer the diversity time based on the length of the tree.

Results

Cumulative number of daily cases of COVID-19 in Africa up to end of May 2020 and number of cases per country.

Based on the data obtained from the Open COVID-19 Data Working Group, we estimated the number of daily cases of COVID-19 in Africa up to the period (end of May 2020) that the data obtained on 17 July 2020 was curated. The number of cases per country was also estimated from the data. This was compared to the number of daily cases collated by the WHO-OCHA (WHO-OCHA, 2020) based on data downloaded on 19 July 2020 that were updated up to end of May 2020.

Table 1 shows the cumulative number of cases per country based on the data from the earlier mentioned sources while the number of daily cases is presented in figure 1 below. The total number of cases of COVID-19 in Africa as of July 3, 2020 is 142,245 based on the WHO-OCHA data. The top 5 countries with highest number of cases are South Africa, Egypt, Nigeria, Algeria, Ghana, and Morocco with 20.7%, 20.1%, 7.9%, 6.8% and 6.0% of all the cases in the continent, respectively. However, estimation of infection rate show that rate of infection is highest in the horn of Africa country of Djibouti, with an infection rate of 388.2 per 100,000 population (see table 1 and figure S1). There was gross under-reporting of number of cases in the data curated

by the COVID-19 Data Working Group when compared to the WHO-OCHA data. Apart from lack of data from some countries, few numbers were also reported for many countries when compared to the WHO-OCHA data. This may be due to the fact that the curated data were often from mixed official and non-official sources (Xu et al., 2020), and may be limited by internet access at the different countries.

From when the first case in Africa was reported on February 15, number of daily cases has progressively increased reaching a record level of over 4500 cases in one day on May 30 as reflected from the WHO-OCHA data [see figure 1.ii]. When viewed from the Open COVID-19 Data Working Group data, the number of daily cases appear to have reduced in the second week of May before picking up again [Figure 1.i]. This however was not the case as the WHO-OCHA data show unrelenting increase in the number of cases.

Number of countries submitting SARS-CoV-2 sequence data

Based on the sequence data from Africa available on GISAID database as of 27 May 2020, we calculated the number of sequences from each country out of the total 227 sequence data available as of that date. Results are shown in table 2 below. Only 9 countries from Africa have submitted SARS-CoV-2 genetic sequence data to GISAID as of the date of interest. The bulk of sequence data (n=133; ≈58.6%) were from the Democratic Republic of Congo (DRC) despite having lower number of cases compared to countries with very high cases at the period. Similarly, the country with second highest number of sequence data, Senegal (n=23; ≈10.1%), had fewer number of cases compared to the top five countries with highest number of cases at this period. This may reflect the molecular sequencing capacity that these countries have

developed over the years as part of their involvement in diagnosing other infectious diseases in the continent, especially Ebola in the DRC.

Age and gender distribution of cases for which genetic sequence are available

We estimated the median age of cases for which genetic sequences are available and the gender distribution of the cases. Based on the available data, the median age for all represented cases is 38 years (range: 3years to 87 years), of which median age for females is 38.5 years (range: 5 years to 75 years) and that for male is 42.5 years (range: 3 years to 87 years). For cases in which data are available on gender, 82 (42.9%) are female, while 109 (57.1%) are male. The age and gender distribution of the cases is shown in figure 2. Most of the available sequences are obtained from individuals aged between 30 to 50 years, reflecting the age group that are more infected with SARS-CoV-2 in Africa. Also, there are more sequence data for men. This supports the well-known observation that men are more affected by COVID-19 than women for obvious reasons.

Clade distribution of SARS-CoV-2 isolates from Africa

Based on the metadata on GISAID platform, we estimated the number of genetic sequences from Africa belonging to the different GISAID clades classification for SARS-CoV-2. As can be seen from figure 3 below, most of the SARS-CoV-2 sequenced in Africa belonged to clade G. The exceptions are Algeria and Egypt in which the available 3 and 2 sequences respectively belonged to clade GH, and Senegal where clade GH slightly outnumber clade G.

Output of sequence alignment and phylogenetic tree

The original 227 sequences were manually filtered and partial gene sequence with less than 29000bp were excluded. This now resulted in 221 sequences that were subjected to multiple sequence alignments (MSA). Result of MSA produced aligned sequences with Mean sequence length of 29783.9 (S.D=115.6) bp (Min. = 29302bp and Max. = 29903) with base coverage of 29.3%, 18%, 19.3%, and 32.1% for adenine (A), Cytosine (C), Guanine (G) and Thymidine (T) respectively. About 129,551 (2.0% of overall alignment) positions contained gaps.

Maximum likelihood statistics was applied to the aligned sequence to find the best DNA model with which to build a maximum likelihood tree. Results of 24 model fits were generated (see table S2) and model was selected based on the lowest BIC score (Bayesian Information Criterion) which is considered to optimally describe the best substitution pattern among the bases. All positions containing gaps and missing data were eliminated.

The evolutionary tree obtained (see figure 4 below) was inferred using the Maximum Likelihood method and General Time Reversible (GTR) model. Branch length of the tree is based on the number of substitutions per site. The tree reflected the period that Africa began to record cases of COVID-19 and the diversity that the virus has undergone within the continent. The earliest available sequences clustered closely to the reference Wuhan isolate, indicating multiple introductions of virus closely related to the reference strain into the continent from outbreak clusters outside the continent (see area A). However, there is also possibility that the virus was already circulating in the continent even before the first confirmed case was announced on February 15, 2020. This can be inferred from the very close affinity between three virus strains from the continent - *Ghana/1659_S14/2020|EPI_ISL_422405*, *DRC/KN0054/2020|EPI_ISL_417437*, and *South_Africa/R05475/2020|EPI_ISL_435059* from Ghana,

DRC, and South Africa respectively – with the Wuhan reference strain. The clustering of the virus at the longer tips of the tree reflects geographic adaptation of the virus to the new location and local circulation of adapted strains with less evidence of notable viral mutations. Given that most of the isolates are from DRC, there is a nodal cluster of viruses from DRC occupying a tree branch (see area B). This reflects ongoing local transmission within the country as the effect of lockdown imposed in most countries might not have allowed importation of cases from outside the country. Generally, the viral divergence has remained stable within the continent and within individual countries since the earlier multiple introductions from different sources from outside the continent.

Discussion

The COVID-19 pandemic has remained relentless since the world awoke to the knowledge of a novel strain of coronavirus in the twilight of the 2020 decade. Effort to control it has been met with mixed successes and failures, with countries hailed to have made commendable strides in tackling it at the initial stage latter facing waves of resurgent outbreaks (WHO, 2020a). Nevertheless, scientists in various fields are not relenting in their effort to provide information that could help both in controlling and stopping the pandemic (Seemann et al., 2020).

In this work, we have employed data generously made available by scientists from different laboratories around the world to the GISAID initiative to interrogate the genetic epidemiology of the novel virus in Africa. This is intended among others to understand how outbreaks in Africa, which were seeded from other continents, especially Europe (Makoni, 2020), have interacted to produce the pattern currently predominant in Africa. And help to understand viral

274 divergence occurring in a location that could be important for other locales in monitoring
275 imported cases. Although our analysis did not include viruses outside Africa, the result
276 obtained however provided certain insight into how the virus has adapted in Africa after
277 multiple introductions from outside the continent at the onset of the outbreak in the continent
278 (Githinji, 2020; Oluniyi, 2020).

279 Our result showed that the number of cases of COVID-19 has increased progressively since the
280 first case was detected in Egypt on February 15, 2020. The number of available sequence data
281 submitted to GISAID was however not proportionate to the number of cases in different
282 countries. Most of the sequence data on GISAID platform from Africa are from the DRC. This
283 could be a reflection of the genetic diagnostic and sequencing capability the country has
284 developed in its long battle with the Ebola virus (Palacios, 2018). Also, the median age of 38
285 years (range: 3years to 87 years) for cases in which there is sequence data is in agreement with
286 the median age of the infected of 36 years reported by the World Health Organisation (WHO,
287 2020c). Moreover, the preponderance of clade G in the sequences from Africa agrees with the
288 observation of O'Toole (2020) with respect to sequences from DRC.

289 Phylogenetic analysis of the sequence data support evidence for multiple introduction into
290 Africa of SARS-CoV-2 strains from different continents at the beginning of the pandemic in
291 Africa (Githinji, 2020; Makoni, 2020 ; Oluniyi, 2020). This is similar to observations in other
292 continents too (Adebalİ et al., 2020; Seemann et al., 2020). However, the affine clustering of some
293 viruses from Africa - *Ghana/1659_S14/2020|EPI_ISL_422405*, *DRC/KN0054/2020|*
294 *EPI_ISL_417437*, and *South_Africa/R05475/2020|EPI_ISL_435059* from Ghana, DRC and South
295 Africa respectively - to the reference Wuhan strain, could suggest an earlier circulation of the
296 virus in the continent prior to the reporting of the first case in February 15 2020. In any case,

we could not be highly confident in this conclusion as we do not have access to a comprehensive epidemiological data detailing the travel and medical history of the cases in which these viruses were isolated. Generally, our result shows that the viruses circulating in the continent have remained stable both within the continent and within individual countries since the earlier multiple introductions from different sources outside the continent.

There are certain limitations that could bias the conclusions of this study. Firstly, there is underreporting of cases of COVID-19 in Africa both from official and unofficial sources (Mbow et al., 2020). This is also reflected in one of our data sources. Secondly, we did not undertake critical evaluation of nucleotide changes in the viruses employed in the study. Thus, any inference to genetic diversity or viral adaptation should be made with caution. Thirdly, we did not include viruses from other continents apart from the reference Wuhan strain that was used as ancestral lineage. Therefore, we could not make categorical statement on the exact origin (s) of the virus circulating in Africa. Lastly, there appear to be some intra-continental mixing in the circulating strains. However, we could not make a conclusive statement on this because we do not have detailed information on travel and medical history of persons in which the viruses were isolated.

Conclusion and recommendation

The current study was conceptualised to evaluate the genetic epidemiology of COVID-19 infection in Africa by analysing sequence data submitted to GISAID database from laboratories in Africa. The original intention also included linking information on GISAID database with those contained in the data curated by the COVID-19 Open Data Working Group. The latter could not be done because certain key datapoints were lacking in the curated data. Also, the apparent

underreporting of the number of infections in Africa in the data from COVID-19 Open Data Working Group warranted a comparison with data from another source (WHO-OCHA). While the result of other demographically related data agrees with what has been reported elsewhere, the outcome of this study however provided evidence to support multiple introductions of SARS-CoV-2 into Africa, and further suggests that the virus may have already been circulating in the continent prior to official recording of the first case in the continent. Moreover, there is strong impression to infer certain genetic adaptation of the virus in the continent that has informed the close clustering of less distant isolates from the continent. Therefore, it is important for public health authorities to keep monitoring the genetic sequence data for early detection of unique mutations or external introduction of 'foreign strain' into the continent. Other researchers should undertake in-depth profiling of the genetic sequences to detect any nucleotide changes that may signal geographical adaptation. This will go a long way in developing interventions that are not generic but tailored to the need of the continent.

Acknowledgement

The authors are grateful to the following persons and institution for their assistance:

Xin Chen (Jessie), for tutorial assistance and support in producing phylogenetic tree.

Aye Moa of the Biosecurity Research Program in The Kirby Institute for her excellent organisation and coordinating capacity.

The Australia Awards Scholarship programme for their scholarship funding to CJN.

All the scientists and curators who contributed data to the GISAID database and the Open COVID-19 Data Working Group repository (details available in supplementary table S1).

Ethical consideration

We confirm that the ethical policies of the journal, as noted on the journal's author guidelines page, have been adhered to. No ethical approval was required for this work as de-identified data obtained from publicly accessible databases were used.

Data Availability

The data that support the findings of this study are available in github repository of the Open COVID-19 Data Working Group at https://github.com/beoutbreakprepared/nCoV2019/blob/master/latest_data/latestdata.csv, as well as genetic data available in the public domain of the GISAID initiative: GISAID's EpiFlu™ Database, www.gisaid.org. These were supplemented by COVID-19 infection data from WHO-OCHA available at https://data.humdata.org/dataset/covid19_africa_infected-recovered-deceased (WHO-OCHA, 2020) and population data from the World Bank's Health Nutrition and Population Statistics database available at https://data.worldbank.org/indicator/SP.POP.TOTL?locations=ZG&name_desc=false. (World Bank, 2020).

Conflicts of interests

The authors declare no conflict of interests.

References

Adebali, O., Bircan, A., Çirci, D., İşlek, B., KilinÇ, Z., SelÇuk, B., & Turhan, B. (2020). Phylogenetic analysis of SARS-CoV-2 genomes in Turkey. *Turkish journal of biology = Turk biyoloji dergisi*, *44*(3), 146-156. doi:10.3906/biy-2005-35

Africa CDC. (2020). *Outbreak Brief #27: Coronavirus Disease 2019 (COVID-19) Pandemic*. Retrieved from https://africacdc.org/download/outbreak-brief-27-covid-19-pandemic-21-july-2020/

Bedford, T., Neher, R., Hadfield, J., Hodcroft, E., Sibley, T., Huddleston, J., . . . Grubaugh, N. (2020). Nextstrain: Real-time tracking of pathogen evolution. Retrieved from https://nextstrain.org/

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792-1797. doi:10.1093/nar/gkh340

GISAID. (2020). EpiCoV™: Pandemic coronavirus causing COVID-19. Retrieved from https://www.epicov.org/epi3/frontend#30d146

Githinji, G. (2020). Introduction and local transmission of SARS-CoV-2 cases in Kenya. Retrieved from https://virological.org/t/introduction-and-local-transmission-of-sars-cov-2-cases-in-kenya/497

Gorbalenya, A. E., Baker, S. C., Baric, R. S., de Groot, R. J., Drosten, C., Gulyaeva, A. A., . . . Coronaviridae Study Group of the International Committee on Taxonomy of, V. (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology*, *5*(4), 536-544. doi:10.1038/s41564-020-0695-z

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., . . . Neher, R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, *34*(23), 4121-4123. doi:10.1093/bioinformatics/bty407

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution*, *35*, 1547 - 1549.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., . . . Feng, Z. (2020). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *New England Journal of Medicine*, *382*(13), 1199-1207. doi:10.1056/nejmoa2001316

Llanes, A., Restrepo, C. M., Caballero, Z., Rajeev, S., Kennedy, M. A., & Lleonart, R. (2020). Betacoronavirus Genomes: How Genomic Information has been Used to Deal with Past Outbreaks and the COVID-19 Pandemic. *International Journal of Molecular Sciences*, *21*(12), 4546. doi:10.3390/ijms21124546

Makoni, M. (2020). Africa Contributes SARS-CoV-2 Sequencing to COVID-19 Tracking. *THE SCIENTIST*. Retrieved from https://www.the-scientist.com/news-opinion/africa-contributes-sars-cov-2-sequencing-to-covid-19-tracking-67348

Mbow, M., Lell, B., Jochems, S. P., Cisse, B., Mboup, S., Dewals, B. G., . . . Yazdanbakhsh, M. (2020). COVID-19 in Africa: Dampening the storm? *Science*, *369*(6504), 624. doi:10.1126/science.abd3902

Nei, M., & Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. New York: Oxford University Press.

Nextstrain. (2020). Situation Report Hiatus. Retrieved from https://nextstrain.org/narratives/ncov/sit-rep/2020-05-15?n=2

O'Toole, Á. (2020). Phylogenetic analysis of SARS-CoV-2 in DRC. Retrieved from https://virological.org/t/phylogenetic-analysis-of-sars-cov-2-in-drc/528

Oluniyi, P. (2020, May 29 2020). SARS-CoV-2 Genomes from Nigeria Reveal Community Transmission, Multiple Virus Lineages and Spike Protein Mutation Associated with Higher Transmission and Pathogenicity. *Novel 2019 Coronavirus Genome Report*. Retrieved from https://virological.org/t/sars-cov-2-genomes-from-nigeria-reveal-community-transmission-multiple-virus-lineages-and-spike-protein-mutation-associated-with-higher-transmission-and-pathogenicity/494

Palacios, G. (2018). DRC-2018-Viral Genome Characterization. Retrieved from https://virological.org/t/drc-2018-viral-genome-characterization/230

Rambaut, A. (2018). FigTree v1.4.4 2006-2018. Retrieved from http://tree.bio.ed.ac.uk/

Seemann, T., Lane, C. R., Sherry, N. L., Duchene, S., Gonçalves da Silva, A., Caly, L., . . . Howden, B. P. (2020). Tracking the COVID-19 pandemic in Australia using genomics. *medRxiv* doi:10.1101/2020.05.12.20099929v1

Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*, *22*(13). doi:10.2807/1560-7917.es.2017.22.13.30494

WHO-OCHA. (2020). COVID-19 Africa: infected, recovered and diseased. Retrieved from https://data.humdata.org/dataset/covid19_africa_infected-recovered-deceased. https://data.humdata.org/dataset/covid19_africa_infected-recovered-deceased

WHO. (2020a). *Coronavirus disease 2019 (COVID-19) Situation Report – 89*. Retrieved from Geneva, Switzerland:

WHO. (2020b). *Coronavirus disease (COVID-19) Situation Report – 181*. Retrieved from Geneva, Switzerland: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200719-covid-19-sitrep-181.pdf?sfvrsn=82352496_2

WHO. (2020c). *COVID-19: Situation update for the WHO African Region - 1 July 2020*. Retrieved from Brazzavile:

WHO. (2020d). Naming the coronavirus disease (COVID-19) and the virus that causes it. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it

WHO. (2020e). *Novel Coronavirus (2019-nCoV) SITUATION REPORT - 1 21 JANUARY 2020* (1). Retrieved from Geneva, Switzerland: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf?sfvrsn=20a99c10_4

World Bank. (2020). Health Nutrition and Population Statistics. Retrieved from https://data.worldbank.org/indicator/SP.POP.TOTL?locations=ZG&name_desc=false. https://data.worldbank.org/indicator/SP.POP.TOTL?locations=ZG&name_desc=false

Xu, B., Kraemer, M. U. G., Xu, B., Gutierrez, B., Mekaru, S., Sewalk, K., . . . Kraemer, M. (2020). Open access epidemiological data from the COVID-19 outbreak. *The Lancet Infectious Diseases*. doi:10.1016/s1473-3099(20)30119-5

Tables

Figures

Appendices

Supporting information