Real Time Viral Sub-Strains Discovery in Emerging Infectious Disease Situation – The African Perspective

Moses Effiong Ekpenyong (mosesekpenyong@uniuyo.edu.ng)
University of Uyo https://orcid.org/0000-0001-6774-5259
Faith-Michael Uzoka
Mount Royal University
Mercy Edoho
University of Uyo
Udoinyang G. Inyang
University of Uyo
Ifiok J Udo
University of Uyo
Nseobong P. Uto
Saint Andrews University
Itemobong S. Ekaidem
University of Uyo
Anietie E. Moses
University of Uyo
Enoabasi D. Anwana
University of Uyo
Youchou M. Tatfeng
Niger Delta University
Geoffery Joseph
University of Uyo
Emmanuel A. Dan
University of Uyo

Research article

Keywords: cognitive mining, genome expression pattern, open-source framework, COVID-19, SARS-CoV-2 sub-strains diversity, infectious disease, intelligent prediction, self-organizing map, transmission pathway

Posted Date: January 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-147634/v1

Real Time Viral Sub-Strains Discovery in Emerging Infectious Disease Situation – The African Perspective

Moses Ekpenyong[1*], Faith-Michael Uzoka[2], Mercy Edoho[1], Udoinyang Inyang[1], Ifiok Udo[1], Nseobong Uto[3], Itemobong Ekaidem[4], Anietie Moses[4], Enoabasi Anwana[5], Youchou Tatfeng[6], Geoffery Joseph[1], Emmanuel Dan[1], Juliana Ndunagu[7]

[1]Department of Computer Science, University of Uyo, Nigeria
[2]Department of Mathematics and Computing, Mount Royal University, Canada
[3]School of Mathematics and Statistics, University of Saint Andrews, United Kingdom
[4]College of Health Sciences, University of Uyo, Nigeria
[5]Department of Botany and Ecological Studies, University of Uyo, Nigeria
[6]College of Health Sciences, Niger Delta University, Nigeria
[7]Department of Computer Science, National Open University, Nigeria

* – Corresponding Author

Abstract

Background: The increased number of accessible genomes has prompted large-scale comparative studies for decerning evolutionary knowledge of infectious diseases, but challenges such as non-availability of close reference sequence(s), incompletely assembled or large number of genomes, preclude real time multiple sequence alignment and sub-strain(s) discovery. This paper introduces a cooperatively inspired open-source framework, for intelligent mining of severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) genomes. We situate this study within the African context, to drive advancement on state-of-the-art, towards intelligent infectious disease characterization and prediction. The outcome is an enriched Knowledge Base, sufficient to provide deep understanding of the viral sub-strains' identification problem. We also open investigation by gender, which to the best of our knowledge has been ignored in related research. Data for the study came from the Global Initiative on Sharing All Influenza Data database (https://gisaid.org) and processed for precise discovery of viral sub-strains transmission between and within African countries. To localize the transmission

route(s) of each isolate excavated and provide appropriate links to similar isolate strain(s), a cognitive solution was imposed on the genome expression patterns discovered by unsupervised self-organizing map (SOM) component planes visualization. The Freidman-Nemenyi's test was finally performed to validate our claim.

Results: Evidence of inter- and intra-genome diversity was noticed. While some isolates (or genomes) clustered differently, implying different evolutionary source (or high-diversity), others clustered closely together, indicating similar evolutionary source (or less-diversity). SOM component planes analysis revealed multiple sub-strains patterns, strongly suggesting local- or intra-community and country to country transmissions. Cognitive maps of both male and female isolates revealed multiple transmission routes. Freidman's test results showed highly significant difference ($p<0.01$) among the various isolate groups. Nemenyi's test revealed groups that differed in their isolates.

Conclusion: The proposed framework offers explanations to SARS-CoV-2 diversity and provides real time identification to disease transmission routes, as well as rapid decision support for facilitating inter- and intra-country contact tracing of infected case(s). Intermediate data produced in this paper are helpful to enrich the genome datasets for intelligent characterization and prediction of COVID-19 and related pandemics, as well as the construction of intelligent device for accurate infectious disease monitoring.

Key words: cognitive mining, genome expression pattern, open-source framework, COVID-19, SARS-CoV-2 sub-strains diversity, infectious disease, intelligent prediction, self-organizing map, transmission pathway.

Background

The coronavirus disease 2019 (COVID-19)–the disease caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), is a data-driven pandemic, because massive data and information are constantly being released and shared at very unprecedented scale. Artificial Intelligence (AI) has recently aided this process, as a growing amount of data and tools are constantly being explored to repurpose traditional approaches toward intelligent techniques, for early disease detection, real time contact tracing, new diagnostic tools development, informed policy formulation and implementation, and swift public health response, recovery and resilience. Taxonomically, SARS-CoV is one of the 36 coronaviruses in the family Coronaviridae within the order Nidovirales. Members of this family are known to cause respiratory or intestinal infections in humans and

other animals. However, despite a marked degree of phylogenetic divergence from other known coronaviruses, SARS-CoV together with bat SARS-CoV are now considered group 2b beta-coronavirus[1]. It has since been confirmed that two strains (the L- and S-strains) of the new coronavirus are spreading around the world today[2], and the fact that the L-type is more prevalent suggests that it is "more aggressive" than the S-strain. To date, the most widely spread is the G-strain, while the L- and S-strains are fast disappearing[3]. Although a lot is yet to be understood about SARS-CoV-2, we know that it is among the six known species that can infect humans, with one of the species assuming two different genetic variants, making a total of seven species. Out of these species, four produce generally mild symptoms, while the remaining three produce potentially severe symptoms[4].

A greater proportion of research progress on SARS-CoV-2 utilize the biotechnology dimension[5,6], strictly focusing on species characterization and variants analysis through features extraction. However, AI and Machine Learning (ML) methods are expanding biotechnology capacity into the bioinformatics realm, through further features probing for precise classification and prediction. Nonetheless, most of these methods lack cooperatively inspired flavors to drive intelligent solutions to the problem. In general, AI/ML research on SARS-CoV-2 has permeated four key areas of healthcare services, discussed within the African context as follows:

Screening and treatment–Real time reverse transcriptase polymerase chain reaction (RT-PCR) appears the "gold standard" for detecting SARS-COV-2 in Africa[7], but prolonged latency is associated with this test from when samples are collected to when they are processed, for results to be made available. Although serological tests that detect specific antibodies to the virus appear a quicker alternative to the RT-PCR test, they are not completely free of limitations[8]. Recently, several rapid diagnostic tests (RDTs) have been developed and tested in single studies, but none has so far been validated or commercially available.

Contact tracing–Placing contact tracing in the hands of all is certain to enhance contact identification[9]. Model-based approaches[10] as well as smartphone-enabled approaches[11], constitute some recent methods for managing the COVID-19 spread. These methods are cost-effective, given the pressure on healthcare systems in a global pandemic, but however, incipient in Africa.

Prediction and forecasting–Combination of clinical indicators are known to predict symptoms of infected persons. Hence, studies exploit predictive and forecasting systems[12-14], to effectively manage the spread/transmission of SARS-CoV-2. Cabore et al.[15] combined virus transmission characteristics and country specific socioecological factors to predict the most likely outcome of widespread and sustained community transmission of SARS-CoV-2. Using Markov chain model, with the transition states and country specific

probabilities derived from available knowledge, they found that sparsely populated cities, including Algeria, South Africa and Cameroon showed high risks of exposure. Nigeria was predicted as having the largest number of infections, followed by Algeria and South Africa. Mauritania was predicted to have the least cases of infection, followed by Seychelles and Eritrea. Sun et al.[16] assuming conservative estimates of African cases, utilized the reported number of COVID-19 cases traced to 12 major countries in Europe and America to Singapore, as well as flight data, to estimate the number of cases imported into Africa. They propagated the uncertainty in imported case count estimates to simulate onward spread of the virus, until 10000 cases were reached. The sensitivity of their results was accomplished using 1000 simulation runs under 4 different parameter combinations. They found that Morocco, Algeria, South Africa, Egypt, Tunisia, and Nigeria had the largest number of COVID-19 cases.

Drugs and vaccine discovery–The use of machine learning and simulations[17-19] as well as nanoparticles[20-21], are helping the repurposing and discovery of drugs, for efficient drug delivery. Furthermore, biotechnology- and bioinformatic-based approaches such as genome sequencing in identifying potential therapeutic options for vaccine discovery and for controlling COVID-19, have heightened. Initially, most genomes of African cases were sequenced abroad, but recently we have witnessed the setting up of molecular laboratories to handle this process within. Some African countries have also intensified efforts on alternative remedies to COVID-19, but there is need for established standards on their efficacy and safety.

This paper proposes an open source framework for advancing infectious disease research in Africa. The specific objectives of the paper are:

- To exhume relevant literature and assertions on SARS-CoV-2, for proper examination of its prevalence and transmission.
- To perform complete genome analysis by gender, for proper understanding and inference of SARS-CoV-2 diversity, and genome expression patterns across African countries.
- To mine cognitive knowledge from discovered patterns, for efficient (contact) tracing of the viral transmission routes.
- To report our findings and corroborate existing literature/assertions.

The contributions of this paper to knowledge include:

- *Open Source Framework*–Most of the biotechnology and bioinformatic tools are ‘black boxes’ and not open to contributions by the research community. This paper therefore encourages reproducible

research by introducing a set of rapid prototype modules capable of generating intermediate results that provide further insights into the prevalence and transmission of the pandemic.

- *Effectual Tracing of Undocumented Source of Infection*–Community transmission of viral and anti-viral treatments could engender novel mutations in the virus, leading to potentially evolving sub-strains with high mortality resistance. Consequently, tracing the routes of infection for efficient documentation of COVID-19 cases is very essential. Unsupervised genome pattern clustering and cognitive modeling are achieved in this study, to explain the genome diversity of the SARS-CoV-2 sub-stains as well as provide real time solution to the disease transmission pathway.
- *Intelligent Genome Surveillance*–It has been observed that when this virus transmits from one person to another over few months, it may acquire random sequence variations of its genetic material which serves as distinctive genomic "fingerprints". This paper enables the accurate mining of newly infected patients, to know which sub-strain of the virus is spreading within a country or been acquired from a different country. By combining machine learning techniques with cognitive knowledge mining, hidden sub-strains are revealed and different expression patterns followed, for seamless navigation of specific disease prevalence and surveillance of global infectious disease threats.
- *Misinformation/Disinformation Management*–Establishing transmission pathways would help minimize the growing trend of misinformation/disinformation, as country specific/global transmissions and spread of the virus could easily be contained. This paper guarantees the identification of possible infection routes by comparing genome sequences from different locations, to discover genetic diversity among sub-strains, and future potentials for investigating its fatal nature and spread.
- *Inputs to Novel Vaccine Development*–Understanding infection and transmission pathways could provide meaningful contributions to vaccine development and the discovery of clinically active variants and prototype drugs/vaccines for curative purposes. This paper does not only discover the SARS-CoV-2 sub-strains but also computes dissimilarity/variability in emerging sub-strains–an essential variable for vaccine development. Probing underlying genetic variations of infected individuals by gender would certainly enhance comprehensibility of the viral strain patterns, impact on the affected cells and aid the development of both preventive and therapeutic vaccine prototypes for the disease.

SARS-COV-2: Existing Assertions and Developments

On March 11, 2020, the World Health Organization (WHO) declared the coronavirus disease a global pandemic. Although the disease is still spreading, the rate of spread has greatly declined. Hence, almost all the countries had reopened their economies after compulsory national lockdowns, and currently adapting to local circumstances, with reduced rate of contact tracing and follow-up. Accelerated research developments and competing demands to contain the virus however have opened several opportunities for clinicians and researchers to exploit available avenues for developing suitable treatment and vaccines. Consequently, a plethora of publications flooded the scientific and medical domains/journals, with majority of the contributions received from the Asian countries and China–the very source of the pandemic (https://clinicaltrails.gov). Several studies and investigations have resulted in the following assertions and developments:

1) WHO claimed that most transmissions of COVID-19 are attributed to symptomatic persons than asymptomatically infected persons, with asymptomatic persons practically incapable of transmitting or spreading the virus; but recently, persistent replication of SARS-CoV-2 variant has been derived from an asymptomatic individual[22]. Furthermore, whole genome sequencing of the persistently replicating strain shows diversity in nucleotide positions leading to 6 non-synonymous ORF1ab protein substitutions.
2) Confirmed cases of COVID-19 have surpassed those of SARS[22-23]. Its genetic diversity in most countries is similar to what obtains globally, suggesting repeated inter- and intra-country spread by infected persons rather than by "patient zero." While some studies claim that mutation of new strains of SARS-CoV-2 potentially escalate severity of the pandemic[24], further analysis have confirmed premature conclusions–as there is currently not enough evidence to support the claim that mutation significantly impacts spread of the virus.
3) Non-pharmaceutical interventions including physical distancing, isolation, and the use of mask are the best approach to contain the outbreak and may assist flatten the peak in communities. However, the challenge of compliance resulting in the allege fear of increased number of infection, especially in low and medium income countries or resource limited settings, such as Africa, remains an unresolved puzzle, as poor health facilities and confusable symptoms continue to becloud the true evidence of infected cases.
4) The question of how Africa has survived the COVID-19 surge thus far may lie in the herbal remedies that abound within the continent's biodiversity-rich ecosystems[25], widely used in most African

communities and typified by the recently announced Madagascan COVID-19 remedy[26]. While the unsubstantiated remedy still requires medical scrutiny to prove its efficacy by globally acclaimed standards stipulated by the WHO, it is also a pointer that Africa may be in a position to provide alternative solution to disease management in moments of distress such as the present pandemic.

5) Amid conspiracy theories, it has since been inferred that SARS-CoV-2 is not a laboratory engineered virus but a natural process, after a comparative analysis of the SARS-CoV-2 genomic data and related (reference) viruses was conducted–as the distinct features of mutation in the receptor-binding domain portion of the virus spike protein usually targets the outer cell (of humans) involved in regulating blood pressure; and the lack of evidence of the virus being engineered from previously known viruses, debunk the notion of SARS-CoV-2 from being biologically engineered.

6) All the three human CoVs (SARS, MERS and SARS-2) are the result of recombination among CoVs[27], as recombination has been found to affect patterns of common variants as well as substitutions.

7) Like SARS-CoV and MERS-CoV, SARS-CoV-2 appears to be a zoonotic virus which is transmitted to humans through animals such as bats, because genomic sequences of SARS-CoV-2 isolates from patients share significant sequence identity with very high degree of certainty that suggests a host shift from bats into humans[14][28].

8) Clinical specimens used for viral ribonucleic acid (RNA) detection of COVID-19 as reported in the literature include nasopharyngeal aspirates, throat and nose swabs, saliva, sputum, endotracheal aspirates, feces, and urine. Of these, saliva yields greater detection sensitivity and consistency with high viral load concentration[29-32].

9) At present, the sensitivity of clinical nucleic acid detection appears limited without clear pointers to genetic variation. However, studies such as[29] specifically identified nucleotides at different sites to infer genotypic/genomic variants of SARS-CoV-2, hence, suggesting multiple outbreak and source of transmission. Further, the presence of more samples in certain sites may indicate increased transmissibility.

10) Emerging trend of the virus may impact human health outcomes, demanding close monitoring and characterization of the viral genetic patterns. However, this view has opened series of inconclusive debates, with many scientists arguing that the prevalence of genetic mutations could have increased as a result of random (stochastic) processes without increased fitness. A more formal analysis of the

frequency of mutation recently suggests decreased transmissibility and the fact that the position of the spike protein does not reside within the receptor-binding domain, nullifies existing notions that mutation confers greater transmissibility.

11) Although majority of the mutations arising from viral replication have shown very negligible effect on the virus, with no possibility of infection; analysis of mutations in the spike protein of SARS-CoV-2 suggests increased mutation frequency[30]. However, mutation information is appropriate to track new variants of the virus with unique mutant genomes, improve understanding of transmission and quicken determination of whether new mutations are changing the virus properties.

12) The presence of near real time whole-genome sequence analysis has provided reliable assessments on the extent of SARS-CoV-2 transmission in communities, hence, facilitating early decision making to control the local spread of the virus.

13) The sudden appearance of various sub-strains of the virus may not be unconnected with the fact that the virus is influenced by the new physical or biochemical environment it finds itself and/or in its ability to adapt to such a new and changing environment. Consequently, studies have successfully traced the SARS-CoV-2 of infected patients using molecular and phylogenetic methods[33]–as most phylogenetic inferences substantially prove that the virus has evolved into several sub-strains or variants specific to regions of transmission. Some studies have also shown high similarity between strains in different countries–as genotyping analysis of SARS-CoV-2 isolates around the globe reveals that specific multiple mutations are predominant during similar pandemics. Hence, comparing genome sequences from different locations allows for the analysis of the genetic diversity among viral sub-strains, its fatal nature, pathogenicity, origin and spread.

14) Although people of all ages are prone to infection by this virus, elderly people with co-morbidities (underlying health conditions and compromised immune system) are more susceptible to severe infection and death. Presence of genetic variants among young men with severe COVID-19 have been confirmed in[34]–using whole-exome sequencing performed to identify potential monogenic cause. But uncertainty did set in among medical practitioners on whether COVID-19 is a viral disease or the response to a person's immune system that invariably damages a patient's organs. Also, confusion in treating diseases presenting COVID-19 symptoms, instigated difficulty for physicians to determine with confidence, the optimal means of caring for critically infected patients. Howbeit, available data

informs the role of immune system in either diminishing or aggravating the infection and optimal measures for resolving confusable symptoms.

15) Confidence in how to treat COVID-19 has tremendously grown, but uncertainty remains[35]. At the outset of the pandemic, there appeared to be no definite treatment and the fear as to whether physicians themselves would get sick griped almost all the health providers/centers, the world over; as some diagnosed with the virus were asymptomatic (showed no symptom). Currently, most COVID-19 patients now have mild symptoms; but two important questions still linger: Will there be a next phase of the pandemic? Has most of the various communities suddenly reached "herd immunity"?
16) Development of high-throughput sequencing has contributed high quality datasets including whole genome sequences of viral isolates to the public domain. Analysis of genome sequence also provides insights into global spread patterns, genetic diversity, as well as the dynamics of sub-strains evolution. With continuous availability of new data, deeper investigation into new methods towards efficient candidate vaccines discoveries for emerging and re-emerging COVID-19 and related pandemics is ongoing.
17) Emergency lockdowns have returned to parts of Europe, as France and Germany struggle with a second wave while the COVID-19 outbreak resurges rapidly. The United States is experiencing a third wave, and the UK is set to impose another lockdown. As this uncertainty window reopens, deaths are flat, with cross-border countries becoming apprehensive.

SARS-COV-2 Genome Analysis of African Isolates

In this section, we review existing works on SARS-CoV-2 analysis conducted on African genomes and present in Table 1, a summary of the viral isolates, their transmission history, intra-country sub-strains discovery and additional information about local transmission, mutation and spread.

Egypt: Sekizuka et. al.[36] characterized the possible origin of 10 SARS-CoV-2 positive travelers from Egypt together with their close contacts. The viral genome sequences of the 10 travelers were aligned with genome sequence retrieved from GISAID using MAFFT v7.222; two distinct genome lineages circulating mostly in Europe and South America were identified. They concluded that increased cases may complicate the identification of infection routes. The analysis and comparison of 2 Egyptian SARS-CoV-2 isolates using CLC Genomic Workbench version 20[37] yielded at least 99.9% similarity. However, variations occurred at 8658,

15907, 19906 and 18877 nucleotide sites. Comparable with the Wuhan reference genome, 5 mutations (C->T: C241T, C->T: C3037T, C->T: C14408T, A->G: A23260G and G->T: G25563T) were observed among the Egyptian sequences. Specifically, the genome sequence of patient 1 was discovered to shared similarities with Taiwanese isolate who traveled in February 2020 to Dubai and Egypt, few mutations were found at sites (C->G: A8658G, G->A: G15907A and C->T: C18877T). Patient 2 recorded similarities but with 3 specific variations (T->C: T4278C, G-T: G18963T, C26692) with the genome sequence of a Japanese on a Nile river ship cruise, who tested positive to SARS-CoV-2 on March 9, 2020. Amidst all the variations, the spike protein of the 2 Egyptian SARS-CoV-2 isolates are identical with G614D but varies from the spike protein of the Wuhan reference genome. Further investigation involved the nucleotide alignment of the Egyptian sequence with other GISAID SARS-CoV-2 genome sequences using BioEdit version 7.0.5.3 and ClustalW. Phylogenetic tree constructed using MEGA7, showed the clustering of Egyptian sequences in clade A2a with Asian Europe, United States, Australian and African sequences.

Kenya: The first reported case of SARS-CoV-2 sequencing and analysis in Kenya consisted of positive samples from symptomatic and asymptomatic patients from Nairobi (20) and Coastal Kenya (102). Seventy-eight global sequences representing countries in Europe, Asia, America and Africa in GISAID were randomly sampled and retrieved. The alignment of the retrieved sequences together with the Kenyan sequences was realized using MAFFT v7.310. A maximum likelihood phylogenetic tree was established through RAXML-NGS v0.9.0 using a GTR+F0+G4m model in 1000 bootstraps run. The assignment of lineages to the Kenyan genome sequences was possible through PANGOLIN toolkit (v1.1.14). Evidently, the phylogenetic tree displayed 10 strains of SARS-CoV-2 circulating in Kenya, which evinces the multiple introduction of the virus into Kenya. Nonetheless, B.1 lineage discerned to be dominant and causing most of the infections in Coastal Kenya. However, all the viral strains were identifiable with the strains circulating globally and none of the strains was distinctive to Kenya.

Morocco: Laamarti et al.[38] studied the molecular distribution of 28 Moroccan SARS-CoV-2 strains isolated between March 3, 2020 and May 15, 2020, with 12 North African (Tunisia (7), Algeria (3) and Egypt (2)) viral sequences downloaded from GISAID and 6 Moroccan genome sequenced for the study. Specifically, 6 sequences were mapped to the Wuhan reference genome using BWA-MEM v0.7.17-r1188, while Minimap v2.12-r847 was used in mapping the GISAID downloaded genomes. The analysis of all Moroccan genome sequences disclosed 61 mutations in comparison with the Wuhan genome: 27 synonymous, 5 intergenic, 27 non-synonymous and 2 lost stops. These mutations were distributed among 5 genes (ORF1ab, S, M, N and

ORF3a); ORF1ab harbored the highest number (37.7%) of non-synonymous mutations. In like manner, the comparison and characterization of the 12 North African genome sequences together with the 28 Moroccan viral genome sequences, against the Wuhan genome, revealed a total of 118 mutations: 58 non-synonymous, 48 synonymous and 12 inter-gene mutations. Regarding non-synonymous mutations, missense, lost stop and stop again were found to contribute 91.38%, 6.90% and 1.72%, respectively. Amongst the 58 non-synonymous mutations, 13 were repeatedly found in more than one genome. The most recurrent of the mutations observed within the four North African countries occurred in the S protein (D614G) and ORF3a (Q57H) with prevalence of 92.5% and 42.5%, respectively. The 11 (T265I, T5020I, K2798R, R203K, D1036E, V2047F, A2637V, T2648I, C4588F, S202N, L84S) other mutations were inconsistently observed among the four countries. Also, in addition to the 5 genes earlier discovered to harbor mutations, 2 more genes (E and ORF8) were discovered. However, ORF1ab remained the leading gene with 67.4% mutations, bearing two-third of the 118 mutations. In addition, with a focus on Morocco, the phylogenetic analysis conducted using 256 representative genomes constituting genome sequences from the 6 continents, the phylogenetic tree disclosed 5 major clades, 2 of which constituted main strains from Asia while the other clades consisted of strains belonging to various continents. This diversity in lineage infers the introduction of SARS-CoV-2 into Morocco from multiple routes. In[39], the molecular analysis of SARS-CoV-2 genome sequences of 22 Moroccan isolates obtained from three laboratories in Morocco as at June 7, 2020 revealed 62 mutations. In comparison with the Wuhan reference genome and 40366 viral genome sequences retrieved from GISAID, the Moroccan genome evinced similar mutations with the Wuhan reference genome and other strains circulating globally. Additional 6 mutations (NSP10_R134S, NSP15_D335N, NSP16_1169L, NSP3_L431H, NSP3_P1292L and Spike_V6F) particular to the Moroccan SARS-CoV-2 genome were also discovered. Their study was realized by performing MAFFT multi-alignment of all retrieved genomes from GISAID and phylogeny analysis of the aligned sequences created by maximum likelihood using IQTREE. The evolutionary analysis revealed 3 clades: 20A, 20B and 20C, and authenticates the findings of[40], which used similar methodology to investigate the phylogenesis of 250 SARS-CoV-2 genome sequences from GISAID. Sixteen variants were detected in 6 Moroccan SARS-CoV-2 genome sequences. Among the variants were synonymous (F924F and L4715L), nonsynonymous (D->G: D614G) and intergenic (241C->T). Jouali et al.[41] corroborate the inference of[39] by comparatively studying the SARS-CoV-2 genome sequence of a mildly symptomatic Moroccan patient with other sequences from Morocco. The phylogenetic analysis of the genome was conducted using GIDSAID enabled Nextstrain tool. The genome under study was discovered to belong to clade B11 revealing high similarity with genome sequences from Florida, USA.

Nigeria: The comprehensive knowledge of phyloevolution and comparative discrimination of SARS-CoV-2 molecular characterization can be useful in the critical investigation of the virus pathogenesis, disease control, treatment and vaccine development. The study of SARS-CoV-2 evolution in Nigeria by[42] exhibited a concerted similarity with the Wuhan reference genome, which introduction into Nigeria was inferred to arrive from Wuhan through an Italian traveler. The study involved a 2-step analysis of multiple sequence alignment and phylogenetic analysis of 39 complete genome of SARS-Cov-2 with their various travel history. The constructed phylogenetic tree separated the Nigerian strains into the cluster with a Wuhan subclade. Multiple sequence analysis using ClustalW revealed >70% similarity with the Wuhan reference sequence. In another study, representative whole genome sequences of each of the seven lineages of human SARS-CoV-2 circulating in Nigeria were obtained from GISAID and aligned with all full genomes from Nigeria using MAFFT v7.310[43]. It was found that 4 of the new sequences clustered closely together forming separate clade, strongly suggesting local community transmission. Similarly, other new sequences behaved in same way, revealing a follow-up from same patients. Inter-country analysis of lineages from Nigeria clustered with sequences from Asia, Europe, USA, Middle East, Australia, and other African countries, indicating multiple transmissions.

South Africa: From the consensus genomic sequence of a South African isolate, who travelled back South Africa from Italy, Allam et al.[44] identified 6 non-synonymous variants. This was attained using MAFFT v7.042 from the multiple sequence alignment of 965 SARS-CoV-2 genome sequence, extracted from GISAID together with the isolate's sequence. As at time of the report, the mutations at location 13,620 bp and 21,595 bp were reported to be absent in every other SARS-CoV-2 genome. Furthermore, DUET web server prognosticated the destabilizing and stabilizing effect of D614G variant of the spike protein and P322L mutation of the nsp12 respectively. As a supplemental study to the obstacles encountered during near-real time SARS-CoV-2 genotyping during pandemic, Pillay et al.[45] established 3 lineages (B, B.1 and B.2) from a phylogenetic analysis of 54 near-full-length genome, using Phylogenetic Assignment of Named Global Outbreak LINeages (PANGOLIN) software suite. The 54 genome sequences and 10,959 GISAID reference genomes were aligned using MAFFT v7.313 and a maximum likelihood tree topology constructed in IQ-TREE. From the 54 genome sequences, lineage B.1 had the highest number of 50 samples clustering closely with 99.99% similarities. Moreover, the number of minimal mutations reported was quite minimal and not too divergent from those found in the public sequences.

Uganda: In Uganda, Bugembe et al.[46] reported on the genomic sequences of 14 travelers from SARS-CoV-2 dense region and 6 truck drivers returning to Uganda from Kenya, Tanzania and South Sudan. Phylogeny of

Ugandan genome sequences identified with 6 lineages (A, B, B.1, B.1.1, B1.1.1 and B4) was performed by comparing with globally detected genomes. The B.1 lineage accommodated greater number of sequences circulating in more than 20 countries in Europe, America and Australia and Asia. Although infection routes were mostly from the cargo truck drivers and air travelers, the viral genome of travelers from Dubai associated with 3 lineages (A, B, B.1.1.1) while genomes from the cargo truck drivers entering Uganda from Tanzania belonged to lineage A and B.1. Whilst the sample size is small, the evidence suggests multiple sources of contact. Furthermore, the diversity of the Uganda genome sequence from the Wuhan reference genome occurred at 5-20 nucleotide positions across approximately 30kb genome, including the spike protein. In the spike protein, four viral sequences from lineage A encode D614, whereas sequences belonging the other clades encode G614.

[Table 1, here]

Methods

Studies on SARS-CoV-2 single nucleotide polymorphism and lineage discovery keep surging the literature; mainly exploiting a two-step algorithm (multi sequence alignment and phylogenetic analysis), with the use of common tools and techniques such as MAFFT and maximum likelihood tree topology that target specific nucleotide sites. Although aligning collected genomes against reference genome(s) has helped in the discovery of gene/genetic variability/ diversity, results of evolutionary analysis have consistently shown structured transmission with possible multiple introductions into the population[79]. Furthermore, most of the works on African genome isolates mine data from the Global Initiative on Sharing All Influenza Data: GISAID EpiFlu™ (a database of SARS-CoV-2 partial and complete genome compilations distributed by clinicians and researchers, the world over).

To advance the current practice, an open source framework that combines biotechnology and bioinformatic approaches with AI methodologies, into a hybridized system, is proposed in this section, for in-depth understanding and further works development on SARS-CoV-2. Our framework generates interesting intermediate data including phylogenomic trees, pairwise nucleotide similarity matrices/scores, gene diversity plots, genome expression patterns analysis, essential for enriching the genome datasets, towards intelligent genome characterization and prediction. With this approach, community contribution is guaranteed, and reproducible research possible. Furthermore, intermediate data could be repurposed for building new concepts

and models. The general workflow describing the proposed system framework is shown in Fig. 1, and algorithm implementing the framework presented in Algorithm 1.

Fig. 1. Workflow describing the proposed system framework. The workflow begins with the excavation of FASTA files of human SARS-CoV-2 genome sequences from GISAID. These files are stripped and processed into a genome database (DB) as multiple columns of nucleotide sequence. A series of AI/ML techniques are applied to extract knowledge from the genome datasets as follows: Compute dis(similarities) scores between the various pairs of genome sequences and obtain a genomic tree of highly dis(similar) isolates grouped in the form of a dendrogram/phylogenomic tree. Determine the optimal number of natural clusters–to provide additional knowledge. Separate the viral sub-strains using self-organizing map (SOM) component planes–for transmission pathways visualization. Perform direct pairwise nucleotide alignment of the entire genome sequences–to yield a nucleotide similarity matrix. Generate cognitive map–for intelligent sub-strains contact tracing and prediction.

Algorithm 1. Steps implementing the workflow in Fig. 1.

```
import necessary libraries
set path to current directory
#Genome nucleotide fragments processing
create a list of FASTA files (fasta_list) to process
for file_name in fasta_list:
    open FASTA_file for read
    store a line of genome sequence
    for line in file_name:
        strip line into a list of nucleotide fragments (nucleotide_fragments)
    for line in nucleotide_fragments:
        write nucleotide code into complete genome file (complete_genome)
    close FASTA_file
#Direct nucleotide alignment and similarity scores generation
open complete_genome for read
store a line of nucleotide code
for line in complete_genome:
    align nucleotide pair and compare nucleotide code
build pairwise dis(similarity) matrix using a suitable distance metric (e.g., Euclidean distance)
#AGNES/hierarchical clustering: generate phylogenomic tree and cluster plots
treat observations (nucleotides) as cluster points and compute AGNES distance coefficients between clusters
compute scores between genome isolates clusters
```

22. build and visualize genomic tree
23. discover and validate optimal natural clusters (k) using any k-means based N approaches (N>2) (elbow, silhouettes, gap-statistics etc.).
24. partition the tree into k clusters
25. #Genome expression patterns discovery
26. perform SOM clustering on complete_genome
27. obtain SOM component planes of learned genome expression patterns
28. obtain pairwise correlation coefficients
29. label target (output) classes using dis(similarity) and genome expression clusters–indicating mutant sub-strains and viral expression patterns, to form enriched genome datasets.
30. generate cognitive maps with embedded links of genome isolates.
31. close complete_genome.

Data Source and Genome Sequences Selection

Publicly available datasets of coronavirus deposited between December 2019 and October 12, 2020 were excavated from GISAID for the purpose of this study, and complete genome sequences of human SARS-CoV-2 isolates from selected countries within Africa, collected. Useful metadata on the extracted genome sequences (Country, State, Abbreviation, Accession No., Length, Gender, Age, Travel history, Specimen source, Status, Submitting Lab, Authors), with Gender and Age removed, are documented (see SupplData1_1.xlsx). The preprocessed FASTA files of genome isolates excavated from GISAID, striped and dumped as column sequences for male and female patients are found in (SupplData2_1.xlsx). Although some of the datasets had incomplete information (e.g. age, specimen source and status), gender served as a compulsory criterion for profiling the excavated genomes. Male isolates excavated according to *Country (number of sequence, state(s) extracted)*, include: Algeria (2, 1); Benin (5, 2); Cameroon (1, 1); DRC(15, 4); Egypt (5, 2); Gambia (6, 2); Ghana (4, 1); Kenya (4, 1); Mali (5, 2); Morocco (5, 3); Nigeria (13, 5); Senegal (14, 6); South Africa (65, 20); Tunisia (1, 1). Female isolates excavated according to *country (number of sequence, state(s) extracted)*, include: Algeria (1, 1); Benin (4, 1); DRC(13, 4); Egypt (4, 2); Gambia (4, 2); Ghana (4, 1); Kenya (4, 1); Madagascar (2, 2); Mali (4, 1); Morocco (2, 1); Nigeria (11, 5); Senegal (11, 3); South Africa (80, 24); Tunisia (1, 1). Hence, a total of 290 genome sequences (145 males, 145 females) with genome lengths of over 29000 nucleotides, were excavated. Specimen sources include swabs (nasal, oral, throat, nasal and oral); fluids (bronchoalveolar lavage, saliva, sputum) and unknown. Status of patients include hospitalized, not hospitalized, acute bronchitis, symptomatic,

asymptomatic, alive and unknown. Age range of 2 months and 99 years were considered. In the Male datasets the ages of 9 patients (Ghana (1), Kenya (4), Morocco (1), Mali (1) and Nigeria (2)), were unknown; while in the female datasets, the ages of 10 patients (Kenya (4), Morocco (2), and Nigeria (4)), were unknown. Finally, about 1.75% and 2.11% of errors in sequencing (noise) were noticed in the male and female genome datasets, respectively.

Unsupervised Genome Clustering

Self-organizing map (SOM) has been used extensively in the field of bioinformatics, for visual inspection of biological processes, genes pattern expressions–as maps of (input) component planes analysis. SOM is an unsupervised artificial neural network (ANN), learned to produce a low-dimensional (typically two-dimensional), discretized representation of the training samples input space, known as a map. Patterns exhibited by the different isolates confirm intra- and inter-country transmissions. The SOM algorithm locates a winning neuron, its adjusting weights and neighboring neurons. Using an unsupervised, competitive learning process, SOMs produce a low-dimensional, discretized representation of the input space of training samples, known as the feature map (see Fig. 2). During training, weights of the winning neuron and neurons in a predefined neighborhood are adjusted towards the input vector using equation (1),

$$w_{id}^{t+1} = w_{id}^{t} + rf(i,q)(x_d - w_{id}^{t});\ 1 \le d \le D. \quad (1)$$

where r is the learning rate and $f(i,q)$ is the neighborhood function, with value 1 at the winning neuron q; and decreases as the distance between i and q increases. At the end, the principal features of the input data are retained, hence, making SOM a dimension reduction technique. The batch unsupervised weight/bias algorithm of MATLAB (*trainbu*) with mean squared error (MSE) performance evaluation, was adopted to drive the proposed SOM. This algorithm trains a network with weight and bias learning rules using batch updates. The training was carried out in two phases: a rough training with large (initial) neighborhood radius and large (initial) learning rate, followed by a finetuned training phase with smaller radius and learning rate. The rough training phase can span any number of iterations depending on the capacity of the processing device. In this paper, we kept the number of iterations at 200 with initial and final neighborhood radius of 5 and 2, respectively, in addition to a learning rate in the range of 0.5 and 0.1. The fine training phase also had a maximum of 1000 epochs, and a fixed learning rate of 0.2. Selection of best centroids of the genome feature within each cluster was based on the Euclidean distance criterion. The algorithm configures output vectors into a

topological presentation of the original multi-dimensional data, producing a SOM in which individuals with similar features are mapped to the same map unit or nearby units, thereby creating smooth transition of related genome sequences to unrelated genome sequences over the entire map.

Fig. 2. SOM showing the map topology and interactions between nodes. Each neuron is assigned a vector of weights ($w = w_{i1}, w_{i2}, .. w_{iN}$) with dimension similar to the input vector i ($i = 1, 2, \dots, L$); where L is the total number of neurons in the network. The input nodes have p features, and the output nodes, q prototypes, with each prototype connected to all features. The weight vector of the connections consumes the prototype of each neuron and has same dimension as the input vector. SOMs differ from other artificial neural networks as they apply competitive learning, against error correction learning such as backpropagation, and the fact that they preserve the topological properties of the input space using a neighborhood function.

Cognitive Knowledge Mining

Knowledge mining has served huge benefits for quick learning from big data. We applied Natural Language Processing of the genome datasets to extract knowledge of similar strains of the virus. A simple iterative technique is imposed on the SOM isolates ($i = 1,2,3, \dots, n$), where n is the maximum number of isolates, as follows: For each isolate pattern, compile similar patterns with the rest of the isolates (i.e., $i + 1, i + 2, \dots, n$). Concatenate compiled patterns into a list ($j_1, j_2, \dots, j_m$) where j is an element of the list. Dump the compiled list into $CogMap(k_i \in j_1, j_2, \dots, j_m)$.

Configuration of Computing Device

A HP laptop 15-bs1xx with up to 1TB storage running on Windows 10 Pro Version 10.018326 Build 18362 was used for processing the excavated genome sequences, algorithms/programs and other ancillary data. The system had an installed memory (RAM) of 16 GB with the following processor configuration: 1.60 GHz, 1801 MHz, 4 Core(s) and 8 logical processors. Although our system performed satisfactorily and produced the desired results, higher system configurations would improve the computational speedup.

Results

Low Cystosine to Guanine Transition and High Thymine Content in Human SARS-CoV-2

The RNA sequence is composed of 4 nucleotides (adenine (A), cytosine (C), guanine (G) and thymine (T)), also considered as polymers of 16 (i.e. 2^4) dinucleotides. Yin[54] revealed the frequency of mutations in the spike-protein, RNA polymerase, RNA primase and nucleoprotein. The study algorithm involved the alignment of multiple genome subsets with SARS-CoV-2 reference genome using Clustal Omega and Jaccard distance in the variation among 558 complete genome sequence retrieved from GISAID on March 23, 2020. Comparably, Wang et al.'s[55] homology analysis and sequence alignment established a reference genome from 95 strains obtained from NCBI and GISAID on February 14, 2020. The study as well, found mutations at nt8782 of ORF1a, nt28144 of ORF8 and nt29095 of N region from the analysis at the nucleotide and amino acid levels.

The SARS-CoV-2 reference genome (29903 nucleotides[56], sequence number NC_045512) consists of 29.94% of A, 18.37% of C, 19.61% of G and 32.08% of T nucleotides[57]. Hence, the expected frequency of CG dinucleotide in the viral genome is 3.60% (i.e. 19.61% × 18.37%). Mercatelli and Giorgi[58] analyzed 48635 complete genome sequences spread across geographic regions including, Africa (514), Asia (3340), Europe (31818), North America (10250), Oceania (2127) and South America (575). The obtained sequences were aligned over the Wuhan reference genome sequence (NC_045512.2) using NUCMER version 3.1. The analyzed result exemplifies the nature of mutation across the world, per continent and per country. The number of mutations per sample was reported to be relatively low but with an average mutation rate of 7.23. Although number of mutations per continent did not differ significantly from the average mutation rate, the average number of mutations per country differed significantly. For two out of the three African countries included in the study; Congo had a high mutational burden of 8.30 while Kenya had a low mutation rate of 5.38. Single nucleotide polymorphism substitution accounted for 0.6% of all the observed mutations, making it more prevalent over insertion and deletion mutations. The transition from C->T makes up the 55.1% of all point mutation, A->G is the second leading transition (14.8%) globally and in Africa, America and Europe. The effect of A->G transition on the protein sequence of SARS-CoV-2 formulates the G-clade predominantly found in Africa, Europe, Oceania and South America. Sjaarda et al.[59] studied 25 SARS-CoV-2 genome samples from local cases of COVID-19 collected during the early days of the spread from eastern Ontario, Canada between March 18 and March 30, 2020, with 2 genomes belonging to the S-clade and the remaining 23 belonging to the G-clade of SARS-CoV-2; and contained 45 polymorphic sites with one shared missense and three unique synonymous variants in the gene encoding the spike protein. They found that most of the genomes had between 6 to 8 variants when compared with the NC_045512.2 reference genome. Also, the most common nucleotide substitution was from C->T (25/45 variants), followed by G->T (7/45 variants) and A->G (4/45).

From the genomes excavated for this study, the average frequency count of nucleotides for male and female isolates are roughly similar, as the proportion of each nucleotide has (A=29.8553%; C=18.3830; G=19.6366; T=32.1254) nucleotides for the male isolates and (A=29.3616; C=18.3765; G=19.6365; T=32.1254) nucleotides for the female isolates; with average frequency of 36.1% CG dinucleotide compared to 59.1% CT and 63.1 GT dinucleotides, in viral genome, for both genders. Our result corroborates the findings of[57] on CG reduction in SARS-CoV-2, achieved through C/G nucleotide mutating into A/T (a universally occurring process in all forms of life). Generally, the mutation spectrum of new genome mutants seems enriched in C->T and G->T mutations, as different studies also corroborate our findings of dominant transition and transversion mutants in human SARS-CoV-2 isolates[43][59][60], but no strong evidence supports the claim that the virus rapidly or slowly mutates than expected, as most of the mutations are probably neutral or deleterious to the virus[61]. But while the T nucleotide is the most frequent nucleotide in the genome, its frequency seems to increase further across all samples, and the substitution process appears non-reversible and unbalanced.

Genome Diversity Analysis

Genes cover very large regions of chromosomes with most gene content having almost identical expressions with other chromosomes in the genome. We introduce the density plots, to examine whole genome sequences, for the discovery of variability in nucleotide distribution in male and female isolates, between and within countries.

Inter-country Analysis: Fig. 3 shows density plots for male and female isolates between African countries. Observe that the male isolates exhibit smoother distribution curve with most of the isolates having identical expression patterns; compared to female isolates, which distribution curve is partly influenced by dominant outliers of possible infested sequences from Gambia, Kenya, Mali, Morocco, Nigeria, Senegal and South Africa. The outliers may be as a result of observed sequencing errors, or extensive localized variation in DNA polymorphism and large regions of low gene density, diversity and recombination.

Fig. 3. Density plots of entire genome datasets. A Density plot visualizes the distribution of data over a continuous scale. It is a variation of the histogram that applies kernel smoothing to plot values, enabling smoother distributions by smoothening out the noise. The peaks of the density plot help display where values are concentrated over the interval.

Intra-country Analysis: Density plots revealing genomic diversity between male and female isolates of selected countries with more genome samples are presented in Fig. 4, Fig. 5, Fig. 6 and Fig. 7, for DRC, Nigeria, Senegal and South Africa, respectively. We observe marked variabilities in some nucleotide sequences of Nigerian (Fig. 5), Senegalese (Fig. 6) and South African (Fig. 7) isolates–indicating gene expression patterns differences in some of the isolates.

Fig. 4. Density plots of DRC's isolates. Both plots exhibit similar curve pattern with near aligned isolates, indicating identical genome expression between the isolates.

Fig. 5. Density plots of Nigerian isolates. Although both plots maintain same curve patterns, the male isolates exhibit diversity in most of the genomes. Two female isolates (Fig, 5b) present distincnt variations from the rest of the isolates, indicating significant genome diversity beween the male and female isolates.

Fig. 6. Density plots of Senegalese isolates. Aside the single isolate that fails to align with other male isolates (Fig. 6a), both plots maintain near-aligned isolates.

Fig. 7. Density plots of South African isolates. Both plots maintain same patterns with shades of similar isolates dominating the plot area.

Phylogenomic Analysis

Genes content comparison has become commonplace, but associating its order is challenging. Phylogenomic trees appear not widely used because of computational difficulties (massive data, high processing cost and limited processing infrastructure). In this paper, we exploit complete genome sequences to construct hierarchical

cluster structures (dendrograms) that discriminate inter- and intra-country SARS-CoV-2 isolates. To achieve this, Hierarchical Clustering Analysis (HCA) also known as Agglomerative Nesting (AGNES)[62] was performed on the various isolates. While there are natural structural entities in some datasets that provide information on the number of clusters or classes, others including the dataset containing genome sequences are structured without boundaries. Cluster validation (an unsupervised methodology aimed at unravelling the actual count of clusters that best describes a dataset without any priori class knowledge) is therefore essential. This paper adopted the elbow criteria[63], to validate the number of clusters available in the genome datasets. Hence, yielding 2 and 3 clusters for inter- and intra-country phylogenomic analysis, respectively.

Inter-Country Analysis: Fig. 8 shows phylogenomic trees for male and female patients. Both trees suggest inevitable sub-strains (independent) mutant accumulation in different countries, resulting in highly dense clusters (encircled in red), while few mild divergent strains with specific mutations are geographically different, hence, occupying smaller disparate clusters.

Fig. 8. Phylogenomic trees for African male and female isolates. For full names of country codes, see Additional file 1: SupplData1_1.xlsx.

Intra-country Analysis: In Fig. 9, phylogenomic trees of male and female isolates from DRC are shown. For male patients (Fig. 9a), Haut-Katanga and Sud Kivu isolates cluster differently with isolates from other states. However, Kinshasha isolate clusters closely with Nord Kivu and Kongo Central isolates, while other Kinshasha, Sud Kivu and Haut-Katanga isolates cluster together showing less genome diversity. For female patients (Fig. 9b), Kinshasha isolate clusters differently with isolates from Kongo Central, but clusters together with Sud Kivu isolate. However, the Sud Kivu and Kongo Central isolates (independently) cluster together indicating intra-specific genome similarity.

Fig. 9. Phylogenomic trees for DRC's male and female isolates. For full names of country codes, see Additional file 1: SupplData1_1.xlsx.

In Fig. 10, phylogenomic trees of male and female isolates from Nigeria are shown. For male patients (Fig. 10a), Osun isolate clusters differently with isolates from Lagos, Kwara, Oyo and other Osun isolates, including the unknow (infected) Nigerian isolate (NGA) who travelled to Greece. Oyo isolate clusters differently with isolates from Kwara, Osun and other Oyo isolate, indicating high genome diversity, even between same state. Also, Kwara isolate clusters differently with isolates from Osun and Oyo. However, Osun isolate clusters closely with isolates from Oyo and Kwara, indicating less genome diversity. For female patients (Fig. 10b), the phylogenomic tree present a near-flat structure with Ekiti isolate closely clustering with Osun, Ogun and the unknown Nigerian isolates (NGN)–infected through a local infection, and from Ogun. Furthermore, Ekiti and Ondo isolates cluster differently with isolates from other states.

Fig. 10. Phylogenomic trees for Nigerian male and female isolates. For full names of country codes, see Additional file 1: SupplData1_1.xlsx.

In Fig. 11, phylogenomic trees of male and female isolates from Senegal are presented. For male patients (Fig. 11a), Pikine isolate clusters differently with isolates from the rest of the states, while Diourbel and Dakar isolates cluster differently with other isolates, save Pikine. However, other Dakar isolates cluster closely with remaining isolates from Diourbel, St. Louis, Kolda and Thies, indicating less genome diversity between these isolates. For Female patients (Fig. 11b), Thies isolates cluster differently with isolates from Dakar and Diourbel as well as another isolate from Thies, indicating high genome diversity. However, Diourbel isolates cluster closely with remaining Diourbel isolate, Dakar and Thies isolates, indicating less genome diversity between these isolates.

Fig. 11. Phylogenomic trees for Senegalese male and female isolates. For full names of country codes, see Additional file 1: SupplData1_1.xlsx.

In Fig. 12, phylogenomic trees of male and female isolates from South Africa are presented. In Fig. 12b, female isolates/patients show high diversity as more isolates cluster differently, compared to male isolates (Fig. 12a), which maintain a near-flat tree structure, with the eThekwini isolate clustering differently from all other isolates.

Fig. 12. Phylogenomic trees for South African male and female isolates. For full names of country codes, see Additional file 1: SupplData1_1.xlsx.

Nucleotide Similarity Analysis

Several techniques for biological sequence alignment (multiple or pairwise) have flourished the literature[64], but most of these techniques suffer from the lack of accuracy and partial interpretations. A direct pairwise genome sequence alignments (embedded in Algorithm 1) was carried out to match each nucleotide pair at the exact nucleotide positions of the SARS-CoV-2 genome, extending the alignments across other genome pairs. The output is a matrix of similarity scores. Fig. 13 shows inter- and intra-nucleotide similarities with strong evolutionary relationships highlighted in green color (for more clearer view of the nucleotide similarity matrices, see SupplData2_1.xlsx). For male isolates (Fig. 13a), inter-nucleotide similarities cut across the following countries with strong evolutionary relationships observed for the countries listed beside them: Algeria (Senegal). Benin (DRC; Gambia; Greater Ghana; Kenya; Mali; Tunisia). Cameroon (DRC; Nigeria; Senegal; South Africa). Kenya (Mali; Nigeria; Tunisia). Morocco (South Africa). Nigeria (Senegal; South Africa; Tunisia). Senegal (South Africa). Intra-nucleotide similarities exist for the following isolates, with strong evolutionary relationships between the (states) listed beside them: Algeria (Bilda-Bilda). Benin (Cotonou-Cotonou, Cotonou-Oueme). DRC (Haut-Katanga-Haut-Katanga, Haut-Katanga-Kinshasha, Haut-Katanga-Kongo Central). Egypt (Cairo-Cairo). Gambia (Kombo-West Coast Region). Ghana (Greater Ghana-Greater Ghana). Kenya. Mali (Bamako-Bamako, Bamako-Mopti). Nigeria (Lagos-Osun; Lagos-Oyo, Osun-Oyo). Senegal (Dakar-Kolda; Dakar-St. Louis; Dakar-Thies; Kolda- St. Louis; Kolda-Thies; St. Louis-St. Louis; Thies-Thies). South Africa has the highest number of isolates with intra-nucleotide similarities spread across the various cities. The following states have strong evolutionary relationships with those listed beside them: Amajuba (eThekwini; Ilembe; King Cetshwayo; Kwazulu Natal; Umgungundlovu; Umkhanyakude;

673 Umzinyathi; Uthukela; Uthungulu). Berea (Berea; Kwazulu Natal). Cape Town Metro (eThekwini; Harry
674 Gwala; King Cetshwayo; Uthukela). eThekwini however, exhibits similar nucleotide relationship with Ilembe,
675 King Cetshwayo, Kwazulu Natal, Umgungundlovu, Umkhanyakude, Umzinyathi; and share evolutionary
676 relationships with some of these states (Ilembe; King Cetshwayo; Kwazulu Natal; Umgungundlovu;
677 Umkhanyakude; Umzinyathi; Uthukela; Uthungulu).
678 For female isolates (Fig. 13b), inter-nucleotide similarities cut across the following countries with strong
679 evolutionary relationships observed for the countries listed beside them: Algeria (DRC; Nigeria). Benin
680 (Gambia; Ghana; Kenya; Mali; Morocco; Nigeria; Tunisia). DRC (Nigeria; Senegal; South Africa). Gambia,
681 Ghana and Kenya exhibit similar nucleotide relationship with some of these countries (Kenya; Mali; Nigeria;
682 South Africa and Tunisia). Morocco (Nigeria). Nigeria (South Africa; Tunisia). Senegal (South Africa). Intra-
683 nucleotide similarities exist for the following isolates with strong evolutionary relationships in the states listed
684 beside them: Benin (Cotonou-Cotonou). DRC (Haut-Katanga-Kongo Central, Kinshasha-Kinshasha, Haut-
685 Katanga-Sud Kivu). Gambia (Kombo-West Coast Region). Ghana (Greater Ghana-Greater Ghana). Kenya.
686 Madagascar (Fenoarivo-Antananarivo). Mali (Bamako-Bamako). Morocco (Rabat-Rabat). Nigeria (Ekiti-Ondo;
687 Ekiti-Osun). Senegal (Dakar-Diourbel; Dakar-Thies; Diourbel-Diourbel). South Africa has the highest number
688 of isolates with intra-nucleotide similarities spread across the various cities. The following states have strong
689 evolutionary relationships with those listed beside them: Amajuba (Cape Town; Ilembe; North West; Overport;
690 Umgungundlovu; Uthukela; Uthungulu; Zululand). Berea (North West; Umbilo; Umgungundlovu;
691 Umkhanyakude; Uthungulu). Cape Town Metro (Cape Town; Ilembe; North West; Umgungundlovu; Uthukela;
692 Zululand). EC (eThekwini). eThekwini (Harry Gwala; Ilembe; King Cetshwayo; Kwazulu Natal; Stanger;
693 Uthukela; Umzinyathi). Free State (Free State; Harry Gwala; King Cetshwayo; Mmorningside; North West;
694 Sisonke; Ugu; Umbilo). eThekwini however, exhibits similar nucleotide relationship with Harry Gwala, Ilembe,
695 King Cetshwayo and Kwazulu Natal; and share evolutionary relationships with some of these states (Harry
696 Gwala; Ilembe; King Cetshwayo; Kwazulu Natal; Stanger; Umkhanyakude; Umzinyathi).
697
698
699 **Fig. 13. Nucleotide similarity matrices.** Green colored cells are regions of high similarity that may indicate functional, structural and/or
700 evolutionary relationships between nucleotide sequences.
701
702
703

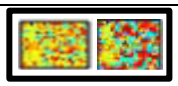
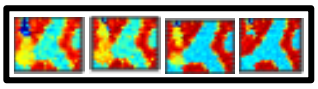
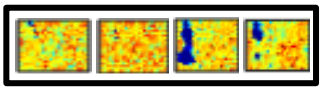
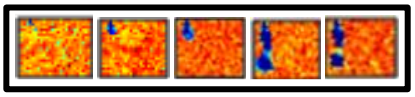
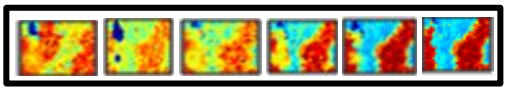
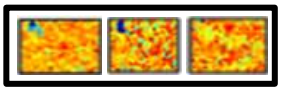
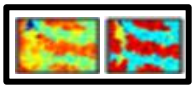
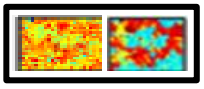
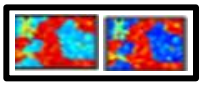
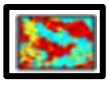
Sub-strain Pattern and Transmission Route Discoveries

Comparing component planes of self-organizing maps (SOMs) can help detect genome expression patterns in identical positions (indicating correlation between the respective components), suitable for the discovery of sub-strains across the respective isolates. Component planes representation can enable the visualization of the relative component distributions of input data. Each component plane having the relative distribution of one data vector component. Local correlations can also occur if two parameter planes are similar in some regions. Furthermore, both linear and non-linear correlations including local or partial correlations between variables are possible[65]. We achieved the correlation hunting automatically, by decoupling the SOM correlations for correlation coefficients of at least 0.60, to explore patterns among pairwise genome samples for distinct identification of transmission pathways or routes. As can be seen in Fig 12, the SOM component planes reveal both inter- and intra-country sub-strains transmission for male and female patients. For male patients, the number of (sub-strains/isolate count) discovered by country include: Algeria (1/2), Benin (3/5), Cameroon (1/1), DRC (4/15), Egypt (3/5), Gambia (3/6), Ghana (2/4), Kenya (2/4), Mali (1/5), Morocco (3/5), Nigeria (5/13), Senegal (5/14), South Africa (7/65), and Tunisia (1). For female patients, the number of (sub-strains/isolate count) discovered by country include: Algeria (1/1), Benin (2/4), DRC (3/13), Egypt (2/4), Gambia (1/4), Ghana (2/4), Kenya (3/4), Madagascar (2/2), Mali (3/4), Morocco (2/2), Nigeria (5/11), Senegal (5/11), South Africa (9/80), and Tunisia (1).

Fig. 14. SOM component planes visualization. Maps are ordered by countries, with at least 1 isolate per country. The isolate numbers (1-145) represent the various states of the country excavated from GSAID (see SupplData1_1.xlsx). The isolates are distributed by countries as follows. Male: Algeria (1-2), Benin (3-7), Cameroon (8), DRC (9-23), Egypt (24-28), Gambia (29-34), Ghana (35-38), Kenya (39-42), Mali (43-47), Morocco (48-52), Nigeria (53-65), Senegal (66-79), South Africa (80-144), Tunisia (145). Female: Algeria (1), Benin (2-5), DRC (6-18), Egypt (19-22), Gambia (23-26), Ghana (27-30), Kenya (31-34), Madagascar (35-36), Mali (37-40), Morocco (41-42), Nigeria (43-53), Senegal (54-64), South Africa (65-144), Tunisia (145).

Discovering transmission routes of a pandemic can be very challenging but could assist both inter- and intra-country contact tracing. Using the Python programming language, cognitive knowledge was mined to localize the transmission routes of every isolate and provide appropriate link(s) to similar isolates with identical pattern(s). We observed multiple inter- and intra-country transmissions, with 10 and 12 sub-strains and their variants. This further knowledge was filtered out from the SOM component planes visualization of the male and

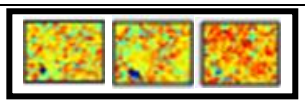
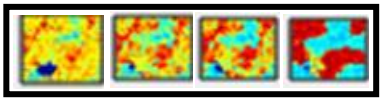
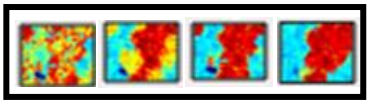
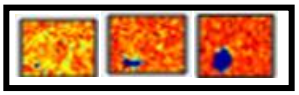
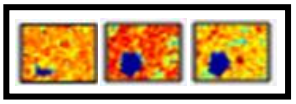
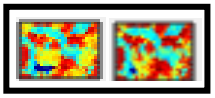
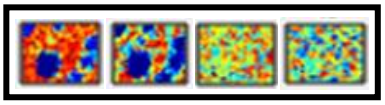
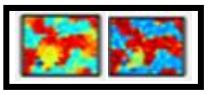
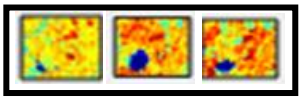
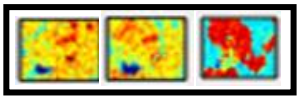
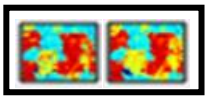
736 female isolates (Fig. 14), and presented in Table 4 and Table 5, respectively. Although there were noise infested
737 genomes (a product of genome sequencing and other unseen defects, which contributed to altering the SOM
738 image, causing semblance of dark blue like clots or stains (e.g., clusters 2-7, of Table 4, and clusters 2-8, 10-12,
739 of Table 5), they did not however, significantly alter the observed pattern(s).
740
741 **Table 4. Discovered SARS-CoV-2 sub-strain clusters and cluster links of male isolates**

Cluster	Filtered pattern	Isolate link cluster
1.		1, 2,68,72
2.		3,29,32,33,34,42,43, 44,45,46,47,65
3.		4,5,6,8,13,14,15,18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144
4.		7,17,28,35,37,38,39,40,41,48,54,55,58,59, 60,62,66,67, 71,74,89,102,112, 113,115,116,122, 125,128,140,142,145
5.		9,10,11,12,73,76,77,78,81,87,88, 90,91,92,93,95,96, 101,104,106,109, 127,129,134,136,
6.		16,21,22,31,36,75,80,84,94,100,105,107,1 14,120,124,133
7.		25,26,27,51,52,103
8.		53,123,126,131
9.		57,85,86,108,110
10.		97,98,121,130

742

743

744 **Table 5. Discovered SARS-CoV-2 sub-strain clusters and cluster links of female isolates**

Cluster	Filtered pattern	Isolate link cluster
1.		1,8,22,48,50,51,55,61,81,86,121,124,128,131,137
2.		2,3,5,23,24,25,26,27,28,33,34,39,41,52,53,145
3.		4,10,15,18,19,20,21,31,40,46,57,58,63,66,67,68,70,71,72,78,79,82,9 3,100,103,104,112,127, 133,136,140,143
4.		6,9,11,12,13,14, 16,17,59,60,62,92, 94,96,97,98,130
5.		7,35,37,38,45, 54,56,64,73,77,101,102,105,107,116,117,129,139
6.		29,30,32,42,108,113
7.		36,69,75,76,122 135,138,142
8.		43,44,47,49,74, 80,99,114,126
9.		65,123,134,141,144
10.		83,84,89,90,91,106,109
11.		85,88,110,111,115,119, 125
12.		87,95,118,120,132

745

[Table 2, here]

[Table 3, here]

Test for Statistical Significance

We conduct Friedman's test[66], a non-parametric statistical test similar to the parametric repeated measures ANOVA and used for detecting differences in treatments across multiple test attempts. The procedure ranks each group of isolates (or block) together, and then considers the values of ranks by country (for inter-country analysis) or by states (for intra-country analysis). The Nemenyi's post hoc test for critical difference (CD)[67] was performed where the overall groups significantly differed from the observed characteristics of the isolates, as detected by the Friedman's test. For the inter-country analysis, we selected countries with up to 13 states (selected randomly), to allow for a balanced block design, hence, resulting in 4 countries (DRC, Nigeria, Senegal and South Africa). For intra-country analysis, we selected the country with the highest number of states have up to 3 samples, to allow for a balanced block design, hence resulting in only 1 country, South Africa.

Inter-country Analysis: From the results of the Friedman test, there is a very highly significant difference in the male isolate groups from the selected countries ($p<0.01$), see Fig. 15a. Moreover, from the Nemenyi post hoc test, the CD plot reveals that isolates in any two countries, from among those listed on the left end of the CD plot (but with the exception of South Africa: Kwazulu Natal isolate) are not significantly different (as evident in the thick horizontal line connecting any pair of lines representing those countries). However, the isolate from South Africa (Kwazulu Natal isolate) is observed to be significantly different from those of Senegal (St. Louis and Diourbel isolates) and Nigeria (Kwara and Oyo isolates), only, among those on the left end of the plot. Similarly, any two countries from among those listed on the right end of the CD plot (with the exception of South Africa: MP isolate) have isolates to be not significantly different. However, Isolates from MP are

777 significantly different from those of Nigeria (unknown) and South Africa (eThekwini), only, among those
778 countries that also appear on the right end of the plot. Moreover, isolates from South Africa (Kwazulu Natal,
779 Harry Gwala and Stanger) are observed to be significantly different from those of each country on the right end
780 of the CD plot, while the isolates from South Africa (Zululand) are significantly different from those of each
781 country on the right end (except Nigeria: unknown isolate) of the plot. Similarly, isolates from South Africa
782 (Ilembe, Cape Town and Amajuba) are significantly different from those of each country on the right end
783 (except Nigeria: unknown isolate, and South Africa: eThekwini isolate) of the plot. Isolates from South Africa
784 (Berea) are significantly different from those of each country on the right end (with the exception of Nigeria:
785 unknown isolate, South Africa: eThekwini isolate, and DRC: Haut-Katanga isolate) of the plot, while isolates
786 from South Africa (King Cetshwayo) are significantly different from those of each country on the right end
787 (with the exception of Nigeria: unknown isolate, and South Africa: eThekwini isolate, DRC: Haut-Katanga
788 isolate, and Nigeria: Osun isolate) of the plot. Isolates from Senegal (Kolda) are observed to be significantly
789 different from each of Daka, South Africa (North West), DRC (Kongo Central) and South Africa (MP), among
790 the countries on the right end of the CD plot; but isolates from each of Senegal (St. Louis and Diourbel), Nigeria
791 (Kwara and Oyo) are significantly different from South Africa (MP), only, among the countries on the right end
792 of the plot.
793 From the results of the Friedman test, there is a very highly significant difference in the female isolates
794 from the selected countries ($p<0.01$), see Fig. 15b. Moreover, from the Nemenyi post hoc test, the CD plot
795 reveals that isolates from any two countries from among those listed on the left end of the CD plot but excluding
796 South Africa (Ugu, Harry Gwala, Amajuba), Nigeria (Osun and Ogun) are not significantly different (as evident
797 in the thick horizontal line connecting any pair of lines representing those countries). However, isolates from
798 any of South Africa (Harry Gwala, Amajuba), Nigeria (Osun and Ogun) are observed to be significantly
799 different from those of each of Senegal (Thies isolate) and South Africa (Ilembe isolate), while the isolates from
800 South Africa (Ugu) are significantly different from the isolates from each of Nigeria (unknow isolate), Senegal
801 (Thies isolate) and South Africa (Ilembe isolate). The isolates from any two countries among those listed on
802 the right end of the CD plot are not significantly different. Moreover, isolates from each of South Africa (Ugu,
803 Harry Gwala, Amajuba), Nigeria (Osun and Ogun) are significantly different from those of any country
804 appearing on the right end of the plot. Similarly, isolates from each of South Africa (Zululand and eThekwini)
805 are observed to be significantly different from those of each country on the right end (except Senegal: Diourbel
806 and Nigeria: Ekiti) of the CD plot, while the isolates from DRC (Haut-Katanga) are significantly different from

those of each country on the right end (except Senegal: Diourbel isolate, Nigeria: Ekiti isolate, DRC: Kinshasha isolate and Senegal: Dakar isolate) of the plot. Isolates from South Africa (King Cetshwayo) are significantly different from those of each country on the right end (except, Senegal: Diourbel isolate, Nigeria: Ekiti isolate, DRC: Kinshasha isolate, Senegal: Dakar and Kongo Central isolates) of the CD plot, while the isolates from each of Nigeria (unknown), Senegal (Thies) and South Africa (Ilembe) differ significantly from those of South Africa (Free State, Cape Town and North West) and Nigeria (Ondo) only, among those on the right end of the plot.

Fig. 15. Inter-country CD plots for male and female patients. For a significance level α the Nemenyi's test determines the critical difference (CD), if the difference between the average ranking of two treatments is greater than CD, then the null hypothesis that the treatments have the same performance is rejected.

Intra-country Analysis: From the results of the Friedman test there is a very highly significant difference in the male isolates from the various isolate groups from selected states ($p<0.01$), see Fig. 16a. Moreover, from the Nemenyi post hoc test, the CD plot reveals that isolates from any two states among Berea, Kwazulu Natal, , Ilembe, Harry Gwala, Umgungundlovu, Zululand, Umkhanyakude and eThekwini (which are the states listed on the left end of the CD plot) are not significantly different (as evident in the thick horizontal line connecting any pair of lines representing those states). Similarly, any two states from among Amajuba, King Cetshwayo, Umzinyathi, Uthungulu, Cape Town and Ugu (the first 6 states on the right end of the CD plot) have their isolates to be not significantly different. However, Isolates from North West are significantly different from those of each state on the right end (except Uthungulu, Cape Town and Ugu) of the CD plot. Moreover, isolates from Berea are found to be significantly different from those of each state on the right end (except Amajuba) of the CD plot, while the isolates from Kwazulu Natal, Ilembe, Harry Gwala and Umgungundlovu are each significantly different from those of each state on the right end (except Amajuba, King Cetshwayo and Umzinyathi) of the CD plot. Similarly, isolates from Zululand are significantly different from those of each state on the right end (except Amajuba, King Cetshawayo, Umzinyathi and Uthungulu) of the plot, while the isolates from Umkhanyakude and eThekwini are each significantly different from those of North West, only.

From the results of the Friedman test there is a very highly significant difference in the female isolate groups from selected states ($p<0.01$), see Fig. 16b. Moreover, from the Nemenyi post hoc test, the CD plot

reveals that isolates from any two states among Umgungundlovu, Harry Gwala , . . ., Umkhanyakude (which are the states listed on the left end of the CD plot) are not significantly different (as evident in the thick horizontal line connecting any pair of lines representing those states). Similarly, any two states from among Uthungulu, eThekwini, . . ., Ilembe (the first 7 states on the right end of the CD plot) have their isolates to be not significantly different. However, Isolates from North West are significantly different from those of each state on the right end (except Capetown, Ilembe and Free State) of the CD plot. Moreover, isolates from Umgungundlovu and Harry Gwala are each found to be significantly different from those of each state on the right end (except Uthungulu, eThekwini, Zululand and Kwazulu Natal) of the CD plot, while the isolates from Berea and Uthukela are each significantly different from those of each of Ilembe, Free State and North West only (which appear on the right end of the plot). Similarly, isolates from Amajuba, King Cetshawayo, Umzinyathi, Ugu and Umkhanyakude are each significantly different from those of each of FS and North West only (both appearing on the right end of the plot).

Fig. 16. Intra-country CD plots for male and female patients. For a significance level α the Nemenyi's test determines the critical difference (CD), if the difference between the average ranking of two treatments is greater than CD, then the null hypothesis that the treatments have the same performance is rejected.

Discussion

Issues of gender and the human genome present several levels of implications at basic scientific research, clinical applications and wider societal investigations[68]. Excluding those with co-morbidities and the aged, males have been found to have worse prognosis during COVID-19 infections[69] with delayed viral clearance compared to females, which show evidence of immune tolerance and slow prognosis[70]. Hence, characterization of the sub-strain clusters by gender clearly explains the diversity of SARS-CoV-2. To the best of our knowledge, research works conducted so far on SARS-CoV-2 genome analysis, aside demographic classification, have ignored the gender dimension. This paper is therefore the first to consider the extent to which SARS-CoV-2 sub-strain transmissions impact on gender. The implication of our findings is most likely to introduce additional information to existing body of knowledge on COVID-19 and aid further research works in this area that would balance the gender dimension.

Understanding the pattern of spontaneous mutation is fundamental in studies of human genome evolution and genetic disease[71]. Mutation diversity therefore appears to be a direct consequence of changing/evolving sub-strains, as it represents the ultimate source of genetic variation and explains the story behind their evolution. However, extensive variability exists among different genes or genome regions in between- and within-species[72,73], and suggests that spontaneous mutation rates are not always constant across the genome–as only a small subset of the new mutation manifests in disease variants[74]. But despite mutation diversity and localized variation in DNA polymorphism diversity and recombination, the pattern of sub-strains is not affected to cause confusion in sub-strain(s) identification. Whole-genome alignment predicts evolutionary relationships at the nucleotide level between two or more genomes[75]. In this paper, homologous pairs of (nucleotide) positions between genome sequences were compared. Aside noise, the genomes are also colinear, as they have not been broken by rearrangement event. Although local alignments between genomes were realized, our alignment algorithm is costly (ran in quadratic time) and requires further improvements, but the similarity scores produced (see Fig. 13) are very useful Knowledge Base component for building intelligent diagnostic systems. Most whole genome algorithms are restricted in the evolutionary relationships captured, as only a subset of homologous relationships may be of interest. Large-scale phylogenomic methodology shows high potential, as application of diverse datasets confirm robustness of the approach.

Wittler[76] proposed a new whole genome-based approach for inferring aligned- and reference-free phylogenies. Their method adopted a colored de Bruijn graph to extract common subsequences for deducing phylogenetic splits, instead of relying on pairwise comparisons to determine distances and tree edges deduction. Similarity in nucleotide sequence is a diversity indicator that measures the relative closeness of gene or genome isolates. Identifying similarities between sequences of special interest is one of the most important goals when working with nucleotide sequences. A distance measure associates the numeric value with a pair of sequences. Direct nucleotide protein sequencing technology[77] have resulted in the explosive growth in the number of known sequences. The results of the nucleotide similarity analysis revealed isolates with strong evolutional relationships between and within countries. Hierarchical clustering that implements agglomerative nesting was adopted in this research for genome-wide ranking instead of focusing on specific subsets. Sub-strains and transmission patterns discoveries confirmed multiple strains with some isolates showing identical sub-strains patterns (with less-diversity), while others showed distinct terminal patterns (without further changes).

SOM results showed reduced sub-strain variants–for increased isolates/states, with disproportionate sub-strains increase or decrease in some states, for male patients (e.g., Gambia=50%, DRC=26.67%,

Nigeria=38.46%, Senegal=35.71% and South Africa=10.77%) and female patients (e.g., Gambia=25%, DRC=23.07%, Nigeria=45.45%, Senegal=45.45% and South Africa=11.25%), hence, establishing a non-linear relationship between mutation and transmission patterns by gender. The generated cognitive maps (Table 2 and Table 3) efficiently associates similar isolate clusters for transmission pathway analysis. The practical implication of this is that early inter- and intra-transmission routes could easily be traced, and immediate contact tracing commenced. Further, countries/states with high prevalence rate could temporarily be locked until satisfactory contact tracing is achieved. The discovered sub-strain variants (10 in male patients and 12 in female patients) reveals the uncertain nature of the SARS-CoV-2 and may be a pointer to a second wave of the virus in Africa.

Finally, the statistical test for significance (Friedman's test) showed highly significant difference for inter- and intra-country analysis by gender, with the Nemenyi's post hoc test revealing significance difference in the countries/states selected.

Conclusions

Infectious disease prediction has significantly benefited from the use of genome mining, which is entirely dependent of computing technology and bioinformatics tools used[78]. The World Health Organization (https://www.afro.who.int/news/covid-19-genome-sequencing-laboratory-network-launches-africa) has underscored the need for application of genome mining in the management of COVID-19 in Africa by collaborating with the Africa Centers for Disease Control and Prevention (Africa CDC) to launch a network of twelve specialized laboratories to facilitate genome sequencing of SARS-CoV-2 virus to track the evolution and mutation of the virus and create an effective mechanism for response to the virus. The grouping of viruses from different countries into lineages has proved useful in establishing the route of virus importation across countries.

In this study, we have introduced a cooperatively inspired open source framework for intelligent mining of SARS-CoV-2 genomes using the unsupervised self-organizing map (SOM), which takes advantage of similarity in genetic behavior of the strains of the SARS-CoV-2 virus. The SOM is among the family of machine learning techniques that facilitate engines that further features probing of genes for precise classification and prediction, which could be useful for screening and treatment, contact tracing, prediction and forecasting, and drugs/vaccine discovery. Our open source framework is a hybridized system that helps an in-depth understanding infectious disease prevalence. Our framework generates phylogenomic trees, pairwise nucleotide similarity

matrices/scores, gene diversity plots, genome expression patterns analysis, essential for enriching the genome datasets, towards intelligent genome characterization and prediction; thus, facilitating community contribution and replicability.

The results of our study show the following: i) Africa countries exhibit varying levels of nucleotide mutation; for example Congo had a high mutation burden (8.30), while Kenya had the least (5.38); ii) The transition from the cytosine to the thymine nucleotide (C>T) accounted for the highest level of mutation, followed by the adenine to guanine (A>G) transition; iii) the average nucleotide count in male and female isolates show approximately similar ratios (A ≈ 32, C ≈ 18, G ≈ 20, T≈ 32); iv) the genome diversity analysis show a smoother distribution curve with male isolates when compared with female isolates; v) the phylogenomic analysis suggests independent sub-strain mutant accumulation in various countries; vi) from the excavated data, various African countries exhibit varying numbers of sub-strain transmissions; South Africa had the highest (male-7, female-9), followed by Nigeria (male-5, female-5), while Tunisia, Algeria had only one sub-strain for both male and female isolates; and vii) multiple inter-country transmissions were observed, with 10 and 12 sub-strains and their variants.

The academic and policy implications of this study are as follows: i) it contributes to a better understanding of the prevalence and transmission of the SARS-CoV-2 virus in Africa; ii) it provides a framework for inter- and intra-country contact tracing, especially in undocumented infection sources; iii) it provides a basis of revealing hidden sub-strains, which could be useful in time varying prediction of infection patterns; iv) the computation of variability in emerging sub-strains by gender isolates could be very useful in the development of appropriate SARS-CoV-2 vaccines.

List of Abbreviations

AGNES	Agglomerative Nesting
AI	Artificial Intelligence
COVID-19	Coronavirus Disease 2019
DNA	Di-Nucleic Acid

GISAID	Global Initiative on Sharing All Influenza Data
HCA	Hierarchical Clustering Analysis
MATLAB	MATrix LABoratory
MERS	Middle East Respiratory Syndrome
MERS-CoV	Middle East Respiratory Syndrome Corona Virus
ML	Machine Learning
NCBI	National Center for Biotechnology Information
ORF	Open Reading Frame
PANGOLIN	Phylogenetic Assignment of Named Global Outbreak LINeages
RDT	Rapid Diagnostic Test
RNA	Ribonucleic Acid
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
SARS	Severe Acute Respiratory Syndrome
SARS-CoV-2	Severe Acute Respiratory Syndrome Corona Virus 2
SOM	Self-Organizing Map

Declarations

None declared

Availability of data and materials

- The datasets used and/or analyzed during the current study are available at: https//gisaid.org
- All data generated or analyzed during this study are included in this published article [and its supplementary information files].

Competing interests

There are no competing interests

Authors' contributions

All authors contributed equally to the final manuscript. Specifically,

M.E.* conceptualized the research idea, contributed to the research methods, preparation of figures, framework/tools design, implementation and interpretation of the results.

M.E.1 provided literature materials, performed critical review as well as data validation.

U.I. contributed to the research methodology, framework/tools design, preparation of figures, implementation and interpretation of results.

F-M.U. structurally edited the original draft and contributed to the tools design component and implementation.

I.U. was involved in critical review of literature and research data validation.

N.U. was involved in the analysis and interpretation of statistical related components.

I.E. contributed to the biotechnology and bioinformatics components of the paper.

A.M. was involved in formal analysis of the study data, research methods and implementation.

E.A. was involved in the critical review as well as a formal analysis of the study data.

Y.T. contributed to the biotechnology components of the paper, including editing of the original draft.

G.J. was involved in data curation, collection/excavation and processing of the human SARS-CoV-2 genomes.

E.D. was involved in data curation, collection/excavation and processing of the human SARS-CoV-2 genomes.

J.N. was involved in critical review of literature and research data validation.

Acknowledgements

We are grateful to the authors, originating and submitting laboratories of the collected sequences from GISAID's EpiFlu database on which this research rests.

References

[1] Fehr AR, Perlman S. Coronaviruses: an overview of their replication and pathogenesis. In Coronaviruses. 2015; (pp. 1-23). Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-2438-7_1.

[2] Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev. 2020; 7: 1012–1023. https://doi.org/10.1093/nsr/nwaa036.

[3] Università di Bologna. The six strains of SARS-CoV-2. ScienceDaily. ScienceDaily, 2020 August 3. https://www.sciencedaily.com/releases/2020/08/200803105246.htm.

[4] Islam H, Rahman A, Masud J, Shweta DS, Araf Y, Ullah MA, Sium SMA, Sarkar, B. A Generalized Overview of SARS-CoV-2: Where Does the Current Knowledge Stand? Electron J Gen Med. 2020; 17 (6): em251. https://doi.org/10.29333/ejgm/8258

[5] Wiechers IR, Perin NC, Cook-Deegan R. The emergence of commercial genomics: analysis of the rise of a biotechnology subsector during the Human Genome Project, 1990 to 2004. Genome Med. 2013; 5(83): 1-9. https://doi.org/10.1186/gm487

[6] Giani AM, Gallo, GR, Gianfranceschi L, Formenti G. Long walk to genomics: History and current approaches to genome sequencing and assembly. Comput. Struct. Biotechnol. 2020; 18. p. 9-19. https://doi.org/10.1016/j.csbj.2019.11.002.

[7] Shey M, Okeibunor JC, Yahaya AA, Herring BL, Tomori O, Coulibaly SO, Gumede-Moeletsi, HN, Mwenda JM, Yoti Z, Wiysonge CS, Talisuna AO. Genome sequencing and the diagnosis of novel coronavirus (SARS-COV-2) in Africa: how far are we?. Pan Afr Med J.2020; 36: 80. https://doi.org/10.11604/pamj.2020.36.80.23723.

[8] Adebisi YA, Oke GI, Ademola PS, Chinemelum, IG, Ogunkola IO, Lucero-Prisno III DE. SARS-CoV-2 diagnostic testing in Africa: needs and challenges. Pan Afr Med J. 2020. 35(2): 4. https://doi.org/10. 11604/ pamj.2020.35.4.22703.

[9] Ekpenyong M, Udo I, Uzoka FM, Attai K. A Spatio-GraphNet Model for Real-time Contact Tracing of CoVID-19 Infection in Resource Limited Settings. In Proceedings of the 4th International Conference on Medical and Health Informatics 2020 pp. 208-217. https://doi.org/10.1145/3418094.3418141.

[10] Das S, Ghosh P, Sen B, Mukhopadhyay, I. Critical community size for CoVID-19--a model-based approach to provide a rationale behind the lockdown. arXiv preprint arXiv. 2020; 2004.03126. https://arxiv.org/pdf/2004.03126.pdf

[11] Maghdid HS, Ghafoor KZ. A Smartphone enabled Approach to Manage CoVID-19 Lockdown and Economic Crisis. arXiv preprint arXiv. 2020; 2004.12240.

[12] Sun L, Liu G, Song F, Shi N, Liu F, Li S, Li P, Zhang W, Jiang X, Zhang Y, Sun L, Chen X, Shi Y. Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19. J Clin Virol. 2020; https://doi.org/10.1016/j.jcv.2020.104431.

[13] Yan L, Zhang HT, Goncalves J, Xiao Y, Wang M, Guo Y, Sun C, Tang X, Jing L, Zhang M, Huang X, Xiao Y, Cao H, Chen Y, Ren T, Wang F, Xiao Y, Huang S, Tan X, Huang N, Jiao B, Cheng C, Zhang Y, Luo A. Mombaerts L, Jin J, Cao Z, Li S., Xu H, Yuan Y. An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell. 2020; https://doi.org/10.1038/s42256-020-0180-7.

[14] Ren, L.L., Wang, Y.M., Wu, Z.Q., Xiang, Z.C., Guo, L., Xu, T., Jiang, Y.Z., Xiong, Y., Li, Ribeiro M. H. D. M., da Silva R. G., Mariani V. C., Coelho L. D. S. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. Chaos, Solitons Fractals. https://doi.org/10.1016/j.chaos.2020.109853.

[15] Cabore JW, Karamagi HC, Kipruto H, Asamani JA, Droti B, Seydi ABW, Titi-Ofei R, Impouma B, Yao M, Yoti Z, Zawaira F. The potential effects of widespread community transmission of SARS-CoV-2 infection in the World Health Organization African Region: a predictive model. BMJ Global Health. 2020;5(5), e002647. doi:10.1136/bmjgh-2020-002647.

[16] Sun H, Dickens BL, Cook AR, Clapham HE. Importations of COVID-19 into African countries and risk of onward spread. BMC Infect Dis. 2020; 20, 598. https://doi.org/10.1186/s12879-020-05323-w.

[17] Ke YY, Peng TT, Yeh TK, Huang WZ, Chang SE, Wu SH, Hung HC, Hsu TA, Lee SJ, Song JS, Lin WH, Chiang TJ, Lin JH, Sytwu HK, Chen CT. Artificial intelligence approach fighting COVID-19 with repurposing drugs. Biomed J. 2020; https://doi.org/10.1016/j.bj.2020.05.001.

[18] Beck BR, Shin B, Choi Y, Park S, Kang K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. Comput Struct Biotechnol J. 2020;18: 784–790. https://doi.org/10.1016/j.csbj.2020.03.025.

1075 [19] Ekins S, Mottin M, Ramos PRP, Sousa BKP, Neves BJ, Foil DH, Zorn KM, Braga RC, Coffee M, Southan
1076 C, Puhl CA, Andrade CH. Déjà vu: stimulating open drug discovery for SARS-CoV-2. Drug Discov Today.
1077 2020; https://doi.org/10.1016/j.drudis.2020.03.019.
1078 [20] Mainardes RM, Diedrich C. The potential role of nanomedicine on COVID-19 therapeutics. Therapeutic
1079 Delivery. 2020; 11(7): 411-414. https://doi.org/10.4155/tde-2020-0069.
1080 [21] Chauhan G, Madou MJ, Kalra S, Chopra V, Ghosh D, Martinez-Chapa, SO. Nanotechnology for COVID-
1081 19: therapeutics and vaccine research. ACS nano. 2020; 14(7), 7760-7782.
1082 https://doi.org/10.1021/acsnano.0c04006.
1083 [22] Caccuri F, Zani, A, Messali S, Giovanetti M, Bugatti A, Campisi G, Filippini F, Scaltriti E, Ciccozzi M,
1084 Fiorentini S, Caruso A. A persistently replicating SARS-CoV-2 variant derived from an asymptomatic
1085 individual. J Transl Med. 2020. 18(1), 1-12. https://doi.org/10.1186/s12967-020-02535-1.
1086 [22] Wu D, Wu T, Liu Q, Yang Z. The SARS-CoV-2 outbreak: what we know. International Journal of
1087 Infectious Diseases. 2020; 94 (2020):44-48. https://doi.org/10.1016/j.ijid.2020.03.004.
1088 [23] Wilder-Smith A, Chiew, CJ, Lee VJ. Can we contain the COVID-19 outbreak with the same measures as
1089 for SARS?. The Lancet Infectious Diseases. 2020; 20: e102–07. https://doi.org/10.1016/S1473-
1090 3099(20)30129-8.
1091 [24] Rabi FA, Al-Zoubi MS, Kasasbeh, GA, Salameh DM, Al-Nasser, AD. SARS-CoV-2 and coronavirus
1092 disease 2019: what we know so far Pathogens. 2020; 9(3), 231. https://doi.org/10.3390/pathogens9030231
1093 [25] Kamatenesi-Mugisha M, Oryem-Origa, H. Traditional herbal remedies used in the management of sexual
1094 impotence and erectile dysfunction in western Uganda. Afr. Health Sci. 2005;5(1): 40–49.
1095 https://www.ajol.info/index.php/ahs/article/view/6896.
1096 [26] Nordling L. Unproven herbal remedy against COVID-19 could fuel drug-resistant malaria, scientist warn.
1097 Science. 2020; https://doi.org/10.1126/science.abc6665.
1098 [27] Li X, Giorgi EE, Marichannegowda, MH, Foley B, Xiao C, Kong XP, Chen Y, Gnanakaran S, Korber B,
1099 Gao F. Emergence of SARS-CoV-2 through recombination and strong purifying selection. Sci. Adv. 2020;
1100 6(27). https://doi.org/10.1126/sciadv.abb9153.
1101 [28] Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song, H, Huang B, Zhu, N, Bi, Y. Genomic
1102 characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor
1103 binding. Lancet Glob Health. 2020; 395(10224): 565-574.

[29] Zhang L, Yang, JR, Zhang Z, Lin Z. Genomic variations of SARS-CoV-2 suggest multiple outbreak sources of transmission. medRxiv. 2020; https://doi.org/10.1101/2020.02.25.20027953.

[30] Korber B, Fischer W, Gnanakaran SG, Yoon H, Theiler J, Abfalterer W, Foley B, Giorgi EE, Bhattacharya T, Parker MD, Partridge DG. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv. 2020; https://doi.org/10.1101/2020.04.29.069054

[31] Wyllie AL, Fournier J, Casanovas-Massana A, et al. Saliva is more sensitive for SARS-CoV-2 detection in CoVID-19 patients than nasopharyngeal swabs. medRxiv.2020; https://doi.org/10.1101/2020.04.16.20067835

[32] Chan KH, Poon LL, Cheng VC, Guan Y, Hung IF, Kong J, Yam LY, Seto WH, Yuen KY, Peiris JS. Detection of SARS coronavirus in patients with suspected SARS. Emerg. Infect. Dis. 2004; 10(2), 294-299. https://doi.org/10.3201/eid1002.030610

[33] Tabibzadeh A, Zamani F, Laali A, Esghaei M, Tameshkel FS, Keyvani H, Makiani MJ, Panahi M, Motamed N, Perumal D, Khoonsari M. SARS-CoV-2 molecular and phylogenetic analysis in COVID-19 patients: a preliminary report from Iran. Infection, Genetics and Evolution. 2020; p.104387. https://doi.org/10.1016/j.meegid.2020.104387.

[34] van der Made CI, Simons A, Schuurs-Hoeijmakers J, van den Heuvel G, Mantere T, Kersten S, van Deuren, RC, Steehouwer M, van Reijmersdal, SV, Jaeger M, Hofste T. Presence of Genetic Variants Among Young Men With Severe COVID-19. JAMA. 2020; 324(7):663-673. https://doi:10.1001/jama.2020.13719.

[35] Torti C, Mazzitelli M, Trecarichi EM, Darius O. Potential implications of SARS-CoV-2 epidemic in Africa: where are we going from now?. BMC Infect Dis. 2020; 20 (412). https://doi.org/10.1186/s12879-020-05147-8.

[36] Sekizuka T, Kuramoto S, Nariai E, Taira M, Hachisu Y, Tokaji A, Shinohara M, Kishimoto T, Itokawa K, Kobayashi Y, Kadokura K. SARS-CoV-2 Genome Analysis of Japanese Travelers in Nile River Cruise. Frontiers in Microbiology. 2020; 11(1316). https://doi.org/10.3389/fmicb.2020.01316.

[37] Kandeil A, Mostafa A, El-Shesheny R, Shehata M, Roshdy WH, Ahmed SS, Gomaa M, El Taweel A, Kayed A E, Mahmoud SH, Moatasim Y, Kutkat O, Kamel MN, Mahrous N, El Sayes M, El Guindy NM, Naguib A, Ali MA. Coding complete genome sequences of two SARSCoV-2 isolates from Egypt. Microbiol. Resour. Announc. 2020; 9: e00489-20. https://doi.org/10.1128/MRA.00489-20.

[38] Laamarti M, Kartti S, Alouane T, Laamarti R, Allam L, Ouadghiri M, Chemao-Elfihri MW, Smyej I, Rahoui J, Benrahma H, Diawara, I. Genetic analysis of SARS-CoV-2 strains collected from North Africa: viral origins and mutational spectrum. bioRxiv, 2020; https://doi.org/10.1101/2020.06.30.181123.

[39] Badaoui B, Sadki K, Talbi C, Tazi L, Salah D. Genetic diversity and genomic epidemiology of sars-cov-2 in Morocco. bioRxiv. 2020; https://doi.org/10.1101/2020.06.23.165902.

[40] Laamarti M, Chemao-Elfihri MW, Kartti S, Laamarti R, Allam L, Ouadghiri M, Smyej I, Rahoui J, Benrahma H, Diawara I, Alouane T. Genome Sequences of Six SARS-CoV-2 Strains Isolated in Morocco, Obtained Using Oxford Nanopore MinION Technology. Microbiol. Resour. Announc. 2020; 9(32): 1-4. https://doi.org/10.1128/MRA.00767-20.

[41] Jouali F, Marchoudi N, El Ansari FZ, Kasmi Y, Chenaoui M, El Aliani A, Azami N, Loukman S, Ennaji MM, Benhida R, Fekkak J. SARS-CoV-2 Genome Sequence from Morocco, Obtained Using Ion AmpliSeq Technology. Microbiol. Resour. Announc. 2020; 9(31): 1-3. https://doi.org/10.1128/MRA.00690-20.

[42] Awoyelu EH, Oladipo EK, Adetuyi BO, Senbadejo TY, Oyawoye OM, Oloke J K. Phyloevolutionary analysis of SARS-CoV-2 in Nigeria. New microbes and new infections. 2020; 36(100717). https://doi.org/10.1016/j.nmni.2020.100717.

[43] Happi C, Ihekweazu C, Oluniyi PE, Olawoye I. New SARS-CoV-2 Genomes from Nigeria Reveals Dominance of Viruses with Spike Protein Mutation (D614G), and Additional Virus Lineages in Circulation. Genome Reports. 2020; https://virological.org/t/new-sars-cov-2-genomes-from-nigeria-reveals-dominance-of-viruses-with-spike-protein-mutation-d614g-and-additional-virus-lineages-in-circulation/527.

[44] Allam M, Ismail A, Khumalo ZTH, Kwenda S, van Heusden P, Cloete R, Wibmer CK, Mtshali PS, Mnyameni F, Mohale T, Subramoney K, Walaza S, Ngubane W, Govender N, Motaze NV, Bhiman J.N. Genome sequencing of a severe acute respiratory syndrome coronavirus 2 isolate obtained from a South African patient with coronavirus disease 2019. Microbiol. Resour. Announc. 2020; 9:e00572-20. https://doi.org/10.1128/MRA.00572-20.

[45] Pillay S, Giandhari J, Tegally H, Wilkinson E, Chimukangara B, Lessells R, Mattison S, Moosa Y, Gazy I, Fish, M, Singh L. Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation During a Pandemic. bioRxiv.2020; https://doi.org/10.1101/2020.06.10.144212.

[46] Bugembe, DL, Kayiwa J, Phan MV, Tushabe P, Balinandi, S, Dhaala, B, Lexow, J, Mwebesa, H, Aceng, J, Kyobe, H, Ssemwanga, D. Main Routes of Entry and Genomic Diversity of SARS-CoV-2, Uganda. Emerging Infectious Diseases. 2020; 26(10): 2411–2415. https://doi.org/10.3201/eid2610.202575.

[47] Hamidouche M. COVID-19 outbreak in Algeria: A mathematical model to predict the incidence. medRxiv. 2020; https://doi.org/10.1101/2020.03.20.20039891.

[48] Kouriba B, Duerr A, Rehn A, Sangare AK, Traoure BY, Bestehorn-Willmann MS, Ouedraogo JL, Heitzer A, Sogodogo E, Maiga A, Walter MC. First phylogenetic analysis of Malian SARS-CoV-2 sequences provide molecular insights into the genomic diversity of the Sahel region. medRxiv. 2020; https://doi.org/10.1101/2020.09.23.20165639

[49] TIBA COVID-19 Pandemic Response Unit (2020). SARS-CoV-2 genomes report for WHO Africa Region [21/08/2020]. http://tiba-partnership.org/tiba/sites/sbsweb2.bio.ed.ac.uk.tiba/files/pdf/SARS-CoV-2%20Genome%20Report%2024.08.2020.pdf

[50] Muyembe-Tamfum, JJ, Ahuka-Mundeke S, Mbala-Kingebeni P, Nkwembe-Mgabana E, Kinganda-Lusamaki E, Amuri-Aziza A, Muyembe-Mawete F, Lokilo-Lofiko E, Claude CCJ, Marceline AO, Bibiche NM. Phylogenetic analysis of SARS-CoV-2 in DRC. ARTIC Network. 2020; https://virological.org/t/phylogenetic-analysis-of-sars-cov-2-in-drc/528

[51] KEMRI-CGMRC Introduction and local transmission of SARS-CoV-2 cases in Kenya. Genome Reports. 2020; https://virological.org/t/introduction-and-local-transmission-of-sars-cov-2-cases-in-kenya/497.

[52] Giandhari J, Pillay S, Wilkinson E, Tegally H, Sinayskiy I, Schuld M, Lourenço J, Chimukangara B, Lessells, RJ, MoosaY, Gazy, I, Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report. medRxiv. 2020; https://doi.org/10.1101/2020.05.29.20116376.

[53] Fares WA, Kais C, Ghedira, M Dorra G, Sondes R, Imene HB, Henda BD Walid T, ,Zina H, Amel M, Nahed S, Imen H, Aurelia A, Veronique K, Guillain H, Valerie M, Jean-Claude C, Nissaf M, Alaya B, Triki H. 2020. First whole genome sequences and phylogenetic analysis of SARS-CoV-2 virus isolates during COVID-19 outbreak in Tunisia, North Africa. Authorea Preprints. https://doi.org/10.22541/au.159137642.26983355.

[54] Yin C. Genotyping coronavirus SARS-CoV-2: methods and implications. Genomics. 2020; 112(5): 3588-3596. https://doi.org/10.1016/j.ygeno.2020.04.016.

[55] Wang C, Liu Z, Chen Z, Huang X, Xu M, He T, Zhang, Z. The establishment of reference sequence for SARS-CoV-2 and Fvariation analysis. Journal of J Med Virol. 2020; 92(6):667-674. https://doi.org/10.1002/jmv.25762.

[56] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao, ZW, Tian JH, Pei,YY, Yuan ML. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798), 265–269. https://doi.org/10.1038/s41586-020-2008-3.

[57] Wang Y, Mao JM, Wang GD, Qiu Z, Yao Q, Chen KP. Human SARS-CoV-2 has evolved to reduce CG dinucleotide in its open reading frames. Sci. Rep. 2020;10(12331). https://doi.org/10.1038/s41598-020-69342-y.

[58] Mercatelli D, Giorgi FM. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Front. Microbiol. 2020; https://doi.org/10.3389/fmicb.2020.01800.

[59] Sjaarda CP, Rustom N, Huang D, Perez-Patrigeon S, Hudson ML, Wong H, Guan TH, Ayub M, Soares CN, Colautti RI, Evans GA. Chasing the origin of SARS-CoV-2 in Canada's COVID-19 cases: A genomics study. bioRxiv. 2020; https://doi.org/10.1101/2020.06.25.171744.

[60] De Maio N, Walker C, Borges R, Weilguny L, Slodkowicz G, Goldman, N. Issues with SARS-CoV-2 sequencing data. 2020; https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.

[61] Kuehn BM. Genetic Analysis Tracks SARS-CoV-2 Mutations in Human Hosts. JAMA. 2020; 323(23), 2363-2363. https://doi.org/10.1001/jama.2020.9825.

[62] Ackermann MR, Blömer J, Kuntze D, Sohler C. Analysis of agglomerative clustering. Algorithmica. 2014 69(1), 184-215. https://doi.org/10.1007/s00453-012-9717-4.

[63] Inyang UG, Akpan EE, Akinyokun OC. A Hybrid Machine Learning Approach for Flood Risk Assessment and Classification. J. Comput. 2020; 19(2), 1-20. https://doi.org/10.1142/S1469026820500121.

[64] Abascal F, Zardoya R, Telford M J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. Spec. Publ. 2010; 38(suppl_2), W7-W13. https://doi.org/10.1093/nar/gkq291.

[65] Vesanto J, Ahola J. Hunting for Correlations in Data Using the Self-Organizing Map. Proceeding of the International ICSC Congress on Computational Intelligence Methods and Applications.1999; p. 279–285.

[66] Friedman M. A comparison of alternative tests of significance for the problem of m rankings. The Annals of Mathematical Statistics. 1940. 11(1):86-92. https://www.jstor.org/stable/2235971.

[67] Pohlert T. The pairwise multiple comparison of mean ranks package (PMCMR). R package. 2014. 1-9.

[68] Chadwick, R. Gender and the human genome. Mens Sana Monographs.2009; 7(1), 10-19. https://doi.org/10.4103 /0973 -1229.44075.

[69] Chen, X., Hu, W., Ling, J., Mo, P., Zhang, Y., Jiang, Q., Ma, Z., Cao, Q., Deng, L., Song, S. and Zheng, R. Hypertension and diabetes delay the viral clearance in COVID-19 patients. medRxiv. (2020). https://doi.org/10.1101/2020.03.22.20040774.

[70] Shastri, A., Wheat, J., Agrawal, S., Chaterjee, N., Pradhan, K., Goldfinger, M., Kornblum, N., Steidl, U., Verma, A. and Shastri, J. Delayed clearance of SARS-CoV2 in male compared to female patients: High ACE2 expression in testes suggests possible existence of gender-specific viral reservoirs. (2020). medRxiv. https://doi.org/10.1101/2020.04.16.20060566.

[71] Schaibley VM, Zawistowski, M., Wegmann, D., Ehm, M. G., Nelson, M. R., Jean, P. L. S., Abecasis, G. R., Novembre, J., Zöllner, S. and Li, J. Z. (2013). The influence of genomic context on mutation patterns in the human genome inferred from rare variants. Genome Res. 23(12): 1974-1984. https://doi.org/doi/10.1101/gr.154971.113.

[72] Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. Genetics. 2000;156(1), 297-304. https://www.genetics.org/content/genetics/156/1/297.full.pdf.

[73] Hodgkinson A, Ladoukakis E, Eyre-Walker A. Cryptic variation in the human mutation rate. PLoS Biol. 2009; 7(2), e1000027. https://doi.org/10.1371/journal.pbio.1000027

[74] Nachman MW. Haldane and the first estimates of the human mutation rate. J. Genet. 2004; 83(3), 231-233. httpa://doi.org/10.1007/bf02717891.

[75] Armstrong J, Fiddes IT, Diekhans M, Paten B. Whole-genome alignment and comparative annotation. Annu Rev Anim Biosci. 2019; 7: 41-64. https://doi.org/10.1038/nmeth.1179.

[76] Wittler, R. Alignment-and reference-free phylogenomics with colored de Bruijn graphs. Algorithms Mol Biol. 2020; 15, 1-12. https://doi.org/10.1186/s13015-020-00164-3.

[77] Feng S, Zhongxi M. Similarity among nucleotide sequences. Acta Biotheor. 2002;50(2):95-99. https://doi.org/10.1023/a:1016376910987.

[78] Amagata T. Natural Products Structural Diversity-II Secondary Metabolites: Sources, Structures and Chemical Biology. Comprehensive Natural Products II. 2010; 2:581-621.

[79] Koyama T, Platt D, Parida L. Variant analysis of SARS-CoV-2 genomes. Bull World Health Organ. 2020; 98(7): 495–504. https://doi.org/10.2471/BLT.20.253591.

1248

1249 **Table 1. Summary of data source, transmission history and identified intra-country sub-strains of African genome isolates**

Author and year	Country/ countries	Isolates considered and source	Transmission history	Identified intra-country sub-strains	Additional information
[47]	Algeria	Information on cases with confirmed COVID-19 infection based on official reports from governmental institutes in Algeria	First case of COVID-19 was reported on when an Italian national tested positive in Ouargla region in the south of the country, later, 2 cases were reported in Blida region in the North of Algeria, following their contacts with two Algerian nationals who came from France.	No information	-
[48]	Mali	21 whole genome sequences were generated from 38 positive isolates	First two COVID-19 cases living in Bamako and Kayes, both returning from France were confirmed. Further 2 lineages, 1 (Asia, Europe, Mali, Oceania) and 2 (Europe, USA, Canada, Northern Africa), were identified	Patients returning from Tunisia, were identified.	The presence of 2 lineages strongly suggests at least two different and independent introduction points of the SARS-CoV-2 infection in the Sahel region.
[49]	15 African countries (Algeria, Benin, Cameroon, Democratic Republic of the Congo (DRC), Gambia, Ghana, Kenya, Madagascar, Mali, Nigeria, Senegal, Sierra Leone, South Africa, Uganda, Zambia.)	SARS-CoV-2 genomes (n=1340) from 15 countries in WHO Africa region (out of 47), excavated from GSAID database on August 17, 2020, and representing ~2% of publicly available sequences globally. Majority of genomes were from South Africa (51%), DRC (20%) and Senegal (10%).	Multiple separate introductions into Africa from other continents, of which 72% came from Europe, 19% from Asia, 6% from South America and 3% from North America.	An estimated 123 introductions came from other continents while 14 introductions were between African countries.	93% of African SARS-CoV-2 genomes have the D614G mutation in spike protein.

Author and year	Country/ countries	Isolates considered and source	Transmission history	Identified intra-country sub-strains	Additional information
[38]	North Africa (Morocco, Tunisia, Algeria, Egypt)	40 SARS-CoV-2 genomes (28 from Morocco, 7 from Tunisia, 3 from Algeria, 2 from Egypt), were downloaded from GISAID between March 03, 2020 and May 15, 2020.	Phylogenetic analysis showed that the Moroccan and Tunisian SARS-CoV-2 strains were closely related to those from different origins (Asia, Europe, North and South America) and distributed in different distinct subclades.	Moroccan and Tunisian strains were closely related to those from different continents, which could indicate different sources of infection with no single dominant strain circulating in Morocco.	2 waves of SARS-CoV-2 infections were recorded, with the first one mostly imported from Europe and the second one dominated by local infections.
[40]	Morocco	6 genome sequences of SARS-CoV-2 strains deposited in DDBJ/ENA/GenBank and GenBank.	The mutation in Morocco was associated with the observed transmission increase in the United States	No information	Mutation analysis revealed the presence of the spike D614G mutation in all six genomes, which is widely present in several genomes around the world.
[39]	Morocco	22 genome sequences reported by three different laboratories in Morocco up to June 7, 2020, as well as 40366 genomes from all around the world.	Introduction of SARS-CoV-2 strains to Morocco came from Belgium, Spain and France at the beginning of the epidemic. Later, strains from USA and Vietnam were noticed after the lockdown, possibly, through sea trades.	Spread in Morocco did not show a predominant SARS-CoV-2 route during analysis.	Different SARS-CoV-2 strains, with different mutation patterns, coexist in Morocco.
[41]	Morocco	1 isolate obtained from Moroccan patient infected in Casablanca.	Sequence belonged a clade which is similar to cases in Forida	No information	Currently circulating strains in Morocco came from different countries with a local evolution.
[50]	DRC	127 SARS-CoV-2 genome sequences excavated from GISAID database	7 distinct lineages which diversity originated from China, are distributed across 4 continents as follows: Asia (China, India, Jordan, Iran);	Ghana constituted one of the large lineage group.	Sampled cases from the DRC are the result of repeated introduction of the virus from a range of locations

Author and year	Country/ countries	Isolates considered and source	Transmission history	Identified intra-country sub-strains	Additional information
			Europe (Italy, Austria, UK); Australia; North America (USA); Africa (Ghana)		followed by local transmission.
[37]	Egypt	2 isolates obtained from Egyptians residing in Upper and Lower Egypt	Egyptian clades fell into A2a denoting strains from Asia, Europe, Africa, Australia and USA	Senegal and DRC were the closely associated strains from Africa.	Sequence analysis showed mutations that differentiate Egyptian strains from the reference strain 2019-nCoV WHU01.
[51]	Kenya	122 genome sequences from Nairobi (n=102) and Coastal Kenya (n=20) collected between 12th March and 30th April 2020.	Evidence of 10 global lineages in China with several global exports to Iran, France, UK and Italy, with multiple introduction of European-centric lineages into the country.	Local transmission, between Nairobi and Mombasa, were found. Infection from the Coastal Kenya was imported from Nairobi	Sequencing of additional SARS-CoV-2 genomes in Kenya will provide a more detailed picture of local transmission patterns.
[43]	Nigeria	24 genome sequences (18 full, 6 partial) were assembled for experiment	7 lineages circulating in Nigeria are from Asia, Europe, USA, Middle East, Australia and Africa.	Strains from Egypt, Senegal and Ghana were revealed through Phylogenetic analysis	Four of the new sequences clustered closely together and formed a separate clade which strongly suggests local community transmission.
[42]	Nigeria	39 complete genomes of SARS-CoV-2 were retrieved from GISAID	First confirmed case in Nigeria came from an infected traveler from Italy.	The strain from Nigeria was found in the Wuhan subclade 3 together with some strains from Congo.	Strain in Nigeria clustered in a monophyletic clade with a Wuhan sublineage.
[52]	South Africa	21 SARS-CoV-2 whole genome, sampled in the first port of entry, KwaZulu-Natal (KZN), during the first month of the epidemic.	First COVID-19 case was a South African citizen returning home from a skiing holiday in Italy.	No information	13 independent introductions in KZN, which lineages revealed imported infection from Europe and North America.

Author and year	Country/ countries	Isolates considered and source	Transmission history	Identified intra-country sub-strains	Additional information
[44]	South Africa	Complete genome sequence of a SARS-CoV-2 isolate obtained from a South African patient.	The patient returned from Italy	No information	This sequence has been deposited in GenBank
[53]	Tunisia	SARS-CoV-2 strains were obtained from imported and locally transmission cases	Patients had travel history from Vietnam, Turkey and France	The Turkey sequence grouped with sequences showing lineage from Egypt	272 imported and 799 locally transmitted cases, have been diagnosed and tested positive for SARS-CoV-2
[46]	Uganda	20 genomic sequences from 14 persons entering Uganda	Miami to Istanbul, UK to Netherlands to Rwanda, Kenya, Tanzania	Kenya, Tanzania, and South Sudan	6 lineages among were imported into Uganda.

Table 2. Cognitive map for male isolates

Algeria	Algeria	Benin	Benin	Benin	Benin	Benin	Cameroon	DRC	DRC
Blida (1)	Blida (2)	Cotonou (3)	Cotonou (4)	Cotonou (5)	Cotonou (6)	Oueme (7)	Yaounde (8)	Ht-Katanga (9)	Ht-Katanga (10)
2,68,72	1,68,72	29,32,33,34,42,43, 44,45,46,47,65	5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144	4,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144	4,5,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,116,122, 125,128,140,142,145	4,5,6,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144	10,11,12,73, 76,77,78,81,87,88, 90,91,92,93,95,96, 101,104,106,109, 127,129,134,136, 137,138	9,11,12,73, 76,77,78,81,87,88, 90,91,92,93,95,96, 101,104,106,109, 127,129,134,136, 137,138
DRC	DRC	DRC	DRC	DRC	DRC	DRC	DRC	DRC	DRC
Ht-Katanga (11)	Kinshasha (12)	Kinshasha (13)	Kinshasha (14)	Kinshasha (15)	Kongo Central (16)	Kongo Central (17)	Kongo Central (18)	Kongo Central (19)	Nord Kivu (20)
9,10,12,73, 76,77,78,81,87,88, 90,91,92,93,95,96, 101,104,106,109, 127,129,134,136, 137,138	9,10,11,73, 76,77,78,81,87,88, 90,91,92,93,95,96, 101,104,106,109, 127,129,134,136, 137,138	4,5,6,8,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144	4,5,6,8,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144	4,5,6,8,13,14, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144	21,22,31,36,75, 80,84,94,100,105, 107,114,120,124,133	7,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,116,122, 125,128,140,142,145	4,5,6,8,13,14,15, 19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144	4,5,6,8,13,14,15, 18,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144	4,5,6,8,13,14,15, 18,19,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144
DRC	DRC	DRC	Egypt	Egypt	Egypt	Egypt	Egypt	Gambia	Gambia
Sud Kivu (21)	Sud Kivu (22)	Sud Kivu (23)	Cairo (24)	Cairo (25)	Cairo (26)	Cairo (27)	Kalyoubia (28)	Kombo (29)	Kombo (30)
16,22,31,36,75, 80,84,94,100,105, 107,114,120,124,133	16,21,31,36,75, 80,84,94,100,105, 107,114,120,124,133	4,5,6,8,13,14,15, 18,19,20,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99,	4,5,6,8,13,14,15, 18,19,20,23,30, 49,50,56,61,63,64, 69,70,79,82,83,99,	26,27,51,52,103	25,27,51,52,103	25,26,51,52,103	7,17,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112,	3,32,33,34,42,43, 44,45,46,47,65	4,5,6,8,13,14,15, 18,19,20,23,24, 49,50,56,61,63,64, 69,70,79,82,83,99,

		111,117,118,119, 132,135,139, 141, 143,144	111,117,118,119, 132,135,139, 141, 143,144				113,115,116,122, 125,128,140,142,145		111,117,118,119, 132,135,139, 141, 143,144
Gambia Kombo (31)	Gambia Kombo (32)	Gambia W.Coast Reg (33)	Gambia W.Coast Reg (34)	Ghana Greater Ghana (35)	Ghana Greater Ghana (36)	Ghana Greater Ghana (37)	Ghana Greater Ghana (38)	Kenya Unknown (39)	Kenya Unknown (40)
16,21,22,36,75, 80,84,94,100,105, 107,114,120,124,133	3,29,33,34,42,43, 44,45,46,46,65	3,29,32,34,42,43, 44,45,46,46,65	3,29,32,33,42,43, 44,45,46,46,65	7,17,28,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,116,122, 125,128,140,142,145	16,21,22,31,75, 80,84,94,100,105, 107,114,120,124,133	7,17,28,35,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,116,122, 125,128,140,142,145	7,17,28,35,37 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,116,122, 125,128,140,142,145	7,17,28,35,37,38, 40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,116,122, 125,128,140,142,145	7,17,28,35,37,38, 39,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,116,122, 125,128,140,142,145
Kenya Unknown (41)	Kenya Unknown (42)	Mali Bamako (43)	Mali Bamako (44)	Mali Bamako (45)	Mali Bamako (46)	Mali Mopti (47)	Morocco Casablanca (48)	Morocco Ouarzazate (49)	Morocco Rabat (50)
7,17,28,35,37,38, 39,40,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,116,122, 125,128,140,142,145	3,29,32,33,34,43, 44,45,46,47,65	3,29,32,33,34,42, 44,45,46,47,65	3,29,32,33,34,42, 43,45,46,47,65	3,29,32,33,34,42, 43,44,46,47,65	3,29,32,33,34,42, 43,44,45,47,65	3,29,32,33,34,42, 43,44,45,46,65	7,17,28,35,37,38, 39,40,41,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,116,122, 125,128,140,142,145	4,5,6,8,13,14,15, 18,19,20,23,24,30, 50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144
Morocco Rabat (51)	Morocco Rabat (52)	Nigeria Kwara (53)	Nigeria Kwara (54)	Nigeria Kwara (55)	Nigeria Lagos (56)	Nigeria Osun (57)	Nigeria Osun (58)	Nigeria Osun (59)	Nigeria Osun (60)
25,26,27,52,103	25,26,27,51,103	123,126,131	7,17,28,35,37,38, 39,40,41,48,55, 58,59,60,62,66,67,	7,17,28,35,37,38, 39,40,41,48,54, 58,59,60,62,66,67,	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,61,63,64,	85,86,108,110	7,17,28,35,37,38, 39,40,41,48,54,55, 59,60,62,66,67,	7,17,28,35,37,38, 39,40,41,48,54,55, 58,60,62,66,67,	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,62,66,67,

			71,74,89,102,112, 113,115,116,122, 125,128,140,142,145	71,74,89,102,112, 113,115,116,122, 125,128,140,142,145	69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144		71,74,89,102,112, 113,115,116,122, 125,128,140,142,145	71,74,89,102,112, 113,115,116,122, 125,128,140,142,145	71,74,89,102,112, 113,115,116,122, 125,128,140, 142,145
Nigeria Oyo (61)	Nigeria Oyo (62)	Nigeria Oyo (63)	Nigeria Oyo (64)	Nigeria Unknown (65)	Senegal Dakar (66)	Senegal Dakar (67)	Senegal Dakar (68)	Senegal Dakar (69)	Senegal Diourbel (70)
4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,66,67, 71,74,89,102,112, 113,115,116,122, 125,128,140,142,145	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144	3,29,32,33,34,42, 43,44,45,46,47	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,67, 71,74,89,102,112, 113,115,116,122, 125,128,140,142,145	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66, 71,74,89,102,112, 113,115,116,122, 125,128,140,142,145	1,2,72	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144
Senegal Diourbel (71)	Senegal Diourbel (72)	Senegal Kolda (73)	Senegal Pikine (74)	Senegal St Louis (75)	Senegal St Louis (76)	Senegal St Louis (77)	Senegal Thies (78)	Senegal Thies (79)	South Africa Amajuba (80)
7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 74,89,102,112, 113,115,116,122, 125,128,140,142,145	1,2,68	9,10,11,12, 76,77,78,81,87,88, 90,91,92,93,95,96, 101,104,106,109, 127,129,134,136, 137,138	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,89,102,112, 113,115,116,122, 125,128,140,142,145	16,21,22,31,36, 80,84,94,100,105, 107,114,120,124,133	9,10,11,12,73, 77,78,81,87,88, 90,91,92,93,95,96, 101,104,106,109, 127,129,134,136, 137,138	9,10,11,12,73, 76,78,81,87,88, 90,91,92,93,95,96, 101,104,106,109, 127,129,134,136, 137,138	9,10,11,12,73, 76,77,81,87,88, 90,91,92,93,95,96, 101,104,106,109, 127,129,134,136, 137,138	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143,144	16,21,22,31,36,75, 84,94,100,105, 107,114,120,124, 133
South Africa Amajuba (81)	South Africa Amajuba (82)	South Africa Amajuba (83)	South Africa Berea (84)	South Africa Berea (85)	South Africa Berea (86)	South Africa Cape Town (87)	South Africa Cape Town (88)	South Africa Cape Town (89)	South Africa Cape Town (90)

9,10,11,12,73, 76,77,78,87,88, 90,91,92,93,95,96, 101,104,106,109, 127,129,134,136, 137,138	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,83,99, 111,117,118,119, 132,135,139, 141, 143,144	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,99, 111,117,118,119, 132,135,139, 141, 143,144	16,21,22,31,36,75, 80,94,100,105, 107,114,120,124,133	57,86,108,110	57,85,108,110	9,10,11,12,73, 76,77,78,88, 90,91,92,93,95,96, 101,104,106,109, 127,129,134,136, 137,138	9,10,11,12,73, 76,77,78,81,87, 90,91,92,93,95,96, 101,104,106,109, 127,129,134,136, 137,138	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,102,112, 113,115,116,122, 125,128,140,142,145	9,10,11,12,73, 76,77,78,81,87,88, 91,92,93,95,96, 101,104,106,109, 127,129,134,136, 137,138
South Africa EC (91)	South Africa eThekwini (92)	South Africa eThekwini (93)	South Africa eThekwini (94)	South Africa eThekwini (95)	South Africa Harry Gwala (96)	South Africa Harry Gwala (97)	South Africa Harry Gwala (98)	South Africa Harry Gwala (99)	South Africa Ilembe (100)
9,10,11,12,73, 76,77,78,81,87,88, 90,92,93,95,96, 101,104,106,109, 127,129,134,136, 137,138	9,10,11,12,73, 76,77,78,81,87,88, 90,91,93,95,96, 101,104,106,109, 127,129,134,136, 137,138	9,10,11,12,73, 76,77,78,81,87,88, 90,91,92,95,96, 101,104,106,109, 127,129,134,136, 137,138	16,21,22,31,36,75, 80,84,100,105, 107,114,120,124,133	9,10,11,12,73, 76,77,78,81,87,88, 90,91,92,93,96, 101,104,106,109, 127,129,134,136, 137,138	9,10,11,12,73, 76,77,78,81,87,88, 90,91,92,93,95, 101,104,106,109, 127,129,134,136, 137,138	98,121,130	97,121,130	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83, 111,117,118,119, 132,135,139, 141, 143,144	16,21,22,31,36,75, 80.84,94,105, 107,114,120,124, 133
South Africa Ilembe (101)	South Africa Ilembe (102)	South Africa Ilembe (103)	South Africa Kg Cetshwayo (104)	South Africa Kg Cetshwayo (105)	South Africa Kg Cetshwayo (106)	South Africa Kg Cetshwayo (107)	South Africa KZN (108)	South Africa KZN (109)	South Africa KZN (110)
9,10,11,12,73, 76,77,78,81,87,88, 90,91,92,93,95,96, 104,106,109, 127,129,134,136, 137,138	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,112, 113,115,116,122, 125,128,140,142,145	25,26,27,51,52	9,10,11,12,73, 76,77,78,81,87,88, 90,91,92,93,95,96, 101,106,109, 127,129,134,136, 137,138	16,21,22,31,36,75, 80,84,94,100, 107,120,124,133	9,10,11,12,73, 76,77,78,81,87,88, 90,91,92,93,95,96, 101,104,109, 127,129,134,136, 137,138	16,21,22,31,36,75, 80,84,94,100,105, 120,124,133	57,85,86,110	9,10,11,12,73, 76,77,78,81,87,88, 90,91,92,93,95,96, 101,104,106, 127,129,134,136, 137,138	57,85,86,108
South Africa	South Africa	South Africa	South Africa	South Africa	South Africa	South Africa	South Africa	South Africa	South Africa

KZN (111)	MP (112)	North West (113)	North West (114)	North West (115)	North West (116)	Stanger (117)	Ugu (118)	Ugu (119)	Ugu (120)
4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 117,118,119, 132,135,139, 141, 143,144	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102, 113,115,116,122, 125,128,140,142,145	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112,115, 116,122,125,128,140, 142,145	16,21,22,31,75, 80,84,94,100,105, 107,120,124,133	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,116,122, 125,128,140,142,145	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,122, 125, 128,140,142,145	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,118,119, 132,135,139, 141, 143,144	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,119, 132,135,139, 141, 143,144	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118, 132,135,139, 141, 143,144	16,21,22,31,36,75, 80,84,94,100,105, 107,124,133

South Africa Ugu (121)	South Africa Umbilo (122)	South Africa Umgungundl (123)	South Africa Umgungundl (124)	South Africa Umgungundl (125)	South Africa Umgungundl (126)	South Africa Umkhanyak (127)	South Africa Umkhanyak (128)	South Africa Umkhanyak (129)	South Africa Umkhanyak (130)
97,98,130	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,116,125, 128,140,142,145	53,126,131	16,21,22,31,36,75, 80,84,94,100,105, 107,120,133	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,116,122, 128,140,142,145	53,123,131	9,10,11,12,73, 76,77,78,81,87,88, 90,91,92,93,95,96, 101,104,106,109, 129,134,136, 137,138	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,116,122, 125,140,142,145	9,10,11,12,73, 76,77,78,81,87,88, 90,91,92,93,95,96, 101,104,106,109, 127,134,136, 137,138	97,98,121

South Africa Umzinyathi (131)	South Africa Umzinyathi (132)	South Africa Umzinyathi (133)	South Africa Umzinyathi (134)	South Africa Uthukela (135)	South Africa Uthukela (136)	South Africa Uthukela (137)	South Africa Uthukela (138)	South Africa Uthungulu (139)	South Africa Uthungulu (140)
53,123,126	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 135,139, 141, 143,144	16,21,22,31,36,75, 80,84,94,100,105, 107,120,124	9,10,11,12,73, 76,77,78,81,87,88, 90,91,92,93,95,96, 101,104,106,109, 127,129,136, 137,138	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132, 139, 141, 143,144	9,10,11,12,73, 76,77,78,81,87,88, 90,91,92,93,95,96, 101,104,106,109, 127,129,134, 137,138	9,10,11,12,73, 76,77,78,81,87,88, 90,91,92,93,95,96, 101,104,106,109, 127,129,134,136, 138	9,10,11,12,73, 76,77,78,81,87,88, 90,91,92,93,95,96, 101,104,106,109, 127,129,134,136, 137	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135, 141, 143,144	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,116,122, 125,128,142,145

South Africa Zululand (141)	South Africa Zululand (142)	South Africa Zululand (143)	South Africa Zululand (144)	Tunisia Bizerte (145)
4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 143,144	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,116,122, 125,128,140,145	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141,144	4,5,6,8,13,14,15, 18,19,20,23,24,30, 49,50,56,61,63,64, 69,70,79,82,83,99, 111,117,118,119, 132,135,139, 141, 143	7,17,28,35,37,38, 39,40,41,48,54,55, 58,59,60,62,66,67, 71,74,89,102,112, 113,115,116,122, 125,128,140,142

Table 3. Cognitive map for female isolates

Algeria Blida (1)	Benin Cotonou (2)	Benin Cotonou (3)	Benin Cotonou (4)	Benin Cotonou (5)	DRC Haut Katanga (6)	DRC Kinshasha (7)	DRC Kinshasha (8)	DRC Kinshasha (9)	DRC Kinshasha (10)
8,22,48,50,51,55, 61,81,86,121,124, 128,131,137	3,5,23,24,25,26, 27,28,33,34,39, 41,52,53,145	2,5,23,24,25,26, 27,28,33,34,39, 41,52,53,145	10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	2,3,23,24,25,26, 27,28,33,34,39, 41,52,53,145	9,11,12,13,14, 16,17,59,60,62,92, 94,96,97,98,130	35,37,38,45, 54,56,64,73,77, 101,102,105,107, 116,117,129,139	1,22,48,50,51,55, 61,81,86,121,124, 128,131,137	6,11,12,13,14, 16,17,59,60,62,92, 94,96,97,98,130	4,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143

DRC Kongo Central (11)	DRC Kongo Central (12)	DRC Kongo Central (13)	DRC Kongo Central (14)	DRC Sud Kivu (15)	DRC Sud Kivu (16)	DRC Sud Kivu (17)	DRC Sud Kivu (18)	Egypt Cairo (19)	Egypt Cairo (20)

6,9,12,13,14, 16,17,59,60,62,92, 94,96,97,98,130	6,9,11,13,14, 16,17,59,60,62,92, 94,96,97,98,130	6,9,11,12,14, 16,17,59,60,62,92, 94,96, 97,98,130	6,9,11,12,13, 16,17,59,60,62,92, 94,96,97,98,130	4,10,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	6,9,11,12,13,14, 17,59,60,62,92, 94,96,97,98,130	6,9,11,12,13,14, 16,59,60,62,92, 94,96,97,98,130	4,10,15,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	4,10,15,18,20,21, 31,40,46,57,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	4,10,15,18,19,21, 31,40,46,57,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143
Egypt Cairo (21)	Egypt Cairo (22)	Gambia Kombo (23)	Gambia Kombo (24)	Gambia Kombo (25)	Gambia West Coast Reg (26)	Ghana Greater Ghana (27)	Ghana Greater Ghana (28)	Ghana Greater Ghana (29)	Ghana Greater Ghana (30)
4,10,15,18,19,20, 31,40,46,57,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	1,8,48,50,51,55, 61,81,86,121,124, 128,131,137	2,3,5,24,25,26, 27,28,33,34,39, 41,52,53,145	2,3,5,23,25,26, 27,28,33,34,39, 41,52,53,145	2,3,5,23,24,26, 27,28,33,34,39, 41,52,53,145	2,3,5,23,24,25, 27,28,33,34,39, 41,52,53,145	2,3,5,23,24,25,26, 28,33,34,39, 41,52,53,145	2,3,5,23,24,25,26, 27,33,34,39, 41,52,53,145	30,32,42,108,113	29,32,42,108,113
Kenya Unknown (31)	Kenya Unknown (32)	Kenya Unknown (33)	Kenya Unknown (34)	Madagascar Antananarivo (35)	Madagascar Fenoarivo (36)	Mali Bamako (37)	Mali Bamako (38)	Mali Bamako (39)	Mali Bamako (40)
4,10,15,18,19,20, 21,40,46,57,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	29,30,42,108,113	2,3,5,23,24,25,26, 27,28,34,39, 41,52,53,145	2,3,5,23,24,25,26, 27,28,33,39, 41,52,53,145	7,37,38,45, 54,56,64,73,77, 101,102,105,107, 116,117,129,139	69,75,76,122 135,138,142	7,35,38,45, 54,56,64,73,77, 101,102,105,107, 116,117,129,139	7,35,37,45, 54,56,64,73,77, 101,102,105,107, 116,117,129,139	2,3,5,23,24,25,26, 27,28,33,34, 41,52,53,145	4,10,15,18,19,20, 21,31,46,57,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143
Morocco Rabat (41)	Morocco Rabat (42)	Nigeria Ekiti (43)	Nigeria Ekiti (44)	Nigeria Ekiti (45)	Nigeria Ogun (46)	Nigeria Ondo (47)	Nigeria Osun (48)	Nigeria Osun (49)	Nigeria Osun (50)

2,3,5,23,24,25,26, 27,28,33,34,39, 52,53,145	29,30,32,108,113	44,47,49,74, 80,99,114,126	43,47,49,74, 80,99,114,126	7,35,37,38, 54,56,64,73,77, 101,102,105,107, 116,117,129,139	4,10,15,18,19,20, 21,31,40,57,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	43,44,49,74, 80,99,114,126	1,8,22,50,51,55, 61,81,86,121,124, 128,131,137	43,44,47,74, 80,99,114,126	1,8,22,48,51,55, 61,81,86,121,124, 128,131,137
Nigeria Osun (51)	Nigeria Unknown (52)	Nigeria Unknown (53)	Senegal Dakar (54)	Senegal Dakar (55)	Senegal Dakar (56)	Senegal Dakar (57)	Senegal Diourbel (58)	Senegal Diourbel (59)	Senegal Diourbel (60)
1,8,22,48,50,55, 61,81,86,121,124, 128,131,137	2,3,5,23,24,25,26, 27,28,33,34,39, 41,53,145	2,3,5,23,24,25,26, 27,28,33,34,39, 41,52,145	7,35,37,38,45, 56,64,73,77, 101,102,105,107, 116,117,129,139	1,8,22,48,50,51, 61,81,86,121,124, 128,131,137	7,35,37,38,45, 54,64,73,77, 101,102,105,107, 116,117,129,139	4,10,15,18,19,20, 21,31,40,46,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	4,10,15,18,19,20, 21,31,40,46,57,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	6,9,11,12,13,14, 16,17,60,62,92, 94,96, 97,98,130	6,9,11,12,13,14, 16,17,59,62,92, 94,96, 97,98,130
Senegal Diourbel (61)	Senegal Thies (62)	Senegal Thies (63)	Senegal Thies (64)	South Africa Amajuba (65)	South Africa Amajuba (66)	South Africa Amajuba (67)	South Africa Amajuba (68)	South Africa Berea (69)	South Africa Berea (70)
1,8,22,48,50,51,55, 81,86,121,124, 128,131,137	6,9,11,12,13,14, 16,17,59,60,92, 94,96, 97,98,130	4,10,15,18,19,20, 21,31,40,46,57,58, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	7,35,37,38,45, 54,56,73,77, 101,102,105,107, 116,117,129,139	123,134,141,144	4,10,15,18,19,20, 21,31,40,46,57,58, 63,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	4,10,15,18,19,20,21, 31,40,46,57,58, 63, 66,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	36,75,76,122 135,138,142	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143
South Africa Berea (71)	South Africa Berea (72)	South Africa Cape Town (73)	South Africa Cape Town (74)	South Africa Cape Town (75)	South Africa Cape Town (76)	South Africa Eastern Cape (77)	South Africa eThekwini (78)	South Africa eThekwini (79)	South Africa eThekwini (80)

4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	7,35,37,38,45, 54,56,64,77, 101,102,105,107, 116,117,129,139	43,44,47,49, 80,99,114,126	36,69,76, 135,138,142	36,69,75,122 135,138,142	7,35,37,38,45, 54,56,64,73, 101,102,105,107, 116,117,129,139	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71,72, 79,82,93,100, 103,104,112,127, 133,136,140,143	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71,72, 78,82,93,100, 103,104,112,127, 133,136,140,143	43,44,47,49,74, 99,114,126
South Africa eThekwini (81)	South Africa Free State (82)	South Africa Free State (83)	South Africa Free State (84)	South Africa Free State (85)	South Africa Harry Gwala (86)	South Africa Harry Gwala (87)	South Africa Harry Gwala (88)	South Africa Harry Gwala (89)	South Africa Ilembe (90)
1,8,22,48,50,51,55, 61,86,121,124, 128,131,137	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71,72, 78,79,93,100, 103,104,112,127, 133,136,140,143	84,89,90,91,106,109	83,89,90,91,92, 106,109	88,110,111,115,119, 125	1,8,22,48,50,51,55, 61,81,121,124, 128,131,137	95,118,120,132	85,110,115,119, 125	83,84,90,91,106,109	83,84,89,91,92, 106,109
South Africa Ilembe (91)	South Africa Ilembe (92)	South Africa Ilembe (93)	South Africa Kg Cetshwayo (94)	South Africa Kg Cetshwayo (95)	South Africa Kg Cetshwayo (96)	South Africa Kg Cetshwayo (97)	South Africa KZN (98)	South Africa KZN (99)	South Africa KZN (100)
83,84,89,90,106,109	6,9,11,12,13,14, 16,17,59,60,62, 94,96,97,98,130	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71,72, 78,79,82,100, 103,104,112,127, 133,136,140,143	6,9,11,12,13,14, 16,17,59,60,62,92, 96,97,98,130	87,118,120,132	6,9,11,12,13,14, 16,17,59,60,62,92, 94,97,98,130	6,9,11,12,13,14, 16,17,59,60,62,92, 94,96,98,130	6,9,11,12,13,14, 16,17,59,60,62,92, 94,96,97,130	43,44,47,49,74, 80,114,126	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71, 72,78,79,82,93 103,104,112,127, 133,136,140,143
South Africa KZN (101)	South Africa LP (102)	South Africa Mmorningside (103)	South Africa Mmorningside (104)	South Africa North West (105)	South Africa North West (106)	South Africa North West (107)	South Africa North West (108)	South Africa Overport (109)	South Africa Sisonke (110)

7,35,37,38,45, 54,56,64,73,77, 102,105,107, 116,117,129,139	7,35,37,38,45, 54,56,64,73,77, 101,105,107, 116,117,129,139	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71, 72,78,79,82,93, 100,104,112,127, 133,136,140,143	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71, 72,78,79,82,93, 100,103,112,127, 133,136,140,143	7,35,37,38,45, 54,56,64,73,77, 101,102,107, 116,117,129,139	83,84,89,90,91, 106,109	7,35,37,38,45, 54,56,64,73,77, 101,102,105, 116,117,129,139	29,30,32,42,113	83,84,89,90,91,92, 106	85,88,111,115,119, 125
South Africa Sisonke (111)	South Africa Stanger (112)	South Africa Ugu (113)	South Africa Ugu (114)	South Africa Ugu (115)	South Africa Ugu (116)	South Africa Umbilo (117)	South Africa Umbilo (118)	South Africa Umbilo (119)	South Africa Umbilo (120)
85,88,110,115,119, 125	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71, 72,78,79,82,93, 100,103,104,127, 133,136,140,143	29,30,32,42,108	43,44,47,49,74, 80,99,126	85,88,110,119, 125	7,35,37,38,45, 54,56,64,73,77, 101,102,105,107, 117,129,139	7,35,37,38,45, 54,56,64,73,77, 101,102,105,107, 116,129,139	87,95,120,132	85,88,110,115,125	87,95,118,132
South Africa Umgungundl (121)	South Africa Umgungundl (122)	South Africa Umgungundl (123)	South Africa Umgungundl (124)	South Africa Umkhanyak (125)	South Africa Umkhanyak (126)	South Africa Umkhanyak (127)	South Africa Umkhanyak (128)	South Africa Umzinyathi (129)	South Africa Umzinyathi (130)
1,8,22,48,50,51,55, 61,81,86,124, 128,131,137	36,69,75,76, 135,138,142	65,134,141,144	1,8,22,48,50,51,55, 61,81,86,121, 128,131,137	85,88,110,115,119	43,44,47,49,74, 80,99,114	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140,143	1,8,22,48,50,51,55, 61,81,86,121,124, 131,137	7,35,37,38,45, 54,56,64,73,77, 101,102,105,107, 116,117,139	6,9,11,12,13,14, 16,17,59,60,62,92, 94,96, 97,98
South Africa Umzinyathi (131)	South Africa Umzinyathi (132)	South Africa Uthukela (133)	South Africa Uthukela (134)	South Africa Uthukela (135)	South Africa Uthukela (136)	South Africa Uthungulu (137)	South Africa Uthungulu (138)	South Africa Uthungulu (139)	South Africa Uthungulu (140)

1,8,22,48,50,51,55, 61,81,86,121,124, 128,137	87,95,118,120	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 136,143	65,123,141,144	36,69,75,76, 122,138,142	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,143	1,8,22,48,50,51,55, 61,81,86,121,124, 128,131	36,69,75,76, 122,135,142	7,35,37,38,45, 54,56,64,73,77, 101,102,105,107, 116,117,129	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,143

South Africa Zululand (141)	South Africa Zululand (142)	South Africa Zululand (143)	South Africa Zululand (144)	Tunisia Ben Arous (145)
65,123,134,144	36,69,75,76, 122,135,138	4,10,15,18,19,20,21, 31,40,46,57,58,63, 66,67,68,70,71,72, 78,79,82,93,100, 103,104,112,127, 133,136,140	65,123,134,141	2,3,5,23,24,25,26, 27,28,33,34,39, 41,52,53

Figures

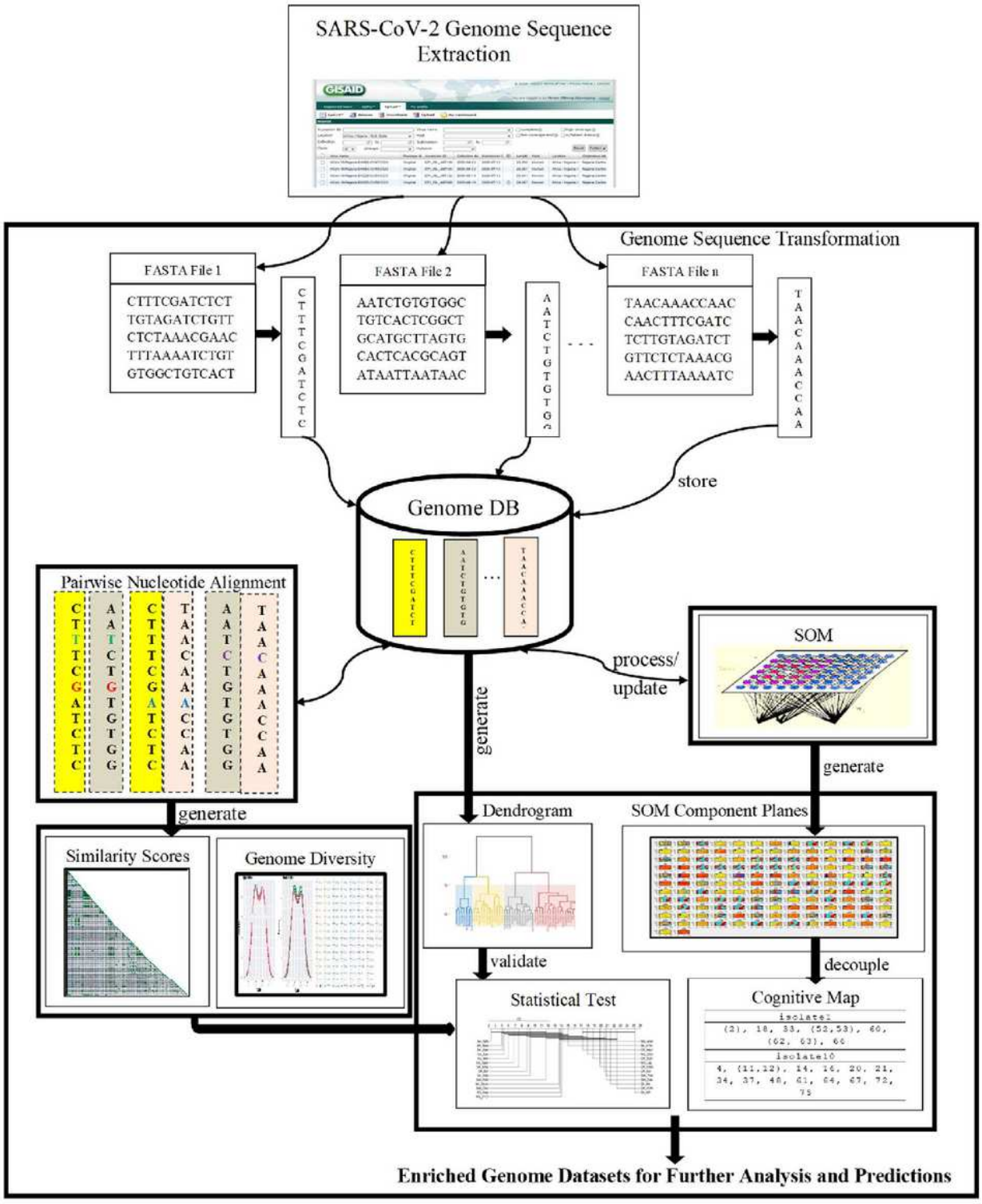


Figure 1

Workflow describing the proposed system framework. The workflow begins with the excavation of FASTA files of human SARS-CoV-2 genome sequences from GISAID. These files are stripped and processed into a genome database (DB) as multiple columns of nucleotide sequence. A series of AI/ML techniques are

applied to extract knowledge from the genome datasets as follows: Compute dis(similarities) scores between the various pairs of genome sequences and obtain a genomic tree of highly dis(similar) isolates grouped in the form of a dendrogram/phylogenomic tree. Determine the optimal number of natural clusters–to provide additional knowledge. Separate the viral sub-strains using self-organizing map (SOM) component planes–for transmission pathways visualization. Perform direct pairwise nucleotide alignment of the entire genome sequences–to yield a nucleotide similarity matrix. Generate cognitive map–for intelligent sub-strains contact tracing and prediction.

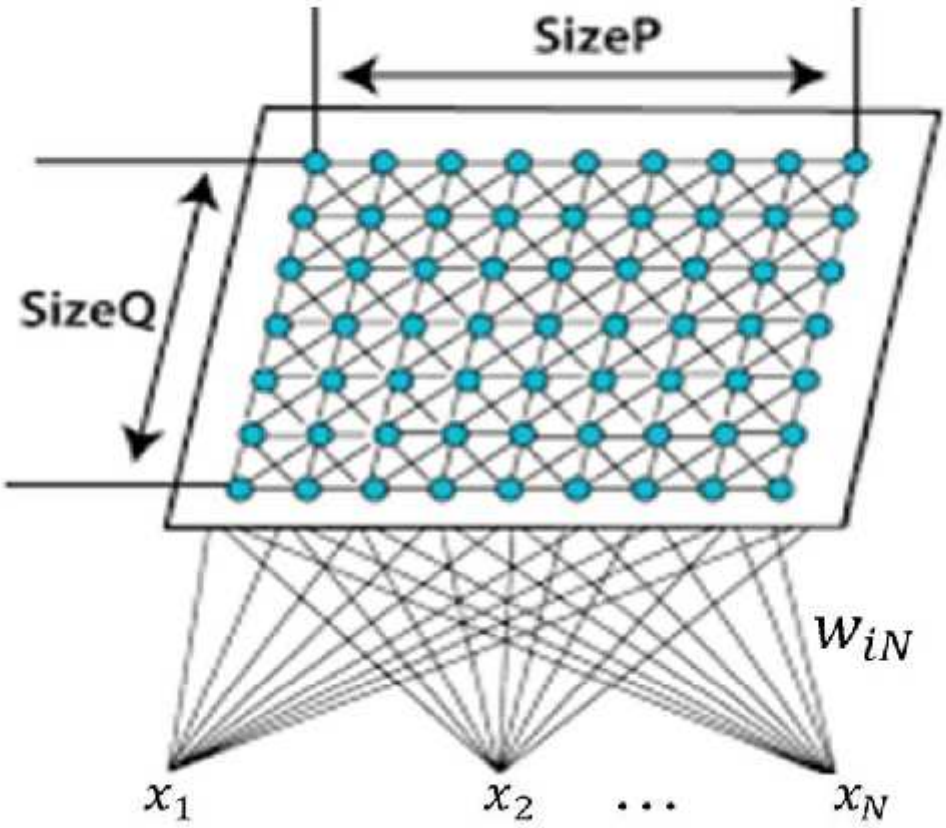


Figure 2

SOM showing the map topology and interactions between nodes. Each neuron is assigned a vector of weights (w=w_i1,w_i2,..w_iN) with dimension similar to the input vector i (i=1,2,…,L); where L is the total number of neurons in the network. The input nodes have p features, and the output nodes, q prototypes, with each prototype connected to all features. The weight vector of the connections consumes the prototype of each neuron and has same dimension as the input vector. SOMs differ from other artificial neural networks as they apply competitive learning, against error correction learning such as backpropagation, and the fact that they preserve the topological properties of the input space using a neighborhood function.

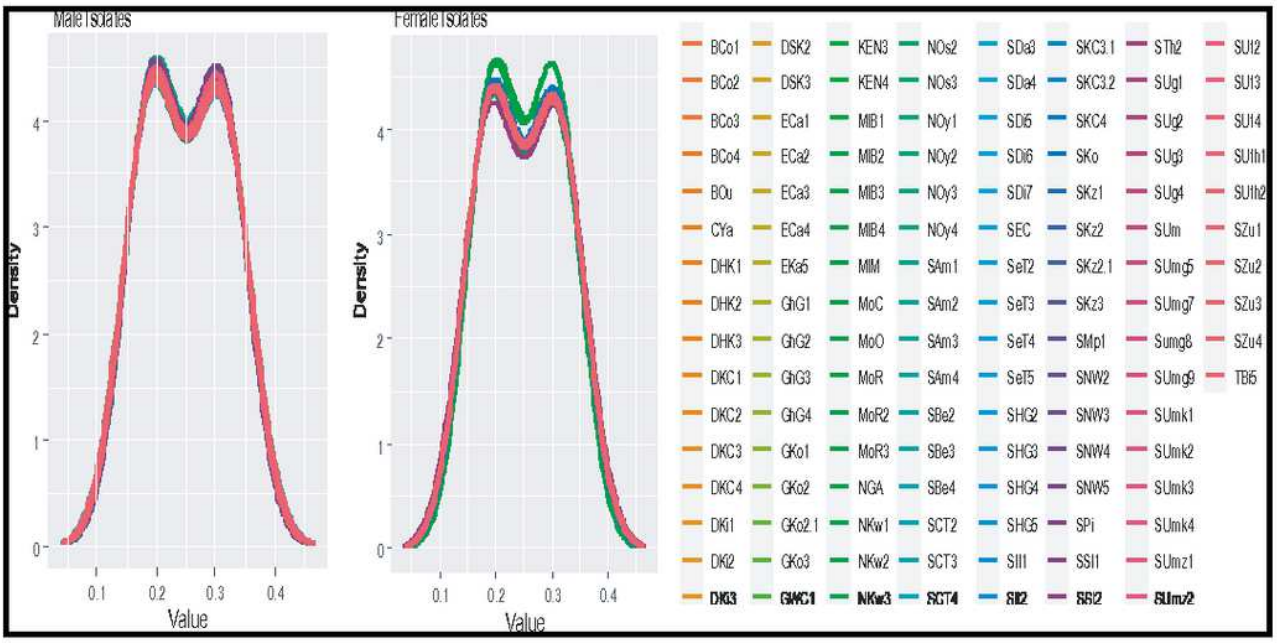
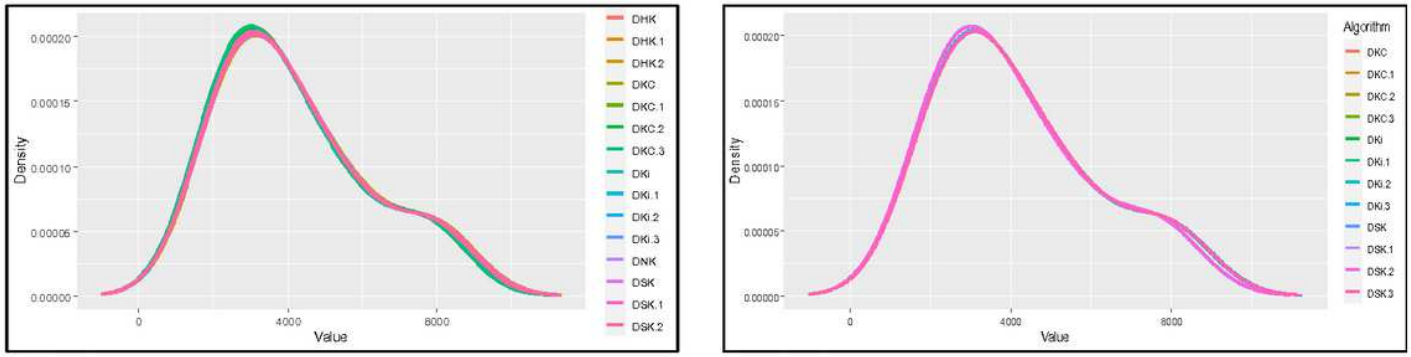


Figure 3

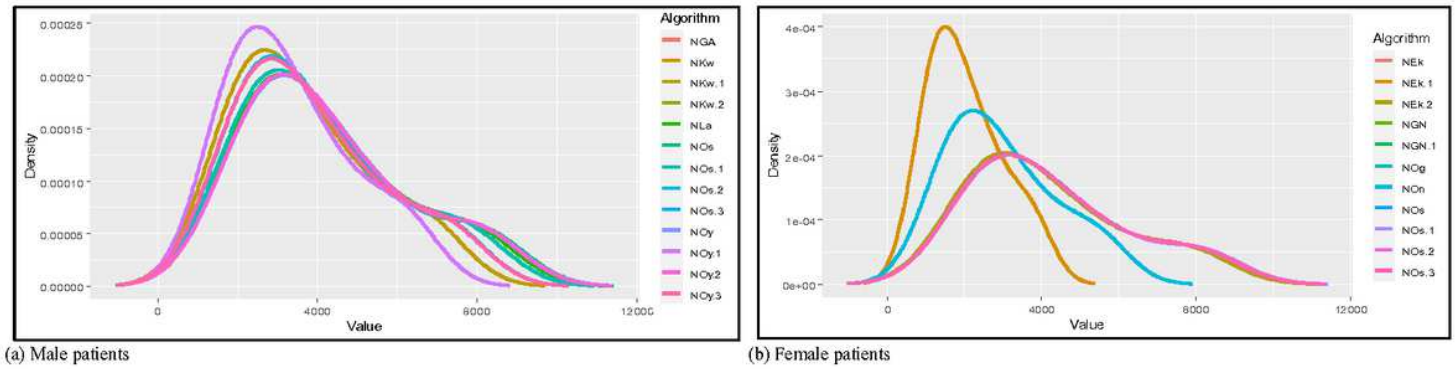
Density plots of entire genome datasets. A Density plot visualizes the distribution of data over a continuous scale. It is a variation of the histogram that applies kernel smoothing to plot values, enabling smoother distributions by smoothening out the noise. The peaks of the density plot help display where values are concentrated over the interval.



(a) Male patients (b) Female patients

Figure 4

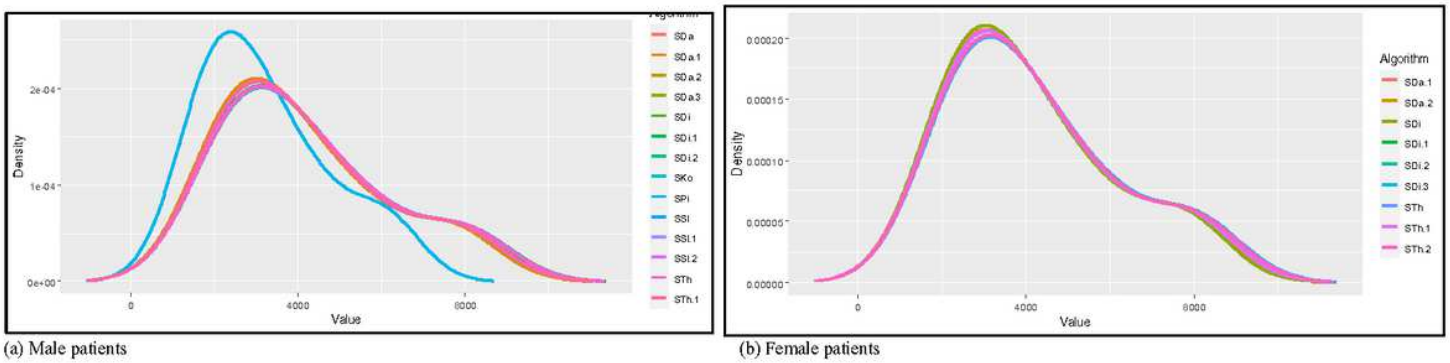
Density plots of DRC's isolates. Both plots exhibit similar curve pattern with near aligned isolates, indicating identical genome expression between the isolates.



(a) Male patients (b) Female patients

Figure 5

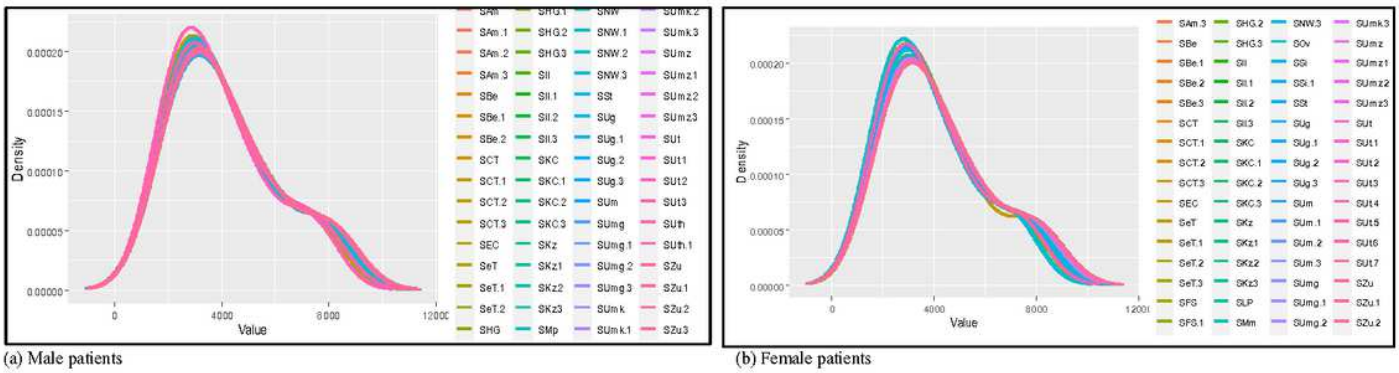
Density plots of Nigerian isolates. Although both plots maintain same curve patterns, the male isolates exhibit diversity in most of the genomes. Two female isolates (Fig, 5b) present distincnt variations from the rest of the isolates, indicating significant genome diversity beween the male and female isolates.



(a) Male patients (b) Female patients

Figure 6

Density plots of Senegalese isolates. Aside the single isolate that fails to align with other male isolates (Fig. 6a), both plots maintain near-aligned isolates.



(a) Male patients (b) Female patients

Figure 7

Density plots of South African isolates. Both plots maintain same patterns with shades of similar isolates dominating the plot area.

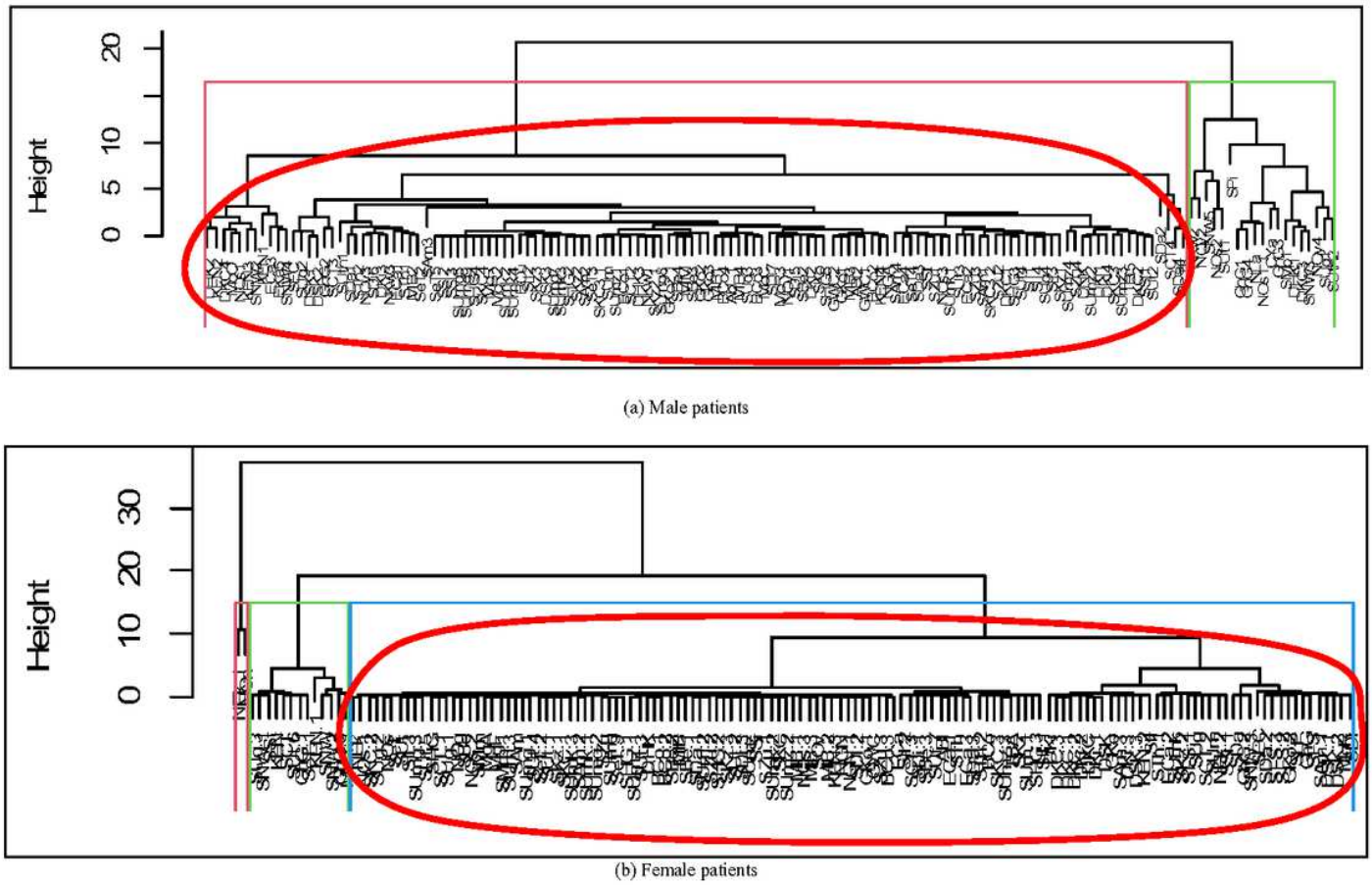


Figure 8

Phylogenomic trees for African male and female isolates. For full names of country codes, see Additional file 1: SupplData1_1.xlsx.

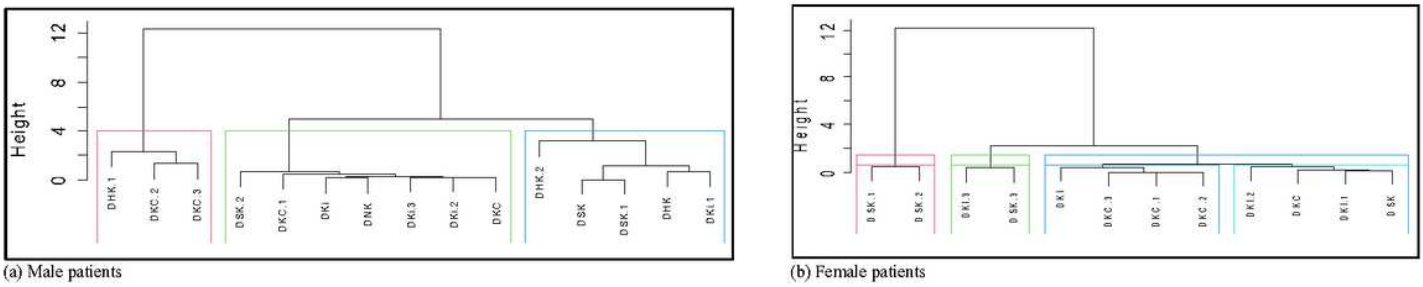
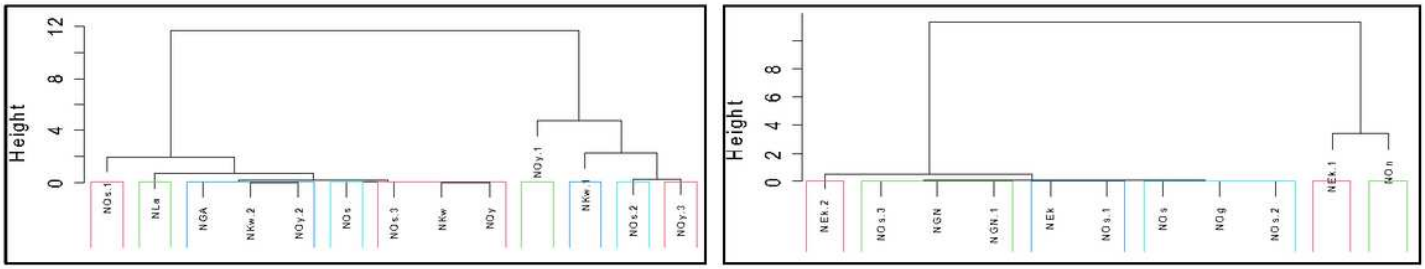


Figure 9

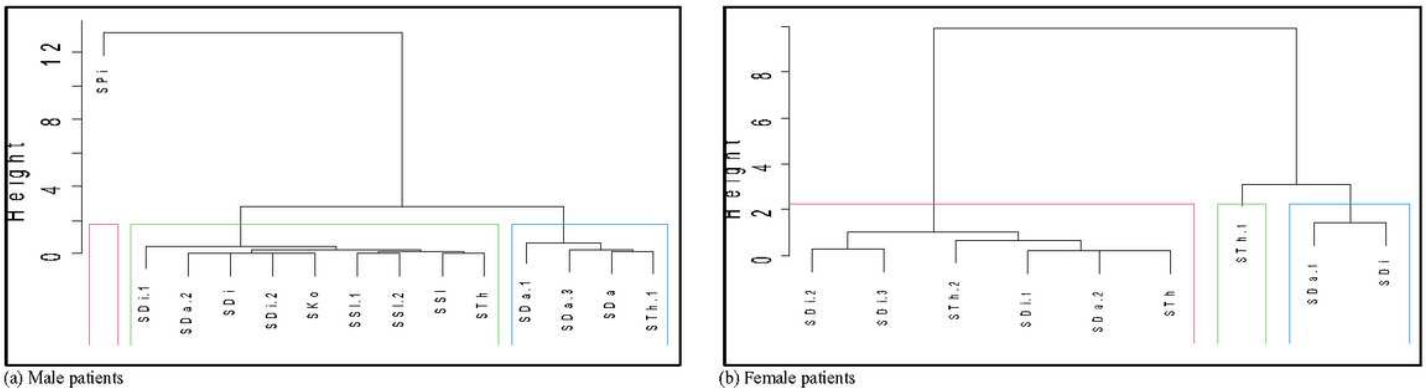
Phylogenomic trees for DRC's male and female isolates. For full names of country codes, see Additional file 1: SupplData1_1.xlsx.



(a) Male patients (b) Female patients

Figure 10

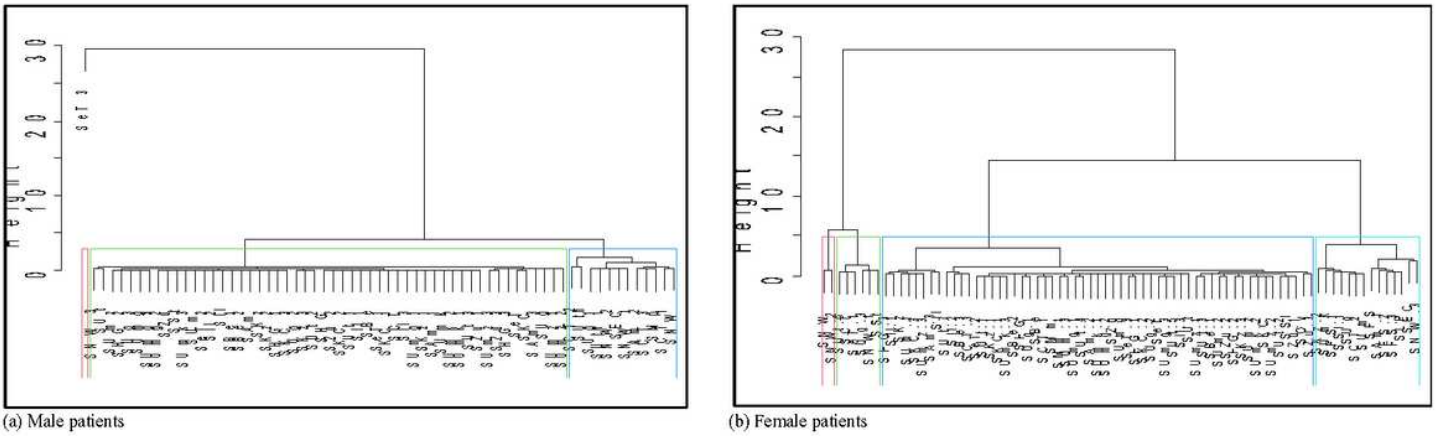
Phylogenomic trees for Nigerian male and female isolates. For full names of country codes, see Additional file 1: SupplData1_1.xlsx.



(a) Male patients (b) Female patients

Figure 11

Phylogenomic trees for Senegalese male and female isolates. For full names of country codes, see Additional file 1: SupplData1_1.xlsx.



(a) Male patients (b) Female patients

Figure 12

Phylogenomic trees for South African male and female isolates. For full names of country codes, see Additional file 1: SupplData1_1.xlsx.

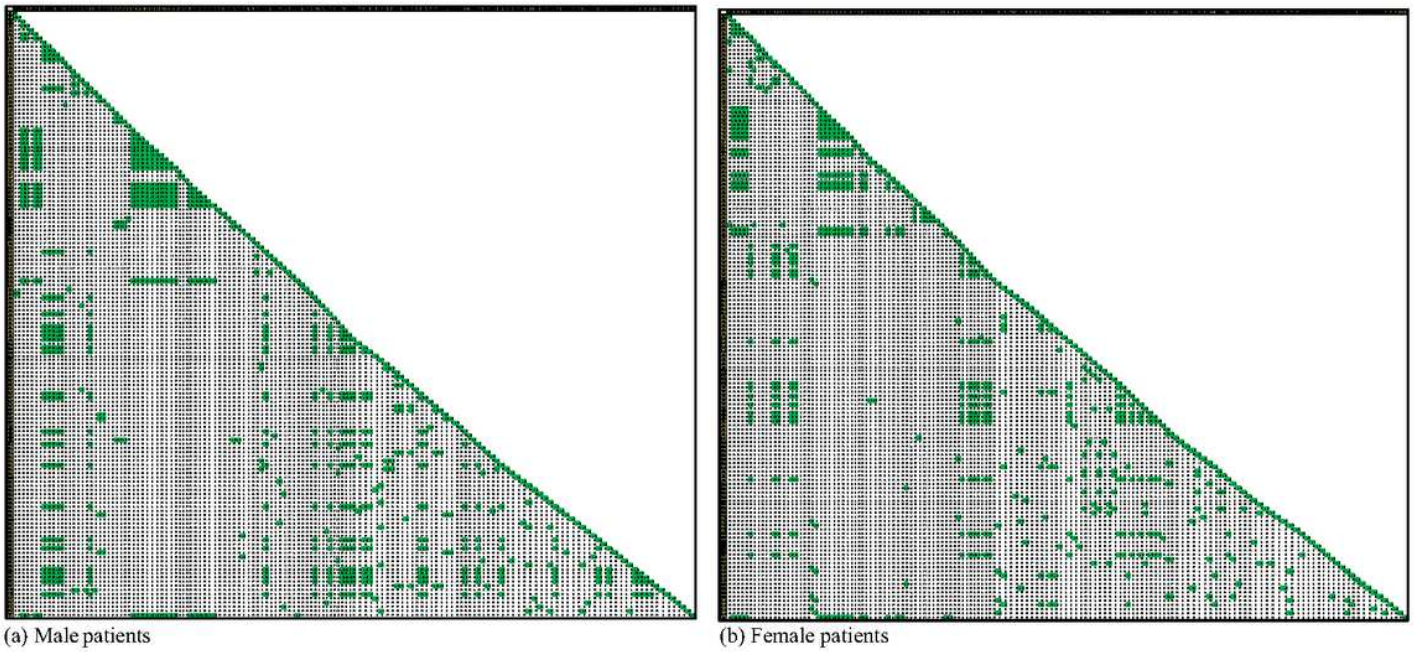


Figure 13

Nucleotide similarity matrices. Green colored cells are regions of high similarity that may indicate functional, structural and/or evolutionary relationships between nucleotide sequences.

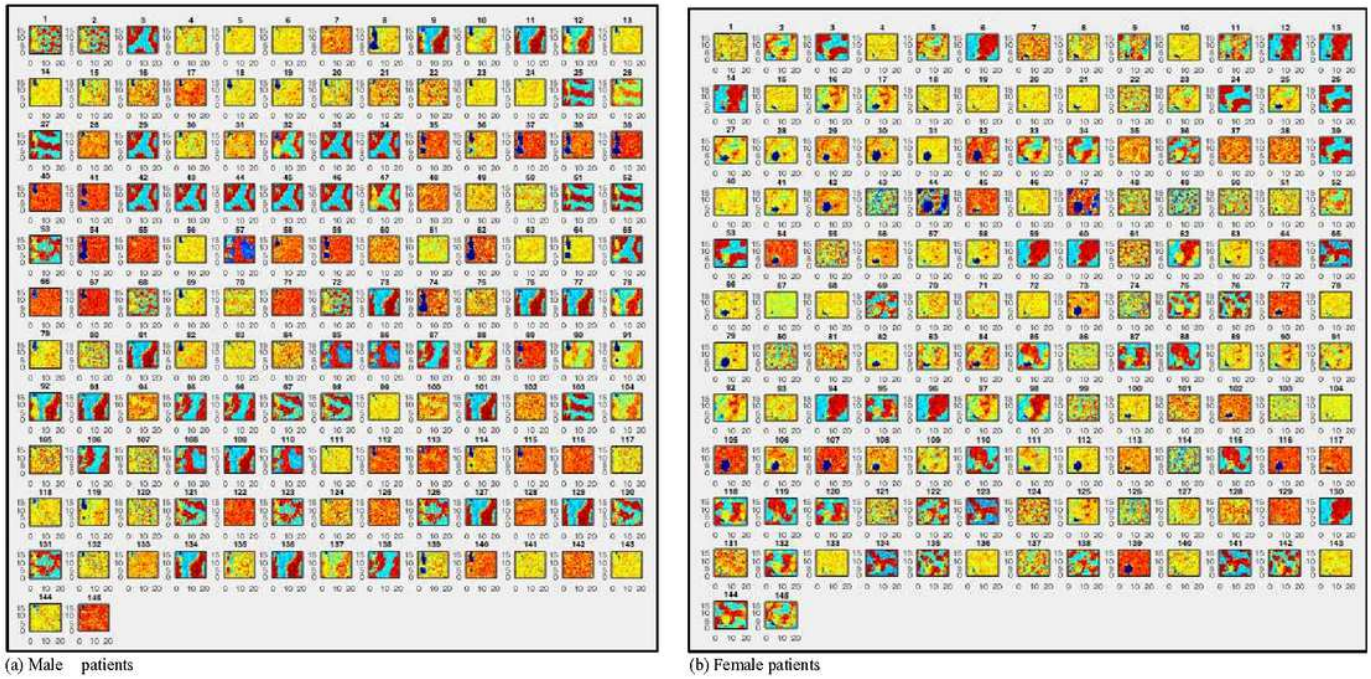


Figure 14

SOM component planes visualization. Maps are ordered by countries, with at least 1 isolate per country. The isolate numbers (1-145) represent the various states of the country excavated from GSAID (see SupplData1_1.xlsx). The isolates are distributed by countries as follows. Male: Algeria (1-2), Benin (3-7), Cameroon (8), DRC (9-23), Egypt (24-28), Gambia (29-34), Ghana (35-38), Kenya (39-42), Mali (43-47),

Morocco (48-52), Nigeria (53-65), Senegal (66-79), South Africa (80-144), Tunisia (145). Female: Algeria (1), Benin (2-5), DRC (6-18), Egypt (19-22), Gambia (23-26), Ghana (27-30), Kenya (31-34), Madagascar (35-36), Mali (37-40), Morocco (41-42), Nigeria (43-53), Senegal (54-64), South Africa (65-144), Tunisia (145).

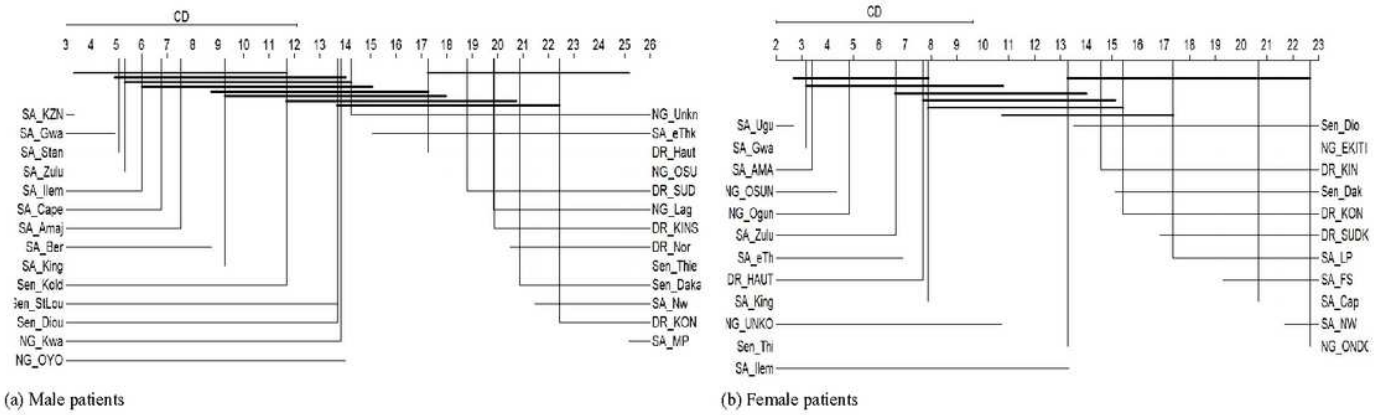


Figure 15

Inter-country CD plots for male and female patients. For a significance level α the Nemenyi's test determines the critical difference (CD), if the difference between the average ranking of two treatments is greater than CD, then the null hypothesis that the treatments have the same performance is rejected.

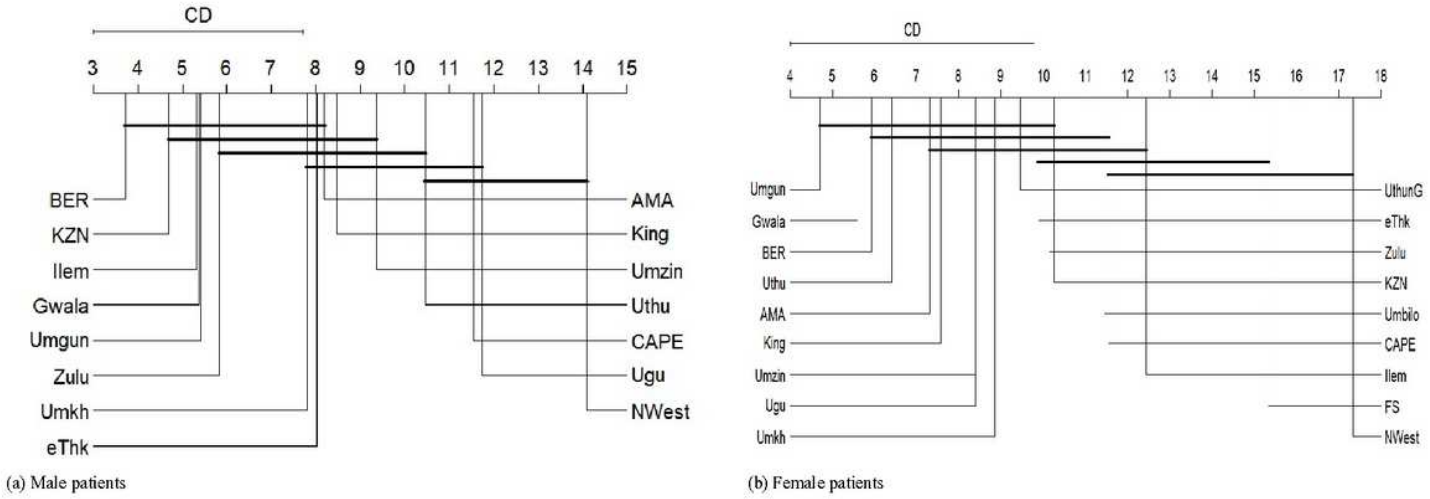


Figure 16

Intra-country CD plots for male and female patients. For a significance level α the Nemenyi's test determines the critical difference (CD), if the difference between the average ranking of two treatments is greater than CD, then the null hypothesis that the treatments have the same performance is rejected.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplData11.xlsx
- SupplData21.xlsx