A Hybrid Computational Framework for Intelligent Inter-continent SARS-CoV-2 Sub-strains Characterization and Prediction

Moses Ekpenyong (✉ mosesekpenyong@uniuyo.edu.ng)
University of Uyo https://orcid.org/0000-0001-6774-5259
Mercy Edoho
University of Uyo
Udoinyang Inyang
University of Uyo https://orcid.org/0000-0001-5086-303X
Faith-Michael Uzoka
Mount Royal University https://orcid.org/0000-0002-2414-5349
Itemobong Ekaidem
University of Uyo https://orcid.org/0000-0003-2312-9969
Anietie Moses
University of uyo https://orcid.org/0000-0001-6881-5466
Martins Emeje
National Institute for Pharmaceutical Research and Development (NIPRD), https://orcid.org/0000-0002-0202-5426
Youtchou Tatfeng
Niger Delta University https://orcid.org/0000-0003-2149-7135
Ifiok Udo
University of uyo
Enoabasi Anwana
University of Uyo https://orcid.org/0000-0001-7711-8291
Oboso Etim
University of Uyo
Geoffery Joseph
University of Uyo
Emmanuel Dan
University of Uyo

Research Article

Keywords: COVID-19, cognitive map, deep learning, hybrid framework, hierarchical clustering, SARS-CoV-2, self-organizing map

Posted Date: November 6th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-88429/v2

Version of Record: A version of this preprint was published at Scientific Reports on July 15th, 2021. See the published version at https://doi.org/10.1038/s41598-021-93757-w.

A Hybrid Computational Framework for Intelligent Inter-continent SARS-CoV-2 Sub-strains Characterization and Prediction

Moses Ekpenyong[1*], Mercy Edoho[1], Udoinyang Inyang[1], Faith-Michael Uzoka[2], Itemobong Ekaidem[3], Anietie Moses[3], Martins Emeje[4], Youtchou Tatfeng[5], Ifiok Udo[1], Enoabasi Anwana[6], Oboso Etim[7], Geoffery Joseph[1], Emmanuel Dan[1].

[1]Department of Computer Science, University of Uyo, Nigeria
[2]Department of Mathematics and Computing, Mount Royal University, Canada
[3]College of Health Sciences, University of Uyo, Nigeria
[4]National Institute for Pharmaceutical Research and Development (NIPRD), Nigeria
[5]College of Health Sciences, Niger Delta University, Nigeria
[6]Department of Botany and Ecological Studies, University of Uyo, Nigeria
[7]Department of Biochemistry, University of Uyo, Nigeria

*Correspondence to: mosesekpenyong@uniuyo.edu.ng.

Abstract
Whereas accelerated attention beclouded early stages of the coronavirus spread, knowledge of actual pathogenicity and origin of possible sub-strains remained unclear. By harvesting the Global initiative on Sharing All Influenza Data (GISAID) database (https://www.gisaid.org/), between December 2019 and August 20, 2020, a total of 157 human SARS-CoV-2 (complete) genome sequences processed by gender, across 6 continents of the world, were analyzed. We hypothesized that data speaks for itself and can discern true and explainable patterns of the disease. Identical genome diversity and pattern correlates analysis performed using a hybrid of biotechnology and machine learning methods corroborate multiple emergence of SARS-CoV-2 sub-strains and explained the diversity of the SARS-CoV-2. Interestingly, some viral sub-strains progressively transformed into new sub-strain clusters indicating varying amino acid and strong nucleotide association derived from same origin. A novel approach to cognitive knowledge mining from enriched genome datasets and output targets labeling, helped intelligent prediction of emerging or new viral sub-strains.

Introduction
The coronavirus disease pandemic had forced complete shutdown on all economies of the world[1,2]. Since then, its breadth and depth have grown exponentially, causing disruptions that require hybrid of computational approaches–to discover the changing nature of the virus as it transmits from country to country. While there exist claims that the virus remained unchanged, a growing number of studies have reported the emergence of several sub-strains[3,4]. The rapid human to human transmission of the pathogenic SARS-CoV-2 to most parts of the world has exhibited differences in disease severity and fatality even within a demographic region of a country. This disparity has been attributed to factors such as gender, age, ethnicity, race and co-morbidities. However, the dissimilarity in genome sequencing of early viral samples obtained from infected individuals from Europe, North America, Asian and Oceanian regions disgorged various studies aimed at analyzing and understanding the evolutionary history and relationships among the different SARS-CoV-2 strains.

The SARS-CoV-2 is a β-coronavirus–an enveloped non-segmented positive-sense RNA virus (subgenus–sarbecovirus, subfamily–Orthocoronavirinae)[5], which proliferation begun in December 2019 in Wuhan China. It has since been confirmed that two strains of the new coronavirus (the L- and S-strains) are spreading around the world today[6], and the fact that the L-type is more prevalent suggests that it is "more aggressive" than the S-type. Greater proportion of research progress on SARS-CoV-2 has taken the biotechnology dimension[7,8], specifically focusing on species characterization and variants analysis through features extraction. However, Artificial Intelligence (AI) and Machine Learning (ML) methods are expanding biotechnology capacity into the bioinformatics realm, through intelligent genome probing for precise classification of the virus. In general, AI/ML research on SARS-CoV-2 has permeated four key areas namely: screening and treatment[9,10,11,12], contact tracing[13], prediction and forecasting[14,15], and drugs and vaccine discovery[16,17,18].

Mining additional knowledge from clinical/experimental data–to support intelligent discoveries and verification could assist complete features extraction, missing information recovery, hidden patterns understanding, and permit output targets labeling for intelligent predictions. Most biotechnology/bioinformatics tools are 'black boxes' and not open to contributions from the research community including reproducible research. Furthermore, extracted features are incomplete to aid meaningful knowledge integration. This paper therefore provides an open source hybrid framework with rapid prototype modules for intelligent SARS-CoV-2 sub-strains prediction. Our framework produces intermediate data/results to aid reproducible research and permits scalability of the design across diverse domains.

SARS-CoV-2 Sub-strains Analysis

Phylogenetic tree and genomic tree (hierarchical clustering) are common determinant measures for representing genetic diversity and evolutionary relationships. While phylogenetic tree reflects slow evolution within the genome (point mutations), hierarchical clustering describes major genetic re-arrangement events (insertions or deletions). However, converting massive amount of data such as complete genome sequences into meaningful biological representations becomes difficult, to aid proper interpretation. Numerous algorithms/tools have evolved to target specific gene sites/locations for "on-the-fly" or online representations such as the phylogenetic tree. But major drawbacks of site-specific analysis include incomplete representation and clustering errors–as different genome sites undergo different evolutionary changes, resulting in disparate multi-dimensional patterns at different sites. Attempts at estimating phylogenies by comparing entire genomes have been made by focusing mainly on gene content and gene order comparisons. While early attempts concentrated on morphological characters with the premise that direct genes comparison makes more sense. However, modern attempts use sequences from homologous genes[19] but are burdened by the fact that a gene's evolutionary history may differ from the evolutionary history of the organism, as some genes sufficiently conserved across the species of interest may escape detection. Furthermore, most of their proves are still at the modelling stage and not yet verified using clinical and experimental data. Li et. al[20] for instance investigated the angiotensin-converting enzyme 2 (ACE2)–the receptor agent for the SARS-CoV-2 virus–a known contributor to viral infections susceptibility and/or resistance[19]. ACE2 generates small proteins by cutting up larger protein angiotensinogen, in turn affecting the nucleotide/protein. They compared ACE2 expression levels across 31 normal human tissues between males and females and between younger and older persons using two-sided student's t-test. By examining the expression patterns, they found that ACE2 was similarly expressed between males and females or between younger and older persons in experimented tissues. Furthermore, men showed worse prognosis than women. Their findings however lacked experimental and clinical data validation.

In Tang et al.[6], a population genetic analysis of 103 SARS-CoV-2 genomes was performed. Their analysis revealed two dominant types of SARS-CoV-2 namely the L type (~ 70%) and S type (~ 30%). In another study, Stefanelli et al.[21] investigated the phylogenesis of 2 patients in Italy; a Chinese tourist from Wuhan and an Italian diagnosed, isolated and hospitalized in January and February 2020. They found the Italian patient's strain to be different from the tourist's strain, as it clustered with strains from Germany and Mexico, while the Chinese tourist's strain was grouped with strains from Europe and Australia. Similarly, Somasundaram et al.[22] systematically explored the phylogenetic and viral clade of 28 Indian isolates of SARS-CoV-2. A total of 449 complete genome samples from USA, Europe, China, East Asia, Oceania, Middle East (Kuwait and Saudi Arabia) and India were collected from Global initiative on Sharing All Influenza Data (GISAID: https://www.gisaid.org/). A phylogenetic analysis by maximum likelihood was achieved using IQ tree. Out of the Indian isolates, 26 samples were equally distributed into 2 clusters (A and B). Cluster A consisted of mostly Oceania/Kuwait and 13 Indian samples, while cluster B contained Europe and some of Middle East/South Asian samples together with another 13 Indian samples. The remaining 2 Indian isolates which neither belonged to cluster A nor cluster B, were present in the cluster with mostly China and East-Asia samples. However, the use of small datasets and the lack of travel history made their findings inconclusive

Hybrid tools that combine biotechnology and ML/AI methodologies, have advanced precision in approach and solution to the pandemic. Lopez-Rincon et al.[23] for instance, combined molecular testing with deep learning for automatic features extraction from SARS-CoV-2 genome sequences. The network's behavior on every sample was analyzed to discover sequences used by the model to classify SARS-CoV-2. Experiments on data from the novel coronavirus resource (2019nCoVR) showed that their approach could correctly classify SARS-CoV-2, and distinguish it from other coronavirus strains, regardless of missing information and errors in sequencing (noise). In Randhawa et al.[24], an intrinsic COVID-19 virus genomic signature was identified and combined with a ML-based alignment-free approach to yield robust classification of complete SARS-CoV-2 genomes. A supervised ML with digital signal processing and decision tree augmentation was used for genome analysis and successive refinements of taxonomic classification. Spearman's rank correlation coefficient analysis was then used for result validation. A large dataset of over 5000 unique viral genomic sequences were finally mined; and their results support the bat origin hypothesis. In Khanday[25], ML and ensemble learning models were used to classify clinical reports into four categories of SARS-related viruses. Feature extraction was achieved and extracted features finally supplied to the ML classifiers. They found that logistic regression and multinomial Naive Bayes yielded better results than other ML algorithms with high testing accuracy. In Melin[26], an analysis of spatial evolution of coronavirus pandemic around the world using self-organizing maps (SOMs) was performed. Data was obtained from the Humanitarian Data Exchange (HDX)[27], from countries where COVID-19 cases have occurred between January 22, 2020 and May 13, 2020. Spatially similar countries were grouped by cases, to analyze which countries adopt similar strategies in dealing with spread of the virus. Villmann et al.[28] investigated SARS-CoV-2 sequences based on alignment-free methods for RNA sequence comparison. Using phylogenetic trees and alignment-free dissimilarity measures plus learning vector quantization classifiers, discrimination and classification of viral types were performed. Learning the vector quantizers provided additional knowledge about the classification decisions. A classifier model was finally obtained after training the classifier with the GISAID datasets. The limitation of their study is its inability to label the SARS-CoV-2 datasets for prediction of new/emerging mutant strains and viral types.

Results

The general workflow describing the proposed computational framework is shown in Fig. 1, and the sequence of steps implementing the workflow is presented as Algorithm 1.

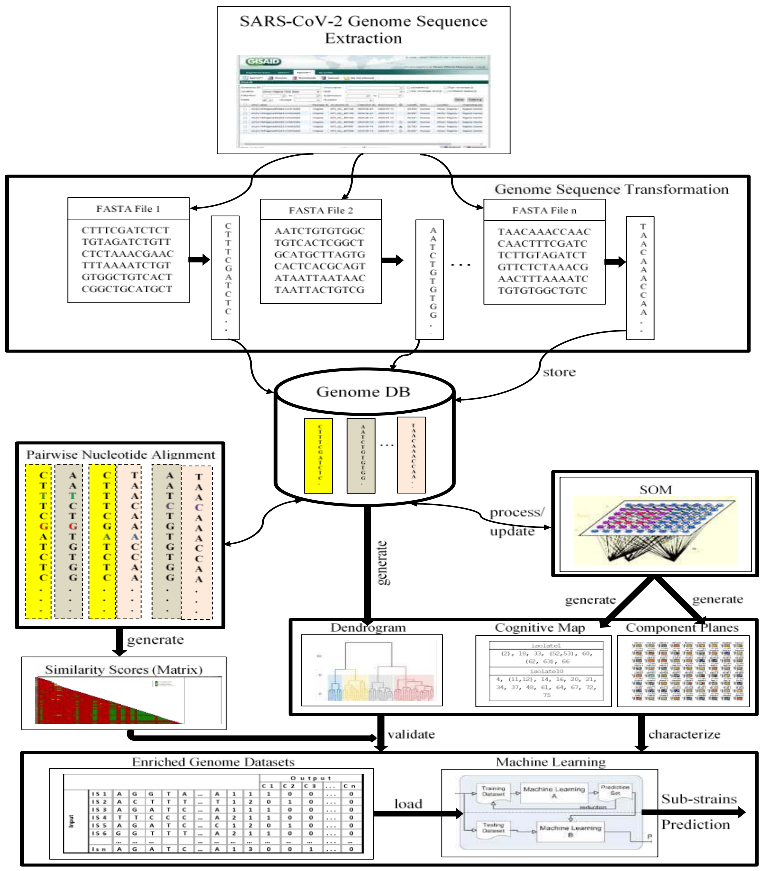


Fig. 1. Workflow describing the proposed hybrid approach. The workflow begins with the excavation of FASTA files of human SARS-CoV-2 genome sequences from GISAID. These files were stripped and processed into a genome database (DB) as multiple columns of nucleotide sequence. A series of AI/ML techniques were applied to extract knowledge from the genome datasets as follows: Using ML techniques, 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/phylogenetic tree. Determine the optimal number of natural clusters–to provide additional knowledge for supervised learning. Separate the viral sub-strains using SOM component planes–for possible 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 prediction. Learn and predict new/emerging sub-strains using ANN.

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

1. import necessary libraries
2. set path to current directory
3. #Genome nucleotide fragments processing

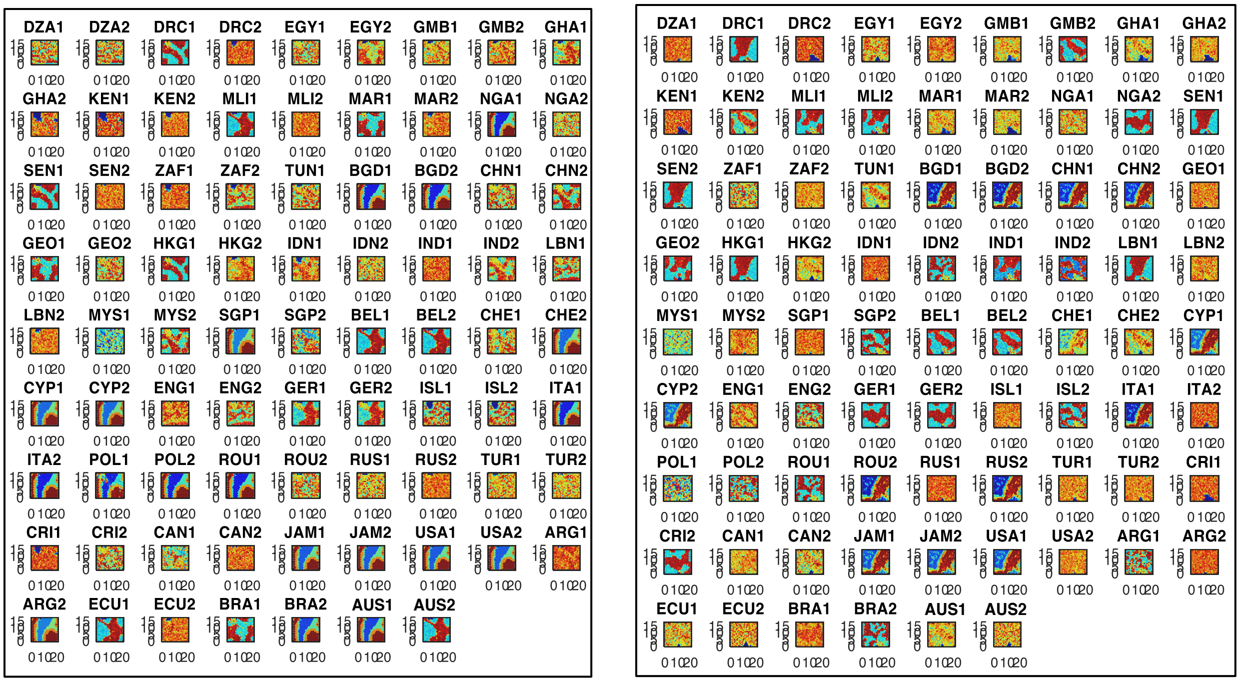
```
4.  create a list of FASTA files (fasta_list) to process
5.  for file_name in fasta_list:
6.      open FASTA_file for read
7.      store a line of genome sequence
8.      for line in file_name:
9.          strip line into a list of nucleotide fragments (nucleotide_fragments)
10.     for line in nucleotide_fragments:
11.         write nucleotide code into complete genome file (complete_genome)
12.     close FASTA_file
13. #Direct nucleotide alignment and similarity scores generation
14. open complete_genome for read
15. store a line of nucleotide code
16. for line in complete_genome:
17.     align nucleotide pair and compare nucleotide code
18. build pairwise dis(similarity) matrix using a suitable distance metric (e.g.,
    Euclidean distance)
19. #AGNES/hierarchical clustering: generate phylogenomic tree and cluster plots
20. treat observations (nucleotides) as cluster points and compute AGNES
    distance coefficients between clusters
21. 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. learn and classify genome (isolate) patterns characterized by (generated)
    cognitive maps, using 3-layered artificial neural network (ANN).
32. predict SARS-CoV-2 sub-strain
33. close complete_genome.
```

Hierarchical Clustering Analysis (Agglomerative Nesting: AGNES)

Results of the distance measures show that the ward method has the highest agglomerative coefficient of (male=0.9936; female=0.9880), indicating more compact clusters; closely followed by complete (male=0.9882; female=0.9754); average (male=0.9874; female=0.9694); and single (male=0.9776; female=0.9392) methods. The HCA or AGNES plots (see Fig. 2 of the methods section) therefore suggests an inevitable sub-strains (independent) mutant accumulation in different countries, while few mild divergent strains with specific mutations are geographically different. To determine if differences exist in the genome sequences between genders, an independent t-test was run on the AGNES dis(similarity) scores. Results showed that males had statistically insignificantly longer genome sequences (0.9727 ± 0.0376) compared to females (0.9673 ± 0.0341), $t(515) = 1.71$, $p = 0.087$. However, there was no statistically significant difference in mean similarity between the nucleotide structures of the two groups at 95% confidence interval, hence, no significant genetic variations were observed. This result corroborates the findings in Ke et al.[20] and validates their claim that no significant genetic variation exists in human SARS-CoV-2 genomes for both groups.

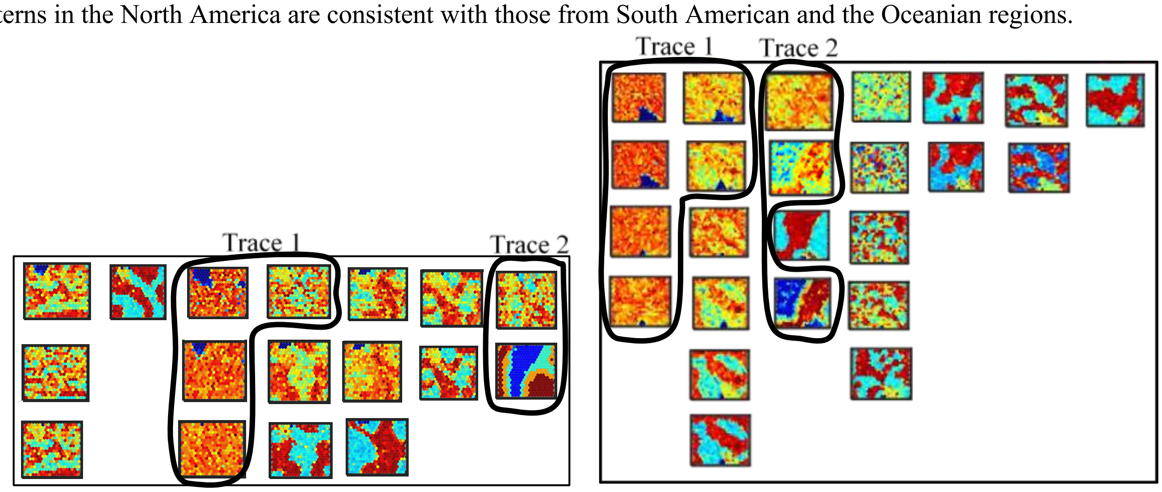
SOM Pattern Analysis
Component planes visualization reveals the distribution of single feature values on a SOM map. Fig. 2 shows SOM component planes generated to study genome dissimilarities between selected countries, for both genders; and was achieved by classifying observed pattern varieties of the main cluster groups. We observe that patients with prolonged prognosis (hospitalized, alive, released) present sharp pattern clusters with well separated boundaries indicating large variance between nucleotide sequences. This implies that, about 50% of the GISAID datasets excavated for this study were severe (late) cases of COVID-19, identified by sharp cluster patterns. Early symptomatic and asymptomatic cases however exhibit no defined cluster patterns, indicating small variance between nucleotide sequence and possible mild mutation[29]. The component planes permit an investigation of continents that share similar sub-strain(s) of SARS-CoV-2 and which sub-strain(s) permeate the different regions.

Viral Sub-strains Discrimination
To compare inter-continent pattern varieties exhibited by male and female genomes, SOMs with dis(similar) pattern characteristics were filtered and presented in Fig. 3. Although both genders exhibit near-dissimilar patterns, some of the viral sub-strains progressively transformed into a next sub-strain indicating the presence of amino acid variants and strong nucleotide association derived from same origin. Aside the general claim that patients with co-morbidities, aged, males have worse prognosis during COVID-19 infections[30], evidence of immune tolerance of females and slow prognosis[31] corroborates our claim of slow pattern transformation for female cases. A distribution of sub-strain clusters across the genome datasets for both genders, explains the diversity of SARS-CoV-2. The arrangement of clusters in this case is not relevant, as some sub-strains evolved from a previous strain(s) showing similar genetic/genomic patterns (Fig. 3). By decoupling the SOMs, we were able to group the SOMs by dominant clusters. Results obtained indicate the presence of 2 dominant clusters for both genders (male vs. female), i.e., (trace 1: 38 [48.10%] vs. trace 1: 37 [47.44%]) and (trace 2: 19 [24.05%] vs. trace 2: 13 [16.67%]).



(a) Male patients **(b) Female patients**

Fig. 2. SOM component planes visualization. Maps are ordered by countries using Alpha-3 code (ISO 3166) representation, with at least 1 isolate per country. We observe single-, double- and multiple-source transmissions and discuss the transmission routes in a combinatorial manner (i.e., order is not a factor). In the African region, females show fewer country transmissions, as 7 out of 12 countries have similar pattern correlates with genomes from the South Asian; West Europe; Asian/Europe and Central American regions; compared to males, where 9 out of 12 countries show similar pattern correlates with genomes from the Southeast and West Asia; Southeast, South-Central and Western Europe; South and Central American; and the Oceanian regions. Furthermore, while male genomes show moderately weak pattern correlates with genomes from Germany, female genomes show very strong pattern correlates, and moderately weak pattern correlates with genomes from Belgium. Genome pattern correlates of males and females in the Asian region are consistent with patterns from Southeast and South-Central Europe; Eastern Mediterranean; North and South American; save Asia/Europe and Oceanian regions, which genome pattern correlates are similar for Asian males and females, respectively. Genome pattern correlates of males and females in South America are consistent with patterns from the North American and the Oceania regions, while genome patterns in the North America are consistent with those from South American and the Oceanian regions.



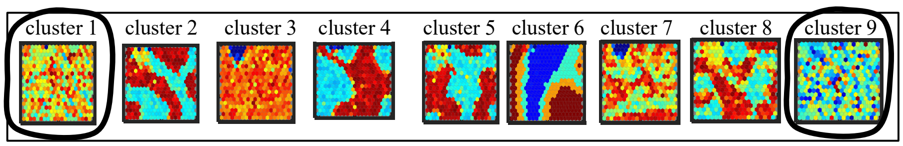
(a) Male patients **(b) Female patients**

Fig. 3. SARS-CoV-2 sub-strains clusters. We observe 7 distinct sub-strain clusters in both cases, with 4 of the sub-strains having more than 2 transitions before yielding sharp distinct clusters/patterns. While male SOM component planes exhibit rapid transitions unto sharp, distinct clusters (Fig. 3a), female SOM plots show slow transitions (Fig. 3b) and a late appearance of sharp, distinct clusters.

Cognitive knowledge extraction for intelligent sub-strains prediction

To intelligently predict the viral sub-strains for both genders, novel cognitive maps that preserve chains of similar sub-strains isolates were generated from SOM component planes using the Python programming language (Fig. 4 and Fig. 5). For male patients, 2 more clusters, 1 and 9 (Fig. 4b), derived from isolates [(1,2), 18, 33, (52,53), 60, (62, 63), 66] and [38, 41, 66], respectively (Fig. 4a) were further separated; where (x,y) are similar isolate pairs from same country, and which countries are as ordered in Fig. 2. For females, 3 more clusters, 1, 2 and 8 (Fig. 5b), derived from isolates [1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75]; [2, (18,19), 29, 35] and [32, 57, 71, 76], respectively (Fig. 5a) were further separated. Additionally, females have fewer inter-similarity links, confirming fewer transmission routes, compared to males, an evidence that our cognitive solution can not only separate hidden or evolving SARS-CoV-2 patterns for precise sub-strains prediction, but also enables efficient inter-country contact tracing. Data file storing generated cognitive maps are found in Data S4 (SupplData4.xlsx).

Learning the enriched genome datasets using a 3-layered ANN, characterized by the generated cognitive maps produced Table 1. In table 1, the ANN was more accurate in predicting female isolates, as the MSE performances for both train and test datasets yield better performance than male isolates. Defuzzification of confusable SOM patterns for performance comparison with our cognitive solution and an intra-country sub-strains prediction are possible future research directions of this paper.

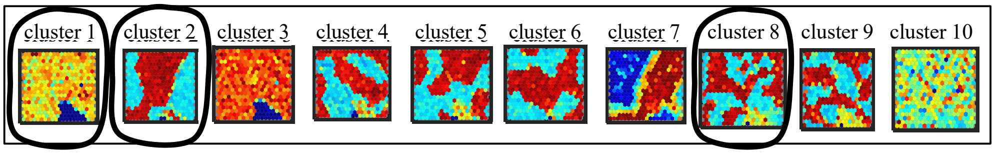


(a) Derived component planes

isolate1	isolate2	isolate3	isolate4	isolate5	isolate6	isolate7	isolate8	isolate9
(2), 18, 33, (52,53), 60, (62, 63), 66	(1), 18, 33, (52,53), 60, (62,63), 66	19, 30	10, (11, 12), 14, 16, 20, 21, 34, 37, 48, 61, 64, 67, 72, 75	(6), 15, 26, 28, 59, 65, 76	(5), 15, 26, 28, 59, 65, 76	(8), 9, 13, 23, 29, 31, (42,43), (50,51), 74, 79	(7), 9, 13, 23, 29, 31, (42,43), (50,51), 74, 79	(7,8), 13, 23, 29, 31, (42,43), (50,51), 74, 79
isolate10	isolate11	isolate12	isolate13	isolate14	isolate15	isolate16	isolate17	isolate18
4, (11,12), 14, 16, 20, 21, 34, 37, 48, 61, 64, 67, 72, 75	4, 10, 12, 14, 16, 20, 21, 34, 37, 48, 61, 64, 67, 72, 75	4, 10, 11, 14, 16, 20, 21, 34, 37, 48, 61, 64, 67, 72, 75	(7,8), 9, 23, 29, 31, (42,43), (50,51), 74, 79	4, 10, (11,12), 16, 20, 21, 34, 37, 48, 61, 64, 67, 72, 75	(5,6), 26, 28, 59, 65, 76	4, 10, (11,12), 20, 21, 34, 37, 48, 61, 64, 67, 72, 75	(24,25), 40, 45, (46,47), (54,55), (56,57), 58, (68,69), (70,71), 73, (77,78)	(1,2), 33, (52,53), 60, (62,63), 66
isolate19	isolate20	isolate21	isolate22	isolate23	isolate24	isolate25	isolate26	isolate27
3, 30	4, 10, (11,12), 14, 16, 21, 34, 37, 48, 61, 64, 67, 72, 75	4, 10, (11,12), 14, 16, 20, 34, 37, 48, 61, 64, 67, 72, 75	36, 49	(7,8), 9, 13, 29, 31, (42,43), (50,51), 74, 79	17, 25, 40, 45, (46,47), (54,55), (56,57), 58, (68,69), (70,71), 73, (77,78)	17, 24, 40, 45, (46,47), (54,55), (56,57), 58, (68,69), (70,71), 73,	(5,6), 15, 28, 59, 65, 76	32, 35, 39, 44
isolate28	isolate29	isolate30	isolate31	isolate32	isolate33	isolate34	isolate35	isolate36
(5,6), 15, 26, 59, 65, 76	(7,8), 9, 13, 23, 31, (42,43), (50,51), 74, 79	3, 19	(7,8), 9, 13, 23, 29, (42,43), (50,51), 74, 79	27, 35, 39, 44	(1,2), 18, (52,53), 60, (62,63), 66	4, 10, (11,12), 14, 16, 20, 21, 37, 48, 61, 64, 67, 72,	27, 32, 39, 44	22, 49
isolate37	isolate38	isolate39	isolate40	isolate41	isolate42	isolate43	isolate44	isolate45
4, 10, (11,12), 14, 16, 20, 21, 34, 48, 61, 64, 67, 72, 75	41, 66	27, 32, 35	17, (24,25), 45, (46,47), (54,55), (56,57), 58, (68,69), (70,71), 73, (77,78)	38, 66	(7,8), 9, 13, 23, 29, 31, (43), (50,51), 74, 79	(7,8), 9, 13, 23, 29, 31, (42), (50,51), 74, 79	27, 32, 35, 39	17, (24,25), 40, (46,47), (54,55), (56,57), 58, (68,69), (70,71), 73, (77,78)
isolate46	isolate47	isolate48	isolate49	isolate50	isolate51	isolate52	isolate53	isolate54
17, (24,25), 40, 45, 47, (54,55), (56,57), 58, (68,69), (70,71), 73, (77,78)	17, (24,25), 40, 45, 46, (54,55), (56,57), 58, (68,69), (70,71), 73, (77,78)	4, 10, (11,12), 14, 16, 20, 21, 34, 37, 61, 64, 67, 72, 75	22, 36, 49	9, 13, 23, 29, 31, (42,43), 51, 74, 79	9, 13, 23, 29, 31, (42,43), 50, 74, 79	(1,2), 18, 33, 53, 60, (62,63), 66	(1,2), 18, 33, 52, 60, (62,63), 66	17, (24,25), 40, 45, 46, (54,55), (56,57), 58, (68,69), (70,71), 73, (77,78)
isolate55	isolate56	isolate57	isolate58	isolate59	isolate60	isolate61	isolate62	isolate63
17, (24,25), 40, 45, (46,47), 54, (56,57), 58, (68,69), (70,71), 73, (77,78)	17, (24,25), 40, 45, (46,47), (54,55), 57, 58, (68,69), (70,71), 73, (77,78)	17, (24,25), 40, 45, (46,47), (54,55), 56, 58, (68,69), (70,71), 73, (77,78)	17, (24,25), 40, 45, (46,47), (54,55), (56,57), (68,69), (70,71), 73, (77,78)	(5,6), 15, 26, 28, 65, 76	(1,2), 18, 33, (52,53), (62,63), 66	4, 10, (11,12), 14, 16, 20, 21, 34, 37, 48, 67, 72, 75	(1,2), 18, 33, (52,53), 60, (63), 66	(1,2), 18, 33, (52,53), 60, (62), 66
isolate64	isolate65	isolate66	isolate67	isolate68	isolate69	isolate70	isolate71	isolate72
4, 10, (11,12), 14, 16, 20, 21, 34, 37, 48, 61, 67, 72, 75	(5,6), 15, 26, 28, 59, 76	(1,2), 18, 33, (52,53), 60, (62, 63)	4, 10, (11,12), 14, 16, 20, 21, 34, 37, 48, 61, 64, 72, 75	17, (24,25), 40, 45, (46,47), (54,55), (56,57), 58, 69, (70,71), 73, (77,78)	17, (24,25), 40, 45, (46,47), (54,55), (56,57), 58, 68, (70,71), 73, (77,78)	17, (24,25), 40, 45, (46,47), (54,55), (56,57), 58, (68,69), 71, 73,	17, (24,25), 40, 45, (46,47), (54,55), (56,57), 58, (68,69), 70, 73, (77,78)	4, 10, (11,12), 14, 16, 20, 21, 34, 37, 48, 61, 64, 67, 75
isolate73	isolate74	isolate75	isolate76	isolate77	isolate78	isolate79		
17, (24,25), 40, 45, (46,47), (54,55), (56,57), 58, (68,69), (70,71), (77,78)	(7,8), 9, 13, 23, 29, 31, (42,43), (50,51), 79	4, 10, (11,12), 14, 16, 20, 21, 34, 37, 48, 61, 64, 67, 72	(5,6), 15, 26, 28, 59, 65	17, (24,25), 40, 45, (46,47), (54,55), (56,57), 58, (68,69), (70,71), 73, 78	17, (24,25), 40, 45, (46,47), (54,55), (56,57), 58, (68,69), (70,71), 73, 77	(7,8), 9, 13, 23, 29, 31, (42,43), (50,51), 74		

(b) Cognitive map

Fig. 4. Derived SOM component planes and cognitive map for male genome isolates. The map provides cognitive solution for sub-strains link modeling and reveal hidden structures which hitherto were subsumed in previously large cluster groups (Fig. 3a), hence, corroborating the possibility of confusable clusters[22].



(a) Derived component planes

isolate1	isolate2	isolate3	isolate4	isolate5	isolate6	isolate7	isolate8	isolate9
3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	(18,19), 29, 35	1, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	(14,15), 16, 20, 65, (73,74), (77,78)	1, 3, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	(7), 8, 11, 22, 30, (41,42), 44, 66	(6), 8, 11, 22, 30, (41,42), 44, 66	(6,7), 11, 22, 30, (41,42), 44, 66	1, 3, 5, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75
isolate10	isolate11	isolate12	isolate13	isolate14	isolate15	isolate16	isolate17	isolate18
1, 3, 5, 9, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	(6, 7), 8, 22, 30, (41,42), 44, 66	(13), 28, 33	(12), 28, 33	4, (15), 16, 20, 65, (73,74), (77,78)	4, (14), 16, 20, 65, (73,74), (77,78)	4, (14,15), 20, 65, (73,74), (77,78)	(49,50), 64	2, (19), 29, 35
isolate19	isolate20	isolate21	isolate22	isolate23	isolate24	isolate25	isolate26	isolate27
2, (18), 29, 35	4, (14,15), 16, 20, 65, (73,74), (77,78)	1, 3, 5, 9, 10, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	(6,7), 8, 11, 30, (41,42), 44, 66	(24), (25,26), 43, (45,46), 53, 58, 60, (67,68), 69	(23,24), (25,26), 43, (45,46), 53, 58, 60, (67,68), 69	(23,24), (26), 43, (45,46), 53, 58, 60, (67,68), 69	(23,24), (25), 43, (45,46), 53, 58, 60, (67,68), 69	1, 3, 5, 9, 10, 21, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75
isolate28	isolate29	isolate30	isolate31	isolate32	isolate33	isolate34	isolate35	isolate36
(12,13), 33	2, (18,19), 35	(6,7), 8, 11, 22, (41,42), 44, 66	1, 3, 5, 9, 10, 21, 27, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	57, 71, 76	(12, 13), 28, 33	40, 48, 52, 56	2, (18,19), 29	1, 3, 5, 9, 10, 21, 27, 31, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75
isolate37	isolate38	isolate39	isolate40	isolate41	isolate42	isolate43	isolate44	isolate45
55	1, 3, 5, 9, 10, 21, 27, 31, 36, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	34, 48, 52, 56	(6,7), 8, 11, 22, 30, (42), 44, 66	(6,7), 8, 11, 22, 30, (41), 44, 66	(23,24), (25,26), (45,46), 53, 58, 60, (67,68), 69	(6,7), 8, 11, 22, 30, (41,42), 66	(23,24), (25,26), 43, (46), 53, 58, 60, (67,68), 69
isolate46	isolate47	isolate48	isolate49	isolate50	isolate51	isolate52	isolate53	isolate54
(23,24), (25,26), 43, (45), 53, 58, 60, (67,68), 69	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 51, 54, 59, (61,62), 63, 70, 72, 75	34, 40, 52, 56	17, (50), 64	17, (49), 64	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 54, 59, (61,62), 63, 70, 72, 75	34, 40, 48, 56	(23, 24), (25,26), 43, (45,46), 58, 60, (67,68), 69	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 59, (61,62), 63, 70, 72, 75
isolate55	isolate56	isolate57	isolate58	isolate59	isolate60	isolate61	isolate62	isolate63
37	34, 40, 48, 52	32, 71, 76	(23,24), (25,26), 43, (45,46), 53, 60, (67,68), 69	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, (61,62), 63, 70, 72, 75	(23,24), (25,26), 43, (45,46), 53, 58, (67,68), 69	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (62), 63, 70, 72, 75	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61), 63, 70, 72, 75	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 70, 72, 75
isolate64	isolate65	isolate66	isolate67	isolate68	isolate69	isolate70	isolate71	isolate72
17, (49,50)	4, (14,15), 16, 20, (73,74), (77,78)	(6,7), 8, 11, 22, 30, (41,42), 44	(23,24), (25,26), 43, (45,46), 53, 58, 60, (68), 69	(23,24), (25,26), 43, (45,46), 53, 58, 60, (67), 69	(23,24), (25,26), 43, (45,46), 53, 58, 60, (67,68)	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 72, 75	32, 57, 76	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 75
isolate73	isolate74	isolate75	isolate76	isolate77	isolate78			
4, (14,15), 16, 20, 65, (74), (77,78)	4, (14,15), 16, 20, 65, (73), (77,78)	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72	32, 57, 71	4, (14,15), 16, 20, 65, (73,74), (78)	4, (14,15), 16, 20, 65, (73,74), (77)			

(b) Cognitive map

Fig. 5. Derived SOM component planes and cognitive map and for female genome isolates. The map provides cognitive solution for sub-strains link modeling and reveal hidden structures which hitherto were subsumed in previously large cluster groups (Fig. 3b), hence, corroborating the possibility of confusable clusters[22].

Table 1. ANN performance analysis

Enriched Genome Datasets	MSE Performance				Overall Accuracy (%)
	Overall	Train	Test	Validation	
Male patients	0.0056	0.0033	0.0024	0.0210	97.50
Female patients	0.0047	0.0025	0.0021	0.0184	98.70

References

1. Mitchell, E. P. Corona Virus: Global Pandemic Causing World-Wide Shutdown. Journal of the National Medical Association. 112(2): 113-114. (2020). https://doi.org/10.1016/j.jnma.2020.03.015.

2. Bedford, J., Enria, D., Giesecke, J., Heymann, D.L., Ihekweazu, C., Kobinger, G., Lane, H.C., Memish, Z., Oh, M.D., Schuchat, A. and Ungchusak, K. COVID-19: towards controlling of a pandemic. The Lancet, 395(10229), 1015-1018. (2020). https://doi.org/10.1016/S0140-6736(20)30673-5.

3. Chen, J., Wang, R., Wang, M., & Wei, G. W. Mutations strengthened SARS-CoV-2 infectivity. arXiv preprint arXiv:2005.14669. (2020). https://arxiv.org/abs/2005.14669

4. Koyama, T., Weeraratne, D., Snowdon, J. L., & Parida, L. Emergence of drift variants that may affect COVID-19 vaccine development and antibody treatment. Pathogens, 9(5), 324: 1-7. (2020). https://doi.org/10.3390/pathogens9050324.
5. Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R. and Niu, P. A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine. 382(8):727–733. (2020). https://doi.org/10.1056/NEJMoa2001017
6. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev. 7: 1012–1023. (2020). https://doi.org/10.1093/nsr/nwaa036
7. Wiechers, I. R., Perin, N. C. & 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 5, 83: 1-9. (2013). https://doi.org/10.1186/gm487
8. Giani, A. M., Gallo, G. R., Gianfranceschi, L., & Formenti, G. Long walk to genomics: History and current approaches to genome sequencing and assembly. Computational and Structural Biotechnology Journal, 18, 9-19. (2020). https://doi.org/10.1016/j.csbj.2019.11.002.
9. Ardakani A.A., Kanafi A.R., Acharya U.R., Khadem N., Mohammadi A. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: results of 10 convolutional neural networks. Comput Biol Med. 121. 103795. (2020). https://doi.org/10.1016/j.compbiomed.2020.103795.
10. Ozturk T., Talo M., Yildirim E.A., Baloglu U.B., Yildirim O., Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med. 121. 103792. (2020). https://doi.org/10.1016/j.compbiomed.2020.103792.
11. 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. 104431. (2020). https://doi.org/10.1016/j.jcv.2020.104431.
12. Wu, J., Zhang, P., Zhang, L., Meng, W., Li, J., Tong, C., Li, Y., Cai, J., Yang, Z., JZhu, J., Zhao, M., Huang, H., Xie, X. and Li, S. (2020). Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. medRxiv Preprint. https//doi.org/10.1101/2020.04.02.20051136.
13. MIT: Covid Tracing Tracker - a flood of coronavirus apps are tracking us. Now it's time to keep track of them. https://www.technologyreview.com/2020/05/07/1000961/launching-mittr-covid-tracing-tracker/ (accessed 20 August, 2020).
14. 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. 109853. https://doi.org/10.1016/j.chaos.2020.109853.
15. Yan L., Zhang H.-T., 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. 1-6. (2020). https://doi.org/10.1038/s42256-020-0180-7.

16. Ke Y.-Y., Peng T.-T., Yeh T.-K., Huang W.-Z., Chang S.-E., Wu S.-H., Hung H.-C., Hsu T.-A., Lee S.-J., Song J.-S., Lin W.-H., Chiang T.-J., Lin J.-H., Sytwu H.-K., Chen C.-T. Artificial intelligence approach fighting COVID-19 with repurposing drugs. Biomed J. (2020). https://doi.org/10.1016/j.bj.2020.05.001.

17. Beck B. R., 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.;18: 784–790. (2020). https://doi.org/10.1016/j.csbj.2020.03.025.

18. Ekins S., Mottin M., Ramos P.R.P.S., Sousa B.K.P., Neves B.J., Foil D.H., Zorn K.M., Braga R.C., Coffee M., Southan C., Puhl C.A., Andrade C.H. Déjà vu: stimulating open drug discovery for SARS-CoV-2. Drug Discov Today. 25(5): 928–941. (2020). https://doi.org/10.1016/j.drudis.2020.03.019.

19. Hussain, M., Jabeen, N., Raza, F., Shabbir, S., Baig, A. A., Amanullah, A., & Aziz, B. Structural variations in human ACE2 may influence its binding with SARS-CoV-2 spike protein. Journal of medical virology. 92:1580–1586. (2020). https://doi.org/10.1002/jmv.25832

20. Li, M. Y., Li, L., Zhang, Y., & Wang, X. S. Expression of the SARS-CoV-2 cell receptor gene ACE2 in a wide variety of human tissues. Infectious diseases of poverty, 9, 1-7. (2020). https://doi.org/10.1186/s40249-020-00662-x.

21. Stefanelli, P., Faggioni, G., Presti, A.L., Fiore, S., Marchi, A., Benedetti, E., Fabiani, C., Anselmo, A., Ciammaruconi, A., Fortunato, A. and De Santis, R. V. Whole genome and phylogenetic analysis of two SARS-CoV-2 strains isolated in Italy in January and February 2020: additional clues on multiple introductions and further circulation in Europe. Eurosurveillance, 25(13), 2000305. (2020). https://doi.org/10.2807/1560-7917.ES.2020.25.13.2000305

22. Somasundaram, K., Mondal, M., & Lawarde, A. Genomics of Indian SARS-CoV-2: Implications in genetic diversity, possible origin and spread of virus. medRxiv. (2020). https://doi.org/10.1101/2020.04.25.20079475

23. Lopez-Rincon, A., Tonda, A., Mendoza-Maldonado, L., Claassen, E., Garssen, J., & Kraneveld, A. D. Accurate identification of sars-cov-2 from viral genome sequences using deep learning. bioRxiv. (2020). https://doi.org/10.1101/2020.03.13.990242.

24. Randhawa, G. S., Soltysiak, M. P., El Roz, H., de Souza, C. P., Hill, K. A., & Kari, L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. Plos one, 15(4), e0232391. (2020). https://doi.org/10.1371/journal.pone.0232391

25. Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., & Din, M. M. U. Machine learning based approaches for detecting COVID-19 using clinical text data. International Journal of Information Technology, 1-9. (2020). https://doi.org/10.1007/s41870-020-00495-9.

26. Melin, P., Monica, J. C., Sanchez, D., & Castillo, O. Analysis of Spatial Spread Relationships of Coronavirus (COVID-19) Pandemic in the World using Self Organizing Maps. Chaos, Solitons & Fractals, 1-7. (2020). https://doi.org/10.1016/j.chaos.2020.109917.

27. The Humanitarian Data Exchange (HDX), [Online] (2020). Available: https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases.

28. Villmann, T., Kaden, M., Bohnsack, K. S., Weber, M., Kudla, M., Gutowska, K., & Blazewicz, J. Analysis of SARS-CoV-2 RNA-Sequences by Interpretable Machine Learning Models. bioRxiv. (2020). https://doi.org/10.1101/2020.05.15.097741

29. Kuehn, B. M. Genetic Analysis Tracks SARS-CoV-2 Mutations in Human Hosts. Jama, 323(23), 2363-2363. (2020). https://doi.org/10.1001/jama.2020.9825.

30. 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.

31. 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.

Authors Contributions:

All authors contributed equally to the final draft of this paper.

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 software design component and implementation.

I.E. contributed to the biotechnology and bioinformatics components of the paper, as well as the research methods.

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

M.E.2 structurally edited the original draft, involved in research data validation and interpretation of results.

Y.T. contributed to the biotechnology/bioinformatics components of the paper and editing of the original draft.

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

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

O.E. contributed to the study background, research methods and biotechnology components of the paper.

G.J. was involved in the 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.

Competing Interest: There are no competing interests.

Methods

Data Source and Genome Sequences Selection

Publicly available datasets of coronavirus cases around the globe deposited between December 2019 and August 20, 2020 were excavated from GISAID (a database of SARS-CoV-2 partial and complete genome compilations distributed by clinicians and researchers, the world over), for the purpose of this study. Four complete genome sequences of human SARS-CoV-2 isolates (2 males, 2 females) from selected countries of the world, across 6 continents, Antartica exempt (as no deposit of SARS-CoV-2 data was found as at the time of excavation), were collected. Useful metadata on the extracted genome sequences (Continent, Region, Country, Abbreviation, Accession No., Length, Gender, Age, Specimen source, Status, Submitting Lab, Authors) were documented (see Data S1: SupplData1.xlsx). No information about the travel history of the patients for extracted genomes were sieved as most of the excavated data lacked this very information. The preprocessed FASTA files of genome isolates excavated from GISAID, striped and dumped as column sequences for male and female patients are found in (Data S2: SupplData2) and (Data S3: SupplData3), respectively. Although incomplete data for some of the datasets such as age, specimen source and status were noticed, gender was a compulsory criterion for profiling the excavated genomes. Aside Tunisia, which had informative deposit about gender of patients for only one pair (1 male, 1 female) and Algeria for (2 males, 1 female), other SARS-CoV-2 isolates had complete pairs (2 males, 2 females). A total of 157 genome sequences (79 males, 78 females) with genome lengths of over 29000 nucleotides, were excavated from 40 different countries distributed across the following continents: Africa, Asia, Europe, North America, South America and Oceania. 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, although the ages of 7 patients from Africa and 3 patients from Asia/Europe were unknown. Finally, about 0.70% and 0.79% of errors in sequencing (noise) were noticed in the male and female genome datasets, respectively.

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.

Genomic Epidemiology of SARS-CoV-2

Due to the naturally expanding genetic diversity of COVID-19, GISAID introduced a nomenclature for grouping major clades based on marker mutations within 6 high-level phylogenetic groupings from the early split of S and L, to the further evolution of L into V and G and later of G into GH and GR; augmented with more detailed lineages assigned by PANGOLIN

(Phylogenetic Assignment of Named Global Outbreak LINeages) tool, to aid the understanding of patterns and determinants of the global spread of the pandemic strain causing COVID-19. Showing 5029 genomes sampled between December 2019 and August 2020, the GR and GH are becoming more dominant with sparse concentration of new deposits, lately, indicating reduced spread of the virus (see Fig. 1a). A more recent effort uses a Year-Letter nomenclature to facilitate the large-scale pattern diversity of COVID-19 and label clades that persist for at least several months with significant geographic spread (see Fig. 1b). Fig 2 reveals that SARS-CoV-2 mutation rate has remained low, signifying good news for vaccine developers.

Comparing genome content has become commonplace but associating its order presents a wider range of problems. Hence, methods for complete genome phylogenetic analysis should show some evidence of robustness against incomplete or inaccurate information. Complete genome-based phylogenetic trees appear not widely used because of computational difficulties (massive data and limited processing infrastructure). In this paper, we exploit complete genome sequences to construct hierarchical cluster structures (dendrograms) that discriminate inter-genetic diversity of SARS-CoV-2 among male and female patients.

Hierarchical Agglomerative Clustering (HAC)

The dataset is configured with observations (nucleotides) represented in rows, while columns are variables (genome sequences ordered by countries). The number of columns corresponds to selected countries while the sequences have varying lengths. The data table is further converted into *as.matrix* format where all values of raster layers objects have columns for each layer and rows for each cells with numeric (continuous) values. In order to make the variables comparable through the elimination of arbitrary variable units, they are transformed (standardized) such that they have mean of zero and standard deviation of unity[32], using equation (1).

$$x(s) = x_i - \frac{mean(x)}{sd(x)}, \tag{1}$$

where $sd(x)$ represents the standard deviation of the feature values.

The procedure for implementing the HAC are as follows: Compute all the pairwise similarities (distances) between observations in the dataset and represent the result as a matrix. The resultant matrix is square and symmetric with diagonal members defined as unity–the measure of similarity between an element and itself. The matrix elements are computed by iterating over each element and calculating its (dis)similarity to every other element. Suppose A is a similarity matrix of size $N \times N$, and B, a set of N elements. A_{ij} is the similarity between elements B_i and B_j using a specified criterion (Euclidean distance, squared Euclidean distance, manhattan distance, maximum distance, Mahalanobis distance, cosine similarity). The selected criterion however depends on the nature of the experimental datasets. This paper adopts the standardized Euclidian distance criterion, as this measure is widely used and has shown good performance in the modeling variances in biological sequences.

HAC Visualization

After calculating the distance between every pair of observation point, the result is stored in a distance matrix. Then, (i) every point is put in its own cluster (i.e., the initial number of clusters corresponds to the number of variables); (ii) the closest pairs of points are merged based on the distances from the distance matrix as the number of clusters reduces by 1; (iii) the distance between the new cluster and the previous ones is recomputed and stored in a new distance matrix; (iv) steps (ii) and (iii) are repeated until all the clusters are merged into one single cluster.

The distance separating the clusters is specified via linkage methods[32] which includes, complete, average, single, and ward. Complete linkage computes the similarities and uses the maximum distance between clusters for merging while calculating cluster distances and adopting minimum inter-cluster distance merging. Average linkage calculates the average distance between groups of genome sequence before merging; while the total within-cluster variance is minimized with ward's method and the pair of clusters with minimum between-cluster distance are merged. We rely on all the four techniques for assessment and adopt the distance measure with the highest agglomerative coefficient for cluster formation. The resultant cluster solution is finally visualized as a tree structure called a dendrogram (or phylogenetic) tree. As the tree is traversed upwards, observations that are similar to each other are combined into branches, which are themselves fused at a higher height. The height of the fusion, provided on the vertical axis, indicates the (dis)similarity between two observations. The higher the height of the fusion, the less similar the observations are. Fig. 2. show cluster plots and genomic plots generated using the ward minimum variance criterion.

Optimal Natural Clusters Selection

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 adopts three widely used criteria to validate the number of clusters in the genome sequence dataset namely, silhouette, elbow[33], and gap-statistics with the aim of minimizing the total intra-cluster variation (total within-cluster sum of square) as given in equation (2).

$$minimize\left(\sum_{i=1}^{k} w(c_k)\right) \tag{2}$$

where c_k is the kth cluster, and, $w(c_k)$ is the within-cluster variation. The total within-cluster sum of squares (wss) measures the compactness of the clustering solution. The following steps are applied to achieve the optimal clusters: (i) Compute clustering algorithm (e.g., k-means clustering) for different values of k; by varying k from 1 to 10 clusters, for instance. (ii) For each k, calculate wss. (iii) plot the curve of wss according to the number of clusters k. (iv) the location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

Silhouette criterion is used to validate the clustering solution using pair-wise difference between the within-cluster distances, and by maximizing the value of this index to arrive at the optimal cluster number[33,34]. Elbow criterion plots the variance resulting from plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. Gap-statistics compares the total intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The reference dataset is generated using Monte Carlo simulations of the sampling process. The silhouette, elbow and gap-statistics methods rely on k-mean algorithm[35]. In this paper, the k-means algorithm is implemented in R script consisting of R functions for the silhouette, elbow and gap-statistics implementation. The decision on the choice of the optimal number of clusters is based on the results of the three methods. The clustering solution is visualized using the *fviz_cluster* function in R programming language for the grouping and extraction of genome sequences and finally represented in tree format using dendrogram.

Direct Sequence Alignment
Several techniques for biological sequence alignment (multiple or pairwise) have flourished the literature[36] and are continually being refined, but most of these techniques suffer from the lack of accuracy and partial interpretations. We perform (direct) pairwise genome sequence alignments that match each nucleotide pair at the exact nucleotide positions of the SARS-CoV-2 genome, extending the alignments across other genome pairs. The results are a set of similarity matrices for male and female patients (Fig. 3a and Fig 3b). The direct sequence alignment algorithm is embedded in Algorithm 1.

Unsupervised Genome Clustering
Several mathematical techniques have been deployed for identifying underlying patterns in complex data. These techniques, which cluster data points differently in multidimensional space are important to discover fundamental patterns of gene expression inherent in data. The clustering technique adopted in this paper is the self-organizing map (SOM) and 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 a neural-network that projects data into a low-dimensional space[37], by accepting a set of input data and then mapping the data onto neurons of a 2D grid (see Fig. 4). 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. During training, weights of the winning neuron and neurons in a predefined neighborhood are adjusted towards the input vector using equation (3),

$$w_{id}^{t+1} = w_{id}^{t} + rf(i,q)(x_d - w_{id}^{t});\ 1 \leq d \leq D. \tag{3}$$

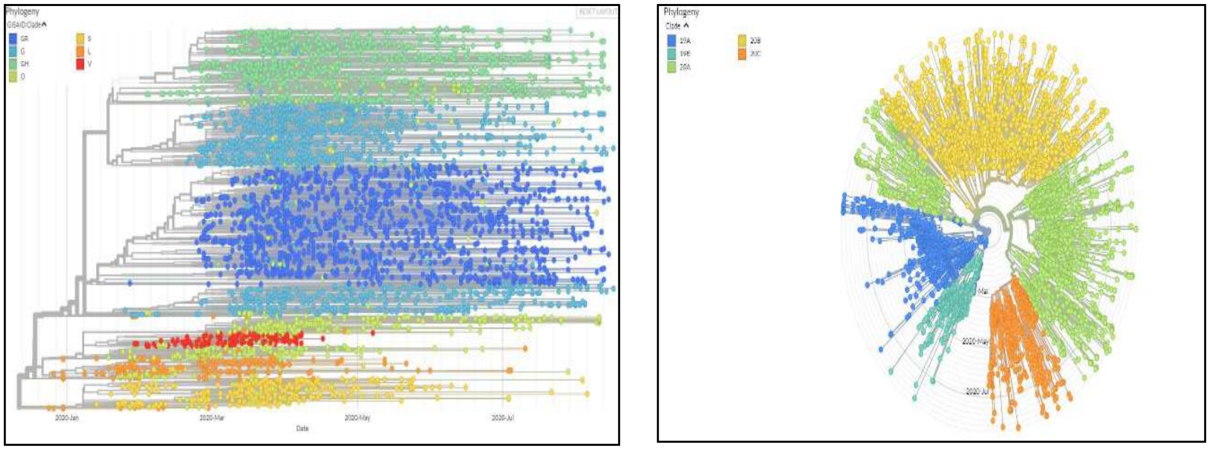
where r is the learning rate and $f(i,q)$ is the neighborhood function, with value 1 at the winning neuron q; and decreases as the distance between i and q increases. At the end, the principal features of the input data are retained, hence, making SOM a dimension reduction technique. The batch unsupervised 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 200 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.

Pattern Correlates Generation: Comparing component planes help detect similar patterns in identical positions indicating correlation between the respective components. Local correlations can also occur if two parameter planes are similar in some regions. Both linear and non-linear correlations including local or partial correlations between variables are possible (34). We achieve the correlation hunting automatically, by decoupling the SOM correlations, to explore patterns among the pairwise genome samples for distinct identification of transmission pathways or routes. The extracted correlation matrices showing pairwise relatedness of the viral sub-strains' transmissions with related pairs ($r^2 \geq 0.60$) colored in green, for male and female patients are presented in Fig. 5a and Fig. 5b, respectively.

Cognitive Knowledge Mining: Knowledge mining has served huge benefits for quick learning from big data. We apply Natural Language Processing of the genome datasets to extract knowledge of similar strains of the virus. A simple iteration technique is imposed on the SOM isolates ($i = 1,2,3, \ldots, n$), 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., $i + 1, i + 2, \ldots, n$). Concatenate compiled isolate(s) into a list ($j_1, j_2, \ldots, j_m$) where j is an element of the list. Dump the compiled list into $CogMap(k_i \in j_1, j_2, \ldots, j_m)$.

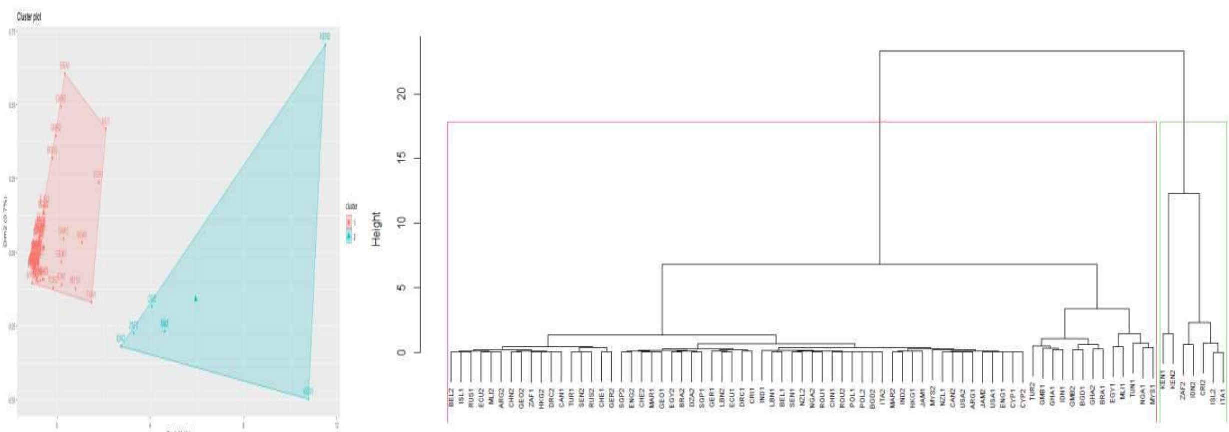
Neural Network Design: Artificial Neural Networks (ANNs) are networks inspired by the neurological structure of the human brain. They are complex computer code written with simple, highly interconnected processing elements inspired by human biological brain structure for simulating the human brain and processing data/information models. Although five core ANN areas have been explored, namely: Multi-Layer Perceptron, Radial Basis Network, Recurrent Neural Networks, Generative Adversarial Networks, and Convolutional Neural Networks; this paper adopts the Multi-Layer Perceptron model (MLP)–a class of feedforward ANNs, with at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training[39].

Fig. 6 shows our ANN architecture. A minimal number of 7 and 5 neurons for hidden layer 1 and hidden layer 2, respectively, is used in this study. The decision for choosing 7 and 5 neurons was informed by the total number of output targets and the evidence that a dropout of neurons represents computationally cheap and remarkably effective regularization method to reduce overfitting and improve generalization error in neural networks of all kinds. The output classes (C1-C7) were derived from the sub-strains discovered from learning the SOM. In training the ANN, the batch training with weight and bias learning rules 'trainb' with the Mean Squared Error (MSE) performance function, was adopted. The 'trainb' trains a network with weight and bias learning rules by continuous batch updates. The weights and biases are updated at the end of an entire pass over the input data. Training progresses according to trainb's training parameters, with a maximum of 1000 training epochs and $1e^{-6}$ minimum performance gradient.

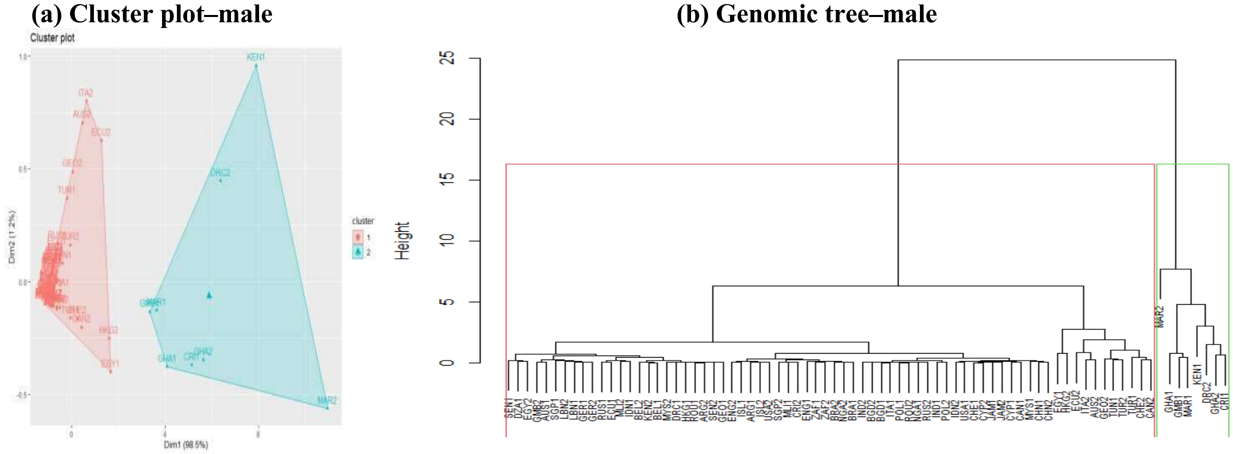


(a) Male patients **(b) Female patients**

Fig. 1. Phylogeny of SARS-CoV-2 genomic epidemiology. The original L strain surfaced in Wuhan in December 2019, with its first mutation, the S strain, appearing at the beginning of 2020; while strains V and G appeared in the mid of January 2020. Currently, strain G which mutated into strains GR and GH at the end of February 2020 are by far most prevalent, presenting 4 different mutations, two of which can change the RNA polymerase and Spike proteins sequence of the virus–a probable characteristic that facilitates the virus spread. Globally, strains G, GH and GR are becoming more dominant, while strain S, L and V are fast disappearing.

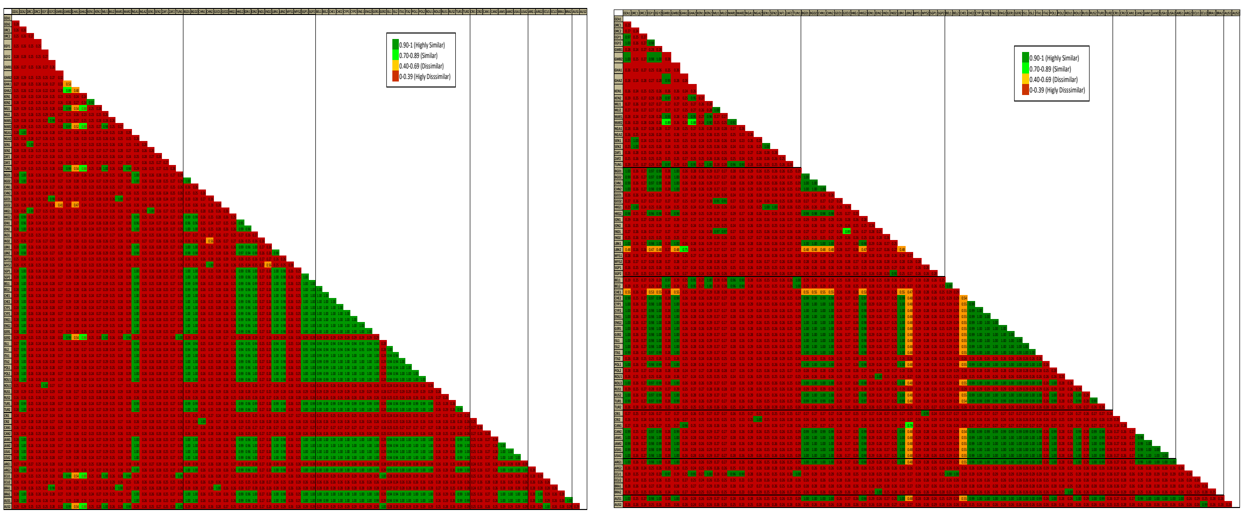


(a) Cluster plot–male **(b) Genomic tree–male**



(c) Cluster plot–female **(d) Genomic tree–female**

Fig 2. Cluster plots and genomic trees. Notice 2 distinct groups A and B separated between closely similar and dissimilar isolates, with the A group having heavy isolates concentration compared to the B group. For males (Fig. 2b), group A consists of 72 isolates with 3 sub-groups. The first sub-group contains 59 isolates from Europe, Africa, Asia, South America, North America and Oceania. The second sub-group contain 10 isolates from Europe, Africa and Asia. The third sub-group contains 3 isolates from Africa. Cluster B consists of 7 isolates distributed among 4 sub-groups namely Africa (2); Africa and Asia (2); Europe (1); and Europe and North America. For females (Fig. 2d), cluster A consists of 70 isolates with 3 sub-groups. The first sub-group contains 59 isolates from Europe, Africa, Asia, South America, North America and Oceania. The second sub-group contain 2 isolates from Africa and Asia. The third sub-group has 9 isolates distributed between North America, South America, Africa, Asia, Europe and Oceania. Cluster B consists of 8 isolates distributed among 4 sub-groups with the first 3 isolates groups from the African region, while the fourth isolate group comes from Africa and Europe regions.



(a) Male patients **(b) Female patients**

Fig. 3. Inter-continent nucleotide similarity matrices. Green colored cells are regions of high similarity that may indicate functional, structural and/or evolutionary relationships between nucleotide sequences.

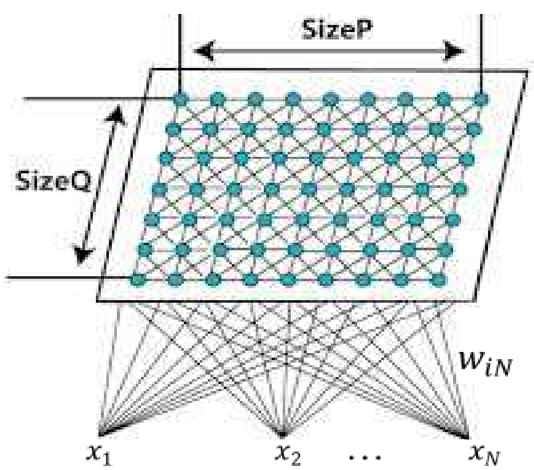
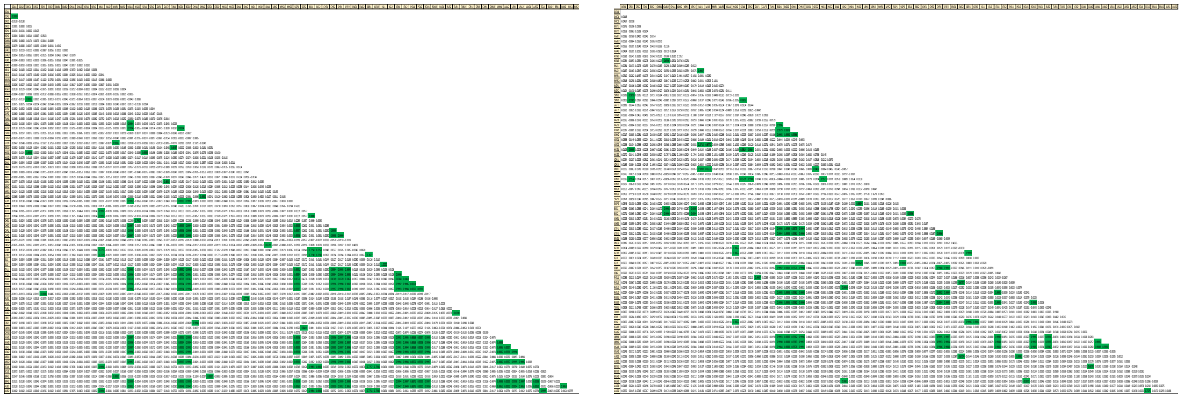


Fig. 4. SOM showing the map topology and interactions between nodes. Each neuron is assigned a vector of weights ($w = w_{i1}, w_{i2}, .. w_{iN}$) with dimension similar to the input vector i ($i = 1, 2, \ldots, L$); where L is the total number of neurons in the network. The input nodes have p features, 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.



(a) Male patients (b) Female patients

Fig. 5. Correlation matrices showing pairwise relatedness of the viral sub-strains' transmissions. Isolate pairs showing **v**iral sub-strains' transmissions ($r^2 \geq 0.60$) are colored in green.

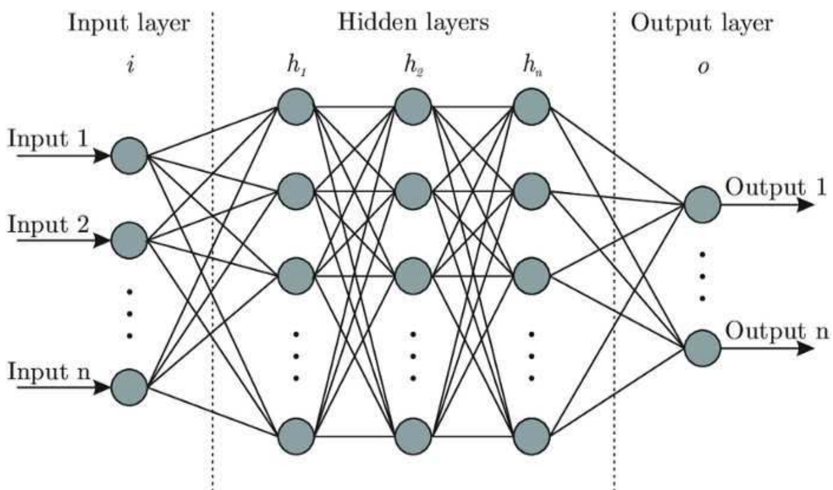


Fig. 6. ANN architecture. A 3-layered network, with one output layer and two hidden layers. The input layer consumes the knowledge-enriched genome datasets comprising of extracted patterns of SOM learning of the respective genome isolates and additional knowledge sieved from analysis of the genome sequences (i.e., number of natural clusters discovered from the genomic tree, discovered SOM sub-strain clusters, and link sequences derived from cognitive maps of the various isolates)

References

32. Inyang, U. G., Eyoh, I. J., Robinson, S. A., & Udo, E. N. Visual Association Analytics Approach to Predictive Modelling of Students' Academic Performance. *International Journal of Modern Education & Computer Science*, *11*(12), 1-13. (2019). https://doi.org/10.5815/ijmecs.2019.12.01
33. Inyang, U. G., Akpan, E. E., & Akinyokun, O. C. A Hybrid Machine Learning Approach for Flood Risk Assessment and Classification. *International Journal of Computational Intelligence and Applications*, *19*(2), 1-20. (2020). https://doi.org/10.1142/S1469026820500121
34. Inyang, U. G., & Joshua, E. E. Fuzzy clustering of students' data repository for at-risks students' identification and monitoring. Computer and Information Science, 6(4), 37-50. (2013). https://doi.org/doi:10.5539/cis.v6n4p37

35. Ekpenyong, M. E., & Inyang, U. G. Unsupervised mining of under-resourced speech corpora for tone features classification. In 2016 International Joint Conference on Neural Networks (IJCNN) (pp. 2374-2381). IEEE. (2016). https://doi.org/10.1109/IJCNN.2016.7727494
36. Abascal, F., Zardoya, R., & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic acids research, 38(suppl_2), W7-W13. (2010). https://doi.org/10.1093/nar/gkq291.
37. Kangas, J., Kohonen, T., & Laaksonen, J. Variants of self-organizing maps. IEEE transactions on neural networks, 1(1), 93-99. (1990). https://doi.org/10.1109/72.80208.
38. Vesanto, J. and Ahola, J. Hunting for Correlations in Data Using the Self-Organizing Map. Proceeding of the International ICSC Congress on Computational Intelligence Methods and Applications, pp. 279–285. (1999).
39. Rafiq, M. Y., Bugmann, G., & Easterbrook, D. J. Neural network design for engineering applications. Computers & Structures, 79(17), 1541-1552. (2001). https://doi.org/10.1016/S0045-7949(01)00039-6

Figures

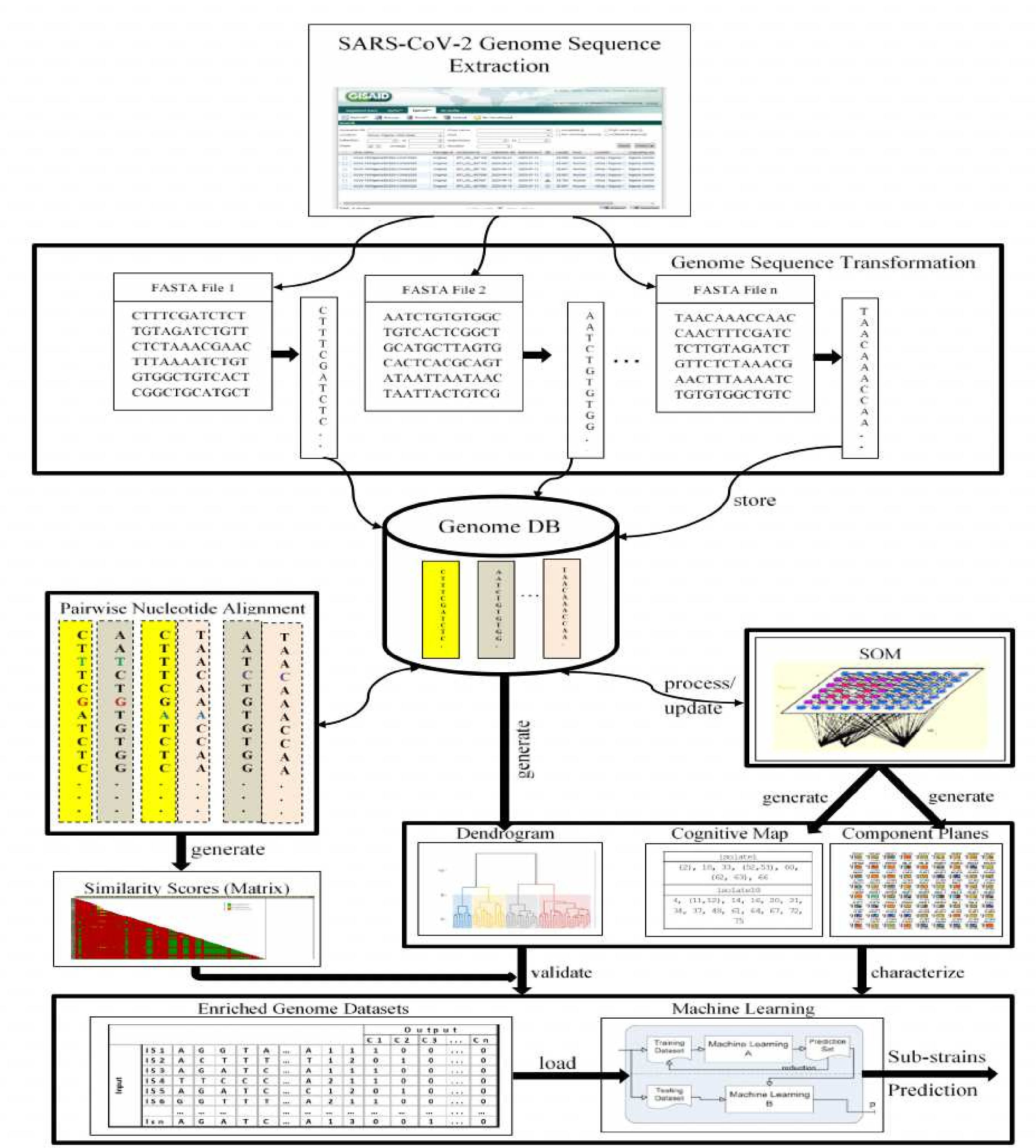
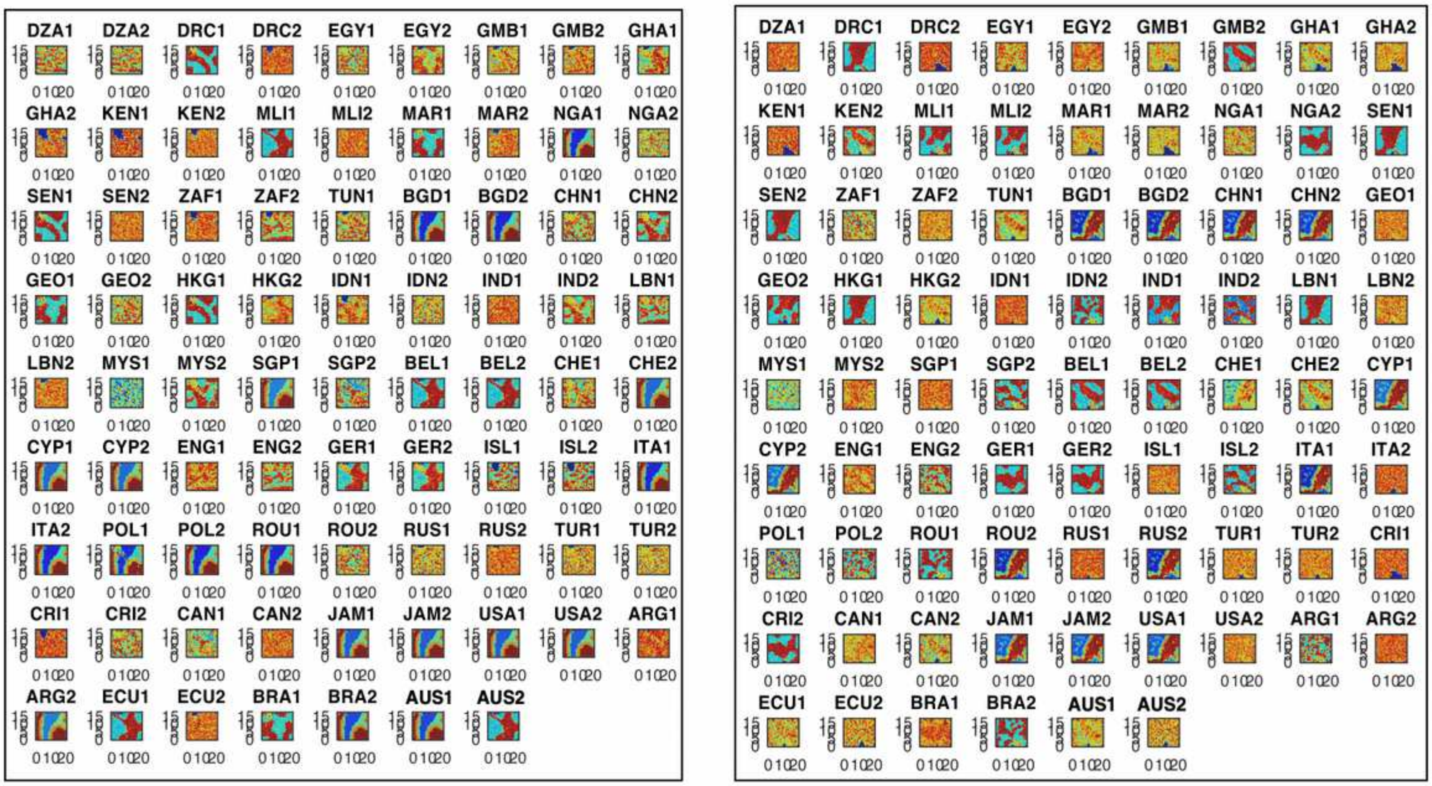


Figure 1

Workflow describing the proposed hybrid approach. The workflow begins with the excavation of FASTA files of human SARS-CoV-2 genome sequences from GISAID. These files were stripped and processed into a genome database (DB) as multiple columns of nucleotide sequence. A series of AI/ML techniques were applied to extract knowledge from the genome datasets as follows: Using ML techniques, 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/phylogenetic tree. Determine the optimal number of natural clusters–to provide additional knowledge for supervised learning. Separate the viral sub-strains using SOM component planes–for possible 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 prediction. Learn and predict new/emerging sub-strains using ANN.



(a) Male patients **(b) Female patients**

Figure 2

SOM component planes visualization. Maps are ordered by countries using Alpha-3 code (ISO 3166) representation, with at least 1 isolate per country. We observe single-, double- and multiple-source transmissions and discuss the transmission routes in a combinatorial manner (i.e., order is not a factor). In the African region, females show fewer country transmissions, as 7 out of 12 countries have similar pattern correlates with genomes from the South Asian; West Europe; Asian/Europe and Central American regions; compared to males, where 9 out of 12 countries show similar pattern correlates with genomes from the Southeast and West Asia; Southeast, South-Central and Western Europe; South and Central American; and the Oceanian regions. Furthermore, while male genomes show moderately weak pattern correlates with genomes from Germany, female genomes show very strong pattern correlates, and moderately weak pattern correlates with genomes from Belgium. Genome pattern correlates of males and females in the Asian region are consistent with patterns from Southeast and South-Central Europe;

Eastern Mediterranean; North and South American; save Asia/Europe and Oceanian regions, which genome pattern correlates are similar for Asian males and females, respectively. Genome pattern correlates of males and females in South America are consistent with patterns from the North American and the Oceania regions, while genome patterns in the North America are consistent with those from South American and the Oceanian regions.

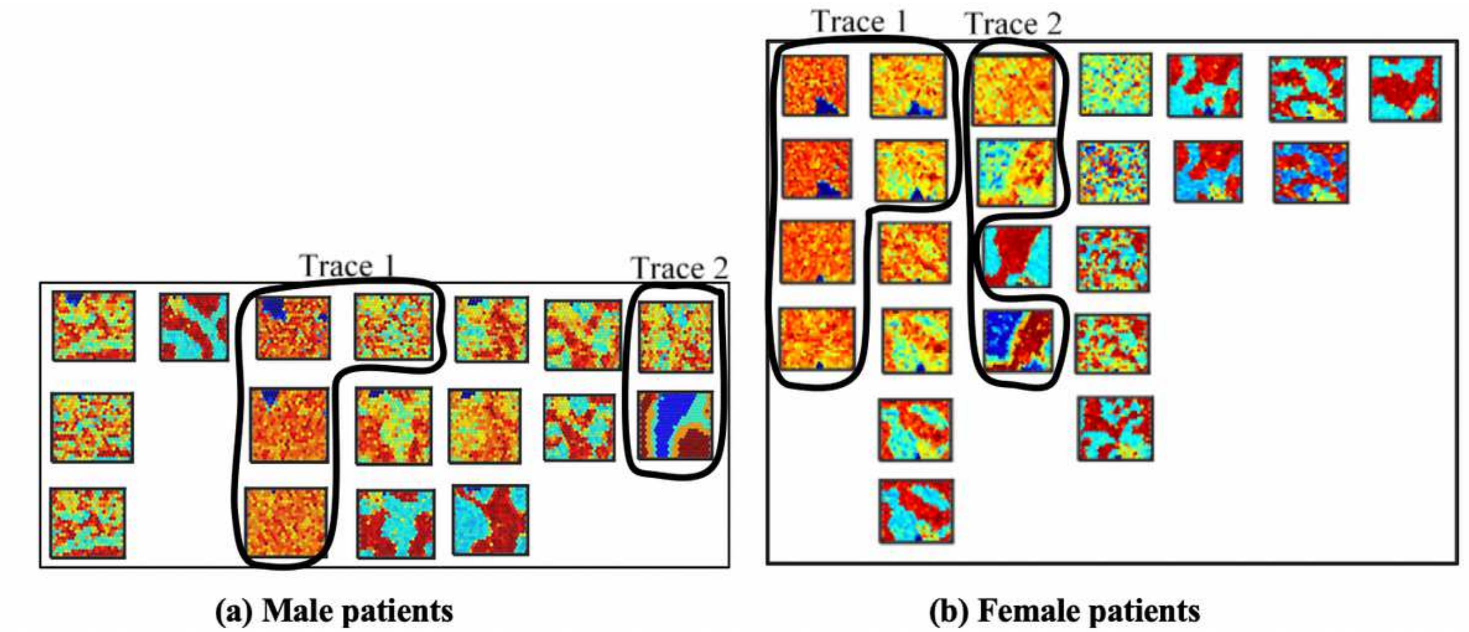
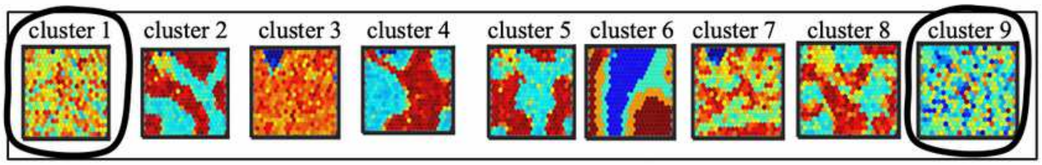


Figure 3

SARS-CoV-2 sub-strains clusters. We observe 7 distinct sub-strain clusters in both cases, with 4 of the sub-strains having more than 2 transitions before yielding sharp distinct clusters/patterns. While male SOM component planes exhibit rapid transitions unto sharp, distinct clusters (Fig. 3a), female SOM plots show slow transitions (Fig. 3b) and a late appearance of sharp, distinct clusters.



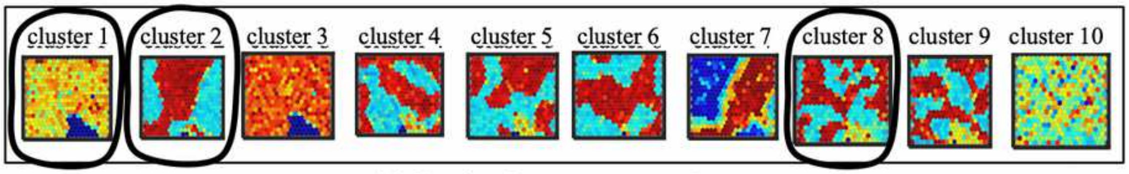
(a) Derived component planes

isolate1	isolate2	isolate3	isolate4	isolate5	isolate6	isolate7	isolate8	isolate9
(2), 18, 33, (52,53), 60, (62, 63), 66	(1), 18, 33, (52,53), 60, (62,63), 66	19, 30	10, (11, 12), 14, 16, 20, 21, 34, 37, 48, 61, 64, 67, 72, 75	(6), 15, 26, 28, 59, 65, 76	(5), 15, 26, 28, 59, 65, 76	(8), 9, 13, 23, 29, 31, (42,43), (50,51), 74, 79	(7), 9, 13, 23, 29, 31, (42,43), (50,51), 74, 79	(7,8), 13, 23, 29, 31, (42,43), (50,51), 74, 79
isolate10	isolate11	isolate12	isolate13	isolate14	isolate15	isolate16	isolate17	isolate18
4, (11,12), 14, 16, 20, 21, 34, 37, 48, 61, 64, 67, 72, 75	4, 10, 12, 14, 16, 20, 21, 34, 37, 48, 61, 64, 67, 72, 75	4, 10, 11, 14, 16, 20, 21, 34, 37, 48, 61, 64, 67, 72, 75	(7,8), 9, 23, 29, 31, (42,43), (50,51), 74, 79	4, 10, (11,12), 16, 20, 21, 34, 37, 48, 61, 64, 67, 72, 75	(5,6), 26, 28, 59, 65, 76	4, 10, (11,12), 20, 21, 34, 37, 48, 61, 64, 67, 72, 75	(24,25), 40, 45, (46,47), (54,55), (56,57), 58, (68,69), (70,71), 73, (77,78)	(1,2), 33, (52,53), 60, (62,63), 66
isolate19	isolate20	isolate21	isolate22	isolate23	isolate24	isolate25	isolate26	isolate27
3, 30	4, 10, (11,12), 14, 16, 21, 34, 37, 48, 61, 64, 67, 72, 75	4, 10, (11,12), 14, 16, 20, 34, 37, 48, 61, 64, 67, 72, 75	36, 49	(7,8), 9, 13, 29, 31, (42,43), (50,51), 74, 79	17, 25, 40, 45, (46,47), (54,55), (56,57), 58, (68,69), (70,71), 73, (77,78)	17, 24, 40, 45, (46,47), (54,55), (56,57), 58, (68,69), (70,71), 73,	(5,6), 15, 28, 59, 65, 76	32, 35, 39, 44
isolate28	isolate29	isolate30	isolate31	isolate32	isolate33	isolate34	isolate35	isolate36
(5,6), 15, 26, 59, 65, 76	(7,8), 9, 13, 23, 31, (42,43), (50,51), 74, 79	3, 19	(7,8), 9, 13, 23, 29, (42,43), (50,51), 74, 79	27, 35, 39, 44	(1,2), 18, (52,53), 60, (62,63), 66	4, 10, (11,12), 14, 16, 20, 21, 37, 48, 61, 64, 67, 72,	27, 32, 39, 44	22, 49
isolate37	isolate38	isolate39	isolate40	isolate41	isolate42	isolate43	isolate44	isolate45
4, 10, (11,12), 14, 16, 20, 21, 34, 48, 61, 64, 67, 72, 75	41, 66	27, 32, 35	17, (24,25), 45, (46,47), (54,55), (56,57), 58, (68,69), (70,71), 73, (77,78)	38, 66	(7,8), 9, 13, 23, 29, 31, (43), (50,51), 74, 79	(7,8), 9, 13, 23, 29, 31, (42), (50,51), 74, 79	27, 32, 35, 39	17, (24,25), 40, (46,47), (54,55), (56,57), 58, (68,69), (70,71), 73, (77,78)
isolate46	isolate47	isolate48	isolate49	isolate50	isolate51	isolate52	isolate53	isolate54
17, (24,25), 40, 45, 47, (54,55), (56,57), 58, (68,69), (70,71), 73, (77,78)	17, (24,25), 40, 45, 46, (54,55), (56,57), 58, (68,69), (70,71), 73, (77,78)	4, 10, (11,12), 14, 16, 20, 21, 34, 37, 61, 64, 67, 72, 75	22, 36, 49	9, 13, 23, 29, 31, (42,43), 51, 74, 79	9, 13, 23, 29, 31, (42,43), 50, 74, 79	(1,2), 18, 33, 53, 60, (62,63), 66	(1,2), 18, 33, 52, 60, (62,63), 66	17, (24,25), 40, 45, 46, (54,55), (56,57), 58, (68,69), (70,71), 73, (77,78)
isolate55	isolate56	isolate57	isolate58	isolate59	isolate60	isolate61	isolate62	isolate63
17, (24,25), 40, 45, (46,47), 54, (56,57), 58, (68,69), (70,71), 73, (77,78)	17, (24,25), 40, 45, (46,47), (54,55), 57, 58, (68,69), (70,71), 73, (77,78)	17, (24,25), 40, 45, (46,47), (54,55), 56, 58, (68,69), (70,71), 73, (77,78)	17, (24,25), 40, 45, (46,47), (54,55), (56,57), (68,69), (70,71), 73, (77,78)	(5,6), 15, 26, 28, 65, 76	(1,2), 18, 33, (52,53), (62,63), 66	4, 10, (11,12), 14, 16, 20, 21, 34, 37, 48, 67, 72, 75	(1,2), 18, 33, (52,53), 60, (63), 66	(1,2), 18, 33, (52,53), 60, (62), 66
isolate64	isolate65	isolate66	isolate67	isolate68	isolate69	isolate70	isolate71	isolate72
4, 10, (11,12), 14, 16, 20, 21, 34, 37, 48, 61, 67, 72, 75	(5,6), 15, 26, 28, 59, 76	(1,2), 18, 33, (52,53), 60, (62, 63)	4, 10, (11,12), 14, 16, 20, 21, 34, 37, 48, 61, 64, 72, 75	17, (24,25), 40, 45, (46,47), (54,55), (56,57), 58, 69, (70,71), 73, (77,78)	17, (24,25), 40, 45, (46,47), (54,55), (56,57), 58, 68, (70,71), 73, (77,78)	17, (24,25), 40, 45, (46,47), (54,55), (56,57), 58, (68,69), 71, 73,	17, (24,25), 40, 45, (46,47), (54,55), (56,57), 58, (68,69), 70, 73, (77,78)	4, 10, (11,12), 14, 16, 20, 21, 34, 37, 48, 61, 64, 67, 75
isolate73	isolate74	isolate75	isolate76	isolate77	isolate78	isolate79		
17, (24,25), 40, 45, (46,47), (54,55), (56,57), 58, (68,69), (70,71), (77,78)	(7,8), 9, 13, 23, 29, 31, (42,43), (50,51), 79	4, 10, (11,12), 14, 16, 20, 21, 34, 37, 48, 61, 64, 67, 72	(5,6), 15, 26, 28, 59, 65	17, (24,25), 40, 45, (46,47), (54,55), (56,57), 58, (68,69), (70,71), 73, 78	17, (24,25), 40, 45, (46,47), (54,55), (56,57), 58, (68,69), (70,71), 73, 77	(7,8), 9, 13, 23, 29, 31, (42,43), (50,51), 74		

(b) Cognitive map

Figure 4

Derived SOM component planes and cognitive map for male genome isolates. The map provides cognitive solution for sub-strains link modeling and reveal hidden structures which hitherto were subsumed in previously large cluster groups (Fig. 3a), hence, corroborating the possibility of confusable clusters22.



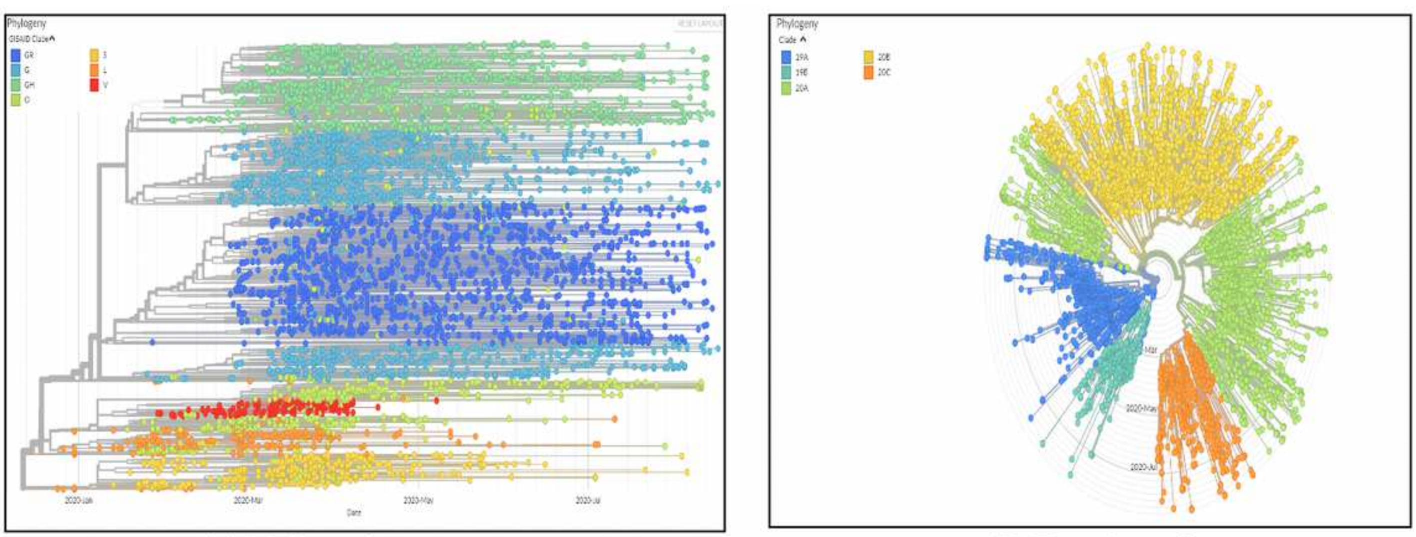
(a) Derived component planes

isolate1	isolate2	isolate3	isolate4	isolate5	isolate6	isolate7	isolate8	isolate9
3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	(18,19), 29, 35	1, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	(14,15), 16, 20, 65, (73,74), (77,78)	1, 3, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	(7), 8, 11, 22, 30, (41,42), 44, 66	(6), 8, 11, 22, 30, (41,42), 44, 66	(6,7), 11, 22, 30, (41,42), 44, 66	1, 3, 5, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75
isolate10	isolate11	isolate12	isolate13	isolate14	isolate15	isolate16	isolate17	isolate18
1, 3, 5, 9, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	(6, 7), 8, 22, 30, (41,42), 44, 66	(13), 28, 33	(12), 28, 33	4, (15), 16, 20, 65, (73,74), (77,78)	4, (14), 16, 20, 65, (73,74), (77,78)	4, (14,15), 20, 65, (73,74), (77,78)	(49,50), 64	2, (19), 29, 35
isolate19	isolate20	isolate21	isolate22	isolate23	isolate24	isolate25	isolate26	isolate27
2, (18), 29, 35	4, (14,15), 16, 20, 65, (73,74), (77,78)	1, 3, 5, 9, 10, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	(6,7), 8, 11, 30, (41,42), 44, 66	(24), (25,26), 43, (45,46), 53, 58, 60, (67,68), 69	(23,24), (25,26), 43, (45,46), 53, 58, 60, (67,68), 69	(23,24), (26), 43, (45,46), 53, 58, 60, (67,68), 69	(23,24), (25), 43, (45,46), 53, 58, 60, (67,68), 69	1, 3, 5, 9, 10, 21, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75
isolate28	isolate29	isolate30	isolate31	isolate32	isolate33	isolate34	isolate35	isolate36
(12,13), 33	2, (18,19), 35	(6,7), 8, 11, 22, (41,42), 44, 66	1, 3, 5, 9, 10, 21, 27, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	57, 71, 76	(12, 13), 28, 33	40, 48, 52, 56	2, (18,19), 29	1, 3, 5, 9, 10, 21, 27, 31, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75
isolate37	isolate38	isolate39	isolate40	isolate41	isolate42	isolate43	isolate44	isolate45
55	1, 3, 5, 9, 10, 21, 27, 31, 36, 39, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 47, 51, 54, 59, (61,62), 63, 70, 72, 75	34, 48, 52, 56	(6,7), 8, 11, 22, 30, (42), 44, 66	(6,7), 8, 11, 22, 30, (41), 44, 66	(23,24), (25,26), (45,46), 53, 58, 60, (67,68), 69	(6,7), 8, 11, 22, 30, (41,42), 66	(23,24), (25,26), 43, (46), 53, 58, 60, (67,68), 69
isolate46	isolate47	isolate48	isolate49	isolate50	isolate51	isolate52	isolate53	isolate54
(23,24), (25,26), 43, (45), 53, 58, 60, (67,68), 69	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 51, 54, 59, (61,62), 63, 70, 72, 75	34, 40, 52, 56	17, (50), 64	17, (49), 64	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 54, 59, (61,62), 63, 70, 72, 75	34, 40, 48, 56	(23, 24), (25,26), 43, (45,46), 58, 60, (67,68), 69	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 59, (61,62), 63, 70, 72, 75
isolate55	isolate56	isolate57	isolate58	isolate59	isolate60	isolate61	isolate62	isolate63
37	34, 40, 48, 52	32, 71, 76	(23,24), (25,26), 43, (45,46), 53, 60, (67,68), 69	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, (61,62), 63, 70, 72, 75	(23,24), (25,26), 43, (45,46), 53, 58, (67,68), 69	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (62), 63, 70, 72, 75	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61), 63, 70, 72, 75	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 70, 72, 75
isolate64	isolate65	isolate66	isolate67	isolate68	isolate69	isolate70	isolate71	isolate72
17, (49,50)	4, (14,15), 16, 20, (73,74), (77,78)	(6,7), 8, 11, 22, 30, (41,42), 44	(23,24), (25,26), 43, (45,46), 53, 58, 60, (68), 69	(23,24), (25,26), 43, (45,46), 53, 58, 60, (67), 69	(23,24), (25,26), 43, (45,46), 53, 58, 60, (67,68)	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 72, 75	32, 57, 76	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 75
isolate73	isolate74	isolate75	isolate76	isolate77	isolate78			
4, (14,15), 16, 20, 65, (74), (77,78)	4, (14,15), 16, 20, 65, (73), (77,78)	1, 3, 5, 9, 10, 21, 27, 31, 36, 38, 39, 47, 51, 54, 59, (61,62), 63, 70, 72	32, 57, 71	4, (14,15), 16, 20, 65, (73,74), (78)	4, (14,15), 16, 20, 65, (73,74), (77)			

(b) Cognitive map

Figure 5

Derived SOM component planes and cognitive map and for female genome isolates. The map provides cognitive solution for sub-strains link modeling and reveal hidden structures which hitherto were subsumed in previously large cluster groups (Fig. 3b), hence, corroborating the possibility of confusable clusters22.



(a) Male patients **(b) Female patients**

Figure 6

Methods Fig. 1. Phylogeny of SARS-CoV-2 genomic epidemiology. The original L strain surfaced in Wuhan in December 2019, with its first mutation, the S strain, appearing at the beginning of 2020; while strains V and G appeared in the mid of January 2020. Currently, strain G which mutated into strains GR and GH at the end of February 2020 are by far most prevalent, presenting 4 different mutations, two of which can change the RNA polymerase and Spike proteins sequence of the virus–a probable characteristic that facilitates the virus spread. Globally, strains G, GH and GR are becoming more dominant, while strain S, L and V are fast disappearing.

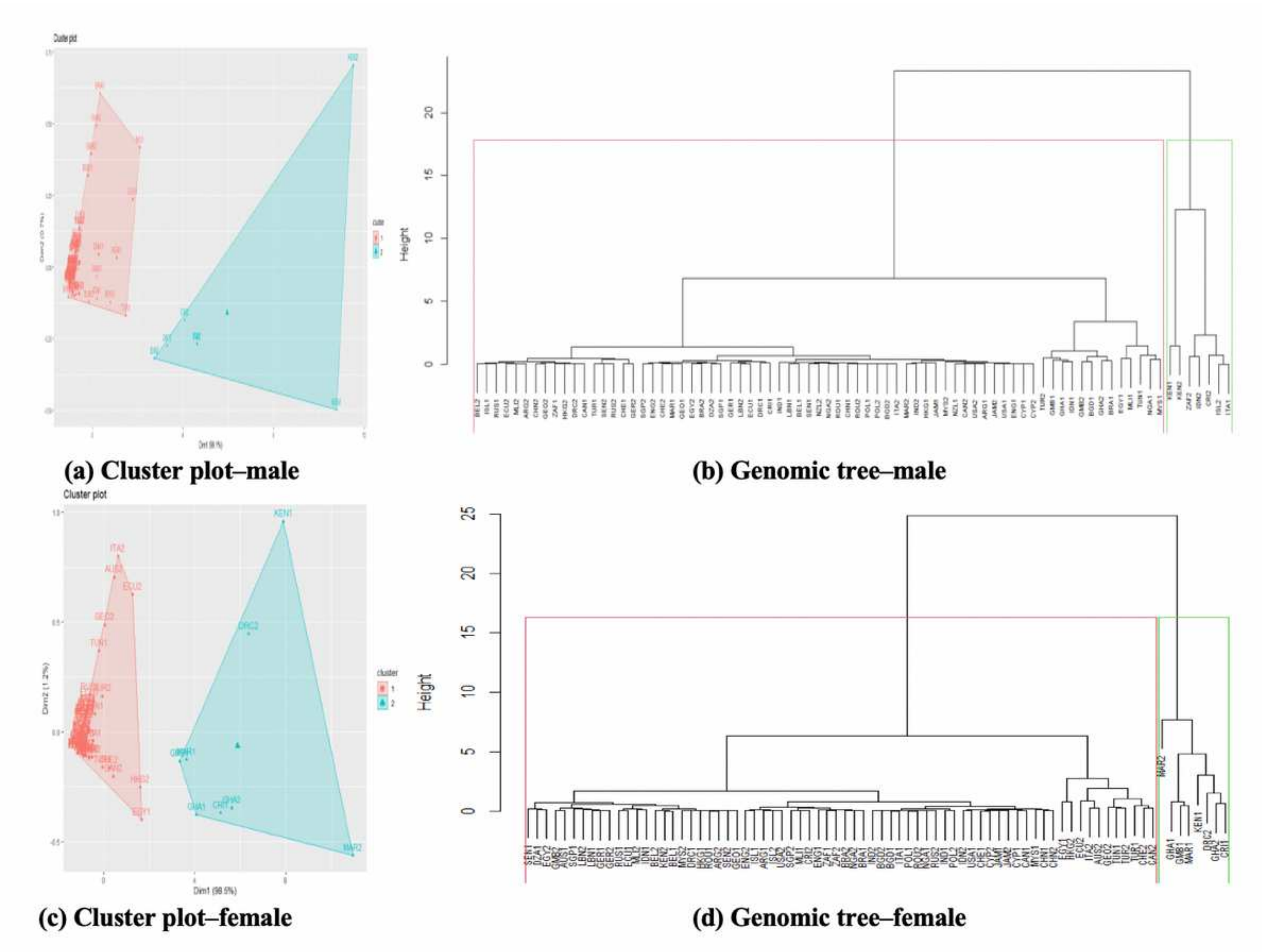


Figure 7

Methods Fig 2. Cluster plots and genomic trees. Notice 2 distinct groups A and B separated between closely similar and dissimilar isolates, with the A group having heavy isolates concentration compared to the B group. For males (Fig. 2b), group A consists of 72 isolates with 3 sub-groups. The first sub-group contains 59 isolates from Europe, Africa, Asia, South America, North America and Oceania. The second sub-group contain 10 isolates from Europe, Africa and Asia. The third sub-group contains 3 isolates from Africa. Cluster B consists of 7 isolates distributed among 4 sub-groups namely Africa (2); Africa and Asia

(2); Europe (1); and Europe and North America. For females (Fig. 2d), cluster A consists of 70 isolates with 3 sub-groups. The first sub-group contains 59 isolates from Europe, Africa, Asia, South America, North America and Oceania. The second sub-group contain 2 isolates from Africa and Asia. The third sub-group has 9 isolates distributed between North America, South America, Africa, Asia, Europe and Oceania. Cluster B consists of 8 isolates distributed among 4 sub-groups with the first 3 isolates groups from the African region, while the fourth isolate group comes from Africa and Europe regions.

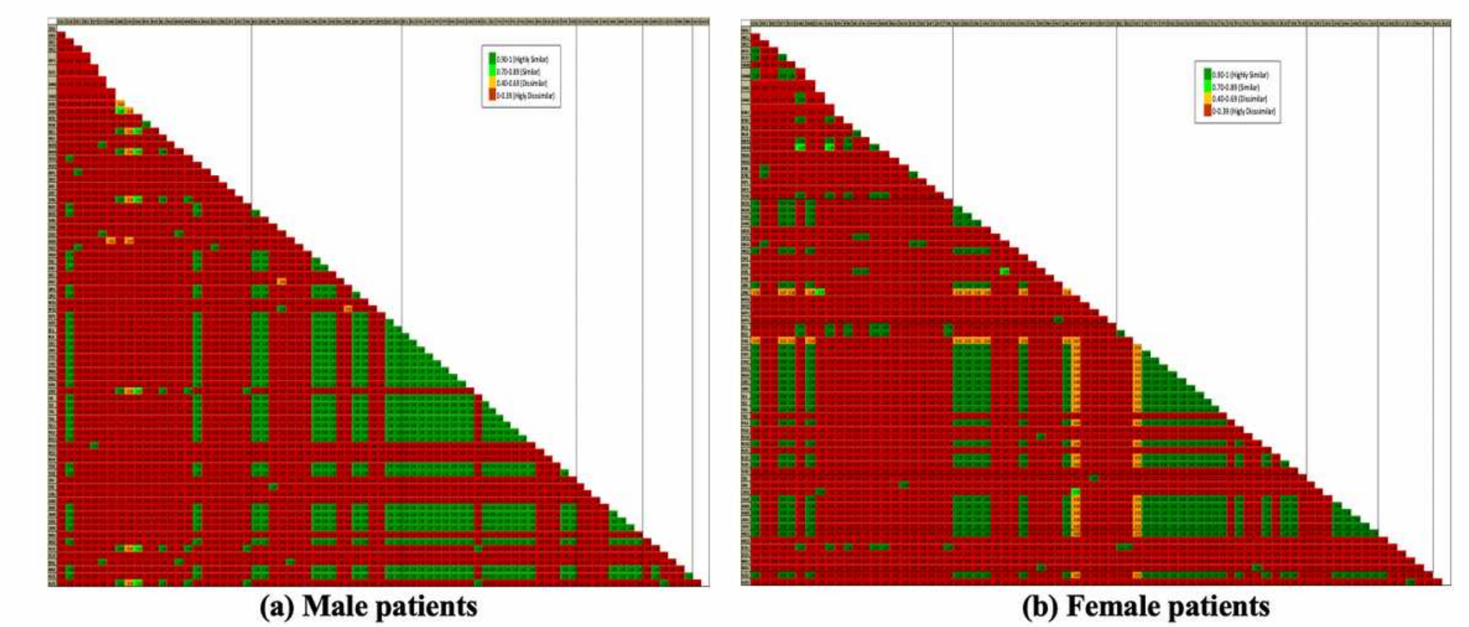


Figure 8

Methods Fig. 3. Inter-continent nucleotide similarity matrices. Green colored cells are regions of high similarity that may indicate functional, structural and/or evolutionary relationships between nucleotide sequences.

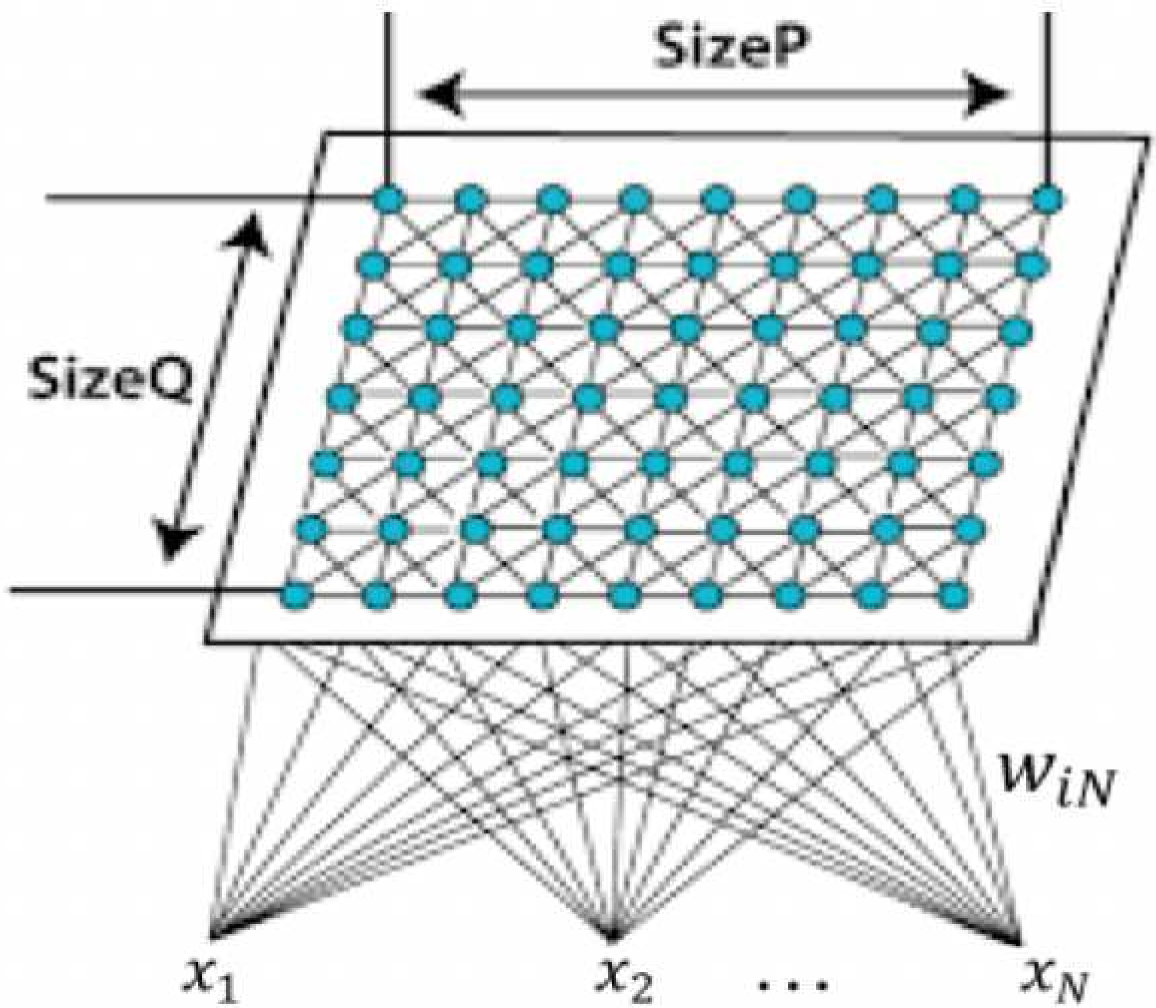


Figure 9

Methods Fig. 4. 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.

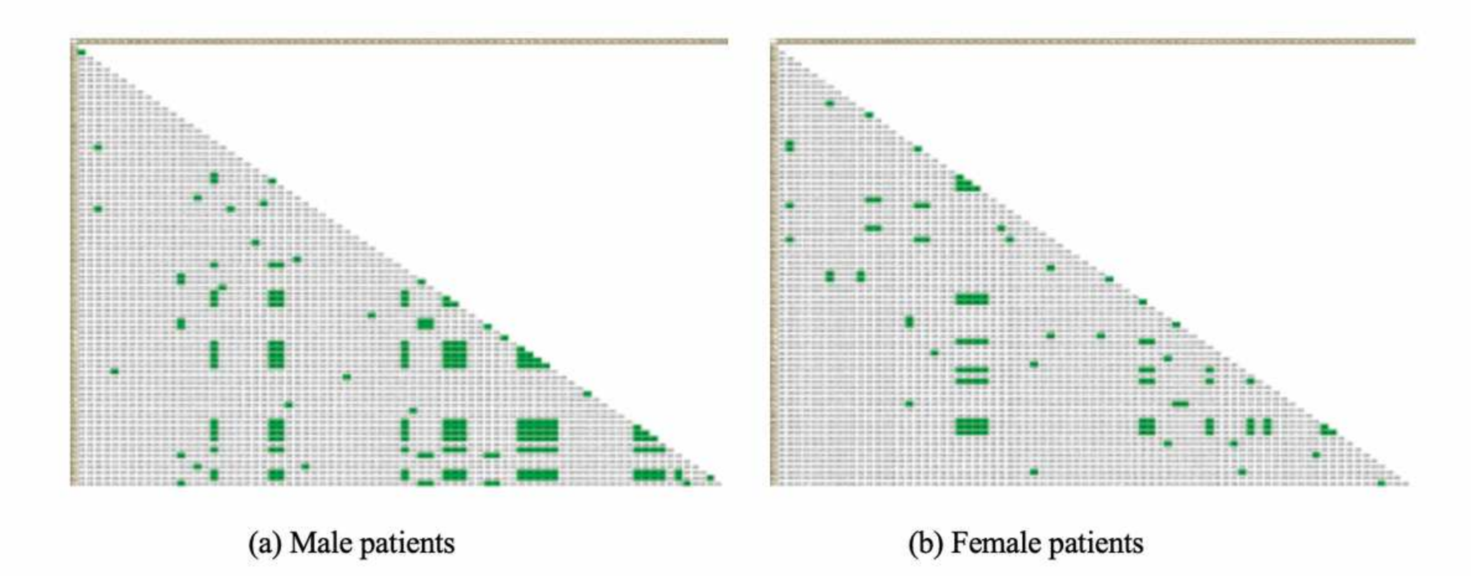


Figure 10

Methods Fig. 5. Correlation matrices showing pairwise relatedness of the viral sub-strains' transmissions. Isolate pairs showing viral sub-strains' transmissions (r^2≥0.60) are colored in green.

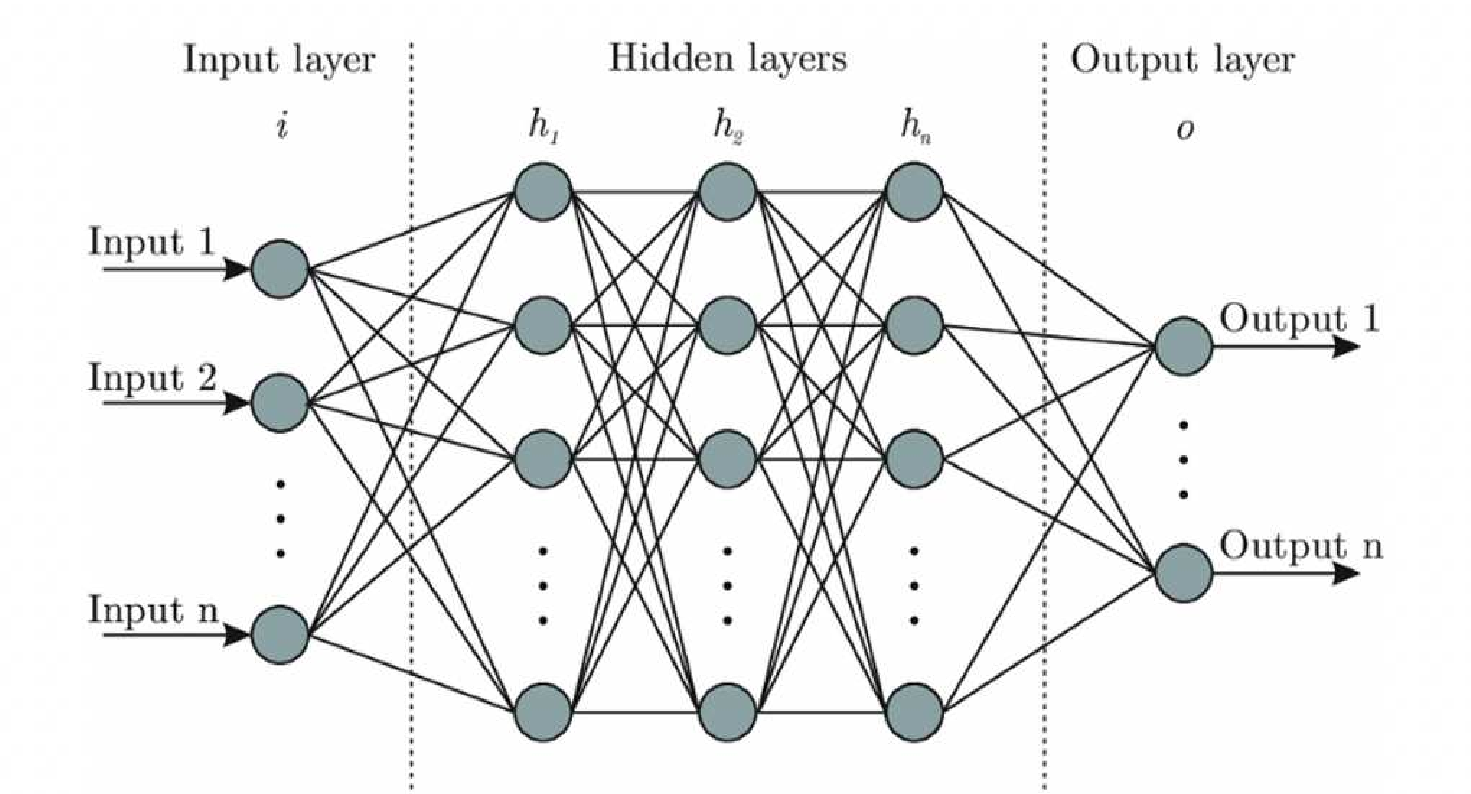


Figure 11

Methods Fig. 6. ANN architecture. A 3-layered network, with one output layer and two hidden layers. The input layer consumes the knowledge-enriched genome datasets comprising of extracted patterns of SOM learning of the respective genome isolates and additional knowledge sieved from analysis of the genome sequences (i.e., number of natural clusters discovered from the genomic tree, discovered SOM sub-strain clusters, and link sequences derived from cognitive maps of the various isolates)

Supplementary Files

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

- SupplData1.xlsx
- SupplData2.xlsx
- SupplData3.xlsx
- SupplData4.xlsx