Virus Research 298 (2021) 198401

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Genome-wide analysis of 10664 SARS-CoV-2 genomes to identify virus strains in 73 countries based on single nucleotide polymorphism

Nimisha Ghosh [a,1], Indrajit Saha [b,1,*], Nikhil Sharma [c,1], Suman Nandi [b], Dariusz Plewczynski [d,e]

[a] *Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India*
[b] *Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India*
[c] *Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India*
[d] *Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland*
[e] *Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Warsaw, Poland*

ARTICLE INFO

Keywords:
Clustering
COVID-19
Multiple sequence alignment
Non-synonymous SNP
SARS-CoV-2

ABSTRACT

Since the onslaught of SARS-CoV-2, the research community has been searching for a vaccine to fight against this virus. However, during this period, the virus has mutated to adapt to the different environmental conditions in the world and made the task of vaccine design more challenging. In this situation, the identification of virus strains is very much timely and important task. We have performed genome-wide analysis of 10664 SARS-CoV-2 genomes of 73 countries to identify and prepare a Single Nucleotide Polymorphism (SNP) dataset of SARS-CoV-2. Thereafter, with the use of this SNP data, the advantage of hierarchical clustering is taken care of in such a way so that Average Linkage and Complete Linkage with Jaccard and Hamming distance functions are applied separately in order to identify the virus strains as clusters present in the SNP data. In this regard, the consensus of both the clustering results are also considered while Silhouette index is used as a cluster validity index to measure the goodness of the clusters as well to determine the number of clusters or virus strains. As a result, we have identified five major clusters or virus strains present worldwide. Apart from quantitative measures, these clusters are also visualized using Visual Assessment of Tendency (VAT) plot. The evolution of these clusters are also shown. Furthermore, top 10 signature SNPs are identified in each cluster and the non-synonymous signature SNPs are visualised in the respective protein structures. Also, the sequence and structural homology-based prediction along with the protein structural stability of these non-synonymous signature SNPs are reported in order to judge the characteristics of the identified clusters. As a consequence, T85I, Q57H and R203M in NSP2, ORF3a and Nucleocapsid respectively are found to be responsible for Cluster 1 as they are damaging and unstable non-synonymous signature SNPs. Similarly, F506L and S507C in Exon are responsible for both Clusters 3 and 4 while Clusters 2 and 5 do not exhibit such behaviour due to the absence of any non-synonymous signature SNPs. In addition to all these, the code, SNP dataset, 10664 labelled SARS-CoV-2 strains and additional results as supplementary are provided through our website for further use.

1. Introduction

SARS-CoV-2 is the causal agent for current ongoing outbreak of disease commonly known as COVID-19 (Zhou et al., 2020) which has proven to have a detrimental effect on the humankind. As a result, medical emergencies have surged and a halt of economic growth has occurred around the globe due to an eccentric impact of SARS-CoV-2. SARS-CoV-2 belongs to the family of Coronaviridae which also houses SARS-CoV-1 and MERS-CoV (van Dorp et al., 2020). First case of Severe Acute Respiratory Syndrome (SARS) was registered way back in 2002-03, which took around 8000 lives.[2] Another pathogenic invasion was reported in 2012, named as Middle East Respiratory Syndrome Coronavirus (MERS-CoV) with a worldwide mortality rate of 35.5% (Ahmed, 2017). However, these two viruses have a significantly low

* Corresponding author.
E-mail address: indrajit@nitttrkol.ac.in (I. Saha).

[1] Equally contributed.

[2] https://en.wikipedia.org/wiki/2002%E2%80%932004_SARS_outbreak.

https://doi.org/10.1016/j.virusres.2021.198401
Received 8 December 2020; Received in revised form 23 February 2021; Accepted 16 March 2021
Available online 26 March 2021

transmission rate among the masses as compared to SARS-CoV-2 (Petersen et al., 2020). On the other hand, SARS-CoV-2 which emerged in late 2019 in China has now spread across the globe. On 11th March 2020, World Health Organization (WHO) declared COVID-19 as a global pandemic due to its high transmission rate and adverse effect on health care systems resulting in 112.2 million affected and 2485k deaths till now (Worldometer, 2021). In this current havoc situation, many organizations such as The University of Oxford (Folegatti et al., 2020), Bharat BioTech, Beijing Institute of Biological Products are witnessing their ongoing trials of designed vaccines respectively to curb the impact of this contagious virus. In fact, in India indigenously produced Covaxin by Bharat BioTech and Covishield (the local name for the Oxford-AstraZeneca vaccine developed in the UK) are already being disseminated among the masses.

Vaccine development is a key component in the prevention of disease spread and reduction in the morbidity and mortality associated with the diseases by evoking an immune response in form of antigens against regions of proteins which are critical for pathogen binding (Amela et al., 2007). Since the emergence of SARS-CoV-2 in December 2019, many mutations such as substitutions and deletions in its coding and non-coding regions have been discovered (Phan, 2020) till date. Genome-wide analysis of 566 Indian SARS-CoV-2 genomes (Saha et al., 2020) revealed numerous mutation points as substitutions, deletions, insertions and single nucleotide polymorphism popularly known as SNP. Among all types of mutations, SNP (Nogales and Dieg, 2019; Yin, 2020) can be considered to be the source of variance in virus genomes giving a way for re-emergence, drug resistance and antibody escape for pathogens. SNPs were used in (Yang et al., 2020) to classify four super spreader (SS) clusters according to relative variants. According to their study, SS1 was widely spread in Asia and US, while SS4 was responsible for the pandemic in Europe. Hence, SNPs play a vital role in tracking virus strains and variations. Analysis of SNP was carried out in (Lo et al., 2018) as well where they suggested a change in the samples of Plasmodium falciparum between the northern and southern regions of Western Kenya; the samples from the southern part showed lesser divergence from each other. Furthermore, clustering has been used by many studies (Sevilla-Reyes et al., 2013; Fischer et al., 2018; Hahn et al., 2020) to understand genetic characteristics for a large set of genomes. To detect genetic diversity of non-structural (NS1) protein of influenza A virus, (Sevilla-Reyes et al., 2013) used clustering on the available protein sequence data and combined it with maximum likelihood phylogenetic RNA reconstruction and consensus WebLogo comparison. In (Fischer et al., 2018), Fischer et al. have used affinity propagation clustering to define clusters for rabies virus (RABV). Most recently, unsupervised cluster analysis of SARS-CoV-2 genomes have been carried out in (Hahn et al., 2020) by using principal component analysis to a similarity matrix to compare all pairs of 2540 nucleotides using Jaccard index. The analysed results were used to illustrate the geographic progression of the virus. Apart from this, for clustering of SARS-CoV-2 genomes, online web servers like GISAID CoVsurver[3] and Pangolin[4] exist. However, these servers take a substantial amount of time while performing clustering in order to generate phylogenetic tree as they consider all substitutions like mutation. This drawback has motivated us to perform the underlying clustering task faster and accurately on smaller and relevant features like SNP in order to identify SARS-CoV-2 virus strains of 10664 SARS-CoV-2 genomes.

To address the clustering task, we have performed genome-wide analysis of 10664 SARS-CoV-2 genomes from 73 countries to determine the SNPs. SNPs represent mutation as substitution that occurs in more than 1% of the virus population for a given genomic position. In this regard, 107 SNPs are identified throughout the genome which are used to prepare a binary dataset. In order to identify the virus strains from this binary dataset, careful application of clustering methods with proper distance functions is very crucial. Thus, from our previous clustering experience, the use of hierarchical clustering such as Average Linkage and Complete Linkage with Jaccard and Hamming distances are appropriate in this context, while for computing the goodness of the clustering, the Silhouette index can be used with the same distance functions. Moreover, the consensus of the clustering results can be an added boost for the identification of the proper number of clusters as virus strains. To the best of our knowledge, this approach has not yet been applied for the identification of SARS-CoV-2 virus strains. This approach results in five major clusters as virus strains. Moreover, their presence in different countries are identified along with the evolution of the virus genomes from January to July 2020 for each of the 73 countries. These outcomes are shown quantitatively and visually through BioCircos (Cui et al., 2016), Visual Assessment of Tendency (VAT) plot (Bezdek and Hathaway, 2002; Kumar and Bezdek, 2020) and Heatmap (Deng et al., 2014). Therefore, the major contributions of this work can be summarised as: SNP identification from 10664 SARS-CoV-2 genomes of 73 countries, binary dataset creation from SNP data to find the number of clusters as virus strains present in 73 countries around the globe and their evolution, identifying signature SNPs in each cluster and determining the structural stability of the non-synonymous signature SNPs to judge the characteristics of the identified clusters.

2. Materials and methods

In this section, collection of SARS-CoV-2 genomes, Visual Assessment of Tendency (VAT) plot and the proposed pipeline with the preparation of SNPs data are discussed. For the ease of understanding for the readers, hierarchical clustering (Tou and Gonzalez, 1974; Devijver and Kittler, 1982) such as Average Linkage (Tou and Gonzalez, 1974; Devijver and Kittler, 1982) and Complete Linkage (Tou and Gonzalez, 1974) with Jaccard (Tou and Gonzalez, 1974; Devijver and Kittler, 1982) and Hamming distance (Tou and Gonzalez, 1974; Devijver and Kittler, 1982) functions, Silhouette index (Rousseeuw, 1987) as cluster validity measure are briefly discussed in supplementary.

2.1. *Collection of SARS-CoV-2 genomes*

Initially, 10664 complete and near complete SARS-CoV-2 genomes were collected from Global Initiative on Sharing All Influenza Data (GISAID)[5] in fasta format while the Reference Genome (NC_045512.2)[6] was collected from National Center for Biotechnology Information (NCBI). These genomic sequences are distributed in 73 countries starting from January till July 2020. This is important to note that GISAID contains many incomplete sequences or virus genomes which we have removed while preparing our sequence dataset. The dataset contains sequence ID and virus genome as a fasta format. The maximum and average length of the 10664 virus genomes are 29,903 and 29,821 bp respectively. Please note that the maximum length has been considered by taking the reference sequence. These 10664 SARS-CoV-2 sequences are aligned using multiple sequencing alignment (MSA) to prepare the SNP dataset. Please note that for the data visualization and editing BioEdit was used. For the alignment of sequences High Performance Computing facility of NITTTR, Kolkata was used and for the identification and preparation of SNP dataset MATLAB R2019b was used.

2.2. *Visual Assessment of Tendency*

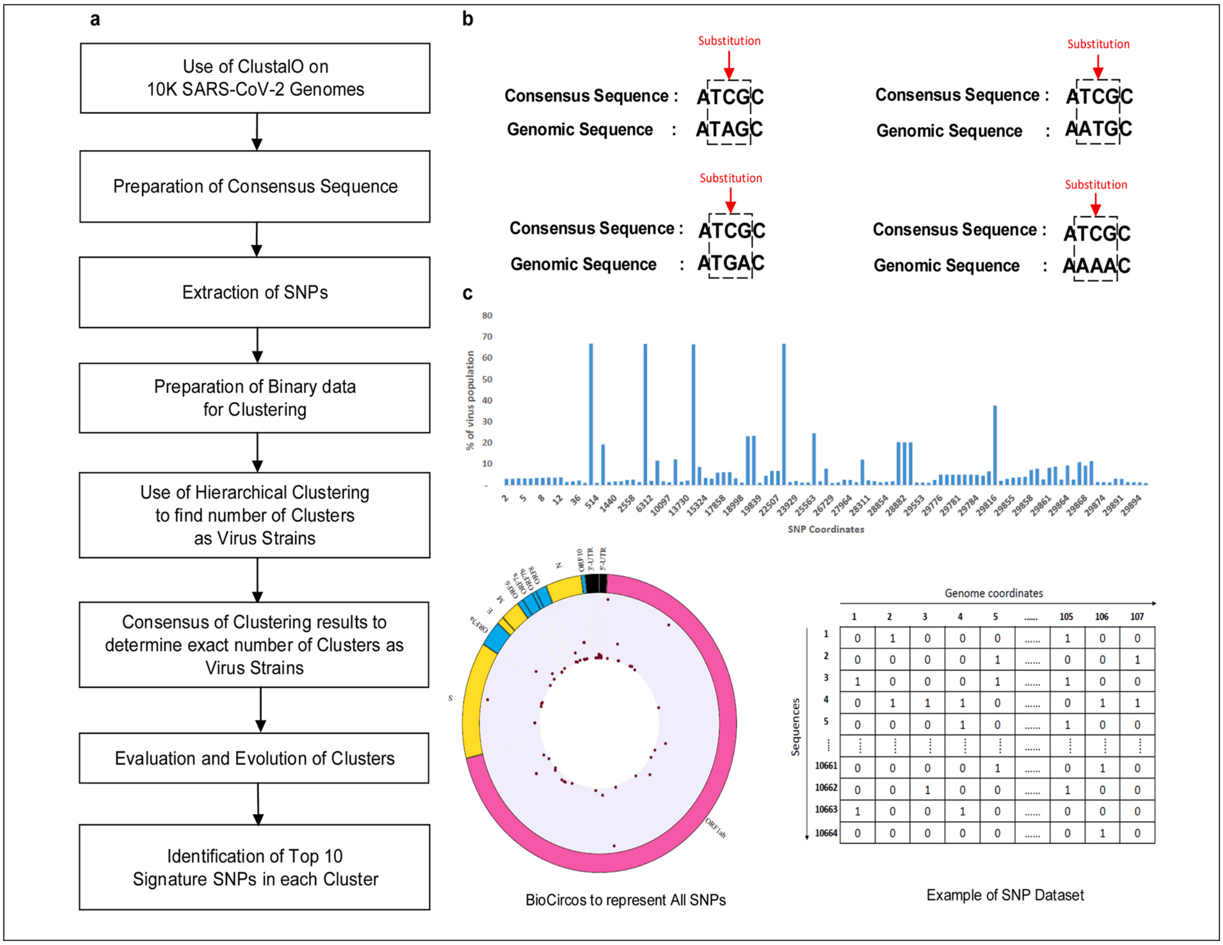
The well-known Visual Assessment of Tendency (VAT) (Bezdek and Hathaway, 2002; Kumar and Bezdek, 2020) representation is used here to visualize the clusters formed by the clustering methods. This

[3] https://www.gisaid.org/epiflu-applications/covsurver-mutations-app/.

[4] https://cov-lineages.org/index.html.

[5] https://www.gisaid.org/.

[6] https://www.ncbi.nlm.nih.gov/nuccore/1798174254.



	1	2	3	4	5	......	105	106	107
1	0	1	0	0	0	......	1	0	0
2	0	0	0	0	1	......	0	0	1
3	1	0	0	0	1	......	1	0	0
4	0	1	1	1	0	......	0	1	1
5	0	0	0	1	0	......	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	......	⋮	⋮	⋮
10661	0	0	0	0	1	......	0	1	0
10662	0	0	1	0	0	......	1	0	0
10663	1	0	0	1	0	......	0	0	0
10664	0	0	0	0	0	......	0	1	0

Example of SNP Dataset

Fig. 1. (a) Pipeline of the Workflow, (b) Detection technique to find mutation as substitution, (c) Bar plot to represent the frequency of SNPs at different genomic positions, BioCircos plot to represent SNPs with corresponding coding regions and example of SNP dataset as binary matrix 1 and 0 represents presence and absence of SNP in any specific sequence.

technique generally represents pairwise dissimilarity information of n data objects as an $n \times n$ image, where the data objects are reordered in such a way so that the resulting image is able to highlight potential cluster structure in the dataset. Therefore, the dataset is sorted first according to the cluster labels obtained after clustering. Subsequently, the distance matrix, e.g., Jaccard or Hamming, is computed for graphical representation. Boxes lying on the main diagonal represent the clusters structure.

2.3. Pipeline of the workflow

The pipeline of the workflow is shown in Fig. 1(a). In order to find the virus strains, it is important to prepare the SNP dataset and then to use that datatset in hierarchical clustering with different distance functions to build a consensus of clustering results so that we can determine robust and deterministic number of clusters as virus strains that are present in the dataset. Initially to prepare the SNP dataset, 10664 SARS-CoV-2 sequences are aligned using multiple sequencing alignment (MSA) technique called Clustal Omega (ClustalO) (Sievers et al., 2011; Sievers and Higgins, 2014) in presence of reference sequence from NCBI. The choice of selecting ClustalO is taken because of its popularity, speed and accuracy. After performing ClustalO, a consensus sequence is built so that the mutation as substitution, so called SNPs that occur in more than 1% of the virus population for a given genomic position i.e. more than 106 viruses can be identified. The detection technique is shown in Fig. 1(b). Once such SNPs are identified, a SNP binary dataset of is prepared based on its presence in the virus sequences as shown in the third figure of Fig. 1(c). Thereafter, such binary dataset is used for hierarchical clustering using Average Linkage and Complete Linkage with the Jaccard and Hamming distance functions separately in iterative manner by considering number of Clusters 2 to 100. In each iteration, Silhouette Index is computed to measure the goodness of the clustering results. Once the number of clusters are determined by each of the four methods, the VAT plots of such clustering results are prepared to see the structure of clusters. Based on the VAT plots, two or more clustering results are compared to create a final consensus clustering solution which gives the robust and stable clusters as virus strains. After finding the clusters as virus strains, the presence of virus strains in different countries is identified as the distribution of the genomic sequences is known. Furthermore, top 10 signature SNPs in each cluster are identified based on their frequency of occurrence in the cluster.

3. Results

The results of the experiments are explained here. Initially, 107 SNPs are identified in both coding and non-coding regions from the 10664 SARS-CoV-2 sequences. SNPs represent mutation as substitution that occurs in more than 1% of the virus population for a given genomic position. Such SNPs are shown using bar and BioCircos plots in the first and second figures of Fig. 1(c). In bar plot, the frequency of SNPs with their genomic locations are shown while the BioCircos plot represents the SNPs with the corresponding coding regions. In the first figure of Fig. 1(c), SNPs at coordinates 241, 3037, 14,408, 23,403 and 29,816

Table 1
Optimal number of clusters produced by hierarchical clustering methods on SNPs data.

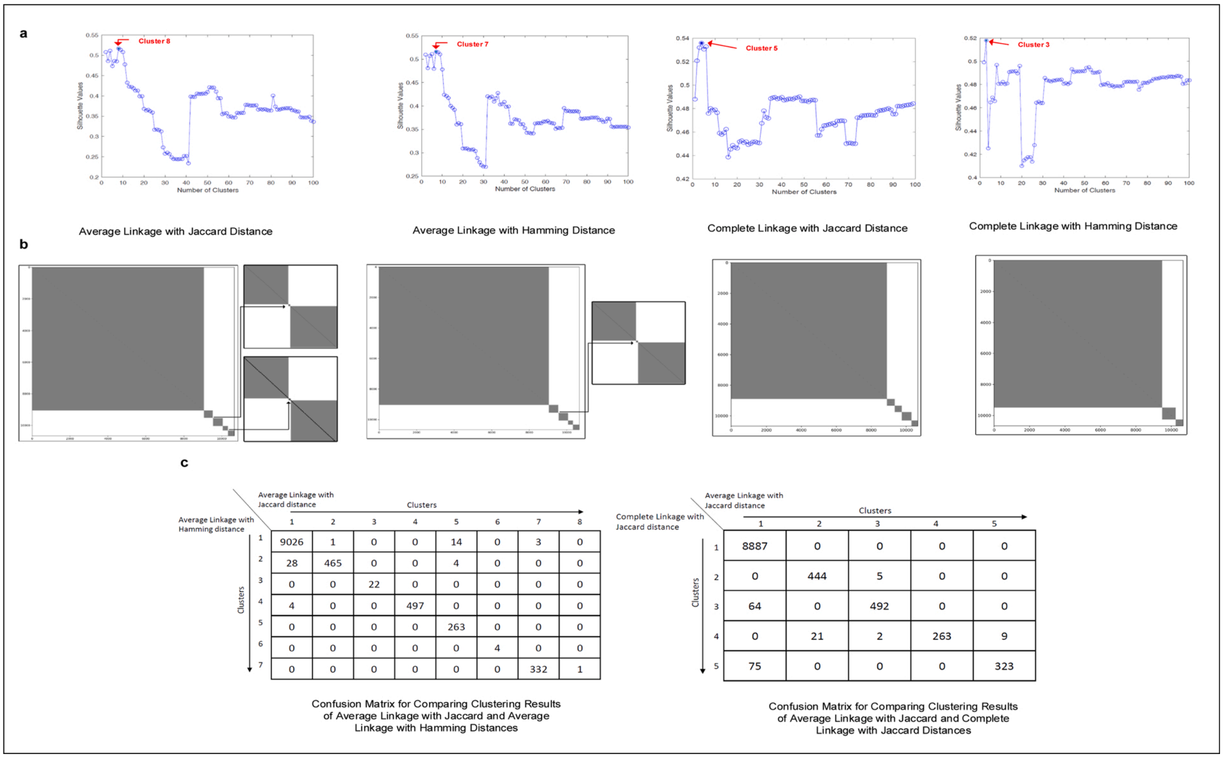
Method	Distance function	Cluster validity index	Number of optimal clusters	Silhouette value
Average Linkage	Jaccard	Silhouette Index	8	0.5163
	Hamming		7	0.5130
Complete Linkage	Jaccard		5	0.5317
	Hamming		3	0.5103

occur in more than 30% of the virus population, that is in more than 3199 genomes out of 10664. The first coordinate belongs to 5′-UTR, the next two are in ORF1ab, 23,403 belongs to Spike, while the last coordinate is in 3′-UTR. Thereafter, by considering the presence or absence of a SNP in the 10664 sequences, the corresponding binary dataset is created of size 10664 × 107 as shown in the third figure of Fig. 1(c). For each of the virus sequence, 1 represents the presence of SNP and 0 represents the absence of SNP at a particular genomic coordinate for such 107 identified SNPs.

Once the SNP binary dataset is prepared, hierarchical clustering such as Average Linkage and Complete Linkage with Jaccard and Hamming distance functions are considered iteratively for number of clusters ranging from 2 to 100 to find the number of clusters as virus strains. These results are presented in Table 1 and Fig. 2(a). It is evident from Fig. 2(a) that the silhouette values are highest at cluster numbers 8, 7, 5 and 3 respectively for Average Linkage with Jaccard distance, Average Linkage with Hamming distance, Complete Linkage with Jaccard distance and Complete Linkage with Hamming distance. The silhouette values for these clusters are reported in Table 1.

Moreover, the mapping of the SARS-CoV-2 genomes to each cluster is reported in Table 2. From Table 1, it can be seen that for Average Linkage with Jaccard and Hamming distance respectively, the silhouette values are 0.5163 and 0.5130 for Clusters 8 and 7. Similarly, for Complete Linkage, the silhouette values are 0.5317 and 0.5103 for Clusters 5 and 3 respectively. For both Average Linkage and Complete Linkage, Jaccard distance shows the higher silhouette value for different number of clusters. Therefore, to take a decision about the number of clusters, it is important to see the size and the structure of the clusters. The size of the clusters is shown in Table 2 by mapping the SARS-CoV-2 genomes to different clusters, while the structure of the clusters is visualised using VAT plots in Fig. 2(b).

From the results of Table 2 and Fig. 2(b), it is found that for Average Linkage with Jaccard distance, Clusters 3, 6 and 8 have very small number of SARS-CoV-2 genomes as compared to the other clusters. Similarly, for Hamming Distance, Clusters 3 and 7 are also found to be



Average Linkage with Hamming distance \ Average Linkage with Jaccard distance (Clusters)	1	2	3	4	5	6	7	8
1	9026	1	0	0	14	0	3	0
2	28	465	0	0	4	0	0	0
3	0	0	22	0	0	0	0	0
4	4	0	0	497	0	0	0	0
5	0	0	0	0	263	0	0	0
6	0	0	0	0	0	4	0	0
7	0	0	0	0	0	0	332	1

Confusion Matrix for Comparing Clustering Results of Average Linkage with Jaccard and Average Linkage with Hamming Distances

Complete Linkage with Jaccard distance \ Average Linkage with Jaccard distance (Clusters)	1	2	3	4	5
1	8887	0	0	0	0
2	0	444	5	0	0
3	64	0	492	0	0
4	0	21	2	263	9
5	75	0	0	0	323

Confusion Matrix for Comparing Clustering Results of Average Linkage with Jaccard and Complete Linkage with Jaccard Distances

Fig. 2. (a) The plots of silhouette values for Average Linkage and Complete Linkage with Jaccard and Hamming distances for number of clusters ranging from 2 to 100, (b) The VAT plots of the clusters produced by Average Linkage and Complete Linkage with Jaccard and Hamming distances for higher silhouette values, (c) The confusion matrices for comparing clustering results of Average Linkage with Jaccard and Hamming distances as reported in Table 2 and Average Linkage and Complete Linkage with Jaccard distance as reported in Table 3.

Table 2
Mapping of SARS-CoV-2 genomes to the different clusters produced by hierarchical clustering methods.

Method	Distance function	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Average Linkage	Jaccard	9058	466	22	497	281	4	335	1
	Hamming	9044	497	26	501	263	332	1	–
Complete Linkage	Jaccard	8898	444	407	545	360	–	–	–
	Hamming	9488	766	410	–	–	–	–	–

Table 3

Re-evaluation of five major clusters produced by Average Linkage and Complete Linkage clustering with Jaccard distance using silhouette value.

Method	Distance function	Cluster validity index	Number of clusters	Silhouette value
Average Linkage	Jaccard	Silhouette Index	5	0.5581
Complete Linkage			5	0.5344

very small in size. On the other hand, for Complete Linkage with Jaccard and Hamming Distances we have 5 and 3 clusters respectively where all the clusters have reasonable number of SARS-CoV-2 genomes. The clustering results of Average Linkage with Jaccard and Hamming distances are compared using confusion matrix in the first figure of Fig. 2(c) in order to build the consensus between them to identify the common number of SARS-CoV-2 genomes in the different clusters. As a result, five major clusters are identified of size 9026, 465, 497, 263 and 332. Thereafter, these results are subsequently compared once again in order to build the consensus between the clustering results of Complete Linkage with Jaccard distance using confusion matrix in the second figure of Fig. 2(c). It also shows five major clusters with 8887, 444, 492, 263 and 323 SARS-CoV-2 genomes. Furthermore, to evaluate the five clusters obtained from Average Linkage and Complete Linkage with Jaccard distances, silhouette values are computed and reported in Table 3.

The results show that the five clusters obtained by applying Average Linkage have a higher silhouette value, 0.5581, in comparison to all the other cases. Thus, after analysis of SNPs data of 10664 SARS-CoV-2 genomes of 73 countries, we can conclude five virus strains are present. After finding these five clusters as virus strains, SARS-CoV-2 genomes in 73 countries are mapped to the five clusters with the corresponding percentage which are reported in Table 4. The same is also visualised through Circos plot in Fig. 3. All the detailed clustering results are reported in Supplementary Table S1.

Furthermore, top 10 signature SNPs from each cluster are identified and reported in Table 5. In unsupervised learning (clustering), selection of features which may define a cluster is a non-trivial task. In this work, this selection has been performed by considering the frequency of a SNP in a cluster. For example, the frequency of occurrence of mutation point 241 is 6088 and thus considered to be the top most signature SNP in Cluster 1. To depict the common signature in the five clusters, visualisation in the form of Venn diagram is shown in Fig. 4. As can be seen from the figure, there are no common SNPs in all the five clusters, thereby confirming the fact that such signature SNPs are features which indeed define the clusters. It is worth mentioning here that multiple changes in nucleotide may lead to multiple amino acid changes as well. For example in Table 5, at mutation point 14,408 there are two nucleotide changes, C>T and C>A. As a consequence, there are two changes in amino acid as well, P323L and P323H. Thus, for 50 signature SNPs in 5 clusters, the total number of non-synonymous signature SNPs are 20, out of which 16 are unique.

4. Discussions

SARS-CoV-2 has turned out to be a worldwide pandemic and has caused disruptions of epic proportions in human lives. In this situation, the identification of virus strains is a very important task. In this regard, we have analysed 10664 SARS-CoV-2 genomes of 73 countries around the world which resulted in some major outcomes: SNP identification from 10664 SARS-CoV-2 genomes of 73 countries, binary dataset creation from SNP data to find the number of clusters as virus strains present in the 73 countries, identification of top 10 signature SNPs for each cluster and the determination of the structural stability of the non-synonymous SNPs to judge the characteristics of the identified clusters.

Initially, we have identified 107 SNPs which are then used to prepare the binary dataset to identify the presence or absence of SNPs. Hierarchical clustering such as Average Linkage and Complete Linkage with Jaccard and Hamming distance functions are then applied on this dataset to the number of clusters as virus strains. These clusters can be used to identify and design peptide based synthetic vaccines viz. epitopes (Ghosh et al., 2021). Also, top 10 signature SNPs from each cluster are identified based on their frequency, the details of which are mentioned in the Results section.

Sequence and structural homology-based prediction of the non-synonymous signature SNPs along with their protein stability are reported in Table 6 using tools like PROVEAN (Protein Variation Effect Analyser) (Choi and Chan, 2015), PolyPhen-2 (Polymorphism Phenotyping) (Adzhubei et al., 2010) and I-Mutant 2.0 (Capriotti et al., 2005) to judge the characteristics of the identified clusters. PROVEAN[7] works on sequence based prediction algorithm while the prediction of Polyphen-2[8] is based on sequence, structural and phylogenetic information pertaining to a SNP. On the other hand, I-Mutant 2.0[9] uses support vector machine (SVM) for the automatic prediction of protein stability changes upon single point mutations. PROVEAN and PolyPhen-2 are used to find the deleterious or damaging non-synonymous SNPs. The threshold value of PROVEAN is set at −2.5. If the PROVEAN score of a SNP is equal to or below this threshold, the corresponding non-synonymous mutation is considered to be deleterious. For Polyphen-2, this range is between 0 to 1. If the score is closer to 1, mutations are more confidently considered to be damaging. From the consensus of both PROVEAN and PolyPhen-2, it can be seen from Table 6 that out of 16 unique non-synonymous signature SNPs, 5 are predicted to be deleterious or damaging, out of which T85I, Q57H and R203M are in NSP2, ORF3a and Nucleocapsid respectively for Cluster 1. For both Clusters 3 and 4, such damaging non-synonymous signature SNPs are F506L and S507C in Exon while Clusters 2 and 3 do not exhibit any such behaviour due to the absence of non-synonymous signature SNPs. All of them are marked in bold in Table 6. Furthermore, protein stability is important to determine the functional and structural activity of a protein. Protein stability dictates the conformational structure of the protein, thereby determining its function. Any change in protein stability may cause misfolding, degradation or aberrant conglomeration of proteins (Hossain et al., 2020). The protein stabilities of the non-synonymous signature SNPs are determined using I-Mutant 2.0. The changes in the protein stability in I-Mutant 2.0 tool is predicted using reliability index (RI) and free energy change values (DDG). The outcome of I-Mutant 2.0 revealed that all of the deleterious 5 unique non-synonymous signature SNPs decrease the stability of the protein structure. These 5 SNPs are shown in their respective protein structures in Fig. 5. The rest of the structures are provided in the Supplementary Figures S1 and S2. All these structures are taken from Zhanglab[10] in the form of respective PDB files. It is to be noted that although some mutations are neutral, their stability can decrease like D614G in Spike protein.

The temporal evolution of SARS-CoV-2 genomes in the five identified clusters for the 73 countries from the month of January till July 2020 is reported in Table 7. For example, in the month of January, 548 genomes are present in Cluster 1. However, the number increased to 6264 in the month of March. In April, there were 1510 genomes while in May, June and July, the numbers are 541, 3 and 21 respectively. Furthermore, to understand the evolution of the SARS-CoV-2 genomes from the month of January till July 2020, their presence in different countries are identified by mapping the genomes to the five major clusters for the 73 countries. This is depicted in the form of pie charts in Table 8. The

[7] https://provean.jcvi.org/index.php.

[8] http://genetics.bwh.harvard.edu/pph2/.

[9] http://folding.biofold.org/i-mutant/i-mutant2.0.html.

[10] https://zhanglab.ccmb.med.umich.edu/COVID-19/.

Table 4
Mapping of SARS-CoV-2 genomes to the five clusters and the corresponding percentage.

Country	Total number of genomes	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
		Number of genome	(%)	Number of genome	(%)	Number of genome	(%)	Number of genome	(%)	Number of genome	(%)
USA	2546	2236	87.82	7	0.27	64	2.51	76	2.99	163	6.40
England	1592	1188	74.62	187	11.75	115	7.22	9	0.57	93	5.84
China	631	585	92.71	1	0.16	11	1.74	30	4.75	4	0.63
Australia	582	345	59.28	186	31.96	29	4.98	14	2.41	8	1.37
Netherlands	568	565	99.47	0	0	2	0.35	0	0	1	0.18
India	566	495	87.46	0	0	20	3.53	48	8.48	3	0.53
Iceland	462	381	82.47	1	0.22	74	16.02	6	1.30	0	0
Scotland	434	406	93.55	0	0	28	6.45	0	0	0	0
Belgium	426	421	98.83	0	0	5	1.17	0	0	0	0
Portugal	349	342	97.99	0	0	7	2.01	0	0	0	0
Spain	267	206	77.15	24	8.99	15	5.62	0	0	22	8.24
Wales	214	158	73.83	33	15.42	15	7.01	2	0.93	6	2.80
Sweden	194	194	100	0	0	0	0	0	0	0	0
France	189	183	96.83	2	1.06	2	1.06	1	0.53	1	0.53
New Zealand	175	160	91.43	0	0	15	8.57	0	0	0	0
Switzerland	164	109	66.46	0	0	16	9.76	38	23.17	1	0.61
Denmark	109	101	92.66	0	0	7	6.42	0	0	1	0.92
Japan	94	89	94.68	0	0	2	2.13	0	0	3	3.19
Brazil	81	80	98.77	0	0	1	1.23	0	0	0	0
Canada	72	67	93.06	1	1.39	4	5.56	0	0	0	0
Luxembourg	71	60	84.51	0	0	7	9.86	1	1.41	3	4.23
Germany	68	67	98.53	0	0	1	1.47	0	0	0	0
Itay	66	55	83.33	1	1.52	7	10.61	3	4.55	0	0
Kazakhstan	49	26	53.06	1	2.04	0	0	22	44.90	0	0
Oman	42	41	97.62	0	0	0	0	1	2.38	0	0
Poland	39	33	84.62	0	0	4	10.26	0	0	2	5.13
South Korea	36	36	100	0	0	0	0	0	0	0	0
Vietnam	31	28	90.32	2	6.45	0	0	1	3.23	0	0
Singapore	28	18	64.29	0	0	2	7.14	1	3.57	7	25
Thailand	28	23	82.14	0	0	3	10.71	2	7.14	0	0
Russia	27	26	96.30	0	0	1	3.70	0	0	0	0
Finland	26	18	69.23	0	0	6	23.08	0	0	2	7.69
Czech Republic	25	25	100	0	0	0	0	0	0	0	0
Mexico	21	19	90.48	0	0	1	4.76	1	4.76	0	0
Norway	20	14	70	0	0	5	25	1	5	0	0
Northern Ireland	19	8	42.11	8	42.11	2	10.53	1	5.26	0	0
Estonia	18	18	100	0	0	0	0	0	0	0	0
Austria	16	14	87.50	0	0	2	12.50	0	0	0	0
Chile	15	14	93.33	0	0	1	6.67	0	0	0	0
DRC	15	12	80	0	0	1	6.67	0	0	2	13.33
Colombia	14	11	78.57	0	0	2	14.29	1	7.14	0	0
Senegal	19	12	63.16	1	5.26	0	0	0	0	6	31.58
Croatia	12	10	83.33	0	0	0	0	2	16.67	0	0
Georgia	11	5	45.45	0	0	0	0	0	0	6	54.55
Kenya	11	11	100	0	0	0	0	0	0	0	0
Malaysia	11	8	72.73	0	0	3	27.27	0	0	0	0
Romania	11	11	100	0	0	0	0	0	0	0	0
South Africa	11	7	63.64	0	0	4	36.36	0	0	0	0
Ireland	10	10	100	0	0	0	0	0	0	0	0
Latvia	10	10	100	0	0	0	0	0	0	0	0
Nigeria	8	8	100	0	0	0	0	0	0	0	0
Kuwait	7	5	71.43	0	0	1	14.29	1	14.29	0	0
Turkey	5	4	80	0	0	0	0	0	0	1	20
Bangladesh	4	4	100	0	0	0	0	0	0	0	0
Greece	4	4	100	0	0	0	0	0	0	0	0
Qatar	4	3	75	0	0	0	0	1	25	0	0
Slovakia	4	4	100	0	0	0	0	0	0	0	0
Algeria	3	3	100	0	0	0	0	0	0	0	0
Argentina	3	3	100	0	0	0	0	0	0	0	0
Belarus	3	3	100	0	0	0	0	0	0	0	0
Hungary	3	3	100	0	0	0	0	0	0	0	0
Saudi Arabia	4	0	0	0	0	1	25	0	0	3	75
Indonesia	2	1	50	0	0	1	50	0	0	0	0
Israel	2	2	100	0	0	0	0	0	0	0	0
Pakistan	2	2	100	0	0	0	0	0	0	0	0
Serbia	2	2	100	0	0	0	0	0	0	0	0
Slovenia	2	2	100	0	0	0	0	0	0	0	0
Cambodia	1	1	100	0	0	0	0	0	0	0	0
Lithuania	1	1	100	0	0	0	0	0	0	0	0

(*continued on next page*)

Table 4 (*continued*)

Country	Total number of genomes	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
		Number of genome	(%)	Number of genome	(%)	Number of genome	(%)	Number of genome	(%)	Number of genome	(%)
Morocco	1	1	100	0	0	0	0	0	0	0	0
Nepal	1	1	100	0	0	0	0	0	0	0	0
Panama	1	1	100	0	0	0	0	0	0	0	0
Peru	1	1	100	0	0	0	0	0	0	0	0

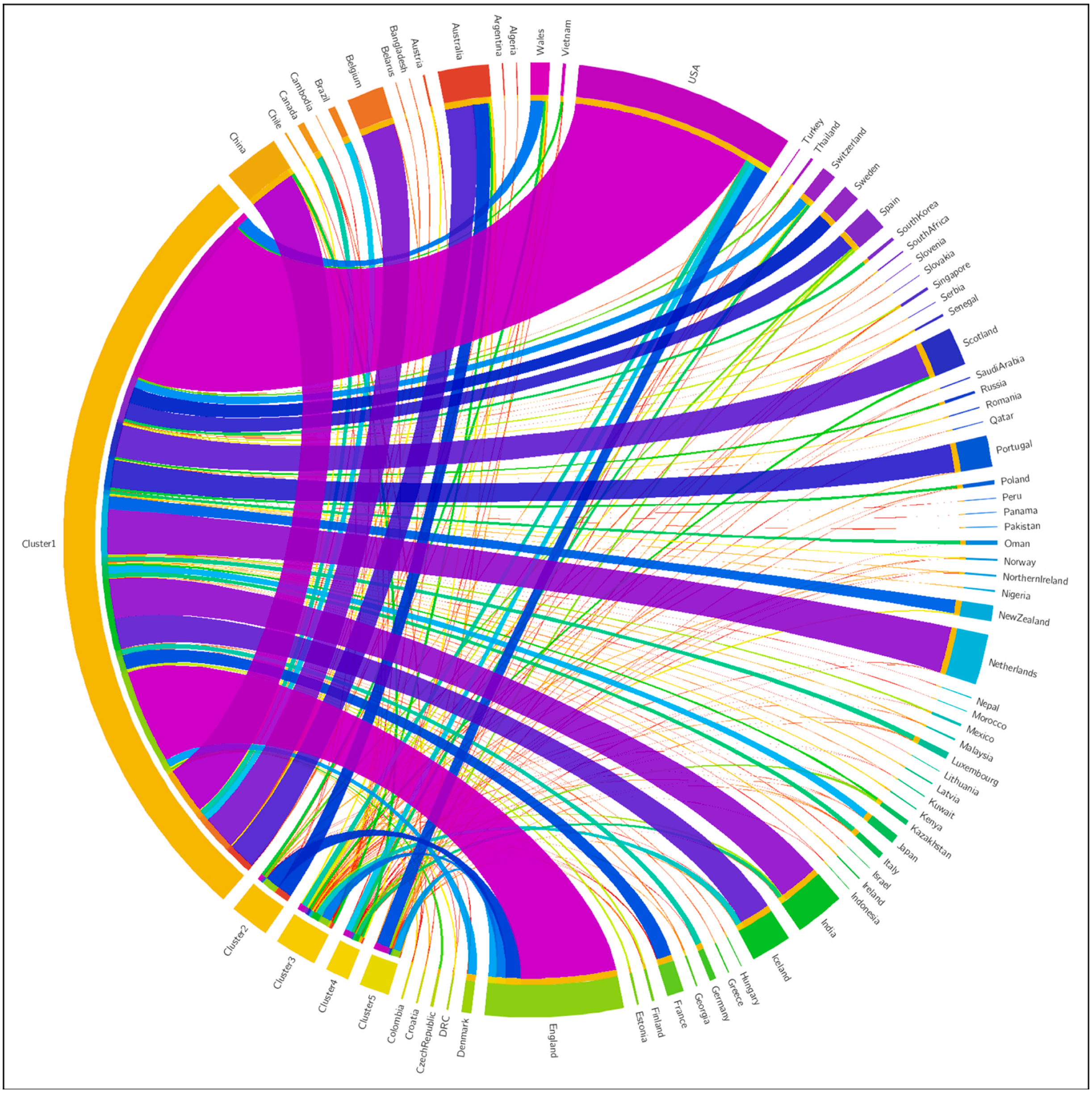


Fig. 3. Circos plot to visualise the mapping of SARS-CoV-2 genomes to five clusters.

corresponding colour representation for the five major clusters and the months are shown in Fig. 6. The evolution of the genomes is quite evident from the pie charts in Table 8. For example, for USA, SARS-CoV-2 genomes mapped to Cluster 1 are mostly in the month of March and almost disappears by the month of May. For Cluster 2, the genomes are not present beyond the month of April. This indicates that the SARS-CoV-2 genomes present in Cluster 1 and Cluster 2 are either not virulent any longer after the months of May and April respectively or through evolution they may have mutated to be a part of some other clusters. Similarly for the other countries as well, one or the other kind of evolutions for the SARS-CoV-2 genomes for the five different clusters are observed.

In order to see the common clusters among 73 countries, heatmap is created and shown in Fig. 7. In the heatmap, deep blue indicates lesser

Table 5
Top 10 signature SNPs in each cluster.

Cluster	Number of sequences in each cluster	Coordinate of signature SNPs	Occurrence of signature SNPs in genome	Change in nucleotide	Change in amino acid	Coordinate of amino acid in protein	Mapped with coding and Non-coding region
Cluster 1	8887	241	6088	C>T	NA	NA	5′-UTR
		1059	1735	C>T	T>I	85	ORF1ab
		3037	6071	C>T	Synonymous	106	ORF1ab
		14,408	6046	(C>T)(C>A)	(P>L), (P>H)	323	ORF1ab
		23,403	6073	A>G	D>G	614	Spike
		25,563	2232	(G>T)(G>C)	Q>H	57	ORF3a
		28,881	1855	(G>A)(G>T)	(R>K) (R>M)	203	Nucleocapsid
		28,882	1848	(G>A)(G>T)	Synonymous, (R>S)	203	Nucleocapsid
		28,883	1847	G>C	G>R	204	Nucleocapsid
		29,816	2613	(T>A)(T>G)	NA	NA	3′-UTR
Cluster 2	444	29,816	416	(T>A)(T>G)	NA	NA	3′-UTR
		29,857	348	(C>A)(C>T)(C>G)	NA	NA	3′-UTR
		29,858	369	(T>A)(T>C)(T>G)	NA	NA	3′-UTR
		29,859	402	(T>A)(T>G)(T>C)	NA	NA	3′-UTR
		29,861	416	(G>A)(G>C)(G>T)	NA	NA	3′-UTR
		29,862	427	(G>C)(G>A)(G>T)	NA	NA	3′-UTR
		29,864	435	(G>A)(G>C)(G>T)	NA	NA	3′-UTR
		29,867	437	(T>A)(T>G)(T>C)	NA	NA	3′-UTR
		29,868	435	(G>A)(G>T)(G>C)	NA	NA	3′-UTR
		29,870	433	(C>A)(C>G)(C>T)	NA	NA	3′-UTR
Cluster 3	492	19,557	475	(T>A)(T>C)(T>G)	(F>L), Synonymous, (F>L)	506	ORF1ab
		19,558	479	(A>G)(A>C)(A>T)	(S>G) (S>R) (S>C)	507	ORF1ab
		22,506	469	(C>A)(C>T)(C>G)	(T>N) (T>I) (T>S)	315	Spike
		29,776	486	(A>G)(A>T)	NA	NA	3′-UTR
		29,779	489	(G>A)(G>T)	NA	NA	3′-UTR
		29,780	487	(A>G)(A>C)	NA	NA	3′-UTR
		29,781	492	(G>A)(G>T)(G>C)	NA	NA	3′-UTR
		29,782	487	(A>G)(A>C)	NA	NA	3′-UTR
		29,783	490	(G>C)(G>T)(G>A)	NA	NA	3′-UTR
		29,784	483	(C>T)(C>A)(C>G)	NA	NA	3′-UTR
Cluster 4	263	19,557	263	(T>A)(T>C)(T>G)	(F>L), Synonymous, (F>L)	506	ORF1ab
		19,558	263	(A>G)(A>C)(A>T)	(S>G) (S>R) (S>C)	507	ORF1ab
		29,858	259	(T>A)(T>C)(T>G)	NA	NA	3′-UTR

(*continued on next page*)

Table 5 (*continued*)

Cluster	Number of sequences in each cluster	Coordinate of signature SNPs	Occurrence of signature SNPs in genome	Change in nucleotide	Change in amino acid	Coordinate of amino acid in protein	Mapped with coding and Non-coding region
		29,859	262	(T>A)(T>G)(T>C)	NA	NA	3′-UTR
		29,860	248	(A>G)(A>C)(A>T)	NA	NA	3′-UTR
		29,861	263	(G>A)(G>C)(G>T)	NA	NA	3′-UTR
		29,862	263	(G>C)(G>A)(G>T)	NA	NA	3′-UTR
		29,863	245	(A>C)(A>T)(A>G)	NA	NA	3′-UTR
		29,864	261	(G>A)(G>C)(G>T)	NA	NA	3′-UTR
		29,867	249	(T>A)(T>G)(T>C)	NA	NA	3′-UTR
Cluster 5	323	3	302	(T>G)(T>A)(T>C)	NA	NA	5′-UTR
		4	306	(A>G)(A>C)(A>T)	NA	NA	5′-UTR
		5	313	(A>T)(A>G)(A>C)	NA	NA	5′-UTR
		6	314	(A>T)(A>C)(A>G)	NA	NA	5′-UTR
		7	320	(G>T)(G>A)(G>C)	NA	NA	5′-UTR
		8	322	(G>A)(G>C)(G>T)	NA	NA	5′-UTR
		10	321	(T>A)(T>C)(T>G)	NA	NA	5′-UTR
		11	319	(T>C)(T>G)(T>A)	NA	NA	5′-UTR
		12	320	(A>C)(A>T)(A>G)	NA	NA	5′-UTR
		29,816	320	(T>A)(T>G)	NA	NA	3′-UTR

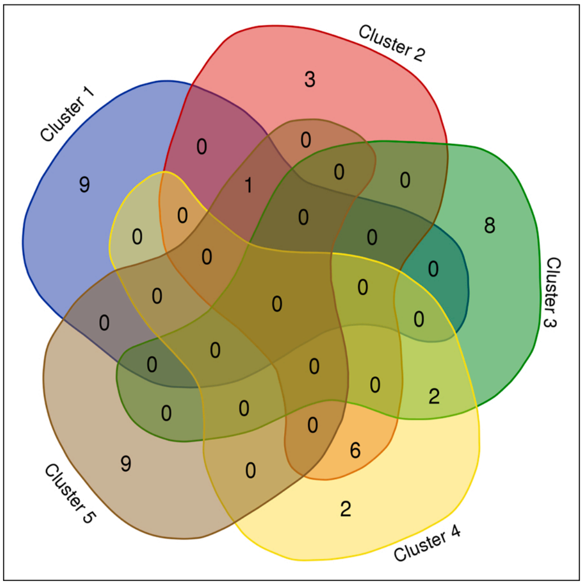


Fig. 4. Venn Digram to represent the common signature SNPs in five clusters.

number of common clusters and yellow shows more number of common clusters. This heatmap is not symmetric. For example, the virus strains from India belong to Clusters 1, 3, 4 and 5 while that for England is distributed among all the five clusters. Thus, if we consider row wise though India shares 100% common clusters as virus strains with England, when considered column wise England shares 80% common clusters as virus strains with India. In this regard, a hypothesis can be drawn that the same vaccine can be effective in those countries with common clusters or virus strains.

It is worth mentioning the advantage of the proposed clustering method as compared to the existing phylogenetic analysers like Maximum-Likelihood and Neighbour-Joining Trees in MEGA-X, GISAID CoVsurver and PANGOLIN. The existing analysers use all substitutions like mutation while the proposed clustering method is more robust as it uses only SNP data, thereby providing faster results. Moreover, SNP data is more relevant in our case as we have small sparse data.

5. Conclusion

In this work, we have analysed 10664 SARS-CoV-2 genomes of 73 countries to identify Single Nucleotide Polymorphisms (SNPs) which comprise of substitution that occurs in more than 1% of the virus population. As a result, 107 SNPs are identified throughout the genome (including coding and non-coding regions) to prepare a binary dataset of SNP. Thereafter, virus strains as clusters are identified from the SNP data. In this regard, hierarchical clustering viz. Average Linkage and Complete Linkage are applied with Jaccard and Hamming distance functions. Additionally, Silhouette Index is used to gauge the goodness of the clusters and also to determine the number of clusters as virus strains with the same distance functions. Jaccard distance gives the better Silhouette value. Thus, the consensus from the two clustering methods using Jaccard distance are used to determine the proper number of clusters which resulted in five major clusters. Moreover, using the presence of the five clusters in the 73 countries, we have also put forth the evolution of the clusters of virus genomes, starting from January till July 2020. Also, top 10 signature SNPs are identified in each cluster and the non-synonymous signature SNPs are visualised in protein structures. The sequence and structural homology-based prediction along with the protein structural stability of these non-synonymous signature SNPs are also determined to judge the characteristics of the identified clusters. Therefore, this work can be summarised as identification of SNPs from 10664 SARS-CoV-2 genomes of 73 countries,

Table 6
Sequence and structural homology-based prediction of non-synonymous signature SNPs along with their protein structural stability.

Cluster	Change in amino acid	Coded protein	PROVEAN		PolyPhen-2		I-Mutant 2.0		
			Prediction	Score	Prediction	Score	Stability	DDG	RI
Cluster 1	**T85I**	**NSP2**	**Deleterious**	**−4.09**	**Probably Damaging**	**0.998**	**Decrease**	**−1.71**	**7**
	P323L	RdRp	Neutral	−0.865	Benign	0.005	Decrease	−0.8	6
	P323H	RdRp	Neutral	−0.865	Benign	0.005	Decrease	−2.09	6
	D614G	Spike	Neutral	0.598	Benign	0.004	Decrease	−1.94	7
	Q57H	**ORF3a**	**Deleterious**	**−3.286**	**Probably Damaging**	**0.966**	**Decrease**	**−1.12**	**7**
	R203K	Nucleocapsid	Neutral	−1.604	Probably Damaging	0.969	Decrease	−2.26	5
	R203M	**Nucleocapsid**	**Deleterious**	**−3.305**	**Probably Damaging**	**0.998**	Decrease	**−1.52**	**4**
	R203S	Nucleocapsid	Neutral	−2.374	Probably Damaging	0.994	Decrease	−2.1	6
	G204R	Nucleocapsid	Neutral	−1.656	Probably Damaging	1	Decrease	0	7
Cluster 3	**F506L**	**Exon**	**Deleterious**	**−5.845**	**Probably Damaging**	**0.98**	**Decrease**	**−2.48**	**5**
	S507G	Exon	Neutral	−2.337	Possibly Damaging	0.662	Decrease	−2.48	8
	S507R	Exon	Neutral	−1.411	Benign	0.015	Increase	−0.39	2
	S507C	**Exon**	**Deleterious**	**−2.759**	**Probably Damaging**	**0.997**	**Decrease**	**−1.51**	**2**
	T315N	Spike	Neutral	−2.206	Probably Damaging	0.998	Decrease	−0.21	2
	T315I	Spike	Neutral	0.365	Probably Damaging	0.999	Decrease	−0.21	4
	T315S	Spike	Neutral	−1.217	Probably Damaging	0.995	Decrease	−0.51	6
Cluster 4	**F506L**	**Exon**	**Deleterious**	**−5.845**	**Probably Damaging**	**0.98**	**Decrease**	**−2.48**	**5**
	S507G	Exon	Neutral	−2.337	Possibly Damaging	0.662	Decrease	−2.48	8
	S507R	Exon	Neutral	−1.411	Benign	0.015	Increase	−0.39	2
	S507C	**Exon**	**Deleterious**	**−2.759**	**Probably Damaging**	**0.997**	**Decrease**	**−1.51**	**2**

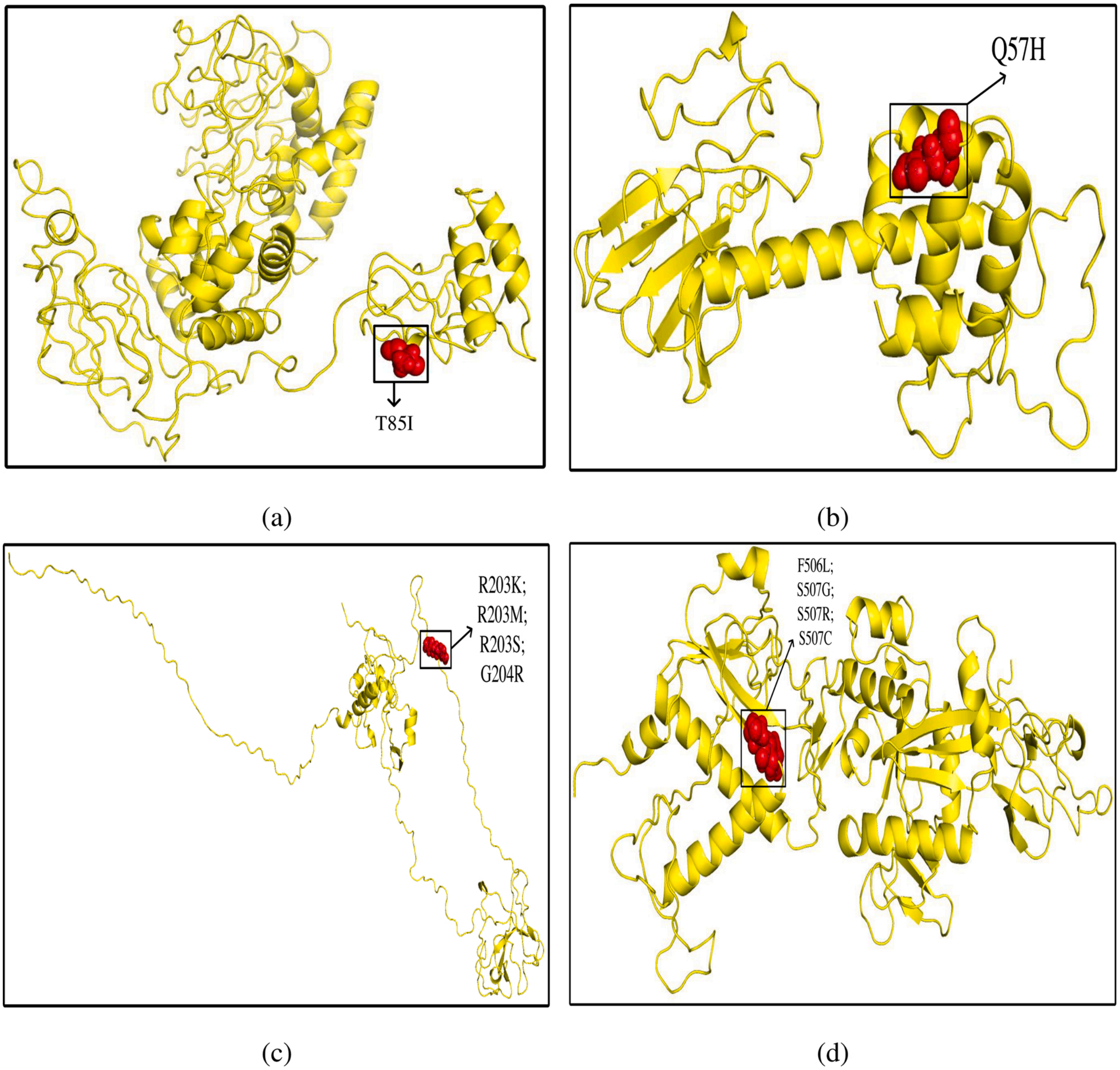


Fig. 5. Non-synonymous signature SNPs highlighted in the structures of (a) NSP2 (b) ORF3a (c) Nucleocapsid and (d) Exon.

Table 7
Temporal evolution of SARS-CoV-2 genomes in the five major clusters from January to July for 73 countries.

Cluster	January	March	April	May	June	July
Cluster 1	548	6264	1510	541	3	21
Cluster 2	13	266	71	94	0	0
Cluster 3	18	276	136	57	0	5
Cluster 4	22	146	71	21	0	3
Cluster 5	12	185	45	81	0	0

creation of binary dataset from SNP data in order to find the number of clusters as virus strains present in 73 countries, identifying signature SNPs in each cluster and determining the structural stability of the non-synonymous signature SNPs to judge the characteristics of the identified clusters.

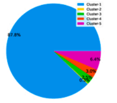
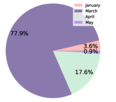
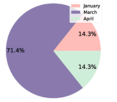
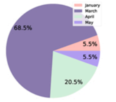
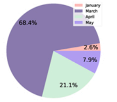
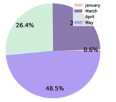
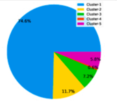
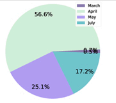
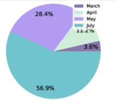
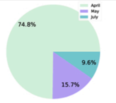
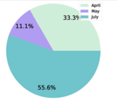
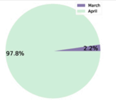
Ethics approval and consent to participate

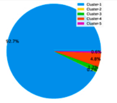
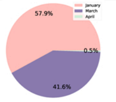
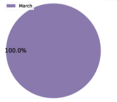
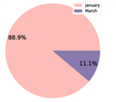
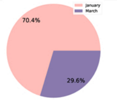
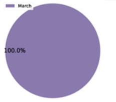
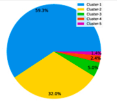
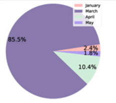
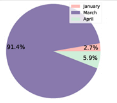
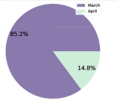
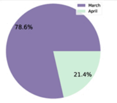
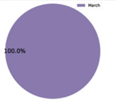
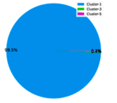
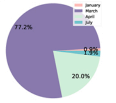
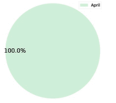
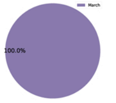
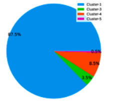
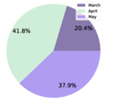
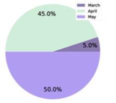
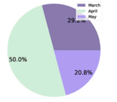
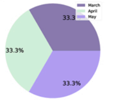
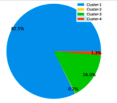
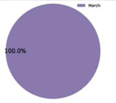
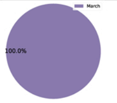
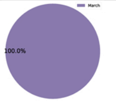
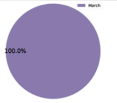
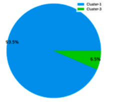
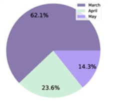
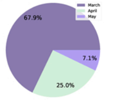
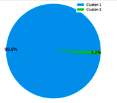
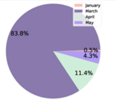
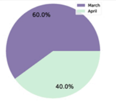
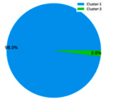
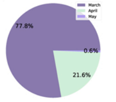
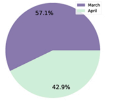
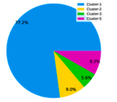
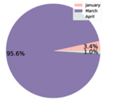
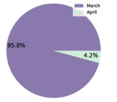
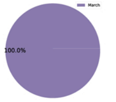
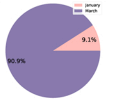
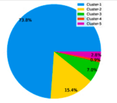
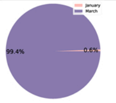
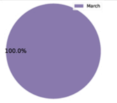
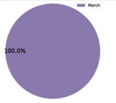
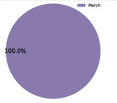
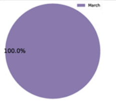
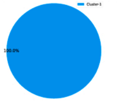
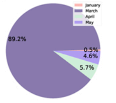
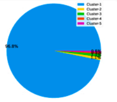
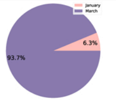
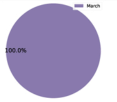
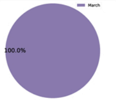
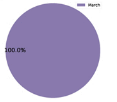
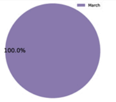
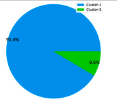
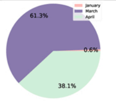
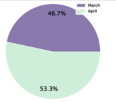
The ethical approval or individual consent was not applicable.

Availability of data and materials

The aligned 10664 SARS-CoV-2 genomes with reference and consensus sequences, SNP dataset and supplementary are available at "http://www.nitttrkol.ac.in/indrajit/projects/COVID-Mutation-10K-Clustering/". Moreover, all the virus genomes used in this work are publicly available at GISAID database.

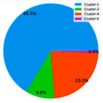
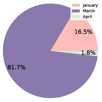
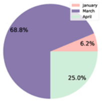
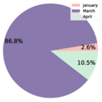
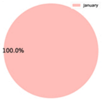
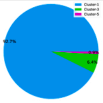
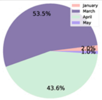
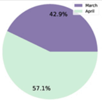
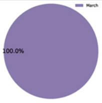
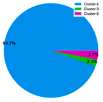
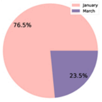
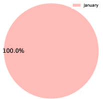
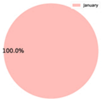
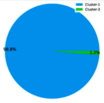
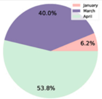
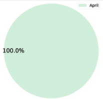
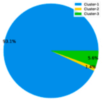
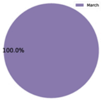
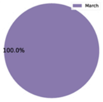
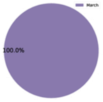
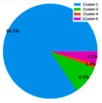
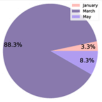
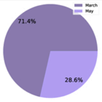
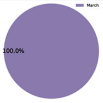
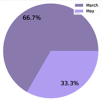
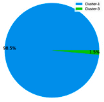
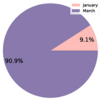
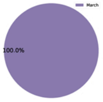
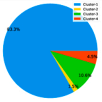
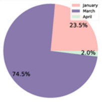
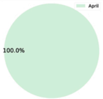
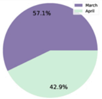
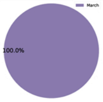
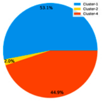
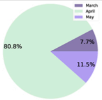
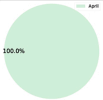
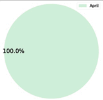
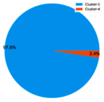
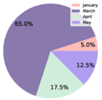
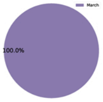
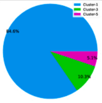
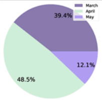
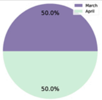
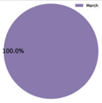
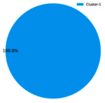
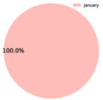
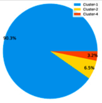
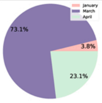
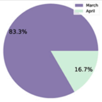
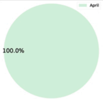
Table 8
Pie charts to represent mapping of SARS-CoV-2 genomes to the five major clusters and the evolution of such genomes from January to July for 73 countries.

Country	All Clusters	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
USA						
England						

Country	All Clusters	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
China						
Australia						
Netherlands						
India						
Iceland						
Scotland						
Belgium						
Portugal						
Spain						
Wales						
Sweden						
France						
New Zealand						

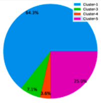
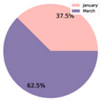
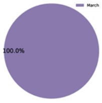
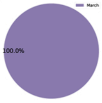
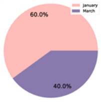
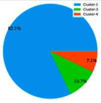
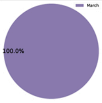
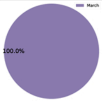
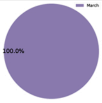
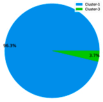
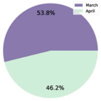
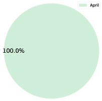
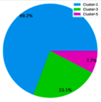
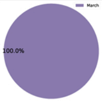
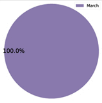
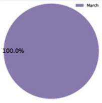
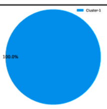
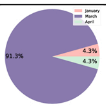
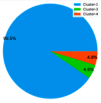
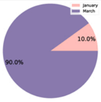
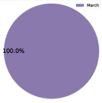
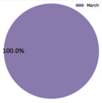
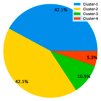
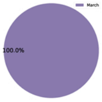
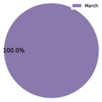
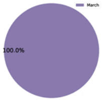
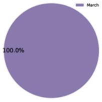
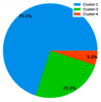
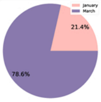
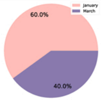
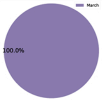
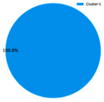
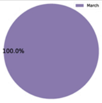
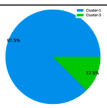
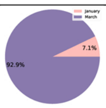
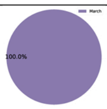
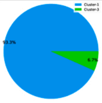
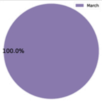
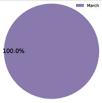
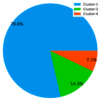
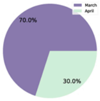
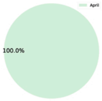
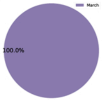
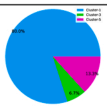
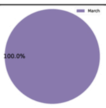
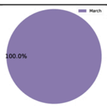
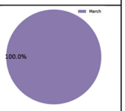
(*continued on next page*)

Table 8 (*continued*)

Country	All Clusters	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Switzerland						
Denmark						
Japan						
Brazil						
Canada						
Luxembourg						
Germany						
Italy						
Kazakhstan						
Oman						
Poland						
South Korea						
Vietnam						

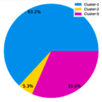
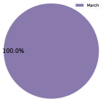
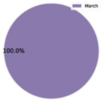
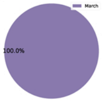
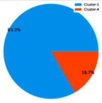
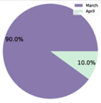
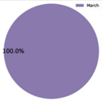
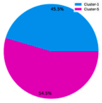
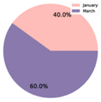
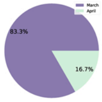
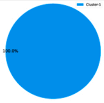
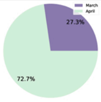
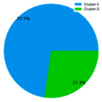
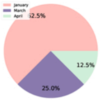
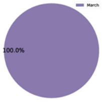
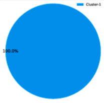
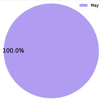
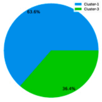
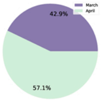
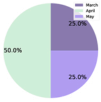
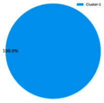
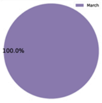
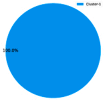
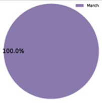
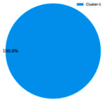
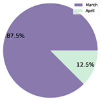
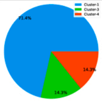
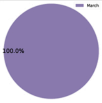
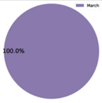
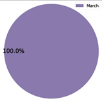
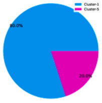
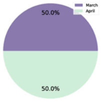
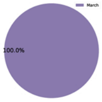
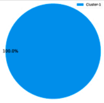
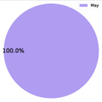
(*continued on next page*)

Table 8 (*continued*)

Country	All Clusters	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Singapore						
Thailand						
Russia						
Finland						
Czech Republic						
Mexico						
Northern Ireland						
Norway						
Estonia						
Austria						
Chile						
Colombia						
DRC						

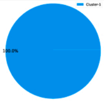
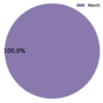
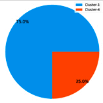
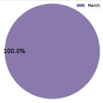
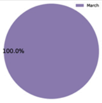
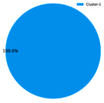
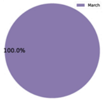
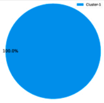
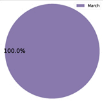
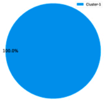
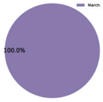
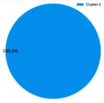
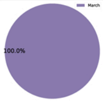
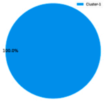
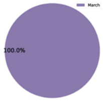
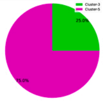
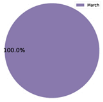
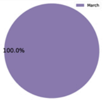
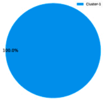
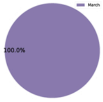
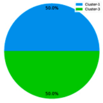
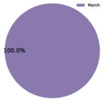
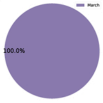
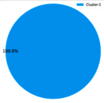
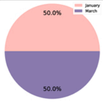
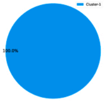
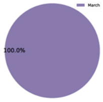
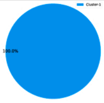
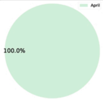
(continued on next page)

Table 8 (*continued*)

Country	All Clusters	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Senegal						
Croatia						
Georgia						
Kenya						
Malaysia						
Romania						
South Africa						
Ireland						
Latvia						
Nigeria						
Kuwait						
Turkey						
Bangladesh						

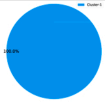
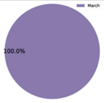
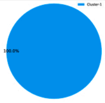
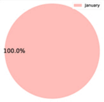
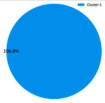
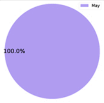
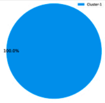
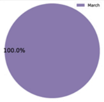
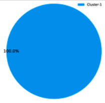
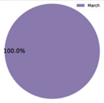
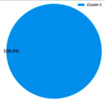
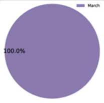
(*continued on next page*)

Table 8 (*continued*)

Country	All Clusters	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Greece						
Qatar						
Slovakia						
Algeria						
Argentina						
Belarus						
Hungary						
Saudi Arabia						
Slovenia						
Indonesia						
Israel						
Pakistan						
Serbia						

(*continued on next page*)

Table 8 (*continued*)

Country	All Clusters	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cambodia						
Lithuania						
Morocco						
Nepal						
Panama						
Peru						

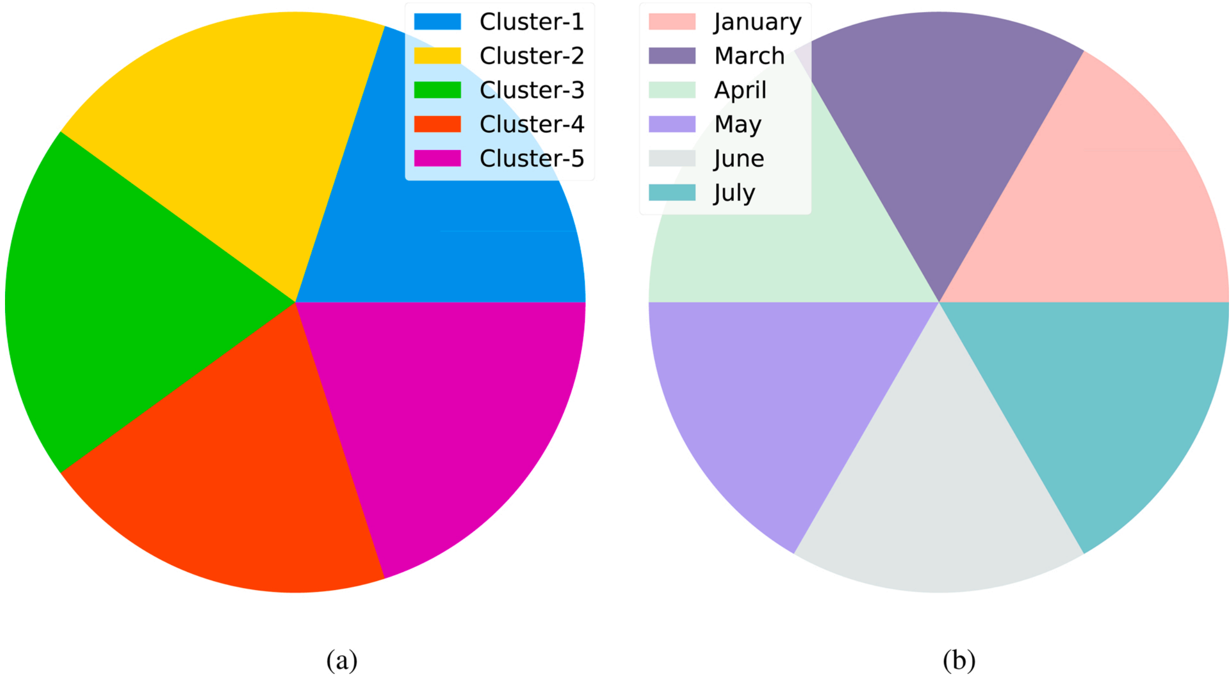


Fig. 6. Colours to represent (a) Five major clusters and b) Months from January to July 2020.

Consent for publication

Not applicable.

Funding

This work has been partially supported by CRGshort term research grant on COVID-19 (CVD/2020/000991) from Science and Engineering Research Board (SERB), Department of Science and Technology, Govt. of India. This work is also partially funded by IDUB against COVID-19 project granted by Warsaw University of Technology under the program Excellence Initiative: Research University (IDUB). It is co-supported by Polish National Science Centre (2019/35/O/ST6/02484), Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund (TEAM to DP).

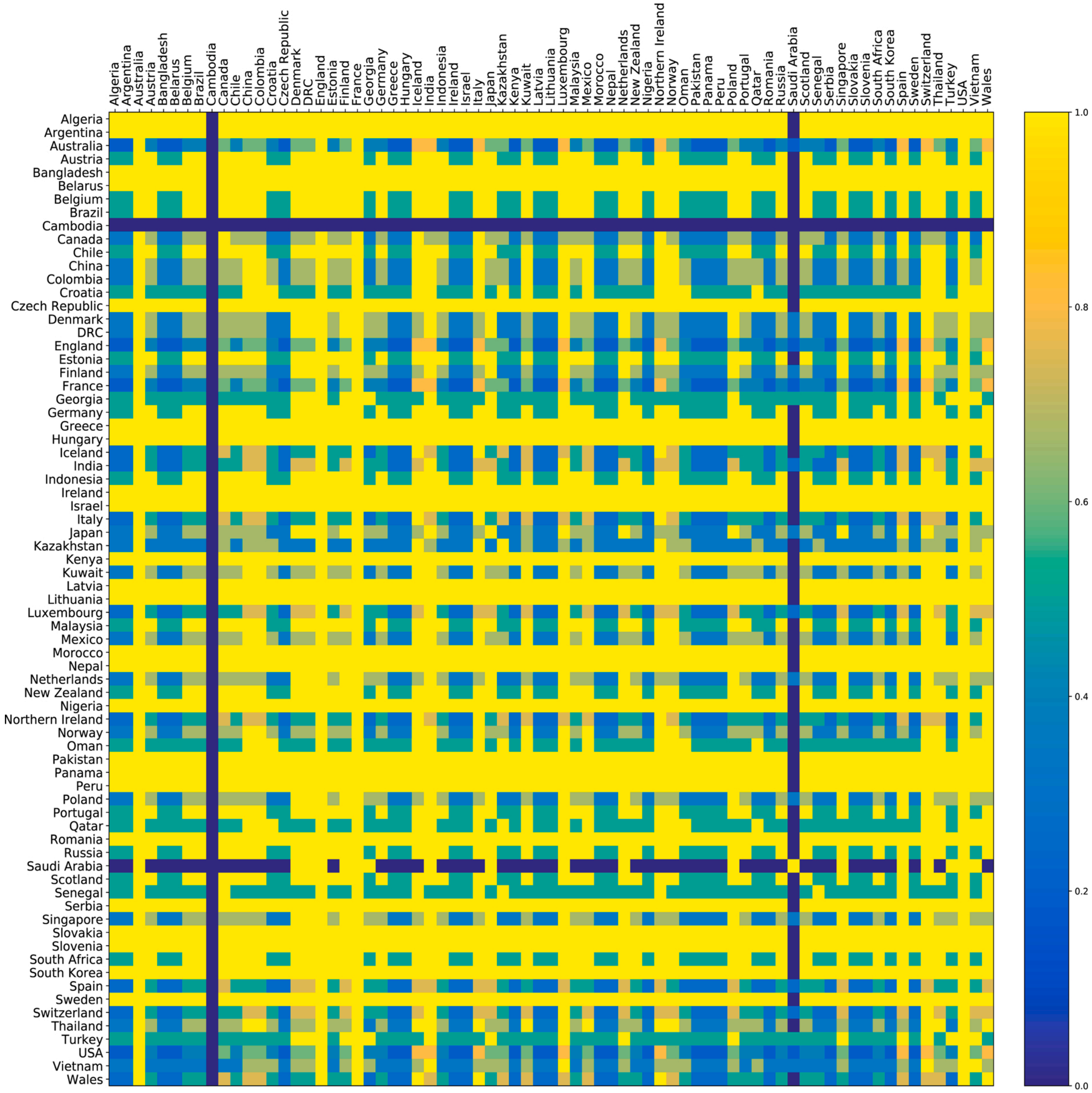


Fig. 7. Heatmap to represent the common clusters as virus strains among 73 countries.

Author contributions

Nimisha Ghosh: Conceptualization; Methodology; Data curation; Formal analysis; Software; Validation; Writing – original draft, **Indrajit Saha**: Conceptualization; Data curation; Supervision; Funding acquisition; Formal analysis; Investigation; Methodology; Web development; Project administration; Resources; Validation; Visualization; Writing – review & editing, **Nikhil Sharma**: Conceptualization; Formal analysis; Software; Validation; Visualization; Writing – review & editing, **Suman Nandi**: Software; Validation; Visualization; Writing – review & editing, **Dariusz Plewczynski**: Conceptualization; Data curation; Supervision; Funding acquisition; Methodology; Project administration; Resources; Validation; Writing – review & editing.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgement

We thank all those who have contributed sequences to GISAID database.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.virusres.2021.198401.

References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R., 2010. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249. https://doi.org/10.1038/nmeth0410-248.

Ahmed, A., 2017. The predictors of 3- and 30-day mortality in 660 mers-cov patients. BMC Infect. Dis. 17, 615. https://doi.org/10.1186/s12879-017-2712-2.

Amela, I., Cedano, J., Querol, E., 2007. Pathogen proteins eliciting antibodies do not share epitopes with host proteins: a bioinformatics approach. PLOS ONE 2, 1–8. https://doi.org/10.1371/journal.pone.0000512.

Bezdek, J.C., Hathaway, R.J., 2002. Vat: a tool for visual assessment of (cluster) tendency. Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290), vol. 3 2225–2230. https://doi.org/10.1109/IJCNN.2002.1007487.

Capriotti, E., Fariselli, P., Casadio, R., 2005. I-mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acid Res. 33, 306–310. https://doi.org/10.1093/nar/gki375.

Choi, Y., Chan, A.P., 2015. Provean web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 31, 2745–2747. https://doi.org/10.1093/bioinformatics/btv195.

Cui, Y., Chen, X., Luo, H., Fan, Z., Luo, J., He, S., et al., 2016. Biocircos.js: an interactive circos javascript library for biological data visualization on web applications. Bioinformatics 32, 1740–1742. https://doi.org/10.1093/bioinformatics/btw041.

Deng, W., Wang, Y., Liu, Z., Cheng, H., Xue, Y., 2014. HemI: a toolkit for illustrating heatmaps. PLoS ONE 9, e111988. https://doi.org/10.1371/journal.pone.0111988.

Devijver, P.A., Kittler, J., 1982. Pattern Recognition: A Statistical Approach. Prentice Hall, London.

Fischer, S., Freuling, C.M., Müller, T., Pfaff, F., Bodenhofer, U., Höper, D., et al., 2018. Defining objective clusters for rabies virus sequences using affinity propagation clustering. PLOS Negl. Trop. Dis. 12, 1–17. https://doi.org/10.1371/journal.pntd.0006182.

Folegatti, P., Ewer, K., Aley, P., Angus, B., Becker, S., Belij, S.R., et al., 2020. Safety and immunogenicity of the chadox1 ncov-19 vaccine against sars-cov-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. Lancet 396, 467–478. https://doi.org/10.1016/S0140-6736(20)31604-4.

Ghosh, N., Sharma, N., Saha, I., Saha, S., 2021. Genome-wide analysis of Indian sars-cov-2 genomes to identify t-cell and b-cell epitopes from conserved regions based on immunogenicity and antigenicity. Int. Immunopharmacol. 91, 107276. https://doi.org/10.1016/j.intimp.2020.107276.

Hahn, G., Lee, S., Weiss, S.T., Lange, C., 2020. Unsupervised cluster analysis of sars-cov-2 genomes reflects its geographic progression and identifies distinct genetic subgroups of sars-cov-2 virus. bioRxiv. https://doi.org/10.1101/2020.05.05.079061.

Hossain, M.S., Roy, A.S., Islam, M.S., 2020. In silico analysis predicting effects of deleterious snps of human rassf5 gene on its structure and functions. Sci. Rep. 10, 14542. https://doi.org/10.1038/s41598-020-71457-1.

Kumar, D., Bezdek, J.C., 2020. Visual approaches for exploratory data analysis: a survey of the visual assessment of clustering tendency (vat) family of algorithms. IEEE Syst. Man Cybern. Mag. 6, 10–48. https://doi.org/10.1109/MSMC.2019.2961163.

Lo, E., Bonizzoni, M., Schroeder, E.H., Ford, A., Janies, D.A., James, A.A., et al., 2018. Selection and utility of single nucleotide polymorphism markers to reveal fine-scale population structure in human malaria parasite plasmodium falciparum. Front. Ecol. Evol. 6, 145. https://doi.org/10.3389/fevo.2018.00145.

Nogales, A., Dieg, L.D., 2019. Host single nucleotide polymorphisms modulating influenza a virus disease in humans. Pathogens 8, 4. https://doi.org/10.3390/pathogens8040168.

Petersen, E., Koopmans, M., Go, U., Hamer, D.H., Petrosillo, N., Castelli, F., et al., 2020. Comparing sars-cov-2 with sars-cov and influenza pandemics. Lancet Infect. Dis. 20, e238–e244. https://doi.org/10.1016/S1473-3099(20)30484-9.

Phan, T., 2020. Genetic diversity and evolution of SARS-CoV-2. Infection. Genet. Evol. 81, 104260. https://doi.org/10.1016/j.meegid.2020.104260.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Saha, I., Ghosh, N., Maity, D., Sharma, N., Sarkar, J.P., Mitra, K., 2020. Genome-wide analysis of Indian sars-cov-2 genomes for the identification of genetic mutation and snp. Infect. Genet. Evol. 104457. https://doi.org/10.1016/j.meegid.2020.104457.

Sevilla-Reyes, E., Chavaro-Pérez, D.A., Piten-Isidro, E., Gutiérrez-González, L.H., Santos-Mendoza, T., 2013. Protein clustering and rna phylogenetic reconstruction of the influenza a virus ns1 protein allow an update in classification and identification of motif conservation. PLOS ONE 8. https://doi.org/10.1371/journal.pone.0063098.

Sievers, F., Higgins, D.G., 2014. Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol. Biol. 1079, 105–116. https://doi.org/10.1007/978-1-62703-646-7_6.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., et al., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol. Syst. Biol. 7 https://doi.org/10.1038/msb.2011.75.

Tou, J.T., Gonzalez, R.C., 1974. Pattern Recognition Principles. Addison-Wesley, Reading.

van Dorp, L., Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., et al., 2020. Emergence of genomic diversity and recurrent mutations in sars-cov-2. Infect. Genet. Evol. 83, 104351. https://doi.org/10.1016/j.meegid.2020.104351.

Worldometer, 2021. Coronavirus Disease 2019 (covid-19) Cases in India (accessed 23.02.20). https://www.worldometers.info/coronavirus/country/india/.

Yang, X., Dong, N., Chan, E., Chen, S., 2020. Genetic cluster analysis of sars-cov-2 and the identification of those responsible for the major outbreaks in various countries. Emerg. Microbes Infect. 9, 1287–1299. https://doi.org/10.1080/22221751.2020.1773745.

Yin, C., 2020. Genotyping coronavirus sars-cov-2: methods and implications. Genomics 112, 3588–3596. https://doi.org/10.1016/j.ygeno.2020.04.016.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., et al., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273. https://doi.org/10.1038/s41586-020-2012-7.