Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Research paper

Topological Analysis for Sequence Variability: Case Study on more than 2K SARS-CoV-2 sequences of COVID-19 infected 54 countries in comparison with SARS-CoV-1 and MERS-CoV

Jnanendra Prasad Sarkar [a,b,1], Indrajit Saha [c,*,1], Arijit Seal [d], Debasree Maity [e], Ujjwal Maulik [b]

[a] Larsen & Toubro Infotech Ltd., Pune, Maharashtra, India
[b] Department of Computer Science and Engineering, Jadavpur University, Kolkata, West Bengal, India
[c] Department of Computer Science and Engineering, National Institute of Technical Teachers' Training & Research, Kolkata, West Bengal, India
[d] Cognizant Technology Solutions, Kolkata, West Bengal, India
[e] Department of Electronics and Communication Engineering, MCKV Institute of Engineering, Howrah, West Bengal, India

ARTICLE INFO

Keywords:
Coronavirsus
COVID-19
Genomic sequences
Global health
MERS-CoV
SARS-CoV-2
Sequence alignment free technique

ABSTRACT

The pandemic due to novel coronavirus, SARS-CoV-2 is a serious global concern now. More than thousand new COVID-19 infections are getting reported daily for this virus across the globe. Thus, the medical research communities are trying to find the remedy to restrict the spreading of this virus, while the vaccine development work is still under research in parallel. In such critical situation, not only the medical research community, but also the scientists in different fields like microbiology, pharmacy, bioinformatics and data science are also sharing effort to accelerate the process of vaccine development, virus prediction, forecasting the transmissible probability and reproduction cases of virus for social awareness. With the similar context, in this article, we have studied sequence variability of the virus primarily focusing on three aspects: (a) sequence variability among SARS-CoV-1, MERS-CoV and SARS-CoV-2 in human host, which are in the same coronavirus family, (b) sequence variability of SARS-CoV-2 in human host for 54 different countries and (c) sequence variability between coronavirus family and country specific SARS-CoV-2 sequences in human host. For this purpose, as a case study, we have performed topological analysis of 2391 global genomic sequences of SARS-CoV-2 in association with SARS-CoV-1 and MERS-CoV using an integrated semi-alignment based computational technique. The results of the semi-alignment based technique are experimentally and statistically found similar to alignment based technique and computationally faster. Moreover, the outcome of this analysis can help to identify the nations with homogeneous SARS-CoV-2 sequences, so that same vaccine can be applied to their heterogeneous human population.

1. Introduction

A worldwide epidemic due to outbreak of a virus disease, COVID-19 caused by a novel virus called Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is first found in Wuhan City, Hubei Province of China on 31st December 2019 as informed by the China Country Office of World Health Organization (WHO) (WHO, 2020; Zhu et al., 2020). The infection and death rate have grown exponentially in subsequent days where more than 43 million infection cases with more than 1 million death cases for coronavirus across globe are reported in "worldometers.info"[2] as on 25th October 2020. According to the medical research community, SARS-CoV-2 transmits human-to-human more rapidly than SARS-CoV-1, but the origin of SARS-CoV-2 is suspected as bat or pangolin (Andersen et al., 2020; Zhang and Holmes, 2020; Zhou et al., 2020). In human-to-human transmission medium, the coronavirus is found transmitting via droplets from one infected person to another individual while the infected person coughs or sneezes over a short distance. The COVID-19 disease seems to be a threat to the human extinct and to control the spreading of the virus, several measures like complete lockdown are taken globally by almost every infected nation.

[*] Corresponding author.
E-mail address: indrajit@nitttrkol.ac.in (I. Saha).
[1] Joint first authors and contributed equally.
[2] https://www.worldometers.info/coronavirus/.

https://doi.org/10.1016/j.meegid.2021.104708
Received 28 July 2020; Received in revised form 27 October 2020; Accepted 31 December 2020
Available online 6 January 2021

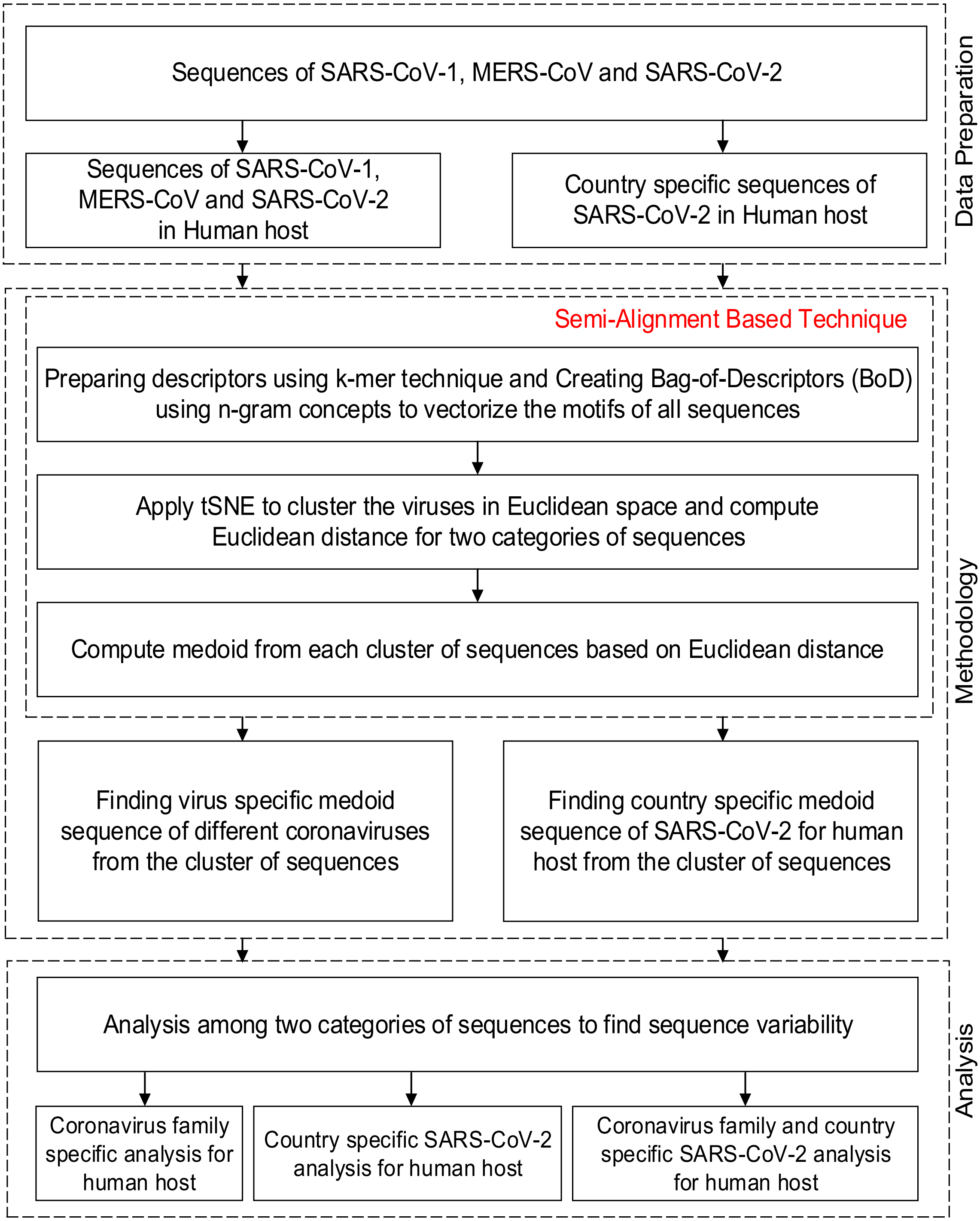

Data Preparation
Sequences of SARS-CoV-1, MERS-CoV and SARS-CoV-2
Sequences of SARS-CoV-1, MERS-CoV and SARS-CoV-2 in Human host
Country specific sequences of SARS-CoV-2 in Human host
Methodology
Semi-Alignment Based Technique
Preparing descriptors using k-mer technique and Creating Bag-of-Descriptors (BoD) using n-gram concepts to vectorize the motifs of all sequences
Apply tSNE to cluster the viruses in Euclidean space and compute Euclidean distance for two categories of sequences
Compute medoid from each cluster of sequences based on Euclidean distance
Finding virus specific medoid sequence of different coronaviruses from the cluster of sequences
Finding country specific medoid sequence of SARS-CoV-2 for human host from the cluster of sequences
Analysis
Analysis among two categories of sequences to find sequence variability
Coronavirus family specific analysis for human host
Country specific SARS-CoV-2 analysis for human host
Coronavirus family and country specific SARS-CoV-2 analysis for human host

Fig. 1. Steps of the workflow.

However, as the long term lockdown is not a permanent solution and has deep impact in global economy (Haleem et al., 2020a), medical research communities of entire world are working to find appropriate vaccine and drug. In this regard, a recent study (Zhu et al., 2020) has found genetic features like potential etiological agents of the SARS-CoV-2 after metagenomic analysis using next-generation sequencing (NGS). A separate research (Wan et al., 2020) has discovered how the spike protein receptor-binding domain (RBD) of SARS-CoV-2 binds with host receptor angiotensin-converting enzyme 2 (ACE2), which is generally considered as primary regulatory agent for transmission of COVID-19 disease in both cross-species and human. Moreover, angiotensin and PPAR family proteins have been found playing vital role in COVID-19 infection (Dey et al., 2020). In this regard, the goal of our earlier research (Saha et al., 2020a; Saha et al., 2020b) was to elucidate the genetic variability of India specific SARS-CoV-2. However, it is important to study the virus genome sequences of different countries in order to understand the sequence variability using topological analysis in faster way, so that the nations with homogeneous sequences can be identified to apply the same vaccine. This fact motivated us to study and experiment the sequence variability in three aspects: (a) sequence similarity measure among three different coronaviruses such as SARS-CoV-1, MERS-CoV and SARS-CoV-2 within same family in human host, (b) SARS-CoV-2 sequence similarity measure in 54 different countries in human host to cover various geographical origins and (c) sequence similarity between coronavirus family and country specific SARS-CoV-2 sequences in human host. For this purpose, as a case study we have considered 340, 291 and 2391 global genomic sequences of SARS-CoV-1, MERS-CoV and SARS-CoV-2 respectively.

The biological experiments to characterize the genetical insights by analyzing genome sequences are typically expensive and the cost grows almost exponentially with the increasing number of sequences. In this regard, computational techniques play an important role. The most widely used sequence comparison method proposed in (Smith and Waterman, 1981) uses dynamic programming to compute optimal local alignment. However, this fails to compare distant sequences effectively. Sequence comparing technique by iteratively pairwise sequence alignment or using multi-sequence alignment techniques are computationally expensive and eventually can be considered as NP-Complete problem. Therefore, heuristic methods like ClustalW (Thompson et al., 1994), Clustal Omega (Sievers et al., 2011) and MUSCLE (Edgar, 2004) gained popularity for multi-sequence alignment to extract valuable information from genome sequences. Subsequently, several other alignment based methods like ARCS (Song et al., 2006), profile Hidden Markov Model (pHMM) (Eddy, 1998), PFam (Punta et al., 2012) are also proposed in this respect. However, major challenge in almost all the alignment based methods is the requirement of correct alignment of multiple sequences. Moreover, computing optimal multi-sequence is NP-Hard problem and it is difficult to compute score correctly for more than two nucleotides. Therefore, less computationally expensive alignment free techniques like k-mer (Manekar and Sathe, 2018; Solis-Reyes et al., 2018) gained popularity. Being an essential part of many bioinformatics processes like genome and transcriptome assembly, metagenomic sequencing, error correction of sequence reads etc. (Melsted and Pritchard, 2011), the importance and superiority of k-mer technique over other techniques like REGA (Pineda-Pena et al., 2013), SCUEAL (Pond et al., 2009), COMET (Struck et al., 2014) etc. have been explained in (Solis-Reyes et al., 2018). Hence, in this article we have proposed semi-alignment based technique using k-mer for sequence analysis. Additionally, t-distributed Stochastic Neighbor Embedding (tSNE) (Hinton and Roweis, 2002) is used on count vector generated through k-mer and n-gram techniques for visualization purpose. After topological study on global sequence variability, we have reported different analytical findings such as (a) genome sequences variability among different SARS-CoV-1, MERS-CoV and SARS-CoV-2 in human host, (b) SARS-CoV-2 sequence variability in human host of different global geographical locations to understand the probable relation and (c) sequence variability between coronavirus family and country specific SARS-CoV-2 sequences in human host. These information can be considered to apply same vaccine to the countries with similar genome sequences.

2. Research objectives

While research communities across world are working on biological and technological advancements (Bahl et al., 2020; Haleem et al., 2020b; Javaid et al., 2020; Singh et al., 2020a; Singh et al., 2020b; Singh et al., 2020c; Vaishya et al., 2020a; Vaishya et al., 2020b) to fight against the challenging pandemic situation of COVID-19, our prime objective of this research is to study the topological genome sequence variability among three intra-family coronaviruses viz. SARS-CoV-1, MERS-CoV and SARS-CoV-2, sequence variability of SARS-CoV-2 in human host for countries in diverse geographical locations as well as sequence variability between coronavirus family and country specific SARS-CoV-2 sequences in human host using semi-alignment based technique. With the analytical findings, the study aims to help medical communities in finding nations with homogeneous sequences for effective application of vaccine and drug. As an additional important objective, the experimental study of this article also shows the similarities between the results produced by both computationally expensive sequence alignment based and computationally less expensive semi-alignment based technique in order to speed up the research in designing prophylactic vaccine and therapeutic drug for SARS-CoV-2.

Table 1
Statistics of the refined genome sequences of coronaviruses in human host.

Virus name	Source of sequence	No. of sequence	Max length of sequence	Avg length of sequence
SARS-CoV-1	NCBI	340	30,311	29,514
MERS-CoV	NCBI	291	30,150	29,983
SARS-CoV-2	GISAID	2391	29,986	29,512

Table 2
Statistics of the country wise refined genome sequences of SARS-CoV-2 in human host.

Country	No. of sequences	Country	No. of sequences	Country	No. of sequences
USA	588	Canada	17	Hungary	2
Iceland	343	Italy	17	Thailand	2
China	321	Singapore	14	Cambodia	1
Netherlands	190	Finland	13	Colombia	1
England	160	South Korea	13	Ecuador	1
Wales	107	Georgia	10	Lithuania	1
Japan	83	Luxembourg	10	Mexico	1
France	75	Denmark	9	Nepal	1
Australia	64	Malaysia	8	Nigeria	1
Belgium	45	New Zealand	8	Northern Ireland	1
Portugal	44	Norway	8	Pakistan	1
India	35	Chile	7	Panama	1
Brazil	34	Ireland	6	Peru	1
Switzerland	31	Vietnam	6	Poland	1
Germany	27	Kuwait	4	Russia	1
Spain	27	Slovakia	4	South Africa	1
Congo	19	Czech Republic	3	Sweden	1
Scotland	18	Saudi Arabia	3	Turkey	1

3. Materials and method

In order to perform experiment and analytical study for research objectives mentioned earlier, first we have collected relevant genome sequences and performed certain important data pre-processing, applied semi-alignment based technique on pre-processed data. Fig. 1 depicts the flow of entire experiment and the following subsections describe the various processes in detail.

3.1. Data preparation

For the experiment, we have downloaded genome sequences of SARS-CoV-1 and MERS-CoV from The National Center for Biotechnology Information (NCBI),[3] whereas genome sequences of SARS-CoV-2 are collected from Global Initiative on Sharing All Influenza Data (GISAID)[4] in fasta format. In order to perform the experiment, it is important to consider the complete or near complete genome sequences of the virus. Therefore, basic data pre-processing is applied to filter the sequences of SARS-CoV-1, MERS-CoV and SARS-CoV-2 having average length more than 29.5 kbp in human host to avoid any incomplete sequence. The Statistics of the refined datasets of coronaviruses are reported in Tables 1 for human host, while the country wise statistics and geoplot of same SARS-CoV-2 sequences are reported in Table 2 and Fig. 2.

[3] https://www.ncbi.nlm.nih.gov/.

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

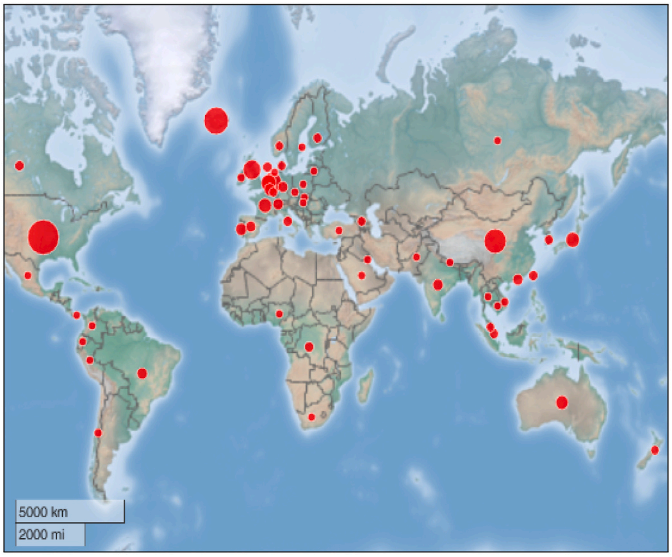


Fig. 2. Geoplot of SARS-CoV-2 sequences in human host for 54 countries.

3.2. *Semi-alignment based technique*

In order to perform the analysis, we have integrated sequence alignment free technique, k-mer (Manekar and Sathe, 2018; Solis-Reyes et al., 2018) together with count vectorization using n-gram and tSNE. k-mer is used to create set of descriptors from complete genome sequence of virus. Subsequently, n-gram is used to build feature by selecting n number of descriptors. Each such n-gram feature set is called Bag-of-Descriptors (BoD). Out of these BoDs, count vectorization technique is used to create numeric feature vector by counting the frequencies of each of n-gram features. Such k-mer generated descriptors and top 10 n-gram BoDs are shown in Fig. 3. Primarily the numeric feature vector is used to perform tSNE with prior information of the cluster of the virus sequences in order to reduce dimension into two. Thereafter, the reduced numerical dataset of virus sequences and the prior cluster information are used to find the medoid sequences of the clusters in euclidean space. The clusters are broadly in two different types where (a) $E^{v,\,m} = \{E_i^{v,\,m} | i = 1, \ldots, 3\}$ is set of medoid reference sequences of three virus wise clusters of SARS-CoV-1, MERS-CoV and SARS-CoV-2 for human host and (b) $E^{c,\,m} = \{E_i^{c,\,m} | i = 1, \ldots, 54\}$ is set of medoid reference sequence of country wise 54 clusters for human host respectively. In this regard, class type (a) is for all three virus sequences while class type (b) is for SARS-CoV-2 in human host only. Here, the euclidean distance based medoid of each cluster represents the reference genome sequence for its population. During the analysis, the similarities among reference sequences are computed based on pairwise alignment, because the size of the set of reference sequences is significantly less as compared to the total number of sequences. Thus, our technique is semi-alignment based while the whole experiment is called as a topological analysis. This technique covers all sequences in less computational time.

4. Results and discussion

Each cluster is an embedded representation of virus sequences as 2D data points generated by tSNE using k-mer generated descriptors and n-gram techniques, where $k = 3$ and $n = 4$ are fixed experimentally. Such embedded representation of clusters of SARS-CoV-1, MERS-CoV and SARS-CoV-2 sequences and SARS-CoV-2 sequences in human host of top 20 countries are shown visually in Figs. 4 (a) and (b). Each cluster is homogeneous in nature, which means the set of sequences represented by 2D data points within same cluster has maximum similarity. On the other hand, two different clusters consisting of sequences represented by 2D data points have less similarities. With these cluster data, $E^{v,\,m} = \{E_i^{v,\,m} | i = 1, \ldots, 3\}$ and $E^{c,\,m} = \{E_i^{c,\,m} | i = 1, \ldots, 54\}$ reference sequences are computed as mentioned in method section. These reference sequences are used for detail experimental analysis, which are described in following subsections. This is to be noted that as a comparative study, the result of alignment based technique are also compared with the result of our semi-alignment based technique.

4.1. *Coronavirus family specific analysis for human host*

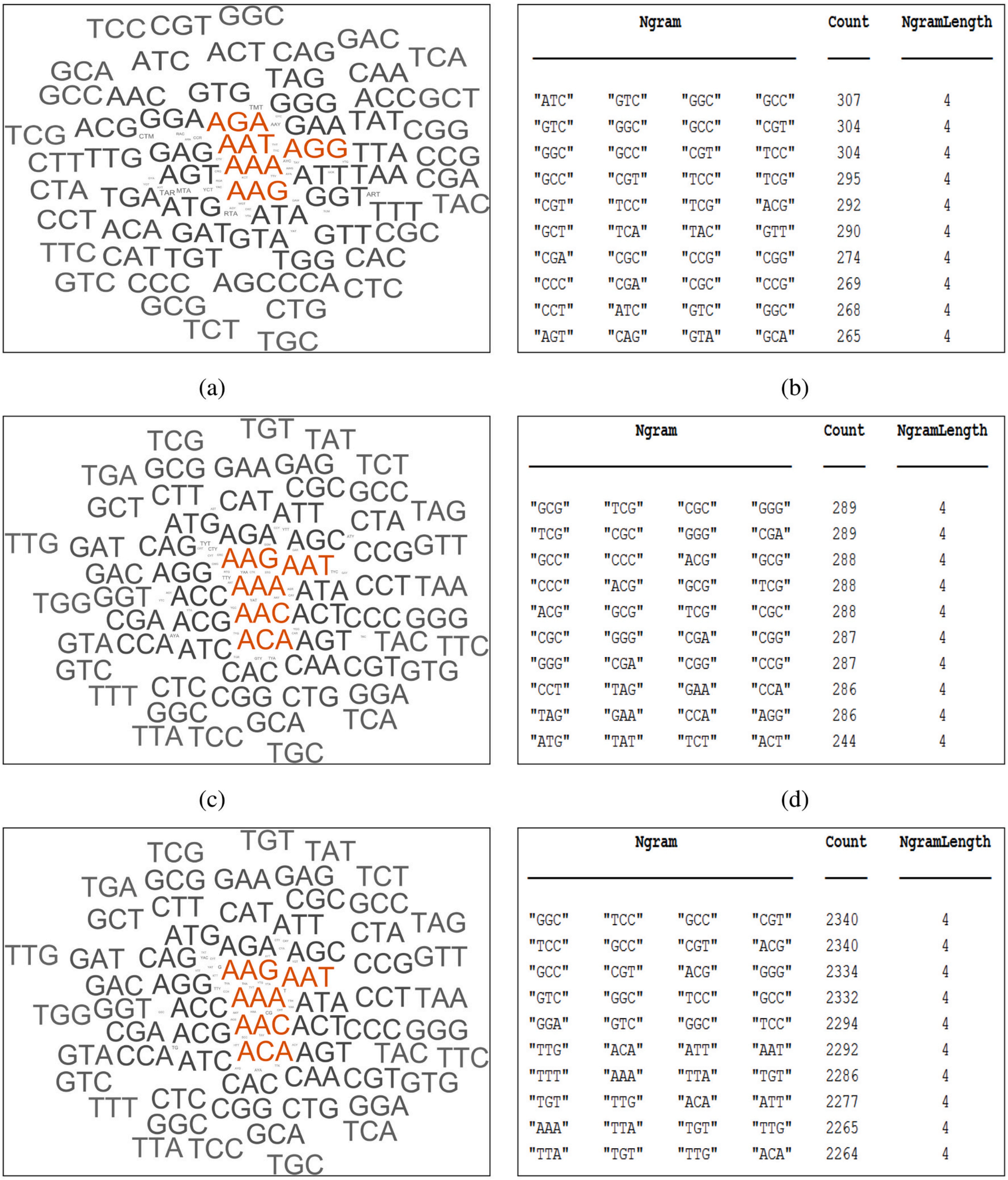
$E^{v,\,m}$ is set of three reference sequences of virus clusters such as SARS-CoV-1, MERS-CoV and SARS-CoV-2, where medoid of each cluster is considered as reference sequence. Table 3 and Fig. 5(a) show that medoid sequence similarities among SARS-CoV-1, MERS-CoV and SARS-CoV-2 computed using semi-alignment based technique, where it is observed that sequence of SARS-CoV-2 is approximately 76.59% similar to SARS-CoV-1 at the nucleotide level, and it is almost dissimilar to the sequence of MERS-CoV with only $\approx$ 36.09% similarity. The research study in (Lan et al., 2020; Wrapp et al., 2020) claim the suitability of SARS-CoV-1 and SARS-CoV-2 for binding with human ACE2 receptor. Hence approximately 76.59% similarity of SARS-CoV-2 with SARS-CoV-1 in our experiment also suggests the correctness of our approach in terms of medoid sequence similarity measures. In this regard, $\approx$ 23.41% dissimilarity between SARS-CoV-2 and SARS-CoV-1 may be the cause of wide spreading of SARS-CoV-2. Additionally, Table 4 and Fig. 5 (b) show the comparative observation of the heatmap results generated by alignment based technique respectively. The comparative observation also suggests the equivalence in results of both the techniques.

4.2. *Country specific SARS-CoV-2 analysis for human host*

In this analysis $E^{c,\,m}$, which is set of reference sequences for 54 countries of SARS-CoV-2 in human host is analyzed to understand the topological sequence variability of all against all nations. In this respect, an embedded representation of country wise SARS-CoV-2 sequence distribution in human host considering medoid data point as reference sequence of each country sequence population is shown in Figs. 6. Subsequently, a heatmap is generated using semi-alignment based technique and shown in Fig. 7(a), where it is observed that Brazil has different variant of sequence as compared to other countries. For further analysis, a bar plot for measuring similarity among country wise SARS-CoV-2 reference sequences as medoid ($E^{c,\,m}$) is also presented in Fig. 7 (b). Here, each country bar represent an aggregated (scaled in range [0,1]) similarity values of all other countries with respect to that particular country. For example, bar value of Australia represents that the Australia sequence population is $\approx$ 90% similar to other countries. As observed from Fig. 7(b), it is evident that almost all nations have maximum inter-country similarities in SARS-CoV-2 reference sequences, except Brazil, Colombia, Ecuador, Ireland and Wales which have comparatively less inter-country similarities in SARS-CoV-2 reference sequences. Moreover, Figs. 8(a) and (b) are generated using results produced by alignment based technique as to compare the similar results of Figs. 7(a) and (b) produced by semi-alignment based technique. In this case, both of the figures suggest that results produced by semi-alignment based technique is almost equivalent to the results produced by alignment based technique.

4.3. *Coronavirus family and country specific SARS-CoV-2 analysis for human host*

Additionally, set of reference sequences ($E^{c,\,m}$) for 54 countries of SARS-CoV-2 in human host is also analyzed with $E^{v,\,m}$, which is set of reference sequences as medoid of SARS-CoV-1, MERS-CoV and SARS-CoV-2 in human host. For this purpose, similarity measures as computed between $E^{v,\,m} = \{E_i^{v,\,m} | i = 1, \ldots, 3\}$ and $E^{c,\,m} = \{E_i^{c,\,m} | i = 1, \ldots, 54\}$ using semi-alignment based and alignment based techniques are reported in Tables 5 and 6 respectively. Subsequently, a visual



(b)

Ngram				Count	NgramLength
"ATC"	"GTC"	"GGC"	"GCC"	307	4
"GTC"	"GGC"	"GCC"	"CGT"	304	4
"GGC"	"GCC"	"CGT"	"TCC"	304	4
"GCC"	"CGT"	"TCC"	"TCG"	295	4
"CGT"	"TCC"	"TCG"	"ACG"	292	4
"GCT"	"TCA"	"TAC"	"GTT"	290	4
"CGA"	"CGC"	"CCG"	"CGG"	274	4
"CCC"	"CGA"	"CGC"	"CCG"	269	4
"CCT"	"ATC"	"GTC"	"GGC"	268	4
"AGT"	"CAG"	"GTA"	"GCA"	265	4

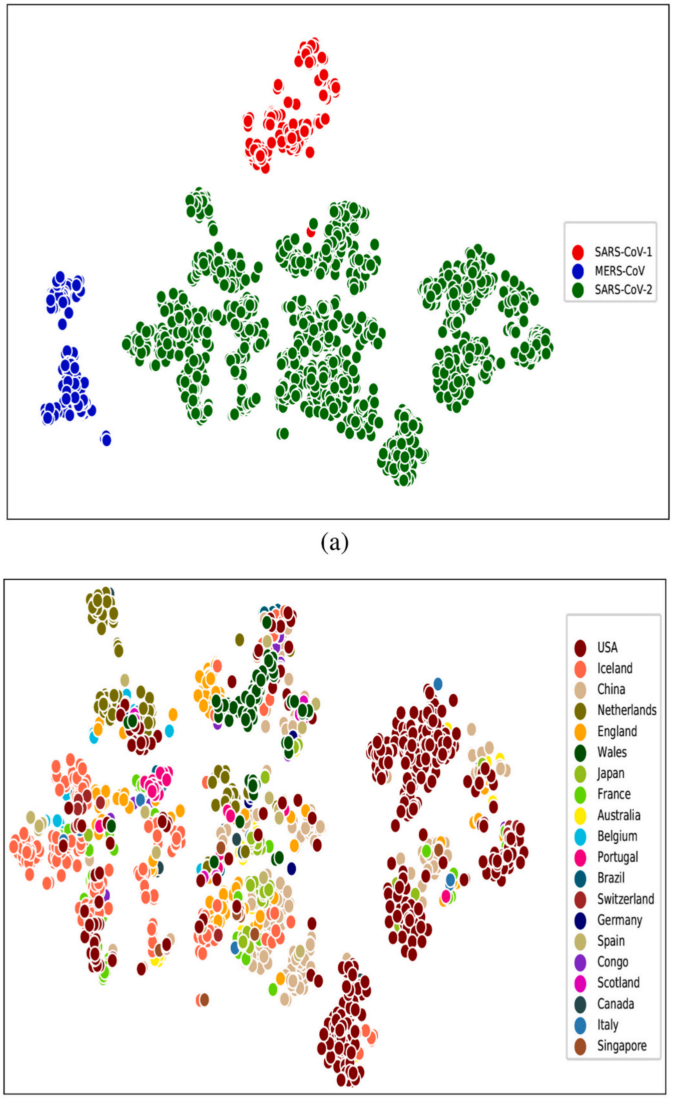
(d)

Ngram				Count	NgramLength
"GCG"	"TCG"	"CGC"	"GGG"	289	4
"TCG"	"CGC"	"GGG"	"CGA"	289	4
"GCC"	"CCC"	"ACG"	"GCG"	288	4
"CCC"	"ACG"	"GCG"	"TCG"	288	4
"ACG"	"GCG"	"TCG"	"CGC"	288	4
"CGC"	"GGG"	"CGA"	"CGG"	287	4
"GGG"	"CGA"	"CGG"	"CCG"	287	4
"CCT"	"TAG"	"GAA"	"CCA"	286	4
"TAG"	"GAA"	"CCA"	"AGG"	286	4
"ATG"	"TAT"	"TCT"	"ACT"	244	4

(f)

Ngram				Count	NgramLength
"GGC"	"TCC"	"GCC"	"CGT"	2340	4
"TCC"	"GCC"	"CGT"	"ACG"	2340	4
"GCC"	"CGT"	"ACG"	"GGG"	2334	4
"GTC"	"GGC"	"TCC"	"GCC"	2332	4
"GGA"	"GTC"	"GGC"	"TCC"	2294	4
"TTG"	"ACA"	"ATT"	"AAT"	2292	4
"TTT"	"AAA"	"TTA"	"TGT"	2286	4
"TGT"	"TTG"	"ACA"	"ATT"	2277	4
"AAA"	"TTA"	"TGT"	"TTG"	2265	4
"TTA"	"TGT"	"TTG"	"ACA"	2264	4

Fig. 3. Word cloud of k-mer ($k = 3$) generated descriptors for (a) SARS-CoV-1 (c) MERS-CoV and (e) SARS-CoV-2 sequences and top 10 n-grams ($n = 4$) of k-mer generated descriptors generated for (b) SARS-CoV-1 (d) MERS-CoV and (f) SARS-CoV-2 sequences in human host.



(a)

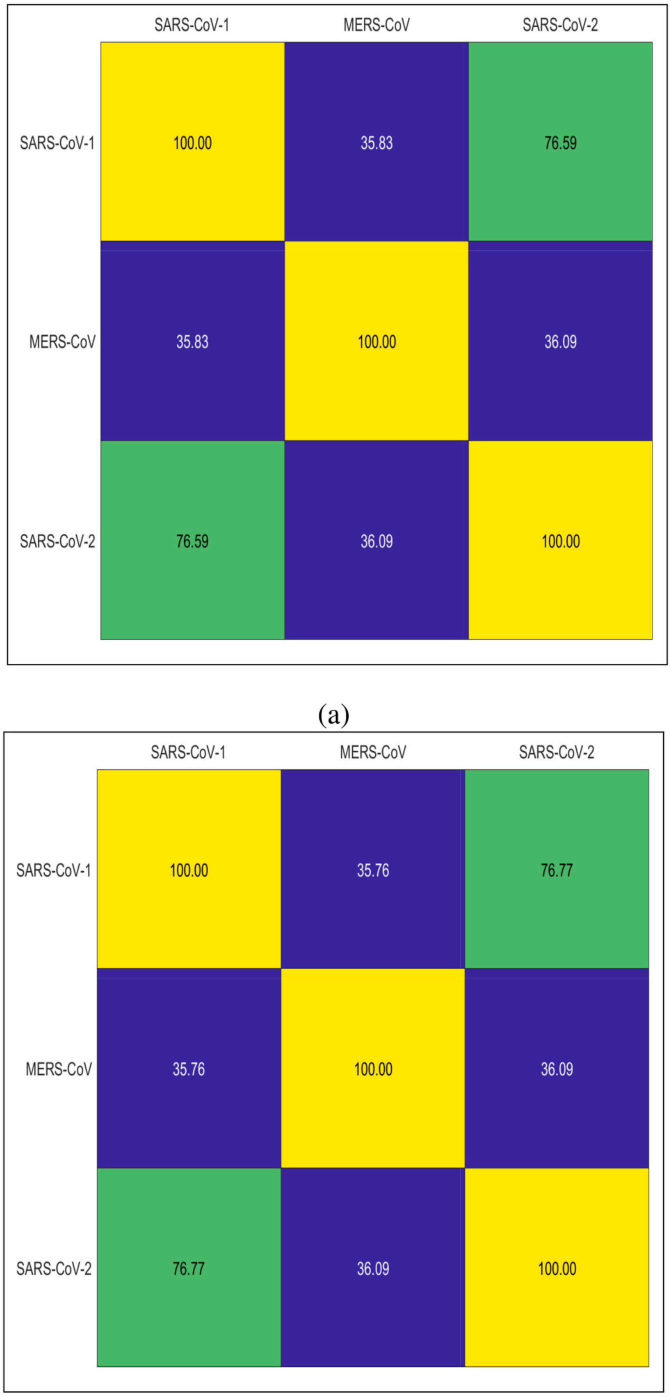
(b)

Fig. 4. Embedded representation of (a) SARS-CoV-1, MERS-CoV and SARS-CoV-2 sequences and (b) SARS-CoV-2 sequences in human host for top 20 countries.

representation of Table 5 is also shown as circos plot in Fig. 9. Interestingly, from both the Tables 5 and 6, the SARS-CoV-2 reference sequence of Brazil is found little different than reference sequence of SARS-CoV-2 family. Rather, the reference sequence of Brazil has comparatively high similarity with reference sequence of SARS-CoV-1 family. For rest of the nations, it is observed that sequences of most of the nations are very much close to sequence of SARS-CoV-2 family. For example, Australia, Belgium, Congo, India, Italy, USA etc. have ≈ 99.94%, ≈ 99.85%, ≈ 98.98%, ≈ 99.75%, ≈ 99.83%, ≈ 98.97% similarity respectively.

Table 3
Similarity measure among virus reference sequences as medoid ($E^{v,\,m}$) of SARS-CoV-1, MERS-CoV and SARS-CoV-2 in human host using semi-alignment based technique.

Virus in human host	SARS-CoV-1	MERS-CoV	SARS-CoV-2
SARS-CoV-1	100.00	35.83	76.59
MERS-CoV	35.83	100.00	36.09
SARS-CoV-2	76.59	36.09	100.00



(a)

(b)

Fig. 5. Heatmap of results produced by (a) semi-alignment based technique and (b) alignment based technique, for measuring similarity among reference sequences as medoid ($E^{v,\,m}$) of SARS-CoV-1, MERS-CoV and SARS-CoV-2 in human host.

Table 4
Similarity measure among virus reference sequences as medoid ($E^{v,\,m}$) of SARS-CoV-1, MERS-CoV and SARS-CoV-2 in human host using alignment based technique.

Virus in human host	SARS-CoV-1	MERS-CoV	SARS-CoV-2
SARS-CoV-1	100.00	35.75	76.77
MERS-CoV	35.75	100.00	36.09
SARS-CoV-2	76.77	36.09	100.00

4.4. *Comparative analysis between semi-alignment based technique and alignment based technique*

While entire analysis is reported with the result generated using semi-alignment based technique, we have also performed the experiment using alignment based technique to understand the equivalence of the outcome between these two techniques. Therefore, it can also establish the fact that the reference sequences that are identified using semi-alignment based and alignment based techniques are equivalent. For this purpose, we have performed two-sample Kolmogorov-Smirnov (KS) (Massey, 1951) test. In this regard, the KS test is performed on the null hypothesis that the sequence similarity results produced by both semi-alignment based and alignment based techniques are same with 5% significant level. This suggests that the null hypothesis is accepted when *p*-value is greater than 0.05. The mean *p*-values of test result are reported in Fig. 10, while the detail results produced by both the techniques are given in supplementary material. From Fig. 10 bar (A) it is evident that p-value of sequence similarity produced by semi-alignment based technique as reported in Table 3 and alignment based technique as reported in Table 4 for intra-virus classes among SARS-CoV-1, MERS-CoV and SARS-CoV-2 in human host is 0.97 which is much greater than 0.05 and signifies that results produced by both the techniques are equivalent. Similarly, p-values of KS test between Fig. 7(a) and Fig. 8(a) and Tables 5 and 6 generated by semi-alignment based and alignment based techniques are 0.40 and 0.97 respectively, which prove that results generated by both the techniques are similar.

Moreover, to compute sequence similarity after performing sequence alignment using Clustal Omega, it takes around 2 days in Intel Core i5-2410M CPU at 2.30 GHz machine with 8GB RAM, whereas with the same configuration of machine it takes less than an hour for complete analysis using semi-alignment based technique. Thus, semi-alignment based technique is found computationally economical and effective for fast outcome.

4.5. *Summary of outcome*

After performing topological experiments for (a) coronavirus family specific analysis for human host, (b) country specific SARS-CoV-2 analysis for human host and (c) coronavirus family and country specific SARS-CoV-2 analysis for human host, we can summarize the findings from our case study and possible suggestions as follows.

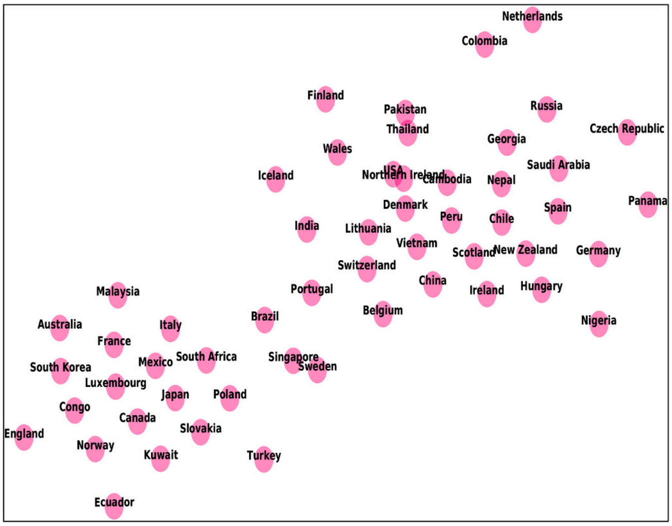


Fig. 6. Embedded representation of country wise reference sequences as medoid ($E^{c,\ m}$) of SARS-CoV-2 in human host.

- SARS-CoV-2 sequence in human host is found approximately 76.59% similar to SARS-CoV-1 sequence and almost dissimilar to MERS-CoV in human host. Moreover, approximately 23.41% dissimilarity between SARS-CoV-2 and SARS-CoV-1 sequences might be the reason of wide spreading capability of SARS-CoV-2.
- It is found that certain nations are having very high sequence similarities, while analyzing inter-country sequence of SARS-CoV-2. Therefore, nations with similar sequences might think of using similar vaccine or drug. There are certain nations like Brazil, Ecuador, Iceland etc. should be paid more attention for detail genetic analysis as these nations have little different sequence than other nations.
- SARS-CoV-2 Sequences of all nations are approximately 99% similar to sequence of SARS-CoV-2 family, whereas approximately 76% and 36% similar to SARS-CoV-1 and MERS-CoV respectively in coronavirus family. However, Brazil is found as little exceptional, where SARS-CoV-2 sequence of Brazil is less similar to sequences of SARS-CoV-2 family. Instead, SARS-CoV-2 sequence of Brazil is approximately 65.40% similar to SARS-CoV-1 family.
- We have used both computationally expensive sequence alignment based technique as well as less computationally expensive semi-alignment based technique for sequence variability analysis. It is experimentally found that outcomes of both the techniques are close to each other. Therefore, semi-alignment based technique can be used to speed up the research on SARS-CoV-2 in critical situation for designing prophylactic vaccine and therapeutic drug (Dey et al., 2017; Nandy and Basak, 2016) of SARS-CoV-2.

5. Conclusions

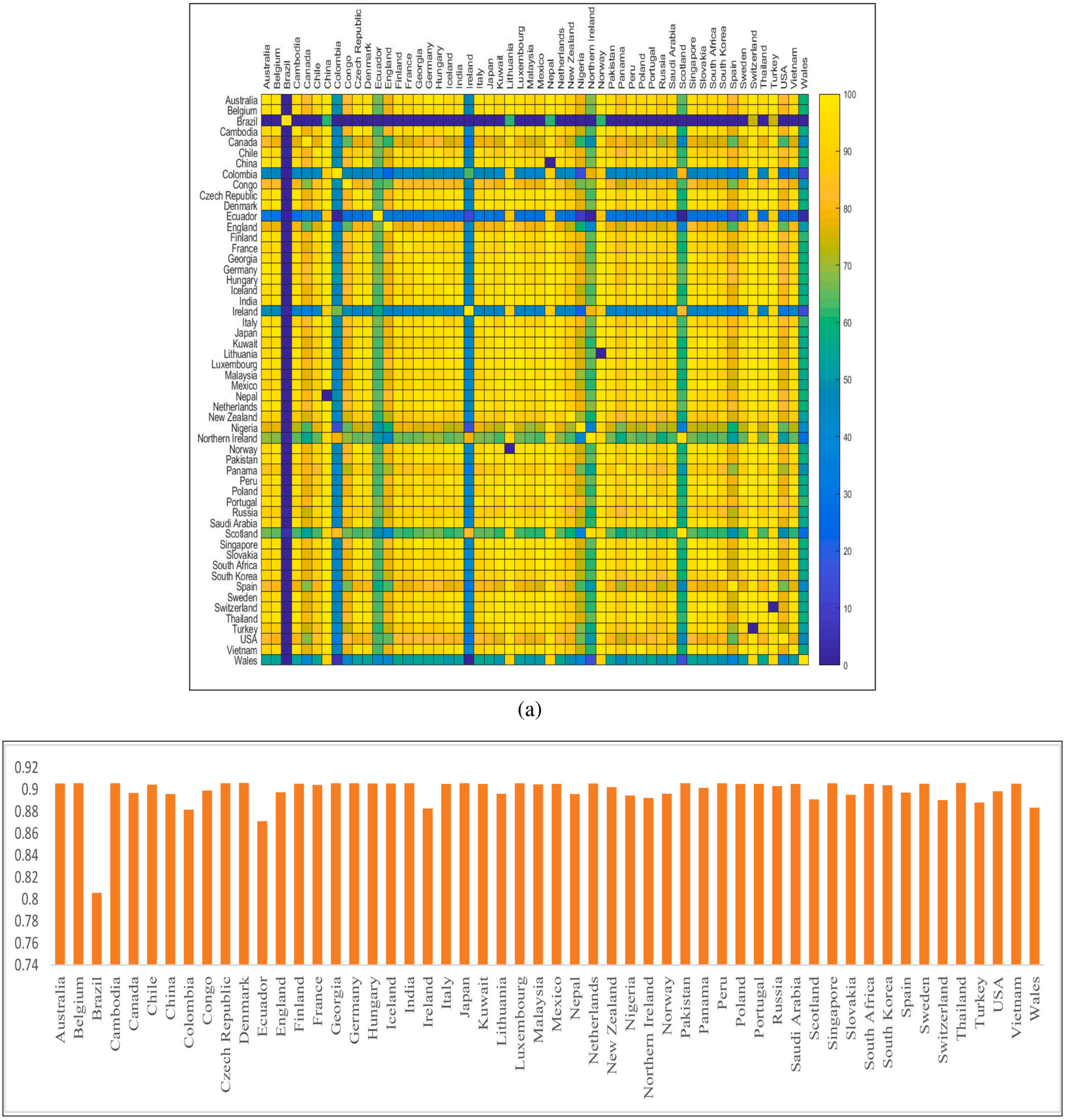
In world health context, COVID-19 disease caused by SARS-CoV-2 has created a pandemic in human population with large number of infection. It is found that person-to-person transmission plays a crucial role in spreading of the COVID-19 disease. Various regulating measures, such as lockdown, are taken worldwide to restrict the transmission. Various researches are being conducted across globe to find appropriate vaccine and drug. Therefore, in this article, we have performed a case study on the genome sequence variability for global health using semi-alignment based technique. We also have conducted experiments using sequence alignment based technique, where it has been found that less computationally expensive semi-alignment based technique has produced results as equivalent as expensive alignment based technique. We have studied primarily focusing on three aspects such as (a) coronavirus family specific sequence variability for human host, (b) country specific sequence variability of SARS-CoV-2 of 54 nations for human host and (c) coronavirus family and country specific SARS-CoV-2 sequence variability for human host. In order to perform the analysis, we have computed medoid reference sequence for each of the categories. It has been observed that the reference sequences across different countries are quite similar with certain exceptional cases like Brazil, Iceland etc. For these exceptional cases, our future research is focused on detail genomic study of SARS-CoV-2 to understand the genome-wide variability for global health. Moreover, our future research will also focus to explore the application of pure alignment free based technique effectively in order to overcome the limitation of current article where we have applied semi-alignment based technique.

Ethics approval and consent to participate

The ethical approval or individual consent was not applicable.

Availability of data and materials

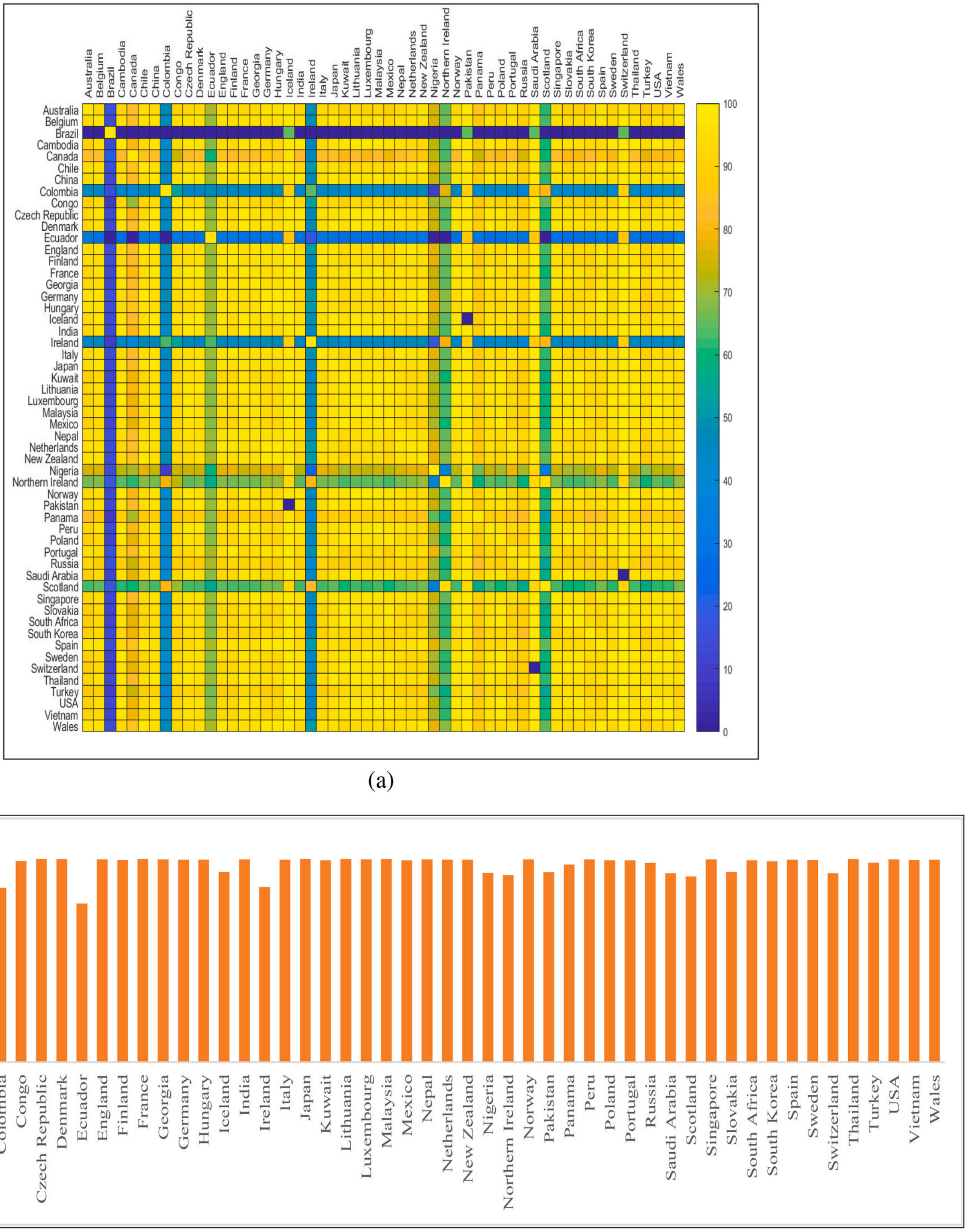
The datasets of SARS-CoV-1, MERS-CoV and SARS-CoV-2 sequences used in our case study, reference sequences identified as mediod in case of virus wise and country wise analysis and software are available at



(a)

(b)

Fig. 7. (a) Heatmap (b) Bar plot representation of aggregated results produced by semi-alignment based technique for measuring similarity among country wise SARS-CoV-2 reference sequences as medoid ($E^{c,\ m}$) in human host.



(a)

(b)

Fig. 8. (a) Heatmap (b) Bar plot representation of aggregated results produced by alignment based technique for measuring similarity among country wise SARS-CoV-2 reference sequences as medoid ($E^{c,\ m}$) in human host.

Table 5

Similarity measure between virus reference sequences as medoid ($E^{v,\ m}$) of SARS-CoV-1, MERS-CoV and SARS-CoV-2 in human host with country wise SARS-CoV-2 reference sequences as medoid ($E^{c,\ m}$) in human host using semi-alignment based technique.

Country	SARS-CoV-1	MERS-CoV	SARS-CoV-2	Country	SARS-CoV-1	MERS-CoV	SARS-CoV-2
Australia	76.55	36.11	99.94	Mexico	76.62	35.93	99.57
Belgium	76.69	36.21	99.85	Nepal	76.66	36.11	99.80
Brazil	65.40	32.53	88.46	Netherlands	76.57	36.16	99.88
Cambodia	76.71	36.11	99.69	New Zealand	76.17	36.05	99.56
Canada	75.70	35.26	98.92	Nigeria	75.53	36.17	98.67
Chile	76.32	35.99	99.76	Northern Ireland	74.57	36.26	98.21
China	76.65	36.08	99.88	Norway	76.75	36.12	99.74
Colombia	73.13	35.98	97.01	Pakistan	76.71	36.16	99.80
Congo	75.68	35.99	98.98	Panama	76.29	35.11	99.21
Czech Republic	76.71	36.13	99.69	Peru	76.75	36.14	99.72
Denmark	76.75	36.23	99.81	Poland	76.64	35.84	99.58
Ecuador	72.66	31.33	96.04	Portugal	76.44	36.25	99.78
England	75.87	35.57	98.82	Russia	76.18	36.10	99.40
Finland	76.53	36.25	97.58	Saudi Arabia	76.62	35.90	99.58
France	76.53	36.13	99.63	Scotland	74.49	36.36	98.07
Georgia	76.65	36.10	99.80	Singapore	76.68	36.02	99.79
Germany	76.57	36.08	99.96	Slovakia	76.63	35.93	99.58
Hungary	76.56	36.13	99.94	South Africa	76.63	35.93	99.57
Iceland	76.68	36.11	99.67	South Korea	76.44	36.08	99.68
India	76.72	36.17	99.75	Spain	75.90	36.18	99.00
Ireland	73.50	35.21	97.05	Sweden	76.67	35.92	99.63
Italy	76.45	36.27	99.83	Switzerland	76.67	36.01	99.61
Japan	76.77	36.18	99.83	Thailand	76.78	36.17	99.78
Kuwait	76.67	36.00	99.60	Turkey	76.45	36.03	99.35
Lithuania	76.75	36.15	99.72	USA	75.59	35.86	98.97
Luxembourg	76.76	36.09	99.77	Vietnam	76.68	35.97	99.62
Malaysia	76.64	35.98	99.62	Wales	74.01	35.94	97.33

Table 6

Similarity measure between virus reference sequences as medoid ($E^{v,\ m}$) of SARS-CoV-1, MERS-CoV and SARS-CoV-2 in human host with country wise SARS-CoV-2 reference sequences as medoid ($E^{c,\ m}$) in human host using alignment based technique.

Country	SARS-CoV-1	MERS-CoV	SARS-CoV-2	Country	SARS-CoV-1	MERS-CoV	SARS-CoV-2
Australia	76.38	35.87	99.67	Mexico	76.76	35.82	99.74
Belgium	76.68	36.11	99.91	Nepal	76.63	36.01	99.81
Brazil	66.04	34.59	88.47	Netherlands	76.54	36.13	99.80
Cambodia	76.87	36.01	99.86	New Zealand	76.53	36.03	99.80
Canada	76.09	35.89	99.06	Nigeria	75.51	36.12	98.56
Chile	76.56	36.14	99.82	Northern Ireland	74.55	36.34	98.07
China	76.78	36.07	99.96	Norway	76.87	36.11	99.88
Colombia	73.11	35.91	96.86	Pakistan	76.77	36.07	99.93
Congo	76.53	36.10	99.71	Panama	76.41	34.99	99.37
Czech Republic	76.81	36.03	99.90	Peru	76.81	36.04	99.89
Denmark	76.77	36.08	99.95	Poland	76.77	35.73	99.75
Ecuador	72.66	31.29	95.91	Portugal	76.46	36.15	99.74
England	76.60	36.17	99.85	Russia	76.26	36.00	99.41
Finland	76.53	36.17	99.77	Saudi Arabia	76.77	35.80	99.75
France	76.78	36.05	99.91	Scotland	74.48	36.22	97.92
Georgia	76.72	36.05	36.40	Singapore	76.77	36.01	99.83
Germany	76.53	36.10	99.80	Slovakia	76.76	35.89	99.75
Hungary	76.52	36.05	99.78	South Africa	76.77	35.82	99.74
Iceland	76.76	36.09	99.97	South Korea	76.48	35.93	99.69
India	76.82	36.02	99.89	Spain	76.53	36.03	99.79
Ireland	73.48	35.09	96.90	Sweden	76.81	35.81	99.79
Italy	76.56	36.15	99.79	Switzerland	76.75	35.90	99.75
Japan	76.76	36.08	99.99	Thailand	76.80	36.06	99.91
Kuwait	76.75	35.89	99.77	Turkey	76.58	35.90	99.52
Lithuania	76.84	36.04	99.89	USA	76.85	35.81	99.83
Luxembourg	76.86	35.82	99.83	Vietnam	76.79	35.86	99.78
Malaysia	76.85	36.06	99.89	Wales	76.55	35.95	99.80

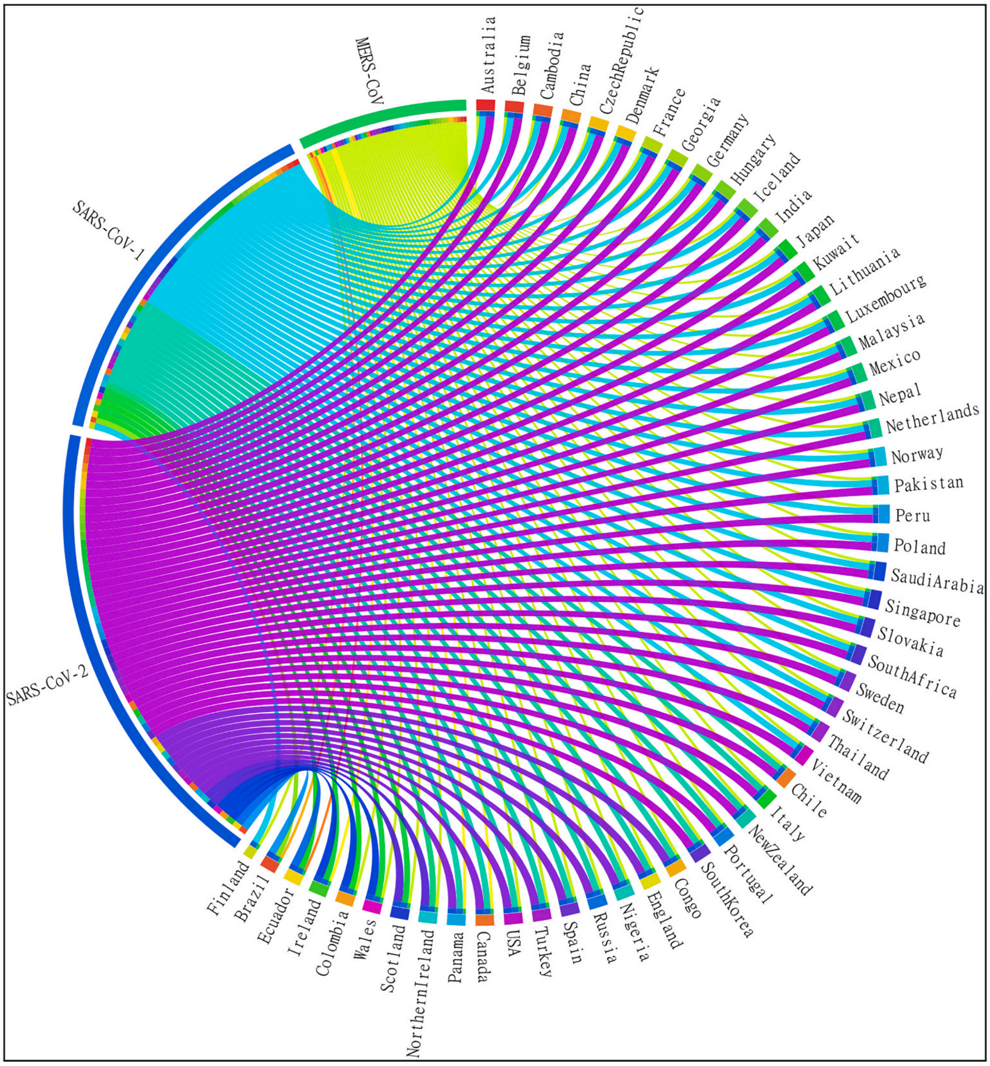


Fig. 9. Circos plot of similarity measure between country wise SARS-CoV-2 reference sequences as medoid ($E^{c,\ m}$) in human host with virus reference sequences as medoid ($E^{v,\ m}$) of SARS-CoV-1, MERS and SARS-CoV-2 in human host using semi-alignment based technique.

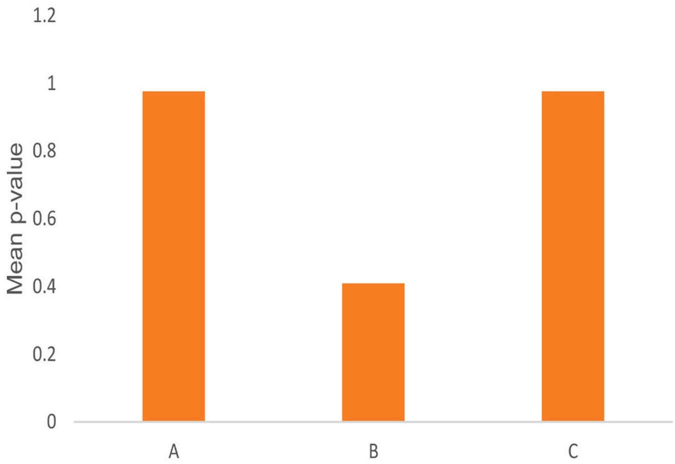


Fig. 10. Mean p-value of two-sample Kolmogorov-Smirnov test between genome sequence similarity results produced by semi-alignment based technique and alignment based technique for (A) virus reference sequences as medoid ($E^{v,\ m}$) in human host (B) country wise SARS-CoV-2 reference sequences as medoid ($E^{c,\ m}$) in human host (C) virus reference sequences as medoid ($E^{v,\ m}$) of SARS-CoV-1, MERS-CoV and SARS-CoV-2 in human host with country wise SARS-CoV-2 reference sequences as medoid ($E^{c,\ m}$) in human host.

“http://www.nitttrkol.ac.in/indrajit/projects/COVID-TopologicalAnalysis-SequenceVariability/”. Moreover, all the virus sequences used in this work are publicly available at NCBI and GISAID databases.

Consent for publication

Not applicable.

Funding

This work has been partially supported by CRG short term research grant on COVID-19 (CVD/2020/000991) from Science and Engineering Research Board (SERB), Department of Science and Technology, Govt. of India.

CRediT authorship contribution statement

Jnanendra Prasad Sarkar: Conceptualization, Formal analysis, Validation, Visualization, Writing - original draft. Indrajit Saha: Conceptualization, Data curation, Supervision, Funding acquisition, Software, Formal analysis, Investigation, Methodology, Web development, Project administration, Resources, Validation, Visualization, Writing - review & editing. Arijit Seal: Conceptualization, Formal analysis, Writing - review & editing. Debasree Maity: Conceptualization, Data curation, Methodology, Writing - review & editing. Ujjwal Maulik: Conceptualization, Methodology, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgment

We thank all those who have contributed sequences to NCBI and GISAID databases. Also thankful to the reviewers for valuable comments to improve the article.

References

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., Garry, R.F., 2020. The proximal origin of SARS-CoV-2. Nat. Med. https://doi.org/10.1038/s41591-020-0820-9.

Bahl, S., Javaid, M., Bagha, A.K., Singh, R.P., Haleem, A., Vaishya, R., Suman, R., 2020. Biosensors applications in fighting COVID-19 pandemic. Apollo Med. 17, 221–223. https://doi.org/10.4103/am.am_56_20.

Dey, S., Nandy, A., Basak, S.C., Nandy, P., Das, S., 2017. A bioinformatics approach to designing a zika virus vaccine. Comput. Biol. Chem. 68, 143–152. https://doi.org/10.1016/j.compbiolchem.2017.03.002.

Dey, A., Sen, S., Maulik, U., 2020. Unveiling COVID-19-associated organ-specific cell types and cell-specific pathway cascade. Brief. Bioinform. https://doi.org/10.1093/bib/bbaa214.

Eddy, S.R., 1998. Profile hidden markov models. Bioinformatics 14, 755–763. https://doi.org/10.1093/bioinformatics/14.9.755.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. https://doi.org/10.1093/nar/gkh340.

Haleem, A., Javaid, M., Vaishya, R., 2020a. Effects of COVID-19 pandemic in daily life. Curr. Med. Res. Pract. 10, 78–79. https://doi.org/10.1016/j.cmrp.2020.03.011.

Haleem, A., Javaid, M., Vaishya, R., Deshmukh, S.G., 2020b. Areas of academic research with the impact of COVID-19. Am. J. Emerg. Med. 38, 1524–1526. https://doi.org/10.1016/j.ajem.2020.04.022.

Hinton, G., Roweis, S.T., 2002. Stochastic neighbor embedding. In: Proceedings of the 15th International Conference on Neural Information Processing Systems, pp. 857–864. https://www.cs.toronto.edu/%20fritz/absps/sne.pdf.

Javaid, M., Haleem, A., Vaishya, R., Bahl, S., Suman, R., Vaish, A., 2020. Industry 4.0 technologies and their applications in fighting COVID-19 pandemic. Diab. Metab. Syndr. Clin. Res. Rev. 14, 419–422. https://doi.org/10.1016/j.dsx.2020.04.032.

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., Wang, X., 2020. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature 581, 215–220. https://doi.org/10.1038/s41586-020-2180-5.

Manekar, S.C., Sathe, S.R., 2018. A benchmark study of k-mer counting methods for high-throughput sequencing. GigaScience 7, 1–13. https://doi.org/10.1093/gigascience/giy125.

Massey, F.J., 1951. The Kolmogorov-Smirnov test for goodness of fit. J. Am. Stat. Assoc. 46, 68–78. https://doi.org/10.2307/2280095.

Melsted, P., Pritchard, J.K., 2011. Efficient counting of k-mers in DNA sequences using a bloom filter. BMC Bioinformatics 12. https://doi.org/10.1186/1471-2105-12-333.

Nandy, A., Basak, S.C., 2016. A brief review of computer-assisted approaches to rational design of peptide vaccines. Int. J. Mol. Sci. 17 https://doi.org/10.3390/ijms17050666.

Pineda-Pena, A.C., Faria, N.R., Imbrechts, S., Libin, P., Abecasis, A.B., Deforche, K., Arley, G.L., Camacho, R.J., Oliveira, T.D., Vandamme, A.M., 2013. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new rega version 3 and seven other tools. Infect. Genet. Evol. 19, 337–348. https://doi.org/10.1016/j.meegid.2013.04.032.

Pond, S.L.K., Posada, D., Stawiski, E., Chappey, C., Poon, A.F., Hughes, G., Fearnhill, E., Gravenor, M.B., Brown, A.J.L., Frost, S.D.W., 2009. An evolutionary modelbased algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. PLoS Comput. Biol. 5, e1000581 https://doi.org/10.1371/journal.pcbi.1000581.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A., Finn, R.D., 2012. The pfam protein families database. Nucleic Acids Res. 40, D290–D301. https://doi.org/10.1093/nar/gkr1065.

Saha, I., Ghosh, N., Maity, D., Sharma, N., Mitra, K., 2020a. Inferring the genetic variability in Indian SARS-CoV-2 genomes using consensus of multiple sequence alignment techniques. Infect. Genet. Evol. 85 https://doi.org/10.1016/j.meegid.2020.104522.

Saha, I., Ghosh, N., Maity, D., Sharma, N., Sarkar, J.P., Mitra, K., 2020b. Genome-wide analysis of Indian SARS-CoV-2 genomes for the identification of genetic mutation and SNP. Infect. Genet. Evol. 85 https://doi.org/10.1016/j.meegid.2020.104457.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D., Higgins, D.G., 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.

Singh, R.P., Javaid, M., Haleem, A., Suman, R., 2020a. Internet of things (IoT) applications to fight against COVID-19 pandemic. Diab. Metab. Syndr. Clin. Res. Rev. 14, 521–524. https://doi.org/10.1016/j.dsx.2020.04.041.

Singh, R.P., Javaid, M., Haleem, A., Vaishya, R., Ali, S., 2020b. Internet of Medical Things (IoMT) for orthopaedic in COVID-19 pandemic: roles, challenges, and applications. J. Clin. Orthopaed. Trauma 11, 713–717. https://doi.org/10.1016/j.jcot.2020.05.011.

Singh, R.P., Javaid, M., Kataria, R., Tyagi, M., Haleem, A., 2020c. Significant applications of virtual reality for COVID-19 pandemic. Diab. Metab. Syndr. Clinical Res. Rev. 14, 661–664. https://doi.org/10.1016/j.dsx.2020.05.011.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197. https://doi.org/10.1016/0022-2836(81)90087-5.

Solis-Reyes, S., Avino, M., Poon, A., Kari, L., 2018. An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. PLoS One 13, e0206409. https://doi.org/10.1371/journal.pone.0206409.

Song, B., Choi, J.-H., Chen, G., Szymanski, J., Zhang, G.-Q., Tung, A.K.H., Kang, J., Kim, S., Yang, J., 2006. ARCS: an aggregated related column scoring scheme for aligned sequences. Bioinformatics 22, 2326–2332. https://doi.org/10.1093/bioinformatics/btl398.

Struck, D., Lawyer, G., Ternes, A.M., Schmit, J.C., Bercoff, D.P., 2014. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. Nucleic Acids Res. 42, e144. https://doi.org/10.1093/nar/gku739.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680. https://doi.org/10.1093/nar/22.22.4673.

Vaishya, R., Haleem, A., Vaish, A., Javaid, M., 2020a. Emerging technologies to combat the COVID-19 pandemic. J. Clin. Exp. Hepatol. 17, 221–223. https://doi.org/10.1016/j.jceh.2020.04.019.

Vaishya, R., Javaid, M., Khan, I.H., Haleem, A., 2020b. Artificial Intelligence (AI) applications for COVID-19 pandemic. Diab. Metab. Syndr. Clin. Res. Rev. 14, 337–339. https://doi.org/10.1016/j.dsx.2020.04.012.

Wan, Y., Shang, J., Graham, R., Baric, R.S., Li, F., 2020. Receptor recognition by the novel coronavirus from wuhan: an analysis based on decade-long structural studies of SARS coronavirus. J. Virol. 94. https://doi.org/10.1128/JVI.00127-20.

WHO (2020). Coronavirus Disease (COVID-19) Pandemic. World Health Organization*, Western Pacific China*,. URL: https://www.who.int/china.

Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.L., Abiona, O., Graham, B. S., McLellan, J.S., 2020. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 367, 1260–1263. https://doi.org/10.1126/science.abb2507.

Zhang, Y.-Z., Holmes, E.C., 2020. A genomic perspective on the origin and emergence of SARS-CoV-2. Cell. https://doi.org/10.1016/j.cell.2020.03.035.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., Zheng, X.-S., Zhao, K., Chen, Q.-J., Deng, F., Liu, L.-L., Yan, B., Zhan, F.-X., Wang, Y.-Y., Xiao, G.-F., Shi, Z.-L., 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.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F., Tan, W., 2020. A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med. 382, 727–733. https://doi.org/10.1056/NEJMoa2001017.