Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Methods 203 (2022) 282–296

Contents lists available at ScienceDirect

Methods

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

Characterisation of SARS-CoV-2 clades based on signature SNPs unveils continuous evolution

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

[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] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland
[c] Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India
[d] Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India

ARTICLE INFO

Keywords:
COVID-19
Clade
Deleterious mutations
Non-Synonymous Signature SNP
SARS-CoV-2

ABSTRACT

Since the emergence of SARS-CoV-2 in Wuhan, China more than a year ago, it has spread across the world in a very short span of time. Although, different forms of vaccines are being rolled out for vaccination programs around the globe, the mutation of the virus is still a cause of concern among the research communities. Hence, it is important to study the constantly evolving virus and its strains in order to provide a much more stable form of cure. This fact motivated us to conduct this research where we have initially carried out multiple sequence alignment of 15359 and 3033 global dataset without Indian and the dataset of exclusive Indian SARS-CoV-2 genomes respectively, using MAFFT. Subsequently, phylogenetic analyses are performed using Nextstrain to identify virus clades. Consequently, the virus strains are found to be distributed among 5 major clades or clusters viz. 19A, 19B, 20A, 20B and 20C. Thereafter, mutation points as SNPs are identified in each clade. Henceforth, from each clade top 10 signature SNPs are identified based on their frequency i.e. number of occurrences in the virus genome. As a result, 50 such signature SNPs are individually identified for global dataset without Indian and dataset of exclusive Indian SARS-CoV-2 genomes respectively. Out of each 50 signature SNPs, 39 and 41 unique SNPs are identified among which 25 non-synonymous signature SNPs (out of 39) resulted in 30 amino acid changes in protein while 27 changes in amino acid are identified from 22 non-synonymous signature SNPs (out of 41). These 30 and 27 amino acid changes for the non-synonymous signature SNPs are visualised in their respective protein structure as well. Finally, in order to judge the characteristics of the identified clades, the non-synonymous signature SNPs are considered to evaluate the changes in proteins as biological functions with the sequences using PROVEAN and PolyPhen-2 while I-Mutant 2.0 is used to evaluate their structural stability. As a consequence, for global dataset without Indian sequences, G251V in ORF3a in clade 19A, F308Y and G196V in NSP4 and ORF3a in 19B are the unique amino acid changes which are responsible for defining each clade as they are all deleterious and unstable. Such changes which are common for both global dataset without Indian and dataset of exclusive Indian sequences are R203M in Nucleocapsid for 20B, T85I and Q57H in NSP2 and ORF3a respectively for 20C while for exclusive Indian sequences such unique changes are A97V in RdRp, G339S and G339C in NSP2 in 19A and Q57H in ORF3a in 20A.

1. Introduction

The first case of SARS-CoV-2 was registered in Wuhan China, 2019 and it quickly took over the normal functioning of human lives. In the initial phases, lock-down was implemented to limit the spread of infection. It is well known that virus mutations take place in the form of single nucleotide variants, deletions and large structural variants [1] mainly due to replication and some hotspot mutations having severe impact on the host. Many cities around the globe have gone through staggered phases of lockdown in order to avoid the spread of the different strains of SARS-CoV-2. Among these strains, B.1.1.7 (Alpha) is found to be highly transmissible [2] and causes more severe pathogenic

* Corresponding author.
E-mail address: indrajit@nitttrkol.ac.in (I. Saha).
[1] Equally contributed.

https://doi.org/10.1016/j.ymeth.2021.09.005
Received 9 April 2021; Received in revised form 3 September 2021; Accepted 15 September 2021
Available online 20 September 2021

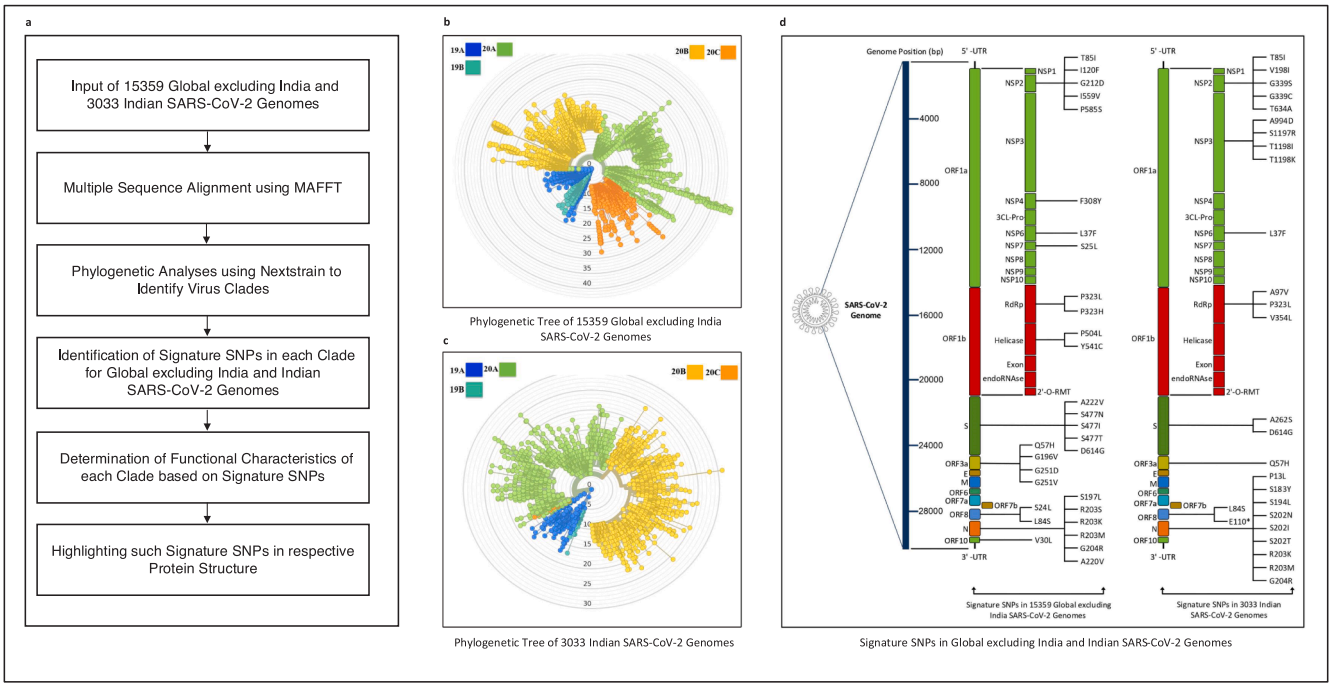

Fig. 1. Pipeline of the work.

infection in young people [3]. Another major variant B.1.351 (Beta) which has emerged from South Africa [4,5] had also led to a sudden surge in the total number of cases. The efficacy of therapeutic monoclonal antibodies (mAbs) are known to be reduced against another variant P.1 (Gamma). It is estimated to be 2.6 times more transmissible [6]. Another variant B.1.617.2 (Delta) is known for the surge in cases in India during the 2nd wave of the pandemic.

SARS-CoV-2 is a 29.9 kb long single-stranded genomic RNA [7–12]. It covers 11 coding regions where ORF1ab occupies majority of the genomic sequence while Spike (S), ORF3a, Envelope, Membrane, ORF6, ORF7a, ORF7b, ORF8, Nucleocapsid and ORF10 constitute the rest of the sequence [8,9,13]. Through various studies it is found that the South African B.1.351 variant strain consists of mutation in three prominent places in Spike (S) protein, they being K417N, E484K and N501Y [4] whereas the UK variant B.1.1.7 which was found to be part of the 20B clade contains multiple mutations with a combination of N501Y in Spike (S) protein [2] and the 69-70del which have been circulating within the community for months. Hence, it is incumbent that such frequent mutations be given special focus by the scientific community to trace and tackle the challenges posed by the mutations.

Tang et al. [14] investigated the extent of molecular divergence between SARS-CoV-2 and other related coronaviruses by analysing 103 SARS-CoV-2 genomes and reported two major lineages, L and S. Several other mutations are also identified in the last few months which demands re-purposing of the current methods to deal with the virus. Wang et al. [15] have proposed a *h*-index mutation ratio criteria to evaluate the non-conserved and conserved proteins with the help of more than 15000 sequences. Consequently, nucleocapsid, spike and papain-like protease are found to be highly non-conserved while envelope, main protease, and endoribonuclease protein are relatively conservative. They have further identified the mutations on 40% of nucleotides in nucleocapsid, thereby indicating potential impacts on the ongoing development of various COVID-19 diagnosis and cure. Such similar analysis conducted by Yuan et al. [16] with 11183 sequences revealed 119 high frequency substitutions or Single Nucleotide Polymorphism (SNP) around the globe. Among the nucleotide changes in SNPs, C to T is the major one indicating adaptation and evolution of the virus in the human host which can pose new challenges. On the other hand, Chen et al. [17] focused on the binding of free energy changes between the angiotensin converting enzyme 2 (ACE2) receptor with the frequently changing Spike protein of SARS-CoV-2 considering algebraic topology-based machine learning model and found 3 sub-type of the virus with slightly high infectivity. 570 SARS-CoV-2 sequences were analysed and 10 distinct hotspot mutations points from China, India, USA, Europe which might affect the replication-relevant proteins were identified by Weber et al. [18]. Further, they found that these mutations can effect the secondary structure of the RNA molecule of SARS-CoV-2 and its repertoire which are essential for viral and cellular proteins. Moreover, Nagy et al. [19] computed the direct effect of mutations over clinical outcome with the help of Chi-square test, in which they found mutations in ORF8, NSP6, ORF3a, NSP4 and nucleocapsid regions are associated with mild effects while inferior outcomes were mapped in spike, RNA-dependent RNA polymerase, ORF3a, NSP3, ORF6 and nucleocapsid. Further, they concluded that mutations in ORF3a and NSP7 can lead to severe outcomes but with low prevalence. Cheng et al. [20] identified 5 major mutation points, C28144T, C14408T, A23403G, T8782C and C3037T in almost all strains for the month of April 2020. Their functional analysis show that these mutations lead to a decrease in protein stability and eventually a reduction in the virulence of SARS-CoV-2 but A23403G mutation increases the Spike-ACE2 interaction leading to an increase in its infectivity. Whole-genome analysis of 837 Indian SARS-CoV-2 genomes were carried out by Sarkar et al. [21] which revealed 33 different mutations, out of which 18 were unique to India. Based on their co-existing mutations, the Indian isolates were classified into 22 groups. Their study highlighted the evolution of divergent SARS-CoV-2 strains and also co-circulation of multiple strains in India.

Motivated by the aforementioned studies, in this work we have analysed 18392 SARS-CoV-2 genomes for 71 countries where 15359 global dataset without Indian (Dataset A) and dataset of 3033 exclusive Indian (Dataset B) SARS-CoV-2 genomes are taken separately to identify clade specific signature SNPs. To achieve this, multiple alignment using fast fourier transform (MAFFT) [22] is used for multiple sequence alignment (MSA) followed by Nextstrain [23] for performing phylogenetic analysis to identify virus clades. As a result, the virus strains are found to be distributed among 5 different clades viz. 19A, 19B, 20A, 20B and 20C. Subsequently, mutation points as SNPs are identified in each

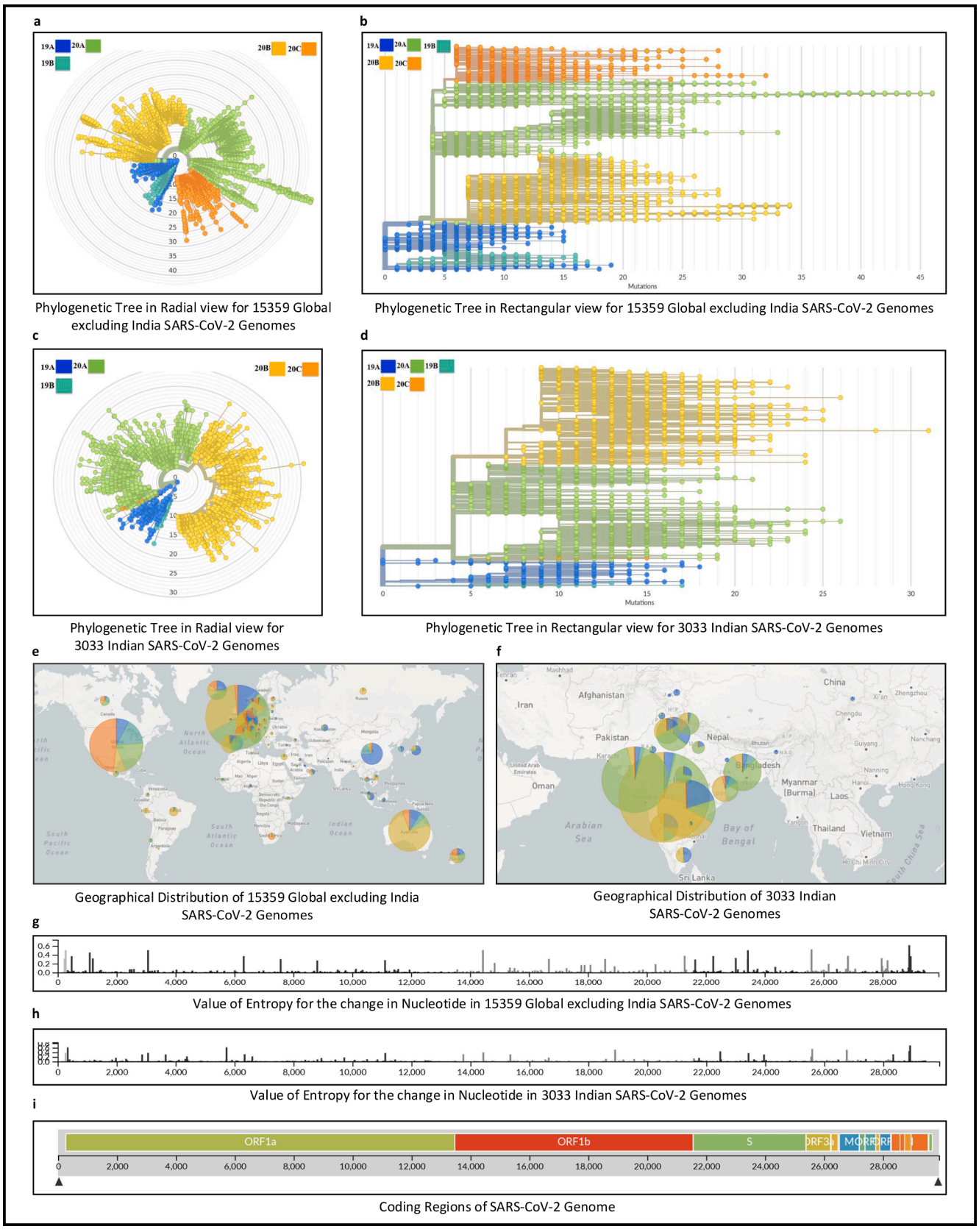

Fig. 2. Phylogenetic analyses of 15359 Global excluding India and 3033 Indian SARS-CoV-2 genomes.

Table 1
Clade wise distribution of 15359 sequences for Global Dataset without Indian sequences (Dataset A).

Country	19A	19B	20A	20B	20C	Country	19A	19B	20A	20B	20C
USA	263	530	707	244	1628	Thailand	2	16	1	3	4
England	336	7	981	1137	43	Northern Ireland	8	0	5	10	0
Australia	190	84	143	1499	116	Norway	8	0	6	1	5
Wales	97	0	1174	176	10	Austria	3	0	8	3	4
Scotland	109	14	514	212	22	Chile	0	8	5	1	1
Netherlands	141	8	241	126	41	Colombia	2	3	6	2	2
Belgium	100	1	185	226	27	Indonesia	12	0	3	0	0
China	394	90	23	12	13	Estonia	0	0	3	8	2
Iceland	95	16	152	91	66	Senegal	3	1	8	0	0
Portugal	33	11	128	189	8	Croatia	1	0	5	2	3
France	33	2	147	18	73	Georgia	4	1	4	1	1
Spain	17	129	97	19	4	Malaysia	10	1	0	0	0
New Zealand	43	10	63	77	56	Romania	0	0	5	6	0
Sweden	9	0	62	113	38	Ireland	3	0	2	5	0
Switzerland	19	0	71	48	25	Kenya	2	0	7	1	0
Italy	7	0	84	35	0	Latvia	5	0	2	2	0
Luxembourg	5	1	79	2	24	Nigeria	7	0	1	0	1
Denmark	2	0	32	9	66	Kuwait	5	0	1	1	0
Japan	71	4	7	24	0	Slovakia	1	0	4	1	0
Canada	7	25	31	16	23	Tunisia	0	0	5	1	0
Brazil	5	2	6	80	1	Bangladesh	1	0	0	3	0
Germany	23	2	5	14	25	Greece	0	1	0	3	0
Singapore	31	1	18	3	1	Qatar	4	0	0	0	0
Russia	0	0	9	41	2	Turkey	2	0	2	0	0
South Africa	1	0	11	5	34	Argentina	0	0	2	1	0
Kazakhstan	26	18	2	0	3	Belarus	2	0	1	0	0
Israel	1	0	8	31	0	Hungary	0	0	2	1	0
Poland	2	0	14	21	3	Saudi Arabia	1	0	2	0	0
Oman	16	0	6	16	1	Slovenia	2	0	1	0	0
Mexico	1	10	15	8	2	Pakistan	2	0	0	0	0
South Korea	17	19	0	0	0	Serbia	0	0	1	1	0
Peru	0	0	2	31	0	Cambodia	1	0	0	0	0
Czech Republic	0	0	9	20	3	Morocco	0	0	0	1	0
Vietnam	5	0	2	22	2	Nepal	1	0	0	0	0
Finland	3	0	13	4	6	Panama	0	0	1	0	0

Table 2
Clade wise distribution of 3033 exclusive Indian sequences (Dataset B).

State	19A	19B	20A	20B	20C
Maharashtra	39	8	289	808	0
Gujarat	16	12	559	21	3
Telangana	94	0	59	311	2
West Bengal	9	14	154	15	0
Delhi	55	1	79	19	2
Karnataka	25	2	25	51	0
Odisha	6	10	28	45	4
Haryana	15	0	44	29	1
Uttarakhand	2	1	40	25	0
Madhya Pradesh	10	0	25	1	0
Tamil Nadu	15	0	1	15	0
Uttar Pradesh	4	0	16	1	0
Rajasthan	4	0	2	0	0
Punjab	4	0	1	0	0
Ladakh	5	0	0	0	0
Bihar	2	0	0	0	0
Assam	2	0	0	0	0
Andhra Pradesh	1	0	0	1	0
Jammu and Kashmir	1	0	0	0	0
Total	309	48	1322	1342	12

clade. Furthermore, top 10 signature SNPs based on their frequency are then identified in each clade resulting in a total of 50 such SNPs each for Dataset A and Dataset B. Out of 50 signature SNPs for Dataset A, 39 unique SNPs are identified among which 25 non-synonymous signature SNPs resulted in 30 amino acid changes in protein while for Dataset B, 22 non-synonymous signature SNPs are identified from 41 unique SNPs resulting in 27 amino acid changes. These 30 and 27 amino acid changes for the non-synonymous signature SNPs are visualised in their respective protein structure as well. Furthermore, in order to judge the characteristics of the identified clades, the non-synonymous signature SNPs are considered to evaluate the changes in proteins as biological functions with the sequences using PROVEAN and PolyPhen-2 while I-Mutant 2.0 is used to evaluate their structural stability. As a consequence, for Dataset A, G251V in ORF3a in clade 19A, F308Y and G196V in NSP4 and ORF3a in 19B are the unique amino acid changes which are responsible for defining each clade as they are all deleterious and unstable. Such changes which are common for both Datasets A and B are R203M in Nucleocapsid for 20B, T85I and Q57H in NSP2 and ORF3a respectively for 20C while for Dataset B such unique changes are A97V in RdRp, G339S and G339C in NSP2 in 19A and Q57H in ORF3a in 20A.

2. Material and Methods

In this section, the collection of the dataset for SARS-CoV-2 genomes and the proposed pipeline are discussed.

2.1. Data acquisition

The collection of the dataset can be summarised as below:

- For phylogenetic analyses, Dataset A with 15359 sequences and Dataset B with 3033 SARS-CoV-2 genomes are collected from Global Initiative on Sharing All Influenza Data (GISAID)[2] while the Reference Genome (NC_045512.2)[3] is collected from National Center for Biotechnology Information (NCBI).
- The 18392 SARS-CoV-2 sequences for 71 countries are mostly distributed from December 2019 to December 2020. The average and

[2] https://www.gisaid.org/
[3] https://www.ncbi.nlm.nih.gov/nuccore/1798174254

Table 3
List of Signature SNPs in each clade for 15359 Global Dataset without India (Dataset A) and 3033 exclusive Indian (Dataset B) SARS-CoV-2 Genomes.

Clade	Signature SNPs in 15359 Global excluding India sequences					Signature SNPs in 3033 Indian sequences				
	Genomic Coordinate	Frequency	Change in Nucleotide	Change in Amino Acid	Mapped with Coding and Non-Coding Region	Genomic Coordinate	Frequency	Change in Nucleotide	Change in Amino Acid	Mapped with Coding and Non-Coding Regions
	11083	939	G>T	L37F	NSP6	13730	274	C>T	A97V	RdRp
	26144	751	G>A, G>T	G251D, G251V	ORF3a	11083	268	G>A, G>T	Synonymous, L37F	NSP6
	14805	655	C>T	Synonymous	RdRp	28311	264	C>T	P13L	Nucleocapsid
	17247	308	T>C	Synonymous	Helicase	6312	263	C>T, C>A	T1198I, T1198K	NSP3
19A	2558	235	C> T	P585S	NSP2	23929	261	C>T	Synonymous	Spike
	2480	215	A>G	I559V	NSP2	19524	60	C>T	Synonymous	Exon
	28144	199	T>C	L84S	ORF8	6310	55	C>A, C>T	S1197R, Synonymous	NSP3
	29742	188	G>A, G>T	Not Present	3'-UTR	1820	26	G>A, G>T	G339S, G339C	NSP2
	8782	166	C>T	Synonymous	NSP4	1397	21	G>A	V198I	NSP2
	1440	163	G>A	G212D	NSP2	28688	21	T>C	Synonymous	Nucleocapsid
	28144	1010	T>C	L84S	ORF8	28144	48	T>C	L84S	ORF8
	8782	993	C>T	Synonymous	NSP4	8782	47	C>T	Synonymous	NSP4
	18060	638	C>T	Synonymous	Exon	28878	45	G>A, G>T, G>C	S202N, S202I, S202T	Nucleocapsid
	17858	626	A>G	Y541C	Helicase	22468	44	G>T	Synonymous	Spike
19B	17747	610	C>T	P504L	Helicase	29742	44	G>A, G>C	Not Present	3'-UTR
	9477	190	T>A	F308Y	NSP4	7945	13	C>T	Synonymous	NSP3
	14805	190	C>T	Synonymous	RdRp	2705	7	A>G	T634A	NSP2
	28657	189	C>T	Synonymous	Nucleocapsid	14500	7	G>T	V354L	RdRp
	28863	187	C>T	S197L	Nucleocapsid	29830	6	G>T, G>C	Not Present	3'-UTR
	25979	184	G>T	G196V	ORF3a	24358	6	C>A	Synonymous	Spike
	14408	5133	C>T, C>A	P323L, P323H	RdRp	23403	1313	A>G	D614G	Spike
	23403	5131	A>G	D614G	Spike	241	1295	C>T	Not Present	5'-UTR
	241	5128	C>T	Not Present	5'-UTR	3037	1294	C>T	Synonymous	NSP3
	3037	5123	C>T	Synonymous	NSP3	14408	1248	C>T	P323L	RdRp
20A	21255	1870	G>A, G>T, G>C	Synonymous, Synonymous, Synonymous	2'-O-RMT	18877	633	C>T	Synonymous	Exon
	26801	1869	C>T, C>G	Synonymous, Synonymous	Membrane	26735	624	C>T	Synonymous	Membrane
	22227	1864	C>T	A222V	Spike	25563	611	G>A, G>T, G>C	Synonymous, Q57H, Q57H	ORF3a
	6286	1863	C>T	Synonymous	NSP3	28854	506	C>T	S194L	Nucleocapsid
	29645	1863	G>T	V30L	ORF10	22444	473	C>T	Synonymous	Spike
	28932	1862	C>T	A220V	Nucleocapsid	15324	281	C>T	Synonymous	RdRp
	241	4623	C>T	Not Present	5'-UTR	241	1341	C>T	Not Present	5'-UTR
	23403	4623	A>G	D614G	Spike	3037	1340	C>T	Synonymous	NSP3
	28882	4621	G>A, G>T	Synonymous, R203S	Nucleocapsid	23403	1340	A>G	D614G	Spike
	28883	4621	G>A, G>C	G204R, G204R	Nucleocapsid	28881	1337	G>A, G>T	R203K, R203M	Nucleocapsid
20B	28881	4620	G>A, G>T	R203K, R203M	Nucleocapsid	28882	1337	G>A	Synonymous	Nucleocapsid
	3037	4614	C>T	Synonymous	NSP3	28883	1337	G>A, G>C	G204R, G204R	Nucleocapsid
	14408	4613	C>T, C>A	P323L, P323H	RdRp	14408	1331	C>T	P323L	RdRp
	1163	1486	A>T	I120F	NSP2	5700	923	C>A	A994D	NSP3
	22992	1421	G>A, G>T, G>C	S477N, S477I, S477T	Spike	313	912	C>T	Synonymous	NSP1
	18555	1395	C>T	Synonymous	Exon	4354	170	G>A	Synonymous	NSP3
	1059	2389	C>T	T85I	NSP2	241	12	C>T	Not Present	5'-UTR
	14408	2388	C>T, C>A	P323L, P323H	RdRp	1059	12	C>T	T85I	NSP2
	3037	2387	C>T	Synonymous	NSP3	3037	12	C>T	Synonymous	NSP3
	23403	2387	A>G	D614G	Spike	14408	12	C>T	P323L	RdRp
20C	25563	2381	G>T, G>C	Q57H, Q57H	ORF3a	23403	12	A>G	D614G	Spike
	241	2379	C>T	Not Present	5'-UTR	25563	12	G>A, G>T, G>C	Synonymous, Q57H, Q57H	ORF3a
	27964	380	C>T	S24L	ORF8	16260	6	C>T	Synonymous	Helicase
	11916	190	C>T	S25L	NSP7	28821	6	C>A	S183Y	Nucleocapsid
	29553	130	G>A	Not Present	3'-UTR	22346	4	G>T	A262S	Spike
	29540	126	G>T, G>A	Not Present	3'-UTR	28221	2	G>T	E110*	ORF8

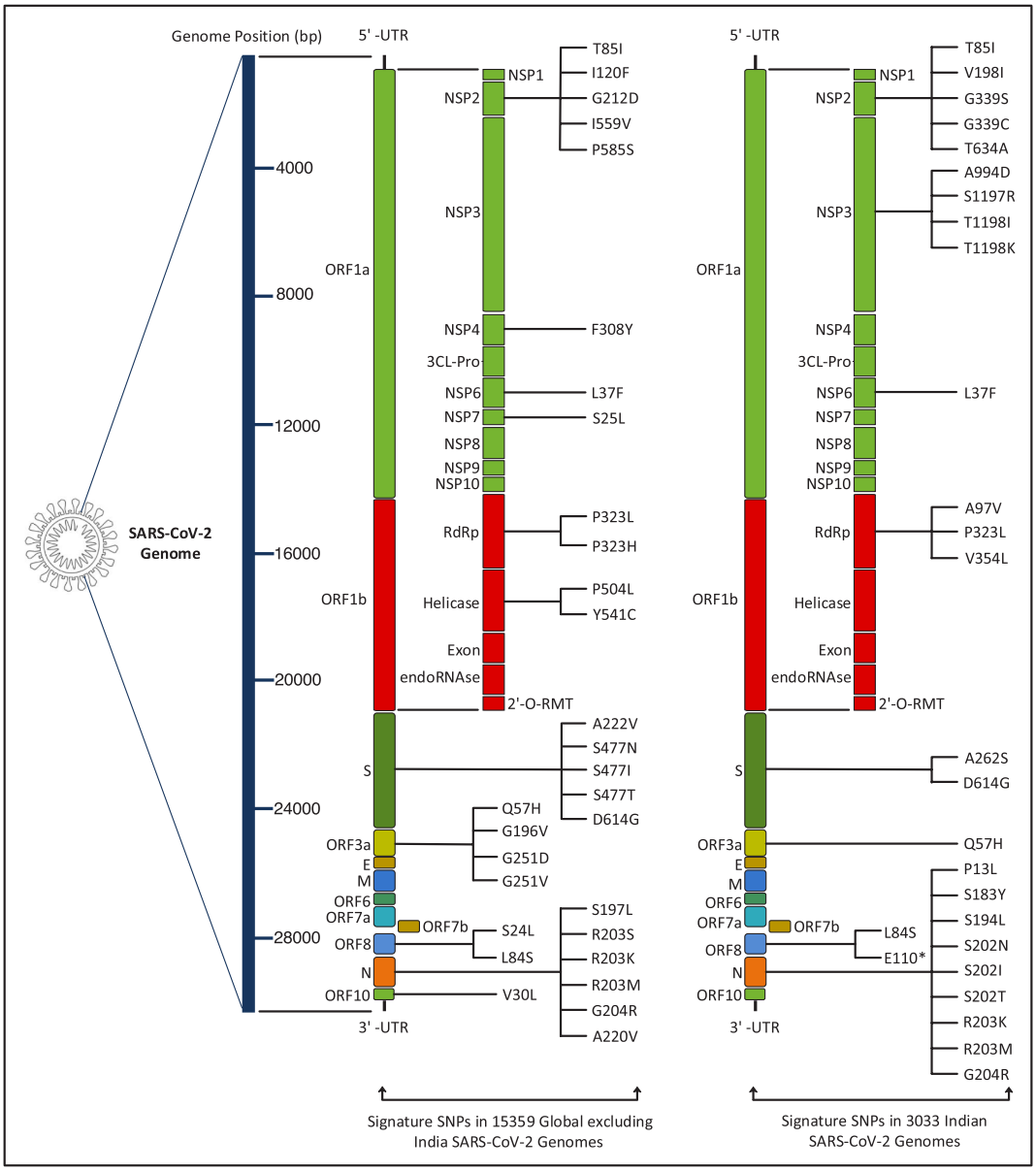

Fig. 3. Amino acid changes in the proteins for the non-synonymous signature SNPs of Global excluding India and Indian SARS-CoV-2 genomes.

maximum lengths of the sequences are 29820 and 29903 respectively.

- Further, to map the changes of amino acid in proteins, PDB of the proteins are collected from Zhang Lab[4] and other reliable sources.
- All these analysis are performed on High Performance Computing facility of NITTTR, Kolkata and the codes are written in MATLAB R2019b.

2.2. *Pipeline of the work*

The pipeline of the work is provided in Fig. 1(a). Initially, multiple sequence alignment of Datasets A and B are carried out using MAFFT followed by phylogenetic analyses using Nextstrain which resulted in the identification of virus clades. The corresponding phylogenetic trees are shown in Fig. 1(b) and Fig. 1(c) respectively. MAFFT has two novel techniques: Fast Fourier Transform (FFT) rapidly identifies homologous regions and a simple scoring system to reduce the CPU time [22]. MAFFT merges local and global algorithms for MSA and uses two different heuristic methods such as progressive (FFT-NS-2) and iterative refinements (FFT-NS-i). FFT-NS-2 is used to calculate all-pairwise distances to create a provisional MSA from which refined distances are calculated. FFT-NS-i is then performed to get the final MSA. The use of fast fourier transform in MAFFT makes it outperform other alignment techniques [22]. Thus, MAFFT is used in this work for multiple sequence alignment.

On the other hand, Nextstrain is a collection of open-source tools which is useful for understanding the evolution and spread of pathogen, particularly during an outbreak. Using Nextstrain, proper and meaningful visualisation of a large number of virus sequences can be achieved. It consists of “auspice” which is a web-based visualisation program used to present and interact with phylogenomic and phylogeographic data. There are a substantial number of tools in Nextstrain which perform phylodynamic analysis [24] which ranges from subsampling, alignment, phylogenetic inference to temporal dating of ancestral nodes and discrete trait geographic reconstruction and inference of the most likely transmission events. The spread and evolution of virus genomes can be visualised at nextstrain.org using auspice. By

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

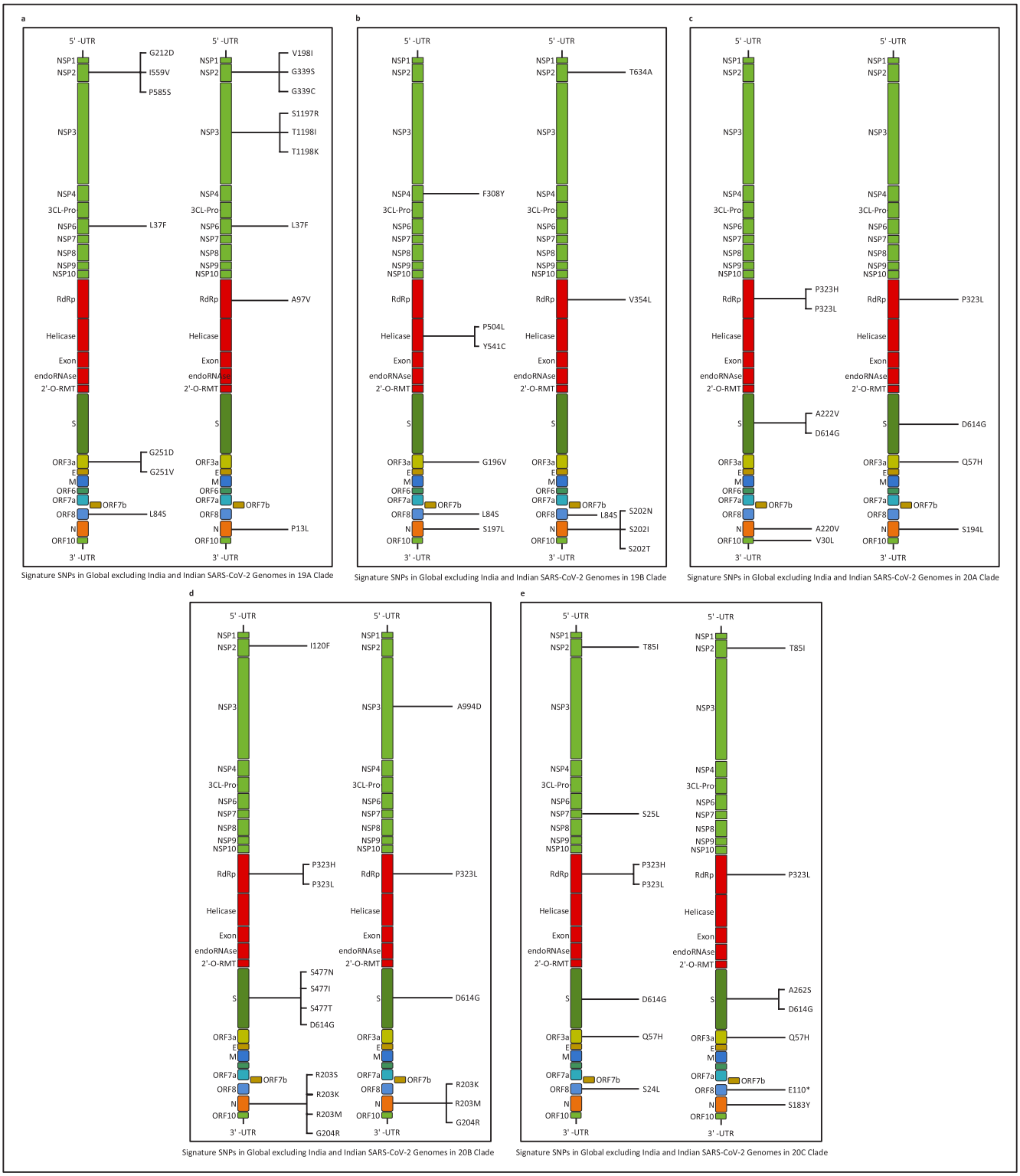

Fig. 4. Amino acid changes in the proteins for the non-synonymous signature SNPs of Global excluding India and Indian SARS-CoV-2 genomes in 5 clades.

taking the advantage of this tool, in this work the evolution and geographic distribution of SARS-CoV-2 genomes are visualised by creating the metadata in our High Performance Computing environment.

Once the virus clades are identified using Nextstrain, clade specific aligned sequences are used to identify the mutation points as substitutions specifically SNPs in each clade. Following this, amino acid changes in the virus proteins corresponding to the SNPs are identified using the codon table. Thereafter, individually for Datasets A and B, top 10 signature SNPs are identified in each clade based on their frequencies or number of occurrences in the virus genome. Such frequencies can also be quantified by considering entropy values as well, the calculation of which is described in details in [25]. It is to be noted that the amino acid changes for the SNPs can be either synonymous or non-synonymous. Thereafter, the amino acid changes in the non-synonymous signature SNPs are considered to evaluate their functional characteristics. These

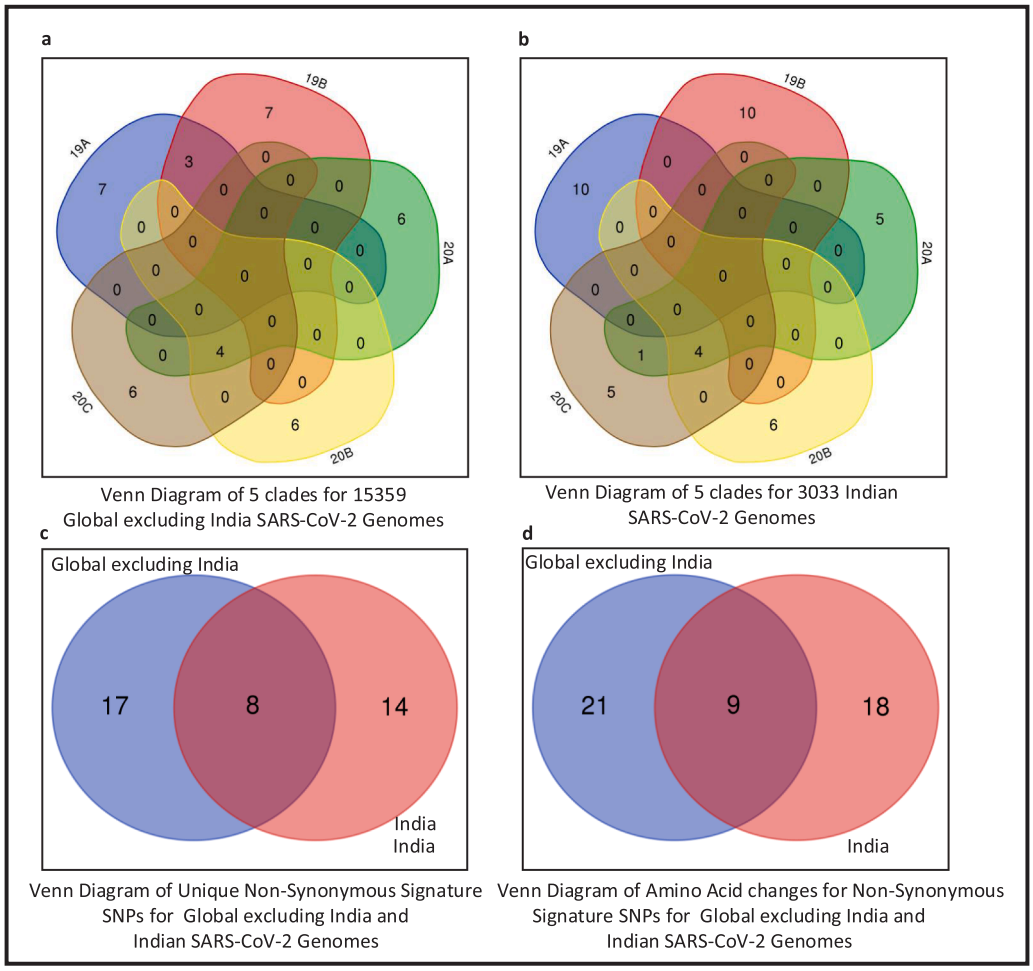

Fig. 5. Venn diagrams of Global excluding India and Indian genomes to represent common signature SNPs.

amino acid changes are visualised in the respective protein structure as well. The amino acid changes for the non-synonymous signature SNPs in the different coding regions for both Datasets A and B are visualised graphically in Fig. 1(d).

3. Results

3.1. Phylogenetic analyses

The experiments in this work are carried out according to the pipeline as provided in Fig. 1(a). In this regard, initially multiple sequence alignment of Dataset A with 15359 and Dataset B with 3033 SARS-CoV-2 genomic sequences are carried out using MAFFT followed by their phylogenetic analysis using Nextstrain. The results from the phylogenetic analyses are as follows:

- As a result of phylogenetic analysis by Nextstrain, 5 clades are identified viz. 19A, 19B, 20A, 20B and 20C. Subsequently, mutation points as SNPs are identified in each clade for both Dataset A and Dataset B.
- As a result, 2060 mutation points as SNPs are identified in clade 19A for 2194 genomic sequences for Dataset A. 1015 sequences belonging to clade 19B have 865 SNPs while for 5134, 4627 and 2389 sequences in clades 20A, 20B and 20C respectively, the number of SNPs are 4292, 3695 and 2280. The corresponding phylogenetic tree in radial and rectangular view is shown in Fig. 2(a)-(b).
- For Dataset B, 467 and 125 SNPs in clades 19A and 19B respectively are identified covering 309 and 48 genomic sequences while 2212 and 2311 SNPs are covered in clades 20A and 20B for 1322 and 1342 genomic sequences respectively. Finally, clade 20C consists of 12 sequences and has 33 SNPs. The phylogenetic tree for Dataset B is shown in Fig. 2(c)-(d).
- The clade wise distribution of 15359 and 3033 sequences for 70 countries in Dataset A and statewise for India in Dataset B are reported in Tables 1,2 respectively. For example, as reported in Table 1, USA has 263, 530, 707, 244 and 1628 sequences distributed in clades 19A, 19B, 20A, 20B and 20C respectively. Thus, most of the variants in USA belongs to clade 20C.
- For India as given in Table 2, the most dominant clade in Maharashtra is 20B while for Gujarat it is 20A. It can be further concluded from Table S2 that most of the variants in India belong to clades 20A and 20B.
- These clade wise distributions of Datasets A and B are visualised in Fig. 2(e) and (f) respectively.
- The clade wise evolution of all 18392 global including India (country wise) and separately 3033 Indian (state wise) SARS-CoV-2 genomes for each month is shown in the form of pie charts in supplementary Tables S1 and S2 respectively.
- The month wise evolution of such genomes for each clade is reported in supplementary Tables S3 and S4 respectively. The corresponding colour representation for the five major clades and the months are shown in supplementary Figure S1.

Moreover, the entropy values for the nucleotide changes for Datasets A and B are shown respectively in Fig. 2(g)-(h). Furthermore, the coding regions of the SARS-CoV-2 genome are visualised in Fig. 2(i).

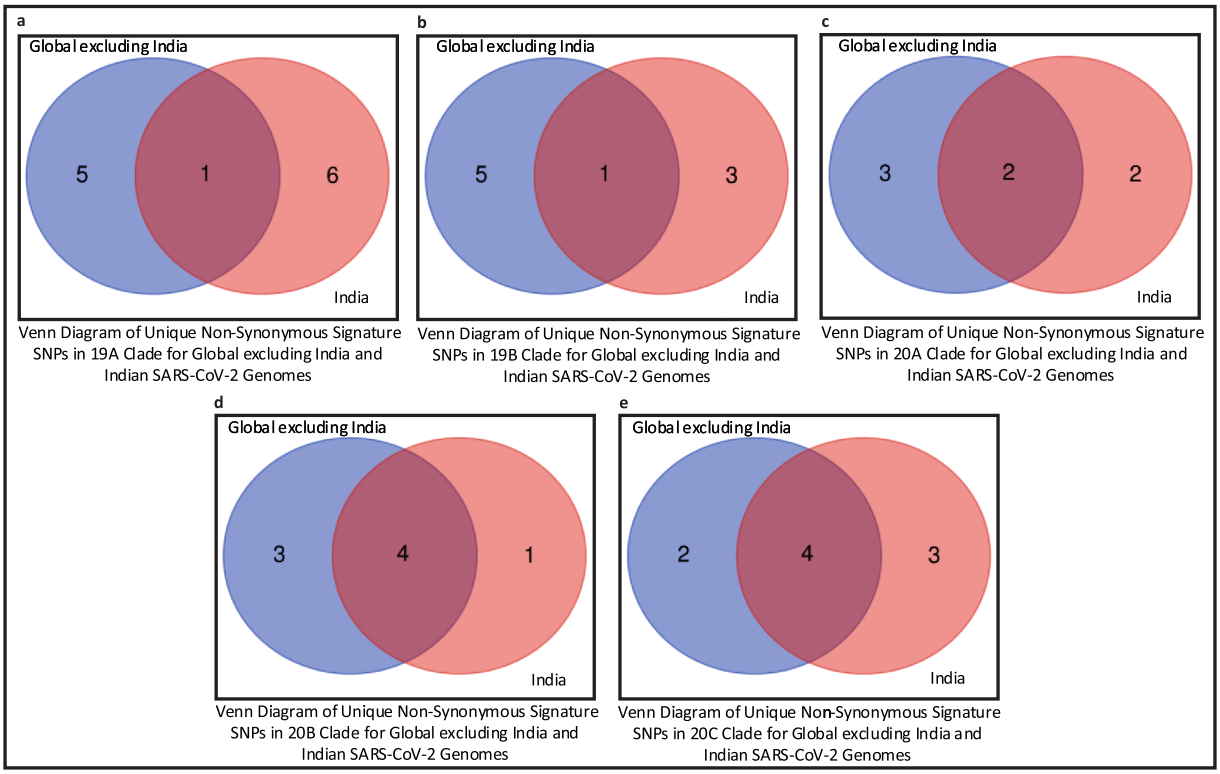

Fig. 6. Venn diagrams of Global excluding India and Indian genomes to represent common signature SNPs in clades (a) 19A (b) 19B (c) 20A (d) 20B (e) 20C.

3.2. Signature SNPs in each Clade

Once the mutation points as SNPs are determined, top 10 signature SNPs are identified in each clade for both Datasets A and B, thus resulting in 50 signature SNPs for each category as reported in Table 3. For example, for Dataset A, G11083T with a frequency of 939 is the top signature SNP in clade 19A while for Dataset B the top signature SNP is C13730T with a frequency of 274. Thereafter, 39 and 41 unique signature SNPs are identified for each category. For Dataset A, these 39 signature SNPs result in 25 non-synonymous signature SNPs with 30 amino acid changes in protein. On the other hand, for Dataset B, 41 unique signature SNPs have 22 non-synonymous signature SNPs with 27 amino acid changes. The non-synonymous signature SNPs are reported in Table 3 and their amino acid changes in protein are shown in Fig. 3. The corresponding clade-wise distribution is shown in Fig. 4. To depict the common signature SNPs in the five clades for both Datasets A and B, visualisation in the form of Venn diagram is shown in Fig. 5(a) and (b). As can be seen from the figures, there are no common SNPs in all the five clades for both Datasets A and B, thereby confirming the fact that such signature SNPs are features which indeed define each of the clades. For Datasets A and B, the visualisation of unique and common non-synonymous signature SNPs are shown in Fig. 5(c) while unique and common amino acid changes in protein are given in Fig. 5(d). Fig. 5(c) depicts 17 and 14 unique non-synonymous signature SNPs in Datasets A and B while the number of common non-synonymous signature SNPs are 8. Fig. 5(d) shows that there are 21 and 18 unique amino acid changes in Datasets A and B while 9 amino acid changes are common in both. Furthermore, clade wise unique and common non-synonymous signature SNPs are visualised in Fig. 6. It can be observed from the figure that 1, 1, 2, 4 and 4 non-synonymous signature SNPs are common for both Datasets A and B in clades 19A, 19B, 20A, 20B and 20C respectively. In 19A, the number of such unique SNPs are 5 and 6 for Datasets A and B while for 19B they are 5 and 3. For 20A and 20B such statistics are 3, 2, 3 and 1 while for 20C, the number of unique SNPs are 2 and 3. All the amino acid changes for the non-synonymous signature SNPs in the respective protein structure are visualised in Fig. 7. All the detailed results are provided in Supplementary Table S5.

4. Discussion

SARS-CoV-2 has resulted in a mass meltdown throughout the globe. Recently, the mutated variants of the virus are turning out to be another major concern for the researchers. Thus, the identification of the virus strains is very crucial in this scenario. In this regard, we have analysed Datasets A and B with 15359 and 3033 SARS-CoV-2 genomes respectively which resulted in the identification of five major clades for both Datasets A and B and consequently top 10 signature SNPs in each clades.

Initially, multiple sequence alignment of Dataset A with 15359 and Dataset B with 3033 SARS-CoV-2 genomes using MAFFT are performed followed by phylogenetic analyses using Nextstrain to identify virus clades. Thereafter, mutation points as SNPs in each clade are identified. Subsequently, top 10 signature SNPs with high frequency are identified in each clade, the details of which are provided in the Results section.

Structural changes in amino acid residues often lead to alterations in the protein translations which can lead to functional instability of the proteins. In this regard, to judge the characteristics of the identified clades, non-synonymous signature SNPs of Datasets A and B are considered to evaluate the changes in proteins as biological functions using PROVEAN (Protein Variation Effect Analyser) [26] and PolyPhen-2 (Polymorphism Phenotyping) [27] while I-Mutant 2.0 [28] is used to evaluate their structural stability. The results are reported in Table 4. PROVEAN[5] works on sequence based prediction algorithm while the prediction of Polyphen-2[6] is based on sequence, structural and phylogenetic information of a SNP. On the other hand, I-Mutant 2.0[7] uses support vector machine (SVM) for the automatic prediction of protein

[5] https://provean.jcvi.org/index.php
[6] http://genetics.bwh.harvard.edu/pph2/
[7] http://folding.biofold.org/i-mutant/i-mutant2.0.html

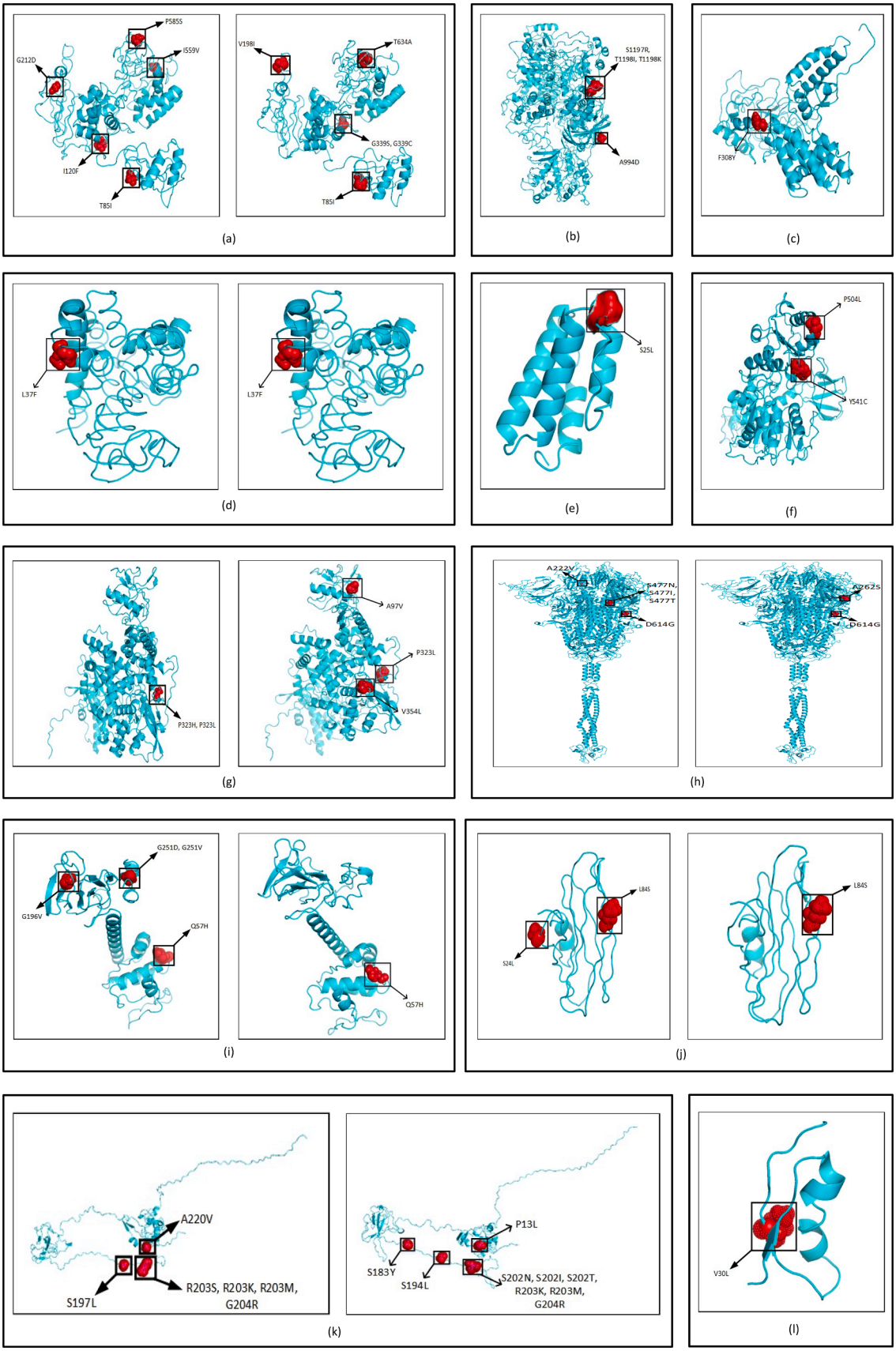

Fig. 7. Highlighted amino acid changes in the protein structures for the non-synonymous signature SNPs of (a) NSP2 for Global excluding India and India, (b) NSP3 for India, (c) NSP4 for Global excluding India, (d) NSP6 for Global excluding India and India, (e) NSP7 for Global excluding India, (f) Helicase for Global excluding India, (g) RdRp for Global excluding India and India, (h) Spike for Global excluding India and India, (i) ORF3a for Global excluding India and India, (j) ORF8 for Global excluding India and India, (k) Nucleocapsid for Global excluding India and India, (l) ORF10 for Global excluding India.

Table 4
Characteristics of non-synonymous signature SNPs for Global Dataset without India (Dataset A) and exclusive Indian (Dataset B) SARS-CoV-2 genomes.

Non-synonymous signature SNPs for Global excluding India sequences								
Clade	Change in	Mapped with	PROVEAN		PolyPhen-2		I-Mutant 2.0	
	Amino Acid	Coding Regions	Prediction	Score	Prediction	Score	Stability	DDG
	L37F	NSP6	Neutral	-1.369	Benign	0.027	Decrease	-0.05
	G251D	ORF3a	Deleterious	-6.933	Probably Damaging	1.000	Increase	0.02
	G251V	**ORF3a**	**Deleterious**	**−8.581**	**Probably Damaging**	**1.000**	**Decrease**	**−0.54**
19A	P585S	NSP2	Neutral	0.442	Benign	0.005	Decrease	−1.58
	I559V	NSP2	Neutral	0.444	Benign	0.003	Decrease	−0.28
	L84S	ORF8	Neutral	2.333	Benign	0.002	Decrease	−2.87
	G212D	NSP2	Neutral	0.704	Benign	0.013	Decrease	−1.15
	L84S	ORF8	Neutral	2.333	Benign	0.002	Decrease	−2.87
	Y541C	Helicase	Deleterious	-8.863	Probably Damaging	1.000	Increase	0.67
19B	P504L	Helicase	Deleterious	-8.158	Probably Damaging	0.993	Increase	0.16
	F308Y	**NSP4**	**Deleterious**	**−2.663**	**Probably Damaging**	**0.998**	**Decrease**	**−0.68**
	S197L	Nucleocapsid	Neutral	-2.221	Probably Damaging	0.994	Increase	0.26
	G196V	**ORF3a**	**Deleterious**	**−6.581**	**Probably Damaging**	**1.000**	**Decrease**	**−0.80**
	P323H	RdRp	Neutral	1.146	Benign	0.005	Decrease	−2.09
	P323L	RdRp	Neutral	-0.865	Benign	0.005	Decrease	−0.80
20A	D614G	Spike	Neutral	0.598	Benign	0.004	Decrease	−1.94
	A222V	Spike	Neutral	-0.096	Benign	0.000	Increase	0.48
	V30L	ORF10	Deleterious	-3.000	Not Generated	Not Generated	Decrease	−1.31
	A220V	Nucleocapsid	Neutral	-0.140	Probably Damaging	0.999	Increase	0.76
	D614G	Spike	Neutral	0.598	Benign	0.004	Decrease	−1.94
	R203S	Nucleocapsid	Neutral	-2.374	Probably Damaging	0.994	Decrease	−2.10
	G204R	Nucleocapsid	Neutral	-1.656	Probably Damaging	1.000	No change	0.00
	R203K	Nucleocapsid	Neutral	-1.604	Probably Damaging	0.969	Decrease	−2.26
20B	**R203M**	**Nucleocapsid**	**Deleterious**	**−3.305**	**Probably Damaging**	**0.998**	**Decrease**	**−1.52**
	P323H	RdRp	Neutral	1.146	Benign	0.005	Decrease	−2.09
	P323L	RdRp	Neutral	-0.865	Benign	0.005	Decrease	−0.80
	I120F	NSP2	Neutral	-1.333	Benign	0.393	Decrease	−1.85
	S477N	Spike	Neutral	-0.034	Benign	0.014	Increase	0.01
	S477I	Spike	Neutral	-1.310	Probably Damaging	0.531	Increase	0.34
	S477T	Spike	Neutral	-0.336	Benign	0.066	Decrease	−0.49
	T85I	**NSP2**	**Deleterious**	**−4.090**	**Probably Damaging**	**0.998**	**Decrease**	**−1.71**
	P323H	RdRp	Neutral	1.146	Benign	0.005	Decrease	−2.09
20C	P323L	RdRp	Neutral	-0.865	Benign	0.005	Decrease	−0.80
	D614G	Spike	Neutral	0.598	Benign	0.004	Decrease	−1.94
	Q57H	**ORF3a**	**Deleterious**	**−3.286**	**Probably Damaging**	**0.966**	**Decrease**	**−1.12**
	S24L	ORF8	Neutral	-1.833	Benign	0.013	Increase	0.53
	S25L	NSP7	Deleterious	-4.272	Probably Damaging	0.600	Increase	0.21
Non-synonymous signature SNPs for Indian sequences								
Clade	Change in	Mapped with	PROVEAN		PolyPhen-2		I-Mutant 2.0	
	Amino Acid	Coding Regions	Prediction	Score	Prediction	Score	Stability	DDG
	A97V	**RdRp**	**Deleterious**	**−3.611**	**Probably Damaging**	**0.990**	**Decrease**	**−0.53**
	L37F	NSP6	Neutral	-1.369	Benign	0.027	Decrease	-0.05
	P13L	Nucleocapsid	Neutral	-1.230	Probably Damaging	1.000	Increase	0.11
	T1198I	NSP3	Neutral	-0.085	Probably Damaging	0.998	Decrease	-0.72
19A	T1198K	NSP3	Neutral	-0.353	Not generated	Not generated	Decrease	-1.37
	S1197R	NSP3	Neutral	-0.835	Not generated	Not generated	Decrease	-0.88
	G339S	**NSP2**	**Deleterious**	**−3.130**	**Probably Damaging**	**1.000**	**Decrease**	**−1.57**
	G339C	**NSP2**	**Deleterious**	**−4.874**	**Probably Damaging**	**1.000**	**Decrease**	**−1.91**
	V198I	NSP2	Neutral	0.307	Benign	0.006	Increase	0.18
	L84S	ORF8	Neutral	2.333	Benign	0.002	Decrease	-2.87
	S202N	Nucleocapsid	Neutral	-0.404	Probably Damaging	0.994	Decrease	-0.80
19B	S202I	Nucleocapsid	Deleterious	-3.308	Probably Damaging	0.998	Increase	0.22
	S202T	Nucleocapsid	Neutral	−1.428	Probably Damaging	0.986	Decrease	-0.53
	T634A	NSP2	Neutral	-0.004	Benign	0.106	Decrease	-1.13
	V354L	RdRp	Deleterious	-2.581	Probably Damaging	0.997	Decrease	-1.95
	D614G	Spike	Neutral	0.598	Benign	0.004	Decrease	-1.94
20A	P323L	RdRp	Neutral	-0.865	Benign	0.005	Decrease	-0.80
	Q57H	**ORF3a**	**Deleterious**	**−3.286**	**Probably Damaging**	**0.966**	**Decrease**	**−1.12**
	S194L	Nucleocapsid	Deleterious	-4.272	Probably Damaging	0.994	Increase	0.45
	D614G	Spike	Neutral	0.598	Benign	0.004	Decrease	-1.94
	R203K	Nucleocapsid	Neutral	-1.604	Probably Damaging	0.969	Decrease	-2.26
20B	**R203M**	**Nucleocapsid**	**Deleterious**	**−3.305**	**Probably Damaging**	**0.998**	**Decrease**	**−1.52**
	G204R	Nucleocapsid	Neutral	-1.656	Probably Damaging	1.000	No change	0.00
	P323L	RdRp	Neutral	-0.865	Benign	0.005	Decrease	-0.80
	A994D	NSP3	Neutral	-1.103	Not generated	Not generated	Decrease	-0.78
	T85I	**NSP2**	**Deleterious**	**−4.090**	**Probably Damaging**	**0.998**	**Decrease**	**−1.71**
	P323L	RdRp	Neutral	-0.865	Benign	0.005	Decrease	-0.80
20C	D614G	Spike	Neutral	0.598	Benign	0.004	Decrease	-1.94
	Q57H	**ORF3a**	**Deleterious**	**−3.286**	**Probably Damaging**	**0.966**	**Decrease**	**−1.12**

(*continued on next page*)

Table 4 (*continued*)

Non-synonymous signature SNPs for Global excluding India sequences								
Clade	Change in	Mapped with	PROVEAN		PolyPhen-2		I-Mutant 2.0	
	Amino Acid	Coding Regions	Prediction	Score	Prediction	Score	Stability	DDG
	S183Y	Nucleocapsid	Deleterious	-2.750	Probably Damaging	0.998	No change	0.00
	A262S	Spike	Neutral	0.154	Not generated	Not generated	Decrease	-0.95

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. Considering the predictions of both PROVEAN and Polyphen-2, for Dataset A it can be seen from Table 4 that out of 30 unique amino acid changes, 11 unique changes are damaging or deleterious while for Dataset B, 9 unique changes are damaging out of 27 unique changes. 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 [29]. The protein stabilities corresponding to 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 free energy change values (DDG). A negative value of DDG indicates that the stability of a protein is decreasing. The result from I-mutant 2.0 concludes that out of the 11 and 9 unique damaging changes for Datasets A and B, 6 changes for both decrease the stability of the protein structures respectively. Consequently, for Dataset A, G251V in ORF3a in clade 19A, F308Y and G196V in NSP4 and ORF3a in 19B are the unique amino acid changes which are responsible for defining each clade as they are all deleterious and unstable. Such changes which are common for both Datasets A and B are R203M in Nucleocapsid for 20B, T85I and Q57H in NSP2 and ORF3a respectively for 20C while for Dataset B such unique changes are A97V in RdRp, G339S and G339C in NSP2 in 19A and Q57H in ORF3a in 20A. All of them are marked in bold in Table 4. It is to be noted that for Indian non-synonymous signature SNPs, there are 26 amino acid changes as opposed to 27 such changes in Table 3, the discarded change being E110* in ORF8 as the amino acid change leads to a stop codon.

Table 5 provides a comparative study of all the signature SNPs identified in our work with that of literature [30,16,31,18,19,32,33,20,34–36]. In [30], the authors have performed whole-genome sequencing of 303 Indian isolates and have reported 11 important genetic mutations as a part of Clade I/A3i which are dominant in most of the states in India. Yuan et al. [16] have performed genomic analysis of 11,183 genomes from around the globe and have reported 9 SNPs with high frequency. In their work, Goswami et al. [31] have identified 18 high frequency genomic co-ordinates viz. hot-spots to investigate inverted repeat loci and CpG islands. They concluded that these points are indicative of genomic instability of SARS-CoV-2. Genomic analysis of 570 SARS-CoV-2 genomes from China, Europe, US and India have been carried out in [18] where they have identified at least 10 hotspot mutations which are present in more than 80% of viral genomes. In [19], Nagy et al. have mapped 3733 non-silent mutations to amino acid changes for 4566 patients where they identified 17 mutations related to hospitalisation, severe and deadly outcomes as well. Rahimi et al. [32] have reported 17 high frequency mutations that are reported in other literatures as well. In [33], the authors have worked with 1566 SARS-CoV-2 genome sequences across ten Asian countries and have clustered and characterized them based on the clade they belong to. Cheng et al. [20] have considered 1809 SARS-CoV-2 genomes and identified 1017 and 512 non-synonymous and synonymous mutations respectively. In their work, they have reported 7 dominant mutations for each month from January to April 2020. In [34], Hamdan et al. have considered 11 SARS-CoV-2 isolates from Lebanon to identify new mutations which have not been reported till then in Lebanon. Yang et al. [35] have performed phylodynamic analysis of 247 genomic sequences to identify four genetic clusters called super-spreaders. SS1 was widely circulating in Asia and the US whereas SS4 was responsible for the pandemic in Europe. In [36] have reported 9 newly evolved SARS-CoV-2 SNPs that have undergone a rapid increase or decrease in frequency for 30–80% in the initial four months. It can be seen from the table that out of the 48 unique genomic coordinates identified in the literature, 29 signature SNPs are common with our work. Out of these 29 common signature SNPs, C14805T, T17247C, C17747T, A17858G, C18060T and G26144T are present in Dataset A. Such signature SNPs present only in Indian genomes are C6310A, C6312A, C13730T, G22346A, C28311T, T28688C, C28854T and G28878A while C241T, C1059T, G1397A, C3037T, C8782T, G11083T, C14408T, A23403G, G25563T, T28144C, G28881A, G28882A, G28883C and G29742A are common in both Datasets A and B. Furthermore, for Dataset A, G26144T which corresponds to G251V in ORF3a in clade 19A is damaging and shows a decrease in stability while C13730T which corresponds to A97V in RdRp for 19A, C1059T corresponding to T85I in NSP2 for clade 20C and G25563T corresponding to Q57H in ORF3a for 20C are damaging and exhibits shrinking stability for both Datasets A and B.

5. Conclusion

In this work, multiple sequence alignment of 18392 SARS-CoV-2 sequences is carried out using MAFFT followed by phylogenetic analyses using Nextstrain where 15359 global dataset without Indian (Dataset A) and dataset of 3033 exclusive Indian (Dataset B) SARS-CoV-2 genomes are considered separately to identify the virus clades. Consequently, the virus strains are found to be distributed among five major clades viz. 19A, 19B, 20A, 20B and 20C. Subsequently, mutation points as SNPs are identified in each clade. Thereafter, clade specific signature SNPs are identified by considering top 10 SNPs with high frequency, resulting in 50 such signature SNPs each for Datasets A and B. Out of each 50 signature SNPs, 39 and 41 unique SNPs are identified among which 25 non-synonymous signature SNPs (out of 39) resulted in 30 amino acid changes in protein while 27 changes in amino acid are identified from 22 non-synonymous signature SNPs (out of 41). These 30 and 27 amino acid changes for the non-synonymous signature SNPs are visualised in their respective protein structures as well. The sequence and structural homology-based prediction of biological functions along with the protein structural stability of such amino acid changes are also determined to judge the characteristics of the identified clades. Consequently, for Dataset A, G251V in ORF3a in clade 19A, F308Y and G196V in NSP4 and ORF3a in 19B are the unique amino acid changes which are responsible for defining each clade as they are all deleterious and unstable. Such changes which are common for both Datasets A and B are R203M in Nucleocapsid for 20B, T85I and Q57H in NSP2 and ORF3a respectively for 20C while for Dataset B such unique changes are A97V in RdRp, G339S and G339C in NSP2 in 19A and Q57H in ORF3a in 20A. Moreover, a comparative study is also put forth to show the correctness of our work. We hope this work will better equip the researchers in their path of designing anti-viral therapeutics to mitigate COVID-19. As a future scope of research, consensus of SNPs can be considered by taking

Table 5
Comparative study of our work with the literature.

Genomic coordinate	Change in Nucleotide	Change in Amino acid	Coordinate of Amino Acid in Protein	Mapped with Coding and Non-coding Region	Banu et al. [30]	Yuan et al. [16]	Goswami et al. [31]	Weber et al. [18]	Nagy et al. [19]	Rahimi et al. [32]	Sengupta et al. [33]	Cheng et al. [20]	Abou-Hamdan et al. [34]	Yang et al. [35]	Zhu et al. [36]	Our work
241	C>T	NA	NA	5'-UTR			✓			✓				✓	✓	✓
1059	C>T	T>I	85	NSP2		✓	✓	✓		✓		✓				✓
1190	C>T	P>S	129	NSP2										✓		
1397	G>A	V>I	378	ORF1a	✓								✓			✓
1440	G>A	G>D	212	NSP2				✓								✓
1605	A>C	N>T	267	NSP 1ab			✓									
1917	C>T	T>I	371	NSP2				✓								
2891	G>R, G>A	A>T	58	NSP3			✓	✓								
3037	C>T	Synonymous	105, 106	NSP3		✓	✓			✓		✓		✓	✓	✓
4402	T>C	Synonymous	561	NSP3										✓		
5062	G>T	L>F	781	NSP3										✓		
6310	C>A	S>R	1197	NSP3					✓							✓
6312	C>A	T>K	1198, 2016	NSP3, ORF1a	✓				✓							✓
6446	G>A	V>I	1243	ORF1ab				✓								
8782	C>T	Synonymous	75, 76	NSP4		✓	✓			✓		✓		✓	✓	✓
9438	C>T	T>I	295	NSP4										✓		
9924	C>T	A>V	3220	ORF1a	✓											
11083	G>T	L>F	37, 3606	NSP6, ORF1a	✓		✓			✓			✓	✓		✓
12053	C>T	L>F	71	NSP7					✓							
13730	C>T	A>V	97	RdRp	✓				✓							✓
14408	C>T	P>L	314, 323	RdRp	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
14805	C>T	Synonymous	446, 455	RdRp			✓							✓		✓
17247	T>C	Synonymous	337	NSP13			✓									✓
17747	C>T	P>L	504	Helicase			✓	✓		✓				✓		✓
17858	A>G	Y>C	541	Helicase			✓	✓		✓				✓		✓
18060	C>T	Synonymous	6, 7	Exon				✓		✓				✓		✓
21724	G>T	L>F	54	Spike					✓		✓					
21859	C>T	Synonymous	99	Spike										✓		
22346	G>A	A>T	262	Spike							✓					✓
22661	G>T	V>F	367	Spike										✓		
23403	A>G	D>G	614	Spike	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
24047	G>A	A>T	829	Spike											✓	
25088	G>T	V>F	1176	Spike					✓							
25563	G>T	Q>H	57	ORF3a		✓	✓		✓	✓	✓	✓	✓			✓
26144	G>T	G>V	251	ORF3a	✓		✓		✓	✓				✓		✓
26149	T>C	S>P	253	ORF3a					✓							
27299	T>C	I>T	33	ORF6					✓							
28144	T>C	L>S	84	ORF8	✓		✓	✓	✓	✓	✓	✓			✓	✓
28311	C>T	P>L	13	Nucleocapsid	✓				✓							✓
28688	T>C	L>L	129	Nucleocapsid										✓		✓
28854	C>T	S>L	194	Nucleocapsid				✓	✓		✓					✓
28878	G>A	S>N	202	Nucleocapsid	✓									✓		✓
28881	G>A	R>K	203	Nucleocapsid		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
28882	G>A	Synonymous	203	Nucleocapsid		✓				✓				✓	✓	✓
28883	G>C	G>R	204	Nucleocapsid		✓			✓	✓	✓		✓	✓	✓	✓
29095	C>T	F>F	274	Nucleocapsid										✓		
29148	T>C	I>T	292	Nucleocapsid					✓							
29742	G>T, G>R	NA	NA	3'-UTR						✓				✓		✓

more than one multiple sequence alignment techniques. Also, investigation of the characteristics of these signature SNPs of SARS-CoV-2 on human hosts can be conducted with the help of virologists. The authors are working in these directions.

Ethics approval and consent to participate

The ethical approval or individual consent was not applicable.

Availability of data and materials

All the files which include dataset (raw and aligned sequences, metadata for Nextstrain and JSON files as outputs of Nextstrain), codes, supplementary PDF and videos of clade specific virus evolution and transmission in 71 countries are available at "http://www.nitttrkol.ac.in/indrajit/projects/COVID-Evolution-SignatureSNPs-18 K/".

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. However, it does not provide any publication fees.

CRediT authorship contribution statement

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, Project administration, Resources, Validation, Writing - review & editing. **Suman Nandi:** Conceptualization, Formal analysis, Software, Validation, Visualization, Writing - review & editing. **Nikhil Sharma:** Conceptualization, Formal analysis, Software, Validation, Visualization, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

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, athttps://doi.org/10.1016/j.ymeth.2021.09.005.

References

[1] A. Nesta, D. Tafur, C. Beck, Hotspots of human mutation, Trends Genet. (2020), https://doi.org/10.1016/j.tig.2020.10.003.
[2] J. Tang, P. Tambyah, D. Hui, Emergence of a new sars-cov-2 variant in the uk, J Infection (2020), https://doi.org/10.1016/j.jinf.2020.12.024.
[3] S. Brookman, J. Cook, M. Zucherman, et al., Effect of the new sars-cov-2 variant b.1.1. 7 on children and young people, Lancet Child Adolescent Health (2021), https://doi.org/10.1016/S2352-4642(21)00030-4.
[4] M. Makoni, South africa responds to new sars-cov-2 variant, The Lancet 397 (2021) 267, https://doi.org/10.1016/S0140-6736(21)00144-6.
[5] S. Kumar, S.K. Saxena, Structural and molecular perspectives of sars-cov-2, Methods (2021), https://doi.org/10.1016/j.ymeth.2021.03.007.
[6] E. Boehm, I. Kronig, R.A. Neher, Novel SARS-CoV-2 variants: the pandemics within the pandemic, Clinical Microbiol Infection 27 (8) (2021) 1109–1117, https://doi.org/10.1016/j.cmi.2021.05.022.
[7] D. Kim, J.-Y. Lee, J.-S. Yang, et al., The architecture of sars-cov-2 transcriptome, Cell 181 (04 2020). doi:10.1016/j.cell.2020.04.011.
[8] P. Zhou, X.L. Yang, X.G. Wang, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (2020) 270–273, https://doi.org/10.1038/s41586-020-2012-7.
[9] D.E. Gordon, G.M. Jang, M. Bouhaddou, et al., A sars-cov-2 protein interaction map reveals targets for drug repurposing, Nature 583 (2020) 459–468, https://doi.org/10.1038/s41586-020-2286-9.
[10] I.-N. Lu, C.P. Muller, F.Q. He, Applying next-generation sequencing to unravel the mutational landscape in viral quasispecies, Virus Res. 283 (2020), 197963, https://doi.org/10.1016/j.virusres.2020.197963.
[11] C. Yin, Genotyping coronavirus sars-cov-2: methods and implications, Genomics 112 (5) (2020) 3588–3596, https://doi.org/10.1016/j.ygeno.2020.04.016.
[12] I. Manfredonia, C. Nithin, A. Ponce-Salvatierra, et al., Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements, Nucleic Acids Res. 48 (22) (2020) 12436–12452, https://doi.org/10.1093/nar/gkaa1053.
[13] I. Saha, N. Ghosh, A. Pradhan, et al., Whole genome analysis of more than 10 000 SARS-CoV-2 virus unveils global genetic diversity and target region of NSP6, Briefings in Bioinformatics 22 (2) (2021) 1106–1121, https://doi.org/10.1093/bib/bbab025.
[14] X. Tang, C. Wu, X. Li, et al., On the origin and continuing evolution of SARS-CoV-2, National Sci. Rev. 7 (6) (2020) 1012–1023, https://doi.org/10.1093/nsr/nwaa036.
[15] R. Wang, Y. Hozumi, C. Yin, et al., Decoding sars-cov-2 transmission, evolution and ramification on covid-19 diagnosis, vaccine, and medicine, J. Chem. Inform. Modeling XXXX (06 2020). doi:10.1021/acs.jcim.0c00501.
[16] F. Yuan, L. Wang, Y. Fang, et al., Global SNP analysis of 11,183 SARS-CoV-2 strains reveals high genetic diversity, Transboundary Emerging Diseases (11 2020). doi:10.1111/tbed.13931.
[17] J. Chen, R. Wang, M. Wang, et al., Mutations strengthened sars-cov-2 infectivity, J. Mol. Biol. 432 (07 2020). doi:10.1016/j.jmb.2020.07.009.
[18] S. Weber, C. Ramirez, W. Doerfler, Signal hotspot mutations in sars-cov-2 genomes evolve as the virus spreads and actively replicates in different parts of the world, Virus Res. 289 (2020), 198170, https://doi.org/10.1016/j.virusres.2020.198170.
[19] A. Nagy, S. Pongor, B. Győrffy, Different mutations in sars-cov-2 associate with severe and mild outcome, Int. J. Antimicrob. Agents 57 (2020), 106272, https://doi.org/10.1016/j.ijantimicag.2020.106272.
[20] L. Cheng, X. Han, Z. Zhu, et al., Functional alterations caused by mutations reflect evolutionary trends of sars-cov-2, Briefings Bioinformatics (2021) 1–9, https://doi.org/10.1093/bib/bbab042.
[21] R. Sarkar, S. Mitra, P. Chandra, et al., Comprehensive analysis of genomic diversity of SARS-CoV-2 in different geographic regions of India: an endeavour to classify Indian SARS-CoV-2 strains on the basis of co-existing mutations, Arch. Virol. 166 (3) (2021) 801–812, https://doi.org/10.1007/s00705-020-04911-0.
[22] K. Katoh, K. Misawa, K. Kuma, et al., MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform, Nucleic Acids Research 30 (14) (2002) 3059–3066. doi:https://doi: 10.1093/nar/gkf436.
[23] J. Hadfield, C. Megill, S. Bell, et al., Nextstrain: real-time tracking of pathogen evolution, Bioinformatics (Oxford, England) 34 (2018), https://doi.org/10.1093/bioinformatics/bty407.
[24] E.M. Volz, K. Koelle, T. Bedford, Viral phylodynamics, PLoS Computer Biol. 9 (3) (2013), e1002947, https://doi.org/10.1371/journal.pcbi.1002947.
[25] I. Saha, N. Ghosh, D. Maity, et al., Genome-wide analysis of indian sars-cov-2 genomes for the identification of genetic mutation and snp, Infection, Genetics and Evolution 85 (2020), 104457, https://doi.org/10.1016/j.meegid.2020.104457.
[26] Y. Choi, A.P. Chan, Provean web server: a tool to predict the functional effect of amino acid substitutions and indels, Bioinformatics 31 (16) (2015) 2745–2747, https://doi.org/10.1093/bioinformatics/btv195.
[27] I.A. Adzhubei, S. Schmidt, L. Peshkin, et al., A method and server for predicting damaging missense mutations, Nature Methods 7 (4) (2010) 248–249, https://doi.org/10.1038/nmeth0410-248.
[28] E. Capriotti, P. Fariselli, R. Casadio, I-mutant2.0: predicting stability changes upon mutation from the protein sequence or structure, Nucleic Acid Res. 33 (2005) 306–310, https://doi.org/10.1093/nar/gki375.
[29] M.S. Hossain, A.S. Roy, M.S. Islam, In silico analysis predicting effects of deleterious snps of human rassf5 gene on its structure and functions, Sci. Rep. 10 (2020) 14542, https://doi.org/10.1038/s41598-020-71457-1.
[30] S. Banu, B. Jolly, P. Mukherjee, et al., A Distinct Phylogenetic Cluster of Indian Severe Acute Respiratory Syndrome Coronavirus 2 Isolates, Open Forum Infectious Diseases 7 (11) (09 2020). doi:10.1093/ofid/ofaa434.
[31] P. Goswami, M. Bartas, M. Lexa, et al., SARS-CoV-2 hot-spot mutations are significantly enriched within inverted repeats and CpG island loci, Briefings in Bioinformatics (12 2020). doi:10.1093/bib/bbaa385.
[32] A. Rahimi, A. Mirzazadeh, S. Tavakolpour, Genetics and genomics of sars-cov-2: A review of the literature with the special focus on genetic diversity and sars-cov-2 genome detection, Genomics 113 (1, Part 2) (2021) 1221–1232. doi:10.1016/j.ygeno.2020.09.059.
[33] A. Sengupta, S.S. Hassan, P.P. Choudhury, Clade gr and clade gh isolates of sars-cov-2 in asia show highest amount of snps, Infection, Genetics Evol. 89 (2021), 104724, https://doi.org/10.1016/j.meegid.2021.104724.

[34] M. Abou-Hamdan, K. Hamze, A.A. Sater, et al., Variant analysis of the first lebanese sars-cov-2 isolates, Genomics (2020) 892–895, https://doi.org/10.1016/j.ygeno.2020.10.021.
[35] X. Yang, N. Dong, E.W.C. Chan, et al., Genetic cluster analysis of sars-cov-2 and the identification of those responsible for the major outbreaks in various countries, Emerging Microbes Infections 9 (1) (2020) 1287–1299, https://doi.org/10.1080/22221751.2020.1773745.
[36] Z. Zhu, G. Liu, K. Meng, et al., Rapid Spread of Mutant Alleles in Worldwide SARS-CoV-2 Strains Revealed by Genome-Wide Single Nucleotide Polymorphism and Variation Analysis, Genome Biology and Evolution 13 (2) (01 2021). doi:10.1093/gbe/evab015.