ELSEVIER

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.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Infection, Genetics and Evolution 85 (2020) 104457

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Research Paper

Genome-wide analysis of Indian SARS-CoV-2 genomes for the identification of genetic mutation and SNP

Indrajit Saha[a,*,1], Nimisha Ghosh[b,1], Debasree Maity[c], Nikhil Sharma[d], Jnanendra Prasad Sarkar[e,f], Kaushik Mitra[g]

[a] *Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India*
[b] *Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Orissa, India*
[c] *Department of Electronics and Communication Engineering, MCKV Institute of Engineering, Howrah, West Bengal, India*
[d] *Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India*
[e] *Larsen & Toubro Infotech, Pune, India*
[f] *Department of Computer Science and Engineering, Jadavpur University, Kolkata, West Bengal, India*
[g] *Department of Community Medicine, Burdwan Medical College, Barddhaman, West Bengal, India*

ARTICLE INFO

Keywords:
Multiple sequence alignment
Point mutation
SNPs
SARS-CoV-2

ABSTRACT

The wave of COVID-19 is a big threat to the human population. Presently, the world is going through different phases of lock down in order to stop this wave of pandemic; India being no exception. We have also started the lock down on 23rd March 2020. In this current situation, apart from social distancing only a vaccine can be the proper solution to serve the population of human being. Thus it is important for all the nations to perform the genome-wide analysis in order to identify the genetic variation in Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) so that proper vaccine can be designed. This fast motivated us to analyze publicly available 566 Indian complete or near complete SARS-CoV-2 genomes to find the mutation points as substitution, deletion and insertion. In this regard, we have performed the multiple sequence alignment in presence of reference sequence from NCBI. After the alignment, a consensus sequence is built to analyze each genome in order to identify the mutation points. As a consequence, we have found 933 substitutions, 2449 deletions and 2 insertions, in total 3384 unique mutation points, in 566 genomes across 29.9 K bp. Further, it has been classified into three groups as 100 clusters of mutations (mostly deletions), 1609 point mutations as substitution, deletion and insertion and 64 SNPs. These outcomes are visualized using BioCircos and bar plots as well as plotting entropy value of each genomic location. Moreover, phylogenetic analysis has also been performed to see the evolution of SARS-CoV-2 virus in India. It also shows the wide variation in tree which indeed vivid in genomic analysis. Finally, these SNPs can be the useful target for virus classification, designing and defining the effective dose of vaccine for the heterogeneous population.

1. Introduction

Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) and its associated disease known as COVID-19, was originated in Wuhan, China (Zhu et al., 2020). It has wreaked havoc on human lives and declared as a pandemic by World Health Organisation on 11th March 2020. Among others, symptoms of COVID-19 include fever, cough and shortness of breath (Chen et al., 2020). In more severe cases, infection may lead to pneumonia (Zhou et al., 2020), kidney failure and eventual death. As of now, no vaccine or medicine has been invented or discovered and the only protective measures are being taken by different countries are through lock downs and social distancing. However, even these extreme measures have not been able to contain the spread of SARS-CoV-2. Everyday thousands of new cases are coming into light. According to the record of 27th June, globally more than 9.8 million people are affected by this deadly virus, with a total death count close to 500 thousand (Worldometer, 2020). India is also fighting hard to keep the citizens safe through lock down and other protective measures. Yet the virus has shown a surge in the country with more than 500 thousand affected cases and death cases of more than 15 thousand

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

https://doi.org/10.1016/j.meegid.2020.104457
Received 10 May 2020; Received in revised form 3 July 2020; Accepted 5 July 2020
Available online 11 July 2020
1567-1348/ © 2020 Elsevier B.V. All rights reserved.

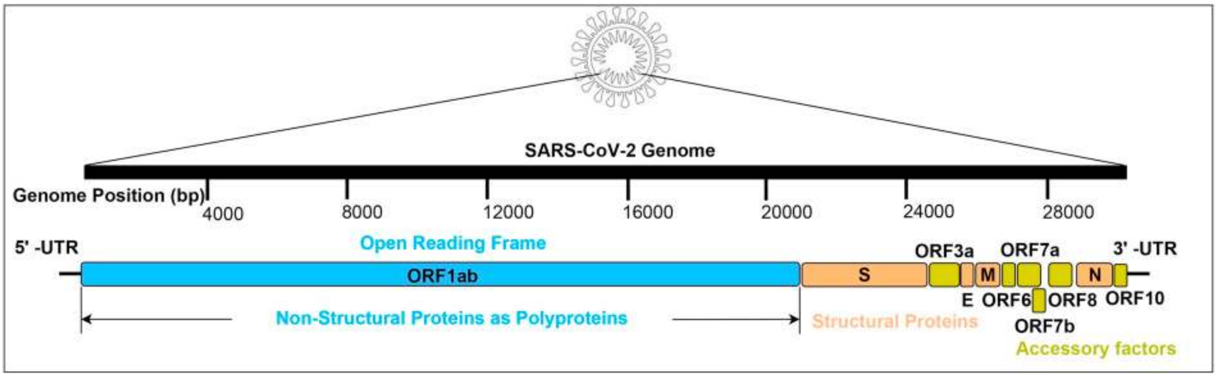

Fig. 1. Genomic orientation of SARS-CoV-2 virus.

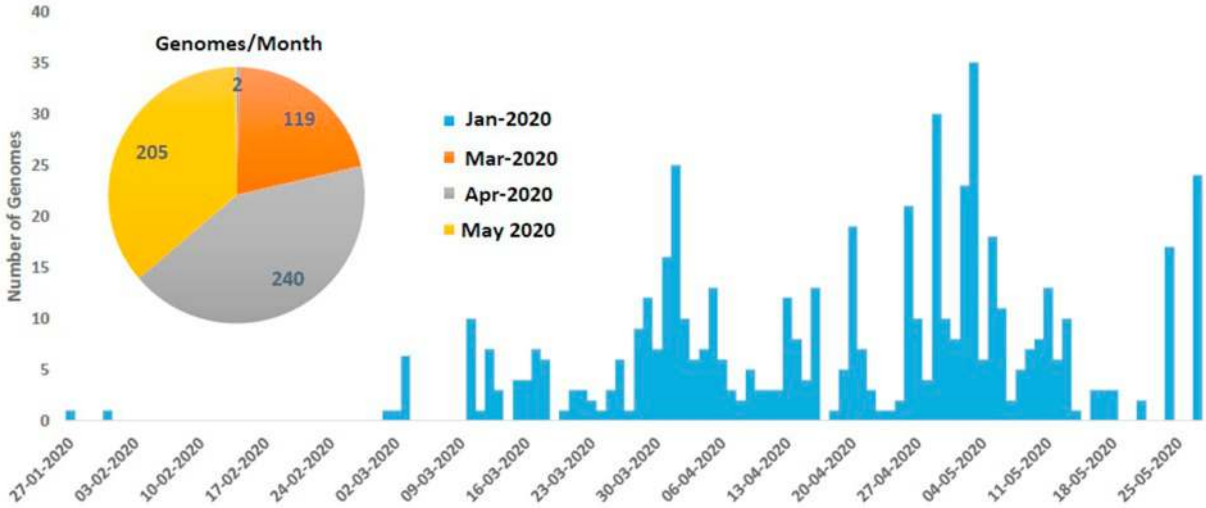

Fig. 2. Bar plot shows number of SARS-CoV-2 genomes per day and pie chart shows number of SARS-CoV-2 genomes per month uploaded during January to May 2020.

Table 1
Coding Regions of SARS-CoV-2 genome after mapping the collected coding regions from NCBI with the Reference Sequence of SARS-CoV-2 genome.

Start coordinate of coding region	End coordinate of coding region	Length of coding region	Name of the coding region
1	265	265	5′-UTR (Non-coding Region)
266	21,555	21,290	ORF1ab
21,563	25,385	3823	Spike (S)
25,393	26,221	829	ORF3a
26,245	26,473	229	Envelope (E)
26,523	27,192	670	Membrane (M)
27,202	27,388	187	ORF6
27,394	27,760	367	ORF7a
27,756	27,888	133	ORF7b
27,894	28,260	367	ORF8
28,274	29,534	1261	Nucleocapsid (N)
29,558	29,675	118	ORF10
29,676	29,905	230	3′-UTR (Non-coding Region)

as of 27th June 2020 (Worldometer, 2020).

The metagenomic analysis using Next-Generation Sequencing (NGS) (Lu et al., 2020) reveals that the SARS-CoV-2 is a single-stranded enveloped RNA virus with a genome length of 29.9 kilobases (Cui et al., 2019; Su et al., 2016; Weiss and Navas-Martin, 2005; Zhou et al., 2020). It has 11 coding regions as reported in NCBI that can encode ORF1ab polyproteins, spike (S) glycoprotein, envelope (E) protein, membrane (M) glycoprotein, nucleocapsid (N) protein and accessory proteins such as ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF10. It has also been reported that several non-structural proteins (nsp) are encoded from Open Reading Frame (ORF). The genomic orientation of SARS-CoV-2 virus is shown in Fig. 1.

The strain of this virus is novel and the understanding the genetic variability as mutation of this virus in different nations is still very limited, especially the coding region of Open Reading Frame (ORF). Generally, the mutation occurs when an error is incorporated in a viral genome (Fleischmann, 1996). It can also be considered to be a coping mechanism with genomic damage. As a consequence, the resultant mutated strain may cause an outbreak in human host like the case with SARS-CoV-2. The DNA mutation can be of three types: base substitution, deletion and insertion. Moreover, if the substitution occurs more than 1% of the population, it can be considered as Single Nucleotide Polymorphism (SNP). Such polymorphism usually different from the mutation as it creates a variant in the population while mutation keeps the population same (Pavlovic-Lazetic et al., 2004). On the other hand, RNA viruses have high mutation rates (Jenkins et al., 2002; Woo et al., 2009). Thus it is difficult to identify the proper string of the virus. Subsequently, designing and define the dose of the vaccine are also very challenging tasks (Paital et al., 2020). In this regard, Chothe et al. used SNP on sequences of Bovine herpesvirus-1 (BoHV-1) (Chothe et al., 2018), which was affecting cattle and causing respiratory illness, to cluster them into three groups with two different vaccine groups and one distinct cluster of field isolates. Based on this information, they

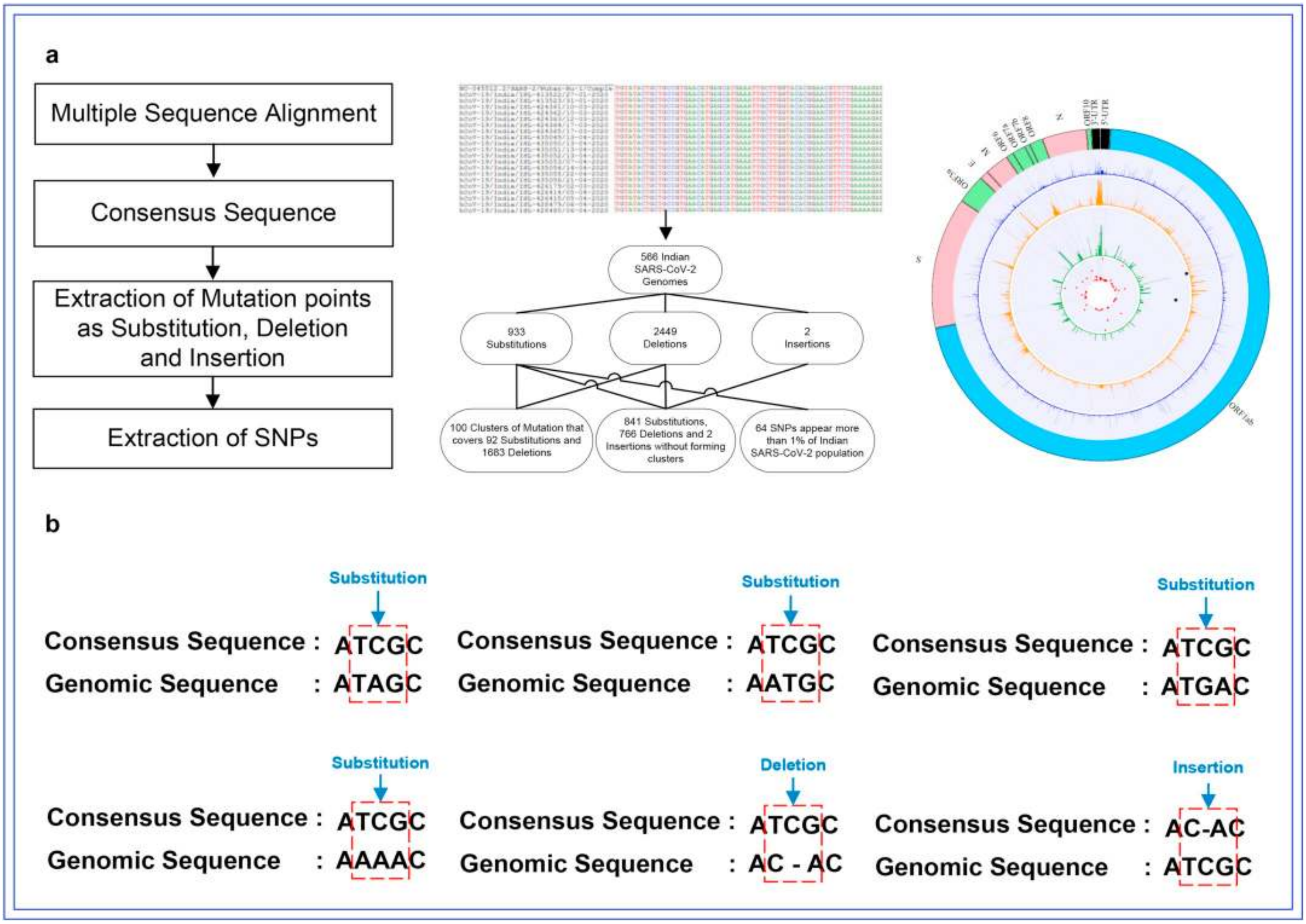
a
Multiple Sequence Alignment
Consensus Sequence
Extraction of Mutation points as Substitution, Deletion and Insertion
Extraction of SNPs
566 Indian SARS-CoV-2 Genomes
933 Substitutions
2449 Deletions
2 Insertions
100 Clusters of Mutation that covers 92 Substitutions and 1683 Deletions
841 Substitutions, 766 Deletions and 2 Insertions without forming clusters
64 SNPs appear more than 1% of Indian SARS-CoV-2 population
ORF10
3'-UTR
5'-UTR
N
ORF8
ORF7b
ORF7a
ORF6
M
E
ORF3a
S
ORF1ab
b
Substitution
Consensus Sequence : ATCGC
Genomic Sequence : ATAGC
Substitution
Consensus Sequence : ATCGC
Genomic Sequence : AATGC
Substitution
Consensus Sequence : ATCGC
Genomic Sequence : ATGAC
Substitution
Consensus Sequence : ATCGC
Genomic Sequence : AAAAC
Deletion
Consensus Sequence : ATCGC
Genomic Sequence : AC - AC
Insertion
Consensus Sequence : AC-AC
Genomic Sequence : ATCGC

Fig. 3. (a) Pipeline of the Workflow (b) Different types of mutation considered during identification of substitution, deletion and insertion in Indian SARS-CoV-2 genomes.

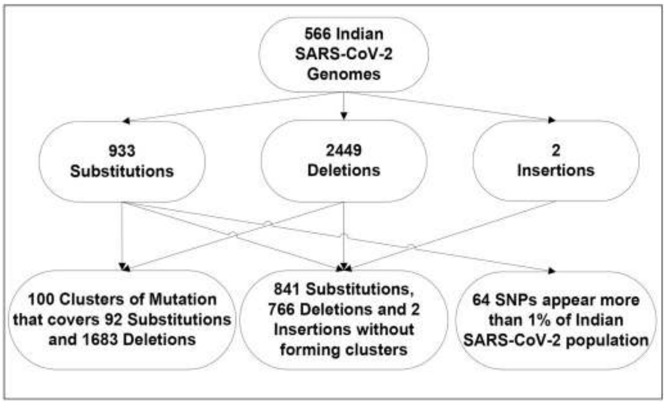
566 Indian SARS-CoV-2 Genomes
933 Substitutions
2449 Deletions
2 Insertions
100 Clusters of Mutation that covers 92 Substitutions and 1683 Deletions
841 Substitutions, 766 Deletions and 2 Insertions without forming clusters
64 SNPs appear more than 1% of Indian SARS-CoV-2 population

Fig. 4. Results of mutation analysis of 566 Indian SARS-CoV-2 genomes.

developed an SNP-based PCR assay to show differentiation between vaccine and clinical strains. Analysis of SNPs for Varicella-zoster virus was carried out in (Jeon et al., 2016) which suggested 24 potential vaccine-specific sites. Taking cues from (Chothe et al., 2018) and (Jeon et al., 2016), we have analyzed 566 Indian SARS-CoV-2 genomes to find the genetic variability in terms of point mutation as substitution, deletion and insertion as well as SNP. It can be helpful for the classification of virus strain and accordingly designing of vaccine and defining the dose of the vaccine can be done effectively.

To address the above facts, we have analyzed publicly available 566 Indian complete or near complete SARS-CoV-2 genomes in order to find the mutation points as substitution, deletion and insertion. For this purpose, Multiple Sequence Alignment (Wallace et al., 2005) is performed in presence of reference sequence from NCBI. Thereafter, a consensus sequence is build to analyze each genome to identify the mutation points. As a result, we have found 933 substitutions, 2449 deletions and 2 insertions, in total 3384 unique mutation points, in 566 genomes across 29.9 kbp. Further, it has been classified into three groups (a) cluster of mutation points if the mutation appears more than two times in consecutive genomic positions (b) point mutations as substitution, deletion and insertion that are not present in clusters (c) Single Nucleotide Polymorphism (SNP) that appeared more than 1% of the population of SARS-CoV-2 used in our study. Finally, 100 clusters of mutation (mostly deletions), 1609 point mutations as substitution, deletion and insertion and 64 SNPs out of categories (a) and (b) have been identified as they appeared more than 1% of the population i.e. 6 times in Indian SARS-CoV-2 genomes. These outcomes are visualized using BioCircos and bar plots as well as plotting entropy value of each genomic location. Moreover, phylogenetic analysis (Stuessy, 2009) has also been performed to see the evolution of SARS-CoV-2 virus in India.

2. Material and methods

In this section, we have discussed the source of data or genomic sequence of virus and methods used in systemic way to accomplish this task of finding mutation points as substitution, deletion, insertion as

Table 2
Mutation of 20 Clusters within Indian SARS-CoV-2 genomes.

Start coordinate of mutation	End coordinate of mutation	Cluster size of mutation	Avg. occurrence of mutation in genomes	Type of mutation	Avg. entropy of cluster of mutation	Mapped with coding region
2	83	82	147	Deletion	0.2757	5′-UTR (Non-coding Region)
3827	3877	51	1	Deletion	0.0150	ORF1ab
6508	6554	47	1	Deletion	0.0161	ORF1ab
6840	6877	38	3	Deletion	0.0363	ORF1ab
7020	7049	30	1	Deletion	0.0151	ORF1ab
7279	7334	56	3	Deletion	0.0341	ORF1ab
8345	8445	101	1	Deletion	0.0146	ORF1ab
19,306	19,354	49	1	Deletion	0.0145	ORF1ab
19,389	19,424	36	1	Deletion	0.0157	ORF1ab
19,497	19,550	54	2	Deletion	0.0267	ORF1ab
21,171	21,208	38	1	Deletion	0.0148	ORF1ab
21,459	21,502	44	2	Deletion	0.0196	ORF1ab
21,547	21,603	57	3	Deletion	0.0367	ORF1ab
23,403	23,405	3	112	Substitution	0.2485	Spike
23,966	24,018	53	3	Deletion	0.0362	Spike
26,340	26,399	60	6	Deletion	0.0540	Envelope
27,541	27,682	142	2	Deletion	0.0284	ORF7a
27,697	27,743	47	2	Deletion	0.0275	ORF7a
28,883	28,885	3	64	Substitution	0.3817	Nucleocapsid
29,742	29,904	163	142	Deletion	0.2406	3′-UTR (Non-coding Region)

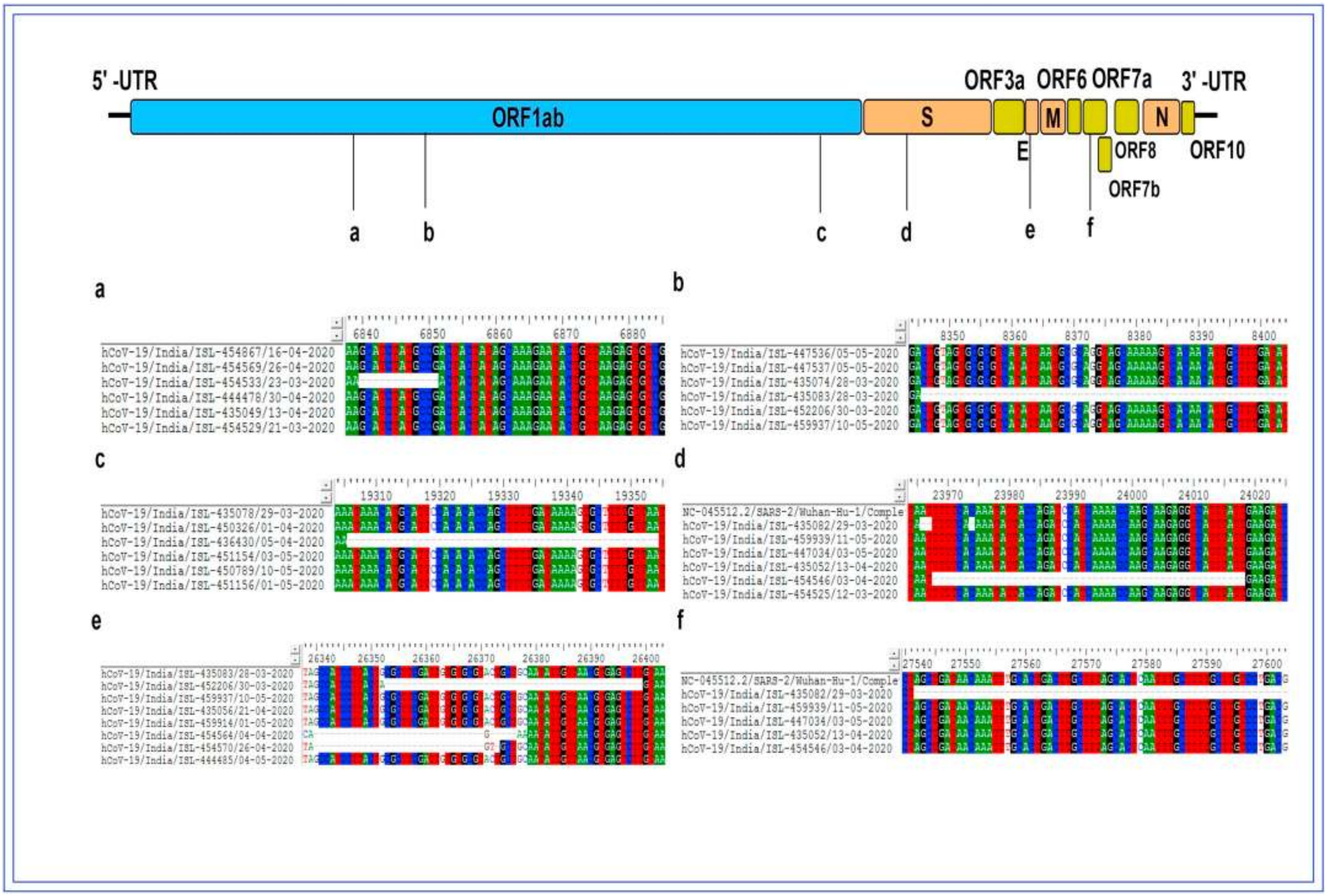
5' -UTR
ORF1ab
S
ORF3a
ORF6
ORF7a
3' -UTR
M
N
E
ORF8
ORF10
ORF7b
a
b
c
d
e
f
a
6840 6850 6860 6870 6880
hCoV-19/India/ISL-454867/16-04-2020
hCoV-19/India/ISL-454569/26-04-2020
hCoV-19/India/ISL-454533/23-03-2020
hCoV-19/India/ISL-444478/30-04-2020
hCoV-19/India/ISL-435049/13-04-2020
hCoV-19/India/ISL-454529/21-03-2020
b
8350 8360 8370 8380 8390 8400
hCoV-19/India/ISL-447536/05-05-2020
hCoV-19/India/ISL-447537/05-05-2020
hCoV-19/India/ISL-435074/28-03-2020
hCoV-19/India/ISL-435083/28-03-2020
hCoV-19/India/ISL-452206/30-03-2020
hCoV-19/India/ISL-459937/10-05-2020
c
19310 19320 19330 19340 19350
hCoV-19/India/ISL-435078/29-03-2020
hCoV-19/India/ISL-450326/01-04-2020
hCoV-19/India/ISL-436430/05-04-2020
hCoV-19/India/ISL-451154/03-05-2020
hCoV-19/India/ISL-450789/10-05-2020
hCoV-19/India/ISL-451156/01-05-2020
d
23970 23980 23990 24000 24010 24020
NC-045512.2/SARS-2/Wuhan-Hu-1/Comple
hCoV-19/India/ISL-435082/29-03-2020
hCoV-19/India/ISL-459939/11-05-2020
hCoV-19/India/ISL-447034/03-05-2020
hCoV-19/India/ISL-435052/13-04-2020
hCoV-19/India/ISL-454546/03-04-2020
hCoV-19/India/ISL-454525/12-03-2020
e
26340 26350 26360 26370 26380 26390 26400
f
27540 27550 27560 27570 27580 27590 27600
NC-045512.2/SARS-2/Wuhan-Hu-1/Comple
hCoV-19/India/ISL-435082/29-03-2020
hCoV-19/India/ISL-459939/11-05-2020
hCoV-19/India/ISL-447034/03-05-2020
hCoV-19/India/ISL-435052/13-04-2020
hCoV-19/India/ISL-454546/03-04-2020

Fig. 5. Some Mutations in coding regions as deletion in Indian SARS-CoV-2 genomes.

well as SNP.

2.1. Data collection

The genomic sequences of Indian SARS-CoV-2 virus was collected from Global Initiative on Sharing All Influenza Data (GISAID)[2] in fasta format on 11th June 2020. The dataset contains 566 genomes with

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

Table 3
Mutation as SNPs in more than 10% of population of Indian SARS-CoV-2 genomes.

Coordinate of mutation	Occurrence of mutation in genomes	Type of mutation	Change in nucleotide	Change in amino acid	Entropy	Mapped with coding region
241	341	Substitution	C>T	S>L	0.7078	5′-UTR (Non-coding Region)
3037	339	Substitution	C>T	S>F	0.7020	ORF1ab
6310	65	Substitution	C>A	A>E,A>K	0.3972	ORF1ab
6312	160	Substitution	C>A	T>K	0.7357	ORF1ab
11,085	204	Substitution	G>T,G>A	C>F,C>-,C>Y	0.74748	ORF1ab
13,732	187	Substitution	C>T	A>V,A>-	0.6561	ORF1ab
14,410	332	Substitution	C>T	P>L	0.7277	ORF1ab
18,879	117	Substitution	C>T	S>F	0.5216	ORF1ab
22,446	69	Substitution	C>T	T>I	0.3703	Spike
23,405	334	Substitution	A>G	D>G	0.7198	Spike
23,931	165	Substitution	C>T	T>I	0.6856	Spike
25,565	122	Substitution	G>T	R>I	0.5207	ORF3a
26,737	112	Substitution	C>T	T>I	0.4969	Membrane
28,313	174	Substitution	C>T	P>L	0.6987	Nucleocapsid
28,856	71	Substitution	C>T	S>L	0.3899	Nucleocapsid
28,883	64	Substitution	G>A	R>K	0.3755	Nucleocapsid
28,884	64	Substitution	G>A	G>N	0.3849	Nucleocapsid
28,885	64	Substitution	G>C	G>T	0.3849	Nucleocapsid

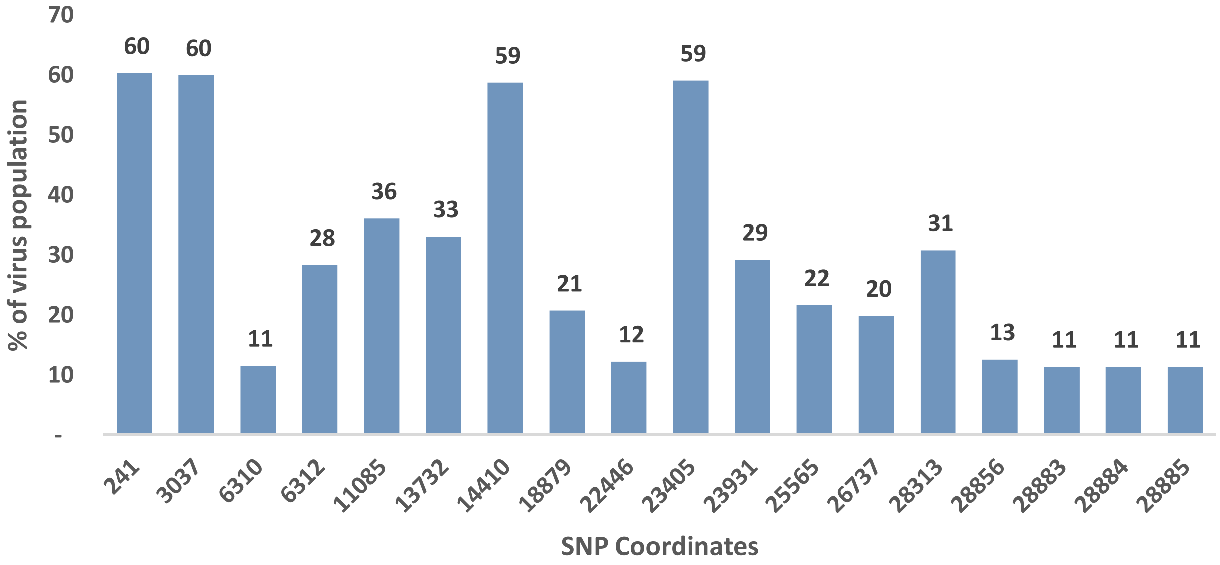

Fig. 6. SNPs present in more than 10\% of population of Indian SARS-CoV-2 genomes.

sequence ID and sequence in fasta format. We have extracted the date from the sequence ID to show the number of sequences uploaded per month. This is shown in Fig. 2. Moreover, the original and shorter sequence IDs (for better visualization) and the length of each genome are mentioned in the supplementary Table S1. The average length of the 566 genomes is 29,831 bp. It is also to be noted that we have considered complete or near complete virus genomes for our study. Further, we have downloaded the Reference Sequence (NC_045512.2)[3] from National Center for Biotechnology Information (NCBI) to conduct the experiment with 566 Indian SARS-CoV-2 genomes. This reference genome is also used to map the coding regions as collected from NCBI. This is also reported in Table 1 and used while mentioning the mutation points in result section. Please note that for the data visualization and editing, BioEdit and MEGA-X have been used.

2.2. Pipeline of the workflow

The pipeline of the workflow is shown in Fig. 3(a). In order to find the mutations in 566 Indian SARS-CoV-2 genomes, the multiple sequencing alignment (MSA) technique called ClustalW Thompson et al. (1994) is used in presence of reference sequence from NCBI. The ClustalW uses the concept of Neighbor-Joining tree where bootstrap size is consider as 1000. It is a widely used MSA technique for aligning any number of homologous nucleotide or protein sequences like in our case. It uses progressive alignment method where the most similar sequences with the best alignment score are aligned first. After performing the alignment, a consensus sequence is built in order to extract the mutation points from each genome as substitution, deletion and insertion. The detection scheme of identifying substitution, deletion and insertion is shown in Fig. 3(b). This is applied for each virus genome. Once we have the list of substitution, deletion and insertion further (a) a list of clusters of mutation points are identified based on their appearance in consecutive genomic positions. Here a cluster is formed if the consecutive mutation in genomic positions is greater than 2. (b) a list of substitution, deletion and insertion that are not present in the clusters (c) Single Nucleotide Polymorphisms (SNPs) that occur as substitution in more than 1% of the population of virus genomes i.e. 566 × 0.01 = ceil(5.66) = 6 in our case. Similarly, if we consider 10% of the virus population, then such substitution occurs in 57 virus

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

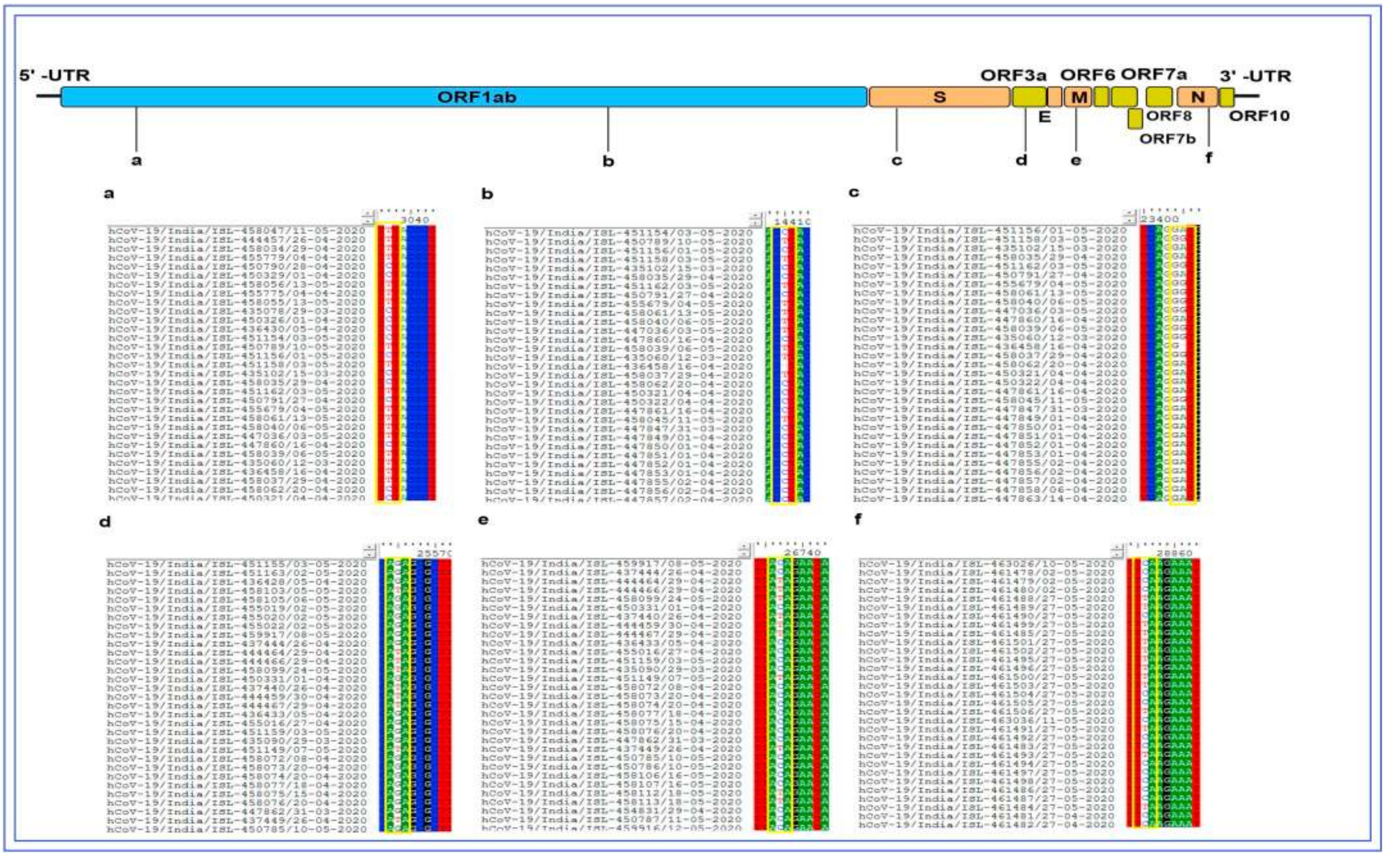

Fig. 7. Mutations in coding regions as SNPs greater than 1% of population of Indian SARS-CoV-2 genomes.

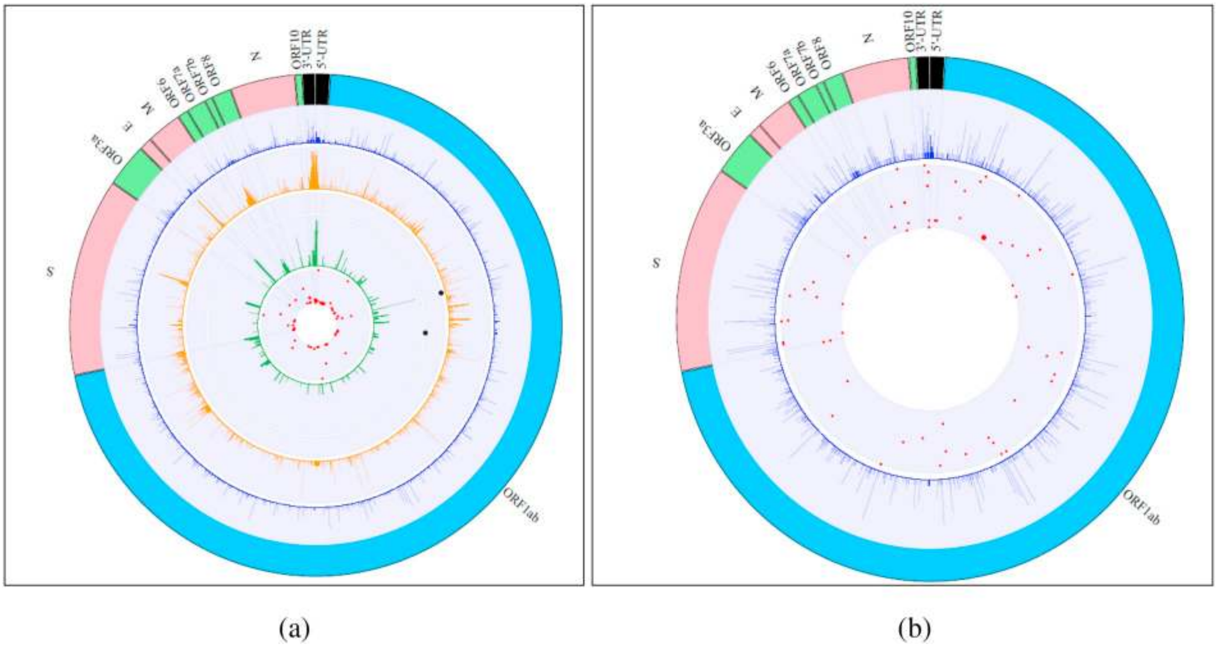

Fig. 8. BioCircos plots to illustrate the frequency of mutations across the Indian SARS-CoV-2 genomes through different tracks: (a) Substitutions (outer track 1), Deletions (track 2), Insertions (track 3), Clusters (track 4) and SNPs (inner track 5) (b) Substitutions (outer track 1) and SNPs (inner track 2).

genomes. The substitution, deletion and insertion in nucleotides is also translated to change in amino acid using codon table.[4] Moreover, to compute the change in nucleotides in virus sequences, entropy ($\mathscr{E}$) is computed as follows:

$$\mathcal{E} = ln\ 5 + \sum \mathscr{V}_{a}^{p}\ [\,ln\,(\mathscr{V}_{a}^{p})\,] \tag{1}$$

where $V_a{}^p$ represents the frequency of each residue a occurring at position p. 5 represents the number of possible residues for nucleic acid (in

[4] https://en.wikipedia.org/wiki/DNA_codon_table

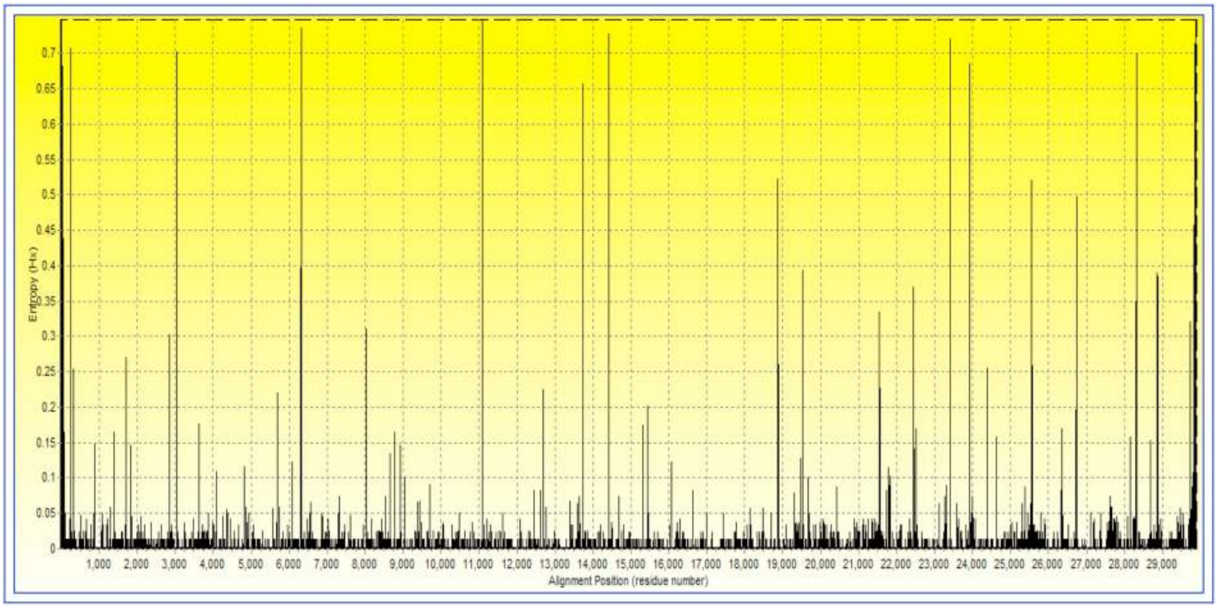

Fig. 9. The change of entropy in Indian SARS-CoV-2 genomes for each genomic location after alignment with Reference Sequence from NCBI.

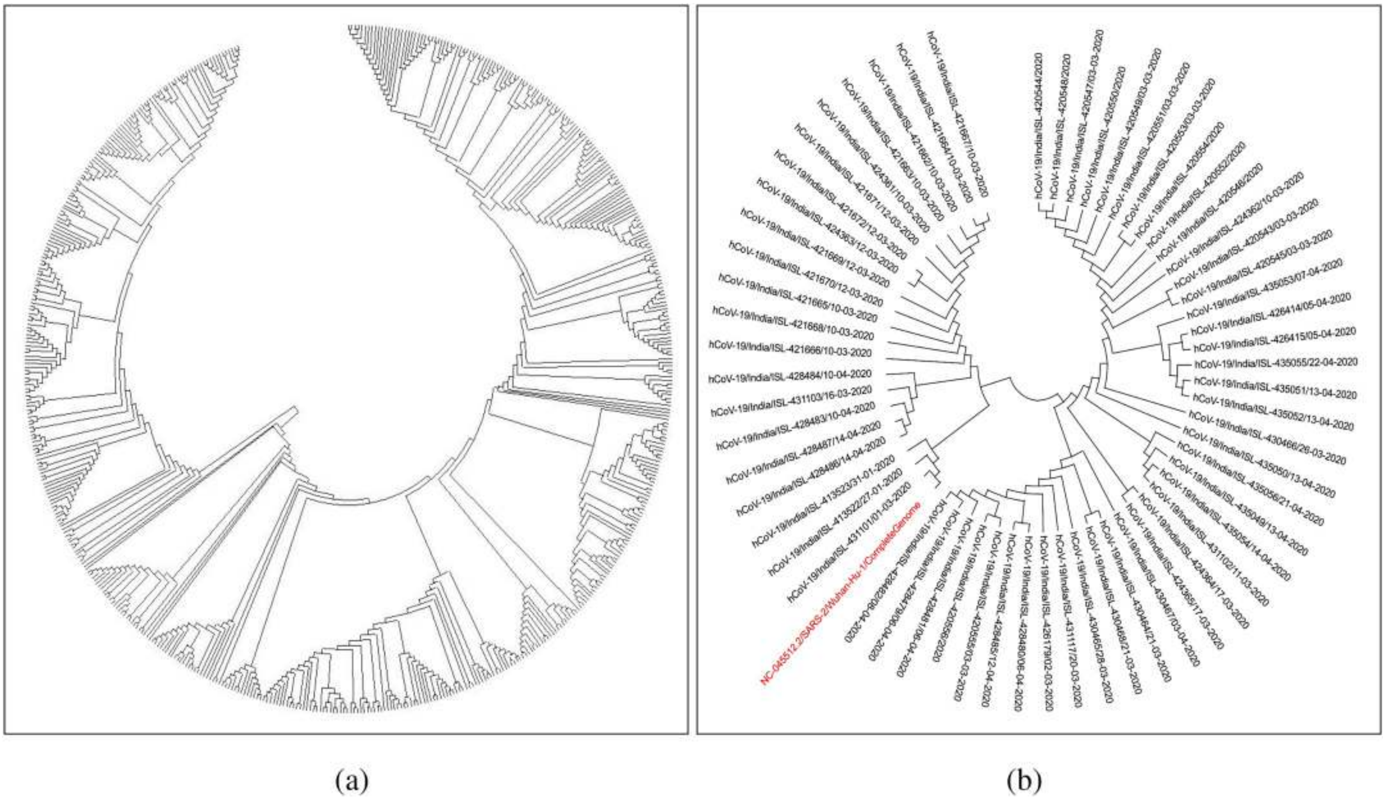
(a)
(b)
hCoV-19/India/ISL-421667/10-03-2020
hCoV-19/India/ISL-421664/10-03-2020
hCoV-19/India/ISL-421662/10-03-2020
hCoV-19/India/ISL-421663/10-03-2020
hCoV-19/India/ISL-424361/10-03-2020
hCoV-19/India/ISL-421671/12-03-2020
hCoV-19/India/ISL-421672/12-03-2020
hCoV-19/India/ISL-424363/12-03-2020
hCoV-19/India/ISL-421669/12-03-2020
hCoV-19/India/ISL-421670/12-03-2020
hCoV-19/India/ISL-421665/10-03-2020
hCoV-19/India/ISL-421668/10-03-2020
hCoV-19/India/ISL-421666/10-03-2020
hCoV-19/India/ISL-428484/10-04-2020
hCoV-19/India/ISL-431103/16-03-2020
hCoV-19/India/ISL-428483/10-04-2020
hCoV-19/India/ISL-428487/14-04-2020
hCoV-19/India/ISL-428486/14-04-2020
hCoV-19/India/ISL-413523/31-01-2020
hCoV-19/India/ISL-413522/27-01-2020
hCoV-19/India/ISL-431101/01-03-2020
NC-045512.2/SARS-2/Wuhan-Hu-1/CompleteGenome
hCoV-19/India/ISL-428482/08-04-2020
hCoV-19/India/ISL-428479/06-04-2020
hCoV-19/India/ISL-428481/06-04-2020
hCoV-19/India/ISL-420556/2020
hCoV-19/India/ISL-420555/03-03-2020
hCoV-19/India/ISL-428485/12-04-2020
hCoV-19/India/ISL-428480/06-04-2020
hCoV-19/India/ISL-426179/02-03-2020
hCoV-19/India/ISL-431117/20-03-2020
hCoV-19/India/ISL-430465/28-03-2020
hCoV-19/India/ISL-430468/21-03-2020
hCoV-19/India/ISL-430464/21-03-2020
hCoV-19/India/ISL-430467/03-04-2020
hCoV-19/India/ISL-424365/17-03-2020
hCoV-19/India/ISL-424364/17-03-2020
hCoV-19/India/ISL-431102/11-03-2020
hCoV-19/India/ISL-435054/14-04-2020
hCoV-19/India/ISL-435049/13-04-2020
hCoV-19/India/ISL-435056/21-04-2020
hCoV-19/India/ISL-435050/13-04-2020
hCoV-19/India/ISL-430466/26-03-2020
hCoV-19/India/ISL-435052/13-04-2020
hCoV-19/India/ISL-435051/13-04-2020
hCoV-19/India/ISL-435055/22-04-2020
hCoV-19/India/ISL-426415/05-04-2020
hCoV-19/India/ISL-426414/05-04-2020
hCoV-19/India/ISL-435053/07-04-2020
hCoV-19/India/ISL-420545/03-03-2020
hCoV-19/India/ISL-420543/03-03-2020
hCoV-19/India/ISL-424362/10-03-2020
hCoV-19/India/ISL-420546/2020
hCoV-19/India/ISL-420552/2020
hCoV-19/India/ISL-420554/2020
hCoV-19/India/ISL-420553/03-03-2020
hCoV-19/India/ISL-420551/03-03-2020
hCoV-19/India/ISL-420549/03-03-2020
hCoV-19/India/ISL-420550/2020
hCoV-19/India/ISL-420547/03-03-2020
hCoV-19/India/ISL-420548/2020
hCoV-19/India/ISL-420544/2020

Fig. 10. Phylogenetic trees of (a) 566 and (b) 60 bootstrap sequences out of 566 Indian SARS-CoV-2 genomes with Reference Genome as marked in red color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

this case 4) plus gap. Further to verify the such genetic variation in Indian SARS-CoV-2 genome phylogenetic analyses is performed so that the evolution can be seen. The phylogenetic analysis is conducted using Maximum Likelihood technique Stuessy (2009) where Neighbor-Joining is used to construct the tree for the visualization of evolution.

3. Results

The results of the experiment are discussed here. Our objective is to identify point mutation as substitution, deletion and insertion initially after performing the multiple sequence alignment. In this regard, we have obtained 933, 2449 and 2 numbers of substitutions, deletions and insertions, respectively. Further, it has been classified as (a) a list of clusters of mutation (b) a list of substitution, deletion and insertion that are not present in the clusters and (c) SNPs that occur more than 1% of the population. The results of the analysis is shown in Fig. 4. As a result after this analysis, we have obtained 100 clusters that cover 92 substitutions and 1683 deletions. The 20 such clusters based on their size and occurrence in genomic sequences are reported in Table 2. The rest of the 80 clusters are reported in supplementary Table S2. Moreover, from this clustering results, 6 deletions in different coding regions are shown in Fig. 5 using BioEdit software.

Once the clusters of mutation points are identified, we have found the mutation points as substitution, deletion and insertion that are not present in the clusters. Consequently 841 substitutions, 766 deletions and 2 insertions are identified and shown in Fig. 4. These 1609 mutation points are reported in supplementary Table S3. Thereafter, 933 substitutions are considered to identify SNPs that are present at least in 6 virus genomes as a clause of 1% of the virus population. As a result, we have found 64 SNPs, out of which 57 SNPs in 6 coding regions while 7 of them are present in 5′-UTR and 3′-UTR. However, in Table 3 and Fig. 6, SNPs that are present in more than 10% of the virus population, i.e. more than 57 virus sequences, are reported and shown as percentage of population having SNP at each genomic location. It describes the size of the virus population for a given genomic position where SNP occurs. The biggest SNPs are identified at 241, 3037, 14,410 and 23,405 genomic coordinates which belongs to the coding regions of ORF1ab and Spike. In these four cases, almost 60% of the virus population are having the change in nucleotide. The rest of the SNPs are shown in supplementary Fig. S1 and reported in supplementary Table S4. Moreover, 6 SNPs in different coding regions are shown in Fig. 7 using BioEdit software. This is to be noted that for each lists of mutations as mentioned in Tables 2 and 3, we have provided genomic coordinates, number of occurrence of mutation in virus genome (frequency of mutation), change in nucleotide, change in amino acid, entropy to measure the change in nucleotide as information contains at that genomic location and mapping with coding region so that mutation point can be identified precisely in Tables 2 and 3. For example, in Table 3 the SNP at 18879 occurs in 117 virus sequences where the change in nucleotide is C > T, change in corresponding amino acid is S > F and the value of entropy is 0.5216. Higher the entropy value signifies that the change in nucleotide is more in sequences.

This is important to mention that the overall results of the mutation as substitution, deletion, insertion, cluster and SNP are shown using BioCircos plots in Fig. 8 where each track shows the frequency of of mutation as histogram using bar and dot plots. Generally, it summarizes all the results visually. Moreover, the computed entropy at each genomic location to have the information of change of nucleotide for the whole population of virus genome is also shown in Fig. 9. This is prepared using BioEdit software. Finally, phylogenetic analysis is shown in Fig. 10 for 566 and its bootstrap samples of 60 virus sequences in order to visualize the variation in trees clearly. It is evident from the trees that the Indian SARS-CoV-2 genomes are having a wide variation which we have also noticed in genomic analysis. These trees are generated using MEGA-X software. In addition to this, the aligned sequences are provided as supplementary[5]for further use.

4. Conclusion

In this paper, we have analyzed 566 Indian SARS-CoV-2 genomes in order to find the mutation as substitution, deletion and insertion as well as SNP. Our analysis has identified 100 clusters of mutations (mostly deletions), 1609 point mutations as substitution, deletion and insertion and 64 SNPs. Out of these 64 SNPs, 57 are present in the 6 coding regions. The purpose of finding SNPs is to identify the genomic locations that can be targeted to classify the virus strain in India. Apart from this, the major advantage is that for personalized vaccine these SNPs could be used to define the dose of the vaccine after identifying the proper strain of the virus. Moreover, for future research, these SNPs can be used to model the proteins and to see its conformational changes so that potential drag can be designed to target such proteins for Indian patients. We are currently working in this direction and also help the other researchers to conduct their research with the use of these SNPs.

Ethics approval and consent to participate

The ethical approval or individual consent was not applicable.

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.

Author contributions

Indrajit Saha: Conceptualization; Data curation; Supervision; Funding acquisition; Formal analysis; Investigation; Methodology; Project administration; Resources; Validation; Visualization; Writing - original draft; Writing - review & editing, Nimisha Ghosh: Formal analysis; Software; Validation; Visualization; Writing - review & editing, Debasree Maity: Conceptualization; Data curation; Methodology; Writing - review & editing, Nikhil Sharma: Methodology; Writing - review & editing, Jnanendra Prasad Sarkar: Methodology; Writing - review & editing, Kaushik Mitra: Conceptualization; 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 GISAID database and reviewers for the valuable comments to improve the article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2020.104457.

References

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., Xia, J., Yu, T., Zhang, X., Zhang, L., 2020. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 395, 507–513. https://doi.org/10.1016/S0140-6736(20)30211-7.

Chothe, S.K., Sebastian, A., Thomas, A., Nissly, R.H., Wolfgang, D., Byukusenge, M., Mor, S.K., Goyal, S.M., Alber, I., Tewari, D., Jayarao, B.M., Kuchipudi, S.V., 2018. Whole-genome sequence analysis reveals unique snp profiles to distinguish vaccine and wild-type strains of bovine herpesvirus-1 (bohv-1). Virology 522, 27–36. https://doi.org/10.1016/j.virol.2018.06.015.

Cui, J., Li, F., Shi, Z., 2019. Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. 17, 181192. https://doi.org/10.1038/s41579-018-0118-9.

Fleischmann, W.R.J., 1996. *Viral Genetics. Galveston* (TX). University of Texas Medical Branch at Galveston.

Jenkins, G.M., Rambaut, A., Pybus, O.G., Holmes, E.C., 2002. Rates of molecular evolution in rna viruses: a quantitative phylogenetic analysis. J. Mol. Evol. 54, 156–165. https://doi.org/10.1007/s00239-001-0064-3.

Jeon, J.S., Won, Y.H., Kim, I.K., Ahn, J.H., Shin, O.S., Kim, J.H., Lee, C.H., 2016. Analysis of single nucleotide polymorphism among varicella-zoster virus and identification of vaccine-specific sites. Virology 496, 277–286. https://doi.org/10.1016/j.virol.2016.06.017.

Lu, I.-N., Muller, C.P., He, F.Q., 2020. Applying next-generation sequencing to unravel the mutational landscape in viral quasispecies. Virus Res. 283, 197963. https://doi.org/10.1016/j.virusres.2020.197963.

Paital, B., Das, K., Parida, S.K., 2020. Inter nation social lockdown versus medical care

[5] http://www.nitttrkol.ac.in/indrajit/projects/COVID-Mutation-India/

against covid-19, a mild environmental insight with special reference to India. Sci. Total Environ. 728. https://doi.org/10.1016/j.scitotenv.2020.138914.

Pavlovic-Lazetic, G.M., Mitic, N.S., Beljanski, M.V., 2004. Bioinformatics analysis of sars coronavirus genome polymorphism. BMC Bioinforma. https://doi.org/10.1186/1471-2105-5-65.

Stuessy, T.F., 2009. Plant Taxonomy: The Systematic Evaluation of Comparative Data. Columbia University Press.

Su, S., Wong, G., Shi, W., Liu, J., Lai, A.C.K., Zhou, J., Liu, W., Bi, Y., Gao, G.F., 2016. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. 24, 490–502. https://doi.org/10.1016/j.tim.2016.03.003.

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.

Wallace, I.M., Blackshields, G., Higgins, D.G., 2005. Multiple sequence alignments. Curr. Opin. Struct. Biol. 15, 261266. https://doi.org/10.1016/j.sbi.2005.04.002.

Weiss, S.R., Navas-Martin, S., 2005. Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus. Microbiol. Mol. Biol. Rev.(4). https://doi.org/10.1128/MMBR.69.4.635-664.2005.

Woo, P.C.Y., Lau, S.K.P., Huang, Y., Yuen, K., 2009. Coronavirus diversity, phylogeny and interspecies jumping. Exp. Biol. Med. 234, 1117–1127. https://doi.org/10.3181/0903-MR-94.

Worldometer, 2020. Coronavirus Disease 2019 (covid-19) Cases in India. https://www.worldometers.info/coronavirus/country/india/.

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