Demographic Analysis of Mutations in Indian SARS-CoV-2 Isolates

Kushagra Agarwal[1], Nita Parekh[1*]
Center for Computational Natural Sciences and Bioinformatics
International Institute of Information Technology, Hyderabad 500032, Telangana, India
*Corresponding author
Email address: nita@iiit.ac.in

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) originally emerged from Hubei province, Wuhan, China, and has caused a worldwide pandemic with millions affected since December 2019. Since the first case of SARS-CoV-2 reported on 27th Jan 2020 in India, over 11 million people have been affected by the virus in the country in the past one year. To understand its demographic distribution, in this study we carried out the early distribution of clades and subclades state-wise based on the analysis of shared mutations in Indian SARS-CoV-2 isolates collected during the period 27th Jan – 27th May 2020. A total of 1279 mutations have been identified in 685 Indian isolates during this period. Principal component analysis on the mutational profile of Indian isolates revealed clusters corresponding to Nextstrain's clades and some state-specific clusters. Phylogenetic analysis of these isolates indicates multiple independent sources of introduction of the virus in the country. It is observed that clade 20A defining mutations C241T (ORF1ab: 5′ UTR), C3037T (ORF1ab: F924F), C14408T (ORF1ab: P4715L), and A23403G (S: D614G) are the predominant mutation set in Indian isolates during this period. Higher number of coronavirus cases were observed in certain states, *viz.*, Delhi, Tamil Nadu, and Telangana. Genetic analysis of isolates from these states revealed a cluster with shared mutations, C6312A (ORF1ab: T2016K), C13730T (ORF1ab: A4489V), C23929T, and C28311T (N: P13L). Contact tracing led the spread of this cluster to a super spreader event, the Tablighi Jamaat religious congregation, and the efficacy of lockdown in containing it. Analysis of region-specific shared mutations carried out to understand the large number of deaths in Gujarat and Maharashtra identified shared mutations defining subclade, I/GJ-20A (C18877T, C22444T, G25563T (ORF3a: H57Q), C26735T, C28854T (N: S194L), C2836T) in Gujarat and two sets of co-occurring mutations C313T, C5700A (ORF1ab: A1812D) and A29827T, G29830T in Maharashtra. From the genetic analysis of mutation spectra of Indian isolates, the insights gained in its transmission, geographic distribution, containment, and impact are discussed.

Introduction

The coronavirus disease 2019 (COVID-19) pandemic is caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), a betacoronavirus belonging to the Coronaviridae family. It exhibits about 79% similarity to SARS-CoV and 50% similarity to MERS-CoV, both of which also originated from bats in 2002 and 2012, respectively [1,2]. Large variation in the rate of infectivity and fatality due to COVID-19 is observed across different countries, with a similar trend observed across various states of India. To understand at the genetic level the role of acquired mutations in the circulating SARS-CoV-2 virus and their possible impact on the spread and virulence, a detailed analysis of 685 Indian SARS-CoV-2 isolates obtained from GISAID [3] during the period 27th Jan – 27th May 2020 is carried out. During the first half of this period (Jan-Mar), positive cases were mainly associated with the travel history of the individuals or their close contacts. Here, the distribution of the virus in

different parts of the country based on mutational analysis of the infected individuals with travel history is discussed. The analysis of genetic variations accumulated is expected to indicate the impact of contact tracing, quarantine, and lockdown in containing the spread of COVID-19 during this period. A detailed state-wise distribution of shared mutations and their global distribution across the world is carried out to understand the transmission and virulence of the disease within and between states. We expect this to help in identifying mutations responsible for the large variation in the number of cases and deaths across different states in India during the early period and identify corresponding subclades of Indian isolates.

Materials and Methods

For this study 705 Indian SARS-CoV-2 viral isolates sequence data was obtained from GISAID [3] corresponding to the period 27th Jan – 27th May 2020. Of these 20 had incomplete metadata or low-quality genomes and were discarded and only 685 isolates data was considered for analysis. Comparison of 685 Indian SARS-CoV-2 isolates with Wuhan-1 isolate (NC_045512.2) as reference revealed a total of 1279 variations during the period 27th Jan – 27th May 2020 (summarized in **Supplementary file S1.xlsx**). Phylogenetic analysis of these isolates was carried out using the resources Augur and Auspice from Nextstrain [4]. For identifying India-specific and state-specific mutations, diversity (entropy) plots that depict the amount of variability at each site in the viral genome, were obtained using the Nextstrain resources. To identify state-specific clusters based on shared mutations during the initial period of the pandemic, principal component analysis (PCA) was performed on the mutational profile of Indian isolates. To assess the impact of India-specific non-synonymous mutations on the stability of the protein, two machine learning based tools, I-Mutant2.0 [5] and MUpro [6], were used. These tools predict both the direction of the change in protein stability and associated $\Delta\Delta G$ values ($\Delta\Delta G = \Delta G$ (mutated protein) - ΔG (wildtype) in Kcal/mol, pH = 7, Temperature = 298K) for a protein structure or the extent of impact of single-site mutation in a protein sequence. The $\Delta\Delta G$ value > 0 indicates increased protein stability, while $\Delta\Delta G < 0$ corresponds to destabilizing effect of the mutation.

Results and Discussion:

Clade Analysis: The phylogenetic tree of Indian isolates constructed with respect to Wuhan-1 isolate as reference using the Nextstrain pipeline and is shown in Figure 1. It is clear from the figure that all the five clades (defined in the new Nextstrain classification scheme) were present in India during this early period with the following distribution: **19A**: 264 samples (38.54%), **19B**: 44 samples (6.42%), **20A**: 300 samples (43.80%), **20B**: 75 samples (10.95%), **20C**: 2 samples (0.29%). The earliest recorded entry of SARS-CoV-2 in the country is 27th Jan 2020 from Kerala with a travel history from Wuhan, China (Acc. ID: EPI_ISL_413522), belonging to the root clade 19A. Another early record is a sample of clade 19B, also from Kerala with a travel history from Wuhan, China on 31st Jan 2020 (Acc. ID: EPI_ISL_413523). That is, by January end the two clades, 19A and 19B, were introduced into the country. By January end, these two clades had globally spread across most parts of the world, including India. The earliest sample corresponding to clade 20A is dated 3rd March, of a tourist from Italy, Europe (Acc. ID: EPI_ISL_420543). Two samples corresponding to clade 20B were observed during the same time, one having contact with another Indian with travel history from Italy in (Acc. ID: EPI_ISL_426179), dated 2nd March, and the other on 29th February (Acc. ID: EPI_ISL_414515), with no state or travel history available. During this period only 2 samples of clade 20C were observed (Acc. ID: EPI_ISL_435051 and EPI_ISL_435052), dated 13th

April in Gujarat with no travel history or contact with anyone with travel history. These results indicate early community spread of the virus in the country.

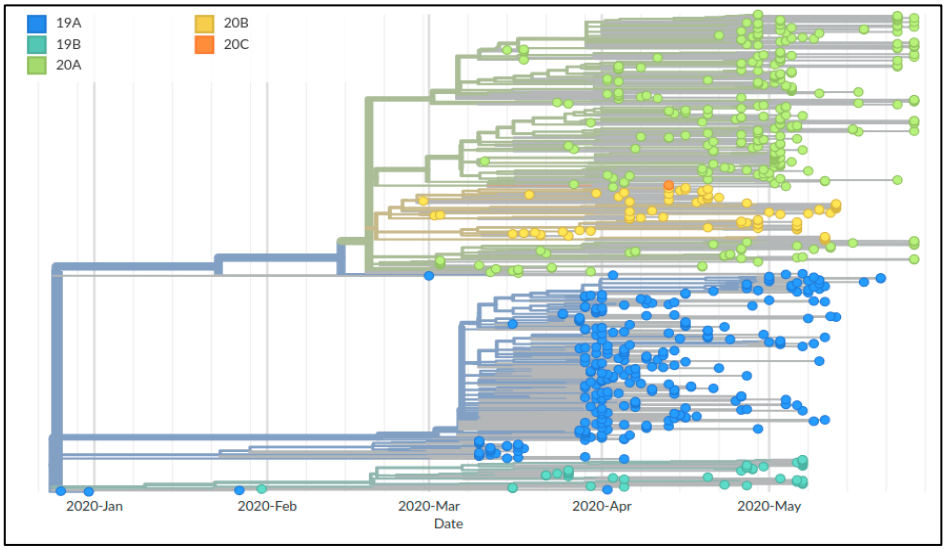


Fig. 1. Divergence of Indian SARS-CoV-2 isolates during the period 27th Jan – 27th May 2020 using Wuhan-1 isolate as reference.

The largest number of isolates belonged to clade 20A (43.80%) in accordance with the global trend, with the next large cluster corresponding to root clade 19A (38.54%). These two clades comprise more than 82% of the reported isolates in India till 27th May. Our analysis revealed a major subcluster of clade 20A, mainly in April, with defining mutations. It would be interesting to study the state-wise distribution of this clade along with travel history to know if there was any community transmission during the lockdown period. Similarly, a large cluster within clade 19A was observed in accordance with an earlier study [7].

State-wise Distribution of Clades: Out of 685 Indian isolates only 658 isolates were considered for state-wise distribution analysis as 27 isolates had no state information available. It is clear from Figure 2 that clade 20A is predominantly observed in Gujarat (178/201), followed by West Bengal (30/45), Odisha (22/46), Madhya Pradesh (16/19), and Maharashtra (16/80), with the earliest reported sample from Gujarat being dated 5th April (Acc. ID: EPI_ISL_426414). Multiple independent entries of this clade are indicated based on clusters with shared mutations (discussed later). The distribution of the second major clade 19A is predominant in Telangana (75/97), followed by Delhi (55/76), Maharashtra (31/80), and Tamil Nadu (19/34). Very few isolates of clade 19B (Odisha (17), Maharashtra (10), Gujarat (8) and West Bengal (5)), clade 20B (Maharashtra (23), Telangana (15), Tamil Nadu (15), and Delhi (7)) and clade 20C (only from Gujarat (2)) are observed. During the analysis it was observed that the genomic sequences of a few isolates from Odisha were shorter in length with missing bases; 1 - 29 bases in 5′ UTR (38 isolates) and 29686 - 29903 bases in 3′ UTR (39 isolates), which are likely to be due to sequencing artifact as the majority of these are from the same sequencing laboratory. The isolates from Odisha also contain other deletion regions, 23842 - 24400 bases in S gene, 26306-26524 bases in E gene, 27527 – 28033 bases covering the genes ORF7b and ORF8 in 8 samples, and 28462 - 28680 bases in the N gene (10 samples), 29000 – 29685 bases covering the genes N and ORF10 (15 samples), which may be region-specific.

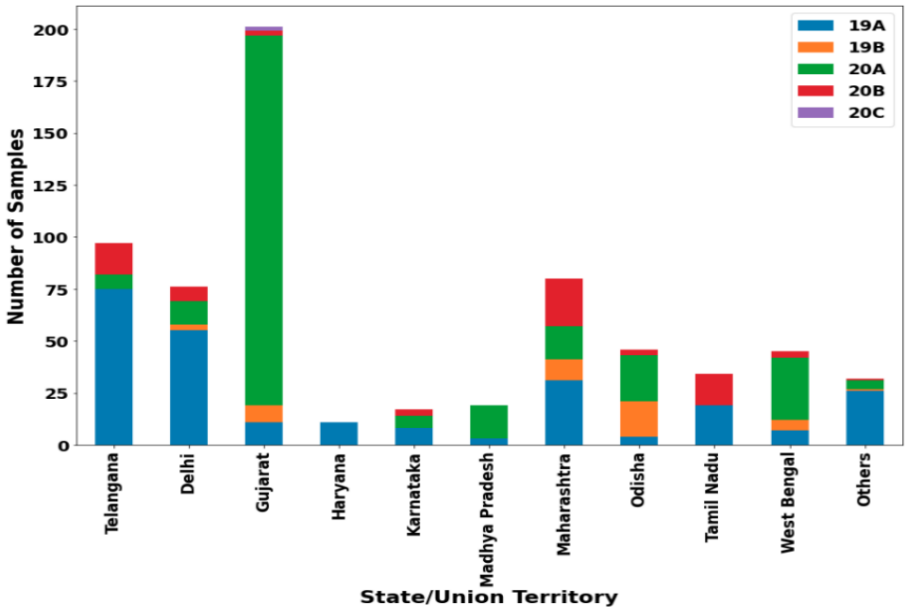


Fig. 2. State-wise distribution of clades. Clade 20A is predominant in Gujarat while the root clade 19A is observed majorly in Telangana, Delhi, and Maharashtra.

Analysis of the Mutational Profile of Indian Isolates

A total of 1279 variations were identified on comparing 685 Indian SARS-CoV-2 isolates with Wuhan-1 isolate as reference State-wise distribution of these variations was carried out to understand at the genetic level the high number of cases in the states of New Delhi, Tamil Nadu, Telangana, and high fatality rates in Gujarat and Maharashtra. To identify India-specific or state-specific variations, diversity (entropy) plots obtained using the Nextstrain resources were analysed at each site in the viral genome.

Most frequent mutations in the Indian isolates: The important mutations in Indian isolates and their distribution state-wise is summarized in Table 1. The mutation C241T in the 5′ UTR region of ORF1ab is the most common mutation with incidence in more than half of the Indian isolates sequenced. State-wise, high prevalence is observed in Gujarat (183/201), followed by Maharashtra (39/80) and West Bengal (33/45). Being a non-genic mutation, the C241T mutation does not cause an amino acid change. However, since it is part of the stem-loop region upstream of the ORF1ab gene, it may involve in differential RNA binding affinity to the ribosome and affect translational efficiency of the virus, though no change in the RNA structure was observed for the mutant form [8]. The study by Chaudhari et al [9] analysed the role of C241T mutation in the interaction of the stem-loop region with the host replication factors, MADP1 Zinc finger CCHC-type and RNA-binding motif 1 (hnRNP1). The molecular docking and molecular dynamics simulations suggest reduced replication efficiency of the virus within the host which may result in less mortality.

Another mutation that has been widely studied globally and has a similar distribution in the Indian samples is the D614G mutation in the spike protein caused by the nucleotide change, A23403G. It was first observed in samples from China and Germany in late January and is now

Table 1. Important mutations in Indian isolates given with their frequency, state distribution, and protein stability analysis scores. (GJ: Gujarat, WB: West Bengal, MH: Maharashtra, TL: Telangana, DL: Delhi, TN: Tamil Nadu).

Variant Description	Nucleotide Mutation	Protein	Amino Acid mutation	No. of Samples in India (%)	State-wise representation (%)	I-Mutant2.0 Score	MUpro Score
Clade 20A	C241T	5' UTR	-	378 (55.2)	GJ (91.0), WB (73.3), MH (48.8)	-	-
MH-specific	C313T	ORF1ab (nsp1)	-	39 (5.7)	MH (25.0)	-	-
Subclade I/GJ-20A	C2836T	ORF1ab (nsp3)	-	51 (7.4)	GJ (25.4)	-	-
Clade 20A	C3037T	ORF1ab (nsp3)	-	374 (54.6)	GJ (90.6), WB (73.3), MH (48.8)	-	-
MH-specific	C5700A	ORF1ab (nsp3)	A1812D	33 (4.8)	MH (26.2)	-0.78	-0.25
Co-occurs with subclade I/A3i	C6310A	ORF1ab (nsp3)	S2015R	44 (6.4)	TL (15.5), DL (10.5)	-0.88	-0.52
Subclade I/A3i	C6312A	ORF1ab (nsp3)	T2016K	228 (33.3)	TL (73.2), DL (69.7), TN (55.9)	-1.37	-1.99
Co-occurs with subclade I/A3i	G11083T	ORF1ab (nsp6)	L3606F	258 (37.7)	TL (72.2), DL (71.0), TN (55.9)	-0.05	-1.29
Subclade I/A3i	C13730T	ORF1ab (nsp12)	A4489V	239 (34.9)	TL (74.2), DL (71.0), TN (55.9)	-0.53	-0.79
Clade 20A	C14408T	ORF1ab (nsp12)	P4715L	367 (53.6)	GJ (88.6), WB (73.3), MH (46.2)	-0.8	0.60
Subclade I/GJ-20A	C18877T	ORF1ab (nsp14)	-	126 (18.4)	GJ (52.2)	-	-
Co-occurs with subclade I/A3i	C19524T	ORF1ab (nsp14)	-	60 (8.8)	TL (20.6), DL (13.2)	-	-
Subclade I/GJ-20A	C22444T	S	-	71 (10.4)	GJ(31.8)	-	-
Clade 20A	A23403G	S	D614G	373 (54.4)	GJ (89.6), WB (73.3), MH (48.8)	-1.94	-1.49
Subclade I/A3i	C23929T	S	-	224 (32.7)	TL (71.1), DL (69.7), TN (47.0)	-	-
Subclade I/GJ-20A	G25563T	ORF3a	Q57H	134 (19.6)	GJ (53.7)	-1.12	-0.92
Subclade I/GJ-20A	C26735T	M	-	123 (18.0)	GJ (50.2)	-	-
Subclade I/A3i	C28311T	N	P13L	236 (34.4)	TL (76.0), DL (68.4), TN (55.9)	0.11	-0.80
Subclade I/GJ-20A	C28854T	N	S194L	73 (10.6)	GJ (32.8)	0.45	-0.67
Clade 20B	G28881A	N	R203K	75 (10.9)	TN(44.1), MH (28.8), TL (15.5)	-2.26	-1.06
Clade 20B	G28882A	N	R203K	75 (10.9)	TN (44.1), MH (28.8), TL (15.5)	-2.26	-1.06
Clade 20B	G28883C	N	G204R	75 (10.9)	TN (44.1), MH (28.8), TL (15.5)	0	-0.68
MH-specific	A29827T	3′ UTR	-	44 (6.4)	MH (55.0)	-	-
MH-specific	G29830T	3′ UTR	-	61 (8.9)	MH (73.8)	-	-

the most prevalent mutation worldwide. A similar trend is observed with over 50% representation (373/685) in Indian isolates, and a high predominance in Gujarat (180/201). The D614G mutation replaced aspartic acid with glycine in the Receptor Binding Domain (RBD) of Spike protein. Glycine being less bulky than aspartic acid, contributes to a more flexible hinge region in the Spike protein, enabling more efficient receptor-ligand binding. This provided a selective advantage to the virus in infection and transmission, making it predominant all over the world [10]. Plante *et al* (2020) [11] characterized the replication of D614 and G614 viruses in a primary human airway tissue culture model and showed that the mutation enhances viral replication through increased virion infectivity and enhances the stability of the virus. A recent study by Daniloski et al (2021) [12] suggests possible mechanism for increased infectivity of the G614 variant is probably due to it being more resistant to proteolytic cleavage. This stability confers G614 variant efficient transmission ability. The protein stability analysis using the two tools, I-Mutant2.0 and MUpro indicate that the mutation has a destabilizing effect due to increased flexibility provided by glycine. The D614G mutation is found to co-occur with C241T, C3037T, and C14408T [13]. The C14408T mutation within the RNA-dependent RNA polymerase (RdRp) encoding region of ORF1ab is a missense mutation that leads to an amino acid change from Proline to Leucine (P4715L). It is found on the Nsp12 region which is involved in replication and pathogenesis and is a potential target for antiviral candidates in coronaviruses. This mutation too is observed in more than half of the Indian samples with predominance in Gujarat (178/201), Maharashtra (37/80), and West Bengal (33/45). From Table 1 it may be noted that I-Mutant2.0 analysis predicts decreased stability for the mutation C14408T while MUpro scores indicate stabilizing effect. The nucleotide mutation C3037T is a silent mutation in the Nsp3 protein of ORF1ab (F924F) and no functional significance of this mutation is reported. These four mutations, *viz.*, C241T, C3037T, C14408T, and 23403G are the defining mutations for clade 20A in Nextstrain and are predominantly found in North America and Europe. The haplotype defined by these four co-occurring mutations is now globally the dominant form and a similar trend is observed in India. This haplotype was proposed to be related to the pathogenicity of the virus and correlated with high death rates in Europe [14]. High prevalence of these mutations in Gujarat with similar death rates suggests probable transmission of the virus to Gujarat from North America and Europe, and its containment in the state during lockdown.

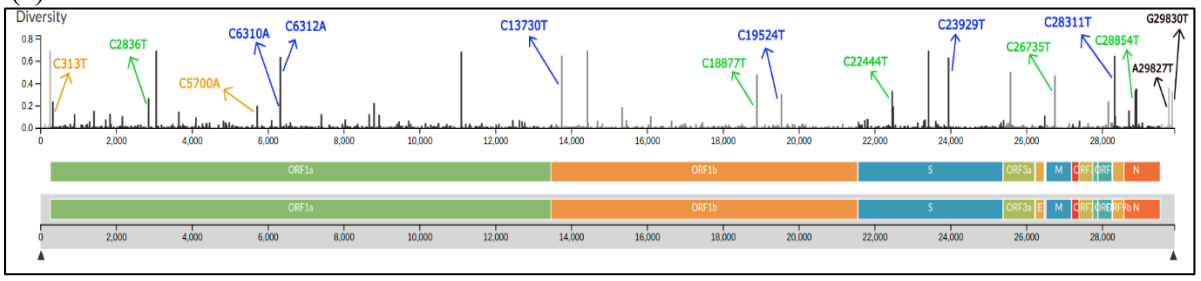
Clade 20B defining mutations, G28881A, G28882A, and G28883C, results in two adjacent amino acid changes R203K and G204R in the nucleocapsid (N) protein. This trinucleotide-bloc mutation, 28881-28883: GGG > AAC, is reported to result in two sub-strains of SARS-CoV-2, *viz.*, SARS-CoV-2g and SARS-CoV-2a. The AAC genotype is observed from March onwards globally and in India (75 samples) its prevalence is mainly observed in the states of Maharashtra (23/80), Tamil Nadu (15/34), and Telangana (15/97). The RG > KR mutation is observed to disrupt the S-R motif by the introduction of Lysine between them. This may affect the phosphorylation of the SR-rich domain that plays an important role in the cellular localization and translation inhibitory function of the N protein [15]. Earlier experimental work on SARS-CoV-1 has shown reduced pathogenicity on deleting part of the SR domain [16]. These mutations also reduce the number of miRNA binding sites from seven to three and thereby increase the likelihood of viral infection [17]. The stability analysis of these three variants using I-Mutant2.0 and MUpro indicate decrease in protein stability from Table 1.

Mutations specific to Indian isolates. The diversity (entropy) plots for SARS-CoV-2 isolates from India and globally are given in Figure 3 (a) and 3 (b), respectively. Sites exhibiting higher entropy in Indian isolates compared to global isolates are marked in Figure 3 (a) and the results are also summarized in Table 1. Four of these mutations, C6312A (T2016K) and C13730T

(A4489V) in ORF1ab, C23929T (synonymous) in Spike protein, and C28311T (P13L) in N gene have been reported in an earlier study as India specific mutations defining subclade I/A3i [7]. In Figure 3, two more India-specific mutations, *viz.*, C6310A (S2015R) and C19524T (synonymous) in ORF1ab are identified to be associated with this subclade. These two mutations correspond to two branch-points of the subclade I/A3i and are mainly found in a subset of samples from Delhi and Telangana.

Our analysis also revealed two other mutations C313T (synonymous) and C5700A (A1812D) in ORF1ab which co-occur with high frequency in India compared to their global presence. These mutations define a unique cluster of viral isolates predominantly observed in Maharashtra and Telangana states. The mutation C5700A is observed to branch out of clade 20B primarily, but some samples from clade 20A also exhibit this mutation, as seen in Figure 4. This mutation is unique to the Indian region (~5% frequency) with no presence globally (Figure 3(a) and 3(b)). It co-occurs with the synonymous mutation, C313T, and is observed in 33 Indian samples: Maharashtra (21), Telangana (8) and Gujarat (4). The first observed sample with this mutation was on 4th April 2020 in Maharashtra and its presence has also been reported in an earlier study on 90 sequences from western India [18]. The percentage of samples carrying this mutation is observed to be increasing and by July 2020, about 10% of Indian samples are found to be carrying it, indicating local transmission. The mutation C5700A lies in the Nsp3 region of ORF1ab, which forms multi-subunit assemblies with other non-structural proteins for the formation of replication-transcription complex [19]. It has the ability to alter surface electrostatic environment in the proximity of Nsp3's viral protease domain [20]. The mutation is seen to have destabilizing effect on the protein from I-Mutant2.0 and MUpro scores in Table 1. In Figure 3(a), the mutations shown in green, C2836T, C18877T, C22444T, C26735T, C28854T are specific to Gujarat viral isolates while those in black, A29827T, and G29830T are observed in Maharashtra. The significance of these mutations is discussed in detail below.

(a)



(b)

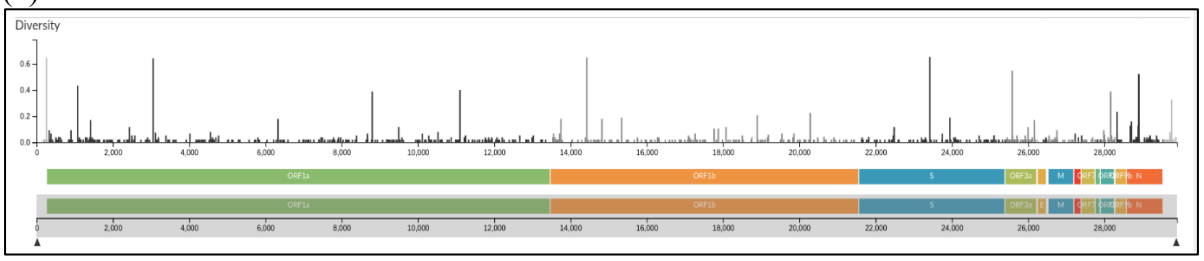


Fig. 3. Diversity plots shown for isolates from (a) India and (b) World. In (a) 15 mutations predominant in India are marked in: Blue (clade 19A): C6310A, C6312A, C13730T, C19524T, C23939T and C28311T, Green (clade 20A): C2836T, C18877T, C22444T, C26735T, C28854T, Orange (clade 20B): C313T, C5700A, and Black (observed in more than one clade): A29827T, G29830T.

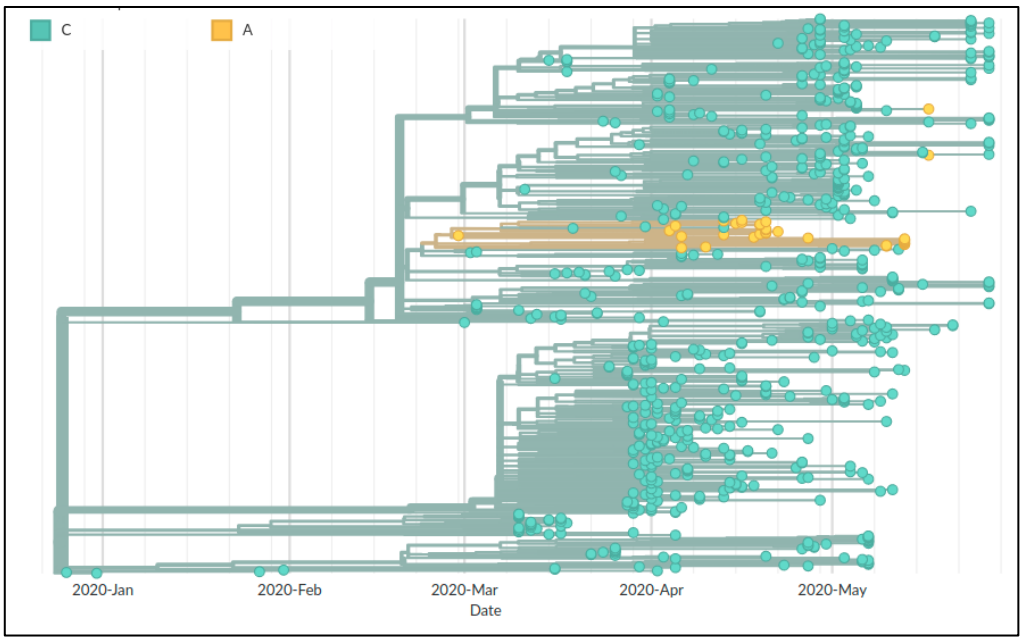


Fig. 4. Distribution of nucleotide mutation C5700A shown on the phylogenetic tree of Indian SARS-CoV-2 isolates.

Geographical distribution of Indian subclade I/A3i. We observe about one-third of the Indian isolates (219/685) defined by the co-occurring mutations C6312A, C13730T, C23929T, and C28311T, i.e., the subclade I/A3i. It is shown as a yellow cluster branching out of clade 19A in Figure 5. From the state-wise distribution of this subclade depicted in Figure 6(a), its predominance is observed in Telangana (69/97, ~71%), Delhi (52/76, ~68%), Haryana (11/11) and Tamil Nadu (16/34) states. The mutation C13730T (ORF1ab: A4489V) lies in the RdRp protein's NiRAN domain, which is involved in RNA binding and nucleotidylation, which is essential for viral replication [21]. The nucleocapsid protein (N) is required for viral RNA replication, transcription, as well as genome packing [22, 23]. The mutation C28311T (N: P13L) is located in the protein's intrinsically disordered region and may impair the terminal domain's RNA-binding function [24, 25]. A previous study evaluated how this mutation may alter protein-protein interaction and proposed its impact on virus stability, potentially contributing to lower pathogenesis [26]. From the protein stability analysis using I-Mutant2.0 and MUpro given in Table 1, the two non-synonymous mutations representing subclade I/A3i (C6312A and C13730T) decrease the stability. The effect of third mutation (C28311T) is not conclusive as I-Mutant results in positive score indicating increase in stability while MUpro resulted in negative score indicating decrease in stability. Mutation C6310A which is also a part of this subclade is also observed to decrease stability.

Analysis of this subclade is important both scientifically and epidemiologically as the mutations defining it are found in ~32% of the Indian samples sequenced, while outside India its distribution is very low (~3.5%). This clearly hints at early community transmission due to some super spreader event during March-April, as it is highly unlikely that around one-third of the samples sharing the same set of mutations could have arisen by multiple independent entries with international travel history, especially when its presence globally is negligibly small. The first reported entry of this clade in India was on 16th March 2020 in Telangana from an Indonesian citizen visiting India and globally the first incidence of this subclade was observed in a sample from Saudi Arabia, dated 7th March 2020, and its distribution is predominantly found in Asia (mostly Singapore and Malaysia) as shown in Figure 6(b). The coincidence of

subclade I/A3i in the country following Tablighi congregation held during early March, the first reported case of this subclade from Saudi Arabia on 7th March, and several Indonesian citizens identified with these mutations in Delhi during that period, all hint at the Tablighi congregation event as the likely cause of spread of this subclade. No isolates belonging to this subclade were observed after May, except for 5 samples in India with the last one dated 13th June 2020 (according to data available in Nextstrain). This indicates that the spread of subclade I/A3i had been largely contained during lockdown with efforts of contact tracing and quarantining the infected individuals. Similarly, no further global spread of this subclade is observed, probably due to the ban on air travel and the efficacy of contact tracing and quarantining of COVID-19 positive individuals. Thus, we see that based on genetic analysis it is possible to identify the chain of transmission of infection and the success of measures used in its containment.

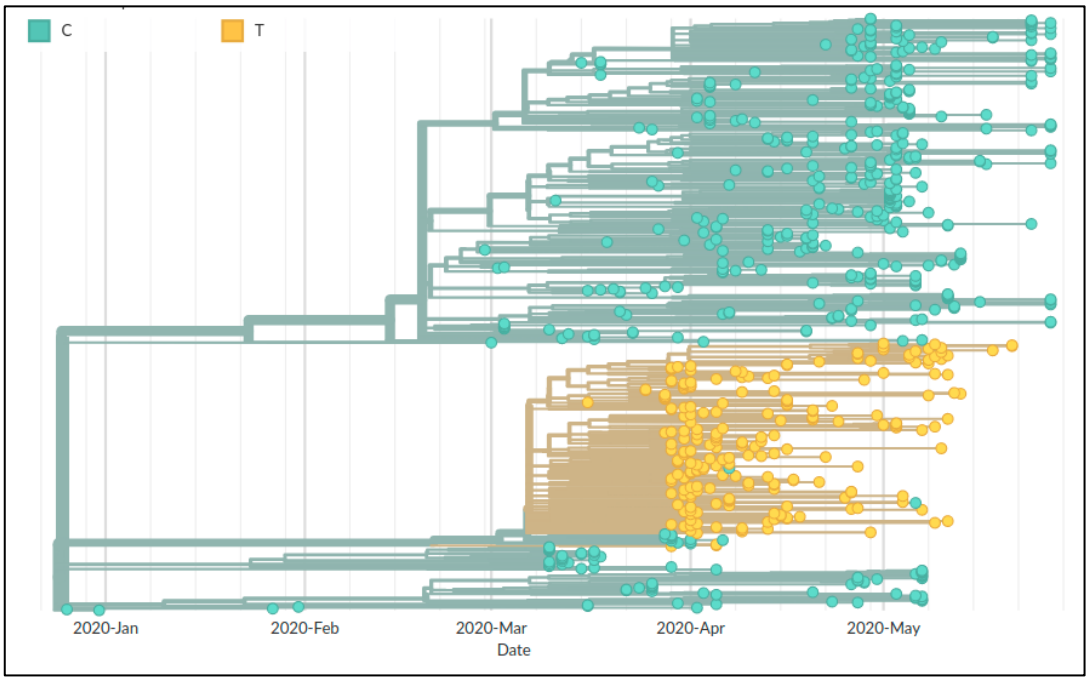


Fig. 5. Viral isolates carrying the synonymous mutation C23929T in Spike protein is depicted in yellow colour, corresponding to subclade I/A3i.

Analysis of samples with travel history from foreign nations. With the ban on international flights were banned and lockdown imposed in the country from March 23rd, 2020, no new transmissions were likely from outside the country. Further, because of lockdown samples may accumulate region-specific mutations and the analysis of such shared mutation sets can help in understanding the extent of local transmission and containment of the circulating strains in a localised region. The analysis of 14 samples with travel history from Iran, 11 from Italy, 8 from Indonesia, and 2 from China has been discussed earlier in introducing clades 19A (by EPI_ISL_413522) and 19B (by EPI_ISL_413523) in India. In the 14 samples from Iran, 13 samples had the mutations, G1397A, T28688C, G29742T. Amongst them, 11 also had the mutations, C884T, G8653T, A29879G, A29883T, A29901T. These mutations were observed in very few other Indian isolates; Ladakh (6/6), Kargil (1/1), and Maharashtra (1/80) indicating the efficacy in quarantining and contact tracing during the lockdown period. All the 11 samples from Italy contained the mutations C3037T, C241T, C14408T (ORF1ab: P4175L), and A23403G (S: D614G), which are clade 20A defining mutations (in Nextstrain). These mutations are found in over 50% of the Indian samples and with highest frequency in Gujarat

(~90% samples). Also, all the 8 samples from Indonesia sampled in Delhi around the time the Tablighi Jamaat religious congregation was conducted, contained the four defining mutations of the subclade I/A3i, *viz.*, C23929T, C6312A, C13730T, and C28311T. As discussed above, this mutation set was mainly observed in Telangana, Delhi, and Tamil Nadu but no further increase was seen in the number of samples containing it.

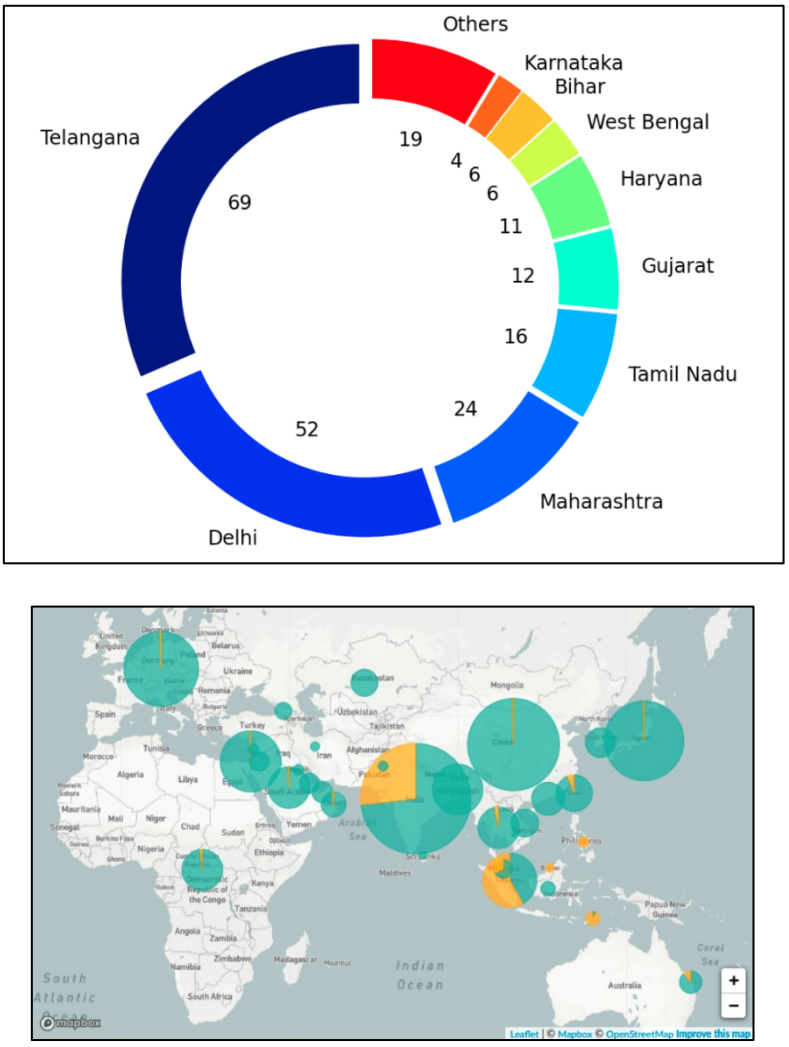


Fig. 6 (a) State wise distribution of subclade I/A3i isolates in India. **(b)** Global distribution (zoomed into Asia) of the mutation C23929T corresponding to subclade I/A3i is shown.

PCA analysis of Indian SARS-CoV-2 isolates

Since travel between states was restricted because of lockdown, the number of samples with shared mutations is expected to increase as a consequence of local community transmission and their analysis can reveal in state-specific mutational profiles. Thus, to see if any state-specific clustering is observed during the initial period of the pandemic, principal component analysis (PCA) on the mutational profile of the Indian isolates was performed. This would provide insights into local community transmission, if any, during the initial period of the pandemic. Figure 7 shows the plot of the first two principal components that captured over 35% variance in this high dimensional data. To identify the mutations that had larger influence on PC1, the mutations were sorted based on their loading scores. Top 10 mutations thus identified are: A23403G (Spike: D614G), C3037T, C241T, C14408T (ORF1ab: P4715L), G11083T

(ORF1ab: L3606F), C28311T (N: P13L), C6312A (ORF1ab: T2016K), C13730T (ORF1ab: A4489V), C23929T, and G25563T (ORF3a: Q57H). This is not surprising as these are also the top 10 most common mutations in the Indian isolates. Thus, we observe that lockdown, contact tracing and quarantine measures were successful in containing the existing strains during the early period of pandemic.

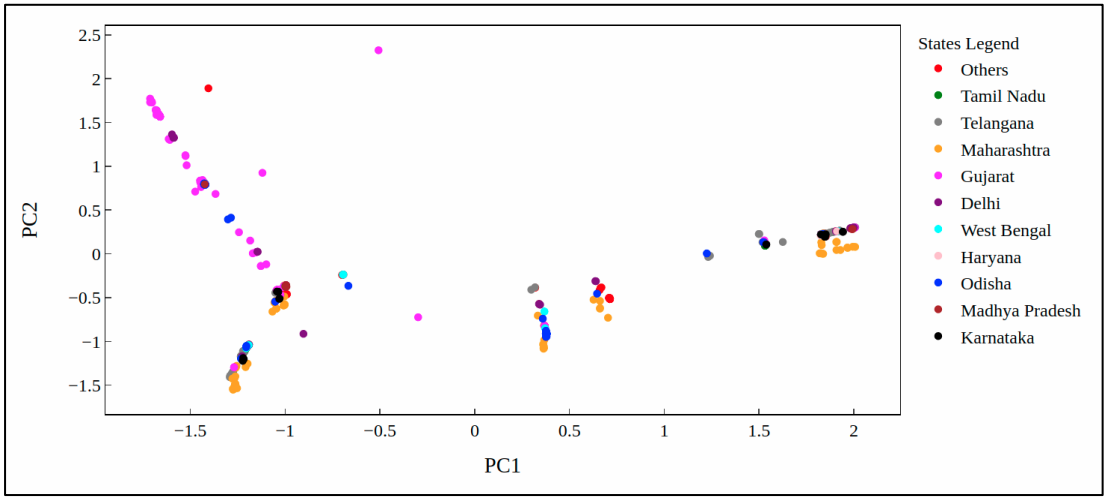


Fig. 7. PCA plot of the 685 Indian viral isolates coloured state-wise.

It may be noted from Figure 7 that the samples from Gujarat shown in 'pink' form a distinct cluster. Also, the samples from Telangana, Delhi, and Haryana states are found to cluster to the right in the plot. This is due to the shared mutation profile of subclade I/A3i and its prevalence in these states. The samples from Maharashtra are observed to be scattered throughout the plot, though in closely grouped clusters. A detailed analysis revealed two sets of co-occurring mutations in the isolates from Maharashtra (discussed later). Majority of the top 10 mutations thus identified are clade defining mutations according to Nextstrain's definition. And this is indeed observed on mapping the clade information on the clusters observed. Thus, on analysing the clade distribution of these clusters, a clear association of the clusters with the clade was observed.

Mutational Profile of Gujarat *vs* Rest of India (RoI)

It is observed that certain countries like Italy, UK, Spain, etc. had large number of mortality cases. This has been pointed to different severe strains of the SARS-CoV-2 virus in these countries and larger aging population. India being a very vast country, it would be interesting to study if a similar pattern is observed across its different states. We analysed the infection and death rates across states and in Table 2 top few states with high severity of COVID-19 cases are listed. We observe that Gujarat state has the highest reported death rate of 5.12%, followed by Maharashtra (4.19%), while the country average is 2.67%, as on 11th July 2020. The recorded number of deaths in Gujarat is 2008 out of a total of 39194 cases while Karnataka (31105), Telangana (30946), and Uttar Pradesh (32363) with similar number of cases, recorded much fewer fatalities, 486, 331, and 862, respectively. Even some of the worst-hit regions such as Delhi and Tamil Nadu had much lower death rates, 3.29%, and 1.39% respectively, compared to Gujarat. To understand this significantly large difference in the percentage of deaths in Gujarat at the genetic level, a detailed analysis of the mutational profile of sequences from Gujarat with that of Rest of India (RoI) was carried out. Table 3 summarizes mutations resulting in amino acid changes that are found mostly in samples from Gujarat but not with

high frequency in samples from rest of the country. The clade 20A defining mutations were observed to occur in ~ 90% of Gujarat samples. Due to the countrywide lockdown from 25th March 2020, this clade and its sub-clusters were localized in the state, defined by Gujarat-specific mutations given in Table 3. A reverse scenario is also expected, that is, the incidence of other state-specific mutations in Gujarat. These mutations, present in the isolates from RoI but not well represented in Gujarat, are listed in Table 4. It may be noted that subclade I/A3i, unique to Indian samples, and clade 20B are not well represented in Gujarat.

Table 2. Number of deaths *vs* infected cases for few severely affected Indian states is given. (Data as on 11th July 2020 from https://www.mygov.in/corona-data/covid19-statewise-status)

State	Deaths	Total Cases	Death Rate (%)
Gujarat	2008	39194	5.12
Maharashtra	9667	230599	4.19
Delhi	3528	107051	3.29
Tamil Nadu	1765	126581	1.39
Uttar Pradesh	862	32363	2.66
Karnataka	486	31105	1.56
Telangana	331	30946	1.06

Table 3. Mutations observed in SARS-CoV-2 isolates from Gujarat (201 samples) with high frequency compared to Rest of India, RoI (484 samples).

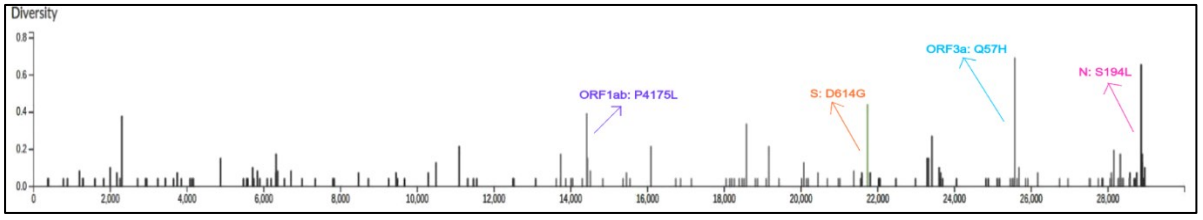
Nucleotide Mutation	Protein	Amino Acid mutation	Count in Gujarat	Frequency in Gujarat (%)	Count in RoI	Frequency in RoI (%)
C241T	5′ UTR	-	183	91.0	195	40.3
C3037T	ORF1ab (nsp3)	-	182	90.6	192	39.7
A23403G	S	D614G	180	89.6	193	39.9
C14408T	ORF1ab (nsp12)	P4175L	178	88.6	189	39.0
G25563T	ORF3a	Q57H	108	53.7	26	5.4
C18877T	ORF1ab (nsp14)	-	105	52.2	21	4.3
C26735T	M	-	101	50.2	22	4.5
C28854T	N	S194L	66	32.8	7	1.4
C22444T	S	-	64	31.8	7	1.4
C2836T	ORF1ab (nsp3)	-	51	25.4	0	0.0

This difference in the mutational profile of isolates from Gujarat with that of the Rest of India is clearly seen in the diversity plots for non-synonymous mutations in isolates in Figure 8 (a) Gujarat, (b) Telangana and (c) India. It may be noted from plots 8(a) and 8(c) that the set of mutations and their frequency is very different in Gujarat isolates compared to that observed in isolates from the whole of India. In contrast, the diversity plots of Telangana (Fig 8(b)) and India (Fig 8(c)) are quite similar. This may be due to the India specific subclade I/A3i being predominant in Telangana. The death rates for Telangana and India are also comparable, while that of Gujarat is strikingly different. This analysis suggests that the characteristic mutations of subclade I/A3i may not be deleterious.

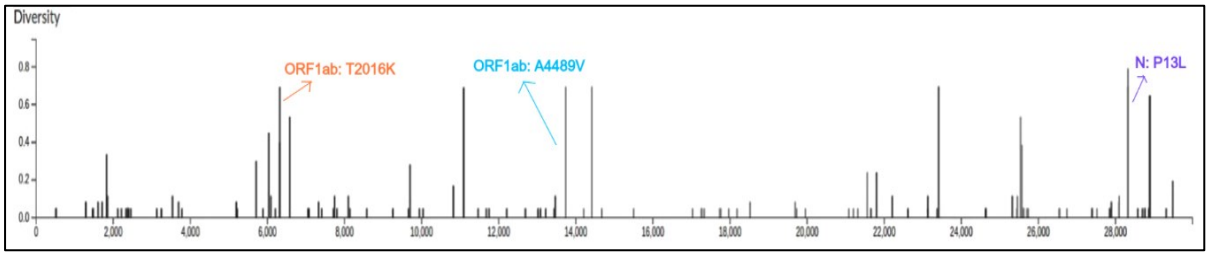
Table 4. Mutations observed in SARS-CoV-2 isolates with higher frequency in Rest of India, RoI (484 samples) but not well represented in Gujarat (201 samples).

Nucleotide Mutation	Protein	Amino Acid mutation	Count in RoI	Frequency in RoI (%)	Count in Gujarat	Frequency in Gujarat (%)
C313T	ORF1ab (nsp1)	-	37	7.6	2	1.0
C5700A	ORF1ab (nsp3)	A1812D	29	6.0	4	2.0
C6310A	ORF1ab (nsp3)	S2015R	38	7.8	6	3.0
C6312A	ORF1ab (nsp3)	T2016K	216	44.6	12	6.0
G11083T	ORF1ab (nsp6)	L3606F	242	50.0	16	8.0
C13730T	ORF1ab (nsp12)	A4489V	227	46.9	12	6.0
C19524T	ORF1ab (nsp14)	-	54	11.2	6	3.0
C23929T	S	-	212	43.8	12	6.0
C28311T	N	P13L	223	46.1	13	6.5
G28881A, G28882A, G28883C	N	R203K, G204R	73	15.1	2	1.0
A29827T	3′ UTR	-	44	9.1	0	0.0
G29830T	3′ UTR	-	61	12.6	0	0.0

A) Gujarat



B) Telangana



C) India

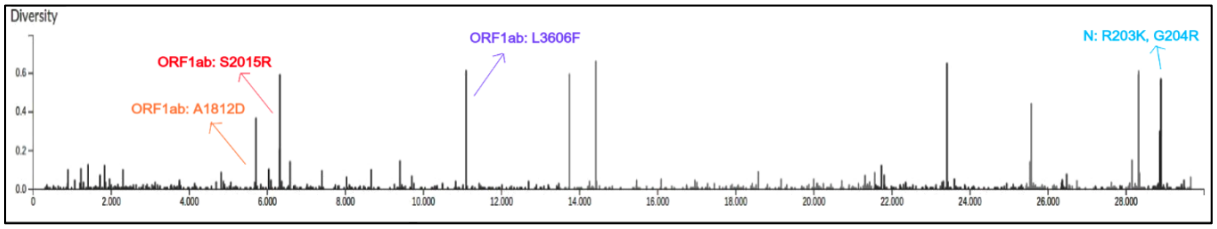


Fig. 8. Diversity plots for non-synonymous mutations in isolates from (a) Gujarat, (b) Telangana, and (c) India clearly exhibit different sets of mutations.

Identification of novel subclade I/GJ-20A and unique mutations in Maharashtra

A detailed analysis of isolates from Gujarat revealed a novel subclade, I/GJ-20A defined by a shared set of mutations C18877T, G25563T (ORF3a: Q57H), and C26735T. This set of mutations are not well represented in the rest of India (RoI) and have been reported in a study by Joshi *et al* (2021) [27]. Analysis of these mutations will likely help in explaining high death rates in Gujarat compared to the other Indian states. These three mutations listed in Table 3, co-occur in 98 samples from Gujarat and form a subclade of clade 20A in the phylogenetic tree. The Q57H mutation caused by G25563T is present on the outer surface of ORF3a protein and has been implicated in inflammation, antiviral responses, and apoptosis control. Mutations in ORF3a are of significance as the ORF3a protein is involved in regulating immunological responses in the host, including "cytokine storm". The mutation results in significant conformational changes to the protein and forms a more stable quaternary structure thereby increasing the viral particle release [26, 28]. It is indicated to increase the transmission but results in reduced deaths. According to I-Mutant2.0 and MUpro analysis, Q57H mutation decreases the stability of the protein. Other mutations that are part of this subclade are C2836T, C22444T, and C28854T (N: S194L). The S194L mutation due to C28854T resides in the central region of the nucleocapsid protein that is essential for oligomerization [29, 30] and is also expected to alter the protein structure [31]. The impact of mutation C28854T on protein stability is inconclusive. It is a stabilizing mutation according to I-Mutant2.0 analysis, while MUpro scores indicate a destabilizing effect for mutation C28854T.

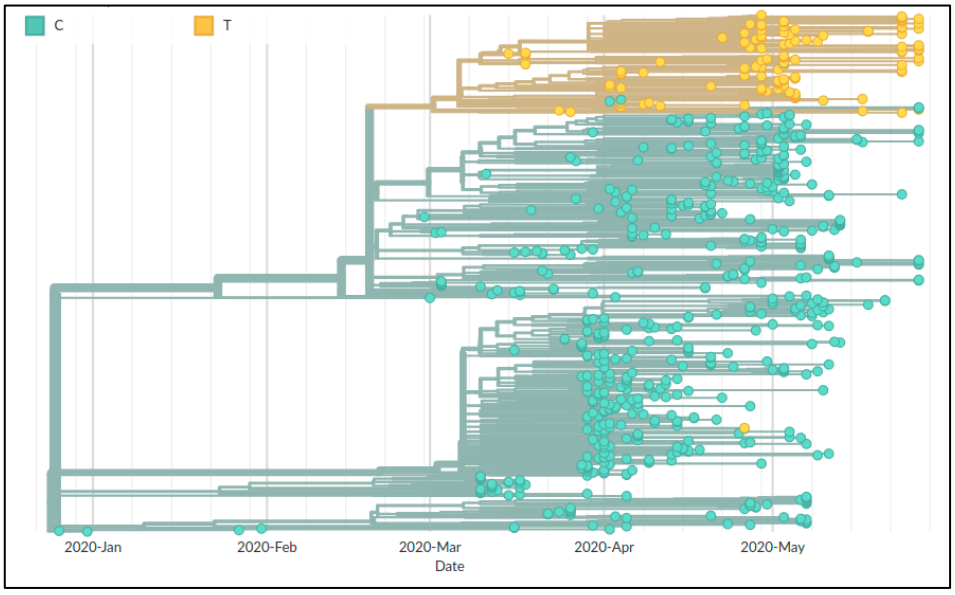


Fig. 9. Subclade I/GJ-20A associated with the co-occurring mutations C18877T, G25563T (ORF3a: Q57H), C26735T, C2836T, C22444T and C28854T (N: S194L).

A similar analysis of the isolates from Maharashtra compared to the isolates from Rest of India (RoI) showed noticeable differences at the genetic level. Maharashtra has the second highest death rate (4.19%) after Gujarat and is the state with the highest number of recorded infections. We observe that two co-occurring mutations in the 3′ UTR, A29827T and G29830T, are present with a prevalence of 55% and 73.75% respectively in the isolates from Maharashtra but their presence was observed in only 2 isolates from the rest of country. Similarly, a second set of co-occurring mutations, C313T and C5700A (ORF1ab: A1812D), were observed in Maharashtra with a high frequency of 25% and 26.25% respectively but with < 3% in isolates from rest of

India. These two distinct sets of mutations form two distinct clusters of samples from Maharashtra in the PCA plot. Among these mutations, only C5700A (A1812D) is a non-synonymous mutation in the Nsp3 protein of ORF1ab and its functional significance is discussed above in the section on India-specific mutations.

Table 5. Mutations observed in SARS-CoV-2 isolates from Maharashtra (MH) (80 samples) with high frequency compared to the Rest of India, R0I (605 samples) are given.

Nucleotide Mutation	Protein	Amino Acid Mutation	Count in MH	Frequency in MH (%)	Count in RoI	Frequency in RoI (%)
C313T	ORF1ab (nsp1)	-	20	25.0	19	3.1
C5700A	ORF1ab (nsp3)	A1812D	21	26.2	12	2.0
A29827T	3′ UTR	-	44	55.0	0	0.0
G29830T	3′ UTR	-	59	73.8	2	0.3

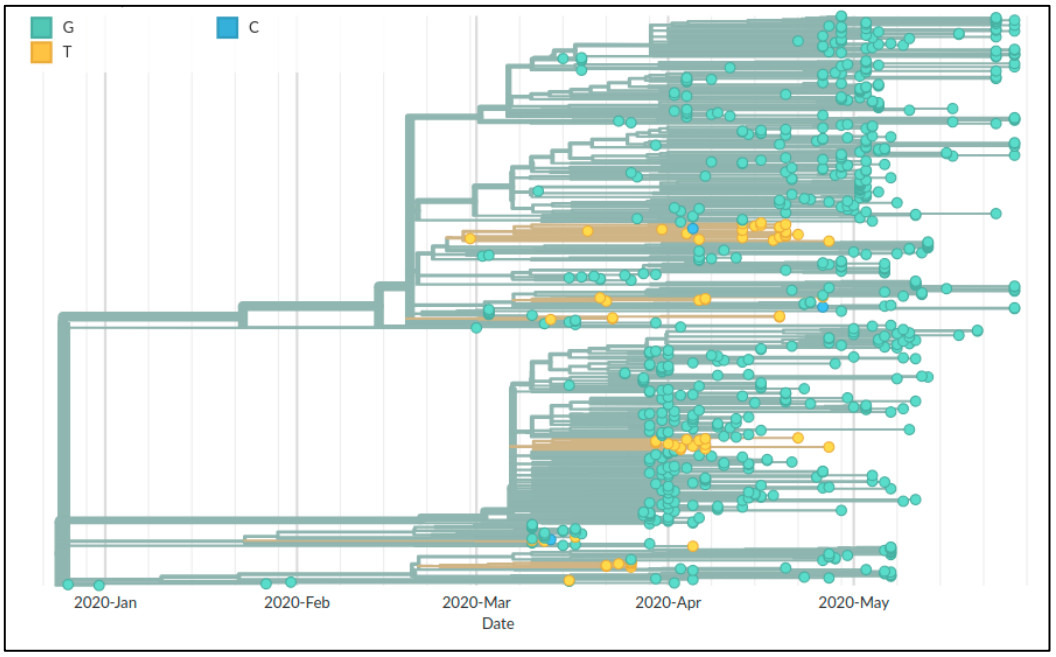


Fig. 10. Indian SAR-CoV-2 isolates carrying the mutation A29827T (co-occurs with G29830T) is depicted in yellow.

Conclusion

In this study, the demographic analysis of mutations occurring in 685 Indian SARS-CoV-2 isolates from 27th Jan 2020 – 27th May 2020 was conducted to understand the viral spread in the country during the early phase of the pandemic and assess the effectiveness of contact tracing, quarantine and lockdown in controlling the spread. The genetic analysis revealed that though lockdown helped in controlling the spread of the virus, region-specific set of shared mutations indicate at local transmission within the states. This is probably the reason for observed variation in the number of infected cases and death rates across states. Region specific sequencing efforts can help in understanding the mutational spectra in local hotspots like Gujarat and Maharashtra and aid in the implementation of state-wise lockdown policies. Due to limited sequencing data from such a vast country, numerous variants are likely to have been missed. A vast country like India needs to improve its sequencing efforts to understand

transmission dynamics of the virus, capture novel variants and assess their clinical impact to follow appropriate containment measures; this would prepare us for the third and fourth waves of COVID-19.

References

1. Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog* **13**, e1006698 (2017).

2. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* **395**, 565–574 (2020).

3. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**, 30494 (2017).

4. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

5. Capriotti, E., Fariselli, P. & Casadio, R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res* **33**, W306-310 (2005).

6. Cheng, J., Randall, A. & Baldi, P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* **62**, 1125–1132 (2006).

7. Banu, S. *et al.* A Distinct Phylogenetic Cluster of Indian Severe Acute Respiratory Syndrome Coronavirus 2 Isolates. *Open Forum Infect Dis* **7**, ofaa434 (2020).

8. Alam, A. S. M. R. U. *et al.* Dominant Clade-featured SARS-CoV-2 Co-occurring Mutations Reveals Plausible Epistasis: An in silico based Hypothetical Model. (2021) doi:10.1101/2021.02.21.21252137.

9. Chaudhari, A. *et al.* In-Silico analysis reveals lower transcription efficiency of C241T variant of SARS-CoV-2 with host replication factors MADP1 and hnRNP-1. *Inform Med Unlocked* **25**, 100670 (2021).

10. Zhang, L. *et al.* SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat Commun* **11**, 6013 (2020).

11. Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **592**, 116–121 (2021).

12. Daniloski, Z. *et al.* The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. *Elife* **10**, e65365 (2021).

13. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827.e19 (2020).

14. Bai, Y. *et al.* Comprehensive evolution and molecular characteristics of a large number of SARS-CoV-2 genomes reveal its epidemic trends. *International Journal of Infectious Diseases* **100**, 164–173 (2020).

15. Ayub, M. I. Reporting Two SARS-CoV-2 Strains Based on A Unique Trinucleotide-Bloc Mutation and Their Potential Pathogenic Difference. (2020) doi:10.20944/preprints202004.0337.v1.

16. Tylor, S. *et al.* The SR-rich motif in SARS-CoV nucleocapsid protein is important for virus replication. *Can J Microbiol* **55**, 254–260 (2009).

17. Maitra, A. *et al.* Mutations in SARS-CoV-2 viral RNA identified in Eastern India: Possible implications for the ongoing outbreak in India and impact on viral structure and host susceptibility. *J Biosci* **45**, 76 (2020).

18. Paul, D. *et al.* Phylogenomic analysis of SARS-CoV-2 genomes from western India reveals unique linked mutation. (2020) doi:10.1101/2020.07.30.228460.

19. Lei, J., Kusov, Y. & Hilgenfeld, R. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. *Antiviral Research* **149**, 58–74 (2018).

20. Gupta, A. *et al.* A comprehensive profile of genomic variations in the SARS-CoV-2 isolates from the state of Telangana, India. *J Gen Virol* **102**, (2021).

21. Gao, Y. *et al.* Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* **368**, 779–782 (2020).

22. Hsin, W.-C. *et al.* Nucleocapsid protein-dependent assembly of the RNA packaging signal of Middle East respiratory syndrome coronavirus. *J Biomed Sci* **25**, 47 (2018).

23. Masters, P. S. Coronavirus genomic RNA packaging. *Virology* **537**, 198–207 (2019).

24. Chang, C.-K. *et al.* Multiple nucleic acid binding sites and intrinsic disorder of severe acute respiratory syndrome coronavirus nucleocapsid protein: implications for ribonucleocapsid protein packaging. *J Virol* **83**, 2255–2264 (2009).

25. Chang, C., Hou, M.-H., Chang, C.-F., Hsiao, C.-D. & Huang, T. The SARS coronavirus nucleocapsid protein--forms and functions. *Antiviral Res* **103**, 39–50 (2014).

26. Oulas, A. *et al.* Generalized linear models provide a measure of virulence for specific mutations in SARS-CoV-2 strains. *PLOS ONE* **16**, e0238665 (2021).

27. Joshi, M. *et al.* Genomic Variations in SARS-CoV-2 Genomes From Gujarat: Underlying Role of Variants in Disease Epidemiology. *Frontiers in Genetics* **12**, 292 (2021).

28. Wang, R. *et al.* Characterizing SARS-CoV-2 mutations in the United States. *Res Sq* rs.3.rs-49671 (2020) doi:10.21203/rs.3.rs-49671/v1.

29. Yu, I.-M., Oldham, M. L., Zhang, J. & Chen, J. Crystal structure of the severe acute respiratory syndrome (SARS) coronavirus nucleocapsid protein dimerization domain reveals evolutionary linkage between corona- and arteriviridae. *J Biol Chem* **281**, 17134–17139 (2006).

30. Zhao, P. *et al.* Immune responses against SARS-coronavirus nucleocapsid protein induced by DNA vaccine. *Virology* **331**, 128–135 (2005).

31. Wu, S. *et al.* Effects of SARS-CoV-2 mutations on protein structures and intraviral protein-protein interactions. *J Med Virol* **93**, 2132–2140 (2021).