Comparative Analysis of SARS-CoV-2 Variants Across Three Waves in India

Kushagra Agarwal and Nita Parekh(✉)

Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology, Hyderabad, Telangana 500032, India
nita@iiit.ac.in

Abstract. In this study we carried out a comprehensive analysis of SARS-CoV-2 mutations and their spread in India over the past two years of the pandemic (27th Jan' 2020 – 8th Mar' 2022). The analysis covers four important timelines, viz., the early phase, followed by the first, second and third waves of the pandemic in the country. Phylogenetic analysis of the isolates indicated multiple independent entries of coronavirus in the country, while principal component analysis identified few state-specific clusters. Genetic analysis of isolates during the first year revealed that though lockdown helped in controlling the spread of the virus, region-specific set of shared mutations were developed during the early phase due to local community transmissions. We thus report the evolution of state-specific subclades, namely, I/GJ-20A (Gujarat), I/MH-2 (Maharashtra), I/Tel-A-20B, I/Tel-B-20B (Telangana), and I/AP-20A (Andhra Pradesh) that explain the demographic variation in the impact of COVID-19 across states. In the second year of the pandemic, India faced an aggressive second wave while the third wave was quite mild in terms of severity. Here we also discuss the prevalence and impact of different lineages and Variants of Concerns/Interests, viz., Delta, Kappa, Omicron, etc. observed during this period. From the genetic analysis of mutation spectra of Indian isolates, the insights gained in its transmission, geographic distribution, containment, and impact are discussed.

Keywords: SARS-CoV-2 · Mutations · India

1 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. Large variation in the rate of infectivity and fatality due to COVID-19 is observed across different countries and a similar trend is 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 Indian isolates obtained from GISAID (1) during the period 27th Jan 2020 – 8th March 2022 is carried out in four phases. 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. A detailed state-wise

R. Somashekhar et al. (Eds.): ICBDS 2022, AHSR 58, pp. 104–118, 2023.
https://doi.org/10.2991/978-94-6463-164-7_9

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 before and after the first wave. The severity and number of hospitalizations and death was high after the second wave in Mar-Apr' 2021 and it would be interesting to assess the variants circulating during that period in India. The third wave also saw huge number of infections in the country; however, the impact was mostly mild and asymptomatic. In this study we attempt to examine the emergence of region-specific mutations and their spread across the country over a period of two years across four different phases with three waves of the pandemic in India.

2 Materials and Methods

In this study we provide a comparative analysis of Indian SARS-CoV-2 isolates for four time-points, namely, early phase of COVID-19 from 27th Jan–27th May'2020 (685 samples - Dataset-I), after 1st wave till 11th Jan'21 (4708 samples - Dataset-II), after 2nd wave till 1st Oct'21 (45171 samples - Dataset-III), and after 3rd wave till 8th Mar'22 (101527 samples - Dataset-IV) in India. The data for Datasets I and II was obtained from GISAID, and sequence Wuhan/Hu-1/2019 (EPI_ISL_402125) was used as reference. Any incomplete metadata or low-quality genomes have been discarded. To assess the impact of lockdown during the early phase and after the first wave in India, comparative analysis is carried out between Datasets I and II. The evolution and spread of various lineages are discussed specifically after the second and third waves in the country. For this analysis, the mutation and lineage frequencies in Datasets III and IV were obtained from COVID-19 CG (2).

We conducted phylogenetic tree analysis of the samples in Datasets I and II using the bioinformatic engine Augur and visualization tool Auspice from Nextstrain (3). Nextstrain pipeline was executed using the Snakemake workflow which includes a multiple sequence alignment using MAFFT (4), followed by creating a maximum likelihood tree using IQ-TREE (5) with 1000 bootstraps. The tree was then time-resolved using TimeTree after refinement. The final step included the inference of clades, mutations (both at nucleotide and amino acid levels), and ancestral traits to obtain the Newick format tree visualized using Auspice. The time-resolved phylogenetic tree thus obtained for the first year of COVID-19, corresponding to Dataset-II is shown in Fig. 1.

State-specific clusters based on shared mutations (in Datasets I and II) are identified by performing principal component analysis (PCA) on the mutational profile of Indian isolates. To assess the impact of India-specific non-synonymous mutations on protein function, SIFT (6) and PROVEAN (7) have been used. SIFT (Sorting Intolerant From Tolerant) uses sequence homology and properties of amino acids to predict whether an amino acid substitution would affect the protein function. PROVEAN (Protein Variation Effect Analyzer) predicts whether an amino acid substitution/indel has an impact on the biological function of a protein based on its homologs searched against the NCBI 'nr' database using BLAST and clustered using CD-HIT. Mutations are defined as deleterious if the PROVEAN score < -2.5 and SIFT score < 0.05.

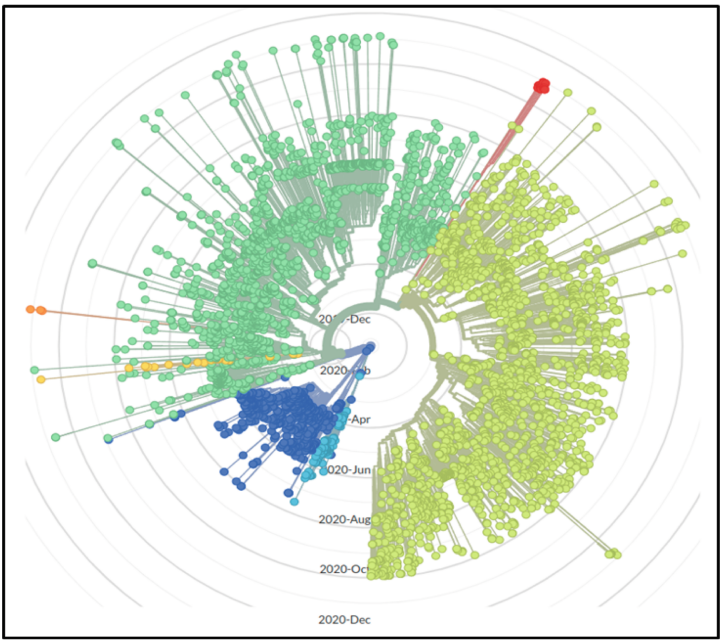

Fig. 1. Time-resolved radial phylogenetic tree for the period (Jan' 2020 – Jan – 2021) with samples colored according to Nextstrain clades: Dark blue: 19A, Sky blue: 19B, Dark green: 20A, Light green: 20B, Yellow: 20C, Orange: 20E (EU1), Red: 20I/501Y.V1.

3 Results

3.1 Clade Analysis

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. 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. By the end of the first year (Dataset-II), 2 new clades (20E, 20I/501Y.V1) were observed with 20B being the most prominent clade followed by 20A and 19A covering $>$ 97% of samples in the country. Of the 4708 samples in Dataset-II, 4678 had state information available (covering 17 states and Union Territories Delhi and Ladakh). Before lockdown Clade 20A was pre-dominant in Gujarat

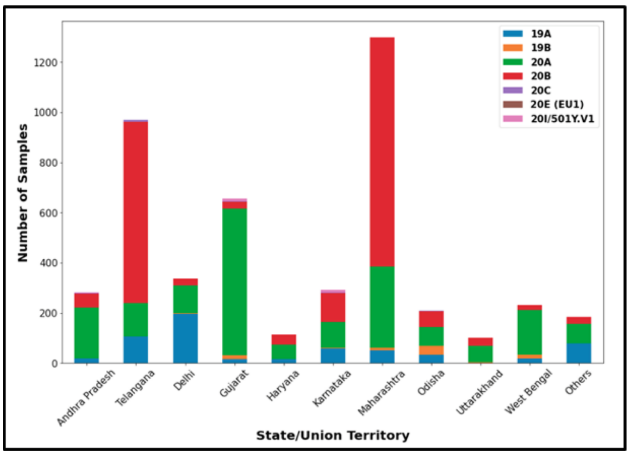

Fig. 2. State-wise distribution of clades across the country shown for the period 26th Dec' 2019 – 21st Jan' 21 (Dataset-II).

(followed by West Bengal, Odissa, Madhya Pradesh), and by Jan'21 (Dataset-II) it was well-represented in majority of states, with higher representation in Gujarat (583/655), Maharashtra (323/1299), Andhra Pradesh (204/281) and West Bengal (178/232). Similarly, Clade 20B with minimal representation in Maharashtra, Tamil Nadu, and Telangana in Dataset-I, showed a tremendous increase in Maharashtra (915/1299) and Telangana (724/970), followed by Karnataka (116/292). The earliest clade 19A (and its India-specific subclade I/A3i (8), which was earlier seen in four states, Telangana, Delhi, Maharashtra, and Tamil Nadu, however, showed a small increase only in Delhi (196/338), suggesting the effect of contact tracing and quarantine during lockdown. Thus, from the distribution of clades across different states in Fig. 2, we can speculate localized transmissions during the early phase and the spread of some of these variants to other states after the lockdown was lifted.

3.2 Most Frequent Mutations

A total of 1279 variations were identified in Dataset-I (685 sequences) and 7126 variations in Dataset-II (4708 sequences) when compared with Wuhan-1 isolate as reference. Top 20 most frequent mutations in Dataset-II along with their amino acid change (if any) and frequencies in Datasets I and II are shown in Fig. 3. The clade 20A mutation A23403G (D614G) in the Receptor Binding Domain (RBD) of Spike protein, first discovered in late January in the West European region (9) was observed in over 50% of samples in Dataset-I (373/685) with a high prevalence in Gujarat (180/201). In accordance with Alai *et al.* (10), it continued to be the most common mutation, occurring in 4002/4708 samples in Dataset-II, with predominance in Maharashtra (1238/1299), Telangana (861/970), and Gujarat (619/655) states. The D614G mutation, present on S2 domain of spike protein, is important for cleavage of S1 by TMPRSS2 enzyme to allow viral spike fusion with the host cell membrane (11). According to previous studies (9, 12), D614G strains are more infectious than the original Wuhan-1 strain. This increased

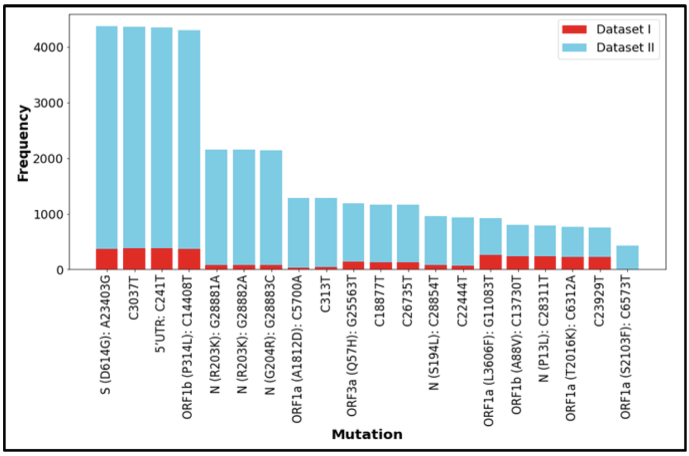

Fig. 3. Top 20 most frequent mutations during the early phase (red) and after first wave (blue) in Indian samples shown.

infectivity is linked to reduced S1 shedding and more S protein incorporation into the pseudovirion. It also increases viral load in the upper respiratory tract in COVID-19 patients. Since D614G mutation falls outside the region responsible for formation of neutralizing antibodies, it has not been a concern for vaccine effectiveness (12). Three other mutations corresponding to clade 20A, C14408T (ORF1b: P314L, RdRP: P323L), C241T (5' UTR of ORF 1ab), and C3037T are observed to co-occur with D614G. According to the predictions of SIFT and PROVEAN, mutations D614G and P314L may not severely affect the protein function (Table 1).

The clade 20B tri-bloc mutation, G28881A (R203K, N protein), G28882A (R203K, N protein), and G28883C (G204R, N protein), observed in 75 samples in the early phase became the second most prevalent mutation set after first wave. Observed in over 2000 samples (~40%), primarily from Maharashtra (914/1299) and Telangana (720/970), clade 20B was the second most dominant clade after clade 20A. The corresponding amino acid mutations R203K and G204R result in an insertion of lysine residue in the SR-rich region of nucleocapsid protein, which is involved in viral capsid formulation (13). Both R203K and G204R are predicted to be neutral by PROVEAN, while SIFT indicates that R203K will be tolerated while G204R may affect protein function.

In Dataset-I, about one-third of the isolates shared the mutations, viz., C6312A (ORF1a: T2016K), C13730T (ORF1b: A88V), C23929T, C28311T (N: P13L) and G11083T (ORF1a: L3606F), branching out of clade 19A. According to SIFT prediction the non-synonymous mutations L3606F, A88V and P13L may affect the protein function, while T2016K may be tolerated. In contrast, PROVEAN predicted all the non-synonymous mutations to be neutral, however, the predicted score for A88V being close to cutoff suggests that it may have some impact on protein function. Since globally the frequency of these 5 mutations is low (~3.5%), Banu *et al.* (8) defined these as India-specific subclade I/A3i. This hints at early community transmission in India due to some super spreader event during March-April as such a high number of isolates with common

Table 1. Frequency of significant mutations circulating in the first year of pandemic. Their functional relevance is assessed using SIFT and PROVEAN. (PROVEAN analysis, N: Neutral, D: Deleterious; SIFT analysis, A: Affects function, T: Tolerated. Del*: Deletion not supported, Failed**: PSI-BLAST could not retrieve enough sequences.)

Mutation	Amino Acid Mutation	Freq in Dataset-I	Freq in Dataset-II	PROVEAN Analysis	SIFT Analysis
A23403G	S: D614G	373 (54.4%)	4002 (85.0%)	0.60 (N)	0.62 (T)
C14408T	ORF1b: P314L	367 (53.6%)	3941 (83.7%)	-0.45 (N)	0.23 (T)
C3037T	-	374 (54.6%)	3984 (84.6%)	-	-
C241T	-	378 (55.2%)	3973 (84.4%)	-	-
G28881A	N: R203K	75 (11.0%)	2077 (44.1%)	-1.60 (N)	0.11 (T)
G28883C	N: G204R	75 (11.0%)	2066 (43.9%)	-1.66 (N)	0.02 (A)
G28882A	N: R203K	75 (11.0%)	2073 (44.0%)	-1.60 (N)	0.11 (T)
G25563T	ORF3a: Q57H	134 (19.6%)	1057 (22.4%)	-3.29 (D)	0 (A)
C26735T	-	123 (18.0%)	1039 (22.1%)	-	-
C18877T	-	126 (18.4%)	1043 (22.1%)	-	-
C28854T	N: S194L	73 (10.7%)	887 (18.8%)		
C22444T	-	71 (10.4%)	859 (18.2%)	-	-
C2836T	-	51 (7.4%)	314 (6.7%)	-	-
C313T	-	39 (5.7%)	1242 (26.4%)	-	-
C5700A	ORF1a: A1812D	33 (4.8%)	1253 (26.6%)	-0.75 (N)	0.41 (T)
G11083T	ORF1a: L3606F	258 (37.7%)	662 (14.1%)	-1.43 (N)	0.01 (A)
C28311T	N: P13L	236 (34.4%)	557 (11.8%)	-1.23 (N)	0 (A)
C13730T	ORF1b: A88V	239 (34.9%)	567 (12.0%)	-2.35 (N)	0 (A)

(*continued*)

Table 1. (*continued*)

Mutation	Amino Acid Mutation	Freq in Dataset-I	Freq in Dataset-II	PROVEAN Analysis	SIFT Analysis
C23929T	-	224 (32.7%)	525 (11.2%)	-	-
C6312A	ORF1a: T2016K	228 (33.3%)	532 (11.3%)	-0.17 (N)	0.59 (T)

set of shared mutations is unlikely to be due to multiple independent entries, especially when international flights were suspended. Coincidence with the Tablighi congregation held in mid-March, earliest reported case of the subclade from Saudi Arabia, and several Indonesian citizens sampled in Delhi found to be carrying these mutations, indicate Tablighi congregation as the probable cause of its spread in the country. Followed by country-wide lockdown, contact-tracing and quarantine, their numbers reduced to ~ 10.5% after first wave. Thus, based on genetic analysis one can identify the chain of transmission of a variant strain and the success of measures for its containment.

3.3 PCA Analysis of Indian SARS-CoV-2 Isolates

Since travel between states was restricted due to lockdown, increase in the number of samples with shared mutations was likely because of local community transmission. To gain insight into state-specific clustering, during the initial period and propagation of shared mutations over the first year of pandemic, principal component analysis (PCA) was performed on the mutational profile comprising 7126 mutations in 4708 isolates (Dataset-II) as can be seen in Fig. 4. Mutations sorted based on their loading scores to assess their impact on PC1 identified 8 out of top 12 most common mutations in Indian isolates after first wave: G28881A, G28882A, G28883C, C313T, C5700A, G25563T, C26735T and C18877T. The remaining 4 mutations (A23403G, C14408T, C241T, C3037T) occur in more than 4000 samples and hence divide the dataset with lower entropy.

A separate 'pink' cluster is observed corresponding to Gujarat-specific subclade I/GJ-20A. In Maharashtra, the cluster corresponding to mutations C313T and C5700A is enriched (Fig. 4 zoomed), while that of A29837T and G29830T mutations reduced from ~ 55% to ~ 3.5%. Other significant clusters correspond to the states Telangana, West Bengal, Andhra Pradesh, and Delhi. Two Telangana state-specific subclades, I/Tel-A-20B and I/Tel-B-20B are observed.

3.4 State Wise Analysis

In the early phase of pandemic, many countries, e.g., Italy, UK, Spain, etc. reported large number of mortality cases. Severity of the circulating strains and large aging population were proposed to be the probable cause. India being a very vast and young country, it would be interesting to study if a similar pattern is observed across its different states. Our

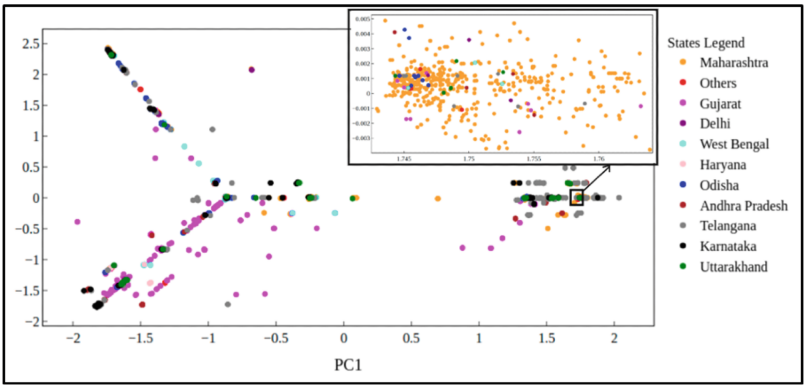

Fig. 4. PCA plot for 4708 samples (Dataset-II) colored state wise.

analysis revealed that few states exhibited high severity of COVID-19 cases. For example, Gujarat (5.12%) and Maharashtra (4.19%) reported higher death rates compared to the country average (2.67%), as on 11th July 2020 (Dataset-I). Number of deaths in Gujarat was 2008 out of 39194 reported 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 (3.29%) and Tamil Nadu (1.39%) had lower fatality rates. By the end of the first year of pandemic, mortality rates in all states fell significantly with the country average dropping to 1.44%. With the highest fatality rate of 2.54%, Maharashtra also recorded maximum number of cases (1971552), while Gujarat despite having much fewer cases (252559), had second highest fatality rate of 1.72% after first wave. Telangana reported the lowest mortality rate of 0.54%. To understand this significantly large difference in mortality rates at the genetic level, a detailed analysis of the mutational profile of sequences from these states was carried out.

Gujarat Analysis

To understand the dense clustering of Gujarat samples in PCA plots and significantly large difference in the percentage of deaths compared to the country average, we compared the mutational profile of Gujarat isolates with those from the Rest of India (RoI). A common set of shared mutations is identified in Gujarat isolates, viz., G25563T (ORF3a: Q57H), C26735T, C18877T, C28854T (N: S194L), C22444T, and C2836T, apart from clade 20A defining mutations (C241T, C3037T, C14408T, A23404G) in Dataset-I. Of these, mutations G25563T, C26735T and C18877T co-occur forming a distinct branch within clade 20A in the phylogenetic tree (Fig. 5). We refer to it as Gujarat-specific subclade I/GJ-20A because of $> 50\%$ presence in Gujarat isolates compared to RoI ($< 5\%$). Mutations C28854T, C22444T and C2836T form sub-branches within this branch. From Table 2, during the first wave I/GJ-20A defining mutations continued to dominate in Gujarat (~56%) and spread to other states (~16%) after lockdown was lifted in agreement with an earlier study (14). Noticeably, C2836T mutation was observed in ~ 42% samples from Gujarat but its representation in RoI was $< 1\%$. Most mutations that were under-represented in Gujarat isolates in Dataset-I continued to be so in Dataset-II,

Table 2. Mutations in Gujarat (GJ) samples with a higher frequency than in Rest of India (RoI) during early phase (Dataset-I) and after first wave (Dataset-II) are given. No. of Gujarat samples: 201 in Dataset-I, 655 in Dataset-II. Rest of India samples: 484 in Dataset-I, 4053 in Dataset-II.

Nucleotide Mutation	Amino Acid Mutation	Dataset-I		Dataset-II	
		Count in GJ	Count in RoI	Count in GJ	Count in RoI
C3037T	-	182 (90.6%)	192 (39.7%)	621 (94.8%)	3363 (83.0%)
A23403G	S: D614G	180 (89.6%)	193 (39.9%)	619 (94.5%)	3383 (83.5%)
C241T	-	183 (91.0%)	195 (40.3%)	617 (94.2%)	3356 (82.8%)
C14408T	ORF1b: P314L	178 (88.6%)	189 (39.0%)	581 (88.7%)	3360 (82.9%)
G25563T	ORF3a: Q57H	108 (53.7%)	26 (5.4%)	369 (56.3%)	688 (17.0%)
C26735T	-	101 (50.2%)	22 (4.5%)	367 (56.0%)	672 (16.6%)
C18877T	-	105 (52.2%)	21 (4.3%)	366 (55.9%)	677 (16.7%)
C22444T	-	64 (31.8%)	7 (1.4%)	293 (44.7%)	566 (14.0%)
C28854T	N: S194L	66 (32.8%)	7 (1.4%)	284 (43.4%)	603 (14.9%)
C2836T	-	51 (25.4%)	0 (0.0%)	274 (41.8%)	40 (1.0%)

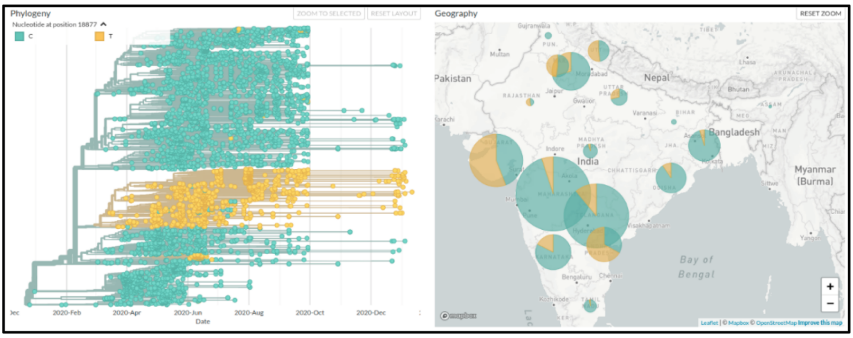

Fig. 5. Subclade I/GJ-20A marked in yellow on the phylogenetic tree. (Generated using Auspice v2.36.0 (https://nextstrain.org/)).

namely, subclade I/A3i mutations, tri-bloc mutation, and C313T, C5700A, and G11083T mutations. Both SIFT and PROVEAN predictions also indicate Q57H mutation has significant impact on ORF3a protein. The other non-synonymous mutation C28854T (N: S194L) results in an altered structure of the nucleocapsid protein. Based on PROVEAN score the mutation is likely to be deleterious, while SIFT score being equal to the threshold score predicts probable impact on the protein function. Notably, both these mutations also have significant impact on key hot spots in respective proteins that can be used as therapeutic targets owing to their major roles in protein-protein interactions.

Maharashtra Analysis

A similar analysis of Maharashtra isolates with that of RoI showed noticeable differences at the genetic level. Maharashtra recorded second highest death rate (4.19%) and highest number of infections. Before the first wave, two co-occurring mutations A29837T and G29830T (in 3'UTR of the viral RNA), with an incidence of 55% and 74% respectively in Maharashtra isolates, were found in only 2 samples outside the state. Interestingly, the occurrence of these mutations reduced to 3.5% and 6.9% respectively after first wave, hinting at the success of lockdown in containing spread of the virus. Another set of co-occurring mutations, C313T and C5700A (ORF1a: A1812D), were identified with frequencies 25% and 26.3% respectively in Maharashtra, but with $< 3\%$ in isolates from RoI in Dataset-I. By Jan '21 (Dataset-II), this set of co-occurring mutations was over-represented (63.12%) in Maharashtra but under-represented in RoI (10.94%). The increased frequency in India (26.6%) indicates its spread to different states after the lockdown was lifted. They form a distinct subclade I/MH-2 in the phylogenetic tree. This mutation set has also been reported by a study on sequences from western India (15). SIFT and PROVEAN analysis however indicates that A1812D mutation may not severely affect the biological function of ORF1a protein.

Telangana Analysis

Two independent subclades emerged from clade 20B in Telangana isolates after lockdown was lifted, viz., I/TelA-20B (G4354A, C6573T (ORF1a: S2103F), C25528T (ORF3a: L46F)), and I/Tel-B-20B (C9693T (ORF1a: A3143V), C16626T, A4372G, G29474T (N: D401Y)). Subclade I/Tel-A-20B defining mutations were found in 393 Telangana samples (40.52%) but in only 17 (0.5%) samples from the rest of country. Similarly, I/Tel-B-20B had over 25% prevalence in the state but in $< 1\%$ samples from rest of the country. Apart from these, two adjacent co-occurring mutations A21550C and A21551T resulting in amino acid change N2695L in ORF1b were also found to be Telangana-specific (but not subset any clade) with more than 11% representation in the state but ~ 0.2% in RoI, in accordance with an earlier work (16). This clearly indicates local community transmission within the state. SIFT analysis indicate that the non-synonymous mutations, S2103F, L46F, D401Y, N2695L may affect the protein function, while PROVEAN score indicates the mutation L46F to be deleterious. According to SIFT mutation D401Y of subclade I/Tel-B-20B, and mutation N2695L may affect the protein function.

Andhra Pradesh Analysis

Isolates from Andhra Pradesh (AP) revealed state-specific subclade, I/AP-20A, defined by the mutations C3267T (ORF1a: T1001I), C21034T (ORF1b: L2523F), G26173T (ORF3a: E261del), G28183T (ORF8: S97I) and T28277C (N: S2P) in Dataset-II. These mutations co-occur and branch out of clade 20A in the phylogenetic tree. It is observed in ~ 53% of AP isolates, with a small presence in the neighbouring states of Telangana (4.6%) and Karnataka (6.9%). The mutation G28183T is predicted to be deleterious according to PROVEAN, while mutation T28277C is predicted to affect protein function according to SIFT.

Apart from these four states, our analysis of Dataset-II revealed noticeable region-specific variations. For example, analysis of 292 sequences from Karnataka revealed 3 mutations, C1218T (ORF1a: S318L), C27110T and T27384C with $> 20\%$ prevalence

in the state and < 1% outside it. Similarly, differences were observed in sequences from West Bengal (G15451A: ORF1ab G662S) and Odisha (A25381C) with ~ 9% prevalence in the respective states and nearly negligible presence in RoI.

4 Discussion

4.1 Prevalence of Mutations Acquired in the Early Phase of Pandemic

It is observed that over the past two years SARS-CoV-2 has evolved in humans, exploring the sequence space, and resulting in the selection of variants with improved replication efficiency and transmissibility. Here we discuss the India-specific variants observed in the first year of pandemic (Datasets I and II) that are still circulating in the country (Dataset-III and IV). Clade 20A defining mutations continue to dominate with more than 97% prevalence in the country till date, while frequency of novel India-specific subclades identified after first wave are significantly reduced, e.g., frequency of I/GJ-20A and I/MH-2 subclades respectively reduced from 56% and 26.6% after first wave (Dataset-II) to 5% and 2% after third wave (Dataset-IV). Similar analysis of Clade 20B defining mutations indicates that its frequency decreased from ~ 44% after first wave to ~ 15% after second wave and then increased again to 23% after third wave. No new incidence of India-specific subclade I/A3i is observed after the early phase (Dataset-I) and its cumulative presence is $\leq$ 1% after second and third waves. However, one of its characteristic mutations, C28311T (P13L), is observed in over 16% of samples after the third wave. The mutation has been reintroduced in the omicron variant, leading to its increased incidence after third wave. A similar decreasing trend is observed for state-specific subclades I/Tel-A-20B, I/Tel-B-20B and I/AP-20A after second and third waves. Thus, the state-specific mutational analysis indicates that during lockdown some variants were confined by state boundaries, and these grew in number after the lockdown was lifted, due to local transmission. Limited travel between states resulted in their numbers restricted majorly to the respective states after first wave. With the advancement of pandemic and introduction of more transmissible variants on opening of international borders, loss of region-specific mutations is observed. The introduction and spread of new lineages of SARS-CoV-2 after the first wave of pandemic in India is discussed.

4.2 Pangolin Lineage Analysis

GISAID categorized the virus into clades during the first year based on unique mutations identified at different positions in the genome. Since SARS-CoV-2 has been evolving very fast, a dynamic nomenclature based on phylogenetic framework is employed to designate lineages with an active spread, known as Pangolin (17). Lineages that posed an increased risk to global public health have been termed 'Variants of Interest' (VOI), or 'Variants of Concern' (VOC) by World Health Organization (WHO), based on the acquired mutations in spike protein receptor binding domain that resulted in a substantial increase in its binding affinity with human ACE2 protein and linked to rapid spread in the population. Early phase of pandemic in India (till May '2020) saw only two lineages B.6 (Clade 19A) and B.1 (Clade 20A), with 30% and 25% incidence, respectively. After

the first wave, lineage B.1.1.306 (Clade 20B) became dominant with ~ 27% prevalence followed by B.1 (14%) and B.6 (10%).

When the world was experiencing huge numbers of COVID-19 cases during summer of 2020, India luckily missed the first wave. Timely screening of international travelers, quarantine and contact tracing, followed by country-wide lockdown on 24th Mar'2020 was helpful. In India first wave peaked around Sept' 2020 and the second wave around Mar-Apr' 2021. Since the second wave was very severe with very high infectivity and fatality, analysis of new emerging variants between the first and second waves is important in the Indian context (18). We observe that Delta variant (B.1.617.2, clade 21A) emerged by Oct' 2020 in India followed by Kappa (B.1.617.1, clade 21B) and Alpha (B.1.1.7, clade 20I) variants, while Alpha variant was still dominant in both UK and USA. Delta variant was first detected in India in late 2020, swept rapidly through the country and reached UK and then USA where it became the dominant variant, accounting for ~ 99% of COVID-19 cases and large number of hospitalizations. It is thought to be twice as contagious with viral loads ~ 1000 times more compared to the previous variants, resulting in enhanced severity, and was flagged as Variant of Concern (VOC) by WHO. Delta variant is characterized by the spike mutations: T19R, del157/158 (in N-terminal domain of RBD), L452R, T478K (in RBD), D614G, P681R, D950N (in S2 region). Further, it has been shown by Planas *et al*. (19), that the vaccine effectiveness is notably lower for Delta variant compared to Alpha variant. This may be because of mutations T19R and del157/158 in the N-terminal domain which provides a 'supersite' for antibodies to latch to the virus, making monoclonal antibodies less effective in treating COVID and increases the Delta variant's ability to escape vaccine-generated antibodies. Delta variant's increased transmissibility and severity lead to the sudden spike in cases and a near breakdown of healthcare system during the second wave in India and other countries.

Another prevalent variant during the second wave in India was Kappa variant (9%), also called 'double mutant' by media and later designated as lineage B.1.617.1 and variant of interest (VOI) by WHO. First observed in Maharashtra in Mar' 2021, it is characterized by the spike mutations: E154K, E484Q, L452R, D614G, P681R and Q1071H. By April 2021, Kappa variant accounted for ~ 35% of all sequenced cases in India and coincided with the rise in daily COVID-19 cases in India. It is an evolutionary ancestor of Delta variant with shared mutations E484Q, L452R, D614G, P681R that provide it increased transmissibility. However, the variant's impact on severity has not been proven. The second wave also saw the introduction of a deescalated VOI, the "Alpha" variant (7%), lineage B.1.1.7 (clade 20B/501Y.V1). It was first observed in UK in Dec' 2020 and has been well characterized for both increased transmissibility (20) and increased severity (21). During the second wave, it was observed that northern part of India had a higher dominance of Alpha, while in the southern and central parts of India, Delta and Kappa variants were rampant (22). Alpha variant is characterized by the spike mutations: A570D, D614G, D1118H, H69del, N501Y, P681H, S982A, T716I, V70del, and Y145del. Apart from these highly infectious variants circulating, some of the factors that fueled the second wave in the country are complacency by the government and the public after first wave, super spreader events such as Kumbh mela, assembly elections,

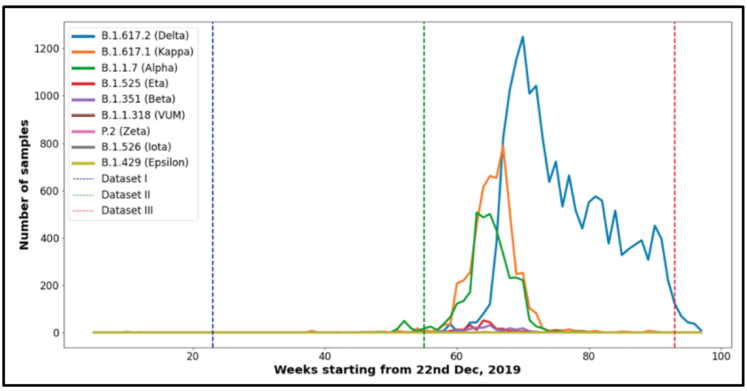

Fig. 6. Week-wise evolution of all the VOCs, VUMs, and former VOIs observed in Indian isolates. Week 0 corresponds to 22nd - 28th Dec 2019 and Week 97 to 25th - 31st October 2021. Dashed vertical lines correspond to the time periods for which data was collected. The data for analysis is obtained from CoV-GLUE (http://cov-glue.cvr.gla.ac.uk/)

weddings and celebrations, and insufficient vaccination (~0.3% of the population with single dose).

With the end of second wave in India, number of deaths reduced drastically. This could also be accredited to large-scale vaccination campaign that took place post-second wave and prepared India for a third wave. The third wave in India (Dec 2021 - February 2022) was characterized by mild infections, very few hospitalizations, and mostly home treatments. This indicates that the variant in circulation, while having increased transmissibility, had reduced severity compared to Delta variant. A comparison of isolates in Datasets III and IV can help in identifying differences between the circulating variants during second and third waves. In Dataset-IV, while Delta variant (B.1.617.2) continued to be the most prevalent lineage (24%) a new variant, BA.2, which is a subset of Omicron variant (VOC) grew in the country (11%). Various studies have discussed the increased transmissibility (23) and reduced severity (24,25) of Omicron variant. An interesting observation by Pulliam *et al.* (26), was that Omicron variant can evade host immunity induced due to prior infection. This was not the case with earlier variants, namely, Alpha, Beta, and Delta, wherein, previous infection robustly prevented reinfection (90%), while that was not true with Omicron (60%) (27). Omicron variant is shown to have a completely new serotype because of which a person infected with Omicron does not have protection against infections due to other variants (28). It contains more than 30 mutations in spike protein compared to Wuhan-1 as reference (26). Along with Omicron variant BA.2, AY.127 (B.1.617.2.127), Kappa variant (B.1.617.1) and AY.112 (B.1.617.2.112), each with 5% prevalence Alpha variant with 4% and BA.1 (subset of Omicron, VOC) with 3% are observed after the third wave. Figure 6 shows the evolution of all the important VOC/VOI/VUM present in the Indian samples across Datasets I, II and III, wherein, Delta, Kappa and Alpha variants can be seen dominating in the Indian samples after the first wave.

Acknowledgments. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Euro Surveill 22, 30494 (2017).
2. Chen, A. T., Altschuler, K., Zhan, S. H., Chan, Y. A. & Deverman, B. E. COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest. eLife 10, e63409.
3. Hadfield, J. et al Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34, 4121–4123 (2018).
4. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30, 3059–3066 (2002).
5. Minh, B. Q. et al IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution 37, 1530–1534 (2020).
6. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. Nat Protoc 11, 1–9 (2016).
7. Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 31, 2745–2747 (2015).
8. Banu, S. et al A Distinct Phylogenetic Cluster of Indian Severe Acute Respiratory Syndrome Coronavirus 2 Isolates. Open Forum Infect Dis 7, ofaa434 (2020).
9. 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).
10. Alai, S., Gujar, N., Joshi, M., Gautam, M. & Gairola, S. Pan-India novel coronavirus SARS-CoV-2 genomics and global diversity analysis in spike protein. Heliyon 7, e06564 (2021).
11. Zhang, L. et al The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. Preprint at bioRxiv, 2020.06.12.148726 (2020).
12. Li, Q. et al The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. Cell 182, 1284-1294.e9 (2020).
13. 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).
14. Joshi, M. et al Genomic Variations in SARS-CoV-2 Genomes From Gujarat: Underlying Role of Variants in Disease Epidemiology. Front Genet 12, 586569 (2021).
15. Paul, D. et al Phylogenomic analysis of SARS-CoV-2 genomes from western India reveals unique linked mutations. Preprint at bioRxiv, 2020.07.30.228460 (2020).
16. 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).
17. Rambaut, A. et al A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 5, 1403–1407 (2020).
18. Sarkar, A., Chakrabarti, A. K. & Dutta, S. Covid-19 Infection in India: A Comparative Analysis of the Second Wave with the First Wave. Pathogens 10, 1222 (2021).
19. Planas, D. et al Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. Nature 596, 276–280 (2021).
20. Davies, N. G. et al Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science 372, eabg3055 (2021).

21. Funk, T. et al Characteristics of SARS-CoV-2 variants of concern B.1.1.7, B.1.351 or P.1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021. Euro Surveill 26, (2021).
22. Yadav, P. D. et al An Epidemiological Analysis of SARS-CoV-2 Genomic Sequences from Different Regions of India. Viruses 13, 925 (2021).
23. Lyngse, F. P. et al Transmission of SARS-CoV-2 Omicron VOC subvariants BA.1 and BA.2: Evidence from Danish Households. Preprint at medRxiv, (2022).
24. Bager, P. et al Reduced Risk of Hospitalisation Associated With Infection With SARS-CoV-2 Omicron Relative to Delta: A Danish Cohort Study. https://papers.ssrn.com/abstract=4008930 (2022).
25. Sheikh, A., Kerr, S., Woolhouse, M., McMenamin, J. & Robertson, C. Severity of Omicron variant of concern and vaccine effectiveness against symptomatic disease: national cohort with nested test negative design study in Scotland. Preprint at Edinburgh Research Explorer, (2021).
26. Pulliam, J. R. C. et al Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa. Preprint at medRxiv, (2022).
27. Altarawneh, H. N. et al Protection against the Omicron Variant from Previous SARS-CoV-2 Infection. N Engl J Med, https://doi.org/10.1056/NEJMc2200133 (2022).
28. Rössler, A., Knabl, L., Laer, D. von & Kimpel, J. Neutralization profile of Omicron variant convalescent individuals. Preprint at medRxiv, 2022.02.01.22270263 (2022).