Original Article

Distribution and Functional Analyses of Mutations in Spike Protein and Phylogenic Diversity of SARS-CoV-2 Variants Emerged during the Year 2021 in India

Vidya Gopalan, Aswathi Chandran, Kishore Arumugam, Monisha Sundaram, Selvakumar Velladurai, Karthikeyan Govindan, Nivetha Azhagesan, Padmapriya Jeyavel, Prabu Dhandapani[1], Srinivasan Sivasubramanian, Satish Srinivas Kitambi[2]

Department of Virology, King Institute of Preventive Medicine and Research, [1]Department of Microbiology, Dr. ALM Post Graduate, Institute of Basic Medical Sciences, University of Madras, Chennai, Tamil Nadu, [2]Department of Translational Sciences, Institute for Healthcare Education and Translational Sciences, Hyderabad, Telengana, India

Abstract

Introduction: Prolonged COVID-19 pandemic accelerates the emergence and transmissibility of severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) variants through the accumulation of adaptive mutations. Particularly, adaptive mutations in spike (S) protein of SARS-CoV-2 leads to increased viral infectivity, severe morbidity and mortality, and immune evasion. This study focuses on the phylodynamic distribution of SARS-CoV-2 variants during the year 2021 in India besides analyzing the functional significance of mutations in S-protein of SARS-CoV-2 variants. **Methods:** Whole genome of SARS-CoV-2 sequences (n = 87957) from the various parts of India over the period of January to December 2021 was retrieved from Global Initiative on Sharing All Influenza Data. All the S-protein sequences were subjected to clade analysis, variant calling, protein stability, immune escape potential, structural divergence, Furin cleavage efficiency, and phylogenetic analysis using various *in silico* tools. **Results:** Delta variant belonging to 21A, 21I, and 21J clades was found to be predominant throughout the year 2021 though many variants were also present. A total of 4639 amino acid mutations were found in S-protein. D614G was the most predominant mutation in the S-protein followed by P681R, L452R, T19R, T478K, and D950N. The highest number of mutations was found in the N-terminal domain of S-protein. Mutations in the crucial sites of S-protein impacting pathogenicity, immunogenicity, and fusogenicity were identified. Intralineage diversity analysis showed that certain variants of SARS-CoV-2 possess high diversification. **Conclusions:** The study has disclosed the distribution of various variants including the Delta, the predominant variant, in India throughout the year 2021. The study has identified mutations in S-protein of each SARS-CoV-2 variant that can significantly impact the virulence, immune evasion, increased transmissibility, high morbidity, and mortality. In addition, it is found that mutations acquired during each viral replication cycle introduce new sub-lineages as studied by intralineage diversity analysis.

Keywords: Clade, COVID-19, India, mutations, phylogeny, severe acute respiratory syndrome coronavirus-2, spike protein

Introduction

Since the identification of a novel severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) virus in December 2019 in China, the world has observed many different variants with multiple mutations in the SARS-CoV-2 genome. The features of high transmission potential and replication of SARS-CoV-2 increase the possibility of accumulation of numerous virus adaptive mutations.[1] The WHO has developed a two major variant classification system such as variants of concern (VOC) and variants of interest (VOI). At present, there are mainly five VOC: Alpha (21A), beta (20H), gamma (20J), delta (21A, 21I, 21J), and Omicron (21K, 21 L) and five VOI: Kappa (21B), Epsilon (21C), Eta (21D), Iota (21F), and Mu (21H) (https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/;.https://covariants.org). COVID-19 pandemic is marked with rapid emergence of numerous variants and phenomenal phylogenic diversity among variants

Access this article online

Quick Response Code:

Website: www.jgid.org

DOI: 10.4103/jgid.jgid_178_22

Address for correspondence: Dr. Satish Srinivas Kitambi, Institute for Healthcare Education and Translational Sciences, 10-2-311, Plot 187, Str 4, Cama Manor, West Marredpally, Secunderabad - 500 026, Telengana, India.
E-mail: satish.kitambi@klife.info

How to cite this article: Gopalan V, Chandran A, Arumugam K, Sundaram M, Velladurai S, Govindan K, *et al*. Distribution and functional analyses of mutations in spike protein and phylogenic diversity of SARS-CoV-2 variants emerged during the year 2021 in India. J Global Infect Dis 2023;15:43-51.

Received: 20 September 2022 **Revised:** 27 November 2022
Accepted: 04 January 2023 **Published:** 17 May 2023

was observed in the year 2021 including the most virulent variant Delta.

The spike protein of SARS-CoV-2 plays a major role in cellular entry by interaction with human angiotensin-converting enzyme 2 (ACE2) receptor. The high-affinity binding of the spike protein receptor-binding domain (RBD) to hACE2 is an essential prerequisite for the rapid transmission of SARS-CoV-2 in humans. The SARS-CoV-2 variants such as Alpha, Beta, Gamma, Delta, and Omicron with mutations at the RBD display increase in viral infectivity and immune evasion.[2] Two main RBD conformations have been described for S-protein, standing-up, and lying-down states, with high and low affinity to ACE2, respectively. These states are influenced by the number and distribution of N-glycosylation and O-glycosylation sites in RBD impacting the interaction of S-protein of SARS-CoV-2 with the host cell and further transmissibility.[3] Observations on the changes in the states have been reported for emerging variants such as Omicron.[4] Preactivation of the S-protein by proteolytic cleavage is essential for viral entry into the host cells and mutations in proteolytic cleavage site of S-protein may affect the viral internalization process thus associated with altered transmissibility, virulence, and cell tropism.[5] There is also a need to closely monitor the antigenic evolution of S-protein in the circulating viruses through identifying the dynamic patterns of mutations indicative of positive selection for S-protein variants. Studies on the distribution of variants belonging to various clades and frequency and functional significance of mutations in S-protein genes of Indian SARS-COV-2 genomes are limited.[6] Hence, in this study, we retrieved 87,957 complete genomes of SARS-CoV-2 deposited from India in Global Initiative on Sharing All Influenza Data (GISAID) during the whole year 2021 and subjected to the studies on clade diversity and phylogenetic analysis to uncover the intra-lineage diversity of the SARS-CoV-2 variants. Besides, this study also focuses on the analyses of mutations and their frequency in the various regions of S-protein from all variants and functional significance of mutations affecting glycosylation patterns, protein stability, immunity, and virulence.

Methods

Genome retrieval and clade analysis

A total of 87,957 annotated SARS-CoV-2 whole genome sequences from the various parts of India deposited as on December 31, 2021 in GISAID (https://www.gisaid.org/) were retrieved. Sequences were aligned using Multiple Alignment using Fast Fourier Transform with SARS-CoV-2 Wuhan-Hu-1 strain (NC_045512.2) used as reference. All the sequences assembled in BioEdit. V.7 for spike gene trimming. The Nextclade-Nextstrain pipeline (https://clades.nextstrain.org/) was used for the clade analysis of SARS-CoV-2 sequences.

Mutation profiling, frequency, and functional analyses

The analyses of mutation profiling, frequency, and functional analyses including stability were performed with reference sequence SARS-CoV-2 Wuhan-Hu-1 strain (NC_045512.2). CoV surver mutation tool (https://www.gisaid.org/epiflu-applications/covsurver-mutations-app) was used to predict the nonsynonymous mutations in the spike gene and sequence analysis pipeline tool (https://cov.lanl.gov/content/sequence/TRACK_MUT/trackmut. html) was used for the analyses of frequency of mutation over time in India (source: GISAID). ESC_Comprehensive resource of immune escape variants in SARS-COV-2 was used to detect the immune escape mutants in S-protein (http://clingen.igib.res.in/esc/). The impact of mutations on S-protein stability was predicted using tools such as sorting intolerant from tolerant (SIFT) (https://sift.bii.astar.edu.sg/www/SIFT_seq_submit2.html), PROVEAN (Protein Variation Effect Analyzer) (http://provean.jcvi.org/seq_submit.php) and DUET (http://biosig.unimelb.edu.au/duet/stability). A SIFT score of 0.0–0.05 indicated a deleterious effect. The functional effects of protein variants were assessed using the PROVEAN web server, using a default threshold value of − 2.5 and the values below and above the threshold value were considered as deleterious and tolerant. The PDB structure (6VXX) was used as the reference for structural divergence and DUET analysis. DUET score displayed the predicted change in folding free energy upon mutation (ΔΔG in kcal/mol), with negative and positive values indicated destabilizing mutation and stabilizing mutation, respectively. RaptorX (http://raptorx.uchicago.edu/) is used to predict the secondary structure for different variants of SARS-CoV-2 spike protein. The impact of mutations in S-protein on Furin cleavage efficiency was studied by using ProP-1.0 Server (https://services.healthtech.dtu.dk/service. php? ProP-1.0) to predict the Furin cleavage site upon subjecting the mutations in the protease cleavage site (PCS) to Pro. P analysis. The score of reference S-protein mutations was compared with variants to predict the efficiency of Furin cleavage.

Phylogenetic analysis of severe acute respiratory syndrome coronavirus-2 variants

The Relative Synonymous Codon Usage (RSCU) of different variants was analysed using MEGA-X. Heat Map was constructed using CIM-miner (https://discover.nci.nih.gov/cimminer/). The average evolutionary divergence rate was estimated using Kimura-2 parameter model and phylogenetic tree construction was performed using MEGA-X. Supplementary files can be obtained by contacting the corresponding author.

Results

Clade analysis

The NEXTCLADE analysis of 87,957 SARS-CoV-2 sequences revealed that these strains were distributed into 21 different clades as represented by Nextstrain [Figure 1]. Some of the variants such as 19A, 19B, 20A, 20B, 20C, and 20D that were found since the early phase of pandemic during 2020 were also distributed in the year 2021. Among these 21 clades, 20A (7.9%) was more prevalent in early 2021 followed by 19B, 20B, 20C, 20D, 20E, and 20G. 19A, the first clade observed in the year 2020, persisted throughout the year and was less frequent (0.19%). The clades 19B (5.9%) and 20A (7.9%) were found to be highly distributed during the initial months (January

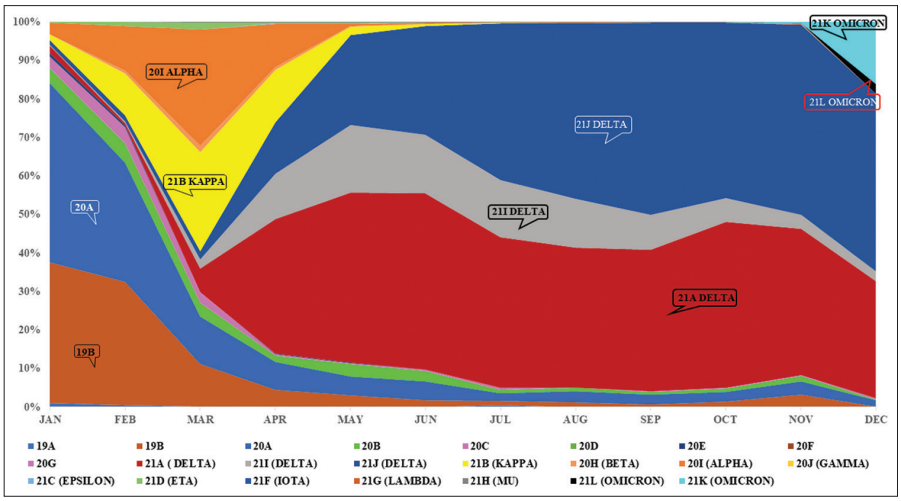

Figure 1: Distribution of SARS-CoV-2 clades in India during the whole year 2021. SARS-CoV-2: Severe acute respiratory syndrome coronavirus-2

to March 2021) but subsequently decreased. The VOC belonging to Alpha (20I) was observed in high prevalence (5.5%) from January to May 2021 and drastically reduced in December 2021. The VOC Beta (20H) was found to be sporadic (0.3%) from January to June 2021 and eventually decreased in the following months. The VOC Gamma (20J) (0.01%) had a negligible presence only from March to May and during August 2021. The VOC delta was the most predominant clade for the year 2021; however, the distribution of variants belonging to Delta clade were insignificantly lower than the other variants during the early months (January to March) of 2021 and a sweeping increase was observed from April 2021. The Delta variant was found to have three sub-lineages such as 21A, 21I and 21J. From January to June 2021, 21A (Delta) (33.3%) was more prevalent which was followed by 21J (26.7%) and 21I (Delta) (10.7%). The sub-lineage 21A (Delta) had decreased from July 2021, but there was a predominant increase in the other two sub-lineages. The distribution of clade 21B (Kappa) (5.6%) was elevated from February to May 2021 and decreased thereafter. The variants of clade 21C (Epsilon) (0.04%), 21F (Iota), and 21H (Mu) (0.05%) appeared in negligible fractions. The clade 21D (Eta) (0.29%) was observed only from January to June 2021. By the end of the year 2021, a new variant omicron had emerged and the study could identify two Omicron sub-lineages such as 21K (BA.1 Omicron) and 21 L (BA.2 Omicron). The presence of 21K (BA.1 Omicron) (0.24%) was first observed in November and there was a sudden spurt in their distribution in December 2021 whereas 21 L (BA.2 Omicron) (0.03%) was less prevalent when compared to 21K and observed only in December [Table 1].

Mutation profile of S-protein

A total of 4639 amino acid mutations were found in S-protein from the 87957 Indian sequences [Supplementary File 1]. There were 3052 and 1550 mutations present in the S1 and S2 domains, respectively with the highest number of mutations in the N-terminal domain (NTD; 1752 mutations) followed by (RBD; 820 mutations), Heptad Repeat 1 (HR1; 170 mutations), Heptad Repeat 2 (HR2; 143 mutations), Cytoplasm domain (CD; 85 mutations), (PCS; 75 mutations), Fusion peptide (FP; 72 mutations), Transmembrane domain (TMD; 57 mutations), and Signal peptide (SP; 37 mutations) [Table 2]. Among the 4639 mutations, 30 mutations were notably more predominant, in which D614G was present in 99% of Indian isolates ($n = 86663$) followed by P681R ($n = 68348$, 78%) of PCS, L452R ($n = 61310$, 70%), T478K ($n = 55694$, 63%) of RBD, T19R ($n = 56027$, 64%) of NTD [Figure 2; Supplementary File 2]. The results of mutation tracking analyses revealed the mutation frequency over time, in which D614 is completely replaced to G614. Variant specific mutations such as L452R, T19R, T478K, P681R, D950N, E156G, T95I, and G142D were originated in January and their presence was recorded throughout the year. Mutations D1118H, E484K, S982A, T716I, A570D, E154K, Q1071H, and H1101D could only be seen from January to May 2021. Among the 30 distinct amino acid mutations found in Omicron variants, 12 mutations notably existed from January 2021 while the other 18 mutations peaked only in December 2021. Few mutations such as G446V, P499R, and S371F were observed to be increased from August, October, and December, respectively. Among the 4,639 mutations, 1,947 mutations were observed once and 2692 mutations have repetition in 87,957 sequences. A total of 1201 mutated sites were found in S-protein of which only 174 had mutated once and the remaining 1027 sites carry more than one mutation.

Mutations affecting glycosylation patterns

Analysis of N-linked glycosylation (NGS) and O-linked Glycosylation (OGS) sites were performed for 87,957 isolates. S-protein carries 22 and 26 amino acid sites as NGS and OGS moieties, respectively. It was found that mutation in these sites resulted in loss of both NGS and OGS moieties. Two deletions were observed in these sites, one at 1194 position (NGS) and another at 1161 position (OGS), respectively. Even though there were 5 VOC and 5 VOI, none of the variant-specific mutation occurred in NGS and OGS moieties except one OGS mutation S982A in Alpha variant [Figure 3; Supplementary File 3].

Immune escape mutations in S-protein

Analysis showed 29 mutations in NTD and 469 mutations in

Table 1: Clade distribution profile of severe acute respiratory syndrome coronavirus-2 isolates (*n*=87,957) in India during the year 2021

Clade	January	February	March	April	May	June	July	August	September	October	November	December
19A	25	16	9	24	6	5	30	4	0	1	2	0
19B	837	1130	646	476	315	133	76	66	44	52	64	2
20A	1062	1082	721	774	538	364	135	170	130	88	69	15
20B	90	176	218	182	336	206	48	49	47	33	32	1
20C	67	147	143	32	32	28	40	7	6	8	1	4
20D	1	2	1	14	4	0	0	0	0	0	1	0
20E	19	16	6	3	2	4	3	2	0	0	1	0
20G	2	2	5	6	0	0	0	0	0	0	0	0
21A (Delta)	44	22	362	3798	4742	3459	2550	2108	2020	1529	772	279
21I (Delta)	7	10	143	1267	1895	1150	961	742	499	222	73	23
21J (Delta)	29	57	126	1457	2508	2128	2661	2645	2744	1618	1007	426
21B (Kappa)	31	382	1503	1481	234	37	8	7	3	0	0	0
20H (Beta)	1	26	94	65	16	6	0	0	0	0	0	0
20I (Alpha)	70	407	1778	1235	61	24	2	7	2	2	2	0
20J (Gamma)	0	0	2	1	1	0	0	1	0	0	0	0
21C (Epsilon)	0	0	1	4	10	1	3	0	2	1	1	0
21D (Eta)	2	34	106	32	13	3	0	0	0	0	0	0
21F (Iota)	0	0	0	0	2	0	0	0	0	0	0	0
21H (Mu)	0	0	1	6	16	3	1	4	1	1	2	0
21L (Omicron)	0	0	0	0	0	0	0	0	0	0	0	22
21K (Omicron)	0	0	0	0	0	0	0	0	0	0	7	148

Table 2: Amino acid substitution mutations observed across various regions of S-protein of Indian severe acute respiratory syndrome coronavirus-2 isolates

S-Protein	Positions	Number of mutations	Number of amino acid sites mutated
Signal peptide	1-13	37	13
N-terminal domain	14-305	1752	292
Receptor binding domain	319-541	820	220
Protease cleavage site	675-692	75	18
Fusion peptide	788-806	72	19
Heptad Repeat1	912-984	170	72
Heptad Repeat2	1163-1213	143	48
Transmembrane domain	1214-1237	57	23
Cytoplasm domain	1238-1273	85	29

RBD were found to have immune escape function. Common mutations among the variants such as N501Y (Alpha, Beta, Gamma, Omicron), K417N (Beta, Omicron), E484K (Beta, Gamma, Iota, Eta), L452R (Delta and Kappa) and T478K (Delta and Omicron) were resistant to neutralizing antibodies and could act as escape mutants. It was observed that Omicron had the highest immune escape potential with 19 escape mutations in S-protein followed by Gamma (8), Beta (5), Kappa (5), Iota (4), Delta (3), Eta (3), and Alpha (2) [Supplementary File 4].

Effect of mutations on stability and structural divergence of S-protein

Among the 4639 mutations, SIFT score predicted 1414 mutations were deleterious and 2771 mutations were neutral. Using PROVEAN score prediction tool, 638 mutations were found to be deleterious and 3547 mutations were neutral. Only 535 amino acid mutations were predicted to have potentially deleterious functional consequences on S-protein by both the mutation score prediction tools. A total of 2668 mutations were predicted to be neutral by SIFT and PROVEAN tools showing that these mutations might exhibit positive selection pressure for virus adaptability. It was observed that only 12 mutations such as T716I, R190S, T1027I, V1176F, D950N, Y145D, L212I, Y505H, T547K, N764K, N856K, and N969K present in different variants of SARS-CoV-2 could display deleterious effect on S-protein. Mutations in the RBD of different variants possess no deleterious effect on S-protein. For DUET score prediction, SARS-CoV-2 spike protein (PDB_ID– 6VXX) was used and the results revealed 6 mutations in NTD and 13 mutations in the RBD of SARS-CoV-2 variants were found to destabilize S-protein. Mutations present in the PCS of variants such as N679K, P681H, and P681R could showed stabilizing effect, whereas Q677H in Eta variant displayed destabilizing effect on S-protein. Mutations were not observed in FP of S-protein of variants except N764K in Omicron variant and it had the stabilizing effect on the protein. The functional effect of several mutations could not be predicted due to the absence of position in PDB structure [Supplementary File 5]. The S-protein of major variants of SARS-CoV-2 in the current study was subjected to secondary structure prediction using the RAPTORX tool including the reference S-protein (NC_045512) [Table 3]. Analysis revealed that the alpha variant had increased β-pleated sheet but reduced coil structure; Delta variant had increased coil but decreased β-sheet structure. Gamma and Omicron variants had increased alpha helix and β-pleated sheet

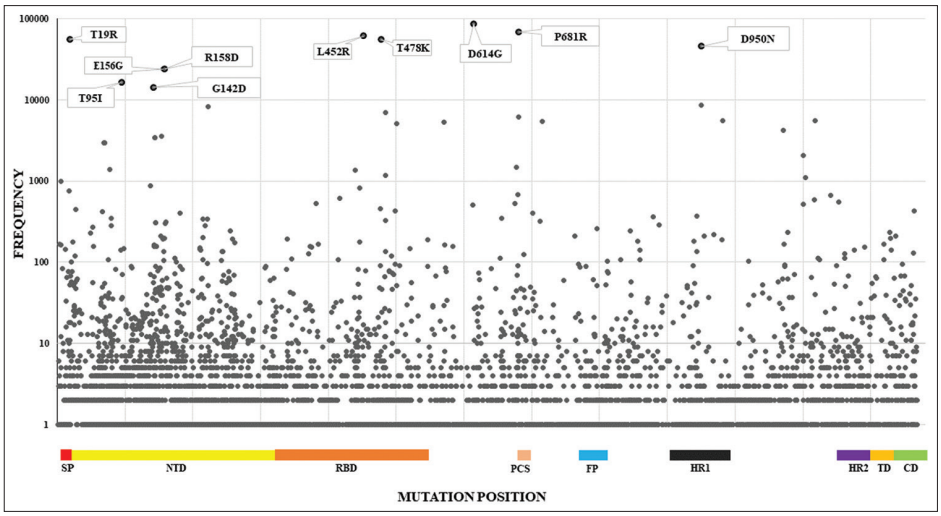

Figure 2: The mutation frequency of S-protein. The mutation frequency were plotted with respective position of S-protein. SP: Signal Peptide, NTD: N Terminal Domain, RBD: Receptor binding domain, PCS: Protease cleavage site, FP: Fusion peptide, HR1: Heptad Repeat 1, HR2: Heptad Repeat 2, TD- Transmembrane Domain, CD: Cytoplasm Domain

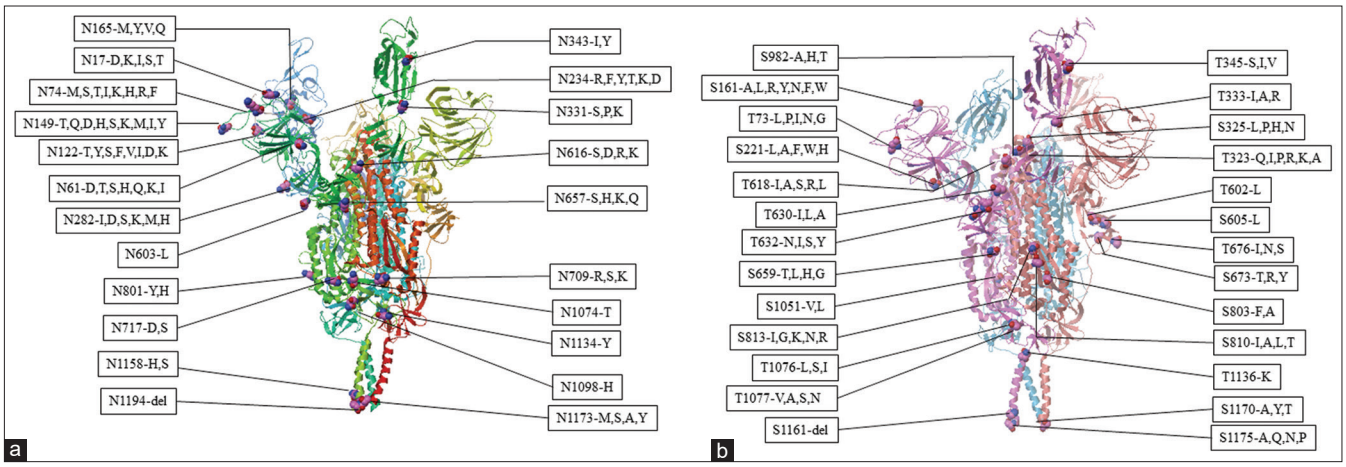

Figure 3: (a) Mutations in N-glycosylation sites; (b) Mutations in O-Glycosylation sites. All glycosylation sites were marked to their respective sites on the trimeric S-protein and the corresponding mutations were marked as single letter amino acid code

and had similar secondary structure conformation to reference S-protein. The PDB structures of S-protein in SARS-CoV-2 variants (PDB-ID-Alpha [7 LWV], Beta [7 LYN], Gamma [7V79], Delta [7V7Q], and Omicron [7TNW]) were retrieved and analyzed for tertiary structure variation in NTD and RBD regions using pair wise structure alignment tool in PDB with S-protein (PDB-ID-6VXX) as reference. The results showed that Omicron had unique α-helix at 142-145 amino acid residues in NTD which was absent in other variants. The Delta and Gamma variants had α-helix at position 436–441 in RBD, but the structure was absent in Alpha, Beta, and Omicron variants. In contrast, Omicron, Alpha, and Beta had α-helix at 346–350 in RBD but were not observed in Delta and Gamma variants [Figure 4].

Influence of mutations on Furin cleavage potential

The S-protein sequences with mutation in the PCS region were subjected to Pro. *P* tool to predict the Furin cleavage potential based on scores. We compared the mutations with score 0.620 of reference S-protein. Mutations with score >0.620 had increased cleavage efficiency and values < 0.620 had decreased cleavage efficiency [Supplementary File 6]. Analysis of 73 mutations revealed that 31 mutations had increased cleavage potential, whereas 32 mutations decreased the cleavage potential. The mutations P681H and P681R had increased furin cleavage function. The mutation in Arginine residue (R685) leads to the absence of the cleavage site. Mutation S686A had increased the cleavage, whereas mutation V687 decreased the cleavage potential. Moreover, deletion at position 679 and 681 also increases the cleavage potential. Mutations present in the neighboring residues of the cleavage site also influences the cleavage potential.

Evolutionary analysis of severe acute respiratory syndrome coronavirus-2 variants

The RSCU was analyzed for all the spike coding sequence of SARS-CoV-2 variants Mega X [Supplementary File 7]. Notably, all the variants were possessing similar codon usage with small difference in the frequency of usage. It was found

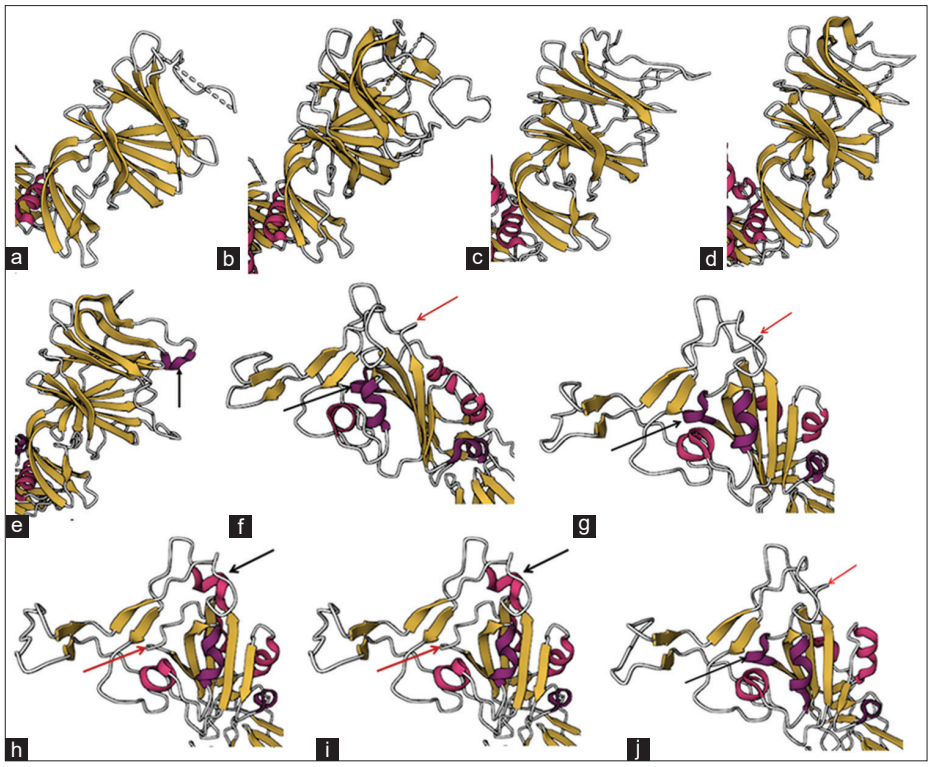

Figure 4: (a) NTD in Alpha variant; (b) NTD in Beta variant; (c) NTD in Gamma variant; (d) NTD in Delta variant; (e) NTD in Omicron variant; (f) RBD in Alpha variant; (g) RBD in Beta variant; (h) RBD in Gamma variant; (i) RBD in Delta variant; and (j) RBD in Omicron variant. Black and red arrows indicate the presence and absence of α-helix respectively. NTD: N Terminal Domain, RBD: Receptor binding domain

Table 3: Proportion of α-helix, β-pleated sheet and coil in tertiary structure of S-protein in Indian severe acute respiratory syndrome coronavirus-2 variants

S-Protein of SARS-CoV-2 variants	Alpha helix (%)	Beta sheet (%)	Coil (%)
Reference (NC_045512)	19	31	48
Alpha spike	19	32	47
Beta spike	19	31	48
Gamma spike	20	31	48
Delta spike	19	30	49
Omicron spike	20	31	48

SARS-CoV-2: Severe acute respiratory syndrome coronavirus-2

that a rare codon CGA which codes for Arginine was present only in Omicron. The total number of codons in Omicron (1265 codons) was less than the sequences of other variants that had 1274 codons, and this could be probably due to the deletion mutations in Omicron variant. No drastic difference was observed in the codon usage of variants, suggesting the genomic conservancy among the variants for their survival and evolution [Figure 5a]. Only 14,937 high quality S-protein genes were taken for studying region-specific average evolutionary divergence. The results showed that PCS had more divergence rate of 1.11×10^{-2} s/s/y followed by NTD (3.27×10^{-3}), RBD (2.52×10^{-3}), and SP (1.53×10^{-3}). The other regions such as FP, HR1, HR2, TM, and CD had very low divergence rate [Figure 5b]. Intra lineage diversity among the SARS-CoV-2 variants were analyzed by random selection of 100 genes of S-protein from each variant and subjected to phylogenetic analysis with NC_045512 as reference. The results revealed that strains belong to Alpha and Beta variants were less diverse suggesting the sequences were closely related among themselves. Highly diverse sequences were observed in Delta, Kappa, Eta, and Omicron variants, with six different clusters in Delta and Omicron, followed by five clusters in Kappa and Eta variants [Figure 6].

Discussion

Progression of COVID-19 pandemic favors the emergence of new SARS-CoV-2 variants through accumulation of adaptive mutations, especially in S-protein resulting in increase in the transmissibility of variants. This underscores the importance of tracking the evolution of S-protein in SARS-CoV-2 by means of mutational, phylogenetic and functional analyses. The changing phylodynamics indicates the necessity to conduct countrywide or regional studies on clade distribution patterns along with mutational analyses that will provide new insights on their epidemiology as well as for evolving therapeutic and prophylatic measures. In this study, we report the phylogenetic distribution of variants based on the SARS-CoV-2 genomes deposited in GISAID from India, mutation frequency and functional analyses of amino acid mutations in S-protein with respect to alteration in glycosylation patterns, immune escape features, protein stability, and evolution.

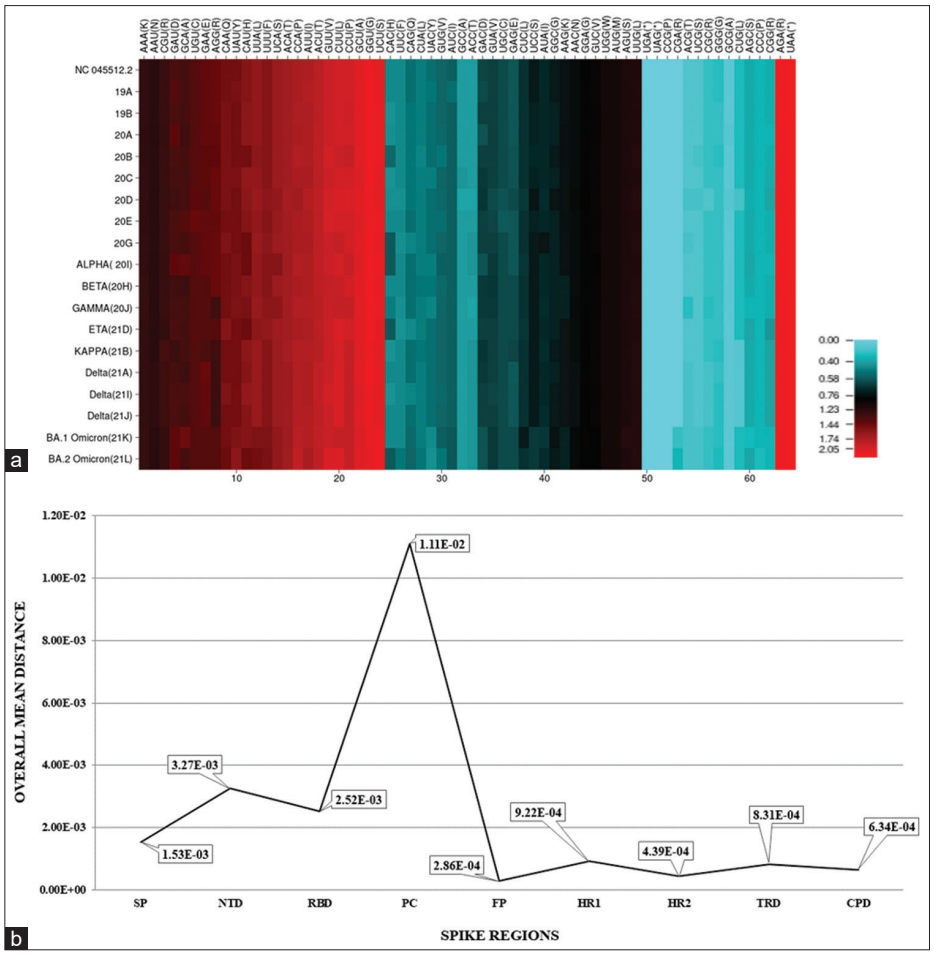

Figure 5: (a) Heat map of RSCU values for the spike coding sequence of SARS-CoV-2 variants. Codons with higher and lower RSCU values are highlighted in red and blue respectively; (b) The average evolutionary rate for different regions of S-protein. SP: Signal Peptide, NTD: N-Terminal Domain, RBD: Receptor Binding Domain, PCS: Protease Cleavage site, FP: Fusion Peptide, HR1: Heptate Repeat 1, HR2: Heptate Repeat 2: TRD- Transmembrane Domain, CPD: Cytoplasm Domain, RSCU: Relative synonymous codon usage

Though large number of SARS-CoV-2 complete genomes from various parts of India have been sequenced and deposited in GISAID, few studies are available on clade distribution, frequency, and functional significance of mutations of SARS-CoV-2 isolates from India.[6,7] Genomic surveillance studies in India had reported the emergence of Delta variant by the end December 2020 and this variant had gradually replaced Alpha variant in May 2021.[8,9] The current study also records the dominancy of Delta variant throughout 2021, and branching of the variant into three sub-lineages namely 21A, 21I and 21J during the year 2021. Sub-lineage 21A (Delta) was more predominantly observed from January to June 2021 which was followed by sub-lineages 21J and 21I. The prevalence of 21A decreased by July 2021 while 21J and 21I remained predominant thereafter. A new variant, Omicron was observed by the end of 2021.

Our previous study had revealed the presence of 557 amino acid substitutions in S-protein of SARS-CoV-2 isolates circulated in India during 2020.[6] The present study observed a total of 4639 mutations in S-protein of all genomes deposited in the year 2021 denoting about 8-fold increase in the rate of mutation occurrence against the previous year. It is clearly observed that there was an increase in the events of mutational occurrence as well as emergence of diverse variants when the pandemic progresses from the early stage. Four hundred and four substitution mutations were found to exist in both years 2020 and 2021. Notably, D614G was profoundly present in 86,663 sequences (99%) in 2021, followed by P681K (78%), L452K (70%), T19K (64%), T478K (63%), D950N (52%), E156G (27%), E157del (27%) and R158 (27%). Among these high frequent mutations, E156G, E157del and R158 are reported to enhance neutralizing antibodies resistance and infectivity.[10] Other mutations that were observed in the study such as L452R, Y453F and N501Y on RBD were found to increase the binding affinity of S-protein with hACE2 receptor resulting in increase in infectivity.[11] The study also reported that these three mutations were found in Beta variant thereby increasing the binding affinity by 3-fold than the primitive strain and the double mutation E484K/N501Y in Mu variant which was observed in low frequency from the present study was found to have stronger binding affinity than single mutation N501Y.[12] A study reported that RBD-specific mutation V367F also has higher affinity to hACE2 receptor.[13]

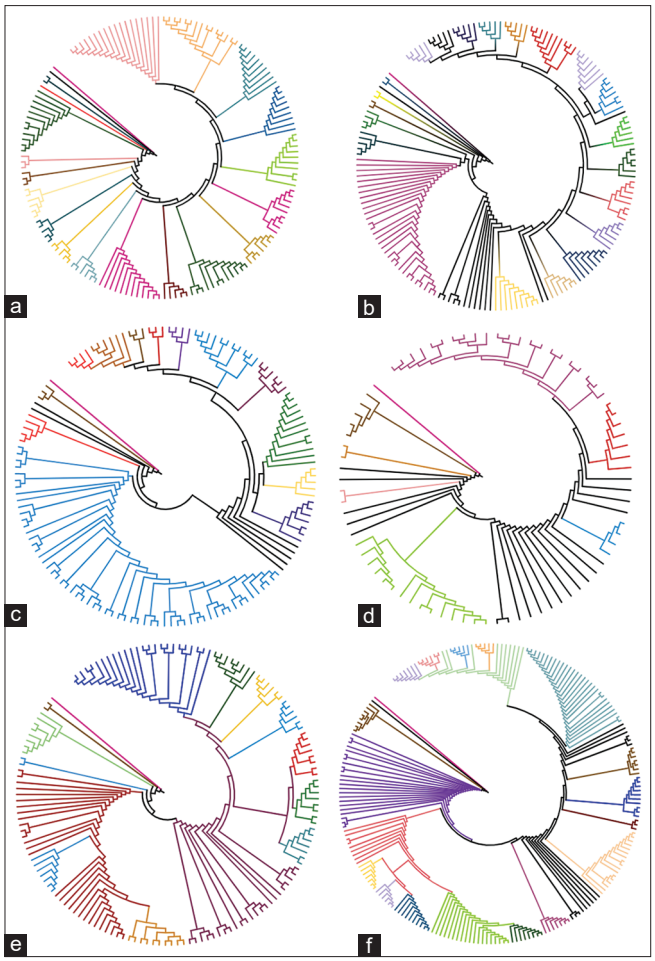

Figure 6: The phylogenetic tree was constructed by Maximum-Likelihood method having the root as SARS-CoV-2 Wuhan-Hu-1 Spike sequence (NC_045512.2). Intra-lineage diversity of (a) Alpha; (b) Beta; (c) Delta; (d) Eta; (e) Kappa; and (f) Omicron variants. SARS-CoV-2: Severe acute respiratory syndrome coronavirus-2

The N-linked glycans play an important role in structural and functional dynamics of RBD of S-protein. Mutations at N165 and N234 glycosylation sites in NTD region can reduce the binding of S-protein with hACE2 receptor due to RBD shift in the down state.[14] The current study observed mutations at both these sites including N165M, N165Y, N165V, N165Q and N165del at position 165, and N234R, N234F, N234Y, N234T, N234K and N234D at position 234. A study reported that mutation at N801 and N1194 in N-glycan site disrupted S-protein trimerization.[15] The present study reports such mutations N801Y, N801H and a stop mutation at position 801, and a deletion at 1194 position.

The study has identified several mutations that have increased cleavage potential at PCS (675–692 amino acid region) essential for viral entry into the host cell. Among the mutations, P681H and P681R exhibited increased furin cleavage function. Variants with mutation P681H are found to have more molecular flexibility for facilitating furin binding to cleavage site.[16] P681R mutation at PCS boosts the cleavage of S1 and S2, accelerates viral fusion and thus leading to increased infectivity of cells.[17,18]

The fitness of SARS-CoV-2 virus for survival was driven by mutations possessing intra and interprotomer interactions. The D614 in prototype S-protein forms H-bonds with K854 and T859 when the RBD is in close conformation and these bonds are lost during the RBD open conformation. This interprotomer H-bonding is absent in D614G mutant leading to reduction of energy required for the conformational transition.[19] It was reported that the Alpha variant with mutation A570D, D614G, S982A and D1118H enables local side chain rearrangements, giving rise to additional interprotomer contacts.[20] The residues S929 and D936 are engaged in side chain H-bonds with S1196 and R1185 of HR2 during the postfusion conformation of the protein and bringing the viral and cellular membrane for fusion. It was observed that mutations in this position D936Y result in loss of inter monomer H-bond which results in reduced protein assembly.[21]

Reports suggest that there are only minor structural differences in S-protein of SARS-CoV-2 variants;[22] however, it is reported that Omicron has higher fraction of α-helix (23.46%) than Delta (22.03%)[23] and the same was observed in the present study. A study discloses that S-protein of SARS-CoV-2 Wuhan reference strain contains mixture of RBD open and closed conformations while S-protein of Omicron has predominantly RBD open conformation. Mutation Q853K in S-protein of Omicron can alter the disordered loop resulting in tighter S1/S2 packing. Further, mutations such as T547K, N764K can promote S1/S2 packing. Overall mutations in S-protein of Omicron introduce new inter-domain and inter-subunit interactions which stabilizes RBD open conformation.[4]

A study on S-protein of SARS-CoV-2 demonstrated the importance of TMD in cellular membrane fusion and virus entry. Any mutation that alter the aromatic and Cysteine rich residues of TMD can affect membrane fusion and entry.[24] In this study, we have observed mutations such as W1214R, Y1215H, C1235F, C1235R, C1236F, C1236V and C1236S in aromatic and Cysteine residues of TMD. New mutations that are acquired during each replication results in the intra lineage diversity of SARS-CoV-2.[25,26] This study reports that Delta and Omicron variants have highly diverse sequences among themselves giving rise to numerous new sub-lineages and sub-variants such as 21A, 21I, 21J in Delta variant and 21 L and 21K in Omicron variant due to accumulation of numerous mutations at high frequency other than clade specific signature mutations. Among the Omicron sub-variants and VOCs, we identified several such mutations that were shared among the isolates belonging to various clades. These unique mutations make the isolates to be distributed diversely within the clade of phylogenetic tree. This suggests that occurrence of such mutations are contributing not only to the antigenic diversity of S-protein but also to facilitate potentially the emergence of more subclades or new subvariants with acquisition of few other mutations. Multiple inter-variant recombination events may have contributed to the shared presence of different mutations between the VOCs. For example, the BA.1 subvariant has three more Alpha-related mutations (del69-70, delY144) than

BA.2, suggesting that Alpha or other unknown variants that carry these mutations may have contributed to the emergence of the BA.1 subvariant.[27] In view of rapid emergence of new variants with phenomenal diversity in the global distribution of variants, continuous monitoring of genomic evolution of SARS-CoV-2 is essential for supporting tasks on vaccine design and development programmes and devising control and preventive measures to manage this infection.

Conclusions

The study has revealed the dynamics of rapidly diversifying SARS-CoV-2 variants and subvariants with a phenomenal observation of shifting of clade predominance within 2 years of the introduction of virus in India. The functional evaluation of several mutations in S-protein, after analyzing all the sequences deposited throughout the year 2021, reveals the significance of various mutations in virulence, immune escape features and disease severity. The findings of the study may support researchers to understand the phylodynamic characteristics, molecular epidemiology and mutation based functional characteristics of variants and sub-variants of SARS-CoV-2.

Research quality and ethics statement

This study was approved by the Institutional Review Board/ Ethics Committee. This work only requires analysis of openly available information and does not require patients or patient samples or data. The authors followed applicable EQUATOR Network (https://www. equator-network. org/) guidelines during the conduct of this research project.

Acknowledgements

We are grateful to all the authors, originating and submitting laboratories from Global Initiative on Sharing All Influenza Data (GISAID's EpiCov database) for enabling the sequences available for use in our study. The authors thank Institute for Healthcare Education and Translational Sciences (www. ihets. info) and Kitambi Foundation for the financial support.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Lou F, Li M, Pang Z, Jiang L, Guan L, Tian L, *et al.* Understanding the secret of SARS-CoV-2 variants of concern/interest and immune escape. Front Immunol 2021;12:744242.
2. Liu J, Chen X, Liu Y, Lin J, Shen J, Zhang H, *et al.* Characterization of SARS-CoV-2 worldwide transmission based on evolutionary dynamics and specific viral mutations in the spike protein. Infect Dis Poverty 2021;10:112.
3. Yao H, Song Y, Chen Y, Wu N, Xu J, Sun C, *et al.* Molecular architecture of the SARS-CoV-2 virus. Cell 2020;183:730-8.e13.
4. Ye G, Liu B, Li F. Cryo-EM structure of a SARS-CoV-2 Omicron spike protein ectodomain. Nat Commun 2022;13:1214.
5. Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, *et al.* Cell entry mechanisms of SARS-CoV-2. Proc Natl Acad Sci U S A 2020;117:11727-34.
6. Sivasubramanian S, Gopalan V, Ramesh K, Padmanabhan P, Mone K, Govindan K, *et al.* Phylodynamic pattern of genetic clusters, paradigm shift on spatio-temporal distribution of clades, and impact of spike glycoprotein mutations of SARS-CoV-2 isolates from India. J Glob Infect Dis 2021;13:164-71.
7. Yadav PD, Nyayanit DA, Majumdar T, Patil S, Kaur H, Gupta N, *et al.* An epidemiological analysis of SARS-CoV-2 genomic sequences from different regions of India. Viruses 2021;13:925.
8. Cherian S, Potdar V, Jadhav S, Yadav P, Gupta N, Das M, *et al.* SARS-CoV-2 spike mutations, L452R, T478K, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India. Microorganisms 2021;9:1542.
9. Dhar MS, Marwal R, Radhakrishnan VS, Ponnusamy K, Jolly B, Bhoyar RC, *et al.* Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India. Science 2021;374:995-9.
10. Mishra T, Dalavi R, Joshi G, Kumar A, Pandey P, Shukla S, *et al.* SARS-CoV-2 spike E156G/Δ157-158 mutations contribute to increased infectivity and immune escape. Life Sci Alliance 2022;5:e202201415.
11. Motozono C, Toyoda M, Zahradnik J, Saito A, Nasser H, Tan TS, *et al.* SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity. Cell Host Microbe 2021;29:1124-36.e11.
12. Laffeber C, de Koning K, Kanaar R, Lebbink JH. Experimental evidence for enhanced receptor binding by rapidly spreading SARS-CoV-2 variants. J Mol Biol 2021;433:167058.
13. Ou J, Zhou Z, Dai R, Zhang J, Zhao S, Wu X, *et al.* V367F mutation in SARS-CoV-2 spike RBD emerging during the early transmission phase enhances viral infectivity through increased human ACE2 receptor binding affinity. J Virol 2021;95:e0061721.
14. Casalino L, Gaieb Z, Goldsmith JA, Hjorth CK, Dommer AC, Harbison AM, *et al.* Beyond shielding: The roles of glycans in the SARS-CoV-2 spike protein. ACS Cent Sci 2020;6:1722-34.
15. Huang HY, Liao HY, Chen X, Wang SW, Cheng CW, Shahed-Al-Mahmud M, *et al.* Vaccination with SARS-CoV-2 spike protein lacking glycan shields elicits enhanced protective responses in animal models. Sci Transl Med 2022;14:eabm0899.
16. Joshi N, Tyagi A, Nigam S. Molecular level dissection of critical spike mutations in SARS-CoV-2 variants of concern (VOCs): A simplified review. ChemistrySelect 2021;6:7981-98.
17. Liu Y, Liu J, Johnson BA, Xia H, Ku Z, Schindewolf C, *et al.* Delta spike P681R mutation enhances SARS-CoV-2 fitness over alpha variant. Cell Rep 2022;39:110829.
18. Saito A, Irie T, Suzuki R, Maemura T, Nasser H, Uriu K, *et al.* Enhanced fusogenicity and pathogenicity of SARS-CoV-2 delta P681R mutation. Nature 2022;602:300-6.
19. Ostrov DA. Structural consequences of variation in SARS-CoV-2 B.1.1.7. J Cell Immunol 2021;3:103-8.
20. Mannar D, Saville JW, Sun Z, Zhu X, Marti MM, Srivastava SS, *et al.* SARS-CoV-2 variants of concern: Sp1ike protein mutational analysis and epitope for broad neutralization. Nat Commun 2022;13:4696.
21. Oliva R, Shaikh AR, Petta A, Vangone A, Cavallo L. D936Y and other mutations in the fusion core of the SARS-CoV-2 spike protein heptad repeat 1: Frequency, geographical distribution, and structural effect. Molecules 2021;26:2622.
22. Cueno ME, Imai K. Structural insights on the SARS-CoV-2 variants of concern spike glycoprotein: A computational study with possible clinical implications. Front Genet 2021;12:773726.
23. Kumar S, Thambiraja TS, Karuppanan K, Subramaniam G. Omicron and delta variant of SARS-CoV-2: A comparative computational study of spike protein. J Med Virol 2022;94:1641-9.
24. Corver J, Broer R, van Kasteren P, Spaan W. Mutagenesis of the transmembrane domain of the SARS coronavirus spike glycoprotein: Refinement of the requirements for SARS coronavirus cell entry. Virol J 2009;6:230.
25. Baj A, Novazzi F, Drago Ferrante F, Genoni A, Tettamanzi E, Catanoso G, *et al.* Spike protein evolution in the SARS-CoV-2 Delta variant of concern: A case series from Northern Lombardy. Emerg Microbes Infect 2021;10:2010-5.
26. Viana R, Moyo S, Amoako DG, Tegally H, Scheepers C, Althaus CL, *et al.* Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in Southern Africa. Nature 2022;603:679-86.
27. Ou J, Lan W, Wu X, Zhao T, Duan B, Yang P, *et al.* Tracking SARS-CoV-2 Omicron diverse spike gene mutations identifies multiple inter-variant recombination events. Signal Transduct Target Ther 2022;7:138.