ORIGINAL ARTICLE

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

Srinivasan Sivasubramanian[1], Vidya Gopalan[1], Kiruba Ramesh[1], Padmapriya Padmanabhan[1], Kiruthiga Mone[1], Karthikeyan Govindan[1], Selvakumar Velladurai[1], Kaveri Krishnasamy[1], Satish Srinivas Kitambi[2,*]

[1] State Viral Research and Diagnostic Laboratory (VRDL) Unit, Department of Virology, King Institute of Preventive Medicine and Research, Chennai, Tamil Nadu, India – 600 032.

2:Institutet for Healthcare Education and Translational Sciences (IHETS), Hyderabad, Telengana -500026.

***Corresponding author.** **Dr. Satish Srinivas Kitambi**

Institute for Healthcare Education and Translational Sciences (IHETS)

10-2-311, Plot 187, Str4, Cama Manor, West Marredpally, Telengana 500026.

Email address: satish.kitambi@klife.info

Running Title: Analyses of SARS-CoV-2 isolates from India

Keywords: SARS-CoV-2, COVID-19, Spike protein, Clade, Phylogeny, Mutations, India

Abstract

Background: The COVID-19 pandemic is associated with high morbidity and mortality, with the emergence of numerous variants. The dynamics of SARS-CoV-2 with respect to clade distribution is uneven, unpredictable and fast changing. **Aims:** Retrieving the complete genomes of SARS-CoV-2 from India and subjecting them to analysis on phylogenetic clade diversity, Spike (S) protein mutations and their functional consequences such as immune escape features and impact on infectivity. **Methods:** Whole genome of SARS-CoV-2 isolates (n=4,326) deposited from India during the period from January 2020 to December 2020 is retrieved from GISAID and various analyses performed using *in silico* tools. **Results:** Notable clade dynamicity is observed indicating the emergence of diverse SARS-CoV-2 variants across the country. GR clade is predominant over the other clades and the distribution pattern of clades is uneven. D614G is the commonest and predominant mutation found among the S-protein followed by L54F. Mutation score prediction analyses reveal that there are several mutations in S-protein including the RBD and NTD regions that can influence the virulence of virus. Besides, mutations having immune escape features as well as impacting the immunogenicity and virulence through changes in the glycosylation patterns are identified. **Conclusions:** The study has revealed emergence of variants with shifting of clade dynamics within a year in India. It is shown uneven distribution of clades across the nation requiring timely deposition of SARS-CoV-2 sequences. Functional evaluation of mutations in S-protein reveals their significance in virulence, immune escape features and disease severity besides impacting therapeutics and prophylaxis.

INTRODUCTION

Analyses of global and Indian SARS-CoV-2 genome sequences (as on December 2020) have revealed that the virus has differentially distributed into at least 10 clades and is continuously evolving.[1] The S-protein of SARS-CoV-2 targets angiotensin-converting enzyme 2 (ACE2) receptor for its entry into target cells. This protein is the major focus of the vaccine development platforms. Changes in the O- and N- linked glycosylation patterns of the S protein have an impact on the immunogenicity and virulence of the virus. Hence, it is important to closely monitor antigenic evolution of the S-protein in the circulating viruses. In this study, we retrieved complete genomes of SARS-CoV-2 from India during the whole year period from GISAID and subjected them to the studies on clade analyses and clade distribution pattern covering all states of the country. Further, mutations in various regions of S-protein, mutation frequency, glycosylation patterns, and the effects on protein structure, immunity and virulence were analysed.

METHODS

Genome data retrieval, phylogenetic and clade analysis

A total of 4326 annotated SARS CoV-2 whole genome sequences (WGS) from various parts of India deposited as on 31st December 2020 in Global Initiative on Sharing All Influenza Data (GISAID)(https://www.gisaid.org/) were retrieved. Sequences were aligned using MAFFT (Multiple Alignment using Fast Fourier Transform) with SARS-CoV-2 Wuhan-Hu-1 strain (NC_045512.2) and GISAID reference sequence (EPI_ISL_402124)[2] used as reference. The Nextclade-Nextstrain pipeline (https://clades.nextstrain.org/) was used for studies on phylogenetic analysis and clustering patterns of the S gene.[3] Further, the Average evolutionary divergence was estimated using Kimura-2 parameter model. Evolutionary analyses and phylogenetic tree construction were performed using MEGA-X.[4]

Frequency and functional evaluation of variants

Frequencies and amino-acid variants were analyzed using COVID CG and Tracking mutation tools (source GISAID) respectively. Functional consequences were predicted using tools like SIFT (Sorting Intolerant from Tolerant) (https://sift.bii.astar.edu.sg/www/SIFT_seq_submit2.html),[5] PROVEAN (Protein Variation Effect Analyzer) (http://provean.jcvi.org/seq_submit.php)[5] and PolyPhen-2 (Polymorphism Phenotyping v2) (http://genetics.bwh.harvard.edu/pph2).[6] A SIFT score of 0.0 to 0.05 indicates 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 are considered as deleterious and tolerant respectively. A PolyPhen threshold scores of > 0.908, >0.446 and ≤ 0.908 and ≤ 0.446 are interpreted as “Probably Damaging”, “Possibly Damaging” and “Benign” respectively. ESC_Comprehensive resource of immune escape variants in SARS-COV-2 was used to detect the escape mutants in S-protein (http://clingen.igib.res.in/esc/).[7]

RESULTS

The retrieved WGS were found to be classified under 7 clusters according to GISAID Clade identification [Figure 1].[8] It was observed that the predominant cluster encompassed 1,755 (40.56%) of genomes that fell under the GR clade [Figure 1a]. Though this clade was represented by samples derived from various states across India, Maharashtra (n = 922) and

Telangana (n = 492) states had the maximum numbers followed by Karnataka (n = 102) [Figure 1b]. The major clade GR was followed by clades G (942; 21.77%), O (783; 18.1%), GH (737; 17.03%), S (82; 1.9%), L (25; 0.58%) and V (3; 0.07%). The clade G was mainly represented by samples from Maharashtra (n = 277), Gujarat (n = 215) and West Bengal (n = 152). The O clade is prevalent in all states of the country. Gujarat state accounts for the highest number of samples under GH clade [Figure 1b]. States such as Andhra Pradesh, Punjab and Rajasthan submitted a smaller number of sequences and the clade diversity pattern could not be clearly deciphered.

The viruses belonging to L, S and O clades were prevalent during the initial months (January to February, 2020) [Figure 1c]. During the starting of the pandemic (March to April), O clade was predominant followed by G, GR, GH, S and L. It is noteworthy that the distribution of S and L clades were drastically reduced during this period and the strains belonged to clades O, S, L and V were remarkably low in numbers during the progress of pandemic. From May to October, GR clade is predominant but becomes second to GH clade during November and December. Notably, the O clade was slowly dominated by GR, G and GH clades in different states during the course of pandemic and there was almost near to complete absence of O clade during November and December. The phylogenetic tree depicting clade diversity throughout the year shows that GR is the dominant clade over the others [Figure 2]. These results suggested spatiotemporal clade diversity and a paradigm shift in phylodynamics of clade distribution.

Mutation analysis of spike protein from Indian strains

A total of 557 amino acid substitution mutations were found in S-protein among the 4,326 Indian strains [Supplementary File 1]. There were 333 and 215 mutations present in the S1 and S2 domains respectively with the highest number of mutations in the N-terminal domain (NTD; 211 mutations) followed by the RBD (63 mutations) [Table 1]. Nine mutations are identified in signal peptide, which is not the component of mature S protein. Among these 557 mutations, D614G was present in 79.99% (n = 3461) of Indian strains followed by L54F (n=111, 2.57%) isolates. The other prominent mutation sites were: Q677 (72), P681 (54), P812 (40), A771 (34), Q675 (30), and L5 (26) [Figure 3]. Besides, 11 types of mutations are found in the 8 sites of highly conserved protease cleavage region (from 675 to 692 of S1 and S2 domains) of the protein. L18F, H69del, V70del, D138Y and Y144del mutations were observed in NTD of S-protein of few isolates and these mutations could enhance the surface electropositivity of the S-protein and thereby facilitating the adhesion of virus to negatively charged lipid raft gangliosides of host cells.[9] It is also observed that two of the study variants possess H69del, V70del and Y144del in NTD and N501Y in RBD suggesting the improved affinity as well as adhesive properties of S-protein due to the concomitant mutations in both regions that synergistically promote virus host interaction.

The frequency of amino acid mutations in S-protein was analyzed using COVID-19 CoV Genetics browser (source: GISAID), and the results showed that non-synonymous mutations were scattered across the S-gene with region specific varying frequency [Supplementary File 2]. Figure 4 shows prevalent mutations such as D614G, Q677H and P681H originated during March, April and July respectively and their appearance was observed till the end of the year 2020. On contrary, L54F as well as K77M and P812L mutations emerged during April and June respectively but absent after few months of their appearance.

Many amino acid mutations were observed to be region specific namely F32Y, T33K and G35Q mutations (in Karnataka); T29I and P681H (Maharashtra); and L7S, L54F, R78M, Q690H, A701T and A879S (Gujarat). These mutations were absent from other states indicating that these mutations might not spread to other states possibly due to effective implementation of lockdown measures throughout the country. Some distinct amino acid variants were observed in Gujarat and Maharashtra (G181A) and V622F in Telangana and Orissa. There were 12 premature stop codons and 8 deletion mutations present in different positions of various S-gene sequences. More than one mutation type can be observed at the same position in the protein. For instance, amino acid A to V, E, S, or K, at position 27, A to G, T, S, or V at position 222. Among the total 419 mutation sites in S-protein of Indian isolates, 114 sites carry more than one mutation. It is noteworthy that there were 190 distinct mutation events that occurred in India first time; among them, 115 mutation events were confined only to India and the rest of 75 mutations were subsequently identified in various countries or occurred independently at different geographical regions across the world [Supplementary File 2].

Immune escape mutations in spike protein

The analyses showed 11 and 17 immune escape mutations in the NTD and RBD of S-protein respectively [Supplementary File 3]. L18F, T19A, D80N, D138Y, Y144del, Y145del, K147E, N148S, W152L, Q218H and S255F were found in NTD, and among them, L18F, Y144del, Y145del and N148S and W152L were shown to display resistance to neutralizing antibodies. Among the mutations in RBD, R346K, N440K, G446V, N450K, V483F, E484K, E484Q, F490S and S494P also showed change in ACE2 binding to the extent of 75% to 90%.[10,11] Variants identified with mutations at sites such as E484 (E484Q), F490 (F490S), Q493 (Q493STOP), and S494 (S494P) in the RBD are presumed to have immune escape features.[12]

Mutations in regions of S-protein other than RBD also can show resistance to antibody and are identified in the present study. It is noteworthy that single amino acid changes such as Y145del, F490S, A831S and double amino acid changes including D614G+A879S, D614G+A879T, and D614G+M1237I were reported to be resistant to convalescent sera or these mutations could confer the S protein monoclonal antibody resistance, whereas V367F of the RBD was reported to have increased sensitivity to neutralizing antibodies.[13] Other mutations M153I, S254F, and S255F identified in the study are found to reduce the affinity between S-protein and antibodies.[14]

Mutations affecting glycosylation patterns

Analysis of both N- linked (NGS) and O-linked glycosylation sites (OGS) was performed for S-protein of 4326 isolates. Among the total 22 and 26 amino acid sites of S-protein carrying with NGS and OGS moieties respectively, it was observed that 7 and 9 of these sites were found to possess mutations that resulted in loss of glycosylation moiety [Figure 5]. All except one variants possessed only one amino acid glycosylation site change either NGS or OGS. One variant (EPI_ISL_479737) had lost both OGS and NGS sites due to mutations such as T602L and N603Y [Table 2]. There were two NGS present in RBD without any mutation; among the four OGS in RBD, only one glycosylation mutation (T323I) was observed.

Functional evaluation of the S protein mutations

Among the total 557 amino acid mutations of S-protein, 531 mutations were taken up for score prediction studies and the remaining 26 mutations observed either as stop codons (STOP) or deletions (del). SIFT score predicted 124 mutations to be deleterious and other mutations to be neutral. Also, PROVEAN score predicted 63 mutations to be deleterious whereas POLYPHEN-2 predicted 213 mutations that could display probably damaging effect. Only 41 amino acid mutations were predicted to result in potentially deleterious functional consequences by all three of the mutation score prediction tools [Supplementary File 4].

Phylogenetic analysis of spike protein

Only 250 S-protein sequences constituting unique mutations were selected for phylogenetic tree construction, and the analysis showed that there was high degree of heterogeneity with multiple clusters and sequences were highly diverged from the reference sequence [Figure 6]. Estimates of Average Evolutionary Divergence of sequence pairs comprising 4312 S-genes showed the evolutionary rate as 5.4 x 10^{-4} substitution/site/year (s/s/y).

DISCUSSION

Continuous monitoring of the virus locally and globally is needed for devising effective measures to handle the pandemic crisis. In this study, we report the molecular epidemiological features of SARS-CoV-2 based on WGS in GISAID deposited from India including the dynamics of clade distribution and diversity, amino acid mutations in S-protein and their impact on virulence, immune evading responses and glycosylation patterns.

The study showed that the GR was predominant and was followed by clades G, O, GH, S, L and V. Though there were only four SARS-CoV-2 clades such as ‘L’, ‘S’, ‘G’, and ‘V’ during the early pandemic phase, swift genetic diversity of the virus and its rapid pace of evolution facilitated GISAID to continuously update the classification by inclusion of three more clades such as GH’, ‘GR’ and ‘GV’. Besides, all unclassified sequences of SARS-CoV-2 strains are grouped as ‘O. It is observed that there are only few studies on phylogenetic analyses of SARS-CoV-2 from India. A recent study from India reported that the major cluster of SARS-CoV-2 was A2a (PANGOLIN lineage B.1/B.1.1/B.1.36) (83%) followed by a distinct A3i clade (PANGOLIN lineage B.6) (11.6%).[5,15] Another phylogenetic study on Indian SARS-CoV-2 revealed the presence of four major clades, *i.e.*, 19A (n = 18.4%), 19B (n = 17%), 20A (n = 34.43%), 20B (n = 28.3%), and one minor clade 20C (n = 1.9%).[16] These reports suggested that Europe and Southeast Asia as two major routes for introduction of the virus in India followed by local transmission. Both the predominant G and GR are European clades and the strains of these clades possess the D614G mutation on the S-protein which is more infectious.[16,17]

The month-wise clade distribution analysis showed that L, S and O clades were prevalent in the country during the early phase of pandemic; subsequently, G, GR, and GH clades became prevalent over them. The prevailing clades in the country could be attributed to the early invasion of strains into India through travelers and subsequent mixing of clades. Few reports with minimal sequences deposited till July 2020 only revealed the presence of few clades such as A2a, A3, B and O in India and among them A2a (related to GISAID clade G) was predominant following A3, O and B.[5,17,18] The present study observes ever-changing genetic diversity with intense clade dynamicity of the virus throughout the year.

Substitution mutations in S protein of all the Indian SARS-CoV-2 sequences were analysed with reference to SARS-CoV-2 Wuhan-Hu-1 strain. The origin of D614G mutation was in China during January, 2020 but the occurrence in India was reported in March and became prevalent afterwards. The first occurrence of L54F was observed in Wuhan in March whereas India reported in April. The protease cleavage region S1/S2 in the S protein is essential for the virus to undergo proteolytic activation of S1 and S2 domains for receptor binding and viral-membrane fusion. The region is highly conserved at sites 685 and 686 where proteolytic cleavage occurs. The study has identified 11 mutations flanking the proteolytic cleavage site. Inferences from the proteolytic cleavage of the S glycoprotein suggest the capability of virus to possess features such as cross species mobility or tropism towards different cells.[19] There are 166 mutation sites observed in Asia with 181 mutation types.[20] However, the present study observes that there are 419 mutation sites in the protein with 557 mutation types meaning that several sites in the protein carry more than one mutation type.

Though D614G is associated with increased infectivity, mutations such as Q239R, T719I, T719S, D839Y, P1263L, mutations in RBD such as I434K and P521S, and D614G+Q675H are reported to have decreased infectivity.[13] Besides, D614G in combination with other mutations such as D614G+L5F (n = 23), D614G+V341I (n = 1), D614G+D936Y (n = 3), D614G+S939F (n = 9) and D614G+S943T (n = 2) in strains of the present study was demonstrated to have increased infectivity compared to Wuhan-1 strain.[13] A recent study has reported that L54F, D614G and V1176F of S-protein, identified in the study, are correlated with severe clinical outcome.[21] It was reported that mutations such as T29I, H49Y, D138Y, E484Q, E484K, A520S, T572I, D614G and H1083Q identified in strains of the study, could increase the stability of S-protein.[6] In contrast, the report suggested that mutations such as L54F, G431S, E471D, G502R, Q506H, P507S, Y508N, E583D and Q675H could weaken the interaction of S-protein with ACE2 receptor; whereas, N440K, E471Q and G504V could improve the binding affinity. Emergence of strains of Variant of Concern (VOC), according to WHO nomenclature, such as Alpha (GISAID clade: GRY), Beta (GH/501Y.V2), Gamma (GR/501Y.V3) and Delta (G/452.V3) as well as Variant of Interest (VOI) such as Eta (G/484K.V3), Iota (GH/253G.V1) and Kappa (G/452R.V3) has been observed during the end of year 2020 and early 2021 worldwide. Though few of these highly transmissible variants identified in India late 2020, the sequences of them were submitted to GISAID only in 2021 except two VOC Alpha strains (EPI_ISL_745197 and EPI_ISL_747244). Hence, the study does not report mutations and their features for these variants including the Delta variant that are likely responsible for the substantial surge in cases that began in the Western state of Maharashtra and spread throughout India from Jan, 2021 onwards.[22] This study observed that only 2 WGS of VOC strain (Alpha) from India were available in GISAID in the year 2020 and were taken for analysis.

Antibodies targeting the RBD are being developed as prophylactics. Determination of mutations in S-protein showing resistance to antibodies is crucial for assessing the antigenic implications of viral evolution. The study has identified immune escape mutations both in NTD and RBD of Indian isolates. Mutations especially in these domains evading the antibody recognition could result in the severity of infection. Presently, most of the SARS-CoV-2 genome is not under positive selection, but if neutralizing antibodies are widely deployed as prophylactics, positive selection pressure that lead to infection-competent viral mutants resulting in resurgence of SARS-CoV-2 infections and pose challenges to prophylactic measures.[11] Virulence of SARS-CoV-2 can be associated with mutations in S-protein such as L18F, H69del, V70del, D138Y and Y144del that confer affinity and adhesive properties for better interaction with host cells through surface electrostatic interaction;[9] besides, these mutations are also reported to evade host immune responses against S-protein.[23] Though the present study particularly focuses on functional features of mutations in S-protein, epistatic interactions involving mutations from other genes can also play a role in clade diversity and spatio-temporal dynamics. Such interactions favor the coevolution of mutations due to selective pressures to form new clades that become dominant. The fitness of mutations in virulence and immune escape features are largely influenced not only by independent mutations in S-protein but also mutations through epistatic interactions. For instance, D614G appears along with 3 other mutations in 5'UTR, NSP3 and NSP12 that form G clade.[24] VOC strains forming distinct clades have virulence features contributed by mutations in S gene and other genes.

Glycosylation of S protein plays a vital role in virulence, S-protein folding, immune sensitivity as well as host immune evasion, and shaping viral tropism.[25] Analysis of both NGS and OGS of the study isolates showed mutations that resulted in loss of glycosylation moiety suggesting the reduced immunogenic potential of S-protein of mutant variants.[26] However, there is no report on the impact of the NGS mutation in the interaction of RBD with ACE receptor. S-protein of SARS-CoV-2 has 22 NGS and several OGS; but, in many strains of this study, several of these sites were lost due to amino acid mutations. There are studies that report absence of mutations at NGS in S-protein. It has been studied that certain mutations incurred in the NGS and OGS increase the stability of the S-protein.[6] Accordingly, in the present study, the observed mutations in the NGS such as N234Y and N603Y and OGS mutations such as S221L, T323I, T602I and T602L are found to stabilize the S-protein. On the contrary, very few mutations at the NGS (N709K) and OGS (T1077I) were found to decrease the stability of S protein.[6] Also, glycosylation mutations such as N149G, N165S, and N709K are reported to increase the sensitivity to neutralizing antibodies and the mutation N234Y is found to reduce the neutralization sensitivity to different set of antibodies. The glycosylation mutation N1074D has been found to decrease the infectivity.[13]

Functional evaluation of 531 mutations in S-protein from Indian isolates reveals 41 amino acid mutations that are predicted to have potential impact on functional consequences. A previous study on Indian SARS-CoV-2 isolates reported scores for D614G mutation in S-protein and several mutations across various proteins with their functional impact.[5] However, the present study reports scores for all mutations occurred in S-protein of Indian isolates that were predicted to be neutral, tolerated, deleterious, benign and probably damaging by means of using mutation score prediction tools. The evolutionary rate of S-gene was estimated to be 5.4 x 10^{-4} substitution/site/year (s/s/y) through analysis of 4312 S-genes. Reports suggest that the genome have the evolutionary rate varying in the range between 1.854 x 10^{-4} and 5.63 × 10^{-3}s/s/y.[27-29] A study reported that the evolutionary rate for S-gene of SARS-CoV-2 was 1.08 x 10^{-3}s/s/y after nine months of pandemic.[30] Another study on Indian SARS-CoV-2 isolates reported the evolutionary rate for S-protein as 3.55 x 10^{-3}s/s/y employing sequences of 1376 isolates.[5]

CONCLUSIONS

The study has revealed a rapidly shifting of clade predominance and uneven distribution within a year of the introduction of SARS-CoV-2 in India. The evaluation of S protein reveals the significance of various mutations in virulence, immune escape features and disease severity besides their impact on therapeutics and prophylaxis.

Acknowledgements

The authors thank Department of Health Research (DHR), Govt of India for supporting State VRDL, King Institute of Preventive Medicine and Research, Chennai in this study. We acknowledge the GISAID EpiCov database and the contributors of genomic data for enabling the sequences available for the study.

Conflicts of interest

None

Ethical approval

None.

Research Quality and Ethics Statement

The authors of this manuscript declare that this scientific work complies with reporting quality, formatting and reproducibility guidelines set forth by the EQUATOR Network. The authors also attest that this clinical investigation was determined to not require the Institutional Review Board / Ethics Committee review, and the corresponding protocol/approval number is not applicable. We also certify that we have not plagiarized the contents in this submission and have done a Plagiarism Check. We also certify that none of the authors is a member of the Editorial board of the Journal of Global Infectious Diseases.

REFERENCES

1. Sun J, He WT, Wang L, Lai A, Ji X, Zhai X, *et al*. COVID-19: Epidemiology, Evolution, and Cross-disciplinary perspectives. Trends Mol Med 2020;26(5):483-95.
2. Elbe S, Buckland□Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall 2017;1(1):33-46.
3. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, *et al*. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34(23):4121-3.
4. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol2018;35(6):1547-1549.
5. Banu S, Jolly B, Mukherjee P, Singh P, Khan S, Zaveri L, *et al*. A distinct phylogenetic cluster of Indian SARS-CoV-2 isolates. Open Forum Infect Dis 2020;7(11):ofaa434.
6. Teng S, Sobitan A, Rhoades R, Liu D, Tang Q. Systemic effects of missense mutations on SARS-CoV-2 spike glycoprotein stability and receptor-binding affinity. Brief Bioinform 2021;22(2):1239-53
7. Rophina M, Pandhare K, Shamnath A, Imran M, Jolly B, Scaria V. ESC - a comprehensive resource for SARS-CoV-2 immune escape variants. bioRxiv 2021.
8. Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-Stroh S, *et al*. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Euro Surveill 2020;25(32):2001410.
9. Fantini J, Yahi N, Azzaz F, Chahinian H. Structural dynamics of SARS-CoV-2 variants: A health monitoring strategy for anticipating Covid-19 outbreaks. J Infect 2021;83(2):197-206.

10. Van Egeren D, Novokhodko A, Stoddard M, Tran U, Zetter B, Rogers M, *et al*. Risk of rapid evolutionary escape from biomedical interventions targeting SARS-CoV-2 spike protein. PLoS One 2021;16(4):e0250780.

11. Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, Loes AN, *et al*. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. Cell Host Microbe 2021;29(1):44-57.e9.

12. Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JC, *et al*. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. Elife 2020;9:e61312.

13. Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, *et al*. The impact of mutations in SARS-CoV-2 Spike on viral infectivity and antigenicity. Cell 2020;182(5):1284-94.e9.

14. Chen J, Gao K, Wang R, Wei GW. Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies. Chem Sci. 2021 Apr 13;12(20):6929-6948.

15. Jacob JJ, Vasudevan K, Veeraraghavan B, Iyadurai R, Gunasekaran K. Genomic evolution of severe acute respiratory syndrome Coronavirus 2 in India and vaccine impact. Ind J Med Microbiol 2020;38(2):210-2.

16. Raghav S, Ghosh A, Turuk J, Kumar S, Jha A, Madhulika S, *et al*. Analysis of Indian SARS-CoV-2 genomes reveals prevalence of D614G mutation in Spike protein predicting an increase in interaction with TMPRSS2 and virus infectivity. Front Microbiol 2020;11:594928.

17. Pattabiraman C, Habib F, Rasheed R, Prasad P, Reddy V, Dinesh P, *et al*. Genomic epidemiology reveals multiple introductions and spread of SARS-CoV-2 in the Indian state of Karnataka. PLoS One 2020;15(12):e0243412.

18. Biswas NK, Majumder PP. Analysis of RNA sequences of 3636 SARS-CoV-2 collected from 55 countries reveals selective sweep of one virus type. Indian J Med Res 2020;151(5):450-8.

19. Menachery VD, Dinnon KH, Yount BL, McAnarney ET, Gralinski LE, Hale A, *et al*. Trypsin Treatment Unlocks Barrier for Zoonotic Bat Coronavirus Infection. J Virol 2020;94(5):e01774-19.

20. Guruprasad L. Human SARS CoV-2 spike protein mutations. Proteins. 2021;89(5):569-76.

21. Nagy A, Pongor S, Gyorffy B. Different mutations in SARS-CoV-2 associate with severe and mild outcome. Int J Antimicrob Agents 2021;57(2):106272.

22. Chatterjee P. Covid-19: India authorises Sputnik V vaccine as cases soar to more than 180 000 a day. BMJ 2021;373:n978.

23. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, *et al*. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol. 2021;19(7):409-424.

24. Banoun H. Evolution of SARS-CoV-2: Review of Mutations, Role of the Host Immune System. Nephron 2021;145(4):392-403.

25. Watanabe Y, Berndsen ZT, Raghwani J, Seabright GE, Allen JD, Pybus OG, *et al*. Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. Nat Commun 2020;11(1):2688.

26. Sanda M, Morrison L, Goldman R. N- and O-Glycosylation of the SARS-CoV-2 Spike protein. Anal Chem 2021;93(4):2003-9.

27. vanDorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, *et al*. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 2020;83:104351.

28. Velazquez-Salinas L, Zarate S, Eberl S, Gladue DP, Novella I, Borca MV. Positive selection of ORF1ab, ORF3a, and ORF8 genes drives the early evolutionary trends of SARS-CoV-2 during the 2020 COVID-19 pandemic. Front Microbiol 2020;11:550674.

29. Motayo BO, Oluwasemowo OO, Olusola BA, Akinduti PA, Arege OT, Obafemi YD, *et al*. Evolution and genetic diversity of SARS-CoV-2 in Africa using whole genome sequences. Int J Infect Dis 2021;103:282-7.

30. Pereson MJ, Flichman DM, Martínez AP, Baré P, Garcia GH, Di Lello FA. Evolutionary analysis of SARS-CoV-2 spike protein for its different clades. J Med Virol 2021;93(5):3000-6.

Legends for Figures

Figure 1. SARS-CoV-2 clade distribution pattern in India. **(a)** Pie chart showing the proportion of various clades of the genomes deposited from India in GISAID; **(b)** Schematic geographical map showing the proportion and distribution of clades from different states of India; **(c)** Month wise clade distribution during the year 2020.

Figure 2. Phylogenetic tree showing clade diversity for SARS-CoV-2 Indian isolates. These isolates fall under 7 genetic clades with the majority falling under GR clade.

Figure 3. Amino acid mutations and their frequency in different regions of S proteins of SARS-CoV-2 isolates from India. Signal peptide (SP), N-terminal domain (NTD), Receptor binding domain (RBD), Protease cleavage site (PC), Fusion peptide (FP), Heptad repeat 1(HR1), Heptad repeat 2 (HR2), Transmembrane domain (TM), Cytoplasm domain (CT).

Figure 4. Distribution and frequency of the most prevalent mutations of S protein of SARS-CoV-2 isolates circulated in India during the year 2020. D614G is predominant throughout the year with high frequency followed by L54F mutation. D614G, Q677H and P681H mutations originated during the first half of the year and their appearance was observed throughout the year; L54F, K77M and P812L mutations emerged during the first half of the year but absent after few months of their appearance.

Figure 5. Frequency of amino acid mutations impacting O- and N-glycosylation patterns. Few sites such as S221, T602 and N1074 are having more than one mutation.

Figure 6. Phylogenetic tree of isolates having distinct mutations in the gene of S protein. The tree was constructed by Maximum-Likelihood method with the tree having the root as SARS-CoV-2 Wuhan-Hu-1 sequence (NC_045512.2).

Legends for Tables

Table 1. Aminoacid substitution mutations observed across various regions of S proteins of Indian SARS-CoV-2 isolates.

Table 2. N- and O-linked glycosylation sites of S protein of SARS-CoV-2 and amino acid mutations at these sites affecting the glycosylation pattern in Indian SARS-CoV-2 variants.

Figure 1

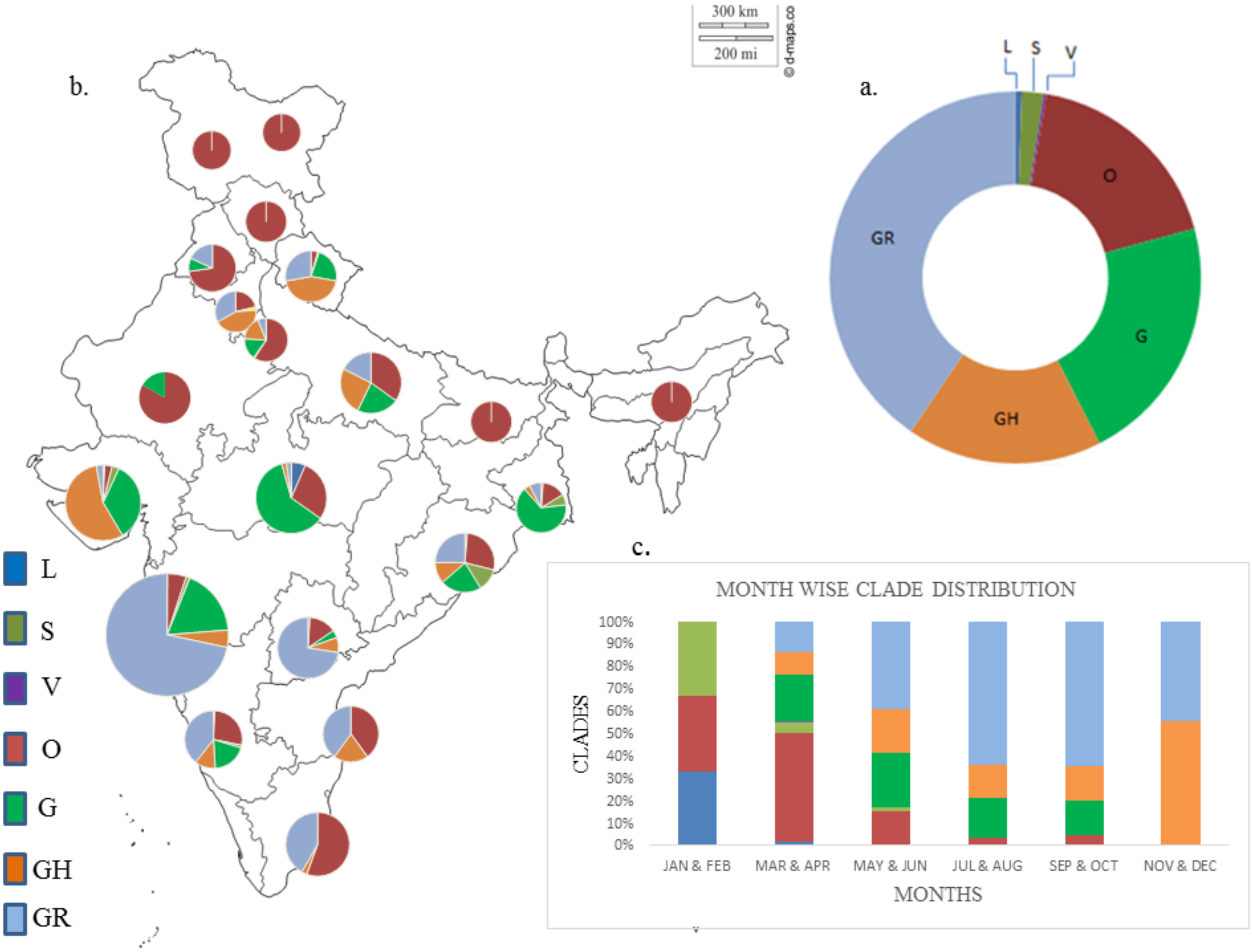

b.
300 km
200 mi
© d-maps.co
a.
L
S
V
O
G
GH
GR
L
S
V
O
G
GH
GR
c.
MONTH WISE CLADE DISTRIBUTION
CLADES
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
JAN & FEB
MAR & APR
MAY & JUN
JUL & AUG
SEP & OCT
NOV & DEC
MONTHS

Figure 2

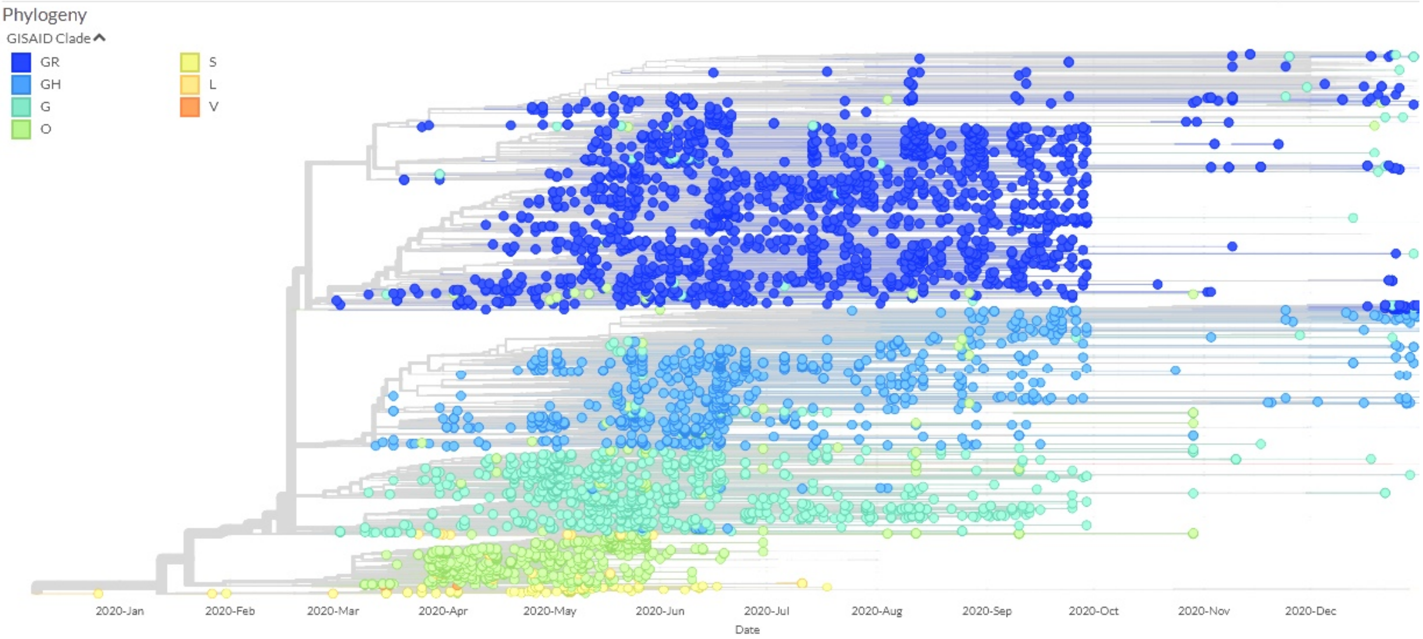

Phylogeny
GISAID Clade
GR
GH
G
O
S
L
V
2020-Jan
2020-Feb
2020-Mar
2020-Apr
2020-May
2020-Jun
2020-Jul
2020-Aug
2020-Sep
2020-Oct
2020-Nov
2020-Dec
Date

Figure 3

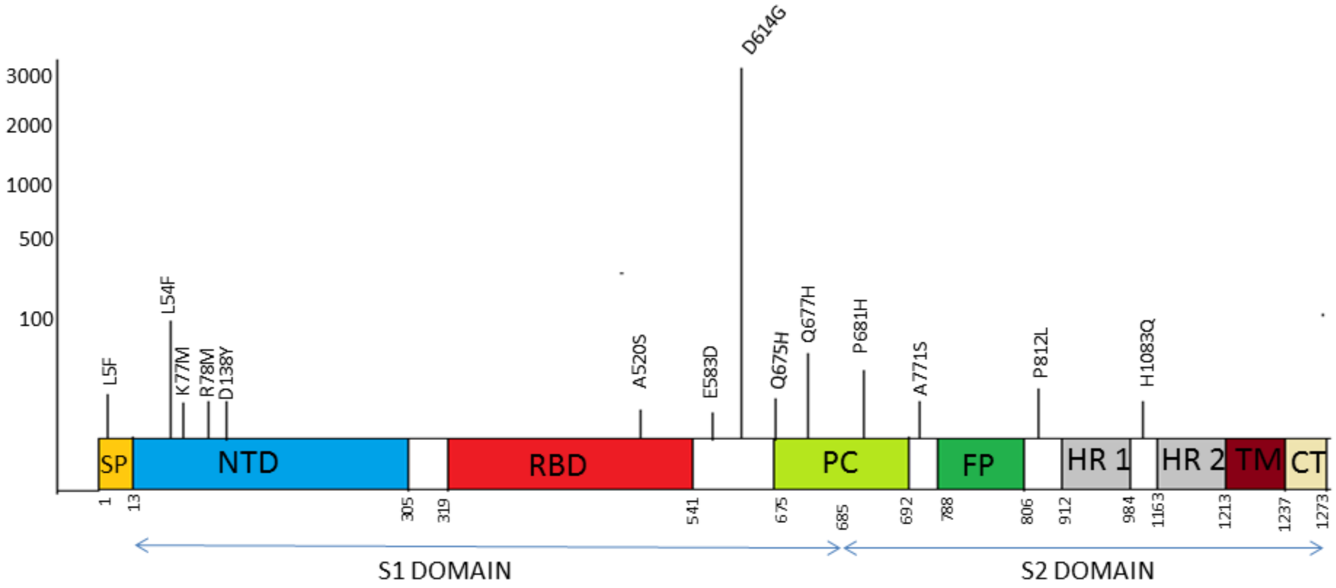
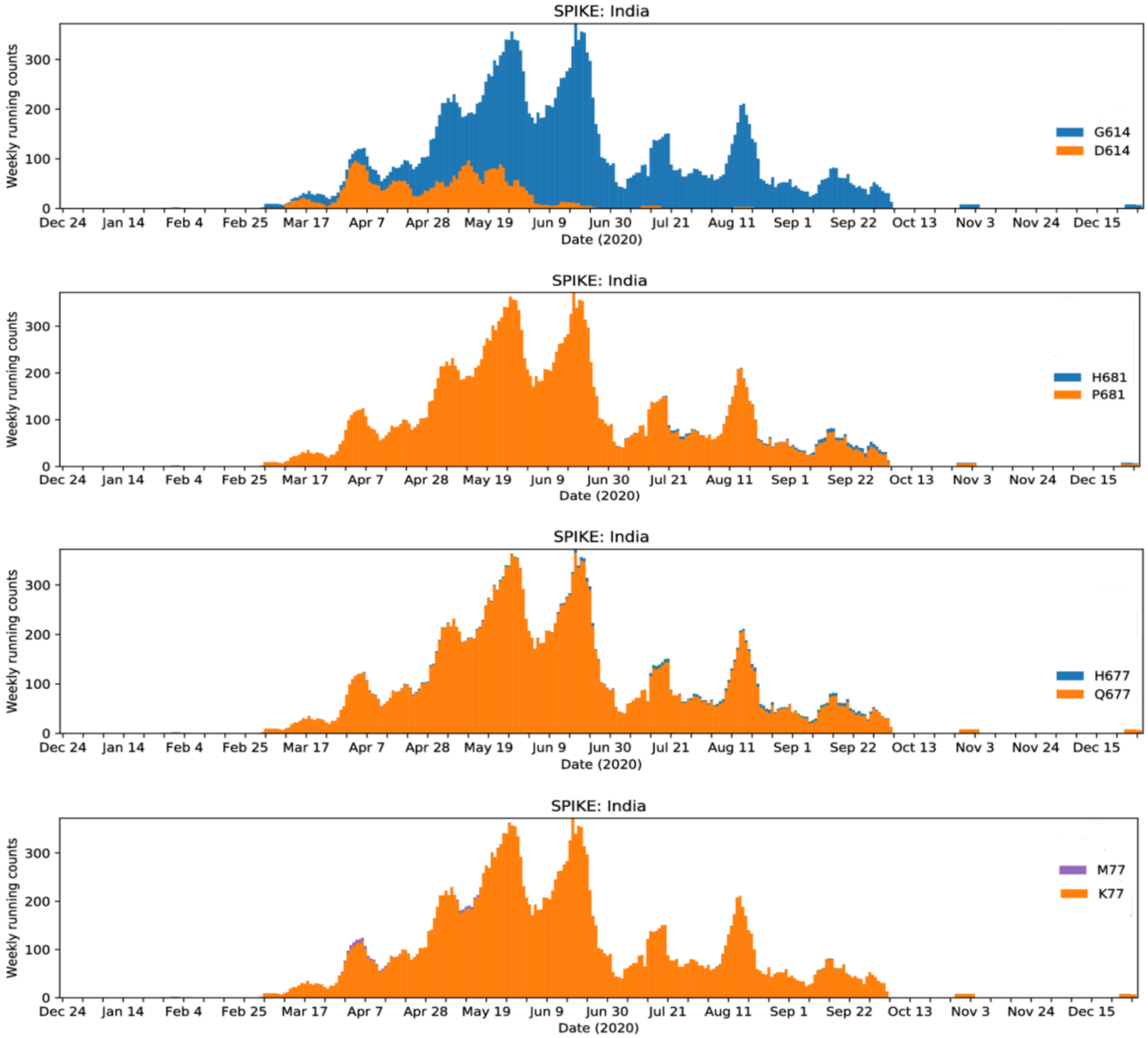

3000
2000
1000
500
100
D614G
L54F
L5F
K77M
R78M
D138Y
A520S
E583D
Q675H
Q677H
P681H
A771S
P812L
H1083Q
SP
NTD
RBD
PC
FP
HR 1
HR 2
TM
CT
1
13
305
319
541
675
685
692
788
806
912
984
1163
1213
1237
1273
S1 DOMAIN
S2 DOMAIN

Figure 4


SPIKE: India
Weekly running counts
0
100
200
300
Dec 24
Jan 14
Feb 4
Feb 25
Mar 17
Apr 7
Apr 28
May 19
Jun 9
Jun 30
Jul 21
Aug 11
Sep 1
Sep 22
Oct 13
Nov 3
Nov 24
Dec 15
Date (2020)
G614
D614
SPIKE: India
Weekly running counts
H681
P681
Date (2020)
SPIKE: India
Weekly running counts
H677
Q677
Date (2020)
SPIKE: India
Weekly running counts
M77
K77
Date (2020)

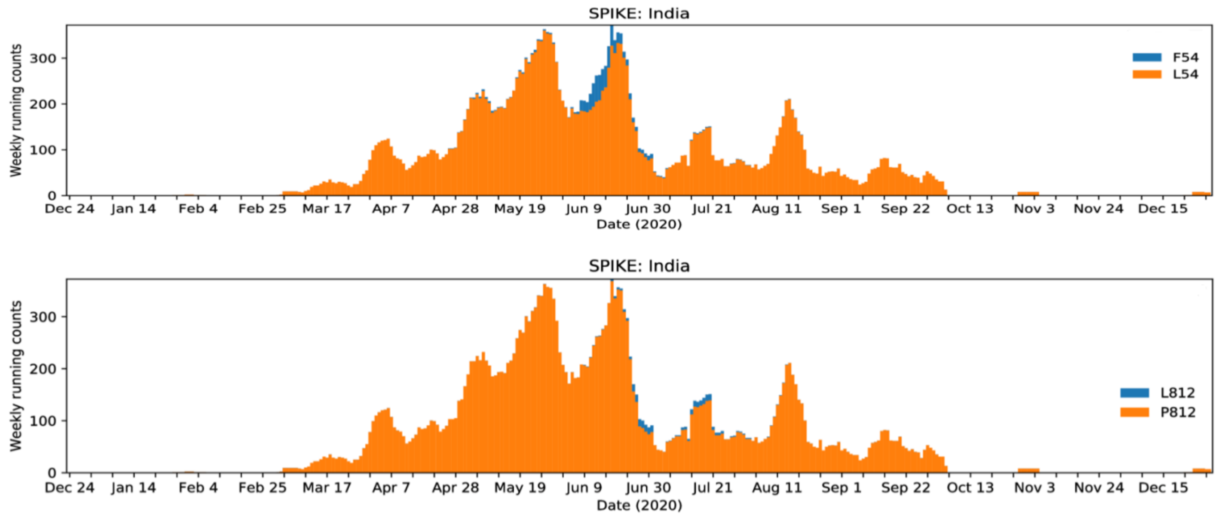
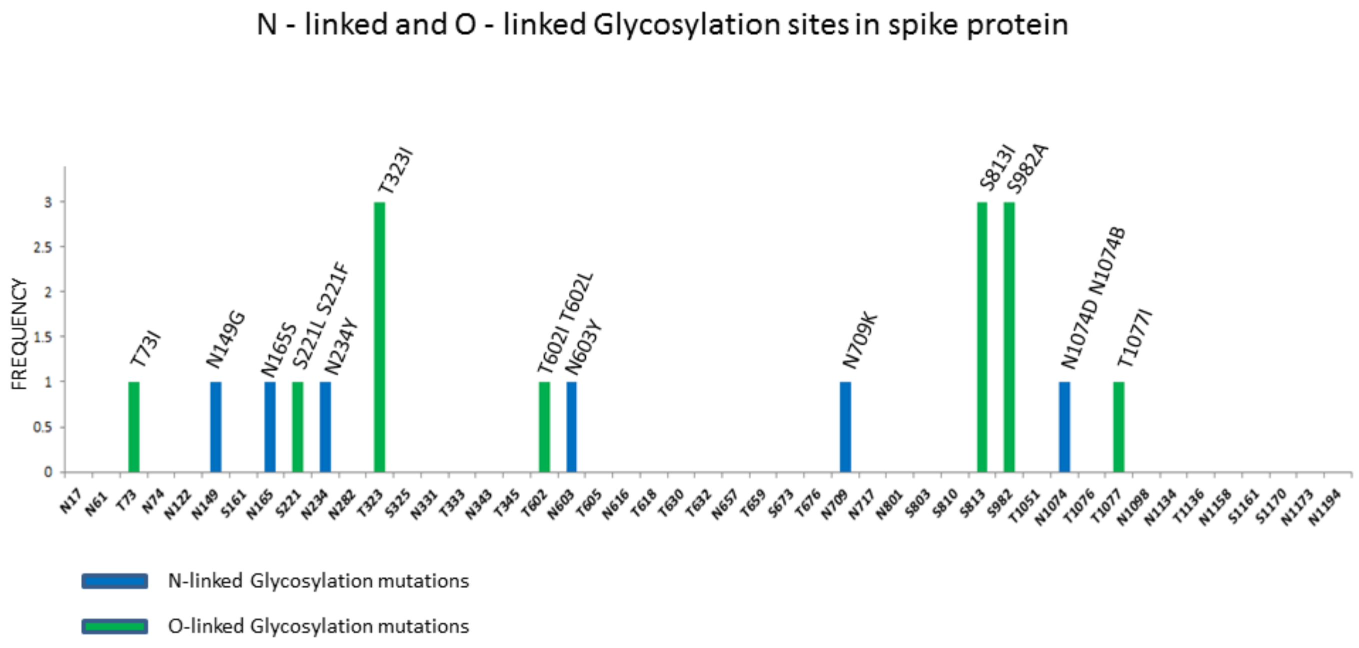

SPIKE: India
Weekly running counts
0
100
200
300
F54
L54
Dec 24
Jan 14
Feb 4
Feb 25
Mar 17
Apr 7
Apr 28
May 19
Jun 9
Jun 30
Jul 21
Aug 11
Sep 1
Sep 22
Oct 13
Nov 3
Nov 24
Dec 15
Date (2020)
SPIKE: India
Weekly running counts
0
100
200
300
L812
P812
Dec 24
Jan 14
Feb 4
Feb 25
Mar 17
Apr 7
Apr 28
May 19
Jun 9
Jun 30
Jul 21
Aug 11
Sep 1
Sep 22
Oct 13
Nov 3
Nov 24
Dec 15
Date (2020)

Figure 5


N - linked and O - linked Glycosylation sites in spike protein
FREQUENCY
0
0.5
1
1.5
2
2.5
3
T73I
N149G
N165S
S221L S221F
N234Y
T323I
T602I T602L
N603Y
N709K
S813I
S982A
N1074D N1074B
T1077I
N17
N61
T73
N74
N122
N149
S161
N165
S221
N234
N282
T323
S325
N331
T333
N343
T345
T602
N603
T605
N616
T618
T630
T632
N657
T659
S673
T676
N709
N717
N801
S803
S810
S813
S982
T1051
N1074
T1076
T1077
N1098
N1134
T1136
N1158
S1161
S1170
N1173
N1194
N-linked Glycosylation mutations
O-linked Glycosylation mutations

.

Figure 6

Region	Position	Number of mutation sites	Number of mutations

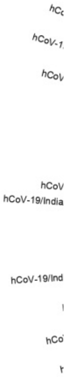


Table 1

Signal Peptide	1 - 13	7	9
N-Terminal Domain	14 - 305	144	211
Receptor Binding Domain	319 - 541	53	63
Protease Cleavage Site	675 - 692	8	11
Fusion Peptide	788 - 806	6	6
HR1	912 - 984	24	31
HR2	1163 - 1213	15	18
Transmembrane Domain	1214 - 1237	13	16
Cytoplasm Domain	1238 - 1273	12	13

Table 2

N-linked	Mutation	O-linked	Mutation

Glycosylation Site (NGS)		**Glycosylation Site (OGS)**	
N17	-	T73	T73I
N61	-	S161	-
N74	-	S221	S221, S221F
N122	-	T323	T323I
N149	N149G	S325	-
N165	N165S	T333	-
N234	N234Y	T345	-
N282	-	T602	T602I, T602L
N331	-	T605	-
N343	-	T618	-
N603	N603Y	T630	-
N616	-	T632	-
N657	-	T659	-
N709	N709K	S673	-
N717	-	T676	-
N801	-	S803	-
N1074	N1074D, N1074B	S810	-
N1098	-	S813	S813I
N1134	-	S982	S982A
N1158	-	T1051	-
N1173	-	T1076	-
N1194	-	T1077	T1077I
		T1136	-
		S1161	-
		S1170	-
		S1175	-