# Structural and Functional Implications of Non-synonymous Mutations in the Spike protein of 2,954 SARS-CoV-2 Genomes

**Shijulal Nelson-Sathi*, Umasankar PK*, E Sreekumar, R Radhakrishnan Nair, Iype Joseph, Sai Ravi Chandra Nori, Jamiema Sara Philip, Roshny Prasad, Navyasree KV, Shikha Ramesh, Heera Pillai, Sanu Ghosh, Santosh Kumar TR and M. Radhakrishna Pillai**

**Corona Research & Intervention Group, Rajiv Gandhi Centre for Biotechnology, Thiruvananthapuram, India**

**Corresponding Authors*: *shijulalns@rgcb.res.in*, *umasankarpk@rgcb.res.in*

## Abstract

SARS-CoV-2, the causative agent of COVID-2019 pandemic is an RNA virus prone to mutations. Information on mutations within the circulating strains of the virus is pivotal to understand disease spread and dynamics. Here, we analyse the mutations associated with 2,954 globally reported high quality genomes of SARS-CoV-2 with special emphasis on genomes of viral strains from India. Molecular phylogenetic analysis suggests that SARS-CoV-2 strains circulating in India form five distinct phyletic clades designated R1-R5. These clades categorize into the previously reported S, G as well as a new unclassified subtype. A detailed analysis of gene encoding the Spike (S) protein in the strains across the globe shows non-synonymous mutations on 54 amino acid residues. Among these, we pinpoint 4 novel mutations in the region that interacts with human ACE2 receptor (RBD). Further *in silico* molecular docking analyses suggest that these RBD mutations could alter the binding affinity of S-protein with ACE2 that may lead to changes in SARS-CoV-2 infectivity. Strikingly, one of these RBD mutations (S438F) is unique to a subset within the R4 clade suggesting intrinsic S-protein variations in strains currently circulating in India. Together, our findings reveal a unique pattern of SARS-CoV-2 evolution that may alert vaccine and therapeutic development.

## Keywords

SARS-CoV-2, Spike protein, Receptor binding domain, mutation, virus evolution

## Introduction

COVID-19, the highly transmittable and pathogenic viral infection caused by SARS-CoV-2, belongs to betacoronavirus genus that is known to cause acute respiratory distress syndrome, coagulation dysfunction and septic shock (Yang, P. and Wang, X. 2020). The rate of COVID-19 spread appears to be more than SARS-CoV and MERS-CoV (Chen, J. 2020) with 2,654,209 confirmed cases and 185,062 death cases reported in 210 countries and territories around the world (https://www.worldometers.info/coronavirus). This increased transmission was recently correlated with the high mutation frequency of SARS-CoV-2 strains (Yin, C. 2020).

The single stranded RNA genome of COVID-19 has 2,9891 nucleotides, possesses 14 ORFs encoding 29 proteins which include four structural proteins: Envelope (E), Membrane (M), Nucleocapsid (N) and Spike (S) proteins, 9 accessory proteins and 16 non-structural proteins (Wu, A. et al. 2020, Gordon et al. 2020). A recent study shows that mutations in the S-protein that mediates viral entry can modulate viral pathogenesis (Shang et al., 2020). The S-protein is functionally bipartite with a proximal S1 segment that binds the host angiotensin-converting enzyme2 (ACE2) receptor (Wang et al., 2020) and a distal S2 region helps in virus-host cell fusion separated by S1-S2 linker containing protease cleavage sites. S1 subunit consists of a signal peptide (SP) followed by the N-terminal domain of unknown function and the C-terminal domain consisting of receptor binding domain (RBD) and a receptor binding motif (RBM) within. S2 subunit comprises a fusion peptide (FP), heptad repeats (HR1&HR2), transmembrane domain (TM) and a short cytoplasmic domain (CP). This viral entry process requires the S protein priming, which is facilitated by TMPRSS2, a serine protease that is produced by the host cell (Hoffmann et al., 2020). Together, these features make S-protein the primary target for the development of antibodies, entry inhibitors and vaccines (Du, L. et al.,2009; Wang, Q. et al.,2016). Also, SNPs identified on S-gene may have an important role in its host range and pathogenicity (Yin, C. 2020). Furthermore, identifying a complete set of variations in S-protein of SARS-CoV-2 and their impact on human ACE2 affinity is much needed for the development of therapeutic countermeasures. Thus, studying S- protein and their evolution can enhance our understanding of host receptor affinity variations and virulence levels.

To study the evolutionary pattern of SARS-CoV-2 strains circulating globally and in Indian subcontinent and also to reveal structural and functional implications of amino acid variations present in the S proteins, we analysed 2,954 SARS- CoV-2 complete genomes. Our analysis reveals that there are at least 5 distinct clades of strains circulating in India. In addition, the global strains appear to have acquired 54 mutations that are non-synonymous in the gene encoding S-protein among which 4 important mutations are on the RBD that may have direct implications on the infectivity of SARS-CoV-2. Most importantly, we identify that one of these RBD mutations is unique to a set of strains currently circulating in India.

## Materials and Methods

### Genome analysis

Complete and high coverage genomes were downloaded on April 6th, 2020 from the GISAID database. This comprises a total of 3,060 genomes (>29,000 bp) and countries with less than ten genomes were discarded for further analysis. In addition, we also added 30 Indian genomes downloaded on April 15th, 2020 for the analysis. The RefSeq genome (NC_045512) from Wuhan is taken as the reference. The genes were predicted using Prokka (Seemann T; 2014), and the complete sequences of structural proteins such as Spike, Membrane and Nucleocapsid proteins were extracted. The alignments of the structural proteins were done using Mafft (*maxiterate* 1,000 and global pair-*ginsi*) (Katoh et al. 2002). The alignments were visualized in Jalview (Waterhouse AM et al., 2009) and the amino acid substitutions in each position were extracted using custom python script. We ignored the substitutions that are present in only one genome and unidentified amino acid X. The mutations that are present in at least two independent genomes in a particular position were further considered. These two criteria were used to avoid mutations due to sequencing errors. The mutated amino acids were further tabulated and plotted as a matrix using R script. The number of amino acid substitutions observed in S proteins were compared against other structural proteins.

### Phylogeny reconstruction

For the Maximum likelihood phylogeny, we have used 30 Indian Genomes (as on April 15th, 2020) with the worldwide genomes sampled 10 per country and with Wuhan RefSeq strain as root. The phylogeny was reconstructed using IQ-Tree (Nguyen et al., 2015), and the best model “TIM2+F+I” was picked according to the ModelFinder (Kalyaanamoorthy et al., 2017). All the data used for reconstructing the phylogeny was downloaded from the GISAID database. In addition, three random samplings were done to check the consistency of the phylogeny.

### Molecular docking analysis

The structural analysis of the mutated spike glycoprotein of SARS-CoV-2 was done to assess the impact of mutations on binding affinity towards human ACE2 receptor. The structures of wild type and mutated Receptor Binding Domain (RBD) of the spike protein was modeled using the Swiss model (Schwede, T. et al., 2003). The target sequence for spike protein was obtained from National Centre for Biotechnology Information (YP_009724390.1) and the crystal structure of SARS-CoV Spike protein was downloaded from Protein Data Bank (PDB ID: 6ACD) and used as the template (76.83% sequence identity) for homology modeling (Dong, S. et al., 2020). The energy minimization of the modeled structure was done by YASARA server (Krieger, E. et al., 2002), and each modeled structure was superimposed with template structure and calculated the RMSD value in Pymol(Delano, W. L., 2002). HADDOCK 2.4 webserver (De Vries et al., 2010) was used for docking of the wild and mutated spike proteins against Human ACE2 protein (PDB ID: 6LZG_A), and the binding affinity of the docked structures was calculated using PRODIGY webserver (Xue, Li C., et al., 2016).

## Results and Discussion

### Genome variations and phyletic pattern of SARS-CoV-2 strains circulating in India

We compared the whole genome of 2924 strains from 29 countries across the globe and 30 Indian strains of SARS-CoV-2 with original Wuhan strain as reference (Wuhan RefSeq strain: YP_009724390.1), all collected from GISAID database. Maximum likelihood phylogenetic analysis of these strains provides distinct evolutionary features. However, country specific, monophyletic pattern was not observed in the global distribution of SARS-CoV-2 strains (**Figure 1**).

Within this global distribution, we identify five major clades- R1, R2, R3, R4 and R5 for strains circulating in India (**Figure 1**). Among these, R1, R3, R4 and R5 belong to previously reported S and G global subtypes whereas R2 stands out as a previously uncharacterized, unique cluster. In-depth phylogenetic analysis shows peculiar features for Indian clades. Clade R1 consists of only one strain from Kerala that belongs to S- subtype and is consistent with epidemiological travel history from Wuhan, China. In addition, R1 defines 5 unique non-synonymous mutations in the orf1ab (I476V, P2079L, Q5538Z), orf8 (L84S) and S gene (A930V). Clade R2 does not belong to any known S, V or G subtypes and possess two non-synonymous substitutions in orf1ab (V378I, L3606F) and one synonymous mutation on N gene (L139L). Clade R3, R4 and R5 belong to G subtype and show clear epidemiological link to Italy but differ in their mutational profile. While two Indian strains in R3 contain non-synonymous substitutions in the M-gene (D3G) and in the orf1ab (Z6697W), R4 possesses non-synonymous mutation in orf1ab (S1515F) with an additional mutation in the RBD region of S-protein (S438F) in 15% of the strains (**Figure 1**). Clade R5 has two non-synonymous substitutions in the N gene (R203K, G204R). Our analysis is currently limited to only 30 genomes from India; hence require further investigation using more genomes. Nevertheless, these findings reveal that along with global strains unique subtypes of SARS-CoV-2 strains are currently circulating in India.

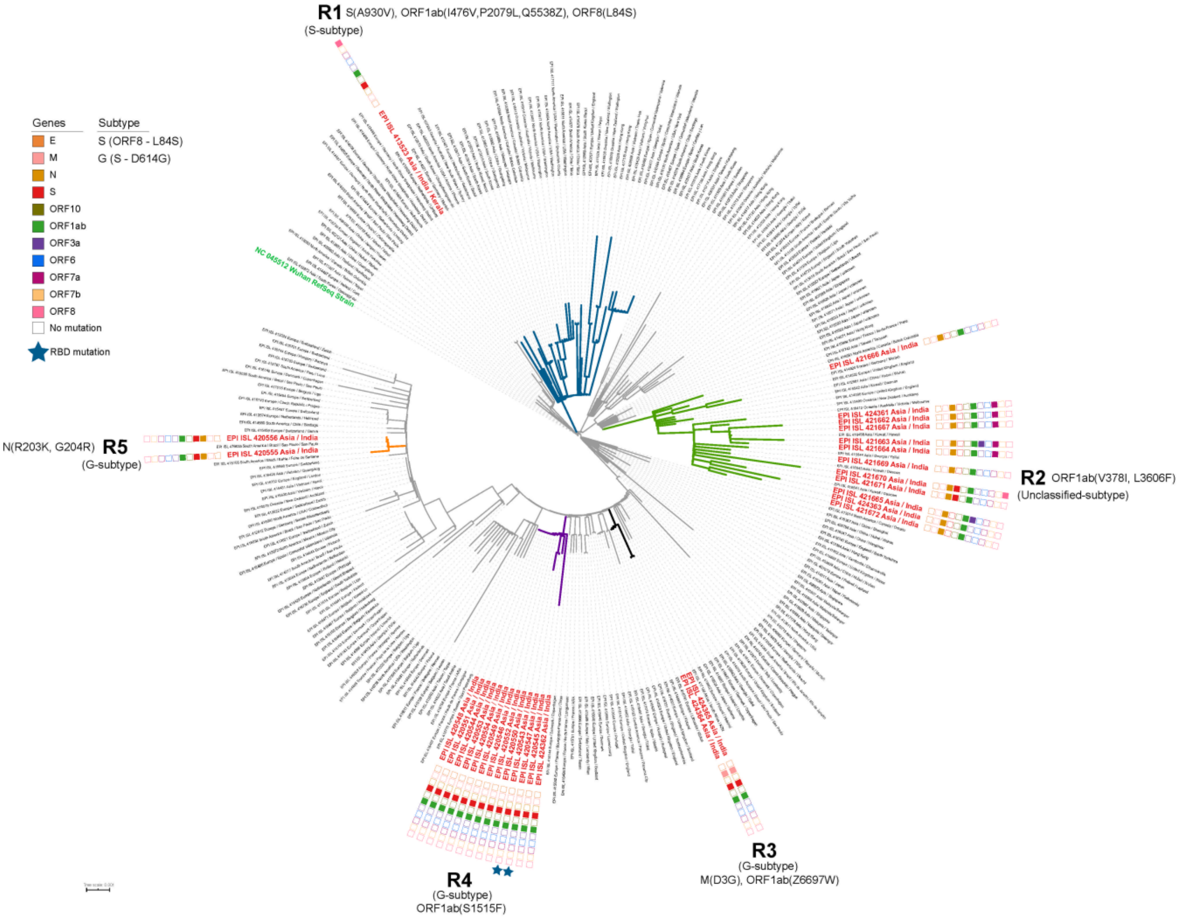


**Figure 1:** Phylogenetic tree of 30 Indian and 226 global SARS-CoV-2 strains sampled from 2954 genomes. Major phyletic clades unique to Indian strains are marked R1-R5 and colour coded with blue, green, black, violet and orange colours respectively. The outgroup Wuhan RefSeq strain is highlighted in green colour. The square boxes in Indian clades represent the presence or absence of synonymous or non-synonymous mutations in structural and non-structural genes. Non-synonymous mutations unique to each clade are given in brackets. Mutations in RBD unique to Indian strains in R4 clade are marked in asterisk.

## Global mutational profile of SARS-CoV-2 S protein.

Since mutations in S-protein have been linked with virus infectivity, we performed a high stringent, in-depth mutational analysis to capture the global variations in the S-protein. Altogether, 54 mutations were identified that belong to 1176 strains from 29 countries. These mutations were found to be distributed across different domains of S-protein (**Figure 2**).

The N-terminally located signal peptide which helps S-protein in ER translocation during its biosynthesis had two mutations L5F and L8V. The N-terminal domain (NTD) of S-protein of certain beta-coronaviruses binds to sugars present on the host cell membrane for attachment. However, the functions of SARS-CoV-2 NTD still need to be determined (Ou, X., et al., 2020) as these sugar binding residues are absent (Li, F., 2015). In our analysis, we identify 20 amino acid substitutions in the NTD region which account for the highest number of mutations within S-protein. Overall, NTD mutations do not appear to perturb the properties of S-protein.

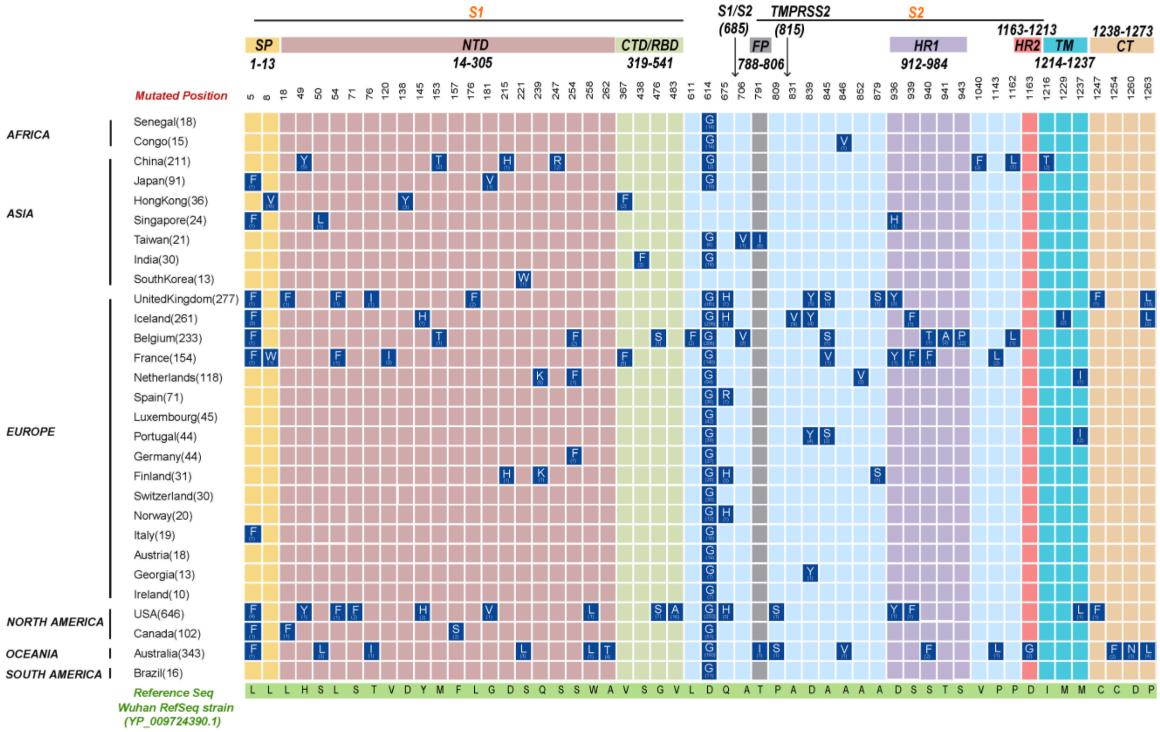
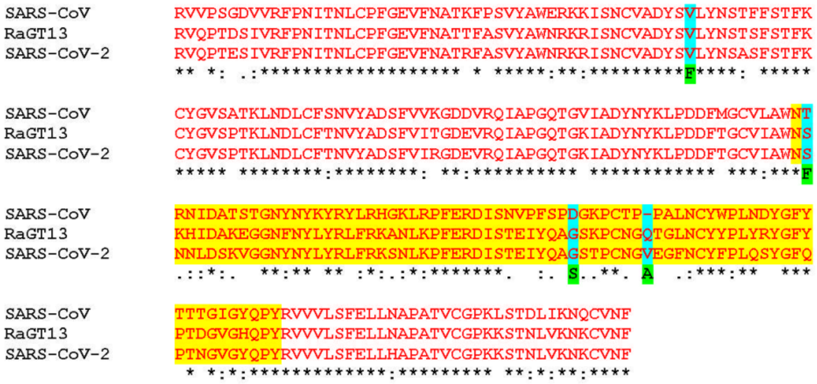


**Figure 2:** Matrix representing amino acid substitutions present in SARS-CoV-2 S protein of 2,954 genomes. Name of countries and the number of genomes sampled are given on the Y-axis and the relevant amino acid residues (single letter code) in the reference strain are given on the X-axis. Mutated amino acid residues and their frequency of mutations are provided in matrix cells. Matrix cells are colour coded based on different domains of S-protein shown at the top. Mutations which are present at least in two independent genomes at the same position are represented in the matrix along with their positions.

RBD comprises of 223 amino acid long peptide in the S-protein is the most variable part of the SARS-like coronavirus genomes (Zhou, P. et al., 2020; Wu, F. et al., 2020, Ortega et al., 2020). In this region, we found mutations in 33 strains from different countries which fall into four amino acid substitutions such as V367F, S438F, G476S, and V483A. V367F mutation is found in strains from Hong Kong and France. A unique S438F mutation is found in the RBD region of only two Indian strains, which is absent in global strains. G476S mutation is seen predominantly in 7 strains from the USA and one from Belgium whereas V483A is found in 16 strains unique to the USA. However, these four RBD mutations together represent only <2% of SARS-CoV-2 strains circulating globally.

Further, to understand the evolutionary significance of these mutations, we compared RBDs from SARS-CoV-2, SARS-CoV (2002) and Bat coronavirus RaTG13 strain, a suspected precursor of SARS-CoV-2 (**Figure 3**). We found that the RBD from SARS-CoV-2 is only 73.4 % identical to SARS-CoV but is 90.1 % identical to RaTG13. Interestingly, none of the amino acid changes observed in SARS-CoV-2 matches with either RaTG13 or SARS-CoV (**Figure 3**). Hence, based on our findings, the possibility of origin of a novel sub-lineage cannot be neglected.



```
SARS-CoV      RVVPSGDVVRFPNITNLCPFGEVFNATKFPSVYAWERKKISNCVADYSVLYNSTFFSTFK
RaGT13        RVQPTDSIVRFPNITNLCPFGEVFNATTFASVYAWNRKRISNCVADYSVLYNSTSFSTFK
SARS-CoV-2    RVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFK
              ** *: .:******************* * *****:**:*********F****: *****

SARS-CoV      CYGVSATKLNDLCFSNVYADSFVVKGDDVRQIAPGQTGVIADYNYKLPDDFMGCVLAWNT
RaGT13        CYGVSPTKLNDLCFTNVYADSFVITGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNS
SARS-CoV-2    CYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNS
              ***** ********:********: **:********** ************ ***:***F

SARS-CoV      RNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCTP-PALNCYWPLNDYGFY
RaGT13        KHIDAKEGGNFNYLYRLFRKANLKPFERDISTEIYQAGSKPCNGQTGLNCYYPLYRYGFY
SARS-CoV-2    NNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQ
              .::*:.  **:** ** :*:.:*:*******.   :.S..**. A .:***:**  ***

SARS-CoV      TTTGIGYQPYRVVVLSFELLNAPATVCGPKLSTDLIKNQCVNF
RaGT13        PTDGVGHQPYRVVVLSFELLNAPATVCGPKKSTNLVKNKCVNF
SARS-CoV-2    PTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNF
               * *:*:*************:********* **:*:**:****
```

**Figure 3:** Conservation of Receptor Binding Domain (RBD) of SARS-COV-2 with its close relatives of SARS-CoV and Bat RATG13. The red colored region shows the Receptor Binding Domain (RBD) and the yellow highlighted region is the Receptor Binding Motif (RBM) in RBD region. The mutated residues are highlighted in light blue color and substitutions are marked below in green.

Interestingly, we found four mutations in the linker region separating S1 and S2 segments of S-protein *viz*., L611F, D614G, Q675H and A706V. Among these, D614G which corresponds to G-subtypes of SARS-CoV-2 was found to be the most abundant mutation globally as well as in Indian subcontinent, accounting for 52.26% (1,544) strains analyzed. This mutation is located outside RBD and near to the furin cleavage site in the S1-S2 junction. The exact functional impact of this mutation remains unclear. Although SARS-CoV-2 possesses a unique furin/ TMPRSS2 cleavage sites, mutations are not observed in this region (**Figure 2**).

Several mutations were found in the S2 region. S2 region consists of proximal fusion peptide (FP) followed by Heptad Repeat-1 ad 2 (HR1 and HR2), transmembrane domain (TM) and short distal cytoplasmic tail (CT) (**Figure 2**). FP along with HR1 and HR2 plays a significant role in the fusion between the virus and target cell membranes (Liu, Shuwen, et al., 2004; Tripet, Brian, et al., 2004). FP is one of the highly conserved regions in S-protein (Cascella, Marco et al., 2020). We found five mutations in the Fusion peptide region- A831V, D839Y, A845S, A846V and A852V. In addition, we identified five mutated positions (D936Y/D936H, S939F, S940F/S940T, T941A, S943P) in the HR1 domain and only one in HR2 (D1163G). The functional implications of these mutations in regulating SARS-CoV-2-host cell fusion need to be determined.

### Structural analysis of RBD mutations

The amino acid substitutions at the RBD can have an influence on the overall binding affinity of the S-protein with ACE2 receptor. Thus, the structural changes and differences in the binding affinity corresponding to each mutation in the RBD (V367F, S438F, G476S and V483A) were analysed in detail (**Figure 4a and b**).

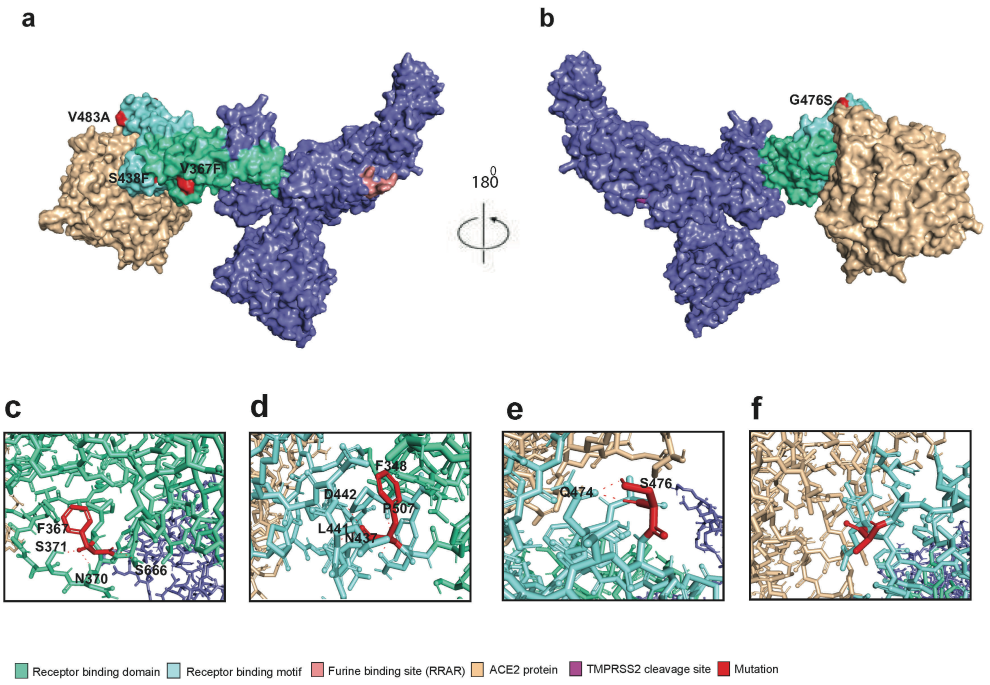


**Figure 4:** Three dimensional structure of modeled SARS CoV-2 S-protein: ACE2 Complex. (a) Surface model of the complex showing RBD mutations (V367F, S438F and V438A) (b) G476S in red color. (c-f) Visualization of possible molecular rearrangements resulting from observed RBD mutations V367F (c), S438F (d), G476S (e) and V483A (f) respectively.

Molecular replacement and docking studies allowed estimation of the binding affinity and dissociation constant of the mutated structure in comparison with wild type. Each of the four mutations invariably exhibit altered binding affinities with ACE2 receptor (**Table-1**). Although no changes were evident in the interactions during conservative substitution of V367F (**Figure 4c**, **Supplementary Figure 1a**), the binding affinity towards ACE2 receptor decreased ($K_d$ = 1.4E-09) when compared to the wild type protein complex ($K_d$ = 4.3E-10). The amino acid substitution that is unique to Indian strains, S438F is associated with alterations in hydrogen bonding with D442. Also, a new hydrogen bond with N437 is formed in the mutant (**Figure 4d, Supplementary Figure 1b**). These structural alterations are consistent with the reduced binding affinity observed in S438F mutant ($K_d$ = 1.2E-09). The amino acid mutation at position 476, G to S also made some changes to its interacting residues. In wild type protein complex, G476 is forming two hydrogen bonds with T478 while this interaction is abolished in the mutant. Also, the single hydrogen bond interaction with Q474 in the wild type was replaced to double bond in mutant (**Figure 4e, Supplementary Figure 1c**) likely reducing the binding affinity with ACE2 ($K_d$ = 2.5E-09). Though V483A mutant does not exhibit considerable alterations in hydrogen bonding, reduced binding affinity was observed ($K_d$ = 1.9E-09). We are currently performing additional molecular docking studies to understand the RBD: ACE2 interface alterations in mutants. In nutshell, our findings might shed light on understanding severity and dynamics of COVID 19 and also in developing the right vaccines and therapeutics.

| | Wildtype | Mutation in 367 | Mutation in 438 | Mutation in 476 | Mutation in 483 | Mutation in 614 |
|---|---|---|---|---|---|---|
| HADDOCK Score | -130.8 +/- 4.5 | -126.1 +/- 6.1 | -117.6 +/- 3.4 | -124.0 +/- 5.2 | -126.2 +/- 1.6 | -128.1 +/- 11.5 |
| Binding affinity ΔG (kcal mol-1) | -12.8 | -12.1 | -12.2 | -11.7 | -11.9 | -12.9 |
| Dissociation Constant Kd (M) | 4.3E-10 | 1.4E-09 | 1.2E-09 | 2.5E-09 | 1.9E-09 | 3.5E-10 |

**Table-1: Summary of** molecular docking analysis comparing binding affinities of RBD mutants *vs* wildtype.

**Acknowledgements**

The authors wish to acknowledge John B Johnson, Mahendran KR and Sara Jones for critical comments. The work was supported by the Department of Biotechnology, Government of India.

## References

1. Cascella, Marco, et al. "Features, evaluation and treatment coronavirus (COVID-19)." *StatPearls [Internet]*. StatPearls Publishing, 2020.
2. Chen, J. (2020). Pathogenicity and transmissibility of 2019-nCoV—a quick overview and comparison with other emerging viruses. Microbes and infection.
3. Chinese SMEC. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. Science. 2004;303(5664):1666–9.
4. De Vries, Sjoerd J., Marc Van Dijk, and Alexandre MJJ Bonvin. "The HADDOCK web server for data-driven biomolecular docking." Nature protocols 5.5 (2010): 883.
5. Delano, W. L. TePyMol Molecular Graphics System. Proteins Structure Function and Bioinformatics. 30, 442–454 (2002).
6. Dong S, Sun J, Mao Z, Wang L, Lu YL, Li J. A guideline for homology modeling of the proteins from newly discovered betacoronavirus, 2019 novel coronavirus (2019-nCoV). Journal of Medical Virology. 2020 Mar 17.
7. Du, L., He, Y., Zhou, Y. et al. The spike protein of SARS-CoV — a target for vaccine and therapeutic development. Nat Rev Microbiol 7, 226–236 (2009). https://doi.org/10.1038/nrmicro2090
8. Hoffmann M, Kleine-Weber H, Krüger N, Mueller MA, Drosten C, Pöhlmann S. The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. BioRxiv. 2020 Jan 1.

9. S. Kalyaanamoorthy, B.Q. Minh, T.K.F. Wong, A. von Haeseler, L.S. Jermiin (2017) ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat. Methods, 14:587-589.
10. Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." Nucleic acids research 30, no. 14 (2002): 3059-3066.
11. Krieger, E., Koraimann, G. &Vriend, G. Increasing the precision of comparative models with YASARA NOVA—a selfparameterizing force feld. Proteins: Structure, Function, and Bioinformatics. 47, 393–402 (2002)
12. Liu, S. et al. Interaction between heptad repeat 1 and 2 regions in spike protein of SARS-associated coronavirus: implications for virus fusogenic mechanism and identification of fusion inhibitors. *Lancet* **363**, 938–947 (2004).
13. Liu, Shuwen, et al. "Interaction between heptad repeat 1 and 2 regions in spike protein of SARS-associated coronavirus: implications for virus fusogenic mechanism and identification of fusion inhibitors." *The Lancet* 363.9413 (2004): 938-947.
14. Liu, Zhixin, et al. "Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2." *Journal of medical virology* (2020).
15. Lu, L. et al. Structure-based discovery of Middle East respiratory syndrome coronavirus fusion inhibitor. *Nat. Commun.* **5**, 3067 (2014).
16. L.-T. Nguyen, H.A. Schmidt, A. von Haeseler, B.Q. Minh (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol. Biol. Evol., 32:268-274.
17. Ortega, Joseph Thomas, et al. "Role of changes in SARS-CoV-2 spike protein in the interaction with the human ACE2 receptor: An in silico analysis." *EXCLI journal* 19 (2020): 410.
18. Ou, Xiuyuan, et al. "Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV." *Nature communications* 11.1 (2020): 1-12.
19. Seemann, Torsten. "Prokka: rapid prokaryotic genome annotation." Bioinformatics 30, no. 14 (2014): 2068-2069.
20. Shafique L, Ihsan A, Liu Q. Evolutionary trajectory for the emergence of novel coronavirus SARS-CoV-2. Pathogens. 2020 Mar;9(3):240.
21. Steinhauer, D. A., & Holland, J. J. (1987). Rapid evolution of RNA viruses. Annual Reviews in Microbiology, 41(1), 409-431.
22. Torsten Schwede, Jürgen Kopp, Nicolas Guex, Manuel C. Peitsch, SWISS-MODEL: an automated protein homology-modeling server, Nucleic Acids Research, Volume 31, Issue 13, 1 July 2003, Pages 3381–3385, https://doi.org/10.1093/nar/gkg520
23. Tripet, Brian, et al. "Structural characterization of the SARS-coronavirus spike S fusion protein core." *Journal of Biological Chemistry* 279.20 (2004): 20836-20849.
24. Wang, Qihui, et al. "MERS-CoV spike protein: Targets for vaccines and therapeutics." *Antiviral research* 133 (2016): 165-177.
25. Waterhouse, Andrew M., James B. Procter, David MA Martin, Michèle Clamp, and Geoffrey J. Barton. "Jalview Version 2—a multiple sequence alignment editor and analysis workbench." Bioinformatics 25, no. 9 (2009): 1189-1191.
26. Worldometer. COVID-19 coronavirus outbreak, 23 April 2020. Available at: https://www.worldometers.info/coronavirus/#countries. Accessed on 23 April 2020
27. Wrapp, Daniel, et al. "Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation." *Science* 367.6483 (2020): 1260-1263.

28. Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., ... & Sheng, J. (2020). Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell host & microbe

29. Wu, Fan, et al. "A new coronavirus associated with human respiratory disease in China." *Nature* 579.7798 (2020): 265-269.

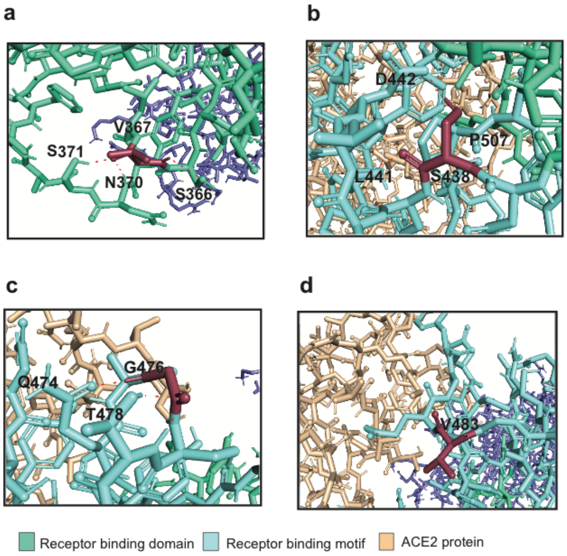
30. Xia, Shuai, et al. "Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion." *Cell research* (2020): 1-13.

31. Xue LC, Rodrigues JP, Kastritis PL, Bonvin AM, Vangone A. PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. Bioinformatics. 2016 Dec 1;32(23):3676-8.

32. Yang, P., & Wang, X. (2020). COVID-19: a new challenge for human beings. Cellular & Molecular Immunology, 1-3

33. Yin, C. (2020). Genotyping coronavirus SARS-CoV-2: methods and implications. arXiv preprint arXiv:2003.10965.

34. Zhou, P., Yang, X., Wang, X. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579,** 270–273 (2020). https://doi.org/10.1038/s41586-020-2012-7



**Supplementary Figure 1**: SARS-Cov2 wild type S protein structure and its complex with ACE2 receptor. Enlarged view of (a) V367 and its interacting residues (S371, N370 and S366) (b) S438 and the interacting residues (D442, L441, and P507) (c) G476 and its interacting residues Q474 and T478 (d) V483