Structural and Functional Implications of Spike Protein Mutational Landscape in SARS-CoV-2

Shijulal Nelson-Sathi*, Perunthottathu K Umasankar*, E Sreekumar, R Radhakrishnan Nair, Iype Joseph, Sai Ravi Chandra Nori, Jamiema Sara Philip, Roshny Prasad, Kolaparamba V Navyasree, Shikha T 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-19 pandemic, is an RNA virus prone to mutations. Interaction of SARS-CoV-2 Spike (S) protein with the host cell receptor, Angiotensin-I Converting Enzyme 2 (ACE2) is pivotal for attachment and entry of the virus. Yet, natural mutations acquired on S protein during the pandemic and their impact on viral infectivity, transmission dynamics and disease pathogenesis remains poorly understood. Here, we analysed 2952 SARS-CoV-2 genomes across the globe, and identified a total of 1815 non-synonymous mutations in the S-protein that fall into 54 different types. We observed that six of these distinct mutations were located in the Receptor Binding Domain (RBD) region that directly engages host ACE2. Molecular phylogenetic analysis revealed that these RBD mutations cluster into distinct phyletic clades among global subtypes of SARS-CoV-2 implying possible emergence of novel sublineages of the strain. Structure-guided homology modelling and docking analysis predicted key molecular rearrangements in the ACE2 binding interface of RBD mutants that could result in altered virus-host interactions. We propose that our findings could be significant in understanding disease dynamics and in developing vaccines, antibodies and therapeutics for COVID-19.

Importance

COVID-19 pandemic shows considerable variations in disease transmission and pathogenesis globally, yet reasons remain unknown. Our study identifies key S-protein mutations prevailing in SARS-CoV-2 strain that could alter viral attachment and infectivity. We propose that the

interplay of these mutations could be one of the factors driving global variations in COVID-19 spread. In addition, the mutations identified in this study could be an important indicator in predicting efficacies of vaccines, antibodies and therapeutics that target SARS-CoV-2 RBD-ACE2 interface.

Introduction

COVID-19, the highly transmissible and pathogenic viral infection that leads to acute respiratory distress syndrome, coagulation dysfunction and septic shock is caused by SARS-CoV-2 (1). The rate of SARS-CoV-2 spread appears to be more than SARS-CoV and MERS-CoV (2). This increased transmission was recently correlated with high mutation frequency of the SARS-CoV-2 strain (3). Spike (S) is a major protein in SARS-CoV-2 responsible for viral entry (4, 5). S-protein consists of N-terminal domain and C-terminal domain in the S1 region that binds to host ACE2, and fusion peptide (FP), heptad repeats (HR1&HR2), transmembrane domain (TM) and a short cytoplasmic domain (CP) in the S2 region that fuses with host cell membrane (6, 7). Together, these features make S-protein the primary target for the development of antibodies, entry inhibitors and vaccines (8). Recent studies show that mutations in the S-protein can modulate viral pathogenesis (9). Taken together, we hypothesized that S-protein of SARS-CoV-2 strain must have acquired genetic changes that correlate with global variations in disease spread.

Methods

Genome analysis

All available complete and high coverage SARS CoV-2 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 not considered in our analysis. In addition, we added 30 Indian genomes downloaded on April 15th, 2020 to the analysis. The Wuhan RefSeq genome (NC_045512) is taken as the reference for our mutational analysis. The genes were predicted using Prokka (14), 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*) (15). The alignments were visualized in

Jalview (16) and the amino acid substitutions in each position were extracted using custom python script. We ignored the substitutions that were 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. Also, we performed an independent random sampling of another 2,000 S proteins from 26,567 genomes to identify new RBD mutations using the above criteria, and those additional RBD mutations present were further considered in our study.

Phylogeny reconstruction

For the Maximum likelihood phylogeny reconstruction, we have sampled 10 SARS COv-2 genomes per country from the worldwide strains with Wuhan RefSeq strain as root. Sequences were aligned using Mafft (*maxiterate* 1,000 and global pair-*ginsi*), and phylogeny was reconstructed using IQ-Tree (17). The best evolutionary model (TIM2+F+I) was picked using the ModelFinder program (18). In addition, three independent random samplings of genomes were done and repeated the analysis to check the consistency of the phylogeny.

Structural analysis

The structural analysis of the mutated spike glycoprotein of SARS-CoV-2 RBD domain was done to assess the impact of mutations on binding affinity towards the human ACE2 (hACE2) receptor. The crystal structure of the SARS-CoV-2 RBD-hACE2 receptor complex was downloaded from Protein Data Bank (PDB ID:6LZG) and the mutagenesis analysis was performed using Pymol (19). Homology modelling of Mouse ACE2 (mACE2) structure was performed in Swiss-Model (20) using SARS-CoV-2 RBD-hACE2 as template. The YASARA server (21) was used for the energy minimization of analysed structures. The Z-dock webserver (22) was used for docking the mACE-2 and the spike protein RBD of SARS-CoV-2. The binding affinity of the wild, mutated and docked structures was calculated using PRODIGY web server (23). The hydrogen bond and salt bridge interactions were calculated using Protein Interaction Calculator (24) and the Vander Waals interactions were calculated using Ligplot (25). All the visualizations were done using Pymol (19).

Results and Discussion

Non-synonymous mutational profile of SARS-CoV-2 S protein

To capture S-protein variations globally, we searched for non-synonymous mutations in S-protein sequences translated from 2,954 SARS-CoV-2 genomes deposited in GISAID till 6th April, 2020. Altogether, 1,815 non-synonymous mutations were identified that belong to 1,176 genomes from 29 countries. These mutations fall into 54 distinct types with 26 different amino acid substitutions present in the S1 segment, 24 in the S2 segment and 4 in the S1-S2 junction of the S-protein (**Table 1**). D614G, which marks the high virulent A2 subtypes of SARS-CoV-2 was found to be the most abundant mutation accounting for 52.26% (n=1,544) of the genomes analyzed.

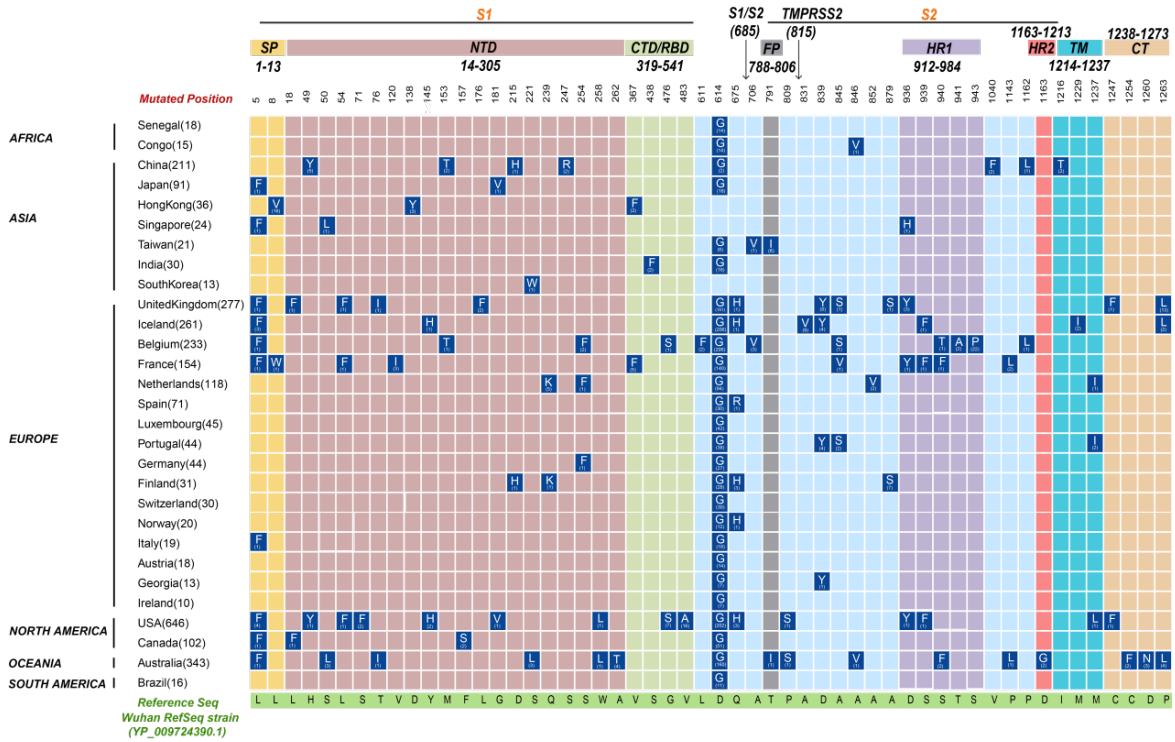


Table 1: Matrix representing amino acid substitutions present in SARS-CoV-2 S protein of 2,954 genomes. Name of countries and the number of genomes sampled are given on the Y-axis and the relevant amino acid residues (single letter code) in the reference strain are given on the X-axis. Mutated amino acid residues and their frequency of occurrence 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.

Non-synonymous RBD mutations

RBD comprises of 223 amino acid long peptide in the S1-region that connects SARS-like coronaviruses to various hosts by directly binding to cellular ACE2 (10). To cover all significant mutations in RBD we expanded our analysis to additional 2000 SARS-CoV-2 genomes. We found 6 distinct types of amino acid substitutions in RBD *viz*; V367F, P384L, S438F, K439N, G476S, and V483A (**Figure 1A**). Overall, these mutations were distributed in 39 isolates from 7 different countries in the world- V367F from Hong Kong and France, P384L from Russia and USA, S438F from India, N439K from UK, G476S from USA and Belgium and V483A from USA. However, the RBD mutations accounted for only ~2% of the total S-protein non-synonymous mutations identified in our analysis. Notably, several non-synonymous mutations were also observed in single independent isolates across the world (Q321L, T323I, P330S, V341I, A344S, A348T, N354D, D364Y, K378R, Q409E, Q414R, A435S, V445A, K458N, K458R, D467V, I468F, I468T, I472V, A475V, S477G, S477I, N481H, P491R, Y508H, R509K, V510L, E516Q, H519P, H519Q, A520S and A522V). Though these mutations got filtered out in our stringent sampling criteria, they may potentially get enriched in future analysis.

Evolutionary pattern of RBD variants

RBD from SARS-CoV-2 is only 73.4 % identical to SARS-CoV (2002) but is 90.1 % identical to Bat coronavirus RaTG13, a suspected precursor of SARS-CoV-2.To see the evolutionary trend of non-synonymous mutations in RBD, we compared RBDs from SARS-CoV-2, SARS-CoV and RaTG13. None of the RBD substitutions observed in SARS-CoV-2, except N439K, were identical to the cognate residues in RaTG13 or SARS-CoV (**Figure 1B**) RBDs. This suggests RBD variants observed in our analysis do not reflect evolutionary reversion. Next, we performed phylogenomic analysis to see the evolutionary pattern of RBD variants within the currently circulating SARS-CoV-2 strain. We observed that the RBD mutations cause the isolates to cluster into 6 distinct phyletic clades that can be classified under 4 major SARS-CoV-2 subtypes **(Figure 1C)**. All the isolates containing S438F, N439K, P384L and one isolate possessing G476S co-clustered with the A2a subtype of SARS-CoV-2. In V367F variants, two isolates formed an independent cluster with the B subtype whereas four isolates did not cluster with any existing SARS-CoV-2 subtypes. Sixteen isolates of V483A and six isolates of G476S co-existed

with the signature mutations defining the B1 subtype **(Figure 1C)**. Together, the clustering pattern of RBD variants shows a likelihood of emergence of sub-lineages within the SARS-CoV-2 strain.

Figure 1: A. Matrix representation of amino acid substitutions present in the RBD region of SARS-CoV-2 Spike protein. Countries where RBD variants are present are shown in the Y-axis and the relevant amino acid residues (single letter code) in the reference strain are represented on the X-axis. **B.** Conservation of Receptor Binding Domain (RBD) of SARS-CoV-2 with its close relatives, SARS-CoV and Bat RaTG13. The blue colored region shows RBD and the yellow highlighted region is the Receptor Binding Motif (RBM) within RBD. The mutated residues are highlighted in light blue and substitutions are marked below. Non-conserved residues are in grey color **C.** Maximum Likelihood Phylogenetic tree of 289 SARS-CoV-2 strains sampled from 5000 genomes. RBD variants are highlighted in red, and the corresponding mutation is marked in blue. Branches are coloured based on the known SARS-CoV-2 subtypes. The phylogeny is rooted with Wuhan RefSeq strain and highlighted in green.

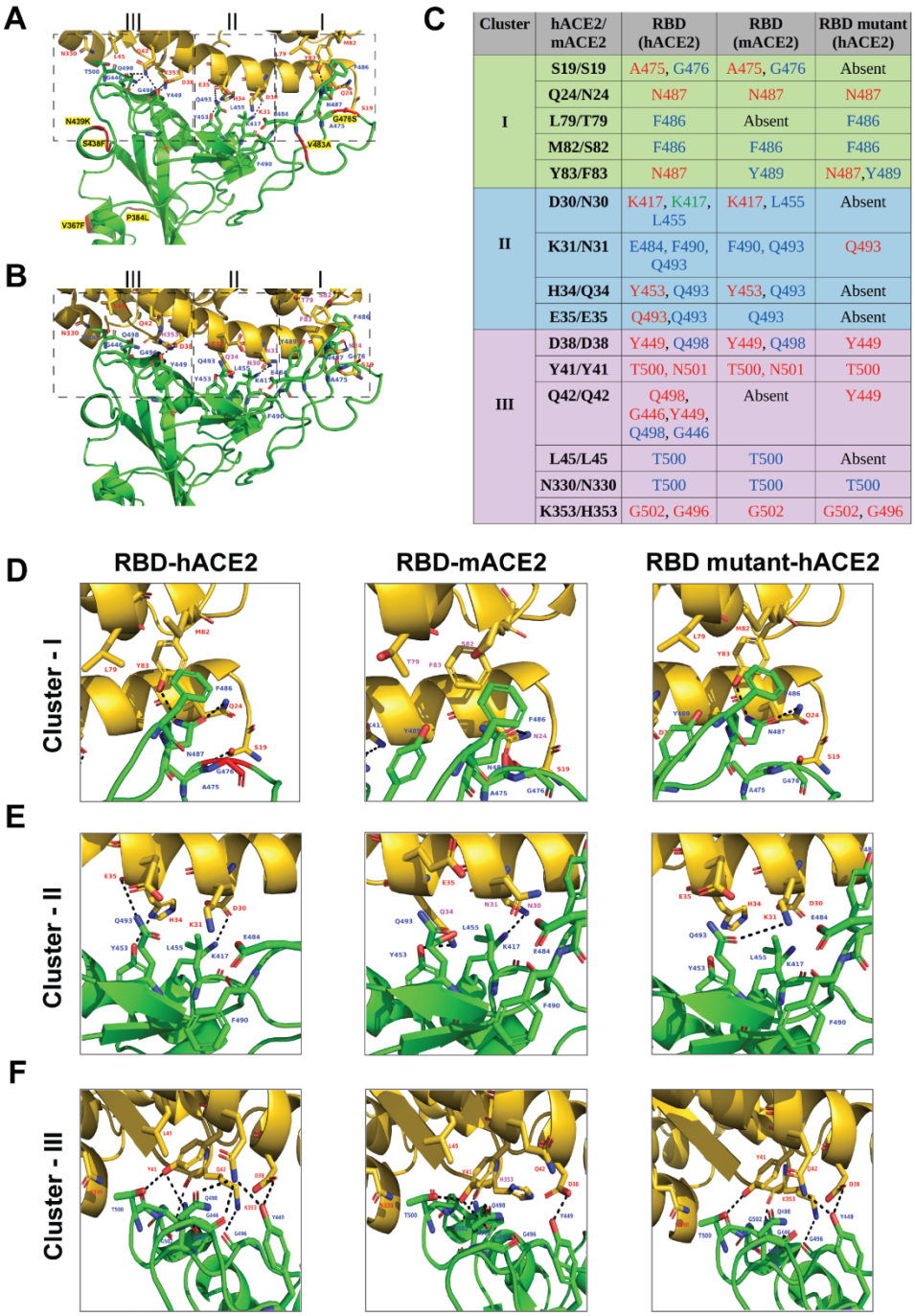
Structural and Functional implications of RBD mutations

We analyzed the effect of natural RBD variants on the overall binding affinity of S-protein with ACE2. Structural information from three recently reported crystal structures of SARS-CoV-2 RBD-ACE2 bound complex was consolidated to predict binding affinities (6, 11, 12). Two RBD mutants, V367F and N439K showed higher binding affinity to ACE2 than the wild type RBD ($\Delta G = -12.5$) with a respective two-fold and three-fold increase in dissociation constants (K_d). On the contrary, the other RBD mutants, P384L, S438F, G476S and V483A substitutions showed lower binding affinity than the wild type with ~ten-fold increase in K_d values (**Figure 1A**).

To gain insight into possible molecular mechanisms underlying altered binding affinities of RBD mutants, we analyzed molecular interactions at the RBD-ACE2 binding interface. It has been proposed that mouse cannot function as a natural reservoir of SARS-CoV as the key ACE2 residues responsible for high affinity binding to RBD in other species are mutated in mouse (13). Homology modeling and molecular docking of mouse ACE2 on SARS-CoV-2 RBD- human ACE2 structure allowed us to visualize key ACE2-RBD residues and resulting molecular arrangements that allow or abolish virus-host interaction (**Figure 2A** and **2B**). Thus we considered RBD- human ACE2 as a positive control and RBD- mouse ACE2 as a negative control in our structural analysis and compared molecular rearrangements induced by RBD mutations.

Crystal structure of RBD comprises of mostly random coil with few β-sheet scaffolds in between which likely allows simultaneous conformational flexibility and stability. On the other hand, the major binding residues on ACE2 are located on a single α- helix (**Figure 2A and 2B**). Further, most RBD-ACE2 molecular interactions are polar in nature and comprises of several hydrogen bonds and vander Waals' forces. We divided RBD-ACE2 interface into three clusters based on spatial arrangement of interacting residues- Cluster-I, II and III (**Figure 2A, 2B and 2C**). Mapping of mutated residues on RBD structure with respect to clusters revealed interesting correlations between spatial arrangements and binding affinities. None of the mutated RBD residues except G476 were directly involved in ACE2 interaction. G476S and V483A were present in the near vicinity of Cluster-I, N439K and S438F were close to Cluster-III. These four mutations were in the receptor binding motif (RBM) within the RBD (**Figure 1B**). V367F and P384L were outside RBM near to Cluster-III. No mutations were found at or near Cluster-II.

Strikingly, residues that showed high affinity substitutions (V367F and N439K) were found to be located on the RBD β-sheet (**Figure 2A**). These mutations also induced additional hydrogen bonds in the Cluster II interface (K31-F490 and K31-Q493) which are not present in wild type or other mutants likely supporting increase in binding affinity (**Figure 2C and Table S1**).



Cluster	hACE2/ mACE2	RBD (hACE2)	RBD (mACE2)	RBD mutant (hACE2)
I	S19/S19	A475, G476	A475, G476	Absent
	Q24/N24	N487	N487	N487
	L79/T79	F486	Absent	F486
	M82/S82	F486	F486	F486
	Y83/F83	N487	Y489	N487,Y489
II	D30/N30	K417, K417, L455	K417, L455	Absent
	K31/N31	E484, F490, Q493	F490, Q493	Q493
	H34/Q34	Y453, Q493	Y453, Q493	Absent
	E35/E35	Q493,Q493	Q493	Absent
III	D38/D38	Y449, Q498	Y449, Q498	Y449
	Y41/Y41	T500, N501	T500, N501	T500
	Q42/Q42	Q498, G446,Y449, Q498, G446	Absent	Y449
	L45/L45	T500	T500	Absent
	N330/N330	T500	T500	T500
	K353/H353	G502, G496	G502	G502, G496

Figure 2: Molecular rearrangements in RBD-ACE2 interface. **A**. SARS-CoV-2 RBD-human ACE2 interface showing key residues (in red and blue) and clusters (boxed region) of interaction. RBD mutations identified are

highlighted in yellow. **B**. SARS-CoV-2 RBD-mouse ACE2 docked interface. Mutated ACE2 residues in mouse are marked in purple. **C**. List of cluster specific molecular interactions of hACE2, mACE2, mutated RBD-ACE2 complexes. Hydrogen bonds are marked in red, vander Waal's in blue and salt bridges in green. **D-F.** Structural visualization of key interactions listed in **C**. SARS-CoV-2 RBD is represented in green and ACE2 receptors in gold. The interacting residues of hACE2 and RBD shown as dotted lines.

On the contrary, the residues having low affinity replacements (P384L, S438F, G476S and V483A) were present in the random coil region (**Figure 2A**). We observed that at least one of the several molecular interactions present in human (S19-A475, D30-K417, H34-Y453 and Y41-N501, L45-T500 hydrogen bonds) and abolished in mouse (E35-Q493, Q42-Q498, Q42-G446 hydrogen bonds; D30-K417 salt bridge; and K31-E484, Q42-Q498, Q42-G446 vander Waal's interaction) were absent in each of these four mutants, a likely reason for their low binding affinity to ACE2 (**Figure 2C and Table S1**). A structural representation of each interacting clusters in mutant RBD-ACE2 interface revealed structures resembling human RBD-ACE2, mouse RBD-ACE2 or a hybrid of both further confirming our notion (**Figure 2D-F**).

Since RBD-ACE2 binding can be directly linked to viral infectivity, we think these mutations may have an influence on the clinical outcome of patients and on disease spread. Clinical metadata associated with GISAID database showed severe hospitalization history for patients having V367F variants likely indicating this possibility. However, similar correlations were not possible for other RBD mutants due to unavailability of clinical metadata.

Our study showed the mutational landscape prevailing in SARS-CoV-2 S-protein especially on RBD and revealed the possible structural and functional implications of these mutations. We think that the interplay of these S-protein mutations may provide clues regarding global variations in viral infectivity and disease spread. At the least, our study could serve as a molecular directory for experimental biologists to validate the impact of S-protein mutations.

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. Yang, P., & Wang, X. (2020). COVID-19: a new challenge for human beings. Cellular & Molecular Immunology, 1-3
2. Chen, J. (2020). Pathogenicity and transmissibility of 2019-nCoV—a quick overview and comparison with other emerging viruses. Microbes and infection.
3. Yin, C. (2020). Genotyping coronavirus SARS-CoV-2: methods and implications. arXiv preprint arXiv:2003.10965.
4. 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
5. Gordon et al., (2020). A SARS-CoV-2 Protein Interaction Map Reveals Targets for Drug Repurposing. Nature.
6. Wang et al., (2020). Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. Cell.
7. Hoffmann et al., (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell,181(2):271-280
8. Du, L., He, Y., Zhou, Y. et al. (2009). The spike protein of SARS-CoV — a target for vaccine and therapeutic development. Nat Rev Microbiol 7, 226–236.
9. Shang et al., (2020). Cell Entry Mechanisms of SARS-CoV-2. Proc Natl Acad Sci U S A. 117(21):11727-11734.
10. 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).
11. Lan et al., (2020). Structure of the SARS-CoV-2 Spike Receptor-Binding Domain Bound to the ACE2 Receptor. Nature. 581(7807):215-220.
12. Yan et al., (2020). Structural Basis for the Recognition of SARS-CoV-2 by Full-Length Human ACE2. Science. 367(6485):1444-1448.
13. Li, W. et al. (2005). Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. EMBO J 24, 1634-1643.
14. Seemann, T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30.14: 2068-2069.
15. Katoh, et al. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* 30.14: 3059-3066.
16. Waterhouse, Andrew M., et al. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25.9 (2009): 1189-1191.
17. Nguyen, et al. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* 32.1: 268-274.
18. Kalyaanamoorthy, et al. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods* 14.6: 587.
19. DeLano, W. L. (2002) The PyMOL molecular graphics system. DeLano Sci. *San Carlos, CA* 30: 442-454.
20. Schwede, et al. (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic acids research* 31.13: 3381-3385.

21. Krieger, et al., (2002) Increasing the precision of comparative models with YASARA NOVA—a self-parameterizing force field. *Proteins: Structure, Function, and Bioinformatics* 47.3: 393-402.
22. Pierce, et al. (2014) ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics* 30.12: 1771-1773.
23. Xue, Li C., et al. (2016) PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics* 32.23: 3676-3678.
24. Tina, K. G., Rana Bhadra, and Narayanaswamy Srinivasan. (2007) PIC: protein interactions calculator. *Nucleic acids research* 35.suppl_2: W473-W476.
25. Wallace, Andrew C., Roman A. Laskowski, and Janet M. Thornton. (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein engineering, design and selection* 8.2: 127-134.

	hACE2	RBD	V367F	P384L	S438F	N439K	G476S	V483A
Cluster- I	S19	A475 G476	Absent	A475	Absent	A475	A475	A475
	Q24	N487	N487	N487	N487	N487	N487	N487
	L79	F486	F486	F486	F486	F486	F486	F486
	M82	F486	F486	F486	F486	F486	F486	F486
	Y83	N487	N487 Y489	N487 Y489 Y489	N487 Y489 Y489	N487 Y489	N487 Y489	N487 Y489
Cluster- II	D30	K417 K417 L455	K417 K417 L455	K417 K417 L455	Absent	K417 L455	K417 K417 L455	K417 K417 L455
	K31	E484 F490 Q493	F490 Q493 F490 L492 Q493	Q493 F490 L492 Q493	Q493 Q493	Q493 F490 L492 Q493	E484	F490 Q493
	H34	Y453 Q493	Q493	Q493	Q493	Q493	Absent	Q493
	E35	Q493 Q493	Q493	Q493	Q493	Q493	Absent	Absent
Cluster- III	D38	Y449 Q498	Y449	Y449	Y449	Y449 Q498	Y449	Y449
	Y41	T500 N501	T500 N501	T500 N501	T500	T500 N501	T500	T500
	Q42	G466 Y449 Q498 G446 Q498	Y449	Y449	Y449	Y449	Y449	Y449
	L45	T500	T500	T500	T500	Absent	T500	Absent
	N330	T500	T500	T500	T500	T500	T500	T500
	K353	G496 G502	G496 G502	G496 G502	G496 G502	G496 G502	G496 G502	G496 G502

Table S1: List of cluster specific molecular interactions of hACE2 with RBD and six RBD mutants. Hydrogen bonds are marked in red, vander Waal's in blue and salt bridges in green.