Natural variants in SARS-CoV-2 S protein pinpoint structural and functional hotspots; implications for prophylaxis strategies

Suman Pokhrel[1], Benjamin R. Kraemer[1] and Daria Mochly-Rosen*

Department of Chemical and Systems Biology, Stanford University School of Medicine, Stanford, CA

[1]These authors contributed equally to the study

*Address correspondence to

Daria Mochly-Rosen

Department of Chemical & Systems Biology, School of Medicine, Stanford University, CA, USA

Email: mochly@stanford.edu

Tel: 650-724-8098

Abstract:

The SARS-CoV-2 viral genome mutates incessantly as it spreads in the world and the gene for the Spike (S) protein, critical for viral transmission into humans, is no exception. Analysis of 4,517 variants isolated from humans identified regions with few mutations, thus pinpointing important functional and structural sites in the S protein. This information can guide the development of effective prophylactic agents to arrest the spread of the COVID-19 pandemic.

Main text

To curb the COVID-19 pandemic, prophylaxis development has focused on preventing entry of the virus by inhibiting the interaction of SARS-CoV-2 with its human receptor, angiotensin converting enzyme 2 (ACE2)[1]. The S protein on the virus is responsible for the interaction with ACE2. Proteases cleave the S protein into S1 and S2 subunits, to enable viral binding to ACE2[2] and viral entry by membrane fusion[3]. S1 contains the receptor binding domain (RBD) in either the 'open' (active) or 'closed' (inactive) conformations[4,5,6] (Supplementary Fig. 1a). Four types of prophylaxis strategies have been employed: 1. Preventing S1 protein proteolysis[7]; 2. Competing with S1 binding to ACE2, using S1 or ACE2 protein fragments or peptides[1]; 3. Generating monoclonal or polyclonal antibodies against SARS-CoV-2 S protein or RBD, to be used as passive vaccines[8]; or 4. Active vaccines, generating an immune response, usually to the S protein, in humans at risk for exposure[9,10,11]. In addition to the RBD, regions such as the trimer interface of S, furin proteolysis sites, glycosylation sites and linoleic acid (LA)-binding site are likely important for maintaining structural integrity, entry, and transmission of the virus and therefore are expected to be more conserved relative to other regions.

To identify areas in the S protein that are the least divergent as the virus evolves in humans, we used an online database[12], that as of November 11, included 189,704 individual

virus sequences isolated throughout the world. As compared with Wuhan (EPI_ISL_402124) sequence of February 2020[13], the 1,273 amino acid S protein[5] had 4,517 variants, excluding in-frame insertions. On average, each position in the protein sequence has approximately 4 variants (Fig. 1a). However, some regions harbor 10 variants per amino acid position whereas others have no variants (Figs. 1a; Supplementary Figs. 1b,c).

Regions in S protein with no more than 2 variants/position are more prevalent in the structurally critical trimer interface (38% of the amino acids; Fig. 1b, Supplementary Fig. 1b,c), the RBD (37%, Fig. 1) and the small LA-binding site (65%; Fig. 2a,b). The trimer interface is less accessible and therefore unlikely to be druggable, but the LA pocket could be a potential target for therapeutics with small molecules that stabilizes the S protein in closed/inactive conformation; the fatty acid-binding pocket in the inactive conformation of S protein is conserved in other coronaviruses[6] and 82% of the variants among the 20 amino acids that make this pocket are predicted to have a neutral effect (Extended Data Table 1).

Much of the therapeutic efforts are focused on the RBD (amino acids 331-524). The RBD contains ten invariant amino acids (Fig. 1c-e) and only 7% of the RBD variants are predicted by PROVEAN software[14] to be structurally or functionally damaging (Extended Data Table 1). Therefore, drugs targeting the RBD are expected to be effective prophylactics to most SARS-CoV-2 variants. We computationally determined the change in affinity of each variant relative to the Wuhan variant. Of the amino acids that are within 4.5 Å of ACE2, 26 residues had 81 different variants; 29 variants had calculated increased affinity, of which 7 had >1kcal/mol increased affinity (Extended Data Table 2). (We calculated the impact of one variant at a time, noting that combinations of variants may have different effects on RBD/ACE2 affinity.) Whether these potential changes in affinity result in altered fitness of the virus for infectivity remains to be determined.

We also identified another so called ‘hot’ or less variable region between residues 541-

612 (Figs. 1a, 2c-e). This invariable region is relatively hydrophobic, yet a substantial number of residues remain exposed in the open and closed conformations (Supplemental Fig. 1d). Very recently, Q564 within this region has been proposed to act as a 'latch', stabilizing the closed conformation of S[15]. We suggest that this hydrophobic domain may participate in membrane fusion. Determining the role of this invariable region is important, as it may be another Achilles heel to target for anti SARS-CoV-2 treatment.

We next examined other regions in the S protein for which functions have been assigned. Furin proteolysis at the S1-S2 boundary (681-685) and in S2 (811-815) exposes the RBD to enable ACE2 binding, and the S2 – to initiate membrane fusion[2]. Yet both the consensus sites and the critical arginine are not conserved (Fig. 2f), in agreement with a prior analysis of furin site 1[16]. This supports the reports that other proteases may contribute to S protein maturation[17]. Therefore, drugs that focus on inhibiting furin specifically or any single protease may not be effective preventive treatment against all SARS-CoV-2 variants.

The S protein has also 66 glycosylation sites in each trimer, which facilitate protein folding and may lead to host immune system evasion[18]. Surprisingly, none of the asparagine residues serving as a glycosylation site are conserved, suggesting that not all the glycosylation sites are essential (Supplementary Fig. 2a, b).

Interaction between NRP1 and S protein was proposed to regulate SARS-CoV-2 transmission[19]. However, this interaction site in S1 is not conserved (Supplementary Fig. 2c); 73% of positions had more than 2 variants. Although all the variants are predicted to have a neutral effect (Extended Data Table 1), this may not be a good target for prophylactic treatment against all SARS-CoV-2 variants.

Our analysis of the frequency of variants throughout the S protein of SARS-CoV-2 identified regions of high and low divergence, which may aid in developing effective prophylactic treatments. In this analysis of mutations in S, we did not consider the frequency of a

particular mutation nor in how many countries the mutation was found. Such analysis, as was done for D614G[20], may further aid in determining the potential fitness acquired by a particular mutation. Our data suggest that it may be beneficial to develop passive and active vaccines that target the RBD, instead of the entire S protein, as well as small molecules that lock the closed conformation, such as those mimicking the LA. However, considering the high rate of mutation throughout the S protein, resistance to treatment with monoclonal antibodies is likely to occur. Finally, drugs and antibodies targeting region 541-612, a relatively conserved and exposed region on the protein's surface, may also have prophylactic benefit.

Acknowledgements

Supported in part by the 2020 COVID-19 Response: Drug and Vaccine Prototyping Grant from the Innovative Medicines Accelerator, Stanford University to D. M.-R. We gratefully thank the many investigators throughout the world that provided the SARS-CoV-2 sequences to this public database.

Author Contributions

S.P. and B.R.K provided data analysis, visualization and draft writing. D. M.-R. conceived the project, supervised the analysis and writing.

Competing Interests Statements

The authors declare no competing interests.

References

1. Yang, J. *et al.* Molecular interaction and inhibition of SARS-CoV-2 binding to the ACE2 receptor. *Nat. Commun.* **11**, (2020).
2. Hoffmann, M. *et al.* SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**, 271-280.e8 (2020).
3. Xia, S. *et al.* Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. *Cellular and Molecular Immunology* **17**, 765–767 (2020).
4. Benton, D. J. *et al.* Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature* (2020) doi:10.1038/s41586-020-2772-0.
5. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2019).
6. Toelzer, C. *et al.* Free fatty acid binding pocket in the locked structure of SARS-CoV-2 spike protein downloaded from. *Science* **370**, 725–730 (2020).
7. Cheng, Y. W. *et al.* Furin inhibitors block SARS-CoV-2 spike protein cleavage to suppress virus production and cytopathic effects. *Cell Rep.* **33**, (2020).
8. Liu, L. *et al.* Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* **584**, 450–456 (2020).
9. Jackson, L. A. *et al.* An mRNA vaccine against SARS-CoV-2 — preliminary report. *N. Engl. J. Med.* **383**, (2020).
10. Mulligan, M. J. *et al.* Phase I/II study of COVID-19 RNA vaccine BNT162b1 in adults. *Nature* **586**, 589–593 (2020).
11. Walsh, E. E. *et al.* Safety and immunogenicity of two RNA-based Covid-19 vaccine candidates. *N. Engl. J. Med.* (2020) doi:10.1056/nejmoa2027906.
12. https://mendel.bii.a-star.edu.sg/METHODS/corona/beta/MUTATIONS/hCoV-19_Human_2019_WuhanWIV04/hCoV-19_Spike_new_mutations_table.html.

13. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).

14. Choi, Y. & Chan, A. P. PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745–2747 (2015).

15. Peters, M. H., Bastidas, O., Kokron, D. S. & Henze, C. E. Static all-atom energetic mappings of the SARS-Cov-2 spike protein and dynamic stability analysis of 'Up' versus 'Down' protomer states. *PLoS One* **15**, e0241168 (2020).

16. Xing, Y., Li, X., Gao, X. & Dong, Q. Natural polymorphisms are present in the furin cleavage site of the SARS-CoV-2 spike glycoprotein. *Front. Genet.* **11**, (2020).

17. Seyran, M. *et al.* The structural basis of accelerated host cell entry by SARS-CoV-2. *FEBS J.* (2020) doi:10.1111/febs.15651.

18. Watanabe, Y. *et al.* Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nat. Commun.* **11**, (2020).

19. Cantuti-Castelvetri, L. *et al.* Neuropilin-1 facilitates SARS-CoV-2 cell entry and infectivity. *Science* **370**, 856–860 (2020).

20. Fang, S. *et al.* GESS: a database of global evaluation of SARS-CoV-2/hCoV-19 sequences. *Nucleic Acids Res.* (2020) doi:10.1093/nar/gkaa808.

Natural variants in SARS-CoV-2 Figures

Figure 1. Functional regions in S protein and the RBD-ACE2 interaction site. **a)** The number of variants per position across entire sequence of S protein, highlighting specific functional regions. **b)** Spike homotrimer with ribbons colored according to legend, bound to ACE2 (red). Black dotted outline shown in **c**. **c**) RBD-ACE2 interface. **d)** RBD-ACE2 interface highlighting residues in RBD within 4.5Å from ACE2. **e)** The number of variants per position across RBD domain.

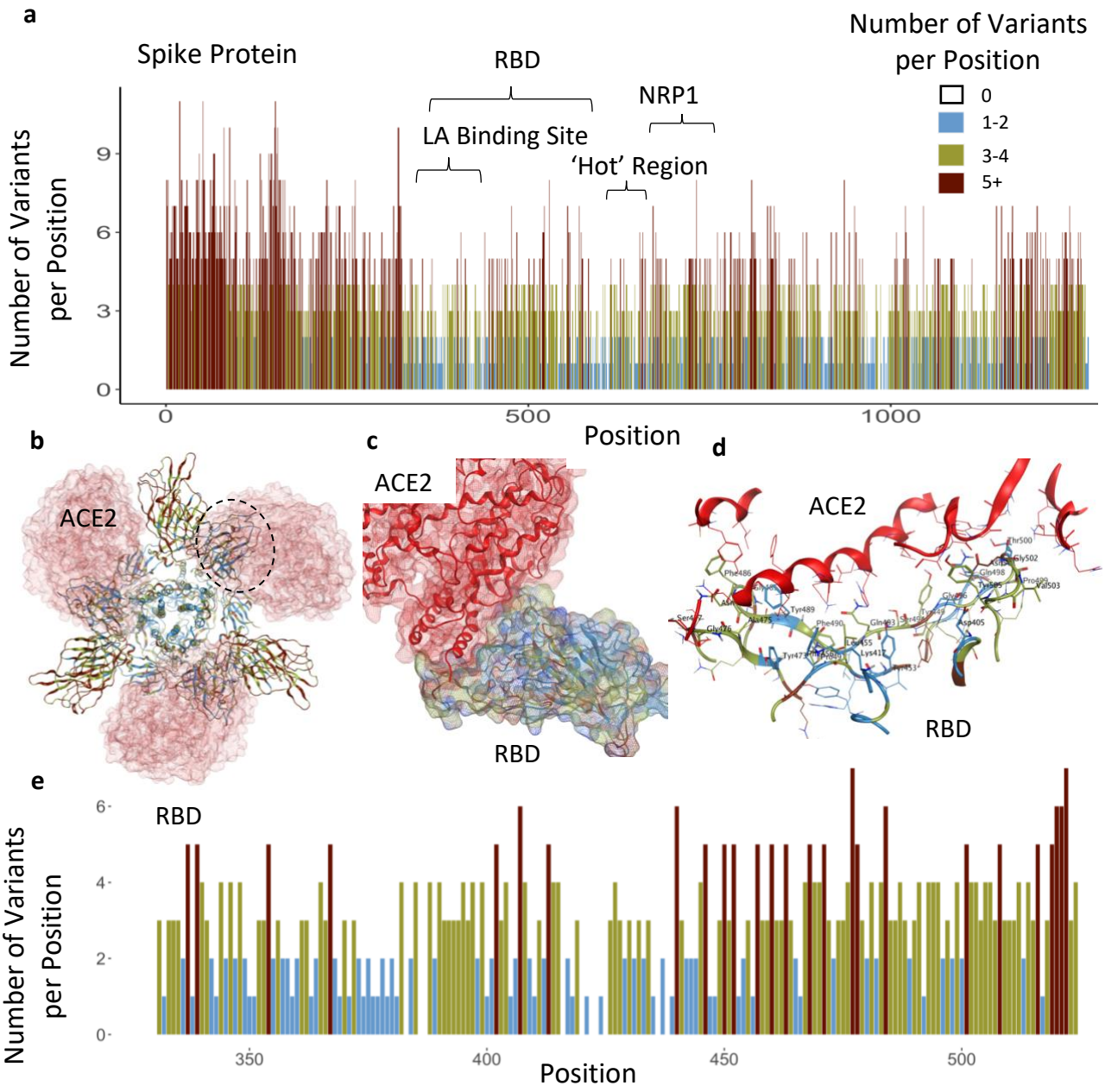

Figure 2. LA-binding site, furin cleavage sites, and hotspots with no known function. **a)** LA bound to fatty acid pocket in S protein. **b)** The number of variants per position across the LA-binding site; circles indicate the LA-pocket. **c)** The number of variants per position across the less-variable, hot regions with un-assigned functions. The star identifies the proposed 'latch', Q564 residue. **d** & **e**) The hot region identified in the 3-D structure of S (open conformation). **f**) Furin cleavage sites in S protein. Amino acid indicated below graphs are of the sequence in the Wuhan isolate. Variants in other SARS-CoV-2 are indicated within the bars of the graphs using one letter abbreviations for the amino acids.