D936Y and Other Mutations in the Fusion Core of the SARS-Cov-2 Spike Protein Heptad Repeat 1 Undermine the Post-Fusion Assembly

D936Y and Other Mutations in the Fusion Core of the SARS-Cov-2 Spike Protein Heptad Repeat 1 Undermine the Post-Fusion Assembly

Luigi Cavallo[1] and Romina Oliva[2,*]

[1]King Abdullah University of Science and Technology (KAUST), Physical Sciences and Engineering Division, Kaust Catalysis Center, Thuwal 23955-6900, Saudi Arabia.
[2]Department of Sciences and Technologies, University Parthenope of Naples, Centro Direzionale Isola C4, I-80143, Naples, Italy.

Email: romina.oliva@uniparthenope.it

KEYWORDS: COVID-19, membrane fusion, mutations, HR1, helical bundle, molecular modeling, salt-bridge, aromatic, infectivity.

Abstract

The iconic "red crown" of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is made of its spike (S) glycoprotein. The S protein is the Trojan horse of coronaviruses, mediating their entry into the host cells. While SARS-CoV-2 was becoming a global threat, scientists have been accumulating data on the virus at an impressive pace, both in terms of genomic sequences and of three-dimensional structures. On April 21st, the GISAID resource had collected 10,823 SARS-CoV-2 genomic sequences. We extracted from them all the complete S protein sequences and identified point mutations thereof. Six mutations were located on a 14-residue segment (929-943) in the "fusion core" of the heptad repeat 1 (HR1). Our modeling in the pre- and post-fusion S protein conformations revealed, for three of them, the loss of interactions stabilizing the post-fusion assembly. On May 29th, the SARS-CoV-2 genomic sequences in GISAID were 34,805. An analysis of the occurrences of the HR1 mutations in this updated dataset revealed a significant increase for the S929I and S939F mutations and a dramatic increase for the D936Y mutation, which was particularly widespread in Sweden and Wales/England. We notice that this is also the mutation causing the loss of a strong inter-monomer interaction, the D936-R1185 salt bridge, thus clearly weakening the post-fusion assembly.

Introduction

Coronavirus Disease 2019 (COVID-19) is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). SARS-CoV-2 is a novel virus belonging to the β genus coronaviruses, which also include two highly pathogenic human viruses identified in the last two decades, the severe acute respiratory syndrome coronavirus (SARS-CoV) and the Middle East respiratory syndrome coronavirus (MERS-CoV) (1-3).

Coronaviruses are named after the protruding spike (S) glycoproteins on their envelope, giving a crown (*corona* in latin) shape to the virions (4). Of the four structural proteins of coronavirues, S, envelope (E), membrane (M), and nucleocapsid (N), the S protein is the one playing a key role in mediating the viral entry into the host cells (5-7), making it one of the main targets for the development of therapeutic drugs and vaccines (8-14). Comprised of two functional subunits, S1 and S2, it first binds to a host receptor through the receptor-binding domain (RBD) in the S1 subunit and then fuses the viral and host membranes through the S2 subunit (7,15). In the pre-fusion conformation, the SARS-CoV-2 S protein forms homotrimers protruding from the viral surface, where its RBD binds to the angiotensin-converting enzyme 2 (ACE2) receptor on the host cell surface (1) (like the SARS-CoV homolog (16), and differently from MERS-CoV S, which recognizes a different receptor, the dipeptidyl peptidase 4 (17)). Receptor binding and proteolytic processing by cellular proteases then cause S1 to dissociate and S2 to undergo large-scale conformational changes towards a stable structure, bringing viral and cellular membranes into close proximity for fusion and infection (7,15,18).

While the outbreak of COVID-19 was rapidly spreading all over the world, affecting millions of people and becoming a global threat, laboratories worldwide promptly started to sequence a large number of SARS-CoV-2 genomes. All the available genomic data is accessible through the Global Initiative on Sharing All Influenza Data (GISAID) website, an invaluable open access resource (19,20). Simultaneously, crucial structural knowledge has been achieved on SARS-CoV-2, especially regarding the S protein. 3D structures are now available from the Protein Data Bank (PDB) (21) for the SARS-CoV-2 S protein in the pre-fusion conformation, also bound to the

ACE2 receptor (22-28), and for the post-fusion core of its S2 subunit in the post-fusion conformation (29).

On April 21st 2020, 4 months after the first sequencing (30), 10,823 genomic sequences of SARS-CoV-2 were available from GISAID. Therefore, we considered the time ripe for an assessment of the mutational spectrum of the SARS-CoV-2 spike protein. To this aim, we extracted all the complete S protein sequences from the GISAID 21st April dataset and identified all the mutations occurring in at least 2 identical sequences (see Table S1). From this analysis, a 14-amino acid segment in the fusion core of the heptad repeat 1 (HR1) emerged as a hotspot for mutations. While the mutations we identified corresponded to a 1 mutation every 12 positions along the protein sequence, as many as 6 amino acids were found to be mutated in the above 14-amino acid segment: S929, D936, L938, S939, S940 and S943.

After the proteolytic processing, in the post-fusion conformation, the S protein HR1 and HR2 motifs interact with each other to form a six-helix bundle (6-HB), which promotes initiation of the viral and cellular membranes fusion. The HR1 "fusion core" is named after its role in giving many interactions with HR2 in the post-fusion conformation, thus playing a key role in the virus infectivity (31). Based on the structural location of the above highly concentrated mutations and on their non-conservative nature, we considered them of particular interest and decided to further investigate their structural basis, both in the pre- and post-fusion conformation, as well as their sequencing dates and geographical distribution. As we show in the following, as many as three of them are responsible for the loss of inter-monomer H-bonds in the post-fusion conformation, while one of them, S943P, would introduce unexpected structural strain in the pre-fusion conformation.

A search in the GISAID resource updated to May 29th showed a significant increase in occurrences especially for one mutant, D936Y, unreported to date, which has become a common variant in some European countries, especially Sweden. It is also the mutant having the most significant structural role, causing the loss of an inter-monomer salt bridge in the post-fusion assembly.

Methods

Identification of mutations

We downloaded the 10,823 genomic sequences available from GISAID on April 21st 2020. From these sequences, we extracted the nucleotide sequences of the spike protein and translated them to protein sequences with in-house scripts. Nucleotides sequences featuring an internal stop codon, having at least one undefined ("N") nucleotide or resulting in spike proteins of length different from 1,273 amino acids were discarded. Sequences annotated as pangolin, bat or canine were also discarded. The remaining 7,692 protein sequences were further analysed. First, we clustered them in sets of identical sequences with CD-HIT (32), obtaining 120 clusters of at least 2 sequences and 245 unique sequences. As a reference system for further analyses, we used the first dated (on December 24th 2019) genomic sequence in GISAID, isolated and sequenced in Wuhan (Hubei, China) (30). Then, upon alignment to the reference sequence, we identified point mutations in all the sets of at least two sequences.

We downloaded again the 34,805 genomic sequences available from GISAID on May 29th 2020 (gisaid_hcov-19_2020_05_29_14) and followed the above pipeline to extract 23,332 complete 1273-residue long S protein sequences. We then recorded the presence and frequency in them of any mutation occurring in the fusion core of the HR1 (residues 929-949) with in-house scripts.

Mutants modelling and analysis

Mutants 3D models were built using the mutate_model module of the Modeller 9v11 program (33). This is an automated method for modelling point mutations in protein structures, which includes an optimisation procedure of the mutated residue in its environment, beginning with a conjugate gradients minimisation, continuing with molecular dynamics with simulated annealing and finishing again by conjugate gradients. The used force field is CHARM-22, for details see Reference (34). Models for mutants in the pre-fusion conformation were built starting from the EM structure of the pre-fusion trimeric conformation (PDB ID: 6VSB, resolution 3.46 Å, (22)). Models for mutants in the post-fusion conformation were built starting from the X-ray structure of the S2 subunit fusion core (PDB ID: 6LXT, resolution 2.90 Å, (29)). Molecular models were analysed and visually inspected with Pymol (35). The

COCOMAPS web server (36) was used to analyse the inter-chain contacts and H-bonds as well as the residues accessibility to the solvent.

Results and Discussion

We downloaded all the SARS-CoV-2 genomic sequences from the GISAID resource on April 21st 2020, extracted from them 7,692 complete S protein sequences and identified all the point mutations occurring in at least two identical sequences (see Methods). The 111 mutations we identified, occurring at 105 different positions spread all over the protein sequence, are reported in Table S1, with the relative number of occurrences.

While the mutations we identified were spaced on average 12 positions along the protein sequence, a segment of 14 amino acids harboured 6 mutations, at positions 929, 936, 938, 939, 940 and 943, proposing itself as a mutational hotspot. This sequence segment is part of the "fusion core" of the HR1, in the protein S2 subunit. The HR1 of coronaviruses S proteins undergoes one of the most notable rearrangements within the protein between the pre- and post-fusion conformations. In the post-fusion conformation, in fact, it experiences a refolding of the pre-fusion multiple helices and intervening regions into a single continuous helix (Figure 1). As already mentioned, three of these long helices then form a 6HB with three HR2 helical motifs (18,29,31). The HR1 and its "fusion core" in particular thus play a crucial role in the virus infectivity.

HR1 "fusion core" mutations: update on April 21st

The following 6 mutations were identified in the fusion core of the HR1 on April 21st 2020: S929I, D936Y, L938F, S939F, S940F, S943P. Two of these mutations, D936Y and S943P, were among the most frequent in the ensemble of mutations we identified. Besides the widespread D614G, now dominant over the original D614 variant (37,38), only 5 other mutations (two of them being very peripheral, L5F and P1263L) recurred indeed in $\geq$ 20 sequences (see Table S1). S943P was also reported in (38), where it was hypothesized to be spreading via recombination.

The D936Y mutation was found in 25 sequences overall. In 22 sequences it was associated to the D614G mutation, while in 2 sequences it was associated to both the D614G and the A1020V mutations. The first D936Y/D614G variant was reported as a single sequence in USA (Washington) on March 15th, then, starting from March 17th in England (7 sequences, March 17th to 31st), Wales (7 sequences, March 17th to 30th), the Netherlands (4 sequences, March 18th to 29th) and Iceland (3 sequences, March 19th to 28th). The 2 D936Y/D614G/A1020V variants were both reported from Wales, on March 26th and 30th, therefore it might be hypothesized that they originated from the D936Y/D614G variant, already circulating in Wales at the time. In addition, the D936Y mutation was found in a unique S protein sequence, from France, dated March 18th, where it was not associated to the D614G variant.

The 22 sequences featuring the S943P mutation on April 21st were all from Belgium. Twenty of them were associated to the D614G mutation and were reported between March 1st and March 20th. In addition, two unique sequences featured the S943P mutation.

The S939F mutation was found associated to the D614G mutation in 8 sequences. It was first reported in 1 sequence from France on March 4th, then in 1 sequence from Iceland on March 16th, then again in 5 sequences from USA (Utah) between March 19th and March 29th, then, finally, in 1 sequence from the Netherlands on April 2nd. In addition, this mutation was found in a unique S protein sequence, from Switzerland, dated February 26th, where it was not associated to the D614G variant.

The L938F mutation was a particularly late one; it was found in 2 sequences, associated to the D614G mutation, both from England and dated March 29th.

The S929I mutation was found in 2 sequences from USA (Washington), dated March 12th and 27th, associated to the D614G mutation.

Finally, the S940F mutation had a unique geographical distribution, as it was found in 2 sequences from Australia (New South Wales) dated February 28th and March 4th, not associated to the D614G mutation. In addition, it was found in 1 single sequence from France, dated March 20th, where it was associated to the D614G mutation.

In conclusion, with the exception of S940F, which was found in Australia, all the mutations in the HR1 core fusion were spread in two continents, Europe and/or North America.

Furthermore, most of them originated from the D614G variant. This is in agreement with them seeming to be quite late mutations, sequenced starting from the end of February/March 2020, i.e. over two months after the first Wuhan variant dated December 24th 2019 (30) (Table 1). The frequency and geographical distribution of D936Y especially are noteworthy, considering that it was first sequenced only on March 15th.

Update on May 29th

On May 29th 2020, we analysed the frequency of the above mutations and the emergence of other possible mutations in the HR1 fusion core in the updated GISAID dataset, containing 34,805 genomic sequences, from which we extracted 23,332 complete S protein sequences, thus roughly tripling the originally analysed April 21st dataset. The result was someway surprising.

A positive selection pressure seemed to emerge for the D936Y mutation, which passed from the 25 cases of the April 21st dataset to the 213 cases of the May 29th dataset, corresponding to a ≈9-fold increment. Of the novel occurrences of the D936Y mutation, only 2 were found in USA (Utah and Minnesota), while most of them came from Europe, especially from UK, 68 from England, 56 from Wales and 1 from Scotland, and from Sweden, 55. Notably, the total number of occurrences of the D936Y mutation amounted to the 20.5% of all the 273 sequences available from Sweden, and to the 2.7% and 1.4% of all the sequences available from Wales and England (5397 and 2374, respectively). The remaining ones came from Denmark, 5, and Poland, 1.

The ≈3-fold increment in frequency of the S929I and S939F mutants was in line with the increment of the sequences in the dataset. The three additional occurrences of the S929I mutation were from USA (Washington), Wales and England. A novel S929T mutation was also reported twice from Scotland. Additional occurrences of the S939F mutation were instead from USA, 7, England, 2, Kazakhstan, 1, and UAE, 1. As for the L938F and S940F mutants, their increment was significantly lower than the increment of the sequences in the dataset. A positive selection thus clearly hasn't emerged to date for these mutations. The only additional occurrence of L938F was

from Denmark, while the 2 additional occurrences of S940F were from France and USA (Washington).

The S943P mutation represented a special case. Most of the sequences harbouring such a mutation were indeed modified between the April 21st and the May 29th datasets, so that they do not feature anymore the mutation to proline. However, 3 novel occurrences of the same mutation, S943P, emerged from China (Beijing). In addition, 3 sequences from Scotland presented the novel S943I mutation.

As for the remaining positions of the HR1 fusion core, to May 29th, either they were fully conserved (S937, K933, A942, A944, L945, G946, K947 and Q949), or they hosted one single occurrence of mutation (to valine for A930, to aspartate for I931 and G932, to histidine for Q934 and D935, to alanine for T941 and to arginine for L948). Because of the rarity of such mutations, we will not discuss them here. However, we will continue to monitor them over time.

Sequence conservation

All the amino acids undergoing mutations in the SARS-CoV-2 S protein are conserved in the bat coronavirus RaTG13 S protein (sharing an overall sequence identity of 97% with SARS-CoV-2 S protein), while as many as five of them are mutated in the SARS-CoV-2 S protein (overall 76% sequence identical to the SARS-CoV-2 homolog) (see Figure 1). Four of these mutations are however conservative (aspartate to glutamate, serine to threonine), except S929, which is a lysine in SARS-CoV. It has been proposed that such mutations in the SARS-CoV-2 HR1 may be associated with enhanced interactions with the HR2, further stabilizing the 6-HB structure and maybe leading to increased infectivity of the virus (29). It is noteworthy that no one of the point mutations we identified restored the corresponding SARS-CoV amino acid.

Effect of the mutations on the protein pre-fusion conformation

In the pre-fusion conformation, all the mutated positions, but S943, are located on the second of four non-coaxial helical segments composing the HR1 (Figure 1). Four of them, S929, D936, S939 and S940, are exposed to the solvent (Table 2), and can be

modelled as larger (mostly aromatic) residues without causing any structural strain (see Figure 2). These mutations are not expected to cause relevant changes in the pre-fusion structure, although they could have a destabilizing effect as a consequence of posing large aromatic residues in direct contact with the solvent instead of smaller apolar (leucine), polar (serine in 2 cases) or even charged (aspartate) residues. In addition, S940 involves its side-chain in a H-bond with the main-chain of D936, 4 residues upstream. The loss of this H-bond in the S940F mutant also points to a slight destabilization of the pre-fusion conformation. As for L938, it is buried in the pre-fusion conformation, pointing towards a three-stranded anti-parallel β-sheet made of the S711-P728 segment from the S1 subunit and of the Y1047-P1053 and G1059-A1078 segments from the S2 subunit, without directly contacting it (distances above 5 Å). It can also be modelled as a large phenylalanine without causing sterical strain. Upon mutation, it seems to optimize the hydrophobic interactions with the neighboring residues, especially I726 and A944.

Finally, S943 is located on a turn immediately downstream the helical segment hosting the above five mutations, between the second and third helical segments. The wild-type residue S943 features ϕ and ψ dihedral angles of 58.5° and 24.5°, respectively, which fall in an unfavourable region for prolines. In the S943P model we generated, the P943 ϕ and ψ dihedrals assume the values of 3.0° and 68.2°, placing the residue in an outlier region (39). The favoured ϕ angle for prolines is indeed restricted to the value of -63 ± 15°, (40) characteristic of α-helices. A proline at such a position would therefore introduce an anomaly in the pre-fusion conformation, while strongly promoting the transition to the post-fusion single continuous helical conformation. It is also worth noticing that this would be the only mutation among those we identified so far, introducing a proline residue in the SARS-CoV-2 S protein (Table S1). In light of the analysis of the GISAID May 29th updated, we also modelled the S943I mutation. Being isoleucine compatible with the S943 dihedral values, this mutation does not result in any structural strain.

Effect of the mutations on the protein post-fusion conformation

When looking at the post-fusion conformation of the SARS-CoV-2 spike protein S2 subunit, these mutations appear more revealing. Three of the wild-type residues, S929, D936 and S943, are indeed engaged in side-chain to side-chain H-bonds with

the HR2 segment of an adjacent monomer. In particular, S929, D936 and S943 (HR1 on Chain A) are H-bonded to S1196, R1185 and E1182, respectively (HR2 on Chain C, Figure 3). These are all strong H-bonds, especially the one between S943 and E1182, involving a negatively charged residue, and the one between D936 and R1185, being actually a salt bridge (estimated to contribute an additional 3-5 kcal/mol to the free energy of protein stability as compare to a neutral H-bond (41)). All these three H-bonds are lost upon mutation, which points to a weakening of the post-fusion assembly.

Of the remaining three mutations, S939F is completely exposed to the solvent and therefore, like in the pre-fusion conformation, expected to act unfavourable on the protein solvation energy. On the contrary, in case of L938F and S940F, which are substantially buried within the structure, mutation to a large aromatic phenylalanine seems even to optimize the network of the hydrophobic interactions; in case of F940, with the aliphatic parts of the side-chains of E1182 and R1185 on an adjacent monomer, and, in case of F938, with V1189 and A1190 on the same monomer and with other F938 residues on both the adjacent monomers.

When comparing the effect of the mutations on the pre- and post-fusion structures, it emerges that the S929I, D936Y and S943I mutations strongly destabilize the post-fusion conformation, while having a marginal impact on the stability of the pre-fusion one. On the contrary, S940F seems to favour the post-fusion conformation over the pre-fusion one. As for S938F and S939F, they seem to have a comparable effect on both the conformations, slightly stabilizing and destabilizing, respectively. Finally, the S943P mutation would strongly destabilize both the pre- and post-fusion conformations.

Conclusions

Based on a thorough analysis of the S protein sequences, that we extracted from the genomic sequences of SARS-CoV-2 reported in GISAID on April 21st, we identified the fusion core of the HR1 as a mutational hotspot. The D936Y and S943P mutations were the most numerous, being among the most frequently occurring mutations overall at the time. Other, less frequent, mutations were S939F and then S929I, L938F and S940F. Overall, such mutations appeared to be late ones, emerging starting from

the end of February or even mid March 2020, and were mainly localized in Europe and USA. Based on their frequency, on their location in a protein region playing a key role in the post-fusion conformation and also on the non-conservative nature of the mutations themselves, we decided to further investigate the structural basis of such mutations, finding out that they all can play a role in tuning the stability of the pre- and/or post-fusion S protein conformation.

A search of the GISAID dataset updated to May 29th revealed a ≈9-fold increase of the D936Y mutation over time (*versus* a ≈3-fold increase in the dataset sequences of 203%), thus indicating a possible positive selection for it. Notably, the D936Y variant represented the 20.5% of all the sequences reported from Sweden and the 2.7 and 1.4 % of all the sequences form Wales and England, respectively.

Other potentially interesting mutations are S929I and S939F, whose number of occurrences underwent a ≈2/3-fold increase. On the other hand, the increment in the occurrence of L938F and S940F was marginal, posing less emphasis on such mutations, which will be nonetheless useful to continue monitoring.

Finally, the S943P mutation, although still reported in few cases, underwent a dramatic reduction of occurrences, due to modification of the original sequences where they were first reported. At the same time, a S943I mutation emerged, that will also be worth continuing to monitor. We remind here that a proline at position 943 would cause a significant destabilization on the S protein pre-fusion conformation.

It is also worth noticing that the 2 mutations significantly increasing their frequency over time, D936Y and S929I, were also those that, together with S943P/I, caused the loss of a inter-monomer H-bond in the post-fusion conformation of the protein. Interestingly, the now emerging S943I mutation gets the same effect without destabilizing the pre-fusion conformation. The most frequently occurring mutation in the HR1 "fusion core", common in Sweden and UK on May 29th, is also the one causing the loss of a strong inter-monomer salt bridge. Our structural analyses provide a rationale for such mutations, pointing to a weakening of the post-fusion assembly. However, only experiments on cellular systems will clarify whether this may be a virus strategy for reducing its membrane fusion capacity, thus lowering its virulence.

Acknowledgements

We gratefully acknowledge all the Authors from the Originating laboratories responsible for obtaining the specimens and the Submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this research is based.

R.O. thanks MIUR-FFABR (Fondo per il Finanziamento Attività Base di Ricerca) for funding; L.C. acknowledge King Abdullah University of Science and Technology (KAUST) for support and the KAUST Supercomputing Laboratory for providing computational resources.

Authors' contribution

L.C. participated in the study's design and carried out the analyses. R.O. conceived of the study, participated in its design, carried out the analyses and drafted the manuscript. All authors read and approved the final manuscript.

Competing Interests

Authors declare no competing interests.

References

1. Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L. *et al.* (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, **579**, 270-273.
2. Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y. *et al.* (2020) A new coronavirus associated with human respiratory disease in China. *Nature*, **579**, 265-269.
3. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X. *et al.* (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*, **395**, 497-506.
4. Li, F. (2015) Receptor recognition mechanisms of coronaviruses: a decade of structural studies. *J Virol*, **89**, 1954-1964.
5. Belouzard, S., Chu, V.C. and Whittaker, G.R. (2009) Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. *Proc Natl Acad Sci U S A*, **106**, 5871-5876.
6. Millet, J.K. and Whittaker, G.R. (2014) Host cell entry of Middle East respiratory syndrome coronavirus after two-step, furin-mediated activation of the spike protein. *Proc Natl Acad Sci U S A*, **111**, 15214-15219.

7. Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A. and Li, F. (2020) Cell entry mechanisms of SARS-CoV-2. *Proc Natl Acad Sci U S A*, **117**, 11727-11734.
8. Salvatori, G., Luberto, L., Maffei, M., Aurisicchio, L., Roscilli, G., Palombo, F. and Marra, E. (2020) SARS-CoV-2 SPIKE PROTEIN: an optimal immunological target for vaccines. *J Transl Med*, **18**, 222.
9. Du, L., He, Y., Zhou, Y., Liu, S., Zheng, B.J. and Jiang, S. (2009) The spike protein of SARS-CoV--a target for vaccine and therapeutic development. *Nat Rev Microbiol*, **7**, 226-236.
10. Yi, C., Sun, X., Ye, J., Ding, L., Liu, M., Yang, Z., Lu, X., Zhang, Y., Ma, L., Gu, W. *et al.* (2020) Key residues of the receptor binding motif in the spike protein of SARS-CoV-2 that interact with ACE2 and neutralizing antibodies. *Cell Mol Immunol*, **17**, 621-630.
11. Tian, X., Li, C., Huang, A., Xia, S., Lu, S., Shi, Z., Lu, L., Jiang, S., Yang, Z., Wu, Y. *et al.* (2020) Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. *Emerg Microbes Infect*, **9**, 382-385.
12. Wang, C., Li, W., Drabek, D., Okba, N.M.A., van Haperen, R., Osterhaus, A., van Kuppeveld, F.J.M., Haagmans, B.L., Grosveld, F. and Bosch, B.J. (2020) A human monoclonal antibody blocking SARS-CoV-2 infection. *Nat Commun*, **11**, 2251.
13. McKee, D.L., Sternberg, A., Stange, U., Laufer, S. and Naujokat, C. (2020) Candidate drugs against SARS-CoV-2 and COVID-19. *Pharmacol Res*, **157**, 104859.
14. Amanat, F. and Krammer, F. (2020) SARS-CoV-2 Vaccines: Status Report. *Immunity*, **52**, 583-589.
15. Heald-Sargent, T. and Gallagher, T. (2012) Ready, set, fuse! The coronavirus spike protein and acquisition of fusion competence. *Viruses*, **4**, 557-580.
16. Li, F., Li, W., Farzan, M. and Harrison, S.C. (2005) Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science*, **309**, 1864-1868.
17. Wang, N., Shi, X., Jiang, L., Zhang, S., Wang, D., Tong, P., Guo, D., Fu, L., Cui, Y., Liu, X. *et al.* (2013) Structure of MERS-CoV spike receptor-binding domain complexed with human receptor DPP4. *Cell Res*, **23**, 986-993.
18. Walls, A.C., Tortorici, M.A., Snijder, J., Xiong, X., Bosch, B.J., Rey, F.A. and Veesler, D. (2017) Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. *Proc Natl Acad Sci U S A*, **114**, 11157-11162.
19. Elbe, S. and Buckland-Merrett, G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall*, **1**, 33-46.
20. Shu, Y. and McCauley, J. (2017) GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill*, **22**.
21. Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr*, **58**, 899-907.
22. Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.L., Abiona, O., Graham, B.S. and McLellan, J.S. (2020) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, **367**, 1260-1263.

23. Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T. and Veesler, D. (2020) Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*, **181**, 281-292 e286.
24. Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L. *et al.* (2020) Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, **581**, 215-220.
25. Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y. and Zhou, Q. (2020) Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science*, **367**, 1444-1448.
26. Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., Lu, G., Qiao, C., Hu, Y., Yuen, K.Y. *et al.* (2020) Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell*, **181**, 894-904 e899.
27. Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., Geng, Q., Auerbach, A. and Li, F. (2020) Structural basis of receptor recognition by SARS-CoV-2. *Nature*, **581**, 221-224.
28. Yuan, M., Wu, N.C., Zhu, X., Lee, C.D., So, R.T.Y., Lv, H., Mok, C.K.P. and Wilson, I.A. (2020) A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science*, **368**, 630-633.
29. Xia, S., Liu, M., Wang, C., Xu, W., Lan, Q., Feng, S., Qi, F., Bao, L., Du, L., Liu, S. *et al.* (2020) 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 Res*, **30**, 343-355.
30. Ren, L.L., Wang, Y.M., Wu, Z.Q., Xiang, Z.C., Guo, L., Xu, T., Jiang, Y.Z., Xiong, Y., Li, Y.J., Li, X.W. *et al.* (2020) Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. *Chin Med J (Engl)*, **133**, 1015-1024.
31. Xia, S., Zhu, Y., Liu, M., Lan, Q., Xu, W., Wu, Y., Ying, T., Liu, S., Shi, Z., Jiang, S. *et al.* (2020) Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. *Cell Mol Immunol*.
32. Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150-3152.
33. Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, **234**, 779-815.
34. Feyfant, E., Sali, A. and Fiser, A. (2007) Modeling mutations in protein structures. *Protein Sci*, **16**, 2030-2041.
35. DeLano Scientific, L. (2002) http://www.pymol.org.
36. Vangone, A., Spinelli, R., Scarano, V., Cavallo, L. and Oliva, R. (2011) COCOMAPS: a web application to analyze and visualize contacts at the interface of biomolecular complexes. *Bioinformatics*, **27**, 2915-2916.
37. Eaaswarkhanth, M., Madhoun, A.A. and Al-Mulla, F. (2020) Could the D614 G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality? *Int J Infect Dis*.
38. Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Foley, B., Giorgi, E.E., Bhattacharya, T., Parker, M.D. *et al.* (2020) Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv*, 2020.2004.2029.069054. PREPRINT: NOT PEER REVIEWED

39. Lovell, S.C., Davis, I.W., Arendall, W.B., 3rd, de Bakker, P.I., Word, J.M., Prisant, M.G., Richardson, J.S. and Richardson, D.C. (2003) Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins*, **50**, 437-450.
40. MacArthur, M.W. and Thornton, J.M. (1991) Influence of proline residues on protein conformation. *J Mol Biol*, **218**, 397-412.
41. Anderson, D.E., Becktel, W.J. and Dahlquist, F.W. (1990) pH-induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry*, **29**, 2403-2408.

FIGURE LEGENDS

Figure 1. *Top*: Cartoon representation of the SARS-CoV-2 S protein HR1 and its fusion core (insets) in the pre- and post-fusion conformations (PDB IDs: 6VSB and 6LXT). Discussed mutations are colored in a purple-to-pink scale, depending on their frequency, and labelled. *Bottom*: Sequence alignment of the HR1 fusion core (framed) and 10 residues up- and down-stream in the S protein of SARS-CoV-2, bat coronavirus RaTG13 (protein_ID: QHR63300.2) and SARS-CoV (protein_ID: AAP13441.1).

Figure 2. Models of mutants in the pre-fusion conformation. *Right:* Cartoon representation of the SARS-CoV-2 S protein in its pre-fusion trimeric conformation (the three monomers are colored in silver, gold and copper; PDB ID: 6VSB), with the structure of the RBD bound to the ACE2 receptor (in blue; PDB ID: 6M0J) superimposed on its chain A. All the mutated positions we identified in GISAID on April 21st are colored on a purple-to-pink scale, depending on their relative frequency. Mutations in the HR1 fusion core are shown in a dots representation for chain A. *Left:* Focus on the structural context of each wild-type residue (silver sticks) and corresponding mutant (purple-to-pink sticks). Contacting residues (within 5 Å) are shown in a dots representation and H-bonds are shown as red dashed lines.

Figure 3. Models of mutants in the post-fusion conformation. *Right:* Cartoon representation of the SARS-CoV-2 S protein in its post-fusion trimeric conformation (the three monomers are colored in silver, gold and copper; PDB ID: 6LXT). The color code is the same of Figure 2. Mutations in the HR1 fusion core are shown in a dots representation for chain A. *Left:* Focus on the structural context of each wild-type residue (silver sticks) and corresponding mutant (purple-to-pink sticks). Contacting residues (within 5 Å) are shown in a dots representation and H-bonds are shown as red dashed lines.

Table 1. Occurrences of mutations on the HR1 "fusion core".

	# Genomic/S protein sequences	S929I	D936Y	L938F	S939F	S940F	S943P
April 21st	10,823[a]/7,692[b]	2	25	2	9	3	22
May 29th	34,805/23,332	5[c]	213	3	20	5	3[d]
% delta	233/203	150	752	50	122	67	-633

[a] As downloaded from GISAID.

[b] Complete S proteins we extracted and translated (no undefined nucleotide, no internal stop codon, no insertion/deletion).

[c] In 2 additional sequences it is mutated to T.

[d] In 3 additional sequences it is mutated to I.

Table 2. Solvent accessibility of mutated residues in the pre- and post-fusion conformations.

Amino acid	Pre-fusion	Post-fusion
I929	exposed	partly buried (18.6%)[a]
Y936	exposed	partly buried (19.0%)
F938	buried (95.8%)	buried (95.3 %)
F939	exposed	exposed
F940	exposed	buried (62.3 %)

[a] Percentage of buried surface upon complex formation.

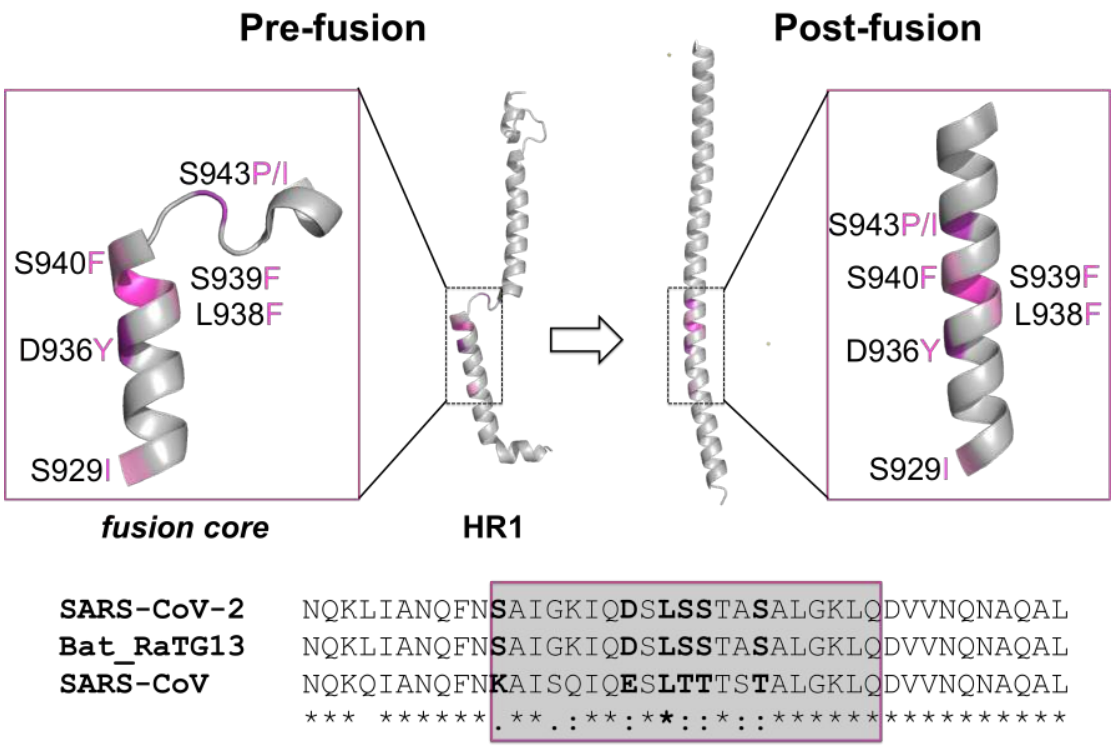

Figure 1

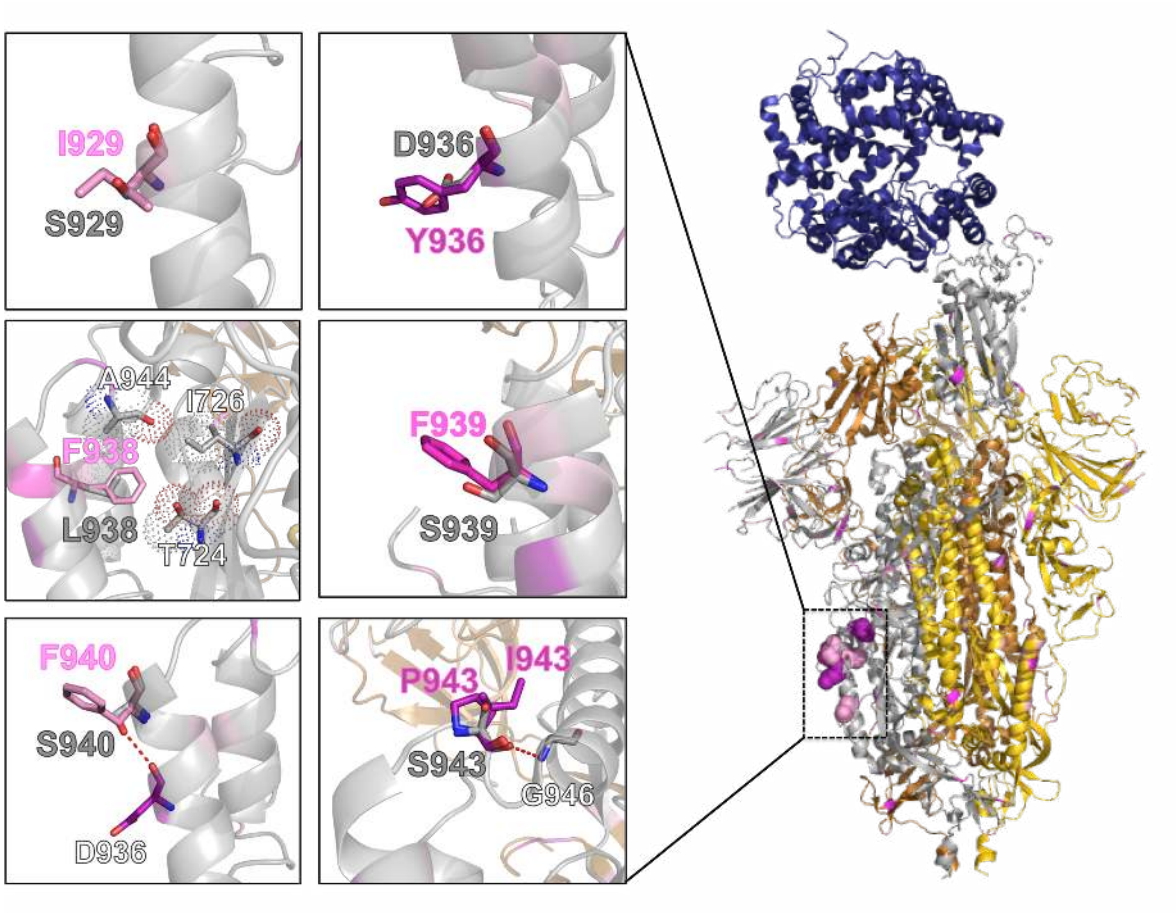

Figure 2

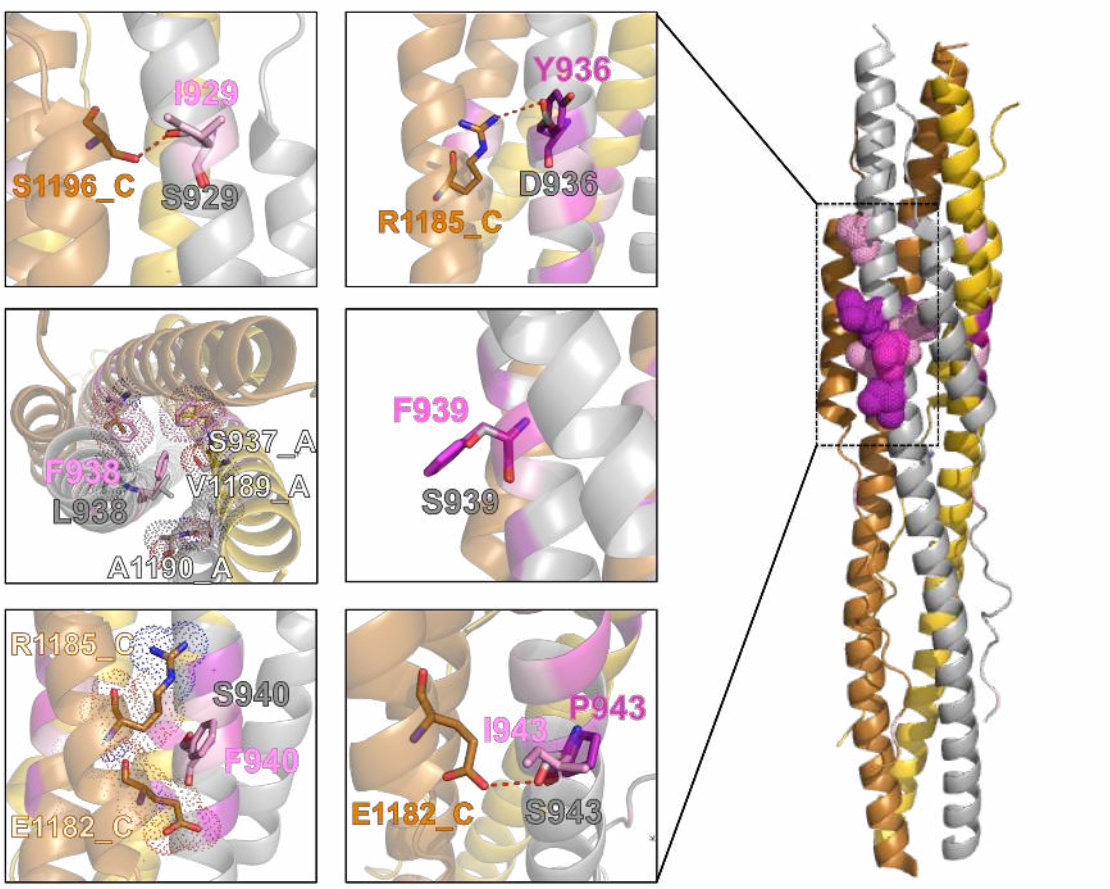

Figure 3

Table S1. List of mutations identified in GISAID in at least 2 identical sequences on April 21st 2020, in sequential order.

	SAP	Total sequences[a]	Structural domain[b]
1	L5F	44	SP
2	L8V	14	SP
3	P9S	2	SP
4	L18F	3	NTD
5	R21I	8	NTD
6	T22I	2	NTD
7	A27S	2	NTD
8	A27V	2	NTD
9	T29I	7	NTD
10	H49Y	11	NTD
11	S50L	4	NTD
12	L54F	4	NTD
13	W64L	2	NTD
14	H69Y	2	NTD
15	S71F	4	NTD
16	D80Y	2	NTD
17	V90F	2	NTD
18	T95I	2	NTD
19	S98F	5	NTD
20	V120I	3	NTD
21	D138H	4	NTD
22	D138Y	3	NTD
23	G142A	2	NTD
24	Y145H	3	NTD
25	H146R	2	NTD
26	H146Y	2	NTD
27	M153T	2	NTD
28	F157S	2	NTD
29	M177I	2	NTD
30	G181V	2	NTD
31	I197V	2	NTD
32	R214L	2	NTD
33	D215H	2	NTD
34	L216F	2	NTD
35	S221L	4	NTD
36	Q239K	8	NTD
37	A243S	2	NTD
38	S247R	3	NTD
39	S254F	4	NTD
40	S255F	3	NTD
41	W258L	2	NTD
42	G261V	4	NTD
43	A262T	4	NTD
44	Q271R	2	NTD
45	E281V	2	NTD
46	A288S	2	NTD
47	F338L	3	RBD
48	V367F	11	RBD
49	Q414E	6	RBD

50	S438F	2	RBM
51	N439K	2	RBM
52	G476S	7	RBM
53	V483A	22	RBM
54	A520S	2	RBD
55	A522S	2	RBD
56	A522V	3	RBD
57	E583D	2	
58	G594S	2	
59	L611F	3	
60	D614G	4,404	
61	V615I	8	
62	A626V	5	
63	P631S	2	
64	H655Y	6	
65	Q675H	15	
66	Q677H	4	
67	A706V	6	
68	I714L	2	
69	R765H	4	
70	R765L	4	
71	T791I	6	FP
72	P809S	3	
73	L821I	2	IFP
74	A831V	28	IFP
75	D839Y	22	
76	A845S	4	
77	A846V	4	
78	A852V	3	
79	N856S	2	
80	A879S	10	
81	S929I	2	HR1
82	D936Y	24	HR1
83	L938F	2	HR1
84	S939F	8	HR1
85	S940F	2	HR1
86	S943P	20	HR1
87	A1020V	2	
88	V1040F	2	
89	L1063F	2	
90	V1065L	2	
91	A1078S	5	CD
92	D1084Y	3	CD
93	P1112S	2	CD
94	G1124V	19	CD
95	P1143L	2	
96	P1162L	4	
97	D1163G	2	HR2
98	D1165G	2	HR2
99	V1176F	3	HR2
100	L1203F	4	HR2
101	I1216T	2	TM
102	G1219C	2	TM
103	G1219V	2	TM
104	V1228L	2	TM

105	M1229I	5	TM
106	V1230L	2	TM
107	M1237I	5	TM
108	C1247F	2	CT
109	C1254F	6	CT
110	D1260N	3	CT
111	P1263L	42	CT

[a] Counts all the sequences where the mutation was found, alone or in combination with other mutations. Unique sequences (having no identical sequence in the dataset) are not included in this analysis.

[b] SP: signal peptide, NTD: N-terminal domain, RBD: receptor-binding domain, RBM: receptor-binding motif, FP: fusion peptide, IFP: internal fusion peptide, HR1: heptad repeat 1, CD: connector domain, HR2: heptad repeat 2, TM: transmembrane domain, CP: cytoplasmic tail.