Evolution of SARS-CoV-2 spike glycoprotein

Antoni Wrobel (antoni.wrobel@crick.ac.uk)
Francis Crick Institute
Donald Benton (donald.benton@crick.ac.uk)
Francis Crick Institute
Pengqi Xu
Precision Medicine Center, The Seventh Affiliated Hospital, Sun Yat-sen University
Chloë Roustan
Francis Crick Institute
Stephen Martin
Francis Crick Institute
Peter Rosenthal
Francis Crick Institute
John Skehel
Francis Crick Institute
Steven Gamblin (steve.gamblin@crick.ac.uk)
Francis Crick Institute

Research Article

Keywords: cryoEM, Virology, Structural Biology, SARS-CoV-2

Posted Date: May 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-29398/v1

Version of Record: A version of this preprint was published on July 9th, 2020. See the published version at https://doi.org/10.1038/s41594-020-0468-7.

Abstract

The spike glycoprotein (S) of SARS-CoV-2 mediates attachment of the virus to cell surface receptors and fusion between virus and cell membranes[1]. The receptor for SARS-CoV-2, like that for SARS-CoV, is the human cell-surface membrane protein ACE2[2–4]. Membrane fusion activity, as for other class-1 fusion glycoproteins, requires S to be proteolytically cleaved into S1 and S2 that remain associated following cleavage[4–7]. SARS-CoV-2 is thought to have emerged from bats, possibly via a secondary host[8,9]. To better understand the transmission of SARS-CoV-2 we have determined the structure of its furin-cleaved S by cryoEM, which shows that cleavage at this polybasic amino-acid site increases the structural plasticity of the receptor binding region and facilitates the adoption of an open conformation that is required for it to bind to the ACE2 receptor. To investigate relationships between S proteins of SARS-CoV-2 and of the most closely related bat virus, RaTG13[8], we have determined and compared their structures and characterised biochemically their affinities for ACE2 and their relative stabilities. Whilst the overall structures are similar, there are key differences likely pertinent to virus infectivity. These include a more stable pre-cleavage form of human S, about 1000-fold tighter binding of SARS-CoV-2 to human receptor, and a higher proportion of S in the conformation required for binding ACE2 upon protease cleavage.

Main Text

Coronaviruses infect a range of mammalian and avian species1. SARS-CoV-2, the agent of the COVID-19 pandemic8,9, belongs to Sarbecovirus subgenus of betacoronaviruses, members of which mostly infect bats10,11. Hence, bat coronaviruses were identified as a likely evolutionary precursor of SARS-CoV-28,9. It remains unknown how SARS-CoV-2 could have evolved to infect humans but two mechanisms have been proposed: selection in an animal host before zoonotic transfer (possibly via an intermediate host such as the Malayan pangolin (*Manis javanica*)12); or natural selection in humans following a zoonotic transfer including direct transmission from bats13. In order to better understand this zoonosis, we have characterised the spike protein of SARS-CoV-2 and its closest relative RaTG13. In addition to substitutions in the RBD, a second difference between the spike proteins from human and bat viruses is the presence of a four-amino-acid polybasic cleavage site, PRRA, between S1 and S2 domains2. Similar cleavage sites have been found in related coronaviruses including HKU1 and MERS that infect humans6,7,14 and their acquisition is associated with increased pathogenicity in other viruses such as influenza15.

To examine the evolutionary origin of SARS-CoV-2 we first characterised the spike (S) protein of the furin-cleaved human pandemic SARS-CoV-2 virus by cryoEM (**Fig. 1a**). We produced a form of the human S protein with the furin-cleavage site intact. This protein, which we expressed in mammalian cells, was secreted in a partially cleaved form, presumably owing to the naturally expressed proteases within these cells6 (**Fig. S1A**). We further cleaved this protein using exogenous furin for structural and biochemical characterisation (**Fig. S1A**). The particles analysed from cryo-electron micrographs fell into three populations; a closed form (34%), an intermediate form (39%) and an open form (27%) with an upright

Receptor Binding Domain (RBD) (**Fig. 1a**). The overall structure of the closed conformation of the S trimer is three-fold symmetric and similar to structures described previously using uncleaved material2,16 (**Fig. 1a**). In the closed conformation, the surface of the RBD, which would interact with the ACE2 receptor, is buried inside the trimer and not accessible for receptor binding. In the intermediate form (**Fig. 1a**) two of the three RBDs maintain a similar interaction to the closed form but the third RBD displays increased mobility and has shifted slightly away from the trimer axis. In the open form (**Fig. 1a**), two of the RBDs remain fairly closely associated, as in the closed and intermediate forms but, the third RBD rotates approximately ~60˚ such that the ACE2 interacting surface is now fully exposed at the top of the assembly. The changes in domain orientations between the closed and open forms are shown for a selected monomer in **Fig. 1b**.

In this protease-cleaved material, there is a higher proportion of the S proteins in an open conformation: 27% compared to 17% in the uncleaved human S trimer described below. The observation here of a substantially populated (39%) intermediate form, where one of the RBDs has separated from the other two of the trimer, also suggests that this conformation, possibly transient, will also lead to a receptor-binding competent form. Thus, protease cleavage is likely a selected feature of the human virus in that it leads to a higher proportion of S proteins on the virus surface capable of binding to receptor. Although the loop containing the cleavage site (residues 676-689) is disordered, in both cleaved and uncleaved forms, cleavage likely introduces additional conformational plasticity in this part of the structure. This plasticity is propagated through the molecule by successively larger domain rearrangements resulting, finally, in the facilitation of the ~60˚ rotation of the RBD.

Next, we determined the cryoEM structure of S from the closest known bat virus (RaTG13) as well as uncleaved human S (**Fig. S1B**). The bat protein was expressed in mammalian cells but was found to be unstable during preparation of EM grids and required chemical cross-linking to produce particles for data collection and analysis. The resulting micrographs yielded a high-resolution single particle reconstruction at 3.1 Å resolution. The uncleaved human S sample was particularly stable and gave rise to the best quality density maps at 2.6 Å (**Fig. S2**), enabling us to model 15% more of the receptor binding domain (RBD, 100% complete) and 25% more of the N-terminal domain (NTD, 98% complete) than earlier studies2,16, which impacts the appearance of the trimer. The overall structure of the bat S protein is similar to that of the uncleaved human closed form (**Fig. 2a, d**). Presumably the chemical crosslinking required to image bat S is responsible for all particles being in the closed conformation. Comparison of the sequences of this bat S protein with the human one reveals a high degree of conservation overall (97.8% in the ectodomain) but with a relatively high proportion of substitutions in the RBD (89.6% identity) (**Fig. 2b**). The substitutions are clustered at two interfaces; the ACE2 receptor binding surface (considered below), and the RBD/RBD interfaces of the trimeric S. Analysis of the latter interface in the human trimer reveals an extensive network of potential intra-trimer hydrogen bonds; including Arg-403, Gln-493 and Tyr-505 from one subunit interacting with Ser-373, Ser-371 and Tyr-369 from another (**Fig. 2c**). The corresponding residues in the bat structure, and other inter-subunit contacts, suggest a lower surface complementarity. Of note, the bat S protein has an N-glycosylation site at Asn-370, where a bulky fucosylated glycan wedges between adjacent domains (**Fig. S3**). Indeed, surface contact area

calculations show that in the bat S trimer, the monomer/monomer interactions account for 5200 Å2 (of which 485 Å2 between the RBDs) while the equivalent contact area in the closed structure of the SARS-CoV-2 S trimer is 6100 Å2 (with 550 Å2 between the RBDs). To further investigate the relative stability of the human and bat S trimers we carried out thermal denaturation experiments (**Fig. S1c**). These data show that the uncleaved human S trimer has a markedly higher thermal stability than the bat protein while the cleaved human protein has a similar stability to the (uncleaved) bat protein. Perhaps the higher stability of human S is required to offset some of the loss of stability that occurs upon cleavage. These structural and biochemical data together suggest that the human virus acquired an advantage by having a polybasic cleavage site, which facilitates a higher proportion of the open, receptor binding competent, conformation.

As outlined above, the second region with a high proportion of sequence differences between the bat and human RBDs is at the receptor binding site. To quantitate the impact of these differences on binding to the human ACE2 receptor we measured binding with surface biolayer interferometry. Spike protein, either human or bat, was immobilised onto the surface of a sensor and purified ACE2 was flowed over the surface to measure binding. Amplitude analysis suggests that the human S has approximately 1000-fold stronger binding to ACE2 than the bat protein with Kd values of <100 nM and >40 μM respectively (**Fig. 3a**).

Previous studies have determined the structural interaction of the isolated RBD of SARS-CoV-2 S with human ACE217,18. Using this information (PDB: 6VW117) enables us to model and compare the ACE2 domain bound to our human and bat S trimers. In the case of the human S/ACE2 there is a buried surface area of 840 Å2. As well as a series of specific salt and hydrogen bonds, another notable feature is that Phe- 486(HS) inserts into a hydrophobic pocket on the surface of ACE2 formed by residues including Phe-28(ACE2), Leu-79(ACE2), Met-82(ACE2) and Tyr-83(ACE2). In contrast, in the bat spike protein, the hydrophobic Phe-486 is replaced by a less-bulky leucine residue (Leu-486(BS)) (**Fig. 3b**), which accounts in part for the smaller buried surface of the bat S/ACE2 of 760 Å2. Structural comparison also suggests another substitution that likely contributes to the greatly enhanced affinity of human S binding to ACE2; Gln-493(HS) makes a potential hydrogen bond with Glu-35(ACE2) that is salt bridged to Lys-31(ACE2) that in turn salt bridges with Glu-484(HS). In contrast, the equivalent residue to Gln- 493(HS) in the bat is a tyrosine that sterically clashes with Lys-31(ACE2) and does not hydrogen bond to Glu-35(ACE2), while Glu-484(HS) is replaced by a threonine that would not bond to Lys-31(ACE2) (**Fig. 3c**). Moreover, the glutamine at position 498 is replaced by a Thr-234(BS) that could not hydrogen bond to Tyr-41(ACE2).

Together, our structural and biochemical data indicate that a bat virus, similar to RaTG13, would not be able to bind effectively to ACE2 receptor and would be unlikely to infect humans directly. Given the modular nature of the human and bat spike glycoproteins, and the number and structural locations of the amino-acid sequence differences between them, our observations could be interpreted to support the involvement of recombination12 between distinct coronavirus genomes in the generation of SARS-CoV-2. The structure of the SARS-CoV-2 spike protein presented here is at high resolution, and nearly complete with many more external loops included and thus provides important insights for vaccine design. Further,

our study suggests that the presence of the polybasic cleavage site in the S of SARS-CoV-2 leads to enhanced virus transmissibility, as it increases the proportion of RBDs on the virus surface able to bind receptor.

Declarations

Author Contributions

A.G.W., D.J.B, P.X., C.R., S.R.M. performed research, collected and analysed data; A.G.W, D.J.B, P.B.R, J.J.S, S.J.G conceived and designed research and wrote the paper.

Conflict Statement

We have no conflicts of interest to declare.

Data Availability

Maps and models have been deposited in the Electron Microscopy Data Bank, http://www.ebi.ac.uk/pdbe/emdb/ (Accession Nos. XXX,XXX,...). Models have been deposited in the Protein Data Bank, https://www.ebi.ac.uk/pdbe/ (PDB ID codes XXX, XXX, ...). [Accession numbers will be available before publication].

Acknowledgements

We would like to acknowledge Andrea Nans of the Structural Biology Science Technology Platform for assistance with data collection, Phil Walker and Andrew Purkiss of the Structural Biology Science Technology Platform and the Scientific Computing Science Technology Platform for computational support. We thank Peter Cherepanov, George Kassiotis and Svend Kjaer for discussions. This work was funded by the Francis Crick Institute which receives its core funding from Cancer Research UK (FC001078 and FC001143), the UK Medical Research Council (FC001078 and FC001143), and the Wellcome Trust (FC001078 and FC001143). P.X. was also supported by 100 Top Talents Program of Sun Yat-sen University.

Referneces

1. Lai, M., Perlman, S. & Anderson, L. Coronaviridae. in *Fields Virology* 1305– 1335 (2007).
2. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. (2020). doi:10.1016/j.cell.2020.02.058
3. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Microbiol.* **5**, 562–569 (2020).
4. 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).

5. Belouzard, S., Chu, V. C. & Whittaker, G. R. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. *Natl. Acad. Sci. U. S. A.* **106**, 5871–5876 (2009).
6. Millet, J. K. & Whittaker, G. R. Host cell proteases: Critical determinants of coronavirus tropism and pathogenesis. *Virus Res.* **202**, 120–134 (2015).
7. Millet, J. K. & Whittaker, G. R. Host cell entry of Middle East respiratory syndrome coronavirus after two-step, furin-mediated activation of the spike protein. *Natl. Acad. Sci. U. S. A.* **111**, 15214–15219 (2014).
8. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
9. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
10. Li, W. *et al.* Bats are natural reservoirs of SARS-like coronaviruses. *Science (80-.).* **310**, 676–679 (2005).
11. Yang, L. *et al.* Novel SARS-like betacoronaviruses in bats, China, *Emerg. Infect. Dis.* **19**, 989–991 (2013).
12. Lam, T. T. Y. *et al.* Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature* 1–6 (2020). doi:10.1038/s41586-020-2169-0
13. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature Medicine* **26**, 450–452 (2020).
14. Chan, C. M. *et al.* Spike protein, S, of human coronavirus HKU1: Role in viral life cycle and application in antibody detection. *Biol. Med.* **233**, 1527– 1536 (2008).
15. Steinhauer, D. A. Role of hemagglutinin cleavage for the pathogenicity of influenza virus. *Virology* **258**, 1–20 (1999).
16. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science (80-.).* (2020). doi:10.1126/science.aax0902
17. Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature* 1–4 (2020). doi:10.1038/s41586-020-2179-y
18. Yan, R. *et al.* Structural basis for the recognition of SARS-CoV-2 by full- length human ACE2. *Science (80-.).* **367**, 1444–1448 (2020).

Tables

Cryo-EM data collection, refinement and validation statistics

	Uncleavable Closed (EMDB-xxxx) (PDB xxxx)	Intermediate (EMDB-xxxx) (PDB xxxx)	Cleaved Open (EMDB-xxxx) (PDB xxxx)	Bat (EMDB-xxxx) (PDB xxxx)
Data collection and processing				
Voltage (kV)	300	300	300	300
Electron exposure (e–/Å2)	33.6	54.4	54.4	54.4
Defocus range (μm)	-1.5 to -3.0	-1.5 to -3.0	-1.5 to -3.0	-1.5 to -3.0
Pixel size (Å)	1.09	1.08	1.08	1.08
Symmetry imposed	C3	C1	C1	C3
Final particle images (no.)	95 k	107 k	73 k	62 k
Map resolution (Å)	2.6	6.8	3.8	3.1
FSC threshold = 0.143 Map resolution range (Å)	2.4-3.2	6-10	3.5-7.5	3.0-3.8
Refinement Initial model used (PDB code)	6VXX	-	-	-
Model resolution (Å)	2.7	8.2	4.1	3.2
FSC threshold = 0.5				
Map sharpening *B* factor (Å2)	-109.8	-207.0	-61.7	-88.2
Model composition				
Non-hydrogen atoms	26991	25776	25776	26169
Protein residues	3294	3294	3294	3180
Ligands	81	-	-	93
B factors (Å2)				
Protein	29.5	-	-	37.13
Ligand	60.0	-	-	52.15
R.m.s. deviations				
Bond lengths (Å)	0.007	0.003	0.003	0.007
Bond angles (°)	0.819	0.853	0.861	0.828
Validation				
MolProbity score	1.42	1.70	1.79	1.70
Clashscore	2.96	3.28	3.79	5.16
Poor rotamers (%)	0.73	0.91	0.98	0.75
Ramachandran plot				
Favored (%)	95.22	88.82	87.07	93.57
Allowed (%)	4.69	10.78	12.41	6.43
Disallowed (%)	0.09	0.40	0.52	0.00

Figures

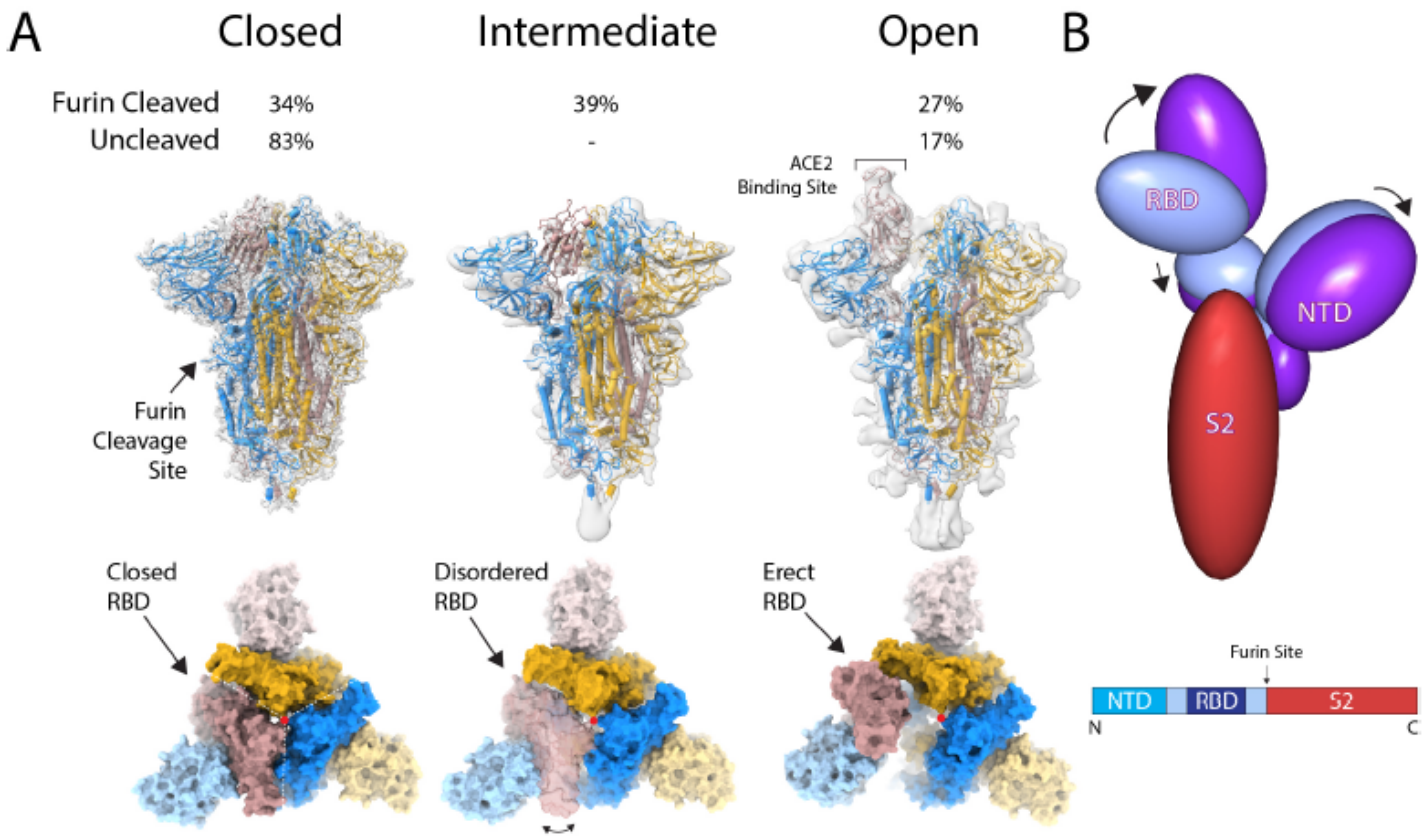

Figure 1

Structure of protease-cleaved SARS-CoV-2 spike glycoprotein. (A) Three structures are calculated from micrographs of furin-cleaved material: closed, intermediate and open forms of approximately equal proportions. In un-cleaved material most of the population represents the closed form with a small proportion in the open conformation. Density maps for the three types of particles, overlaid with a ribbon representation of the built molecular models, as viewed with the long axis of the trimer vertical (middle panel); the three monomers are coloured blue, yellow and brown. An orthogonal view (lower panel) looking down the long axis (indicated by a red dot), the colouring is as in the middle panel with the NTDs in a lighter hue. (B) The changes in domain orientation, between the closed and open forms, shown schematically, for the monomer that undergoes the most substantive change in the RBD position. The image is produced by the CCP4 MG 'bloboid representation' and is calculated from the shape and centre of mass of the molecular model. Also shown is a bar representation of the domains with the protease cleavage site indicated.

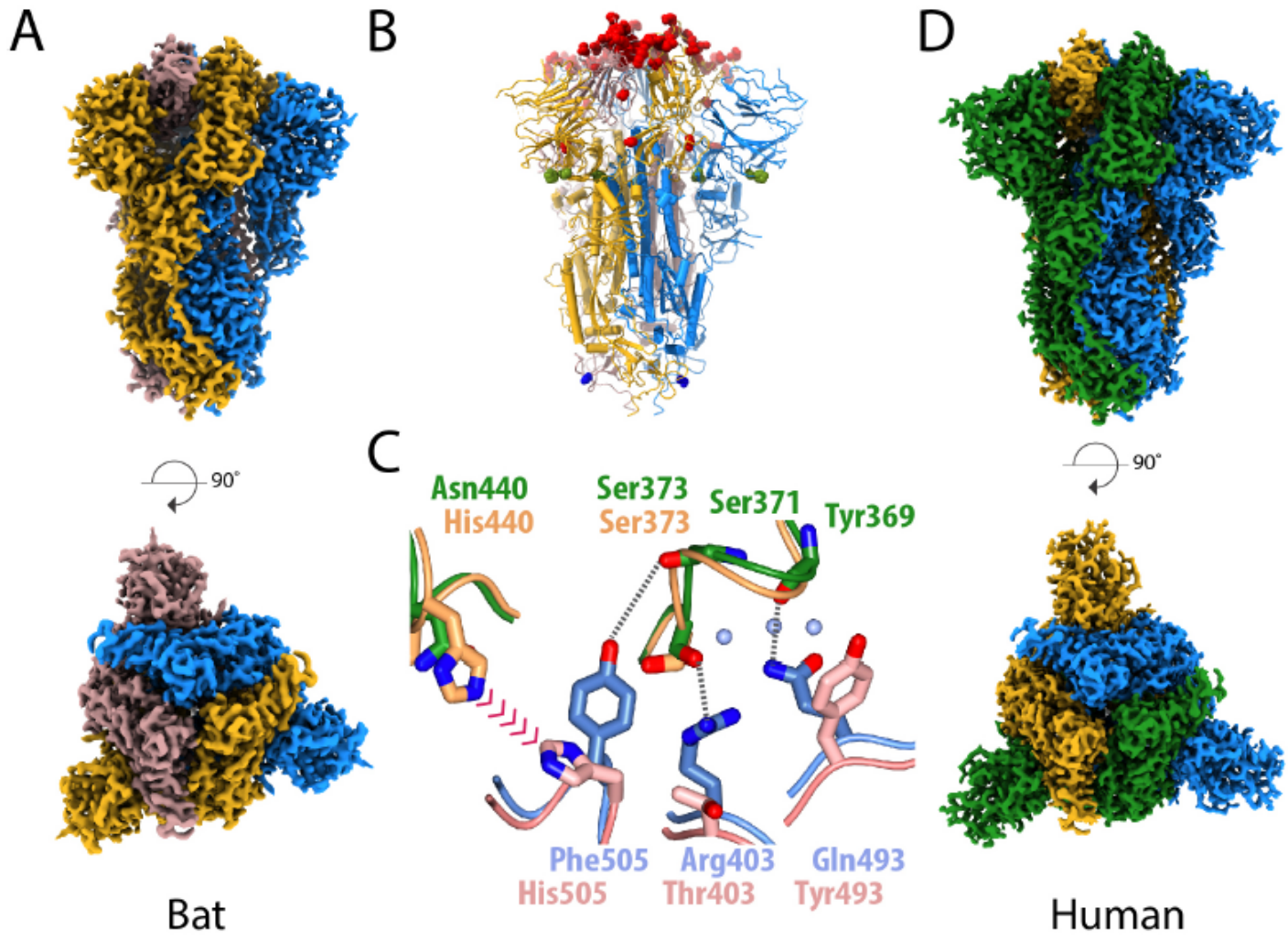

Figure 2

Structural comparison of the spike glycoprotein from the bat virus RaTG13 and human SARS-CoV-2. (A) The density map for the bat trimeric spike is shown with the long axis vertical in the top panel and viewed orthogonally in the bottom panel. All of the particles are in the closed conformation likely because of the cross-linking of the material. The three monomers are coloured blue, yellow and brown. (B) Molecular model of the bat spike protein, coloured as in (A), with substitutions between the bat and human SARS-CoV-2 highlighted. Most of the changes are in the RBD and coloured red, there are four substitutions in S1 outside of the RBD, which are shown in green, and a single substitution in S2 shown in blue. (C) Overlay of the molecular structure of a portion of the RBD/RBD interface; the two bat monomers are coloured gold (upper) and pink (lower) while the two superposed human RBD chains are shown in green (upper) and blue (lower). Analysis suggests that the residues at the interface of the human RBD chains support several additional stabilising interactions and avoid potential steric repulsion between His-505 & His-440 seen in the bat structure. (D) The density map for the uncleaved human S protein, in the closed conformation, shown in the same orientation as (A) with the subunits coloured blue, green and yellow. This sample gave the best quality maps and enabled the most extensive build of the polypeptide chain.

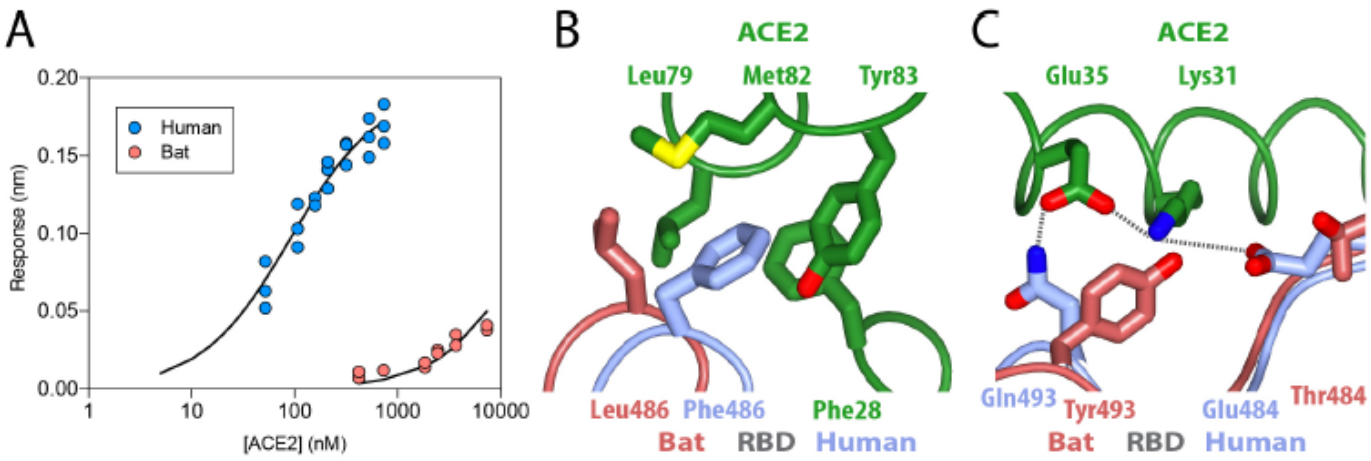

Figure 3

Binding of ACE2 receptor to bat and SARS-CoV-2 spike protein. (A) Plot of surface biolayer amplitude measurement as a function of ACE2 concentration with the data for the human coloured blue and the bat red. The equilibrium dissociation constant was calculated to be 91 ± 18 nM for the human protein and estimated to be >40 μM for the bat protein. The equilibrium dissociation constant for the human protein calculated from kinetic constants (koff = 0.0105 s-1 and kon = 1.56 x 105 M-1s-1) was 67.5 +/- 9 nM. (B & C) Ribbon representation of modelled molecular interactions between ACE2 (green) with human (blue) (both PDB ID: 6VW117) and bat (brown) RBD (from this study). (B) Details of a hydrophobic pocket on ACE2 that accommodates a phenylalanine residue from the human S RBD. (C) Shows two salt bridges and a charged hydrogen bond linking human S RBD to ACE2, while the interface with bat S RBD is not able to make these interactions and presents a potential steric clash between Tyr-493(BS RBD) with Lys-31(ACE2).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalFigures.pdf