Evolutionary relationships and sequence-structure determinants in human SARS coronavirus-2 spike proteins for host receptor recognition

Lalitha Guruprasad*, School of Chemistry, University of Hyderabad, Hyderabad 500046, India.

****Email: lalitha.guruprasad@uohyd.ac.in***

Keywords: Severe acute respiratory syndrome coronavirus-2, complete genomes, spike proteins, multiple sequence alignment, phylogenetic tree, receptor binding domain

Abstract:

Coronavirus disease 2019 (COVID-19) is a pandemic infectious disease caused by novel Severe Acute Respiratory Syndrome coronavirus-2 (SARS CoV-2). The SARS CoV-2 is transmitted more rapidly and readily than SARS CoV. Both, SARS CoV and SARS CoV-2 via their glycosylated spike proteins recognize the human angiotensin converting enzyme-2 (ACE-2) receptor. We generated multiple sequence alignments and phylogenetic trees for representative spike proteins of CoV and CoV-2 from various host sources in order to analyze the specificity in SARS CoV-2 spike proteins required for causing infection in humans. Our results show two sequence regions in the N-terminal domain; "M_{153}ESEFR_{158}" and "S_{247}YLTPG_{252}" that are specific to human SARS CoV-2 and pangolin SARS CoV. In the receptor binding domain (RBD), we report the identification of three sequence loops; V_{445}GGNY_{449} (loop 1), Y_{473}QAGSTPC_{480} (loop 2) and E_{484}GFNCY_{489} (loop 3) in human SARS CoV-2 (NCBI Accession: QHD43416.1) and the equivalent loops; S_{432}TGNY_{436}, F_{460}SPDGKPC_{467} and A_{471}LNCY_{475} in human SARS CoV (NP_828851). A disulfide bridge tethers loops 2 and 3 in SARS CoV-2 and SARS CoV. From our analyses, these loop insertions and the tethered disulfide bridge are present only in the spike proteins of SARS CoV and SARS CoV-2 that recognize ACE-2 receptor and are absent from other SARS CoV. In SARS CoV-2 and SARS CoV, the loops 2 and 3 are connected by NGV and TPP sequences, respectively, that are associated with a beta-turn conformation. The complete genome analysis of representative SARS CoVs from different host sources; bat, civet, pangolin, human and the human SARS CoV-2, identified the bat genome (GenBank code: MN996532.1) and the pangolin SARS CoV genomes as being closest to the novel human SARS CoV-2 genomes. The bat CoV genomes (GenBank codes: MG772933 and MG772934) are evolutionary intermediates in the mutagenesis progression towards becoming human SARS CoV-2.

Introduction:

In the last two decades, zoonotic coronaviruses, Severe Acute Respiratory Syndrome coronavirus SARS CoV (2002) (Dorsten et al., 2003) and Middle East Respiratory Syndrome coronavirus (MERS CoV) (2012) (Azhar et al., 2014a) have caused acute respiratory diseases in humans that have resulted in several deaths. The present coronavirus disease 2019 (COVID-19) is a pandemic respiratory disease caused by the novel SARS CoV-2. The initial infection started in Wuhan, Hubei province, China in December 2019 and very soon became a global outbreak, infecting populations in almost every country in the world causing a total of 5,951,004 coronavirus cases and 363,023 deaths as of 29th May 2020 (https://www.worldometers.info/coronavirus/). Within a short span of time this pandemic has caused major social and economic disruptions. Compared to other coronaviruses, the novel SARS CoV-2 appears to be spreading more rapidly and readily, posing a challenging task before the administrative and scientific communities. The SARS CoV-2 is transmitted from person to person contact via respiratory secretions during coughing and sneezing. Infection of this highly pathogenic virus can cause acute respiratory distress syndrome which impacts the lung and heart functions. The prominent symptoms of this viral infection are flu, severe respiratory, enteric and neurological disorders, resulting in increased white blood cells and kidney failure. There are no vaccines or drugs available to combat this deadly infectious disease and there is no strategic plan to treat the infected patients. Hence, there is a need to develop specific anti-CoV-2 vaccines and drugs to treat infected patients so as to reduce viral shedding and further transmission in populations.

The SARS CoV-2 comprises positive-sense single-stranded RNA genome of size 29-30 kb and belongs to the coronaviridae family and betacoronavirus sub-family. Mammals such as bats are the main reservoir of betacoronaviruses, but due to the zoonotic contacts and viral genomic mutations, SARS CoV-2 has recently crossed species and caused infections in humans (Wu et al., 2020). Research findings have pointed that previous zoonotic CoV infections such as, SARS CoV, that first infected humans in the Guangdong province of southern China in 2002 was transmitted from bats and civets (Xu et al., 2004, Marra et al., 2003, Rota et al., 2003, Ksiazek et al., 2003, Holmes & Enjuanes 2003). The MERS CoV that originated in bats was first identified from camel to human transmission in Saudi Arabia in 2012 (Azhar et al., 2014a, Chan et al., 2015, Sabir et al., 2016, Azhar et al., 2014b, Omrani et al., 2015). These coronaviruses have crossed species and resulted in causing human

infections leading to mortality. It is reported that civet SARS CoV can also infect humans (Wang et al., 2005, Li et al., 2006). Recent reports indicate that SARS CoV-2 related viruses are harbored in pangolins (Han 2020, Lam et al., 2020) and share 85.5% to 92.4% sequence similarity to SARS CoV-2. The SARS-like CoVs from some bats and civets are predicted to result in human infections (Menachery et al., 2015, Wang et al., 2018) due to their changing genomic RNA sequences, importantly in the spike protein regions (Song et al., 2005, Menachery et al., 2016). Since January 2020, several complete genome sequences of viral CoV-2 isolated from infected patients belonging to various geographical locations, such as, Australia, China, Denmark, Finland, Hungary, India, Italy, Japan, South Korea, USA and Vietnam have been deposited in the GenBank (https://www.ncbi.nlm.nih.gov/genbank/). At the genomic level, the nucleotide sequences of SARS CoV and SARS CoV-2 share 79.6% sequence identity (Zhou et al., 2020). The viral RNA stores the genetic information and also serves to translate into structural and non-structural proteins of SARS CoV-2. The SARS CoV uses angiotensin converting enzyme-2 (ACE-2) as receptor for entry into human epithelial cells (Li et al., 2003) to cause the infection. Zhou et al., 2020 have carried out virus infectivity studies on HeLa cells and have shown that SARS CoV-2 also uses ACE-2 as receptor for cellular entry.

It has been reported that the first SARS CoV-2 infection in Wuhan, China is caused from the original host, bats (Zhou et al, 2020), and in less than 5 months transmitted among the human populations in the entire world. The initial contact between the SARS CoV/CoV-2 and human host is via recognition between the trimeric assembly of a heavily glycosylated cell envelope spike protein of the virus and the ACE-2 receptor of the human host resulting in the infection.

In order to understand the specificity and to estimate the extent of similarities and variations in SARS CoV-2 spike proteins required for binding the host receptor, we analyzed the representative spike protein sequences. Further, in order to estimate the evolutionary progression of the bat SARS CoV genome, such that it is able to adapt to human host as a novel coronavirus causing COVID-19, we have analyzed the complete genomes of the bat, civet, pangolin, human SARS CoV and human SARS CoV-2. We have carried out computational analyses on the nucleotide and protein sequences by generating multiple sequence alignments (MSAs), constructing phylogenetic trees and analyzing the three-dimensional structure of the spike proteins to address the above.

Materials and Methods:

The spike proteins were retrieved from the NCBI database, using the sequence similarity search BLAST program (Schäffer et al., 2001) with human SARS CoV-2 spike protein as the query sequence (NCBI code: QHD43416.1) from the genome (GenBank code: MN908947.3) (Wu et al., 2020). The complete genome nucleotide sequences of SARS CoV from bats, civets, pangolins, and SARS CoV and CoV-2 from human host were obtained from NCBI virus database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide) in the FASTA format. Only complete genome sequences without any ambiguity in nucleotide composition were considered for analyses. The redundancy in each dataset was removed using the CD-HIT program (Li & Godzik 2006).

The nucleotide and protein sequence homology analyses based on MSA reveals the substitutions, deletions and insertions at each position along the sequence. To understand the evolutionary relationships between the members, the MSAs were further processed to generate phylogenetic trees - a pictorial representation of the evolutionary relationships between related members of various sequences analyzed. The MSAs and phylogenetic trees of the spike proteins and complete genomes were generated using the Next Generation Phylogeny.fr web service available at https://NGPhylogeny.fr (Lemoine et al., 2019). The protocol takes all sequences (nucleotide or protein) as input in FASTA file format and generates the MSA and phylogenetic tree. In the NGPhylogeny server, we have selected FastME 2.0 program that infers phylogenies using a distance approach since it is capable of handling large datasets (Lefort et al., 2015). Based on the input FASTA sequences, a MSA is generated that adopts Multiple Alignment using Fast Fourier Transform (MAFFT) (Katoh & Standley 2013) with gap extension penalty; 0.123 and gap opening penalty; 1.53. The MSA generated is parsed through Block Mapping and Gathering with Entropy (BGME) software for selecting regions suitable for phylogenetic inference (Criscuolo & Gribaldo 2010). This method uses a sliding window size; 3, Maximum entropy threshold; 0.5, gap rate cut-off; 0.5, minimum block size; 5, matrix: PAM250 for DNA and BLOSUM62 for proteins. FastME estimates phylogenies that employ distance-based methods from MSAs using TN93 and LG as substitution models for DNA and proteins, respectively. In the distance-based methods, pairwise distances between all pairs of sequences are generated as a square matrix. The sequence pairs with shorter pairwise distances are clustered together more closely in the phylogenetic tree. Tree refinement was performed using Subtree Pruning and Regrafting

(SPR) with Balanced version of Minimum Evolution (BalME), with a decimal precision for branch length set to 6. Finally, the phylogenetic trees were generated using Interactive Tree Of Life program (iTOL) v4 (Letunic and Bork, 2019).

Results and discussion:

Analyses of the spike proteins of SARS CoV and SARS CoV-2

The RNA inside the virus is protected by an outer layer that is formed by structural proteins, namely; spike, envelope, membrane and nuclear capsid. The spike protein is heavily glycosylated and assembles into trimers on the outer layer of the virus to form a distinctive crown-like appearance or "Corona" that specifically recognizes the human host cell receptor called angiotensin converting enzyme-2 (ACE-2). The SARS CoV and SARS CoV-2 spike proteins retrieved from various host sources have a sequence length ranging between 1240 to 1273 amino acids. Structurally, a spike protein is characterized by three regions; 1) the N-terminal extracellular domain, 2) a transmembrane anchor domain and 3) an intracellular segment. The N-terminal extracellular domain comprises a receptor binding subunit (S1) and a membrane-fusion subunit (S2). The S1 subunit comprises two domains, an N-terminal domain (NTD) and RBD. The sequence analyses of spike proteins from the various host sources is shown in the MSA in supplementary Figure S1 and the phylogenetic tree in Figure 1. From Figure 1, it is observed that the paralogous proteins from individual host sources are associated with a distinct clade. The spike proteins from human SARS CoV-2 share highest sequence similarity according to the least pairwise distances. The orthologous spike proteins from other host species also are highly similar according to the low pairwise distances. Therefore, it is intriguing to see that despite high sequence identity between the spike proteins from various host sources, only some SARS CoVs and SARS CoV-2 are able to bind the human host ACE-2 receptor. In order to understand the above, we analyzed the MSA of the spike proteins.

From the supplementary Figure S1, it is observed that the N-terminal ~500 amino acids, comprising the S1 subunit vary to a moderate extent among all host sources, relative to the later region that shares higher sequence identity. The sequence region between ~300-500 amino acid residues is crucial in spike proteins, as it forms the RBD that recognizes the ACE-2 receptor which allows entry of the virus into the host cells. A sequence motif "$P_{681}RRA_{684}$", (amino acid numbering according to NCBI code: QHD43416) that is gained only in the human SARS CoV-2 spike proteins is referred to as a furin cleavage site (Wang et al., 2020a, Ou et al., 2020). In this work, we identify two, six amino acid insertion sequence regions; "$M_{153}ESEFR_{158}$" and "$S_{247}YLTPG_{252}$" specific to human SARS CoV-2 and the corresponding

sequences in pangolin SARS CoV; “VENEFR” and “SYLTPG”. The bat SARS CoV spike protein QHR63300 (GenBank code: MN996532.1) also comprises the sequence regions identical to human SARS CoV-2. In two bat spike proteins, AVP78042 (MG772934) and AVP78031 (MG772933), a six residues sequence "SIREFA" and a three residues sequence "GDP" are present at equivalent positions, respectively. The insertion sequences in human SARS CoV-2; "M_{153}ESEFR$_{158}$" and "S_{247}YLTPG$_{252}$" are associated with NTD that is distant from the ACE-2 binding site. From the three-dimensional structure of human SARS CoV-2, we infer that this region is likely to be exposed towards the surface of the spike protein in NTD.

In human SARS CoV-2, there are three insertion loop regions; a five residues loop "V_{445}GGNY$_{449}$", an eight residues loop “Y_{473}QAGSTPC$_{480}$” and a six residues loop "E_{484}GFNCY$_{489}$". The bat SARS CoV spike protein QHR63300 has also gained equivalent insertions with the amino acid sequences; “EGGNF”, “YQAGSKPC” and “TGLNCY”, respectively, and similar sequences are also present in the pangolin SARS CoV. It is interesting to note that in the third loop region, the bat SARS CoV QHR63300 and pangolin SARS CoV have already acquired six residues similar to SARS CoV-2, whereas, the equivalent loop region in all the other spike proteins of SARS CoV that recognize the ACE-2 receptor comprise five residues. The Figure 2A was generated by editing the MSA in Figure S1 to depict the sequence regions discussed above in representative spike proteins from the four host sources (bat, civet, pangolin, human SARS CoV and human SARS CoV-2).

The three loop regions; "V_{445}GGNY$_{449}$", “Y_{473}QAGSTPC$_{480}$” and "E_{484}GFNCY$_{489}$" are part of the RBD and involved in recognition of the ACE-2 receptor in human host. Their absence in some of the bat SARS CoV at equivalent positions may be responsible for their inability to bind human ACE-2. To study this, we have analyzed the three-dimensional structures of human SARS CoV (PDB code: 6ACG) (Song et al., 2018) and the RBD domain of human SARS CoV-2 (6LZG) complexed with ACE-2 (Wang et al., 2020b). The structures were superimposed and amino acid residues within 4.5Å distance from the ACE-2 receptor were identified. Despite amino acid mutations in the RBD region in both proteins, the structures are highly superimposable (Figure 2B). The location corresponding to the three insertion loop regions in human SARS CoV and human SARS CoV-2 match with the ACE-2 binding region (Figure 2C). The amino acid residues ≤4.5Å distance from the ACE-2 receptor are mentioned in the legend to Figure 2C. The second and third insertion loops, i.e.,

“Y_{473}QAGSTPC_{480}”, and "E_{484}GFNCY_{489}" in the RBD domain of human SARS CoV-2 are stabilized by a disulfide bridge that connects half-cystines at positions 480 and 488. This disulfide bridge between the insertion loops 2 and 3 is conserved in all the spike proteins that recognize ACE-2. Interestingly, the bat SARS CoV spike protein QHR63300 which has acquired the insertion sequences in RBD also has this disulfide bridge. The three insertion loops and the tethered disulfide bridge in the RBD of spike proteins represent important structural features required for the recognition of human ACE-2 receptor.

The sequence regions identified in this work serve as potential candidate epitopes for the design of antibodies specific for human SARS CoV-2 recognition. Our analyses suggest that the bat spike protein QHR63300 has undergone significant evolutionary changes, such that, it resembles the human CoV-2 spike protein more than the bat CoV which may have led to the transmission of CoV from bat to human as the novel SARS CoV-2. Among all the sequences studied, the pangolin SARS CoV shares the highest sequence identity with human SARS CoV-2 spike proteins. Our observations with regard to the pangolin SARS CoV spike protein sequences lead us to propose that the pangolin SARS CoV could also have probably resulted in the transmission to humans causing the infection. Our results also suggest that the bat SARS CoV spike proteins; AVP78042 and AVP78031, are in progression of acquiring mutations towards becoming SARS CoV-2-like proteins. The phylogenetic tree in Figure 1 showing the proteins; QHR63300, AVP78042 and AVP78031 close to the human SARS CoV-2 are in support of our hypothesis.

Complete genome analyses of SARS CoV and SARS CoV-2

The representative complete genomes of nucleotide sequences from bat, civet, pangolin, human CoV and human CoV-2 genomes were analyzed. The MSA is shown in Figure S2 and the phylogenetic tree in Figure 3. From Figure 3, it is observed that the human SARS CoV-2 genomes cluster into one clade (pairwise distance is lower than 0.002) revealing high identity that suggest their recent evolution. The bat SARS CoV genome (GenBank code: MN996532.1) is also member of this clade (pairwise distance between 0.042-0.043) indicating that it is the closest homolog to the human SARS CoV-2 among the bat genomes. The two bat SARS CoV genomes (GenBank codes: MG772933.1 and MG772934.1) are also close to the human SARS CoV-2 clade. The human and civet SARS CoV genomes cluster into another distinct clade. The pangolin CoV genomes cluster as one clade (pairwise

distance is lower than 0.0024). In comparison, the pangolin genomes are closely related to human SARS CoV-2 genomes (pairwise distance 0.185). The pangolin genomes share greater divergence with human and civet CoVs (pairwise distance ~0.31). The members of bat SARS CoV clade have undergone maximum evolutionary changes as observed in Figure 3. Based on these results, we propose that the bat SARS CoV genomes have diverged the most during the last 18 years (since its detection in 2002) and have evolved closer to civet and human SARS CoV genomes. The pangolin SARS CoV genomes are closer to the human SARS CoV-2. The bat SARS CoV genome (GenBank code: MN996532.1) has diverged significantly into the recent novel human SARS CoV-2 genomes, whereas, the bat CoV genomes (GenBank codes: MG772933 and MG772934) are intermediates during the evolution of bat SARS CoV into human novel SARS CoV-2. Genomes such as MG772933 and MG772934 are likely to undergo further evolutionary mutations and become adaptable to infecting human populations at the opportunistic moment. The availability of complete SARS CoV and CoV-2 genome sequences from various host sources at different collection timelines will be helpful to trace the evolutionary mutations and pathways of these viruses.

Conclusions:

Two sequence regions; "M_{153}ESEFR$_{158}$" and "S_{247}YLTPG$_{252}$" in the NTD of spike protein are specific to human SARS CoV-2 and pangolin SARS CoV. Three insertion loops; "V_{445}GGNY$_{449}$", "Y_{473}QAGSTPC$_{480}$" and "E_{484}GFNCY$_{489}$" in RBD that interact with ACE-2 may be exploited as potential candidates for antibody design. The phylogenetic analyses of the bat, civet, pangolin, human SARS CoV and human SARS CoV-2 genomes and spike proteins show that a bat SARS CoV (GenBank code: MN996532.1) and the pangolin SARS CoV are closest homologs of human SARS CoV-2. The above observations lead us to propose that pangolin SARS CoV like the bat SARS CoV must also have caused transmission to humans. Two other bat SARS CoV genomes (GenBank codes: MG772933 and MG772934) that have gained insertion sequences in NTD are intermediates in the evolution of bat genomes into human SARS CoV-2.

Conflict of interest:

The author declares no potential conflict of interest.

Acknowledgements:

LGP thanks School of Chemistry and CAS, UGC for providing research facilities.

References:

Azhar, E. I., El-Kafrawy, S. A., Farraj, S. A., Hassan, A. M., Al-Saeed, M. S., Hashem, A. M., & Madani, T. A. (2014a). Evidence for camel-to-human transmission of MERS coronavirus. *New England Journal of Medicine*, *370*(26), 2499-2505.

Azhar, E. I., Hashem, A. M., El-Kafrawy, S. A., Sohrab, S. S., Aburizaiza, A. S., Farraj, S. A., ... & Madani, T. A. (2014b). Detection of the Middle East respiratory syndrome coronavirus genome in an air sample originating from a camel barn owned by an infected patient. *MBio*, *5*(4), e01450-14.

Chan, J. F., Lau, S. K., To, K. K., Cheng, V. C., Woo, P. C., & Yuen, K. Y. (2015). Middle East respiratory syndrome coronavirus: another zoonotic betacoronavirus causing SARS-like disease. *Clinical microbiology reviews*, *28*(2), 465-522.

Criscuolo, A., & Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC evolutionary biology*, *10*(1), 210.

Drosten, C., Günther, S., Preiser, W., Van Der Werf, S., Brodt, H. R., Becker, S., ... & Berger, A. (2003). Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *New England journal of medicine*, *348*(20), 1967-1976.

Han, G. Z. (2020). Pangolins Harbor SARS-CoV-2-related Coronaviruses. *Trends in Microbiology*.

Holmes, K. V., & Enjuanes, L. (2003). The SARS coronavirus: a postgenomic era. *Science*, *300*(5624), 1377-1378.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772-780.

Ksiazek, T. G., Erdman, D., Goldsmith, C. S., Zaki, S. R., Peret, T., Emery, S., ... & Rollin, P. E. (2003). A novel coronavirus associated with severe acute respiratory syndrome. *New England journal of medicine*, *348*(20), 1953-1966.

Lam, T. T. Y., Shum, M. H. H., Zhu, H. C., Tong, Y. G., Ni, X. B., Liao, Y. S., ... & Leung, G. M. (2020). Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature*, 1-6.

Lefort, V., Desper, R., & Gascuel, O. (2015). FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Molecular biology and Evolution*, *32*(10), 2798-2800.

Lemoine, F., Correia, D., Lefort, V., Doppelt-Azeroual, O., Mareuil, F., Cohen-Boulakia, S., & Gascuel, O. (2019). NGPhylogeny. fr: new generation phylogenetic services for non-specialists. *Nucleic acids research*, *47*(W1), W260-W265.

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic acids research*, *47*(W1), W256-W259.

Li, W., Moore, M. J., Vasilieva, N., Sui, J., Wong, S. K., Berne, M. A., ... & Choe, H. (2003). Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature*, *426*(6965), 450-454.

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658-1659.

Li, W., Wong, S. K., Li, F., Kuhn, J. H., Huang, I. C., Choe, H., & Farzan, M. (2006). Animal origins of the severe acute respiratory syndrome coronavirus: insight from ACE2-S-protein interactions. *Journal of virology*, *80*(9), 4211-4219.

Marra, M. A., Jones, S. J., Astell, C. R., Holt, R. A., Brooks-Wilson, A., Butterfield, Y. S., ... & Cloutier, A. (2003). The genome sequence of the SARS-associated coronavirus. *Science*, *300*(5624), 1399-1404.

Menachery, V. D., Yount Jr, B. L., Debbink, K., Agnihothram, S., Gralinski, L. E., Plante, J. A., ... & Randell, S. H. (2015). A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nature medicine*, *21*(12), 1508.

Menachery, V. D., Yount, B. L., Sims, A. C., Debbink, K., Agnihothram, S. S., Gralinski, L. E., ... & Swanstrom, J. (2016). SARS-like WIV1-CoV poised for human emergence. *Proceedings of the National Academy of Sciences*, *113*(11), 3048-3053.

Omrani, A. S., Al-Tawfiq, J. A., & Memish, Z. A. (2015). Middle East respiratory syndrome coronavirus (MERS-CoV): animal to human interaction. *Pathogens and global health*, *109*(8), 354-362.

Ou, X., Liu, Y., Lei, X., Li, P., Mi, D., Ren, L., ... & Xiang, Z. (2020). Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nature communications*, *11*(1), 1-12.

Rota, P. A., Oberste, M. S., Monroe, S. S., Nix, W. A., Campagnoli, R., Icenogle, J. P., ... & Tong, S. (2003). Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *science*, *300*(5624), 1394-1399.

Sabir, J. S., Lam, T. T. Y., Ahmed, M. M., Li, L., Shen, Y., Abo-Aba, S. E., ... & Alharbi, N. S. (2016). Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. *Science*, *351*(6268), 81-84.

Schäffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., ... & Altschul, S. F. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic acids research*, *29*(14), 2994-3005.

Song, H. D., Tu, C. C., Zhang, G. W., Wang, S. Y., Zheng, K., Lei, L. C., ... & Zheng, H. J. (2005). Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. *Proceedings of the National Academy of Sciences*, *102*(7), 2430-2435.

Song, W., Gui, M., Wang, X., & Xiang, Y. (2018). Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. *PLoS pathogens*, *14*(8), e1007236.

Wang, M., Yan, M., Xu, H., Liang, W., Kan, B., Zheng, B., ... & Wang, H. (2005). SARS-CoV infection in a restaurant from palm civet. *Emerging infectious diseases*, *11*(12), 1860.

Wang, N., Li, S. Y., Yang, X. L., Huang, H. M., Zhang, Y. J., Guo, H., ... & Hagan, E. (2018). Serological evidence of bat SARS-related coronavirus infection in humans, China. *Virologica Sinica*, *33*(1), 104-107.

Wang, Q., Qiu, Y., Li, J. Y., Zhou, Z. J., Liao, C. H., & Ge, X. Y. (2020a). A unique protease cleavage site predicted in the spike protein of the novel pneumonia coronavirus (2019-nCoV) potentially related to viral transmissibility. *Virologica Sinica*, 1-3.

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., ... & Wang, Q. (2020b). Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell*.

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., ... & Yuan, M. L. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, *579*(7798), 265-269.

Xu, R. H., He, J. F., Evans, M. R., Peng, G. W., Field, H. E., Yu, D. W., ... & Li, L. H. (2004). Epidemiologic clues to SARS origin in China. *Emerging infectious diseases*, *10*(6), 1030.

Yan, R., Zhang, Y., Guo, Y., Xia, L., & Zhou, Q. (2020). Structural basis for the recognition of the 2019-nCoV by human ACE2. *bioRxiv*.

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., ... & Chen, H. D. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, *579*(7798), 270-273.

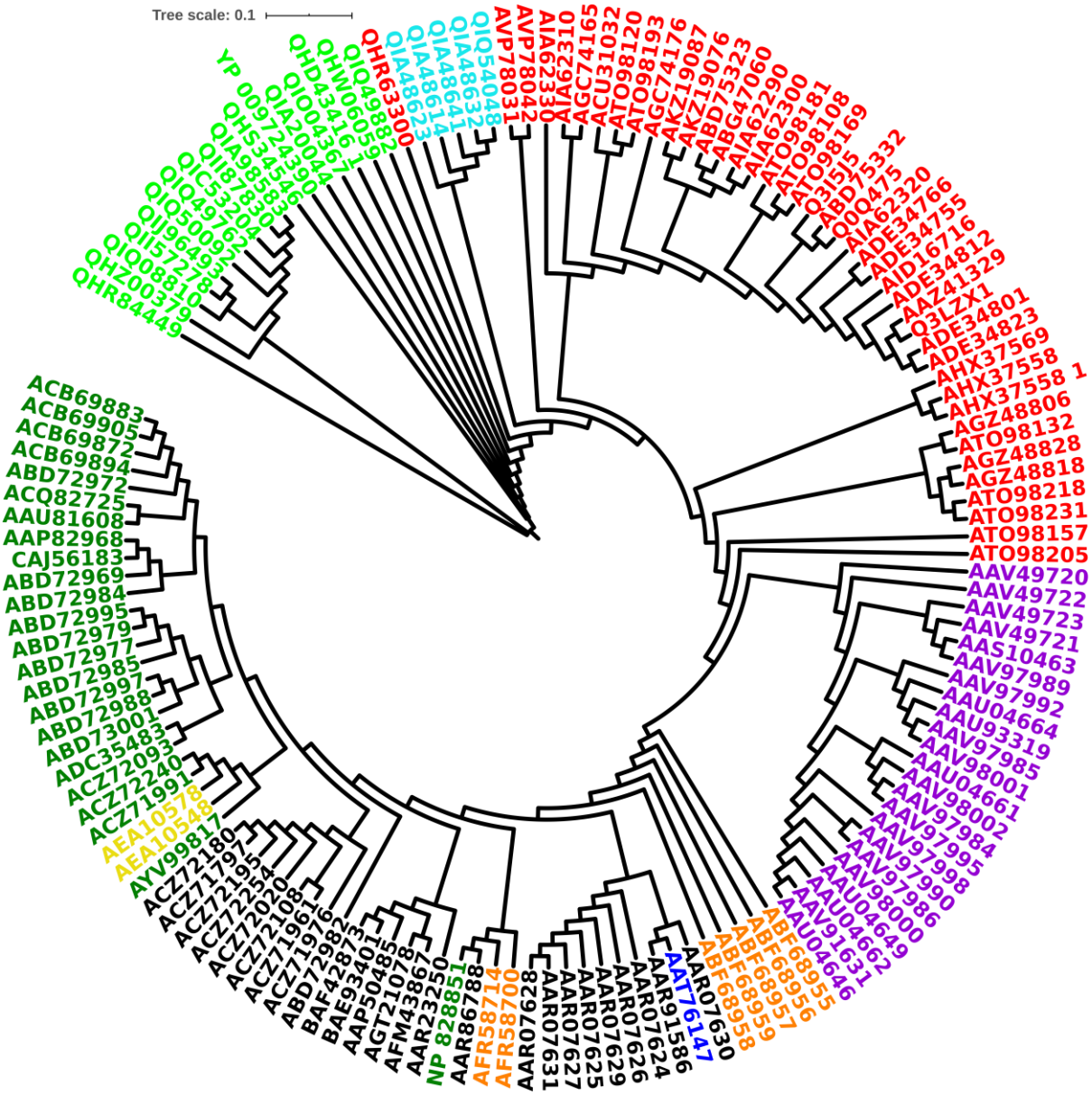

Figure 1. Phylogenetic tree of SARS CoV and SARS CoV-2 spike proteins from human CoV-2 (light green), pangolin CoV (cyan), human CoV (dark green), bat CoV (red), civet CoV (violet), cat CoV (orange), swine CoV (blue), mouse CoV (bright yellow), SARS CoV recombinant spike proteins and from lab adapted cells (black).

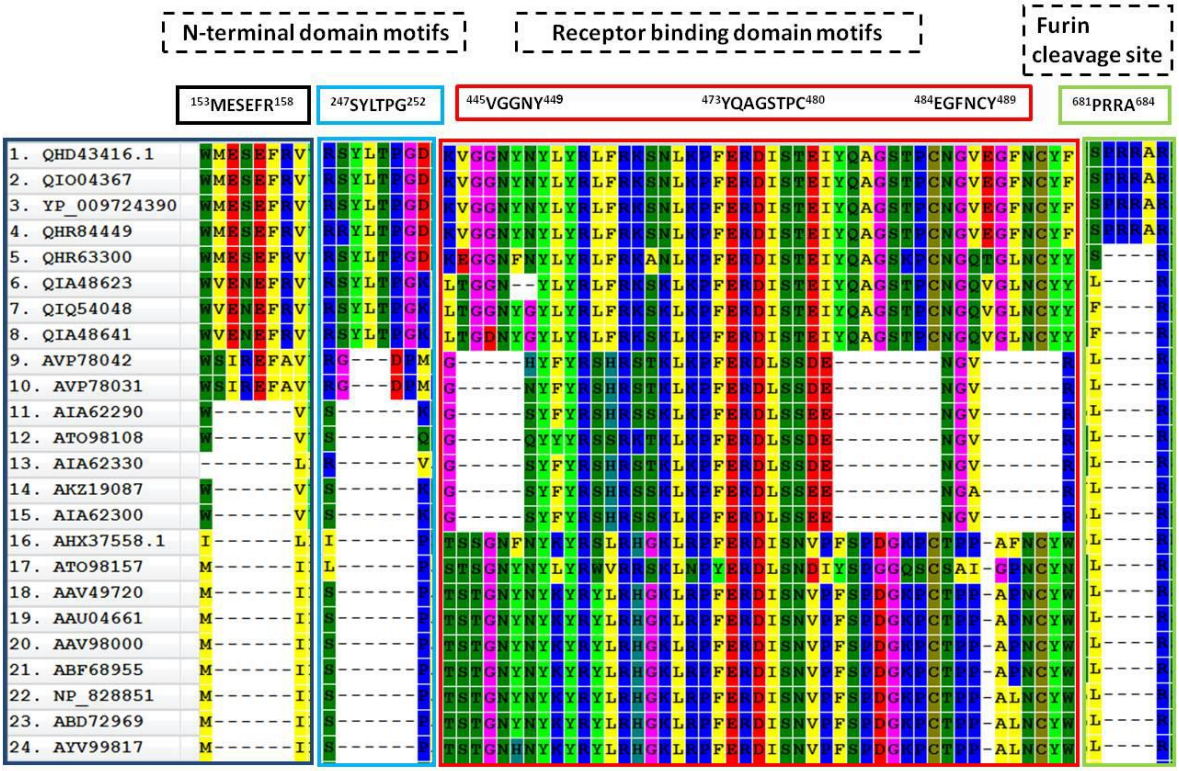

Figure 2A. Portions of the alignment of spike proteins extracted from the multiple sequence alignment (Figure S1) showing the insertion sequences and their locations within the NTD, RBD and furin cleavage sites for human SARS CoV-2 (1-4), bat SARS CoV RaTg13 (5), pangolin SARS CoV (6-8), bat SARS CoV (9-17), civet SARS CoV (18-21), human SARS CoV (22-24).

Figure 2B. Structural superposition of human SARS CoV (PDB code: 6ACG, cyan) and human SARS CoV-2 (6LZG, orange) and the long H1 and H2 helices in ACE-2 are shown in green and magenta, respectively.

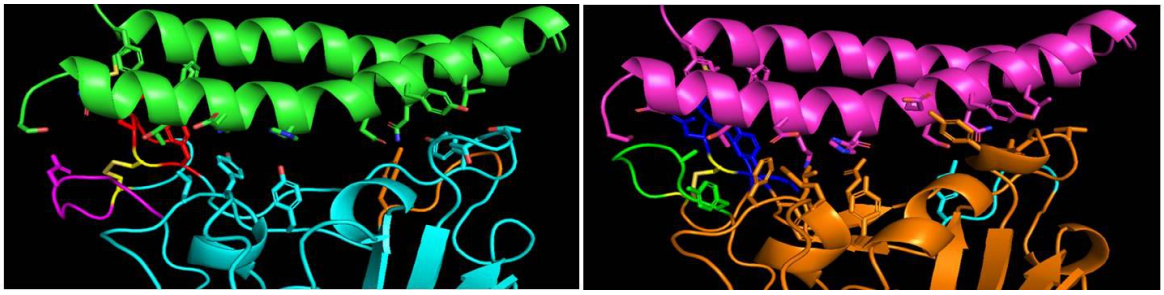

Figure 2C. [Left panel] The location and conformation of the three loops and the tethered disulfide bridge are shown in the structure of human SARS CoV bound to ACE-2 receptor (PDB code: 6ACG). Loop 1 (S_{432}TGNY$_{436}$ in orange), loop 2 (F_{460}SPDGKPC$_{467}$ in magenta) and loop 3 (A_{471}LNCY$_{475}$ in red). Amino acid residue side-chains that make contacts ≤4.5Å between the receptor binding domain of spike protein (Y436, Y440, Y442, L443, D463, L472, N473, Y475, G482, Y484, T486, T487) and ACE-2 receptor (S19, Q24, T27, F28, D30, K31, H34, D38, Y41, Q42, L45, L79, M82, Y83). [Right panel] The location and conformation of the three loops and the tethered disulfide bridge are shown in the structure of human SARS CoV-2 bound to ACE-2 receptor (PDB code: 6LZG). Loop 1 (V_{445}GGNY$_{449}$ in cyan), loop 2 (Y_{473}QAGSTPC$_{480}$ in green), loop 3 (E_{484}GFNCY$_{489}$ in blue). Amino acid residue side-chains that make contacts ≤4.5Å between the receptor binding domain of spike protein (K417, G446, Y449, Y453, L455, F456, Y473, A475, G476, E484, F486, N487, Y489, F490, Q493, G496, Q498, T500, N501, Y505) and ACE-2 receptor (S19, Q24, T27, F28, D30, K31, H34, E35, E37, D38, Y41, Q42, L45, L79, M82, Y83).

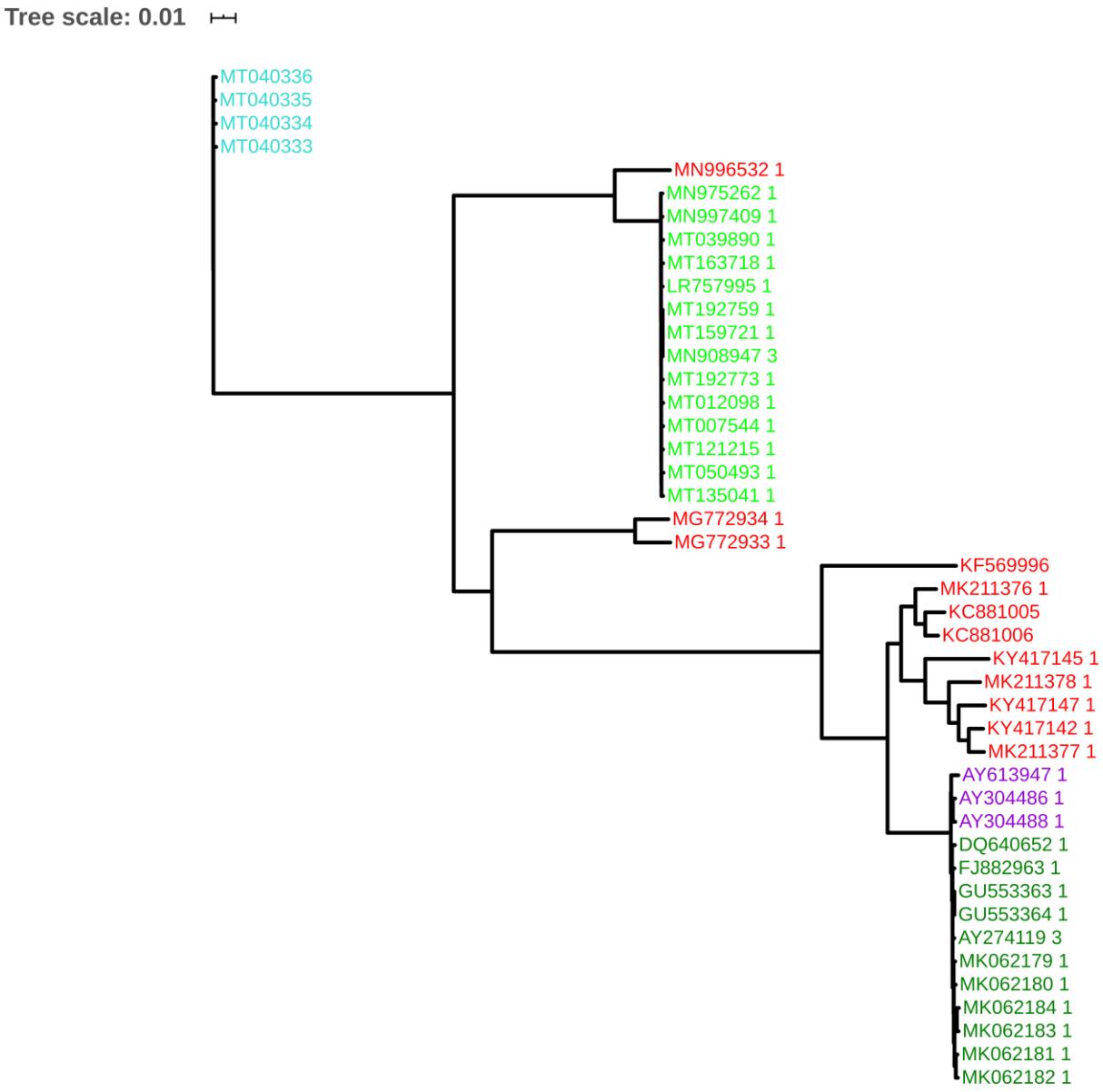

Figure 3. Phylogenetic tree of SARS CoV and SARS CoV-2 complete genomes. Human SARS CoV-2 (light green), human SARS CoV (dark green), pangolin SARS CoV (cyan), bat SARS CoV (red), civet SARS CoV (violet).

Legend to supplementary Figures:

Figure S1. Multiple sequence alignment of spike proteins from SARS CoV and SARS CoV-2 from different host sources.

Figure S2. Multiple sequence alignment of complete genomes from human SARS CoV-2, human SARS CoV, pangolin SARs CoV, bat SARS CoV and civet SARS CoV.