 doi:10.20944/preprints202102.0264.v1

SARS-CoV-2 and the secret of the furin site

Antonio R. Romeu[1] and Enric Ollé[2]

Department of Biochemistry and Biotechnology, University Rovira i Virgili, E-43007 Tarragona, Spain.

[1]: Professor of Biochemistry and Molecular Biology. Corresponding author. Email: antonioramon.romeu@iubilo.urv.cat
[2]: Associated Professor. Email: enric.olle@urv.cat

Abstract

The SARS-CoV-2 high infectivity is due to the functional polybasic furin cleavage site in the S protein. How it was acquired is unknown. There are two challenges to face: (*i*) an evolutionary model, to fit the origin of the coronavirus; and (*ii*) a molecular mechanism for the site acquisition. Here we show genomic fingerprints which are specific of Pangolin-CoVs, Bat-SARS-like (CoVZC45, CoVZXC21), bat RatG13 and human SARS-CoV-2 coronaviruses. This, along with phylogenetic analysis, we found that these species have the same evolutionary origin in the bat, including a genetic recombination of *S* gene between Pangolin-CoV (2017) and RatG13 ancestors. However, this does not explain why SARS-CoV-2 is the only of them with the furin site, which consists in four amino acid (PRRA) motif. The Arginine doublet is encoded by CGGCGG codons. Surprisingly, none of the Arginine doublet of other furin site of viral proteins from several type of viruses, are encoded by the CGGCGG codons. This makes it difficult to consider a virus recombination as mechanism for the PRRA acquisition. The origin of SARS-CoV-2, is the origin of the recognition cleavage site. The bat coronavirus RaTG13 appears to be the closest relative of the SARS-CoV-2, but was isolated in 2013. So, new RatG13 samples would provide insights into the acquisition of the polybasic motif.

Key words

SARS-CoV-2, RaTG13, Furin Site, Molecular Evolution, Bioinformatics.

Introduction

The origin of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the pandemic virus of the coronavirus disease 2019 (COVID-19), is controversial. It is linked to the origin of the polybasic furin clevage site in the sipke glycoprotein (1). Furin is a protease ubiquitously expressed in human cells. In the human genome, the furin gene (*FUR*) is located on chromosome 15. Furin specifically cleaves substrates at single or paired basic residues in normal protein processing (2). However, furin is also involved in protein processing in infectious diseases and cancer (3), and now in COVID-19. The polybasic furin cleavage site is present in many viral proteins from different types of viruses (3). Unlike SARS-CoV and other Betacoronavirus Sarbecovirus (lineage B), the spike protein of SARS-CoV-2 is thought to be uniquely cleaved by the furin (2,4,5). We addressed both the SARS-CoV-2 evolutionary model analysis and the site acquisition, through a bioinformatic approach, based on the available genomic data, as an attempt to fit together the pieces of a puzzle. The reproducibility of the results has been considered essential.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2021 doi:10.20944/preprints202102.0264.v1

Methods

The source of information was the National Center for Biotechnological Information (NCBI) databases and the methodology was based on the bioinformatic resources of NCBI and the European Laboratory of Molecular Biology (EMBL). The nucleotide Basic Local Alignment Search Tool (BLASTn) were performed using the NCBI-BLAST program (6). A reference SARS-CoV-2 genome was used as query against the entire NCBI nucleotide collection. The entire nucleotide collection consists of GenBank+EMBL+DDBJ+PDB+RefSeq sequences. The database is non-redundant. Identical sequences have been merged into one entry. Number of sequences: 65512295. For clarity, search was limited to records that exclude: Severe acute respiratory syndrome coronavirus 2 (taxid:2697049). Multiple sequence alignments were created by Clustal Omega (v.1.2.4) using default parameters (7). The phylogenetic analysis were based on both complete genomes and *S* gene sequences. Because the purpose of the phylogenetic analyses is to make the detailed phylogenetic relationships of closely related SARS-CoV-2 coronaviruses, and to ensure that the analyses are comparable, sequences from the same coronaviruses were used in both phylogenetic analyses. The sequences were those of the SARS-CoV-2 group (Table 1), and selected coronaviruses sequences from the literature (4,8,9,10), for the construction of the phylogenetic tree. It was constructed with the Neighbor Joining method of the Clustal Omega (v.1.2.4) software package, and the iTol Interactive Tree Of Life tool (11). Assessed clustering strength was calculated by bootstrap using 1000 replicates. Tree scale bar stands for the evolutionary distance. A node in the phylogenetic tree may represent both a common ancestor of the homologous sequences that originate from it (defining a cluster), and/or an evolutionary event of speciation.

The coronavirus S protein sequences used in the multiple alignment were the following (GenBank accession number and coronavirus): ADC35483.1, human SARS coronavirus HKU-39849; human sp|P59594| (UNIPROT SPIKE_CVHSA); AAR07630.1, human SARS coronavirus BJ302; AVP78031.1, Bat-SL-CoV ZC45; AVP78042.1, Bat-SL-CoV ZXC21; QIG55945.1, Pangolin-CoV, MP789 (2019); QIQ54048.1, Pang-CoV,GX-P2V; QIA48614.1, Pang-CoV,GX-P4L; QIA48623.1, Pang-CoV,GX-P1E; QHR63260.2, human SARS-CoV-2; QII57208.1 human SARS-CoV-2; QIA98554.1, human SARS-CoV-2; QHR63300.2, bat RaTG13.

The coronavirus genomes used in the phylogenetic analyses were the following (GenBank accession number and coronavirus): MT040333.1-MT040336.1, Pangolin coronavirus, 2017, Sarbecovirus; MT072864.1, Pangolin coronavirus (2018), Sarbecovirus; MG772933.1-MG772934.1, Bat SARS-like coronavirus CoVZC45, CoVZXC21, Sarbecovirus; MT121216.1 Pangolin coronavirus, MP789, 2019, Sarbecovirus; MN996532.2, BatCoV-RaTG13, Sarbecovirus; MT159709.1, SARS-CoV-2 isolate 2019-nCoV/USA-CruiseA-12/2020, Sarbecovirus; MN996528.1, SARS-CoV-2 isolate WIV04, Wuhan, Sarbecovirus; NC045512.2, SARS-CoV-2 isolate Wuhan-Hu-1, Sarbecovirus; DQ022305.2, Bat SARS coronavirus HKU3-1, Sarbecovirus; KF294457.1, Bat SARS-like coronavirus, Sarbecovirus; DQ412042.1, Bat SARS CoV Rf1, Sarbecovirus; NC_014470.1, Bat coronavirus BM48-31/BGR/2008, Sarbecovirus; FJ588686.1, SARS coronavirus Rs_672, Sarbecovirus; DQ071615.1, Bat SARS CoV Rp3, Sarbecovirus; NC_014470.1, Bat coronavirus BM48-31/BGR/2008, Sarbecovirus; MH734115.1, Camel, Middle East respiratory syndrome-related coronavirus isolate MERS-CoV, Merbecovirus; NC_009019.1, Tylonycteris bat coronavirus HKU4, Merbecovirus; NC_009020.1, Pipistrellus bat coronavirus HKU5, Merbecovirus; KF530114.1, Human coronavirus NL63, Alphacoronavirus; KF514433.1, Human coronavirus 229E, Alphacoronavirus; MK303625.1, Human coronavirus OC43, Embecovirus; KF430201.1, Human coronavirus HKU1, Embecovirus.

Results and Discussion

Evidence of a common ancestor of SARS-CoV-2 and its closely related coronaviruses

The basic principles of biology are also fulfilled in the world of viruses. The principle of the Cell Theory of Rudolf Virchow (1858), *Omnis cellula ex cellula* (each cell derived from another pre-existing cell), could be interpreted as "each virus derives from a pre-existing virus". Thus, we were interested in locating an ancestor

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2021 doi:10.20944/preprints202102.0264.v1

of the SARS-CoV-2 within the framework of an evolutionary model with its closely related coronaviruses.

Fortunately, through the BLASTn search, using as query a reference SARS-CoV-2 complete genome sequence, against the entire collection of NCBI nucleotide sequences, we found three SARS-CoV-2 genomic fingerprints that allowed us to identify its closely related coronaviruses, which are: Pangolin-CoVs (2017, 2019), Bat-SARS-like (CoVZC45, CoVZXC21) and bat RatG13 (Figure 1 and Table1). Regarding the genomic fingerprints, with coordinates based on NC_045512.2 SARS-CoV-2 isolate Wu han-Hu-1, complete genome (used as quesry), they are: fingerprint 1, 1923-3956; fingerprint 2, 21577-22539; and fingerprint 3, 27910-28257. More specifically, fingerprint 1 is at the beginning of the genome in the *orf1a* RNA polimerase gene, covering the *nsp2* (the final 796 nucleotides) and *nsp3* (the initial 1237 nucleotides) regions. The fingerprint 2 is at the beginning of *S* gene, covering the part encoding the N-terminal domain and the ACE2 receptor binding domain (RBD). The fingerprint 3 is the *orf8* gene itself. Interestingly, these genomic fingerprints are only specific markers at gene level (RNA sequence), not at the protein level. That is, their encoded amino acid sequences are similar to those of the other Sarbecovirus.

The genomic fingerprints are shared by the closely related SARS-CoV-2 coronaviruses but not by other coronaviruses. It is an evidence that a common ancestor served as a progenitor of them. It is unlikely that these SARS-CoV-2 related species independently acquired identical markers at three different locations in the genome. As further evidence of that close phylogenetic kinship, in the N-terminal domain of the S protein, there are short sequence features (one deletion and three insertions, Figure 2) which are also shared by the SARS-CoV-2 related coronaviruses but not by other Sarbecovirus. Again, It is unlikely that these species independently acquired identical deletion/insertions at four different locations in the *S* gene.

The phylogenetic analysis based on complete genomes, corroborated that Pangolin-CoV (2017), Pangolin-CoV (2019), Bat-SL-CoV (CoVZC45, CoVZXC21), bat RatG13 and human SARS-CoV-2 coronavirus species have the same evolutionary origin in the bat, and have been separated by speciation events. Along the evolutionary process, Pangolin-CoV (2017) species was first diverged from the others (bootstrap support 1000). Clearly, RatG13 is the closest relative of SARS-CoV-2 (8) (Figure 3).

Evidence of S gene recombination between Pangolin-CoV (2017) and BatCov-RatG13 ancestors

A detailed analysis of the N-terminal domain of S protein evidence high similarity between Pangolin-CoV (2017), RatG13 and SARS-CoV-2 sequences (Figure 2). Since at the complete genome level, Pangolin-CoV (2017) is phylogenetically distant from RatG13 and SARS-CoV-2, this already point out to a recombination event, that it has been validated by through further phylogenetic analyses.

We found in the phylogenetic analysis based on the nucleotide sequence of *S* gene region encoding the N-terminal domain and the RBD, that Pangolin-CoVs (2017) on the one hand, and RatG13 and SARS-CoV-2 on the other, were consistently (bootstrap support 1000) grouped together in the same cluster. Pangolin-CoV (2019) and Bat-SL-CoV grouped with the other coronaviruses (Figure 4). Comparing both phylogenetic analyses, the one based on complete genomes (Figure 3), and that based on the *S* gene (Figure 4), there are a conflicting evolutionary history suggesting a recombination or horizontal gene transfer event (12), involving Pangolin-CoV (2017) and RatG13 ancestors. This could be an evidence that these species have undergone recombination of *S* gene. So, the Pangolin-CoV (2017) species has similar N-terminal domain and RBD to that of RatG13 and SARS-CoV-2 (Figure 2 and 6). However, from the current sequence and phylogenetic analyses, it is not possible to elucidate the directionality of that recombination. That is, we cannot know which of them acted as donor and/or receptor on that exchange of *S* gene region.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2021 doi:10.20944/preprints202102.0264.v1

The secret of the SARS-CoV-2 functional polybasic furin clevage site

A key identity mark of SRAS-CoV-2 is the polybasic furin cleavage in the S protein, but is absent in the closely related SARS-CoV-2 coronaviruses (4,5). It is a polybasic recognition motif of the human ubiquitously expressed serine protease furin, that cleaves the S protein in the conserved S1/S2 cleavage region (1,4,5). According the SRAS-CoV-2 biology, this furin site is responsible for its high infectivity and transmissibility, as well as, for the COVID-19 pathogenesis (13,14).

In SARS-CoV-2 S protein, this site is four amino acid PRRA, encoded by the inserted of 12 nucleotides in the S gene. The presence of a doublet of Arginine is a distinctive structural feature of the furin site (15). Taking as reference the S protein of the Bat coronavirus RaTG13 (the closest relative), the PRRA insertion occurred between the Serine-680 (encoded by TCA) and the Arginine-681 (encoded by CGT). However, there are three possible 12-nucleotide fragments that, inserted in a different strategic way, encoded the same PRRA sequence. The Figure 5 shows details of each possible insertion. From the current genomic data, it is now impossible to know what of these cases actually happened in the most decisive evolutionary event in the SARS-CoV-2 speciation.

However, furin sites are present in many viral proteins of all types of viruses (3). Table 2 shows several examples. From the seventh coronavirus known to infect humans (4), the site is also found in the S protein of the Betacoronavirus Embecoviruses (Lineage A) HKU1, OC43; and the Betacoronavirs Merbecovirus (lineage C) Middle East respiratory Syndrome-Related Coronavirus (MERS-CoV); but not in the Betacoronavirus Sarbecovirus (Lineage B) SARS-CoV; and the Alphacoronaviruses NL63 and 229E.

hypotheses for how the furin site could be acquired by SARS-CoV-2 include: (*i*) random insertion mutation (4); (*ii*) recombination (16); and (*iii*) creation in a laboratory (17).

With regard to random insertion mutation. Viral RNA synthesis is performed by the RNA- dependent RNA polymerase (RdRP), including 15–16 non-structural proteins, RNA-modifying enzymes, and a 3′–5′ exonuclease activity that assists RNA synthesis with a unique RNA proofreading function necessary for maintaining the integrity of the >30 kb coronavirus genome (1,18). Moreover, SARS-CoV-2, like the RNA viruses, could be considered a quasispecies, where there are always random point mutations, or error tail, but it is metastable and the mutation rate is below the threshold of error catastrophe (1). Thus, the complexity of RNA replicase and the very structure of the virus genome, reduce the likelihood of a random insertion mutation as mechanism for the origin of the PRRA motif.

Concerning recombination with other viruses. The two Arginines of the PRRA motif, in SARS-CoV-2 are encoded by the CGGCGG codons. So, we have analysed the codon usage of representative furin sites with an Arginine double of a wide variety of viral proteins, of all types of viruses (including, dsDNA, (+)ssRNA, (-)ssRNA) (Table 2). Surprisingly, none of the Arginine doublet is encoded by the CGGCGG codons. Then, we were interested in the Arginine codon usage in SARS-CoV-2 genome. As it is shown in Table 3, out of the six codons of the Aarginine, the CGG codon is the minority. Also surprisingly, out of the 42 Arginines of the SARS-CoV-2 S protein, only two Arginines are encoded by the CGG codon, which are those of the PRRA motif. All recombination event requires a donor. In this case, the donor should be another furin site, probably from another virus, but having the double RR and encoded also by the CGGCGG codons. We have not been able to identify this hypothetical donor. So, this makes it difficult to consider viral recombination as mechanism for PRRA acquisition in SARS-CoV-2.

As concerns a laboratory origin. It is unlikely that someone manipulated these viral changes in a laboratory, but not impossible.

Furthermore, the mechanism of the acquisition of the functional polybasic furin site in SARS-CoV-2 refers to how, but not when and where. Since RatG13 sample was isolates in 2013 (8), it is plausible that a SARS-CoV-2 ancestor acquired the site after this year. Regarding where, we propose two scenarios that can plausibly explain the origin of SARS-CoV-2: (*i*) in animal host before jumping to human, or (*ii*) inside the human host. Assuming the first scenario, the current evolutionary model, tells us that that "animal host" should be the bat.

It is not unreasonable to consider that SARS-CoV-2 could have acquired the furin site within the bat. There are signs that point to this possibility. Part of the optimization that the pandemic virus acquired for binding to the human receptor ACE2 was ad quired within the bat. J. Lan *et al*. (2020) describe the contacting residues of SARS-CoV-2 RBD in direct comparison with the contacting residues of SARS-CoV RBD (19). Interestingly, we found that SARS-CoV-2 RBD optimization is also shown in RatG13 (Figure 6). In the same vein, SARS-CoV-2 sequence analysis, led to predicted the acquisition of three O-linked glycans around the furin site (4), also, SARS-CoV-2 and RAtG13 sequences around the site are 100% identical.

However, if a SARS-CoV-2 ancestor had acquired the furin site in the bat, it does not imply that RatG13 had also acquired the site. This is impossible to know, because we are analysing the RatG13 sample of 2013. Until no new RatG13 genomes, but captured in 2021, would be analysed, the scenario is open.

Taken together, all these lines of evidence and reasoning show that the acquisition of the polybasic furin cleavage site by SARS-CoV-2 is a "missing link" in our understanding of its evolutionary history, that can only be addressed through the discovery of new viruses. The principle of Theodosius Dobzhansky (1973) "Nothing in Biology Makes Sense Except in the Light of Evolution" becomes more relevant in the case of the pandemic virus, not only because there is the hypothesis of a laboratory origin, but also because there is a furin site that has changed the world.

Conclusions

- Genomic fingerprints in *Orf1a*, *S* and *orf8* genes, and short sequence features in the N-ternimal domain of the S proteins; along with phylogenetic analysis based on complete genome, suggest that Pangolin-CoV (2017), Pangolin-CoV (2019), Bat-SL-CoV (CoVZC45, CoVZXC21), bat RatG13 and human SARS-CoV-2 coronavirus species have the same evolutionary origin in the bat, and have been separated by speciation events.

- Sequence and phylogenetic analyses suggest that Pangolin-CoV (2017) and RatG13 coronaviruses ancestors have undergone recombination in the *S* gene region encoding the N-terminal domain and the RBD, before pangolin coronavirus jumped to the pangolin as host animal.

- The CGG CGG codons of the Arginine doublet in the SARS-COV-2 PRRA polybasic furin cleavage site (S protein), have not been identified in any other furin site Arginine doublet, from viral proteins, including S protein, of a wide variety of viruses. This not supports recombination as mechanism for PRRA acquisition in SARS-CoV-2.

- SARS-CoV-2 origin matters. The million-dollar question: would a bat coronavirus RatG13, isolated in 2021, have the functional polybasic furin site?

Acknowledgements

This work has not been awarded grants by any research-supporting institution.

Competing interest declaration

All authors declare that they have no conflicts of interest.

References

1. Britt Glaunsinger. Coronavirus biology. The second lecture in the COVID-19, SARS-CoV-2 and the Pandemic Series. University of California, Berkeley. 2020. Accessed January 30, 2021. https://www.youtube.com/watch?v=r2mOU2qOCYs.

2. NCBI. Human furin gene. Accessed January 30, 2021. https://www.ncbi.nlm.nih.gov/gene/5045

3. Elisabeth Braun, Daniel Sauter. Furin-mediated protein processing in infectious diseases and cancer. Clin. Transl. Immunol. E1073, 2019. PMID: 31406574. doi.org/10.1002/cti2.1073.

4. Kristian G Andersen, Andrew Rambaut, W Ian Lipkin, Edward C Holmes, Robert F Garry. The proximal origin of SARS-CoV-2. Nat. Med. 26:450-452, 2020. PMID: 32284615. doi: 10.1038/s41591-020-0820-9.

5. Javier A Jaimes, Nicole M André, Joshua S Chappie, Jean K Millet, Gary R Whittaker. Phylogenetic Analysis and Structural Modeling of SARS-CoV-2 Spike Protein Reveals an Evolutionary Distinct and Proteolytically Sensitive Activation Loop. J. Mol. Biol. 432:3309-3325, 2020. PMID: 32320687. doi: 10.1016/j.jmb.2020.04.009.

6. Zheng Zhang, Scott Schwartz, Lukas Wagner, Webb Miller. A greedy algorithm for aligning DNA sequences. J. Comput. Biol. 7:203-214, 2000. PMID: 10890397. doi:10.1089/10665270050081478.

7. Fábio Madeira, Young Mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian R N Tivey, Simon C Potter, Robert D Finn, Rodrigo Lopez. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucl. Acids Res. 47(W1):W636-W641, 2019. PMID: 30976793. doi: 10.1093/nar/gkz268.

8. Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Chao-Lin Huang, Hui-Dong Chen, Jing Chen, Yun Luo, Hua Guo, Ren-Di Jiang, Mei-Qin Liu, Ying Chen, Xu-Rui Shen, Xi Wang, Xiao-Shuang Zheng, Kai Zhao , Quan-Jiao Chen, Fei Deng, Lin-Lin Liu, Bing Yan, Fa-Xian Zhan, Yan-Yi Wang, Geng-Fu Xiao, Zheng-Li Shi. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270–273, 2020. PMID: 32015507. doi: 10.1038/s41586-020-2012-7. (Addendum, 588(7836):E6, 2020. PMID: 33199918. doi: 10.1038/s41586-020-2951-z).

9. Tao Zhang, Qunfu Wu, Zhigang Zhang. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. Curr. Biol. 30:1346–1351.e2, 2020. PMID: 32315626. doi: 10.1016/j.cub.2020.03.022.

10. Muhamad Fahmi, Yukihiko Kubota, Masahiro Ito. Nonstructural proteins NS7b and NS8 are likely to be phylogenetically associated with evolution of 2019-nCoV. Infect. Genet. Evol. 81:104272, 2020. PMID: 32142938. doi.org/10.1016/j.meegid.2020.104272.

11. Ivica Letunic, Peer Bork. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucl. Acids Res. 2011, Vol. 39(W475–W478), 2011. PMID: 21470960. doi:10.1093/nar/gkr201.

12. Eugene V Koonin , Tatiana G Senkevich, Valerian V Dolja. The ancient Virus World and evolution of c Biol. Direct 1:29, 2006. PMID: 16984643. doi: 10.1186/1745-6150-1-29. ells.

13. Shuai Xia, Qiaoshuai Lan, Shan Su, Xinling Wang, Wei Xu, Zezhong Liu, Yun Zhu, Qian Wang, Lu Lu, Shibo Jiang. The role of furin cleavage site in SARS-CoV-2 spike protein-mediated membrane fusion in the presence or absence of trypsin. Signal Transduct. Target Ther. 5:92, 2020. PMID: 32532959. doi.org/10.1038/s41392-020-0184-0.

14. Markus Hoffmann, Hannah Kleine-Weber, Stefan Pöhlmann. Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. Mol. Cell 78:779-784, 2020. PMID: 32362314. doi: 10.1016/j.molcel.2020.04.022.

15. Imène Kara, Marjorie Poggi, Bernadette Bonardo, Roland Govers, Jean-François Landrier, Sun Tian, Ingo Leibiger, Robert Day, John W M Creemers, Franck Peiretti. The Paired Basic Amino Acid-cleaving Enzyme 4 (PACE4) Is Involved in the Maturation of Insulin Receptor Isoform B. J. Biol. Chem. 290:2812-2821, 2015. PMID: 25527501. doi: 10.1074/jbc.M114.592543.

16. William R Gallaher. A palindromic RNA sequence as a common breakpoint contributor to copy-choice recombination in SARS-COV-2. Arch. Virol. 165:2341-2348, 2020. PMID: 32737584. doi: 10.1007/s00705-020-04750-z.

17. Rossana Segreto, Yuri Deigin. The genetic structure of SARS-CoV-2 does not rule out a laboratory origin. Bioessays e2000240, 2020. PMID: 33200842. doi: 10.1002/bies.202000240.

18. Philip V'kovski, Annika Kratzel, Silvio Steiner, Hanspeter Stalder, Volker Thiel. Coronavirus biology and replication: implications for SARS-CoV-2. Nat. Rev. Microbiol. Oct 28;1-16, 2020. PMID: 33116300. doi: 10.1038/s41579-020-00468-6.

19. Jun Lan, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang, Xuanling Shi, Qisheng Wang, Linqi Zhang, Xinquan Wang. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature 581:215-220, 2020. PMID: 32225176. doi:10.1038/s41586-020-2180-5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2021 doi:10.20944/preprints202102.0264.v1

Table 1. Group of Betacoronavirus, Sarbecovirus whose genomic sequences match the SARS-CoV-2 genomic fingerprints*

Coronavirus	Isolate	Country	Year	GenBank	% Identity
Bat coronavirus RaTG13	RatG13	China	2013	MN996532.2	96.14
Pangolin coronavirus	MP789	China	2019	MT121216.1	90.11
Bat SARS-like coronavirus	bat-SL-CoVZC45	China	2017	MG772933.1	89.12
Bat SARS-like coronavirus	bat-SL-CoVZXC21	China	2015	MG772934.1	88.65
Pangolin coronavirus	PCoV_GX-P5L	China	2017	MT040335.1	85.98
Pangolin coronavirus	PCoV_GX-P4L	China	2017	MT040333.1	85.97
Pangolin coronavirus	PCoV_GX-P1E	China	2017	MT040334.1	85.95
Pangolin coronavirus	PCoV_GX-P5E	China	2017	MT040336.1	85.95
Pangolin coronavirus	PCoV_GX-P2V	China	2018	MT072864.1	85.94

*: based on the NCBI-BALSTn search described in Figure 1. Percentatge of identity, using the NC_045512.2 SARS-CoV-2 isolate Wuhan-Hu-1, complete genome sequence as a query.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2021 doi:10.20944/preprints202102.0264.v1

Table 2. Virus polybasic furin cleavage sites*

Type	Taxonomy	Virus	Host	Protein	GenBank	Furin site	Codon
dsDNA	Herpesviridae	Betaherpesvirus 5	Human	Envelop. gB	AFR55885.1	THRTRRST 462	ACT CAT AGG ACC AGA AGA AGT ACG
dsDNA	Herpesviridae	Alphaherpesvirus 3	Human	ORF31	QCA47402.1	NTRSRRSV 496	AAT ACC AGA TCC CGA CGA AGC GTG
dsDNA	Herpesviridae	Alphaherpesvirus 3	Human	Glycoprot. B	AXA97875.1	NTRSRRSV 492	AAT ACC AGA TCC CGA CGA AGC GTG
dsDNA	Herpesviridae	Herpesvirus 3	Human	ORF31	AH009994.2	NTRSRRSV 433	AAT ACC AGA TCC CGA CGA AGC GTG
dsDNA	Herpesviridae	Gammaherpesvirus 4	Human	GP110	AKA28433.1	LRRRRRDA 434	CTG AGG CGC CGG AGG CGG GAT GCG
dsDNA	Herpesviridae	Herpesvirus 4 type 2	Human	BALF4	YP_001129508.1	LRRRRRDA 434	CTG AGG CGC CGG AGG CGG GAT GCG
(+) ssRNA	Coronaviridae	Infectious bronchitis	chicken	S	ABG48666.1	TRRFRRSI 539	ACA CGT CGT TTT AGA CGT TCT ATT
(+) ssRNA	Coronaviridae	MHV-3	Mouse	S	ACN89743.1	SRRARRSV 772	TCA CGC AGA GCG CGC CGA TCA GTT
(+) ssRNA	Coronaviridae	MHV	Mouse	S	ABS87264.1	SHRARRSI 760	TCA CAT CGA GCT CGC AGG TCC ATC
(+) ssRNA	Coronaviridae	MHV	Mouse	S	AAD45229.1	SRRARRSV 618	TCA CGC AGA GCC CGC CGA TCA GTT
(+) ssRNA	Coronaviridae	OC43	Human	S	AMK59677.1	TRRSRRAI 766	ACC AGA CGA AGT CGT AGA GCG ATT
(+) ssRNA	Coronaviridae	HKU1	Human	S	AYN64561.1	SRRKRRGI 758	TCT CGG CGT AAG CGT AGA GGT ATT
(+) ssRNA	Coronaviridae	MHV-3	Mouse	S	AFO11517.1	SRRARRSV 630	TCA CGC AGA GCC CGC CGA TCA GTT
(+) ssRNA	Coronaviridae	HKU24	Rat	S	QOE77327.1	TRRAKRDL 765	ACA CGG CGA GCC AAG AGA GAT TTA
(+) ssRNA	Coronaviridae	Equine coronavirus	Horse	S	BAJ52885.1	ARRQRRSP 770	GCA CGT CGT CAG CGT AGG TCA CCT
(+) ssRNA	Coronaviridae	Hemagglut. Enceph.	Pig	S	ASB17086.1	ALRSRRSF 756	GCA CTT AGA TCA CGT AGA TCT TTT
(+) ssRNA	Coronaviridae	Canine resp. cov.	Dog	S	ACL93318.1	KRRSRRSI 770	AAA AGA CGA AGT CGT AGA TCG ATT
(+) ssRNA	Coronaviridae	MERS	Human	S	AHI48550.1	SMLKRRDS 702	TCA ATG CTT AAA CGG CGA GAT TCT
(+) ssRNA	Coronaviridae	Pipistrellus cov. HKU5	Bat	S	AGP04934.1	STRFRRAT 748	TCT ACA CGC TTT CGA CGT GCT ACT
(+) ssRNA	Coronaviridae	Rodent coronavirus	Rat	S	ATP66727.1	ARRKRRAL 755	GCA CGT CGC AAG CGC AGA GCT CTC
(+) ssRNA	Coronaviridae	Longquan Rl rat cov.	Rat	S	QOE77336.1	ARRKRREI 755	GCA CGT CGC AAA CGC AGA GAA ATC
(+) ssRNA	Coronaviridae	Betacoronavirus sp.	Rat	S	AYR18599.1	AHRGRRAL 751	GCG CAT CGC GGC CGC AGA GCT CTG
(+) ssRNA	Coronaviridae	Bovine coronavirus	Calf	S	AGO98871.1	KRRSRRAI 770	AAA AGA CGA AGT CGT AGA GCG ATT
(+) ssRNA	Coronaviridae	Enteric cov. 4408	Human	S	ACT11030.1	KRRSRRAI 770	AAA AGA CGA AGT CGT AGA GCG ATT
(+) ssRNA	Coronaviridae	HKU23	Camel	S	QEY10673.1	KRRSRRAI 770	AAA AGA CGA AGT CGT AGA GCG ATT

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2021 doi:10.20944/preprints202102.0264.v1

(+) ssRNA	Coronaviridae	Bovine cov. E-DB2-TC	Bovine	S	ACT10983.1	KRRSRRAI	770	AAA AGA CGA AGT CGT AGA GCG ATT
(+) ssRNA	Coronaviridae	Bovine cov. DB2	Bovine	S	ABG89288.1	KRRSRRSI	770	AAA AGA CGA AGT CGT AGA TCG ATT
(+) ssRNA	Coronaviridae	Cov. cyc-BetaCoV/2019	Human	S	QLC35798.1	LRRSRRAI	769	AAC AGA CGA AGT CGT AGA GCG ATT
(+) ssRNA	Coronaviridae	Rabbit cov. HKU14	Rabbit	S	YP_005454245.1	QLRSRRAI	769	CAA TTA CGG AGT CGT AGA GCG ATT
(+) ssRNA	Coronaviridae	Giraffe cov. OH3-TC	Giraffe	S	ABP38313.1	KRRSRRSI	765	AAA AGA CGA AGT CGT AGA TCG ATT
(+) ssRNA	Coronaviridae	SARS-CoV-2	Human	S	NC_045512.2	SP**RR**ARSV	687	TCT CCT **CGG CGG** GCA CGT AGT GTA
(+) ssRNA	Flaviviridae	Alkhumra hemor. fever	Sand Tampan	Polyprotein	AFF18429.1	GGRSRRSV	208	GGC GGC AGA AGC AGG AGG TCG GTG
(+) ssRNA	Flaviviridae	Karshi	Mouse	Polyprotein	ABB90671.1	GGRSRRSV	207	GGA GGA CGG TCG CGA AGA TCG GTG
(+) ssRNA	Flaviviridae	Long Pine Key	Mosquito	Polyprotein	ATN29919.1	GRRSRRSV	235	GGC AGG AGG AGC AGG AGA TCG GTG
(+) ssRNA	Flaviviridae	Nhumirim	Mosquito	Polyprotein	YP_009026410.1	HRRSRRSV	224	CAC CGA CGG TCA CGG CGA TCA GTG
(+) ssRNA	Flaviviridae	Nounane	Mosquito	Polyprotein	ACN73462.1	AQRSRRSV	212	GCG CAA CGT TCA AGG AGA TCA GTG
(+) ssRNA	Flaviviridae	Chaoyang	Mosquito	Flav. polyprot.	YP_005454257.1	SRRSRRSV	218	AGT AGA CGC AGC AGA CGA TCT GTT
(+) ssRNA	Flaviviridae	Japanese encephalitis	Human	Polyprotein	AJE59927.1	SRRSRRSV	221	TCC AGG AGA AGT AGA AGA TCT GTG
(+) ssRNA	Flaviviridae	Tick-borne encephalitis	Mouse	Polyprotein	AKC88489.1	GSRTRRSV	208	GGA TCA AGA ACA AGG CGT TCA GTG
(+) ssRNA	Flaviviridae	Louping ill	W. Ptarmig.	Polyprotein	QGA69984.1	GSRTRRSV	207	GGC TCA CGG ACG AGA CGC TCG GTG
(+) ssRNA	Flaviviridae	Dengue virus type 2	Human	Polyprotein	QCZ24972.1	HRREKRSV	207	CAC AGA AGG GAA AAA AGA TCA GTG
(+) ssRNA	Togaviridae	Semliki Forest	Human	Struct. polyprot	ABA29030.1	GTRHRRSV	335	GGA ACA AGA CAC CGG CGC AGC GTG
(-) ssRNA	Bornaviridae	Borna disease 1	Human	Glycoprotein B	VVX76772.1	LKRRRRDT	251	TTG AAA AGG CGG CGT AGG GAT ACC
(-) ssRNA	Filoviridae	Ebola	Macaque	Envelop. gB	ARG43223.1	GRRTRREA	503	GGG AGA AGA ACT CGA AGA GAA GCA
(-) ssRNA	Paramyxoviridae	Newcastle disease	Gull	Fusion prot.	QES91204.1	GRRQRRFI	105	GGA AGG AGA CAG AGA CGT TTT ATA
(-) ssRNA	Paramyxoviridae	Canine morbillivirus	Dog	Fusion prot.	ARQ80424.1	GRRQRRFV	226	GGT AGG AGA CAA AGG CGT TTT GTA
(-) ssRNA	Paramyxoviridae	Bat paramyxovirus	Bat	Fusion prot.	AIF74181.1	SRRRKRFA	112	TCT CGC AGA AGG AAG AGG TTT GCA

* SRARS-CoV-2 is denoted in red

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2021 doi:10.20944/preprints202102.0264.v1

Table 3. Arginine codon usage in NC_045512.2 SARS-CoV-2, isolate Wuhan-Hu-1, genome

Gene	AGG	AGA	CGG	CGA	CGT	CGC	Total
nsp1	0	0	0	1	7	2	10
nsp2	2	5	0	2	7	3	19
nsp3	6	24	3	2	8	2	45
nsp4	2	11	0	0	5	2	20
3C-like proteinase	4	3	0	1	2	1	11
nsp6	1	6	0	0	1	1	9
nsp7	1	1	0	0	0	0	2
nsp8	2	3	0	0	2	0	7
nsp9	2	2	0	1	1	0	6
nsp10	0	0	0	0	1	1	2
nsp12	5	19	2	1	9	7	43
nsp13	2	14	1	2	9	2	30
nsp14A2	1	14	0	0	5	2	22
nsp15-A1	1	4	1	0	2	1	9
nsp16_OMT	2	6	0	0	0	1	9
S	10	20	2	0	9	1	42
ORF3a	1	3	0	0	1	1	6
ORF4	0	1	0	1	1	0	3
ORF5	3	3	0	1	5	2	14
ORF6	1	0	0	0	0	0	1
ORF7a	0	4	0	0	1	0	5
ORF8	0	2	0	0	2	0	4
ORF9	1	10	2	5	6	5	29
ORF10	0	1	0	0	1	0	2
Total	**47**	**156**	**11**	**17**	**85**	**34**	**350**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2021 doi:10.20944/preprints202102.0264.v1

Figure 1. Graphic summary of a SARS-CoV-2 BLASTn search

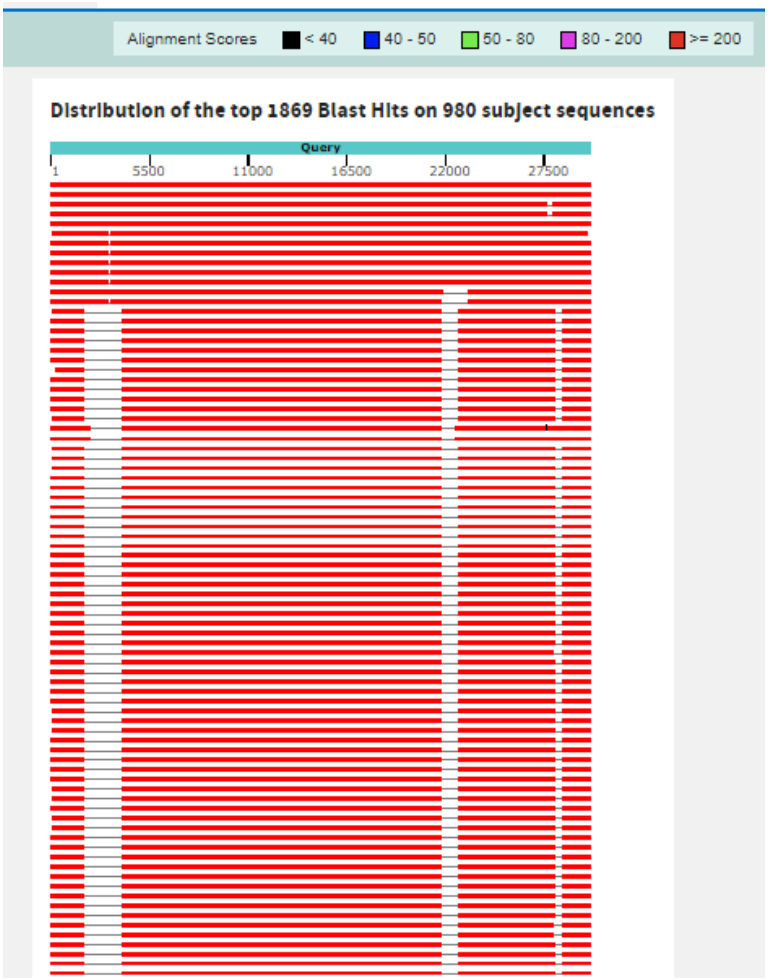

Figure 1. Screenshot of the graphical summary that appears by default in the results of a NCBI-BLAST search. At the top there is a color scale on the alignment scores between query and hit sequences, which are represented for each line (red, maximum value). Also in the upper center, there is a bar with a scale (1-30000) that represents the query sequence that was NC_045512.2 SARS-CoV-2 isolate Wuhan-Hu-1, complete genome sequence (29903 nucleotide length), against the entire NCBI nucleotide collection. Description of the settings and software that was used are included in the Methods. The top hits with a complete query match were (continuous red lines) are (Sequence descripton, GenBanK id, percent identity): Synthetic construct (2019), MT108784.1, 100.00%; Synthetic construct clone icSARS-CoV-2-WT (2020), MT461669.1, 99.99%; Synthetic construct clone icSARS-CoV-2-nLuc-GFP (2020), MT461671.1, 99.99%; Synthetic construct clone icSARS-CoV-2-GFP (2020), MT461670.1, 99.99%; Bat coronavirus RaTG13 (2013), complete genome, MN996532.2, 96.14%; Pangolin coronavirus isolate MP789 (2019), complete genome, MT121216.1, 90.11%; Pangolin coronavirus isolate PcoV_GX-P5L (2017), complete genome, MT040335.1, 85.98%; Pangolin coronavirus isolate PcoV_GX-P4L (2017), complete genome, MT040333.1, 85.97%; Pangolin coronavirus isolate PcoV_GX-P2V (2018), complete genome, MT072864.1, 85.94%; Pangolin coronavirus isolate PcoV_GX-P1E (2017), complete genome, MT040334.1, 85.95%; Pangolin coronavirus isolate PcoV_GX-P5E (2017), complete genome, MT040336.1, 85.95%; Bat SARS-like coronavirus isolate bat-SL-CoVZC45 (2017), complete genome, MG772933.1, 89.12%; Bat SARS-like coronavirus isolate bat-SL-CoVZXC21 (2015), complete genome, MG772934.1, 88.65%. The rest of the hits represent coronavirus genomes that not match in the genomic fingerprints of the SARS-CoV-2 group (discontinuous red lines). So, the vertical white bands on the graphical summary could be seen as the projection of the SARS-CoV-2 related fingerprints (see text).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2021 doi:10.20944/preprints202102.0264.v1

Figure 2. Multiple sequence alignment of the N-terminal domain of S protein

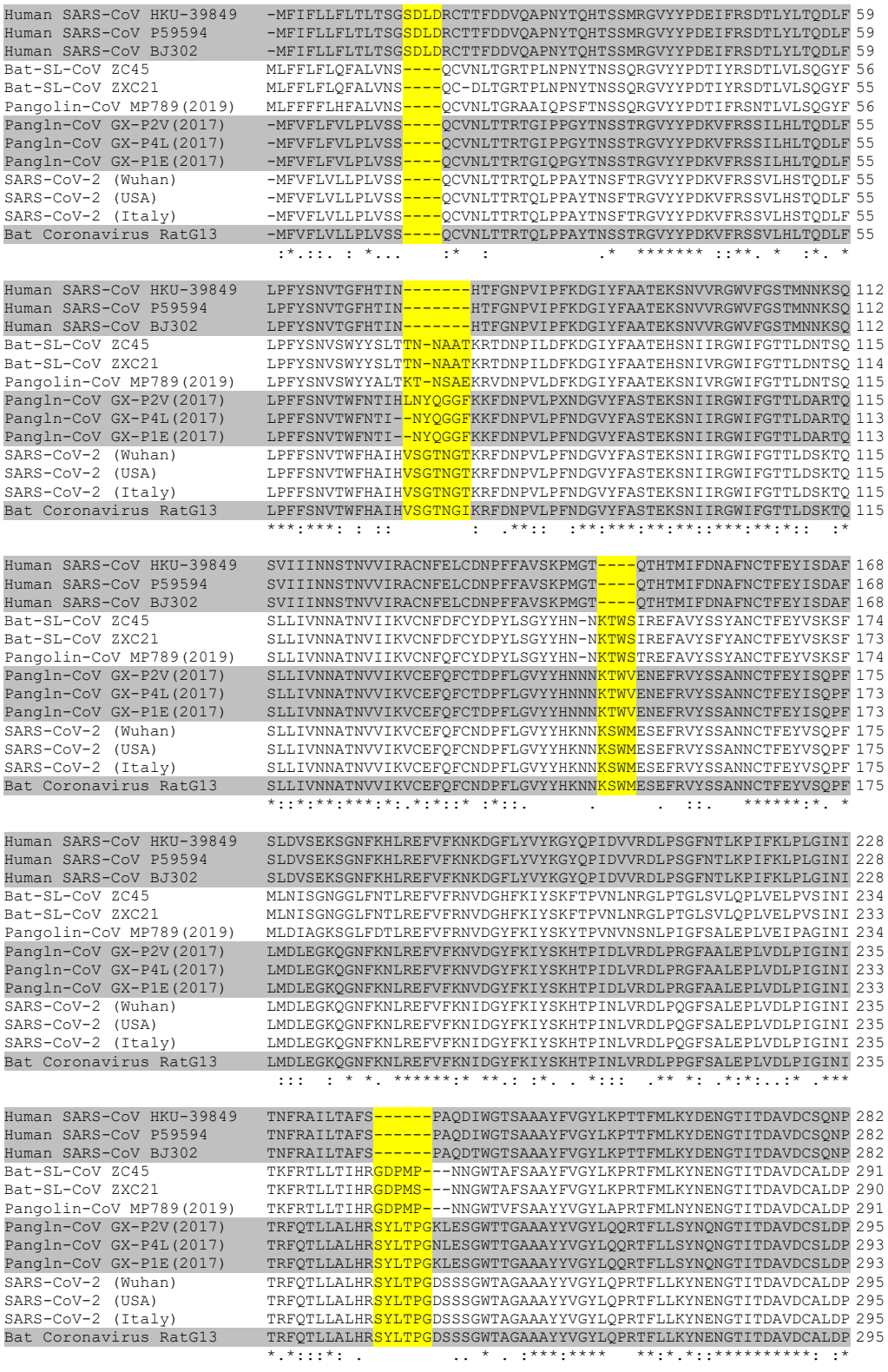

Figure 2. Multialignment of the S protein fragment corresponding to the N-terminal domain. For clarity, only a small group of sequences have been included. Description of the settings and software that was used are included in the Methods. Strictly conserved amino acids are denoted by *, gaps are denoted by -. The position of the amino acids in each sequence is indicated by the numbers to the right. The short deletion and insertions of the SARS-CoV-2 related sequences are highlighted in yellow.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2021 doi:10.20944/preprints202102.0264.v1

Figure 3. Phylogenetic tree of the closely related SARS-CoV-2 coronaviruses based on complete genomes

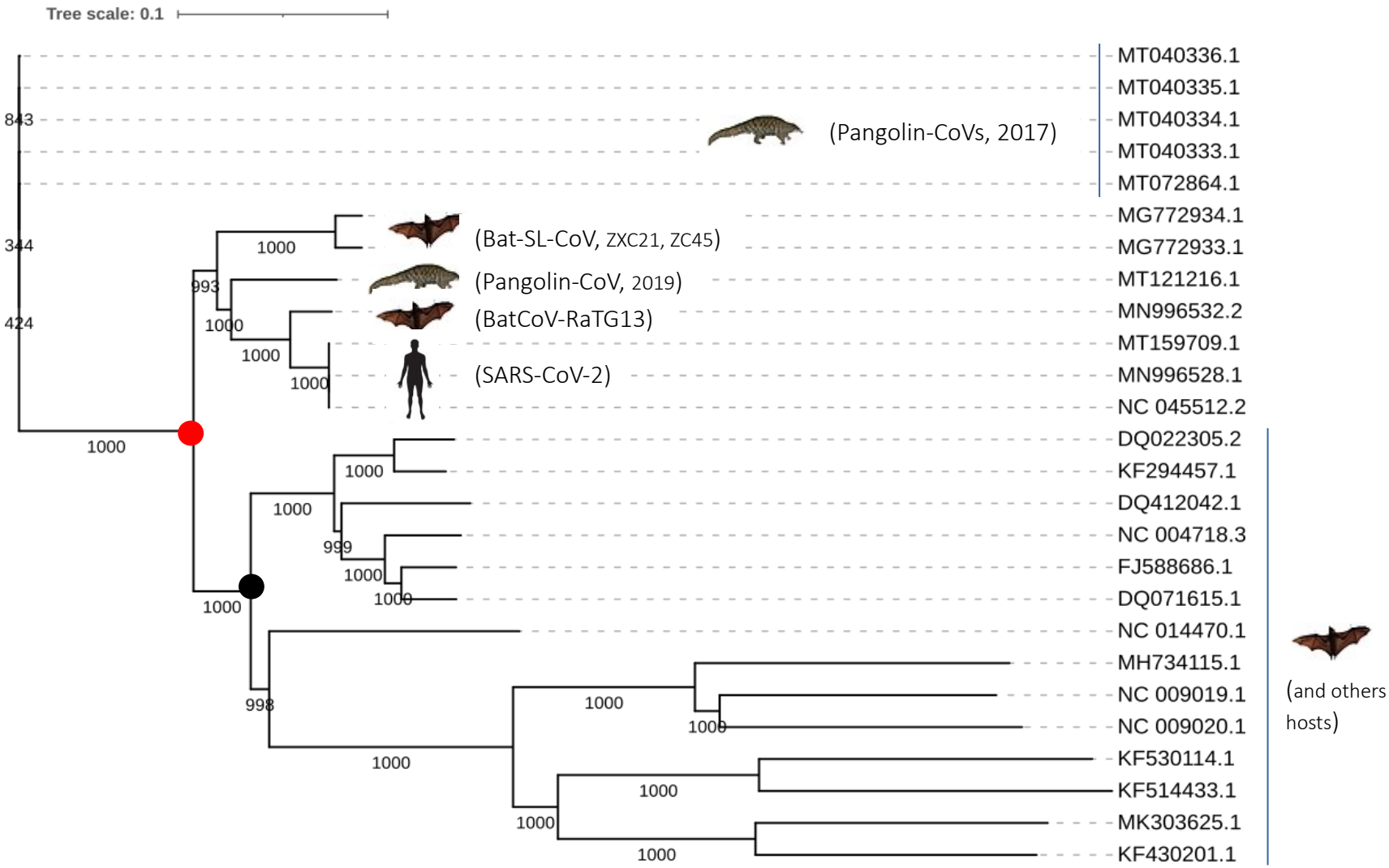

Figure 3. The purpose of the figure is to make the detailed phylogenetic relationship, based on complete genome of the closely related SARS-CoV-2 coronaviruses (Table 1). For better identification, each of these coronavirus is indicated on the corresponding branches of the tree. For simplicity, only three genomes of SARS-CoV-2 have been included. Description of the settings and software that was used are included in the Methods. The black point stands for a common ancestor of the SARS-CoV-2 group, and the red point depicts the divergence of the Pangolin-CoV (2017) from the other SARS-CoV-2 related coronaviruses (bootstrap support 1000). The branches are identifies by the GenBank-id of the corresponding virus genome.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2021 doi:10.20944/preprints202102.0264.v1

Figure 4. Phylogenetic tree of the closely related SARS-CoV-2 coronaviruses based on *S* gene

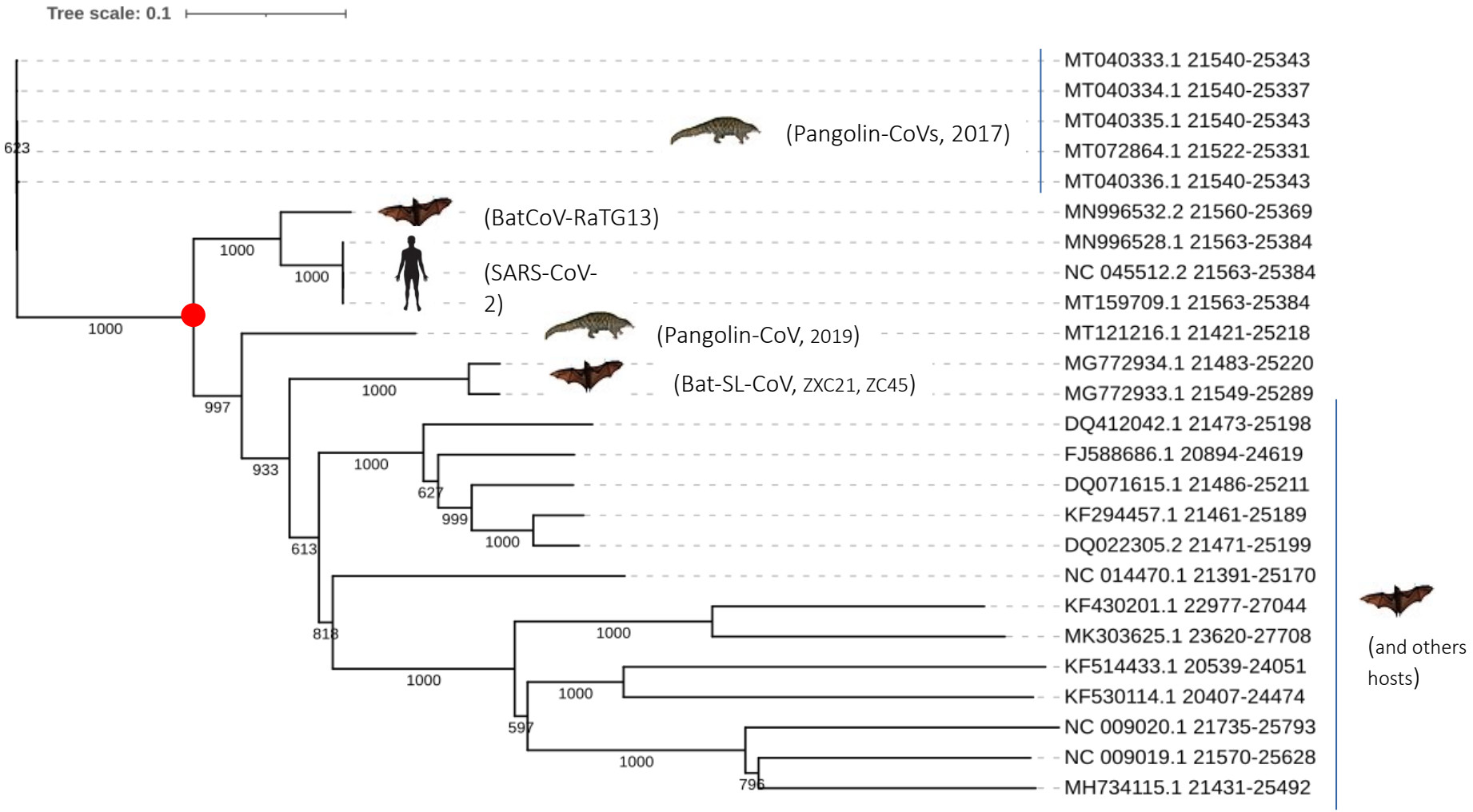

Figure 4. The purpose of the figure is to make the detailed phylogenetic relationships of closely related SARS-CoV-2 coronaviruses (Table 1) based on the nucleotide sequence of the *S* gene region encoding the N-terminal domain and the RBD. Description of the settings and software that was used are included in the Methods. The phylogenetic analysis is in direct comparison with that based on the complete genome (Figure 3). The red point indicates the cluster that groups together Pangolin-CoVs (2017) and, RatG13 and SARS-CoV-2. Branches are identified by the genome coronavirus GenBank accession number, and the *S* gene coordinates.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2021 doi:10.20944/preprints202102.0264.v1

Figure 5. Insertion of 12 nucleotides in the *S* gene encoding the furin site

Figure 5. The S protein of SARSCoV-2 has a functional polybasic furin cleavage site at the S1–S2 boundary through the insertion of 12 nucleotides in the *S* gene. Based on the NCBI-BLASTn pairwise alignment of SARS-CoV-2 (NC_045512.2) and RatG13 (MN996532.2) genomes, it is shown the three possibilities (denoted in yellow) of inserting 12 nucleotides, in strategic positions of RatG13 Serine and Arginine codons to encode PRRA motif in SARS-CoV-2. Point and silent mutations between both genomes, around the furin site, are denoted in pink.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2021 doi:10.20944/preprints202102.0264.v1

Figure 6. Multiple sequence alignment of a fragment of the RBD of S protein

```
Human SARS-CoV HKU-39849    GVIADYNYKLPDDFMGCVLAWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSP 462
Human SARS-CoV P59594       GVIADYNYKLPDDFMGCVLAWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSP 462
Human SARS-CoV BJ302        GVIADYNYKLPDDFMGCVLAWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSP 462
Bat-SL-CoV ZC45             GVIADYNYKLPDDFTGCVIAWNTAKQDV-----GNYFYRSHRSTKLKPFERDLSSDEN-- 464
Bat-SL-CoV ZXC21            GVIADYNYKLPDDFTGCVIAWNTAKQDT-----GHYFYRSHRSTKLKPFERDLSSDEN-- 463
Pangolin-CoV MP789(2019)    GRIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQA 471
Pangln-CoV GX-P2V(2017)     GVIADYNYKLPDDFTGCVIAWNSVKQDALTGGNYGYLYRLFRKSKLKPFERDISTEIYQA 475
Pangln-CoV GX-P4L(2017)     GVIADYNYKLPDDFTGCVIAWNSVKQDALTGGNYGYLYRLFRKSKLKPFERDISTEIYQA 473
Pangln-CoV GX-P1E(2017)     GVIADYNYKLPDDFTGCVIAWNSVKQDALTGGNY--LYRLFRKSKLKPFERDISTEIYQA 471
SARS-CoV-2 (Wuhan)          GKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQA 475
SARS-CoV-2 (USA)            GKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQA 475
SARS-CoV-2 (Italy)          GKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQA 475
Bat Coronavirus RatG13      GKIADYNYKLPDDFTGCVIAWNSKHIDAKEGGNFNYLYRLFRKANLKPFERDISTEIYQA 475
                            * ************ ***:***: : *          **  *  :*:*****:*.

Human SARS-CoV HKU-39849   DGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPATVCGPKLSTDLIK 521
Human SARS-CoV P59594       DGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPATVCGPKLSTDLIK 521
Human SARS-CoV BJ302        DGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPATVCGPKLSTDLIK 521
Bat-SL-CoV ZC45             ------G------VRTLSTYDFNPNVPLEYQATRVVVLSFELLNAPATVCGPKLSTQLVK 512
Bat-SL-CoV ZXC21            ------G------VRTLSTYDFNPNVPLEYQATRVVVLSFELLNAPATVCGPKLSTQLVK 511
Pangolin-CoV MP789(2019)    GSTPCNGVEGFNCYFPLQSYGFHPTNGVGYQPYRVVVLSFELLKAPATVCGPKQSTNLVK 531
Pangln-CoV GX-P2V(2017)     GSTPCNGQVGLNCYYPLERYGFHPTTGVNYQPFRVVVLSXELLNGPATVCGPKLSTTLVK 535
Pangln-CoV GX-P4L(2017)     GSTPCNGQVGLNCYYPLERYGFHPTTGVNYQPFRVVVLSFELLNGPATVCGPKLSTTLVK 533
Pangln-CoV GX-P1E(2017)     GSTPCNGQVGLNCYYPLERYGFHPTTGVNYQPFRVVVLSFELLNGPATVCGPKLSTTLVK 531
SARS-CoV-2 (Wuhan)          GSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVK 535
SARS-CoV-2 (USA)            GSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVK 535
SARS-CoV-2 (Italy)          GSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVK 535
Bat Coronavirus RatG13      GSKPCNGQTGLNCYYPLYRYGFYPTDGVGHQPYRVVVLSFELLNAPATVCGPKKSTNLVK 535
                                      *  *.*  .  : :*  ****** ***:.******** ** *:*
```

Figure 6. Multialignment of the S protein fragment corresponding a region of RBD, with coordinates based on (19). SARS-CoV-2 appears to be optimized for binding to the human receptor ACE2. Based also on (19), the contacting residues in the SARS-CoV RBD and SARS-CoV-2 RBD are denoted in yellow. The purpose of the figure is to highlight that SARS-CoV-2 RBD contacting residues are also shared by Pangolin-CoVs (2017) and RatG13 sequences (highlighted in brown). The sequences are identified by the virus name. Description of the settings and software that was used are included in the Methods.