Evidence of the Recombinant Origin and Ongoing Mutations in Severe Acute Respiratory Syndrome 2 (SARS-COV-2)

Jiao-Mei Huang[1,5], Syed Sajid Jan[2,3,5], Xiaobin Wei[4], Yi Wan[1#] and Songying Ouyang[2,3,5#]

[1]State Key Laboratory of Marine Resource Utilization in South China Sea, Marine College, Key Laboratory of Tropical Biological Resources of Ministry of Education, School of Life and Pharmaceutical Sciences, Hainan University, Haikou, China

[2]The Key Laboratory of Innate Immune Biology of Fujian Province, Provincial University Key Laboratory of Cellular Stress Response and Metabolic Regulation, Biomedical Research Center of South China, Key Laboratory of OptoElectronic Science and Technology for Medicine of the Ministry of Education, College of Life Sciences, Fujian Normal University, Fuzhou, China

[3]Laboratory for Marine Biology and Biotechnology, Pilot National Laboratory for Marine Science and Technology (Qingdao), Qingdao, China

[4]Department of Clinical Laboratory, Haikou People's Hospital, Affiliated Haikou Hospital, Xiangya School of Medicine, Central South University, Haikou, Hainan, China.

[5]Contributed equally to this work

[6]Lead contact

To whom correspondence should be addressed: ouyangsy@fjnu.edu.cn (SO); 993602@hainanu.edu.cn (YW)

SUMMARY

The recent global outbreak of viral pneumonia designated as Coronavirus Disease 2019 (COVID-19) by coronavirus (SARS-CoV-2) has threatened global public health and urged to investigate its source. Whole genome analysis of SARS-CoV-2 revealed ~96% genomic similarity with bat CoV (RaTG13) and clustered together in phylogenetic tree. Furthermore, RaTG13 also showed 97.43% spike protein similarity with SARS-CoV-2 suggesting that RaTG13 is the closest strain. However, RBD and key amino acid residues supposed to be crucial for human-to-human and cross-species transmission are homologues between SARS-CoV-2 and pangolin CoVs. These results from our analysis suggest that SARS-CoV-2 is a recombinant virus of bat and pangolin CoVs. Moreover, this study also reports mutations in coding regions of 125 SARS-CoV-2 genomes suggesting the trajectory of the viral evolution. In short, our findings propose that homologous recombination has been occurred between bat and pangolin CoVs that triggered cross-species transmission and emergence of SARS-CoV-2, and, during the ongoing outbreak, SARS-CoV-2 is still evolving for its adaptability.

Keywords: SARS-CoV-2, pangolin CoVs, recombination, mutations

INTRODUCTION:

The family *Coronaviridae* is comprised of large, enveloped, single stranded, and positive-sense RNA viruses that can infect a wide range of animals including humans (To *et al.,* 2013; Guan *et al.,* 2003). The viruses are further classified into four genera: *alpha, beta, gamma,* and *delta* coronavirus (King *et al.,* 2012). So far, all coronaviruses (CoVs) identified in human belong to the genera *alpha* and *beta*. Among them betaCoVs are of particular importance. Different novel strains of highly infectious betaCoVs have been emerged in human populations in the past two decades that have caused severe health concern all over the world. Severe acute respiratory syndrome coronavirus (SARS-CoV) was first recognized in 2003, causing a global outbreak (Zhong, 2004; Peiris *et al.,* 2004; Cherry, 2004). It was followed by another pandemic event in 2012 by a novel strain of coronavirus designated as Middle East respiratory syndrome coronavirus (MERS-CoV) (Lu *et al.,* 2013). Both CoVs were zoonotic pathogens and evolved in animals. Bats in the genus *Rhinolophus* are natural reservoir of coronaviruses worldwide, and it is presumed that both SARS-CoV and MERS-CoV have been transmitted to human through some intermediate mammalian hosts (Li *et al.,* 2005a; Bolles *et al.,* 2011; Al-Tawfiq and Memish, 2014). Recently, emergence of another pandemic termed as Coronavirus Disease 2019 (COVID-19) by World Health Organization (WHO) caused by a novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been reported (Zhu *et al.,* 2020). To date, more than 174,000 people are infected and over 6,600 death tolls, having transmission clusters worldwide including China, Italy, South Korea, Iran, Japan, USA, France, Spain, Germany and several other countries causing alarming global health concern.

The large trimeric spike glycoprotein (S) located on the surface of CoVs is crucial for viral infection and pathogenesis, which is further subdivided into N-terminal S1 subunit and C-terminal S2 domain. The S1 subunit is specialized in recognizing receptors on host cell, comprising of two separate domains located at N- and C-terminal which can fold independently and facilitate receptor engagement (Masters, 2006). Receptor-binding domains (RBDs) of most CoVs are located on S1 C-terminus and enable attachment to its host receptor (Li *et al.,* 2005b). The host specificity of virus particle is determined by amino acid sequence of RBD and is usually dissimilar among different CoVs. Therefore, RBD is a core determinant for tissue tropism and host range of CoVs. This article presents SARS-CoV-2 phylogenetic trees, comparison and analysis of genome, spike protein, and RBD amino acid

sequences of different CoVs, deducing source and etiology of COVID-19 and evolutionary relationship among SARS-CoV-2 in human.

RESULTS AND DISCUSSION

Phylogenetic classification of SARS-CoV-2 and its closely related CoVs

To determine the evolutionary relationship of the SARS-CoV-2, phylogenetic analysis was performed on whole genomic sequences of different CoVs from various hosts. The Maximum-likelihood (ML) phylogenetic tree is shown in **Figure 1**, which illustrates four main groups representing four genera of CoVs, *alpha, beta, gamma,* and *delta*. In the phylogenetic tree, strains of SARS-CoV-2 (red colored) are cluster together and belong to the genera Betacoronavirus. Among Beta-CoVs, SARS-CoV, Civet SARS CoV, Bat SARS-like CoVs, bat/RaTG13 CoV, and SARS-CoVs-2 clustered together forming a discrete clade from MERS-CoVs. The clade is further divided into two branches and one of the branches comprises all SARS-CoV-2 strains clustered together with Bat/Yunnan/RaTG13 CoV forming a monophyletic group. Bat/Yunnan/RaTG13 exhibited ~96% genomic similarity with SARS-CoV-2. This specifies that SARS-CoV-2 is closely related to Bat/Yunnan/RaTG13 CoV.

The ML phylogenetic tree demonstrates that CoVs from bat source are found in the inner joint or neighboring clade of SARS-CoV-2. This indicates that bats CoVs particularly Bat/Yunnan/RaTG13 are the source of SARS-CoV-2, and they are emerged and transmitted from bats to humans through some recombination and transformation events in intermediate host.

Detection of putative recombination within the spike protein

To explore the emergence of SARS-CoV-2 in humans, we investigated CoVs S-protein and its RBD as they are responsible for determining the host range (**Table 1**). The S-protein amino acid sequence identity between SARS-CoV-2 and related beta-CoVs showed that bat/Yunnan/RaTG13 shares highest similarity of 97.43%. However, the amino acid sequence identity of RBD of SARS-CoV-2 with bat/Yunnan/RaTG13 is 89.57%. On the other hand, Beta-CoVs from pangolin sources (pangolin/Guandong/1/2019 and

pangolin/Guangdong/lung08) revealed highest RBD amino acid sequence identity of 96.68% and 96.08% respectively with SARS-CoV-2. These indication shows the existence of homologous recombination events within the S-protein gene between bat and pangolin CoVs. Similarity plot analysis of CoVs genome sequences from bat, pangolin and human also indicated a possible recombination within S-protein of SARS-CoVs-19 (**Figure S1**).

The amino acid residues change in S-protein of SARS-CoV-2 was further analyzed with SARS-CoV, pangolin and bat CoVs including pangolin/Guandong/1/2019, pangolin/Guangdong/lung08, and bat/Yunnan/RaTG13 (**Figure 2**). Regardless of low homology between SARS-CoV-2 (Wuhan-Hu-1_MN908947) and SARS-CoV (SARS_AAR07630), they had many homologues areas in S-protein. The five key amino acid residues of S-protein at positions 442,472, 479,480, and 487 of SARS-CoVs are described to be at the angiotensin-converting enzyme-2 (ACE2) receptor complex interface and supposed to be crucial for human to human and cross-species transmission (Li *et al.,* 2005b; Wu *et al.,* 2012). **Figure 2b** and **Table S1** describe that all key amino acid residues of RBD (except two positions) are completely homologues between SARS-CoV-2 (Wuhan-Hu-1_MN908947) and pangolin CoVs (pangolin/Guandong/1/2019 and pangolin/Guangdong/lung08), supporting our postulation of recombination event in S-protein gene. Even though, all five crucial amino acid residues of SARS-CoV-2 for binding to ACE2 are different from SARS-CoV, their hydrophobicity and polarity are similar, having same S-protein structural confirmation and identical RBD 3-D structure (Xu *et al.,* 2020). In addition, six critical key residues in MERS-CoV RBD binding to its receptor dipeptidyl peptidase 4 (DPP4) are all different in SARS-CoV and SARS-CoV-2 related coronavirus **(Figure 2a)**.

Ongoing mutations in SARS-CoV-2 during its spread

We also investigated some of the important evolutionary and phylogenetic aspects of SARS-CoV-2 during its spread in human population. Mutation in encoding segments of 125 SARS-CoV-2 genomes obtained from public-domain databases were investigated. In comparison with the first reported SARS-CoV-2 (Wuhan-Hu-1_MN908947), amino acid substitutions were observed at 87 positions of SARS-CoV-2 open reding frames (orfs) (**Figure 3**). Total number of amino acid substitution in corresponding orfs are listed in **Table 2**. Among different orfs of SARS-CoV-2, orf1a was most variable segment with total number of 44

dissimilar amino acid substitutions. It was followed by spike segment S orf with 13 amino acid residue substitutions. However, orf6 and orf7b are the most conserved regions without amino acid changes. In addition, orf10, E, M and orf7a have tended to be more conserved, with only one or two amino acid substitutions.

With the global spread of SARS-CoV-2, its amino acid sequence is also significantly varied **(Figure 3).** Usually, RNA viruses have high rate of genetic mutations, which leads to evolution and provide them with increased adaptability (Lin *et al.,* 2019). To further explore SARS-CoV-2 evolution in human, we have performed phylogenetic analysis based on the aforementioned SARS-CoV-2 in correspondence with their amino acid substitution. **Figure 4.** illustrates the phylogenetic tree of SARS-CoV-2 and associated amino acid changes. Our phylogenetic tree demonstrates that BetaCoV/Chongqing/YC01/2020 is the closest SARS-CoV-2 to bat CoV (Bat/Yunnan/RaTG13) as compared to the first reported SARS-CoV-2 (Wuhan-Hu-1_MN908947). Taking BetaCoV/Chongqing/YC01/2020 as a reference group, South Korea SARS-CoV-2 (BetaCoV/South_Korea/SNU01/2020) has 7 amino acid substitutions in six orfs (S, E, orf1a, orf31, and orf8). BetaCoV/Shenzhen/SZTH-001/2020 has total of 16 amino acid substitutions in four coding regions with highest substitution of 11 amino acid residues in orf1a. United States of America (USA) SARS-CoV-2 (BetaCoV/USA/MA1/2020) has 5 amino acid substitutions in N, orf1a, orf1ab, and orf8. France CoV (BetaCoV/France/IDF0571/2020) has 3 amino acid mutations in orf1a and 1 amino acid mutation in orf8.

CONCLUSION

Based on amino acid and genome sequences analysis and comparison, our results suggest that SARS-CoV-2 is a recombinant virus between bat and pangolin coronaviruses, and the recombination event has been occurred in spike protein genes. Our finding suggest that pangolin is the most possible intermediate SARS-CoV-2 reservoir, which may have given rise to cross-species transmission to humans. These new findings suggest further research to investigate pangolin as a SARS-CoV-2 reservoir. Another important outcome of our analysis is the genetic mutations and evolution of SARS-CoV-2 as it spread globally. These findings are very significant for controlling the SARS-CoV-2 pandemic.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China grants 31770948 and 31570875 and Marine Economic Development Special Fund of Fujian Province (FJHJF-L-2020-2) and the High-level personnel introduction grant of Fujian Normal University (Z0210509). We also gratefully acknowledge the support from Key Research and Development Program (COVID-19) of Hainan (No. ZDYF(XGFY)2020002).

COMPETING INTERESTS: The authors have declared that no competing interests exist.

REFERENCES

Al-Tawfiq, J. A., and Memish, Z. A. (2014). Middle East respiratory syndrome coronavirus: transmission and phylogenetic evolution. Trends Microbiol. 22, 573-579.

Bolles, M., Donaldson, E., and Baric, R. (2011). SARS-CoV and emergent coronaviruses: viral determinants of interspecies transmission. Curr. Opin. Virol. 1, 624-634.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972-1973.

Cherry, J. D. (2004). The chronology of the 2002–2003 SARS mini pandemic. Paediatr. Respir. Rev. 5, 262-269.

Guan, Y., Zheng, B.J., He, Y.Q., Liu, X.L., Zhuang, Z.X., Cheung, C.L., Luo, S.W., Li, P.H., Zhang, L.J., Guan, Y.J. and Butt, K.M., 2003. (2003). Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302, 276-278.

Katoh, K., Rozewicki, J., and Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinformatics 20, 1160-1166.

King, A. M., Adams, M. J., Carstens, E. B., and Lefkowitz, E. J. (2012). Virus taxonomy. Ninth report of the International Committee on Taxonomy of Viruses 486-487.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357.

Li, D., Liu, C. M., Luo, R., Sadakane, K., and Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674-1676.

Li, F., Li, W., Farzan, M., and Harrison, S. C. (2005b). Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science 309, 1864-1868.

Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J.H., Wang, H., Crameri, G., Hu, Z., Zhang, H. and Zhang, J. (2005a). Bats are natural reservoirs of SARS-like coronaviruses. Science 310, 676-679.

Lin, R.W., Chen, G.W., Sung, H.H., Lin, R.J., Yen, L.C., Tseng, Y.L., Chang, Y.K., Lien, S.P., Shih, S.R. and Liao, C. L. (2019). Naturally occurring mutations in PB1 affect influenza A virus replication fidelity, virulence, and adaptability. BMC Biol. 26, 55.

Lole, K.S., Bollinger, R.C., Paranjape, R.S., Gadkari, D., Kulkarni, S.S., Novak, N.G., Ingersoll, R., Sheppard, H.W. and Ray, S. C. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. 73, 152-160.

Lu, L., Liu, Q., Du, L., and Jiang, S. (2013). Middle East respiratory syndrome coronavirus (MERS-CoV): challenges in identifying its source and controlling its spread. Microbes Infect. 15, 625-629.

Masters, P. S. (2006). The molecular biology of coronaviruses. Adv. Virus Res. 66, 193-292.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268-274.

Peiris, J. S. M., Guan, Y., and Yuen, K. Y. (2004). Severe acute respiratory syndrome. Nat. Med. 10, S88-S97.

To, K. K., Hung, I. F., Chan, J. F., and Yuen, K. Y. (2013). From SARS coronavirus to novel animal and human coronaviruses. J. Thorac. Dis. 5, S103.

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189-1191.

Wu, K., Peng, G., Wilken, M., Geraghty, R. J., and Li, F. (2012). Mechanisms of host receptor adaptation by severe acute respiratory syndrome coronavirus. J. Biol. Chem. 287, 8904-8911.

Xu, X., Chen, P., Wang, J., Feng, J., Zhou, H., Li, X., Zhong, W. and Hao, P. (2020). Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. Sci. China, C, Life Sci. 1-4.

Zhong, N. (2004). Management and prevention of SARS in China. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 359, 1115-1116.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R. and Niu, P. (2020). A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med.

MATERIALS AND METHODS

Sequence data collection

One hundred and twenty-five newly sequenced SARS-CoV-2 complete genomes were obtained from Global Initiative on Sharing All Influenza Data EpiFluTM database (GISAID EpiFlu^{TM}) and GenBank. Closely related beta-CoVs genomes sequences from different hosts were also collected and analyzed together with SARS-CoV-2. Open reading frames (orfs) of CoVs genomes were predicted using ORFfinder (v0.4.3) with default parameters ignoring nested orfs.

Raw pair-end reads of pangolin dataset sample (SRR10168377) obtained from NCBI were filtered with bbmap.sh (v38.79) by removing adaptors, trimming low quality reads from both sides (quality value < 20), and reads length less than 50 nt were ignored. Host reference genome (pangolin ManJav1.0, GCF_001685135.1) contaminant reads were removed by bowtie2 (v2.3.5.1) [13]. Pangolin CoV genome fragments were assembled via MEGAHIT (v1.2.9) (Li *et al.,* 2015).

Phylogenetic and recombination analysis

The sequences of CoVs were aligned using multiple sequence alignment MAFFT (v7.450) (Katoh et al., 2019). Aligned sequences were visualized with Jalview (v2.10.3) (Waterhouse *et al.,* 2009). Poorly aligned regions and gaps were removed by trimAL (v1.4.rev22) (Capellagutiérrez *et al.,* 2009). Maximum likelihood (ML) phylogenetic trees of whole genome sequences were constructed in IQ-TREE (v1.6.12) (Nguyen *et al.,* 2015). Support for inferred relationships in the phylogenetic tree was assessed by bootstrap analysis with 1000 replicates and the best-fit substitution model was determined by IQ-TREE model test.

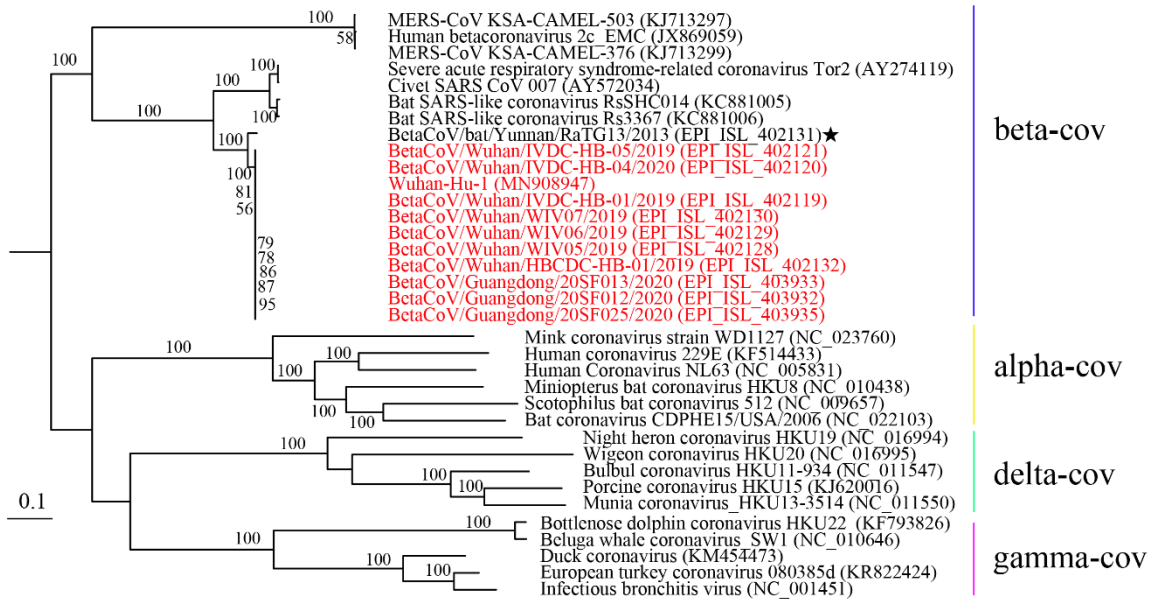
FIGURE LEGENDS

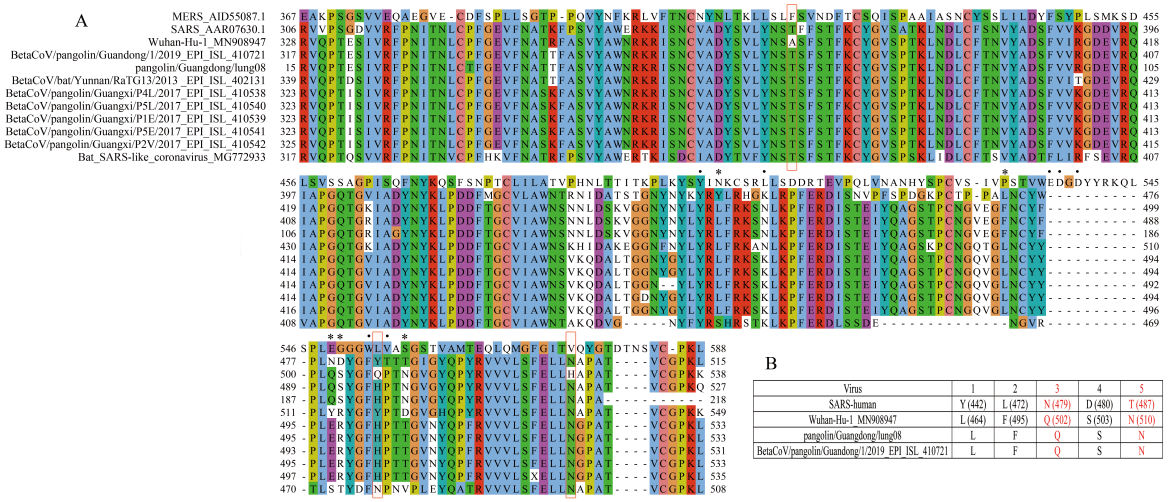
Figure 1. Phylogenetic tree of different CoVs with SARS-CoV-2 based on full genome sequences.

Figure 2. The amino acid residues change in S-protein of CoVs. (A) Amino acid residues change in S1 domain (Asterisks indicate five key amino acid residues of SARS-CoV critical for RBD binding to ACE2； Dots specify critical amino acid residues of MERS-CoV RBD for binding to DPP4). Rectangles specify positions unique in SARS-CoV-2. (B) The combination of five key amino acid residues in SARS-CoVs, SARS-CoV-2, and pangolin CoVs. (C) The amino acid residues change in S2 domain (Rectangles designate changes near two cleavage sites).

Figure 3. Amino acid substitutions in different open reading frames (orfs) of SARS-CoV-2 (125 genomic sequences collected till February 22, 2020).

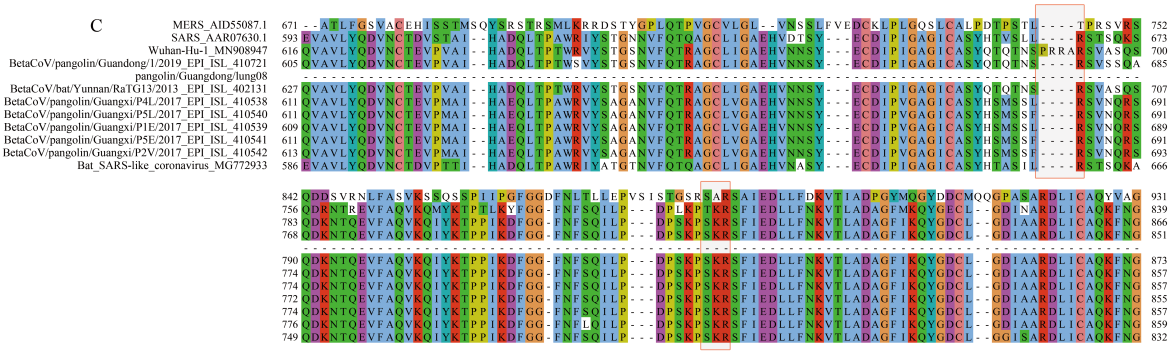
Figure 4. Phylogenetic analysis of SARS-CoV-2 (left) and the number of amino acid substitutions in orfs of corresponding SARS-CoV-2 (right). Bat/Yunnan/RaTG13 was treated as the outgroup.

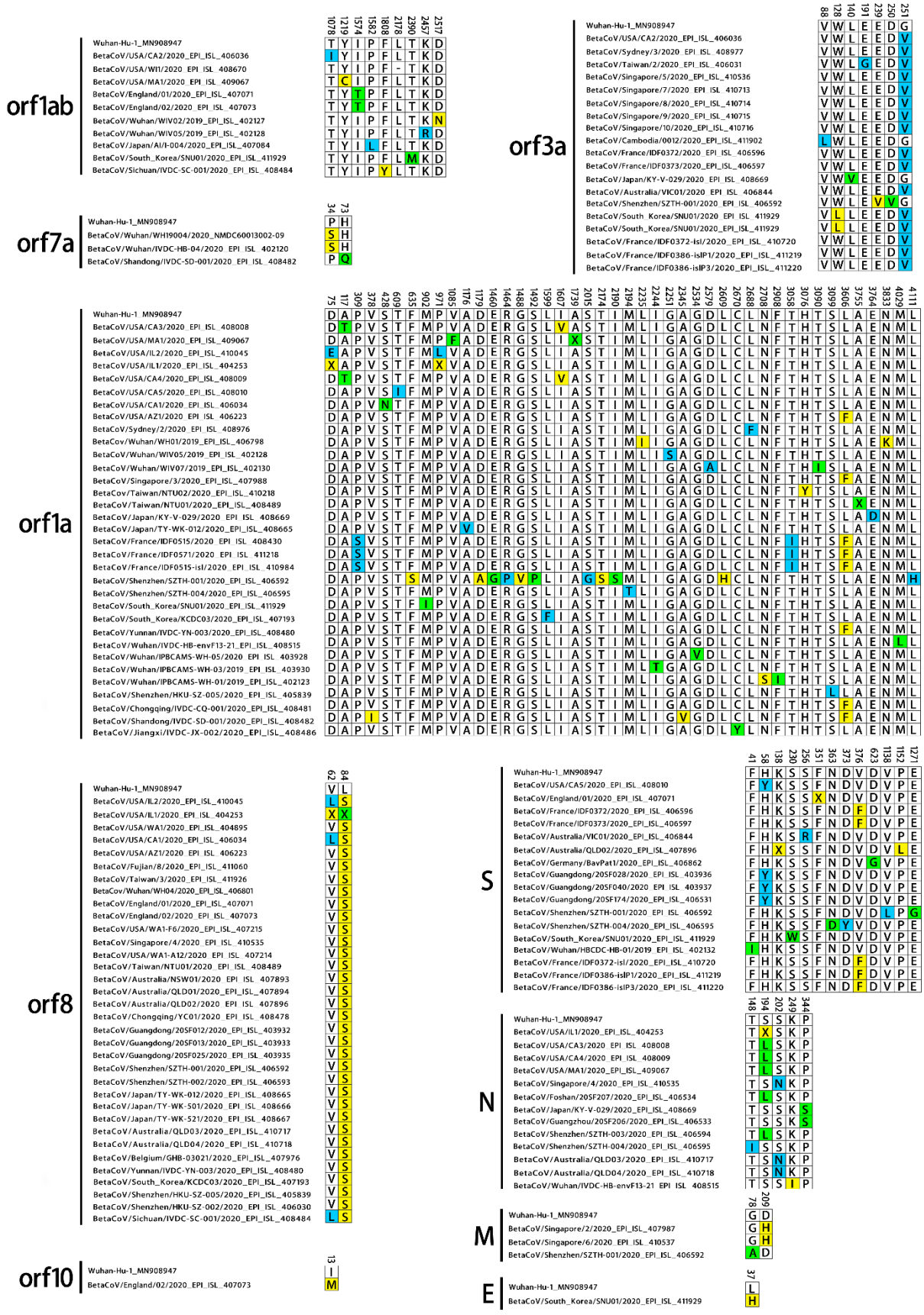


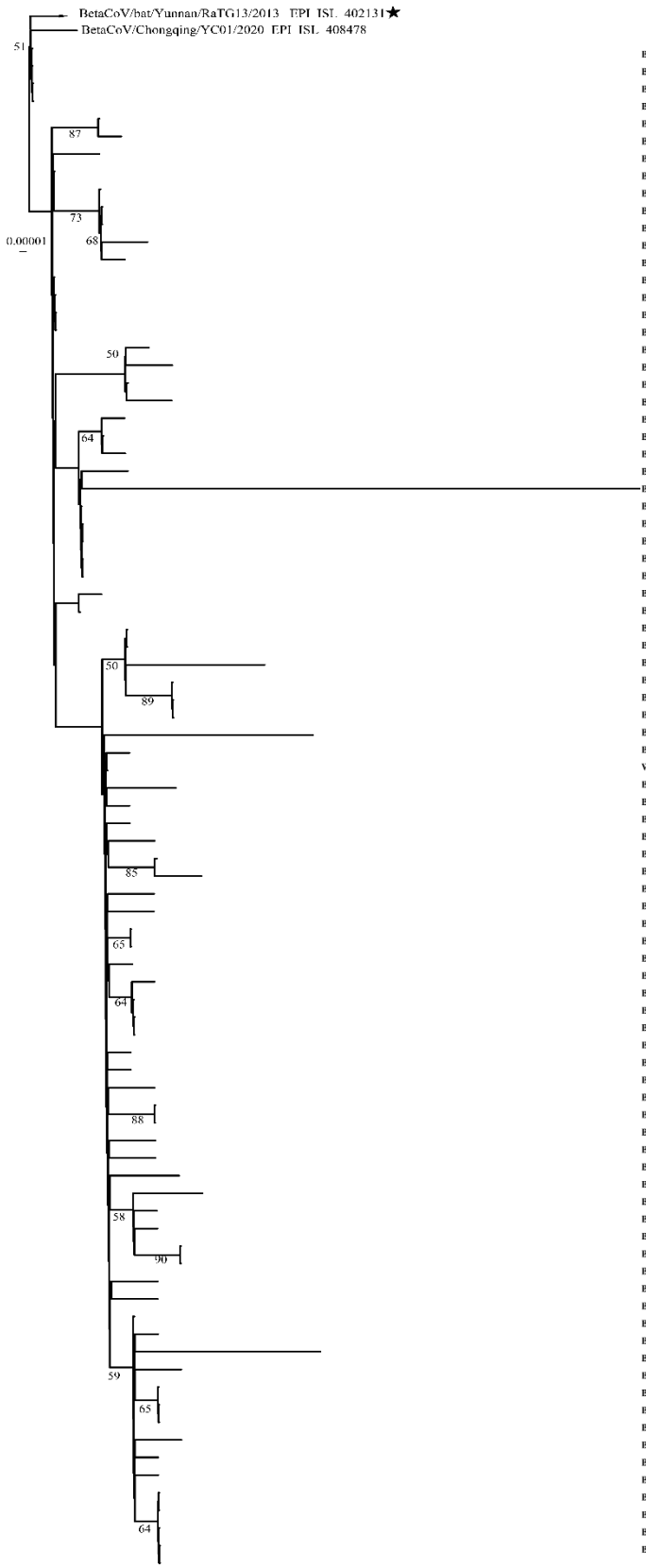


B

Virus	1	2	3	4	5
SARS-human	Y (442)	L (472)	N (479)	D (480)	T (487)
Wuhan-Hu-1_MN908947	L (464)	F (495)	Q (502)	S (503)	N (510)
pangolin/Guangdong/lung08	L	F	Q	S	N
BetaCoV/pangolin/Guandong/1/2019_EPI_ISL_410721	L	F	Q	S	N







	S	E	M	N	orf10	orf1a	orf1ab	orf3a	orf7a	orf8
BetaCoV/Fujian/8/2020_EPI_ISL_411060	0	0	0	0	0	0	0	0	0	0
BetaCoV/USA/WA1/2020_EPI_ISL_404895	0	0	0	0	0	0	0	0	0	0
BetaCoV/USA/WA1-A12/2020_EPI_ISL_407214	0	0	0	0	0	0	0	0	0	0
BetaCoV/USA/WA1-F6/2020_EPI_ISL_407215	0	0	0	0	0	0	0	0	0	0
BetaCoV/England/01/2020_EPI_ISL_407071	1	0	0	0	0	0	1	0	0	0
BetaCoV/England/02/2020_EPI_ISL_407073	0	0	0	0	1	0	1	0	0	0
BetaCoV/South_Korea/KCDC03/2020_EPI_ISL_407193	0	0	0	0	0	1	0	0	0	0
BetaCoV/Taiwan/NTU01/2020_EPI_ISL_408489	0	0	0	0	0	1	0	0	0	0
BetaCoV/Singapore/4/2020_EPI_ISL_410535	0	0	0	1	0	0	0	0	0	0
BetaCoV/Australia/QLD01/2020_EPI_ISL_407894	0	0	0	1	0	0	0	0	0	0
BetaCoV/Australia/QLD04/2020_EPI_ISL_410718	0	0	0	1	0	0	0	0	0	0
BetaCoV/Australia/QLD02/2020_EPI_ISL_407896	2	0	0	1	0	0	0	0	0	0
BetaCoV/Australia/QLD03/2020_EPI_ISL_410717	0	0	0	1	0	0	0	0	0	0
BetaCoV/Belgium/GHB-03021/2020_EPI_ISL_407976	0	0	0	0	0	0	0	0	0	0
BetaCoV/Australia/NSW01/2020_EPI_ISL_407893	0	0	0	0	0	0	0	0	0	0
BetaCoV/Taiwan/3/2020_EPI_ISL_411926	0	0	0	0	0	0	0	0	0	0
BetaCov/Wuhan/WH04/2020_EPI_ISL_406801	0	0	0	0	0	0	0	0	0	0
BetaCoV/Sichuan/IVDC-SC-001/2020_EPI_ISL_408484	0	0	0	0	0	0	1	0	0	1
BetaCoV/USA/CA1/2020_EPI_ISL_406034	0	0	0	0	0	1	0	0	0	1
BetaCoV/USA/IL1/2020_EPI_ISL_404253	0	0	0	1	0	2	0	0	0	2
BetaCoV/USA/IL2/2020_EPI_ISL_410045	0	0	0	0	0	2	0	0	0	1
BetaCoV/Japan/TY-WK-012/2020_EPI_ISL_408665	0	0	0	0	0	1	0	0	0	0
BetaCoV/Japan/TY-WK-501/2020_EPI_ISL_408666	0	0	0	0	0	0	0	0	0	0
BetaCoV/Japan/TY-WK-521/2020_EPI_ISL_408667	0	0	0	0	0	0	0	0	0	0
BetaCoV/Shenzhen/HKU-SZ-005/2020_EPI_ISL_405839	0	0	0	0	0	1	0	0	0	0
BetaCoV/Shenzhen/SZTH-001/2020_EPI_ISL_406592	2	0	1	0	0	11	0	2	0	0
BetaCoV/Guangdong/20SF025/2020_EPI_ISL_403935	0	0	0	0	0	0	0	0	0	0
BetaCoV/Guangdong/20SF012/2020_EPI_ISL_403932	0	0	0	0	0	0	0	0	0	0
BetaCoV/Shenzhen/SZTH-002/2020_EPI_ISL_406593	0	0	0	0	0	0	0	0	0	0
BetaCoV/Guangdong/20SF013/2020_EPI_ISL_403933	0	0	0	0	0	0	0	0	0	0
BetaCoV/Shenzhen/HKU-SZ-002/2020_EPI_ISL_406030	0	0	0	0	0	0	0	0	0	0
BetaCoV/USA/AZ1/2020_EPI_ISL_406223	0	0	0	0	0	1	0	0	0	0
BetaCoV/Yunnan/IVDC-YN-003/2020_EPI_ISL_408480	0	0	0	0	0	1	0	0	0	0
BetaCoV/Chongqing/IVDC-CQ-001/2020_EPI_ISL_408481	0	0	0	0	0	1	0	0	0	1
BetaCoV/Singapore/3/2020_EPI_ISL_407988	0	0	0	0	0	1	0	0	0	1
BetaCoV/Shandong/IVDC-SD-001/2020_EPI_ISL_408482	0	0	0	0	0	3	0	0	1	1
BetaCoV/France/IDF0515/2020_EPI_ISL_408430	0	0	0	0	0	3	0	0	0	1
BetaCoV/France/IDF0515-isl/2020_EPI_ISL_410984	0	0	0	0	0	3	0	0	0	1
BetaCoV/France/IDF0571/2020_EPI_ISL_411218	0	0	0	0	0	3	0	0	0	1
BetaCoV/Shenzhen/SZTH-004/2020_EPI_ISL_406595	2	0	0	1	0	1	0	0	0	1
BetaCoV/Wuhan/HBCDC-HB-01/2019_EPI_ISL_402132	1	0	0	0	0	0	0	0	0	1
Wuhan-Hu-1_MN908947	0	0	0	0	0	0	0	0	0	1
BetaCoV/Wuhan/IPBCAMS-WH-01/2019_EPI_ISL_402123	0	0	0	0	0	2	0	0	0	1
BetaCoV/Wuhan/IPBCAMS-WH-03/2019_EPI_ISL_403930	0	0	0	0	0	1	0	0	0	1
BetaCoV/Cambodia/0012/2020_EPI_ISL_411902	0	0	0	0	0	0	0	1	0	1
BetaCoV/Japan/AI/I-004/2020_EPI_ISL_407084	0	0	0	0	0	0	1	0	0	1
BetaCoV/Guangzhou/20SF206/2020_EPI_ISL_406533	0	0	0	1	0	0	0	0	0	1
BetaCoV/Japan/KY-V-029/2020_EPI_ISL_408669	0	0	0	1	0	1	0	1	0	1
BetaCoV/Wuhan/WIV05/2019_EPI_ISL_402128	0	0	0	0	0	1	1	0	0	1
BetaCoV/Wuhan/WIV07/2019_EPI_ISL_402130	0	0	0	0	0	2	0	0	0	1
BetaCoV/Singapore/2/2020_EPI_ISL_407987	0	0	1	0	0	0	0	0	0	1
BetaCoV/Singapore/6/2020_EPI_ISL_410537	0	0	1	0	0	0	0	0	0	1
BetaCoV/Jiangxi/IVDC-JX-002/2020_EPI_ISL_408486	0	0	0	0	0	1	0	0	0	1
BetaCoV/USA/CA5/2020_EPI_ISL_408010	1	0	0	0	0	1	0	0	0	1
BetaCoV/Guangdong/20SF040/2020_EPI_ISL_403937	1	0	0	0	0	0	0	0	0	1
BetaCoV/Guangdong/20SF028/2020_EPI_ISL_403936	1	0	0	0	0	0	0	0	0	1
BetaCoV/Guangdong/20SF174/2020_EPI_ISL_406531	1	0	0	0	0	0	0	0	0	1
BetaCoV/USA/WI1/2020_EPI_ISL_408670	0	0	0	0	0	0	1	0	0	1
BetaCoV/Wuhan/IPBCAMS-WH-05/2020_EPI_ISL_403928	0	0	0	0	0	1	0	0	0	1
BetaCoV/Wuhan/WIV02/2019_EPI_ISL_402127	0	0	0	0	0	0	1	0	0	1
BetaCoV/Wuhan/IVDC-HB-04/2020_EPI_ISL_402120	0	0	0	0	0	0	0	0	1	1
BetaCoV/Wuhan/WH19004/2020_NMDC60013002-09	0	0	0	0	0	0	0	0	1	1
BetaCoV/Wuhan/IVDC-HB-envF13-21_EPI_ISL_408515	0	0	0	1	0	1	0	0	0	1
BetaCov/Wuhan/WH01/2019_EPI_ISL_406798	0	0	0	0	0	2	0	0	0	1
BetaCoV/Germany/BavPat1/2020_EPI_ISL_406862	1	0	0	0	0	0	0	0	0	1
BetaCoV/USA/MA1/2020_EPI_ISL_409067	0	0	0	1	0	2	1	0	0	1
BetaCoV/Shenzhen/SZTH-003/2020_EPI_ISL_406594	0	0	0	1	0	0	0	0	1	1
BetaCoV/Foshan/20SF207/2020_EPI_ISL_406534	0	0	0	1	0	0	0	0	0	1
BetaCoV/USA/CA3/2020_EPI_ISL_408008	0	0	0	1	0	2	0	0	0	1
BetaCoV/USA/CA4/2020_EPI_ISL_408009	0	0	0	1	0	2	0	0	0	1
BetaCoV/Sydney/2/2020_EPI_ISL_408976	0	0	0	0	0	1	0	0	0	1
BetaCov/Taiwan/NTU02/2020_EPI_ISL_410218	0	0	0	0	0	1	0	0	0	1
BetaCoV/Sydney/3/2020_EPI_ISL_408977	0	0	0	0	0	0	0	1	0	1
BetaCoV/Singapore/7/2020_EPI_ISL_410713	0	0	0	0	0	0	0	1	0	1
BetaCoV/South_Korea/SNU01/2020_EPI_ISL_411929	1	1	0	0	0	1	1	2	0	1
BetaCoV/Australia/VIC01/2020_EPI_ISL_406844	1	0	0	0	0	0	0	1	0	1
BetaCoV/Singapore/9/2020_EPI_ISL_410715	0	0	0	0	0	0	0	1	0	1
BetaCoV/Singapore/10/2020_EPI_ISL_410716	0	0	0	0	0	0	0	1	0	1
BetaCoV/Singapore/5/2020_EPI_ISL_410536	0	0	0	0	0	0	0	1	0	1
BetaCoV/Taiwan/2/2020_EPI_ISL_406031	0	0	0	0	0	0	0	2	0	1
BetaCoV/USA/CA2/2020_EPI_ISL_406036	0	0	0	0	0	0	1	1	0	1
BetaCoV/Singapore/8/2020_EPI_ISL_410714	0	0	0	0	0	0	0	1	0	1
BetaCoV/France/IDF0373/2020_EPI_ISL_406597	1	0	0	0	0	0	0	1	0	1
BetaCoV/France/IDF0386-islP1/2020_EPI_ISL_411219	1	0	0	0	0	0	0	1	0	1
BetaCoV/France/IDF0372-isl/2020_EPI_ISL_410720	1	0	0	0	0	0	0	1	0	1
BetaCoV/France/IDF0372/2020_EPI_ISL_406596	1	0	0	0	0	0	0	1	0	1
BetaCoV/France/IDF0386-islP3/2020_EPI_ISL_411220	1	0	0	0	0	0	0	1	0	1

Table 1. Amino acid sequence identity of S-Protein and RBD of BetaCoVs

Virus	S protein (%)	RBD(%)
BetaCoV/pangolin/Guandong/1/2019_EPI_ISL_410721	90.02	96.68
pangolin/Guangdong/lung08	96.3	96.08
BetaCoV/bat/Yunnan/RaTG13/2013 _EPI_ISL_402131	97.43	89.57
BetaCoV/pangolin/Guangxi/P4L/2017_EPI_ISL_410538	92.18	87.14
BetaCoV/pangolin/Guangxi/P5L/2017_EPI_ISL_410540	92.26	87.14
BetaCoV/pangolin/Guangxi/P1E/2017_EPI_ISL_410539	92.1	86.67
BetaCoV/pangolin/Guangxi/P5E/2017_EPI_ISL_410541	92.18	86.67
BetaCoV/pangolin/Guangxi/P2V/2017_EPI_ISL_410542	92.03	86.67
SARS_AAR07630.1	76.04	73.33
Bat_SARS-like_coronavirus_MG772933	80.69	66.67
MERS_AID55087.1	35	24.24

Table 2. Open reading frame (orf) variation between different SARS-CoV-2

	S	**E**	**M**	**N**	**orf10**	**orf1a**	**orf1ab**	**orf3a**	**orf7a**	**orf8**
Number of dissimilar amino acid substitution	13	1	2	5	1	44	8	7	2	4

SUPPLEMENTAL INFORMATION

Figure S1. Similarity plot of the full genome sequence among Bat/Yunnan/RaTG13, BetaCoV/pangolin/Guandong/1/2019, pangolin/Guangdong/lung08 and Wuhan-Hu-1_MN908947 (as the query). Grey box highlights the Spike protein coding region, supposed to be involved in recombination. Similarity scores between genomic sequences were generated by Simplot (v3.5.1) (Lole *et al.,* 1999).

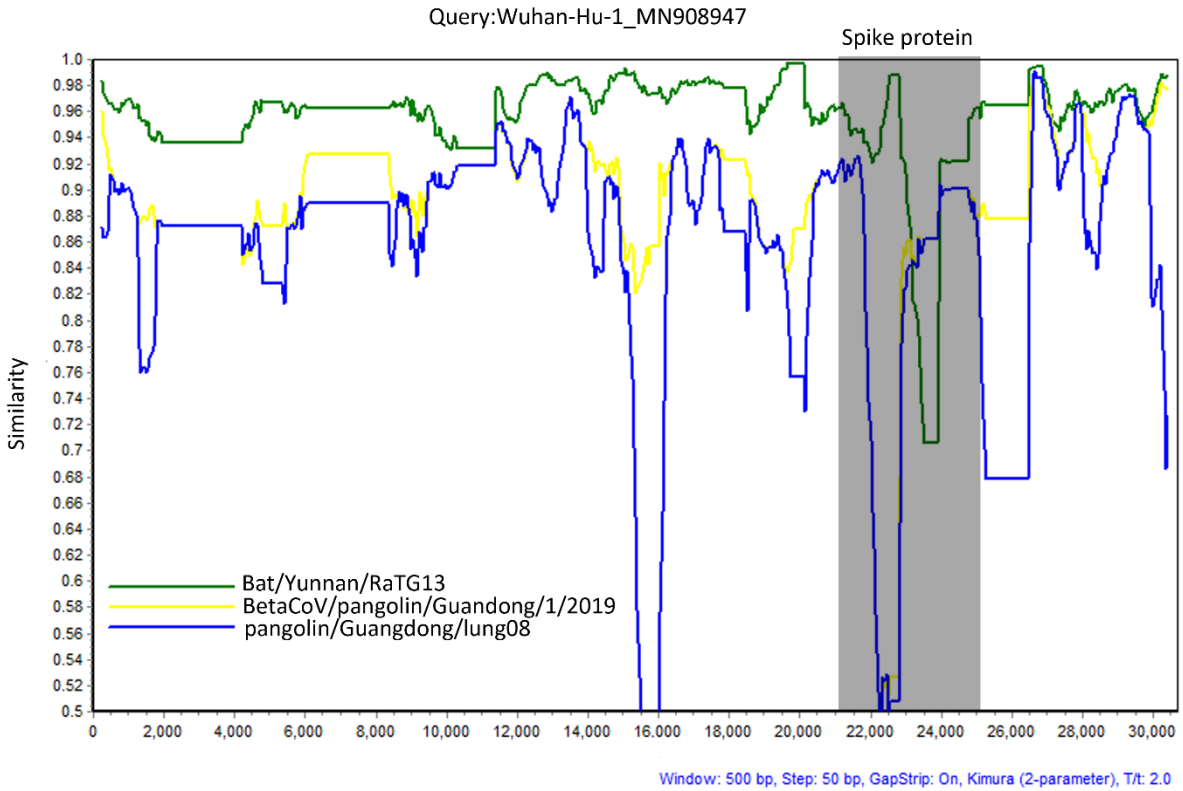


Table S1. The interactions of SARS-CoV RBD with ACE2 and their comparison with RBD of SARS-Cov-2, BetaCoV/pangolin/Guandong/1/2019 and pangolin/Guangdong/lung08.

	SARS-CoV	ACE2	SARS-Cov-2	BetaCoV/pangolin/Guandong/1/2019	pangolin/Guangdong/lung08
Hydrogen bonds	N473(ND2)	Q24(OE1)	N496	N485	N183
		E35(OE1)	Q502	Q491	Q189
	Y491(OH)	E37(OE1)	Y514	Y503	Y201
		E37(OE2)	Y514	Y503	Y201
	Y436(OH)	D38(OD1)	Y458	Y447	Y245
	Y436(OH)	D38(OD2)	Y458	Y447	Y245
	T486(OG1)	Y41(OH)	T509	T498	T296
	T487(N)	Y41(OH)	N510	N499	N297
		Q42(NE2)	Q507	H496	H294
		Q42(NE2)	G455	G444	G242
		Q42(NE2)	Y458	Y447	Y245
	Y436(OH)	Q42(OE1)			
	N473(ND2)	Y83(OH)	N496	N485	N183
	Y475(OH)	Y83(OH)	Y498	Y487	Y185
		Q325(OE1)			
		R426(NH2)			
		N330(ND2)			
	T486(O)	K353(NZ)	Y504	Y493	Y291
		K353(NZ)	G505	G494	G292
	G488(N)	K353(O)	G511	G500	G198
		R393(NH2)	Y514	Y503	Y201
Salt bridges	R426(NH1)	D30(OD2)	K426	R415	R113
		E329(OE2)			
	R426(NH2)	E329(OE2)			

Red color highlights interaction positions of SARS-CoV-2 and pangolin CoVs with different amino acids residues.