Strong evolutionary convergence of receptor-binding protein spike between COVID-19 and SARS-related coronaviruses

Yonghua Wu

Affiliation: School of Life Sciences, Northeast Normal University, 5268 Renmin Street, Changchun, 130024, China

Correspondence: Email: wuyh442@nenu.edu.cn

Abstract

Coronavirus Disease 2019 (COVID-19) and severe acute respiratory syndrome (SARS)-related coronaviruses (e.g., 2019-nCoV and SARS-CoV) are phylogenetically distantly related, but both are capable of infecting human hosts via the same receptor, angiotensin-converting enzyme 2, and cause similar clinical and pathological features, suggesting their phenotypic convergence. Yet, the molecular basis that underlies their phenotypic convergence remains unknown. Here, we used a recently developed molecular phyloecological approach to examine the molecular basis leading to their phenotypic convergence. Our genome-level analyses show that the spike protein, which is responsible for receptor binding, has undergone significant Darwinian selection along the branches related to 2019-nCoV and SARS-CoV. Further examination shows an unusually high proportion of evolutionary convergent amino acid sites in the receptor binding domain (RBD) of the spike protein between COVID-19 and SARS-related CoV clades, leading to the phylogenetic uniting of their RBD protein sequences. In addition to the spike protein, we also find the evolutionary convergence of its partner protein, *ORF3a*, suggesting their possible co-evolutionary convergence. Our results demonstrate a strong adaptive evolutionary convergence between COVID-19 and SARS-related CoV, possibly facilitating their adaptation to similar or identical receptors. Finally, it should be noted that many observed bat SARS-like CoVs that have an evolutionary convergent RBD sequence with 2019-nCoV and SARS-CoV may be pre-adapted to human host receptor ACE2, and hence would be potential new coronavirus sources to infect humans in the future.

Introduction

The 2019 novel coronavirus (2019-nCoV, also called severe acute respiratory syndrome (SARS)-CoV-2) has caused the current outbreak of coronavirus disease (COVID-19), which has emerged as a serious public health concern. The clinical and pathological features caused by 2019-nCoV resemble those seen in SARS [1-3], which is caused by SARS coronavirus (SARS-CoV). Both 2019-nCoV and SARS-CoV have been determined to be of bat origin, with possible intermediate hosts prior to infecting humans [4,5]. Phylogenetic studies have shown that 2019-nCoV and SARS-CoV belong to the subgenus *Sarbecovirus*, but they are distantly related [5-8], with a sequence identity of 79.6% at the whole-genome level [5]. One recent study showed that 2019-nCoV is more similar to a bat coronavirus (RaTG13), with a sequence identity of 96.2% at the whole-genome level, than many other coronaviruses from different hosts, suggesting a phylogenetic affinity of 2019-nCoV to bat coronavirus compared with SARS-CoV [5]. Despite

their relatively distant phylogenetic relationships, 2019-nCoV and SARS-CoV are both known to be capable of infecting humans using the same cell receptor, angiotensin-converting enzyme 2 (ACE2) [5,6,9,10], and their protein structures of receptor-binding protein spike (S) are found to be highly similar to each other [10,11], suggesting their phenotypic convergence. The spike protein is responsible for receptor binding and membrane fusion, and it is important for host tropism and transmission capacity [6]. The spike protein of coronaviruses comprises two subunits, S1 and S2. The S1 subunit contains a receptor-binding domain (RBD), which harbors a receptor-binding motif (RBM) to make complete contact with the receptor (i.e., ACE2) [12,13].

Considering that 2019-nCoV and SARS-CoV are distantly related, but show high similarity in the RBD protein structure and can use the same cell receptor, ACE2 [5-11], an evolutionary convergence may have occurred between them. In the present study, we employ a recently developed molecular phyloecological approach [14-16], which uses a comparative phylogenetic analysis of functional gene sequences to determine the genetic basis of phenotypic evolution, and we examine the possible molecular basis underlying the phenotypic convergence between 2019-nCoV and SARS-CoV. Our results reveal positive selection signals and evolutionary convergent amino acid sites of the spike protein in both 2019-nCoV and SARS-CoV and their related coronaviruses, providing new insights into understanding the evolutionary origin of their phenotypic convergence.

Results and discussion

We used likelihood ratio tests based on the branch and branch-site models implemented in the codeml program of PAML [17] to examine the possible Darwinian selection of all 11 genes annotated in the 2019-nCoV genome (NC_045512). Positively selected genes (PSGs) were found by the branch-site model (Table 1), independent of the initial value variation of parameters (kappa and ω). Specifically, for the branches leading to 2019-nCoV and SARS-CoV, no PSGs were found, nor were PSGs found along the branches leading to their sister coronaviruses. However, we detected PSGs along the ancestral branch (branch C) of 2019-nCoV and its sister taxon, RaTG13, as well as along the ancestral branch (branch K) of SARS-CoV and its sister taxa, WIV16 and Rs4231 (Table 1, Fig. 1). For branch C, three PSGs (*S*, *Orf1ab* and *N*) were found, and for branch K, only one gene (*S*) was found to be under positive selection (Table 1). *Orf1ab* encodes replicase and *N* encodes nucleocapsid [18]. Intriguingly, the *S* gene was subject to Darwinian selection in both branches, C and K. This gene encodes the spike protein, which mediates receptor binding and membrane fusion [6]. The finding of Darwinian selection on the spike protein may suggest its adaptive evolution to the host receptors. To further examine the possible adaptive evolution of 2019-nCoV and SARS-CoV to a human host, we used RELAX [19] to analyze the relative selection intensity change of the 11 genes along the branches leading to 2019-nCoV and SARS-CoV compared with their most recent ancestors, branches C and K, respectively (Tables S1-2). Among the 11 genes examined, gene *S* along the 2019-nCoV branch exhibited a significant selection intensification signal (K = 30.54, p = 0.000, Table S1, Fig. S1), and this remained robust in four independent runs. We also found that the *ORF6* gene showed slight selection intensification (K = 1.81, p = 0.000) along the SARS-CoV branch (Table S2), while its statistical significance only received two supports among five independent runs. For the selectively intensified gene *S* along the 2019-nCoV branch, 0.46% of the amino acid sites (about five amino acids) were under positive selection, while most sites (86.63%) were under purification selection

(Table S1). This may suggest that 2019-nCoV was subject to an adaptive evolution during its adaption to possible intermediate and/or human hosts.

Given the positive selection of the *S* gene in both branches, C and K, we further examined its positive selection signals using branch-site model along all other main branches (Fig. 1) to test whether the positive selection uniquely occurred along the branches related to 2019-nCoV and SARS-CoV. Our results showed positive selection signals along 10 out of 45 branches examined (Fig. 1, Table S3). This may suggest that the *S* gene was widely subject to Darwinian selection in different coronavirus strains, indicating that it may be crucial for the successful survival of coronaviruses. Further analyses of positive selection sites showed that most positive selection sites among all 12 branches under positive selection were located within subunit 1 of *S* gene (Table 1, Table S3), which is used for receptor binding. This may suggest that there are strong selection pressures of different coronavirus strains for their own receptor binding.

Given the selection intensification of *S* gene in 2019-nCoV since its evolutionary divergence from RaTG13, we conducted comparative sequence analyses between the two. We found that there were more than 20 amino acid differences between them, and most of them were located within RBD, especially RBM (Fig. S2), suggesting a high variability of RBM. Given the importance of RBM for receptor binding, we further conducted a comparative sequence analysis among all the coronavirus strains studied to examine its variability. The results showed high sequence variability, with insertion and/or deletion and amino acid substitutions among the coronavirus strains studied (Fig. 2). Despite the high variability of RBM, strikingly, we found that SARS-CoV and its phylogenetic relatives—including WIV16 (KT444582), Rs4231 (KY417146), Rs7327 (KY417151), Rs9401 (KY417152), and BtRs-BetaCoV/YN2018B (MK211376), called SARS-related CoV here, shared many identical or nearly identical amino acids with their phylogenetically distant coronavirus strains, including 2019-nCoV and RaTG13, which we called COVID-19-related CoV (Fig. 2). These shared amino acids were clearly distinct from bat SARS-like CoV that were phylogenetic intermediates between them (Fig. 2). Further analyses showed that such identical amino acids shared between SARS-related CoV and COVID-19-related CoV were not restricted to RBM, but rather, they were scattered throughout the spike protein, with a total of 32 such sites, which were centered on RBD (28 sites in total, Fig. S3). To further examine whether such similarity occurred in other proteins, we analyzed all 11 genes studied among these coronaviruses, and we found that one additional gene, *ORF3a*, contained eight such sites (Fig. 1). The existence of these shared amino acids between SARS-related CoV and COVID-19-related CoV may suggest their high sequence similarity. In support of this, we reconstructed maximum likelihood and neighbor-joining phylogenies using full-length RBD protein sequences, and both showed that SARS-related CoV and COVID-19-related CoV were grouped in the same clade, with relatively high support, which is consistent with two previous studies [6,7], at the same time, their phylogenetic intermediates were clustered in distinct clades (Fig. 3, Fig. S4). The phylogenetic uniting of SARS-related CoV and COVID-19-related CoV provide evidence of their high similarity of RBD protein sequences.

Given their genome-level phylogenetic disparity, the high similarity of RBD protein sequences between SARS-related CoV and COVID-19-related CoV may suggest their evolutionary convergence in the spike protein. To test this possibility, we used an empirical Bayes approach in PAML [17] to reconstruct ancestral

amino acid sequences along internal nodes, and our results showed there were up to 35 evolutionary convergent sites, including 3 convergent and 32 parallel amino acid substitutions that were shared by two ancestral branches leading to SARS-related CoV and COVID-19-related CoV, respectively (Fig. 1). It should be noted that the 35 evolutionary convergent sites of spike protein were apparently underestimated, since those amino acid sites with alignment gaps were not considered by the approach used. Still, these 35 sites represented an unusually high incidence of evolutionary convergence sites, which have rarely been found in previous studies related to molecular convergent evolution [20-26]. This suggests a strong evolutionary convergence between SARS-related CoV and COVID-19-related CoV. On completion of our data analyses, one of the most recent studies showed that the RBD of the spike protein of Pangolin-CoV is nearly identical to that of 2019-nCoV, with only one amino acid difference [27], suggesting that Pangolin-CoV also belongs to the clade of COVID-19-related CoV. It should be noted that the RBD sequences of two other coronavirus strains, BM48-31 (GU190215) and BtKY72 (KY352407), were grouped with SARS-related CoV and COVID-19-related CoV (Fig. 3, Fig. S4), suggesting an evolutionary convergence among them. In addition to these evolutionarily convergent coronavirus strains, intriguingly, we found the evidence of evolutionary convergence of the spike protein between the ancestral branch leading to SARS-related CoV and the ancestral branch of COVID-19-related CoV and its sister strains (CoVZC45 and CoVZXC21), which harbored 9 evolutionary convergent amino acid sites (Fig. S5). These results suggest that the spike protein may have been subjected to a successive evolutionary convergence among ancestral coronavirus strains leading to SARS-related CoV, COVID-19-related CoV and CoVZC45 and CoVZXC21.

Previous studies show that spike protein interacts tightly with a related protein *ORF3a* and they likely coevolved [28-30]. If their coevolution does occur, we may expect that the evolutionary convergence of the spike protein found may have led to the occurrence of the evolutionary convergence of *ORF3a* as well. To test this, we reconstructed ancestral amino acid sequences of *ORF3a* along internal nodes, and our results revealed 6 parallel amino acid substitutions shared between the ancestral branch leading to SARS-related CoV and the ancestral branch of COVID-19-related CoV and its sister strains (CoVZC45 and CoVZXC21) (Fig. S5). And we also detected a parallel amino acid substitution of *ORF3a* between the two ancestral branches leading to SARS-related CoV and COVID-19-related CoV (Fig. S6). Considering that these evolutionary convergent branches of *ORF3a* also showed evolutionary convergence in spike protein as mentioned above (Fig. 1, Fig. S5), it may suggest that spike protein and its partner protein *ORF3a* may have been subjected to a co-evolutionary convergence.

Evolutionary convergence may occur by chance or by Darwinian selection. Our results showed that the evolutionary convergent sites found in this study were mainly restricted to two genes (*S* and *ORF3a*, Fig. 1), and in particular, they were centered within the RBD of the *S* gene. This biased distribution of evolutionary convergent sites is difficult to explain according to chance; rather, Darwinian selection would be favored as a plausible explanation. In support of this, we used CONVERG2 [31] to evaluate the probability of the occurrence of our observed convergent sites of spike protein between the two ancestral branches leading to SARS-related CoV and COVID-19-related CoV, and the results showed high statistical significance (p = 0.000000), regardless of whether the JTT model or Poisson model was used. This result apparently rejects chance or neutral evolution as a possible explanation; rather, it indicates a

predominately strong Darwinian selection. Moreover, we observed an apparently accelerated evolution of RBD of SARS-related CoV and COVID-19-related CoV related to their phylogenetic intermediates (Fig. 3), and we detected a significant Darwinian selection of the *S* gene along two branches (branches C and K) of SARS-related CoV and COVID-19-related CoV (Fig. 1, Table 1). These lines of evidence may strongly support the evolutionary convergence found in this study as a result of adaptive evolution. Regarding the possible adaptive evolutionary convergence, previously proposed causes, such as gene duplication and horizontal gene transfer [21,32,33], are less likely because only single-copy *S* genes were found in all 35 genomes examined and evolutionary convergent sites presented an apparently biased distribution pattern. Parallel and/convergent evolution, which occur through point mutation, could contribute to our observed evolutionary convergence, but it could not account for the unusually high incidence of convergent sites observed in this study, representing a rare finding in previous studies [20-26]. Recent studies have shown a relatively high likelihood of occurrence of homologous recombination in spike protein [7,34,35] , and especially, it is considered that the RBD of 2019-nCov may be derived from a recombination event between that of human SARS-CoV and another (unsampled) SARS-like CoV [35]. If this is the case, the homologous recombination, if any, may have occurred between the ancestors (branches C and K) of SARS-related CoV and COVID-19-related CoV, accounting for their unusually high incidence of convergent sites observed in this study.

Given the strong evolutionary convergence of RBD of spike protein between the two clades, COVID-19-related CoV and SARS-related CoV, the coronaviruses of the two clades may have more likely adapted to similar or the same receptor. To date, 2019-nCoV and SARS-CoV have been known to be capable of using the ACE2 receptor in human host [5,6,9,10], but the receptors of their phylogenetic relatives from the two clades, COVID-19-related CoV and SARS-related CoV, are less clear [4,5,36,37]. Regarding the bat SARS-like CoV, Rs4231 and Rs7327 are known to be able to use human ACE2 receptor [37], while WIV16 is capable of using the ACE2 receptor from humans, civets and Chinese horseshoe bats (*Rhinolophus sinicus*) [4]. The receptors of Rs9401, BtRs-BetaCoV/YN2018B, and BatCoV RaTG13 remain to be explored. Further studies on the receptors of these bat SARS-like CoVs in their natural reservoirs are badly needed to determine whether ACE2 or other candidates, if any, represent their shared cell receptor, leading to their strong evolutionary convergence of spike protein.

Our molecular phyloecological study demonstrates that spike protein shows significant Darwinian selection along two ancestral branches related to SARS-CoV and 2019-nCoV, suggesting their adaptive evolution to recognizing their own cell receptors. Comparative sequence and phylogenetic analyses indicate a high similarity of RBD sequences of spike protein between SARS-related CoV and COVID-19-related CoV. Subsequent ancestral sequence reconstruction and convergent evolution analyses reveal an unusually high incidence of parallel and convergent amino acid substitutions between them, suggesting an extremely strong adaptive evolutionary convergence in spike protein. In addition to spike protein, we also found evolutionary convergence of its partner protein, *ORF3a*, suggesting their possible co-evolutionary convergence. Finally, considering that SARS-CoV and 2019-nCoV have posed serious concerns to public health and safety, it should be noted that many other bat SARS-like CoV strains that were evolutionarily convergent with SARS-CoV and 2019-nCoV recognized in this study may be potential novel coronaviruses to infect humans in the future.

Materials and method

Taxa and sequences

We used 35 coronavirus strain genomes of the subgenus *Sarbecovirus* based on two published studies [5,6], including 2019-nCoV, SARS-CoV, and their phylogenetic relative, bat SARS-like CoV (please see Fig. 1 for details). For all these coronavirus strains, we downloaded their full-length genome sequences from GenBank except for five 2019-nCoV and RaTG13 strains, which were downloaded from GISAID. We aligned these genome sequences using MAFFT (https://mafft.cbrc.jp/alignment/server/). The coding sequences of 11 genes (Fig. 1) annotated in the genome of 2019-nCoV (NC_045512) were used as a reference sequence to obtain the homologous gene sequences of the 35 coronavirus strains. We aligned these homologous gene sequences using the online software webPRANK (http://www.ebi.ac.uk/goldman-srv/webprank/) [38], which is considered to create a more reliable alignment to decrease false-positive results in positive selection analyses [39].

Adaptive evolution analyses

We employed the branch and branch-site models implemented in the codeml program of PAML [17] to examine the adaptive evolution of our focal genes. For this, a codon-based maximum-likelihood method was used to estimate the ratio of non-synonymous to synonymous substitutions per site (dN/dS or ω), and likelihood ratio tests (LRTs) were used to calculate statistical significance. A statistically significant value of $\omega > 1$ suggests positive selection. Upon analysis, an unrooted taxon tree (Fig. 1) was constructed based on two published studies [5,6]. For branch model analysis, we used a two-rate branch model, and our focal branches were labelled as foreground branches, while others were treated as background branches. The two-rate branch model was compared with the one-rate branch model, which assumes a single ω value across the tree, to determine statistical significance. If a statistically significant value of $\omega > 1$ in a foreground branch was detected, the two-ratio branch model was then compared with the two-ratio branch model with a constraint of $\omega = 1$ to further determine whether the $\omega > 1$ of the foreground branch was statistically significant. In addition to the branch model, we also used a branch-site model (Test 2) to detect positively selected sites for a particular branch. Test 2 compares a modified model A with its corresponding null model with a constraint of $\omega = 1$ to determine the statistical significance. Positively selected sites were found using an empirical Bayes method. For result robustness, we evaluated the dependence of the signal of positive selection on parameter variation. For this, we used two different initial values of kappa (kappa = 0.5, 3.0) and of omega ($\omega = 0.5, 2.0$), and eventually, several independent runs were conducted for each of the positively selected genes found.

Selection intensity analyses

We analyzed relative selection intensity using the RELAX [19] program, available from the Datamonkey webserver (http://test.datamonkey.org/relax). RELAX is a hypothesis testing framework, and it can be used to test whether selection strength has been relaxed or intensified along a certain branch or lineage. For analyses, RELAX calculates a selection intensity parameter value (k), and $k > 1$ shows an intensified selection, while $k < 1$ indicates a relaxed selection, assuming a priori partitioning of the test branches and reference branches. We would expect that an intensified selection shows ω categories away from neutrality ($\omega = 1$), while a relaxed selection is expected to show ω categories converging to neutrality ($\omega = 1$). Statistical significance was evaluated by LRT by comparing an alternative model with

a null model. The null model assumes k = 1 and the same ω distribution for both test and reference branches, while the alternative model assumes that k is a free parameter, and the test and reference branches may have different ω distributions.

Phylogenetic analyses

We reconstructed an maximum likelihood (ML) tree and neighbor-joining (NJ) tree using MEGA X [40]. For ML analyses, the WAG+ G model was selected as the best amino acid substitution model according to the Bayesian information criterion. All amino acid sites with an alignment gap were included for analyses. For NJ analyses, JTT+G was selected as the best model. For the ML and NJ analyses, the bootstrap value was set to 1, 000. Other parameters were used as defaults in the program.

Ancestral sequence reconstruction

We used the amino acid-based marginal reconstruction implemented in the empirical Bayes approach in PAML [17] for ancestral sequence reconstruction. In the analyses, the character was assigned to a single interior node and the character with the highest posterior probability was used as the best reconstruction. We used two different amino acid substitution models, JTT and Poisson, to examine the consistency of our results. The JTT model assumes different substitution rates of different amino acids, while the Poisson model assumes the same substitution rate of all amino acids. For the analyses, we obtained the full-length spike protein sequences of our focal coronavirus strains and used their phylogeny, as given in Fig. 1. The amino acid substitutions along our focal branches were analyzed. The results based on the JTT and Poisson models were generally identical; for convenience, only the results based on the JTT model are shown.

Convergent evolution analyses

We used the CONVERG2 [31] program to evaluate the probabilities that the observed convergent and parallel substitutions were due to random chance. A statistical significant *p*-value may suggest the observed evolutionary convergent sites are less likely attributable to random chance, but instead, favor Darwinian selection as a possible explanation. For the analyses, two different amino acid substitution models, JTT and Poisson, were used. The RBD amino acid sequences of our focal 35 coronavirus genomes were abstracted and aligned using CLUSTAL W [41] program. The phylogenetic relationships among the coronavirus strains studied are given in Fig. 1.

Acknowledgements

We thank Hui Wang for helping with the convergent evolution analyses. This research was supported by the National Natural Science Foundation of China (grant number, 31770401) and the Fundamental Research Funds for the Central Universities.

References

1 Guan, W.-j. *et al.* Clinical characteristics of 2019 novel coronavirus infection in China. *medRxiv*, doi:https://doi.org/10.1101/2020.02.06.20020974 (2020).

2 Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. *Journal of Virology*, doi:10.1128/JVI.00127-20 (2020).

3 Xu, Z. *et al.* Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *The Lancet Respiratory Medicine*, doi:https://doi.org/10.1016/S2213-2600(20)30076-X (2020).

4 Yang, X.-L. *et al.* Isolation and characterization of a novel bat coronavirus closely related to the direct progenitor of severe acute respiratory syndrome coronavirus. *Journal of Virology* **90**, 3253-3256 (2016).

5 Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, doi:10.1038/s41586-020-2012-7 (2020).

6 Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* **395**, 565-574 (2020).

7 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature*, 1-5, doi:https://doi.org/10.1038/s41586-020-2008-3 (2020).

8 Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine* **382**, 727-733, doi:10.1056/NEJMoa2001017 (2020).

9 Hoffmann, M. *et al.* The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. *bioRxiv*, doi:https://doi.org/10.1101/2020.01.31.929042 (2020).

10 Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, eabb2507, doi:10.1126/science.abb2507 (2020).

11 Lan, J. *et al.* Crystal structure of the 2019-nCoV spike receptor-binding domain bound with the ACE2 receptor. *bioRxiv*, doi:https://doi.org/10.1101/2020.02.19.956235 (2020).

12 Du, L. *et al.* The spike protein of SARS-CoV—a target for vaccine and therapeutic development. *Nature Reviews Microbiology* **7**, 226-236 (2009).

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

14 Wu, Y. & Wang, H. Convergent evolution of bird-mammal shared characteristics for adapting to nocturnality. *Proceedings of the Royal Society B: Biological Sciences* **286**, 20182185, doi:10.1098/rspb.2018.2185 (2019).

15 Wu, Y., Wang, H. & Hadly, E. A. Invasion of ancestral mammals into dim-light environments inferred from adaptive evolution of the phototransduction genes. *Scientific Reports* **7**, 46542, doi:https://doi.org/10.1038/srep46542 (2017).

16 Wu, Y., Wang, H., Wang, H. & Hadly, E. A. Rethinking the origin of primates by reconstructing their diel activity patterns using genetics and morphology. *Scientific Reports* **7**, 11837, doi:10.1038/s41598-017-12090-3 (2017).

17 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).

18 McBride, R. & Fielding, B. C. The role of severe acute respiratory syndrome (SARS)-coronavirus accessory proteins in virus pathogenesis. *Viruses* **4**, 2902-2923 (2012).

19 Wertheim, J. O., Murrell, B., Smith, M. D., Kosakovsky Pond, S. L. & Scheffler, K. RELAX: detecting relaxed selection in a phylogenetic framework. *Molecular Biology and Evolution* **32**, 820-832 (2015).

20 Hu, Y. *et al.* Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 1081-1086 (2017).

21 Li, Y., Liu, Z., Shi, P. & Zhang, J. The hearing gene Prestin unites echolocating bats and whales. *Current Biology* **20**, R55-R56 (2010).

22 Nagai, H. *et al.* Reverse evolution in RH1 for adaptation of cichlids to water depth in Lake Tanganyika. *Molecular Biology and Evolution* **28**, 1769-1776 (2011).

23 Shen, Y.-Y., Liang, L., Li, G.-S., Murphy, R. W. & Zhang, Y.-P. Parallel evolution of auditory genes for echolocation in bats and toothed whales. *PLoS Genetics* **8**, e1002788, doi:10.1371/journal.pgen.1002788 (2012).

24 Ujvari, B. *et al.* Widespread convergence in toxin resistance by predictable molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 11911-11916 (2015).

25 Zhang, J. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nature genetics* **38**, 819-823 (2006).

26 Zhen, Y., Aardema, M. L., Medina, E. M., Schumer, M. & Andolfatto, P. Parallel molecular evolution in an herbivore community. *Science* **337**, 1634-1637 (2012).

27 Xiao, K. *et al.* Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan Pangolins. *bioRxiv*, doi:https://doi.org/10.1101/2020.02.17.951335 (2020).

28 Tan, Y.-J. The Severe Acute Respiratory Syndrome (SARS)-coronavirus 3a protein may function as a modulator of the trafficking properties of the spike protein. *Virology Journal* **2**, 5, doi:10.1186/1743-422X-2-5 (2005).

29 Yount, B. *et al.* Severe acute respiratory syndrome coronavirus group-specific open reading frames encode nonessential functions for replication in cell cultures and mice. *Journal of Virology* **79**, 14909-14922 (2005).

30 Zeng, R. *et al.* Characterization of the 3a protein of SARS-associated coronavirus in infected vero E6 cells and SARS patients. *Journal of Molecular Biology* **341**, 271-279 (2004).

31 Zhang, J. & Kumar, S. Detection of convergent and parallel evolution at the amino acid sequence level. *Molecular Biology and Evolution* **14**, 527-536 (1997).

32 Li, G. *et al.* The hearing gene Prestin reunites echolocating bats. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 13959-13964 (2008).

33 Stewart, C.-B., Schilling, J. W. & Wilson, A. C. Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* **330**, 401-404 (1987).

34 Ji, W., Wang, W., Zhao, X., Zai, J. & Li, X. Cros-species transmission of the newly identified coronavirus 2019-nCoV. *Journal of Medical Virology* **92**, 433-440 (2020).

35 Patino-Galindo, J. A., Filip, I., AlQuraishi, M. & Rabadan, R. Recombination and convergent evolution led to the emergence of 2019 Wuhan coronavirus. *bioRxiv*, doi:https://doi.org/10.1101/2020.02.10.942748 (2020).

36 Han, Y. *et al.* Identification of diverse bat alphacoronaviruses and betacoronaviruses in China provides new insights into the evolution and origin of coronavirus-related diseases. *Frontiers in Microbiology* **10**, 1900, doi:10.3389/fmicb.2019.01900 (2019).

37 Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathogens* **13**, e1006698, doi:1371/journal.ppat.1006698 (2017).

38 Löytynoja, A. & Goldman, N. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics* **11**, 579, doi:10.1186/1471-2105-11-579 (2010).

39 Fletcher, W. & Yang, Z. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Molecular Biology and Evolution* **27**, 2257-2267 (2010).

40 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution* **35**, 1547-1549 (2018).

41 Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673-4680 (1994).

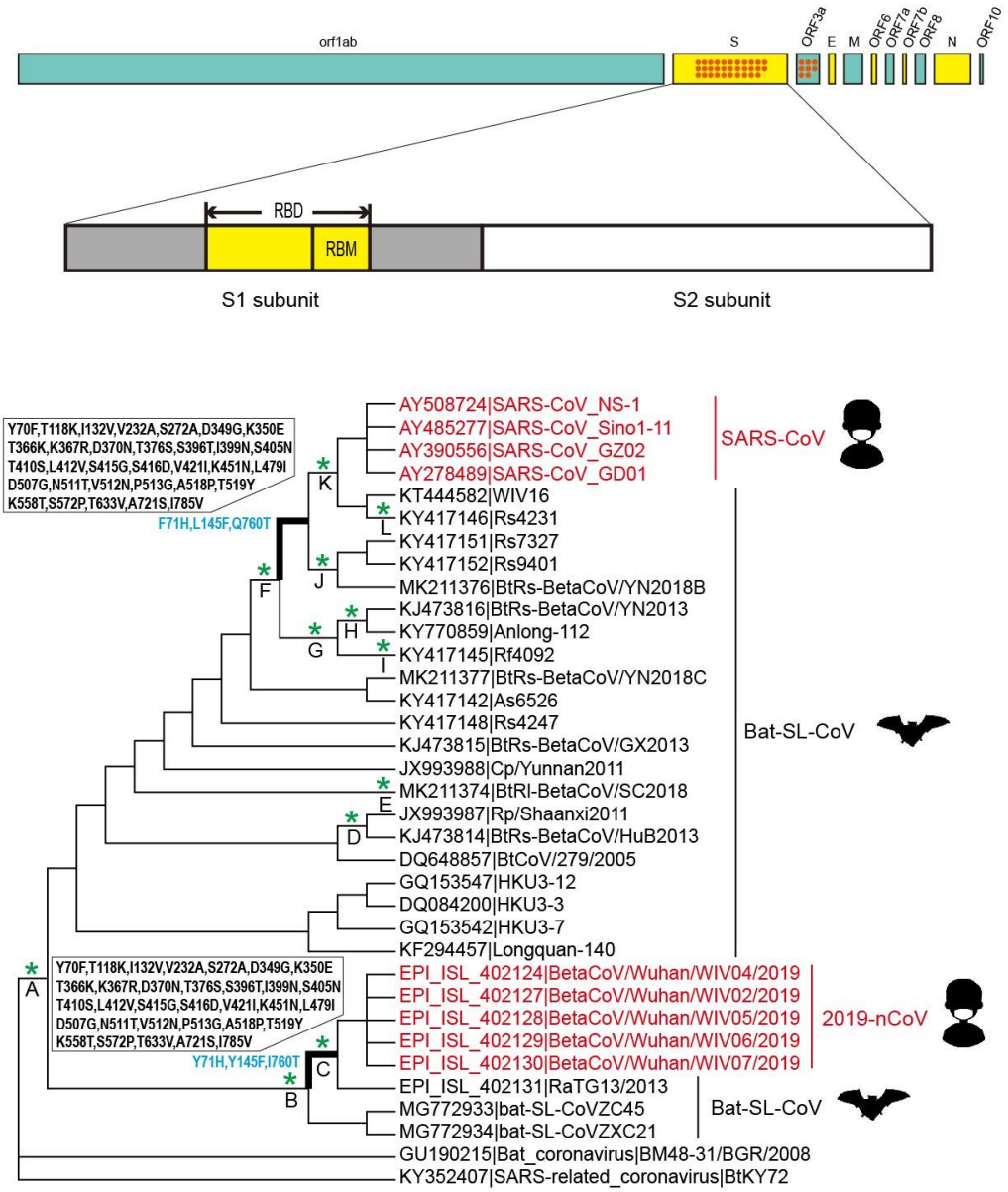

Fig. 1 The phylogeny of subgenus *Sarbecovirus* and 11 genes used in this study. The coronavirus phylogeny follows two published studies [5,6]. 2019-nCoV, 2019 novel coronavirus; SARS-CoV, severe acute respiratory syndrome coronavirus; Bat-SL-CoV, bat-derived severe acute respiratory syndrome (SARS)-like coronaviruses. Genomic organization and the 11 genes annotated in the reference genome of 2019-nCoV (NC_045512) are shown. Red dots represent the numbers of identical or nearly identical amino acid sites found in genes, *S* and *ORF3a*, which are shared between SARS-related CoV and COVID-19-related CoV, but are completely or nearly completely distinct from those of their phylogenetic intermediates. The spike (S) protein structure follows one previous study [12], and its receptor-binding domain (RBD) and receptor-binding motif (RBM) are highlighted. * above branches and their corresponding capital letters (A-L) denote the branches with positive selection signals found in the *S* gene. Two branches (bold) indicate two evolutionary convergent branches with three convergent amino acid substitutions (blue) and 32 shared parallel amino acid substitutions of spike protein.

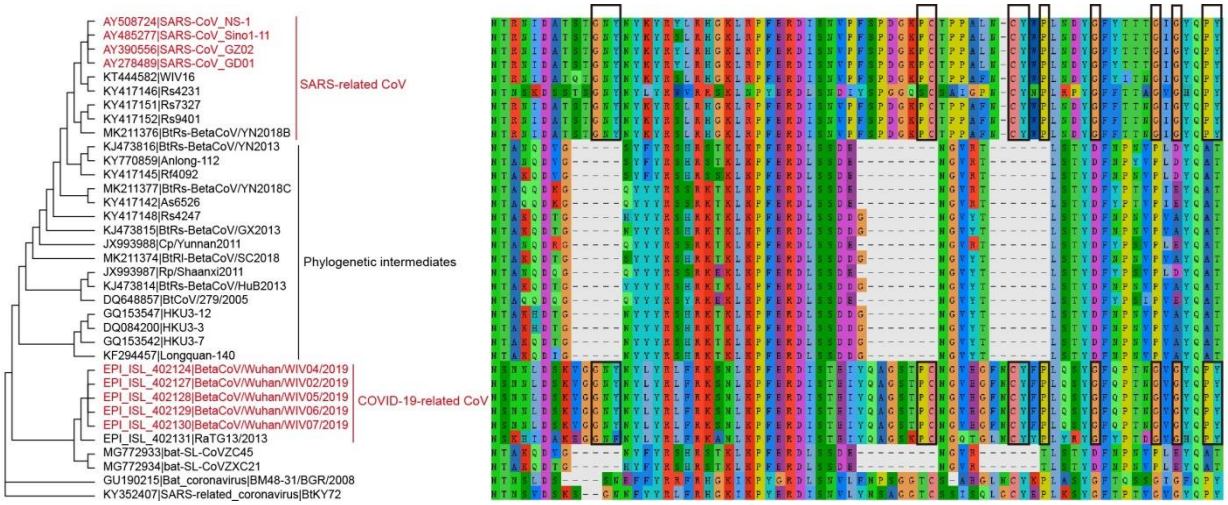

Fig. 2 The identical or nearly identical RBM amino acid sites (rectangle) shared between SARS-related CoV and COVID-19-related CoV. These shared amino acids are distinct from that of their phylogenetic intermediates. The phylogeny is the same as in Fig. 1.

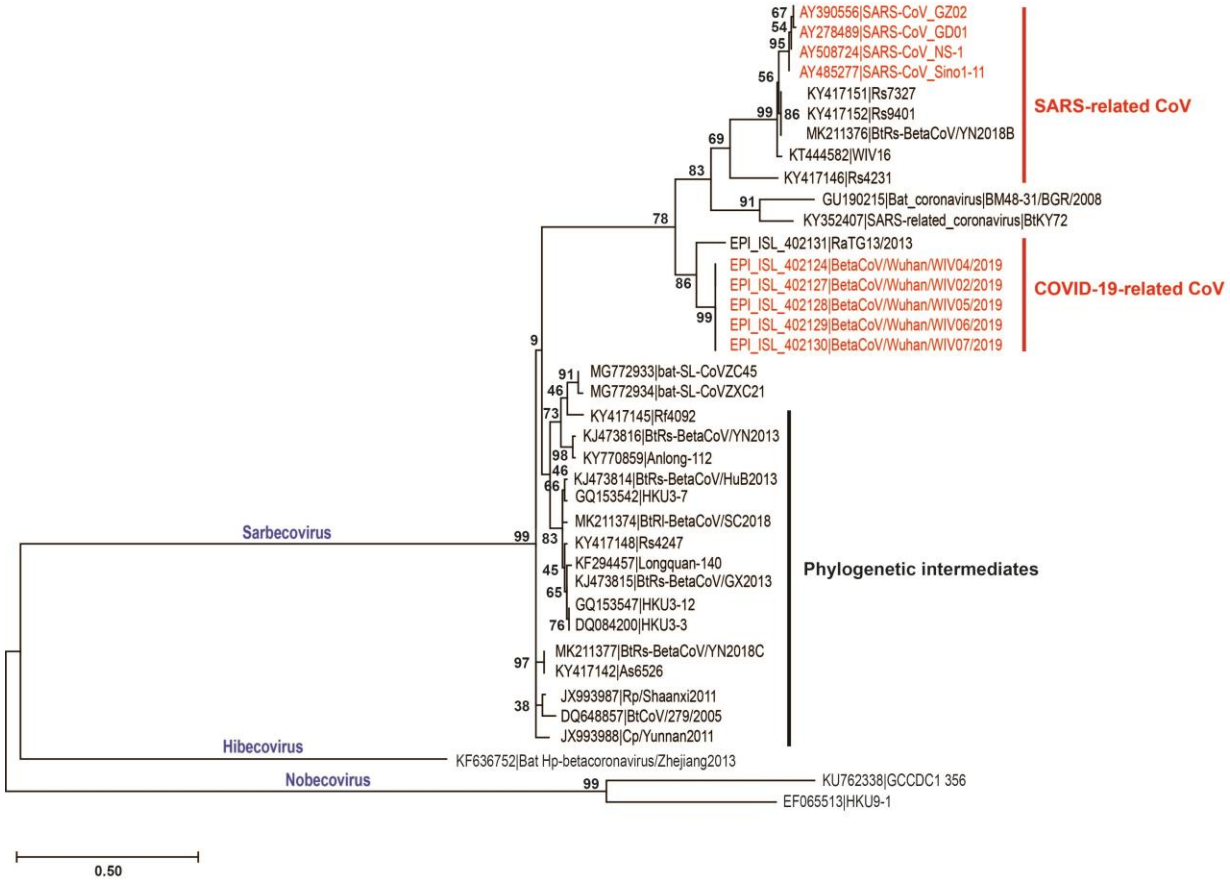

Fig. 3 Maximum likelihood tree of the full-length amino acid sequence of RBD. The WAG+G amino acid substitution model is used. The tree has the highest log likelihood (-3406.88). The node supports are shown in number. *Hibecovirus* and *Nobecovirus* are used as outgroups.

Table 1 Positively selected genes identified based on the branch-site model. Only the ω values of the foreground branches are shown. Positive selection sites located in subunit 1 of gene *S* are shown in grey. Underlining shows positive selected sites located within RBD. Dashes (-) shows alignment gaps in the reference sequence.

Branch/Gene	Parameter estimates	2ΔL	df	*p*-value	Positive selection sites
Branch C					
S	p_0= 0.899 p_1= 0.067 p_{2a}= 0.030 p_{2b}= 0.002	18.76	1	1.475E-05	19Y,67N,90N,187K,210M,289E,294S,351V
	ω_0= 0.029 ω_1= 1.000 **ω_{2a}= 27.064 ω_{2b}= 27.064**				393G,399Y,420K,447T,463F,513T,539Q
					551T,570-,577Q,615L,760A,770A
Orf1ab	p_0= 0.944 p_1= 0.050 p_{2a}= 0.004 p_{2b}= 0.000	56.45	1	5.756E-14	154I,1756G,3909G,3918N,4019H,4222M,4228S
	ω_0= 0.029 ω_1= 1.000 **ω_{2a}= 206.003 ω_{2b}= 206.003**				4287L,4298L,4319V,4446V,4478I,4486L,4502C
					4864S,4960H,5882S,5930N,6128H,6191S,6323I
					6396A,6428S,6436K,6488S
N	p_0= 0.914 p_1= 0.080 p_{2a}= 0.005 p_{2b}= 0.000	5.15	1	0.023	26D,104E,129E,219T,236V,336H,347N,415G
	ω_0= 0.037 ω_1= 1.000 **ω_{2a}= 122.692 ω_{2b}= 122.692**				
Branch K					
S	p_0= 0.924 p_1= 0.069 p_{2a}= 0.004 p_{2b}= 0.000	7.29	1	0.006	11D,19Y,150N,390P
	ω_0= 0.030 ω_1= 1.000 **ω_{2a}= 999.000 ω_{2b}= 999.000**				

2ΔL: twice difference of likelihood values between two nested models; df: degrees of freedom; proportion of sites and their corresponding ω values in four site classes (p_0, p_1, p_{2a} and p_{2b}) are shown.

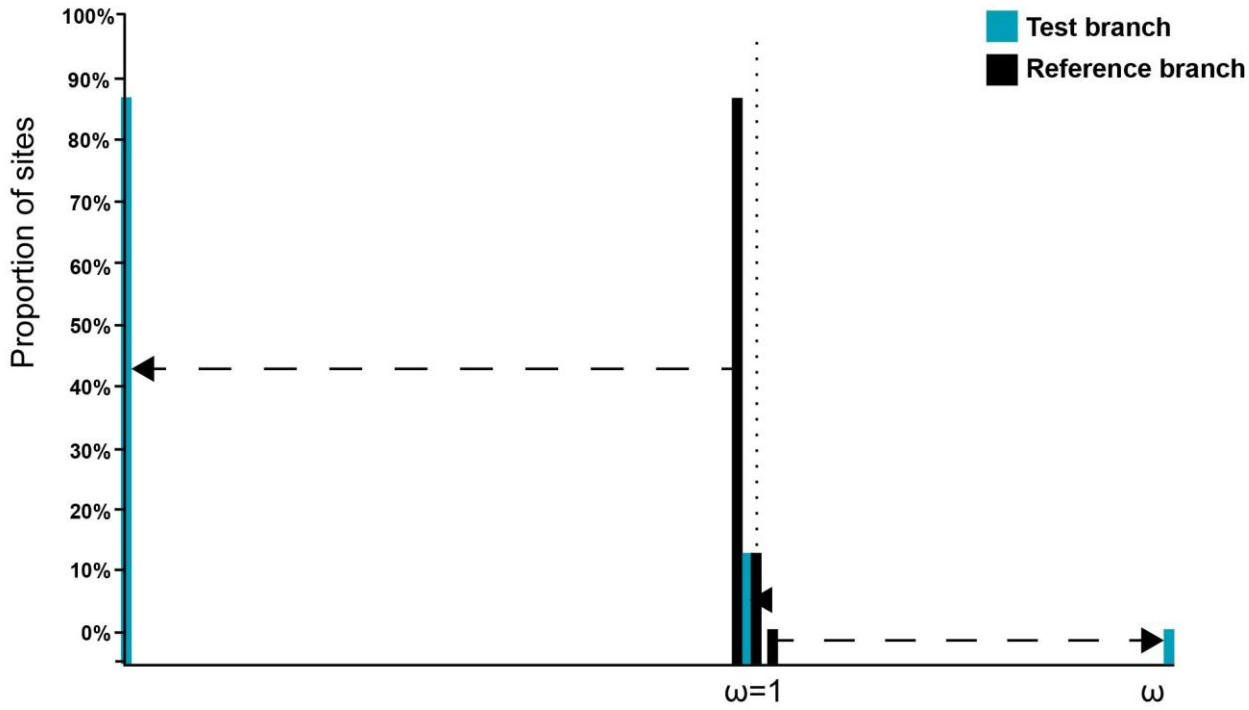

Fig. S1 Selection intensity changes of gene *S* along the 2019- nCoV branch (test branch) compared with the common ancestral branch (reference branch) of 2019 n-CoV and RaTG13. The result shows that the ω categories of the test branch are apparently away from neutrality (ω = 1), indicating an intensified selection along the 2019- nCoV branch.

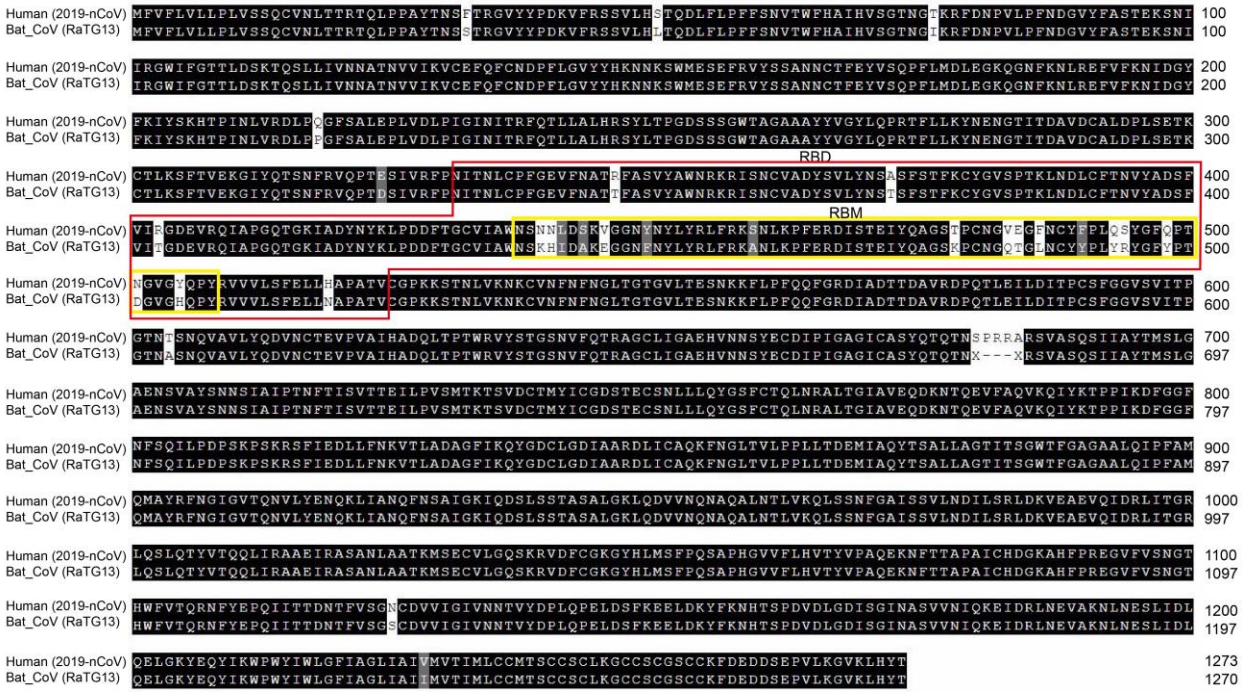

Fig. S2 Amino acid variations of full-length spike protein sequences of 2019-nCoV and RaTG13. RBD, receptor-binding domain; RBM, receptor-binding motif.

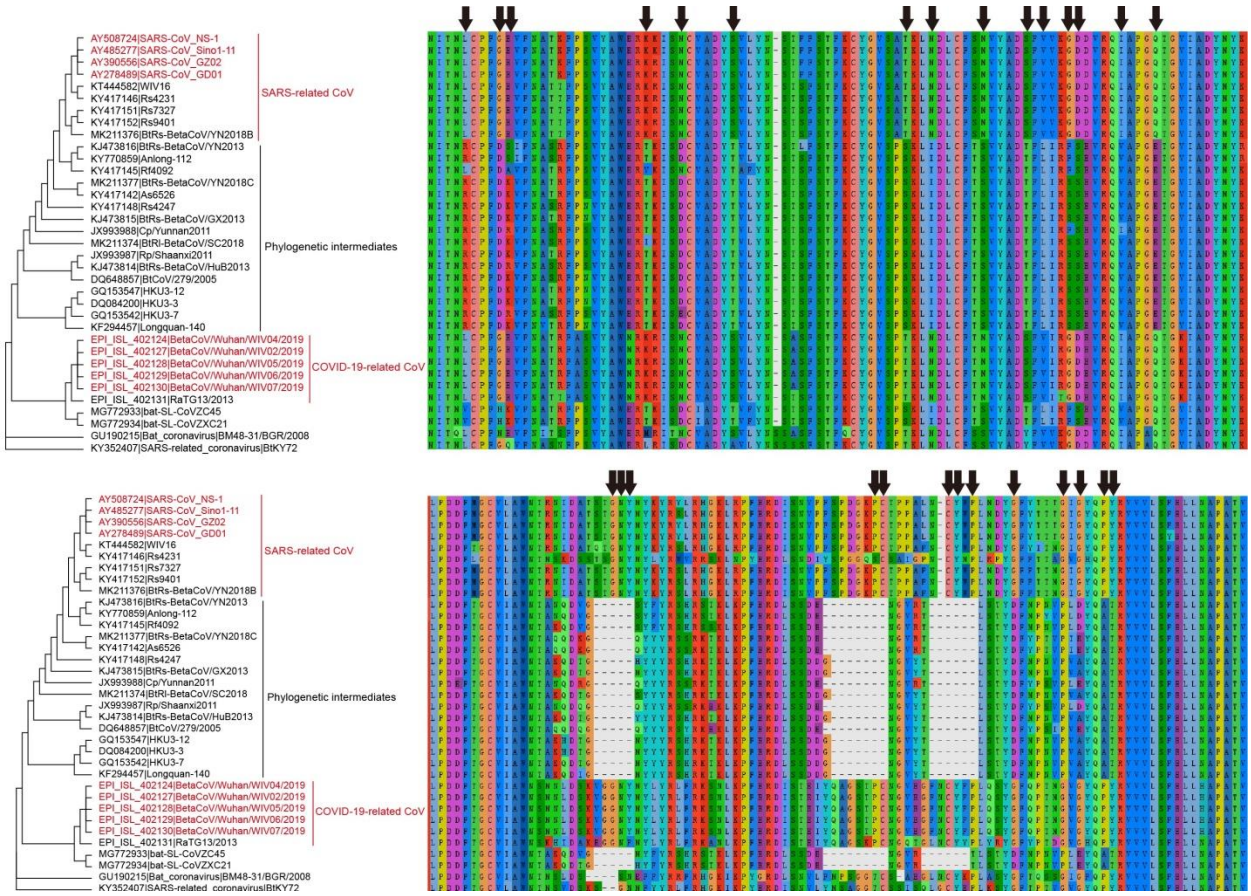

Fig. S3 Twenty-eight identical or nearly identical RBD amino acid sites (arrows) shared between SARS-related CoV and COVID-19-related CoV. These shared amino acids are completely or nearly completely distinct from those of their phylogenetic intermediates. The phylogeny is the same as in Fig. 1.

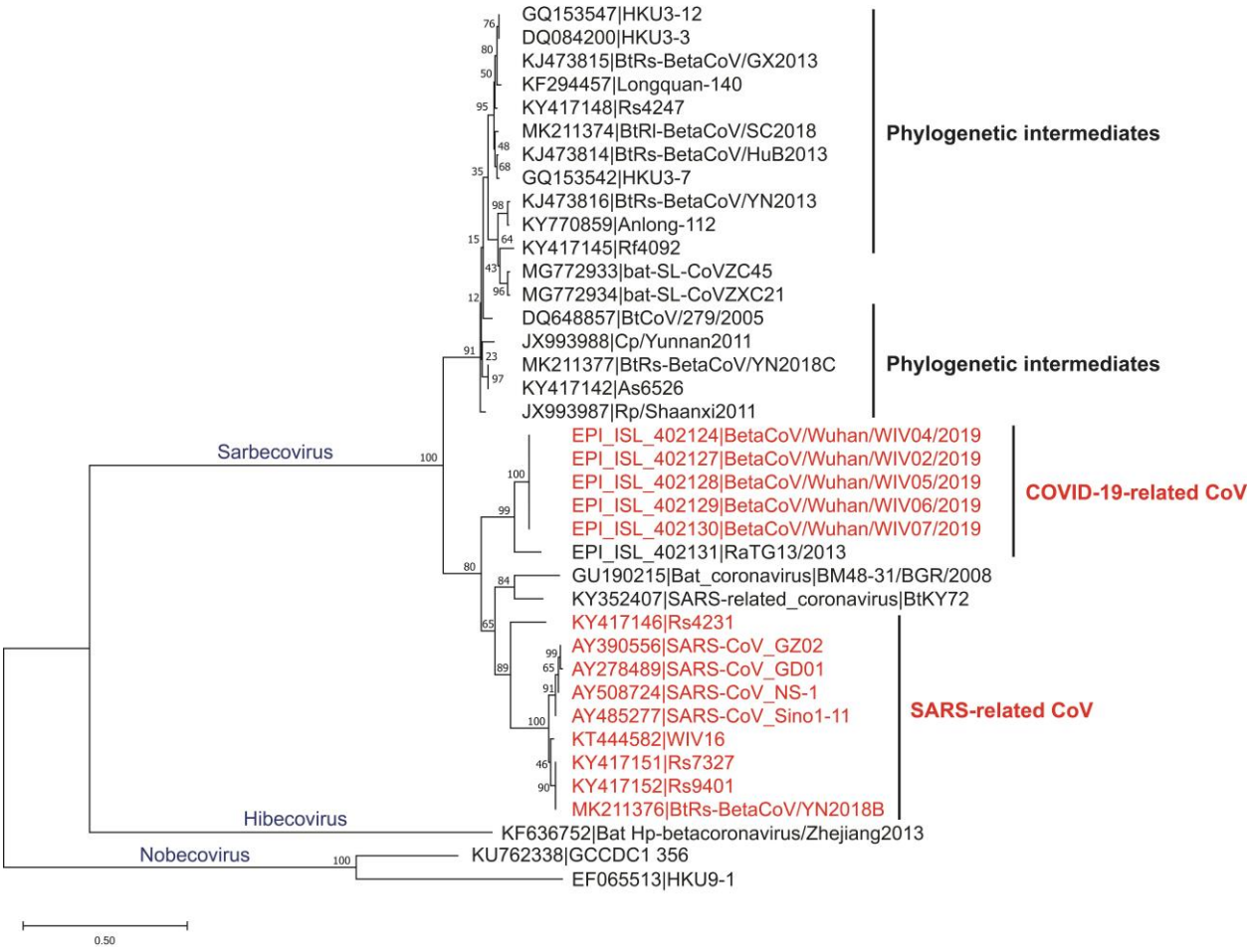

Fig. S4 Neighbor-joining tree based on full-length amino acid sequence of RBD. The JTT+G model is used. The node supports are shown in numbers. *Hibecovirus* and *Nobecovirus* are used as outgroups.

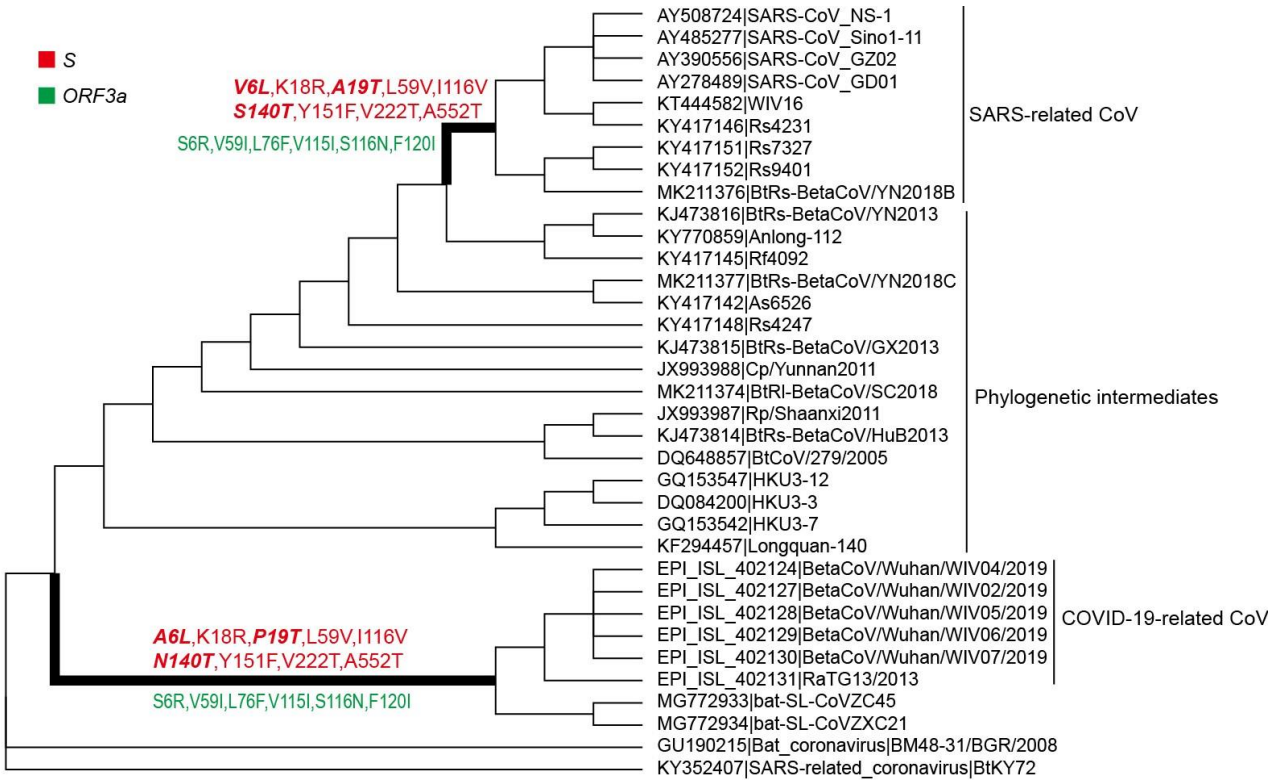

Fig. S5 Convergent (bold italic) and parallel amino acid substitutions of genes *S* and *ORF3a*. The phylogeny is the same as in Fig. 1.

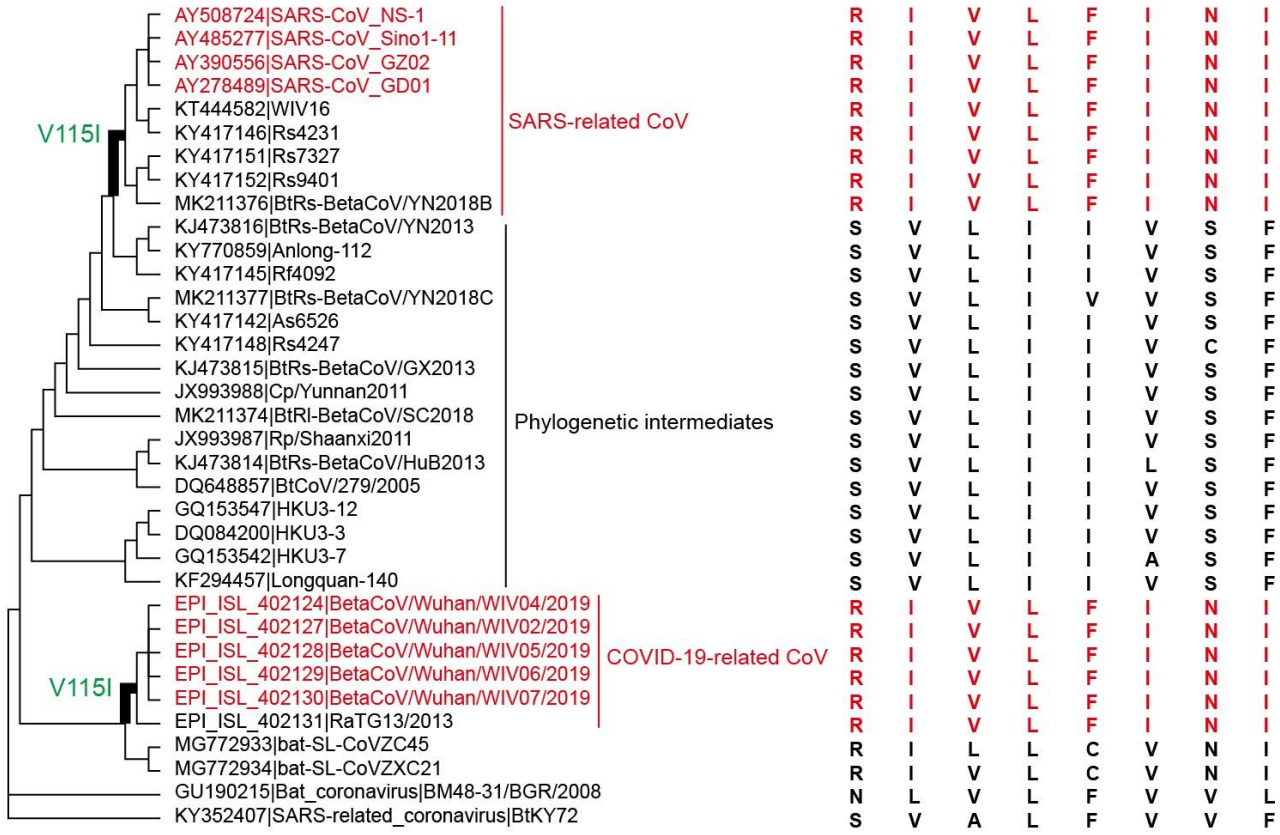

Fig. S6 One parallel amino acid substitution of *ORF3a* protein shared by two ancestral branches (bold) leading to SARS-related CoV and COVID-19-related CoV. This parallel amino acid substitution is only supported as Poisson model was used. Eight identical amino acid sites of *ORF3a* protein shared between SARS-related CoV and COVID-19-related CoV are also shown. These shared amino acids are completely distinct from those of their phylogenetic intermediates. The phylogeny is the same as in Fig. 1.

Table S1 Selection intensity change of 11 genes along the 2019-nCoV branch (test branch) relative to the common ancestral branch (reference branch) of 2019-nCoV and RaTG13.

Gene	Model	log L	# par.	Branch set	ω1	ω2	ω3	K	P-value
E	Alternative	-719.0	67	Test branch	0.00 (100.00%)	3.26 (0.00%)			
				Reference branch	0.01 (100.00%)	1.12 (0.00%)		10.49	0.989
	Null	-719.0	66	Test branch	0.00 (100.00%)	1.13 (0.00%)			
				Reference branch	0.00 (100.00%)	1.13 (0.00%)			
M	Alternative	-3454.2	75	Test branch	0.00 (64.66%)	0.00 (35.34%)	1.07 (0.00%)		
				Reference branch	0.00 (64.66%)	0.00 (35.34%)	1.08 (0.00%)	0.90	0.858
	Null	-3454.2	74	Test branch	0.00 (68.22%)	0.00 (31.78%)	1.13 (0.00%)		
				Reference branch	0.00 (68.22%)	0.00 (31.78%)	1.13 (0.00%)		
N	Alternative	-5824.2	80	Test branch	0.03 (88.84%)	0.03 (10.35%)	119.34 (0.81%)		
				Reference branch	0.07 (88.84%)	0.07 (10.35%)	39.22 (0.81%)	1.30	0.289
	Null	-5824.7	79	Test branch	0.00 (10.17%)	0.07 (88.99%)	65.17 (0.84%)		
				Reference branch	0.00 (10.17%)	0.07 (88.99%)	65.17 (0.84%)		
Orf1ab	Alternative	-119063.9	86	Test branch	0.00 (94.12%)	0.88 (5.36%)	1.74 (0.53%)		
				Reference branch	0.00 (94.12%)	0.40 (5.36%)	60.91 (0.53%)	0.13	0.000***
	Null	-119070.0	85	Test branch	0.01 (96.92%)	0.30 (2.48%)	6.76 (0.60%)		
				Reference branch	0.01 (96.92%)	0.30 (2.48%)	6.76 (0.60%)		
ORF3a	Alternative	-5474.6	80	Test branch	0.07 (100.00%)	0.09 (0.00%)	1.00 (0.00%)		
				Reference branch	0.03 (100.00%)	0.05 (0.00%)	1.00 (0.00%)		
	Null	-5474.9	79	Test branch	0.01 (0.00%)	0.05 (100.00%)	1.01 (0.00%)	0.79	0.434
				Reference branch	0.01 (0.00%)	0.05 (100.00%)	1.01 (0.00%)		
ORF6	Alternative	-974.5	70	Test branch	0.12 (100.00%)	0.14 (0.00%)	1.06 (0.00%)		
				Reference branch	0.10 (100.00%)	0.12 (0.00%)	1.06 (0.00%)		
	Null	-974.7	69	Test branch	0.10 (34.17%)	0.10 (65.83%)	1.13 (0.00%)	0.91	0.561
				Reference branch	0.10 (34.17%)	0.10 (65.83%)	1.13 (0.00%)		
ORF7a	Alternative	-2320.2	76	Test branch	0.00 (73.38%)	0.04 (26.62%)	1.06 (0.00%)		
				Reference branch	0.00 (73.38%)	0.94 (26.62%)	1.00 (0.00%)	50.00	0.092
	Null	-2321.6	75	Test branch	0.00 (92.16%)	1.00 (0.64%)	2.03 (7.20%)		
				Reference branch	0.00 (92.16%)	1.00 (0.64%)	2.03 (7.20%)		
ORF7b	Alternative	-669.0	58	Test branch	0.00 (17.28%)	40.06 (82.72%)			
				Reference branch	0.00 (17.28%)	2.56 (82.72%)		3.92	0.892

	Null	-669.0	57	Test branch	1.00 (0.00%)	30.99 (100.00%)			
				Reference branch	1.00 (0.00%)	30.99 (100.00%)			
	Alternative	-2536.6	77	Test branch	0.61 (80.84%)	0.62 (16.75%)	1.04 (2.40%)		
ORF8				Reference branch	0.00 (80.84%)	0.00 (16.75%)	2.48 (2.40%)	0.05	0.169
	Null	-2537.6	76	Test branch	0.00 (25.76%)	0.00 (71.48%)	24.15 (2.75%)		
				Reference branch	0.00 (25.76%)	0.00 (71.48%)	24.15 (2.75%)		
	Alternative	-227.7	40	Test branch	0.62 (0.06%)	1.10 (99.94%)			
ORF10				Reference branch	0.59 (0.06%)	1.12 (99.94%)		0.91	0.989
	Null	-227.7	39	Test branch	0.54 (0.06%)	1.13 (99.94%)			
				Reference branch	0.54 (0.06%)	1.13 (99.94%)			
	Alternative	-27508.8	84	Test branch	0.00 (86.63%)	0.04 (12.91%)	6.46e+55 (0.46%)		
S				Reference branch	0.00 (86.63%)	0.90 (12.91%)	67.18 (0.46%)	30.54	0.000***
	Null	-27528.3	83	Test branch	0.00 (85.80%)	0.42 (13.32%)	1567.50 (0.89%)		
				Reference branch	0.00 (85.80%)	0.42 (13.32%)	1567.50 (0.89%)		

***P < 0.001

Table S2 Selection intensity change of 11 genes along the SARS-CoV branch (test branch) relative to the common ancestral branch (reference branch) of SARS-CoV and its sister taxa (WIV16 and Rs4231).

Gene	Model	log L	# par.	Branch set	ω1	ω2	ω3	K	P-value
E	Alternative	-719.4	71	Test branch	0.26 (81.92%)	0.33 (9.36%)	4.06 (8.73%)	0.91	0.989
				Reference branch	0.23 (81.92%)	0.30 (9.36%)	4.66 (8.73%)		
	Null	-719.4	70	Test branch	0.21 (75.40%)	0.27 (8.54%)	3.02 (16.06%)		
				Reference branch	0.21 (75.40%)	0.27 (8.54%)	3.02 (16.06%)		
M	Alternative	-3455.2	75	Test branch	0.00 (96.31%)	0.00 (2.62%)	23.90 (1.06%)	1.01	0.607
				Reference branch	0.00 (96.31%)	0.00 (2.62%)	23.47 (1.06%)		
	Null	-3455.3	74	Test branch	0.00 (4.06%)	0.00 (94.89%)	24.22 (1.06%)		
				Reference branch	0.00 (4.06%)	0.00 (94.89%)	24.22 (1.06%)		
N	Alternative	-5823.0	80	Test branch	0.00 (100.00%)	0.63 (0.00%)	1.07 (0.00%)	0.88	0.902
				Reference branch	0.00 (100.00%)	0.59 (0.00%)	1.08 (0.00%)		
	Null	-5823.0	79	Test branch	0.00 (5.09%)	0.00 (94.91%)	1.12 (0.00%)		
				Reference branch	0.00 (5.09%)	0.00 (94.91%)	1.12 (0.00%)		
Orf1ab	Alternative	-119093.0	86	Test branch	0.02 (2.04%)	0.04 (97.96%)	1.00 (0.00%)	0.84	0.435
				Reference branch	0.01 (2.04%)	0.02 (97.96%)	1.00 (0.00%)		
	Null	-119093.3	85	Test branch	0.02 (6.80%)	0.04 (93.20%)	1.10 (0.00%)		
				Reference branch	0.02 (6.80%)	0.04 (93.20%)	1.10 (0.00%)		
ORF3a	Alternative	-5477.9	80	Test branch	1.00 (27.44%)	1.00 (54.72%)	2.40 (17.84%)	0.38	0.117
				Reference branch	1.00 (27.44%)	1.00 (54.72%)	10.18 (17.84%)		
	Null	-5479.1	79	Test branch	1.00 (12.16%)	1.00 (60.29%)	2.10 (27.55%)		
				Reference branch	1.00 (12.16%)	1.00 (60.29%)	2.10 (27.55%)		
ORF6	Alternative	-973.7	66	Test branch	0.00 (98.27%)	138560.78 (1.73%)		1.81	0.000***
				Reference branch	0.00 (98.27%)	698.81 (1.73%)			
	Null	-993.7	65	Test branch	0.00 (98.26%)	26981896.49 (1.74%)			
				Reference branch	0.00 (98.26%)	26981896.49 (1.74%)			
ORF7a	Alternative	-2322.0	76	Test branch	0.15 (14.72%)	0.15 (85.28%)	1.00 (0.00%)	0.96	0.965
				Reference branch	0.14 (14.72%)	0.14 (85.28%)	1.00 (0.00%)		
	Null	-2322.0	75	Test branch	0.14 (0.00%)	0.15 (100.00%)	1.13 (0.00%)		
				Reference branch	0.14 (0.00%)	0.15 (100.00%)	1.13 (0.00%)		
	Alternative	-669.0	62	Test branch	0.62 (100.00%)	0.89 (0.00%)	1.08 (0.00%)		

ORF7b				Reference branch	0.61 (100.00%)	0.89 (0.00%)	1.08 (0.00%)		
	Null	-669.0	61	Test branch	0.62 (100.00%)	0.84 (0.00%)	1.09 (0.00%)	0.99	0.863
				Reference branch	0.62 (100.00%)	0.84 (0.00%)	1.09 (0.00%)		
	Alternative	-2537.8	77	Test branch	0.12 (70.24%)	0.61 (0.00%)	1.95 (29.76%)		
ORF8				Reference branch	0.12 (70.24%)	0.61 (0.00%)	1.94 (29.76%)	1.00	0.993
	Null	-2537.8	76	Test branch	0.12 (70.22%)	0.54 (0.00%)	1.94 (29.78%)		
				Reference branch	0.12 (70.22%)	0.54 (0.00%)	1.94 (29.78%)		
	Alternative	-227.5	40	Test branch	0.58 (0.00%)	53.08 (100.00%)			
ORF10				Reference branch	0.60 (0.00%)	40.22 (100.00%)			
	Null	-227.5	39	Test branch	0.58 (0.00%)	40.24 (100.00%)		1.08	0.945
				Reference branch	0.58 (0.00%)	40.24 (100.00%)			
	Alternative	-27527.7	84	Test branch	0.00 (94.72%)	1.00 (1.29%)	1.00 (3.99%)		
S				Reference branch	0.00 (94.72%)	1.00 (1.29%)	4.61 (3.99%)	0.00	0.053
	Null	-27529.5	83	Test branch	0.01 (96.78%)	1.00 (1.10%)	2.59 (2.12%)		
				Reference branch	0.01 (96.78%)	1.00 (1.10%)	2.59 (2.12%)		

***p < 0.001

Table S3 Branches under the positive selection of the *S* gene. Positive selections are analyzed using branch-site model. For convenience, only the ω values of the foreground branches are shown. Positive selection sites located in subunit 1 of the *S* gene are shown in grey. Underlining shows positive selected sites located in the RBD.

Branch	Parameter estimates	2ΔL	df	*p*-value	Positive selection sites
A	p_0= 0.901 p_1= 0.069 p_{2a}= 0.027 p_{2b}= 0.002	7.13	1	0.007	9F,292R,335Y,376H,393G,399Y,514M
	ω_0= 0.030 ω_1=1.000 ω_{2a}= 23.705 ω_{2b}= 23.705				529F,591Q,607S,673T,675S,774Y,820S,845V
B	p_0= 0.902 p_1= 0.071 p_{2a}= 0.024 p_{2b}= 0.001	32.59	1	1.139E-08	9F,11D,27L,63S,72A,110M,133F,174I
	ω_0= 0.029 ω_1=1.000 ω_{2a}= 49.941 ω_{2b}= 49.941				190V,192M,227S,228Q,234L,238T,240S
					260E,458R,701S,733N, 810Q,999S,1135M
D	p_0= 0.917 p_1= 0.066 p_{2a}= 0.014 p_{2b}= 0.001	12.21	1	0.000	14R,25S,27L,39I,52R,68T,69T,85V
	ω_0= 0.030 ω_1=1.000 ω_{2a}= 657.736 ω_{2b}= 657.736				87N,125H,138H,176K, 216E
E	p_0= 0.922 p_1= 0.067 p_{2a}= 0.009 p_{2b}= 0.000	8.72	1	0.003	25S,27L,37P,52R,93F,149*
	ω_0= 0.030 ω_1= 1.000 ω_{2a}= 123.522 ω_{2b}= 123.522				176K,,216E,438T,473F
F	p_0= 0.919 p_1= 0.067 p_{2a}= 0.011 p_{2b}= 0.000	29.62	1	5.267E-08	9F,27L,40P,56I,65L,110M,125H
	ω_0= 0.030 ω_1= 1.000 ω_{2a}= 108.595 ω_{2b}= 108.595				139N,143V,161S,194L,224V,227S,376H
G	p_0= 0.925 p_1= 0.069 p_{2a}= 0.004 p_{2b}= 0.000	8.03	1	0.004	136V,144D,174I,181L,222D, 229D,376H
	ω_0= 0.030 ω_1= 1.000 ω_{2a}= 226.266 ω_{2b}= 226.266				733N,756D
H	p_0= 0.915 p_1= 0.069 p_{2a}= 0.013 p_{2b}= 0.001	25.31	1	4.882E-07	55V,67N,69T,75F,128R,144D,161S
	ω_0= 0.030 ω_1= 1.000 ω_{2a}= 193.715 ω_{2b}= 193.715				171L,177L,216E,276V,358K,408L
I	p_0= 0.917 p_1= 0.069 p_{2a}= 0.011 p_{2b}= 0.000	13.07	1	0.000	2S,48A,63S,73V,160F,209*,240S
	ω_0= 0.030 ω_1= 1.000 ω_{2a}= 998.996 ω_{2b}= 998.996				259T,270L
J	p_0= 0.922 p_1= 0.070 p_{2a}= 0.005 p_{2b}= 0.000	9.72	1	0.001	75F,84I,112A, 128R,190V,191I
	ω_0= 0.030 ω_1= 1.000 ω_{2a}= 999.000 ω_{2b}= 999.000				
L	p_0= 0.926 p_1= 0.071 p_{2a}= 0.001 p_{2b}= 0.000	4.35	1	0.037	364T,380K,382Y,732I,955Y
	ω_0= 0.030 ω_1= 1.000 ω_{2a}= 31.578 ω_{2b}= 31.578				

2ΔL: twice difference of likelihood values between two nested models; df: degrees of freedom; proportion of sites and their corresponding ω values in four site classes (p_0, p_1, p_{2a} and p_{2b}) are shown. * represents alignment gap in reference sequence.