The discovery of a recombinant SARS2-like CoV strain provides insights into SARS and COVID-2019 pandemics

Xin Li [1$], Xiufeng Jin [1$], Shunmei Chen[2], Liangge Wang[3], Tung On Yau[4]

Jianyi Yang[5], Zhangyong Hong[1], Jishou Ruan[5], Guangyou Duan[6]*,Shan Gao[1]*

[1] College of Life Sciences, Nankai University, Tianjin, Tianjin 300071, P.R.China;
[2] Yunnan Key Laboratory of Stem Cell and Regenerative Medicine, Biomedical Engineering Research Center, Kunming Medical University, Kunming, Yunnan 650500, P.R.China;
[3] Taikang Xianlin Drum Tower Hospital, Nanjing University School of Medicine, Nanjing, Jiangsu 210046, P.R.China;
[4] John Van Geest Cancer Research Centre, School of Science and Technology, Nottingham Trent University, Nottingham, NG11 8NS, United Kingdom;
[5] School of Mathematical Sciences, Nankai University, Tianjin, Tianjin 300071, P.R.China;
[6] School of Life Sciences, Qilu Normal University, Jinan, Shandong 250200, P.R.China

[$] These authors contributed equally to this paper.

* The corresponding authors.

SG：gao_shan@mail.nankai.edu.cn

GD：guangyou.duan@qlnu.edu.cn

Abstract

In December 2019, the world awoke to a new zoonotic strain of coronavirus named severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2). In the present study, we classified betacoronavirus subgroup B into the SARS-CoV-2, SARS-CoV and SARS-like CoV clusters, and the ORF8 genes of these three clusters into types 1, 2 and 3, respectively. One important result of our study is that the recently reported strain RmYN02 was identified as a recombinant SARS2-like CoV strain that belongs to the SARS-CoV-2 cluster, but has an ORF8 from a SARS-like CoV. This result provides substantial proof for long-existing hypotheses regarding the recombination and biological functions of ORF8. Based on the analysis of recombination events in the Spike gene, we propose that the Spike protein of SARS-CoV-2 may have more than one specific receptor for its function as gp120 of HIV has CD4 and CCR5. We concluded that the furin protease cleavage site acquired by SARS-CoV-2 may increase the efficiency of viral entry into cells, while the type 2 ORF8 acquired by SARS-CoV may increase its replication efficiency. These two most critical events provide the most likely explanation for SARS and COVID-2019 pandemics.

Keywords: 5' UTR; furin enzyme; SARS-CoV; ORF8; recombinant

Introduction

A new zoonotic strain of coronavirus named severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) emerged in December 2019. We previously reported several important findings on SARS-CoV-2 for the first time, including the following discoveries in particular: (1) the alternative translation of coding sequences (CDSs) that explain the rapid mutation, multiple hosts and strong host adaptability of betacoronavirus at the molecular level [1]; (2) a furin protease cleavage site (FCS) in the spike (S) protein of SARS-CoV-2 that may increase the efficiency of viral entry into cells [2]; and (3) the use of 5' untranslated-region (UTR) barcoding for the detection, identification, classification and phylogenetic analysis of—though not limited to—coronaviruses [3]. By data mining betacoronaviruses from public databases, we found that more than 50 nucleotides (nts) at the 3' ends of the 5' UTRs in betacoronavirus genomes are highly conserved with very few single nucleotide polymorphisms (SNPs) within each subgroup of betacoronaviruses. We defined 13~15-bp sequences of 5' UTRs including the start codons (ATGs) of the first open reading frames (ORFs) as barcodes to represent betacoronaviruses. Using 5' UTR barcodes, 1265 betacoronaviruses were clustered into four classes, matching the C, B, A and D subgroups of betacoronavirus [3], respectively. In particular, SARS-CoV-2 and SARS-CoV have the same 5' UTR barcode—GAAAGGTAAG(ATG)—laying the foundation to rename 2019-nCoV as SARS-CoV-2.

In the subsequent studies, InDels (insertions/deletions) at six sites were used to classify betacoronavirus subgroup B into two classes (**Results and Discussion**). Using the InDels at six sites, we identified two recently detected betacoronavirus strains RmYN01 and RmYN02 from a bat [4] as belonging to the second and first classes, respectively. Moreover, we also discovered that RmYN02 was a recombinant SARS2-like CoV strain. This recombination event occurred in a gene named open reading frame 8 (ORF8), existing only in the genomes of betacoronaviruses from subgroup B. The ORF8 gene of SARS-CoV is considered to have played a significant role in adaptation to human hosts following interspecies transmission [5] via the modification of viral replication [6]. A 29-nt deletion in SARS-CoV (GenBank: AY274119) was reported and considered to be associated with attenuation during the early stage of human-to-human transmission [6]. A 382-nt deletion during the early evolution of SARS-CoV-2 (GISAID: EPI_ISL_414378) was also reported [7]. In conjunction with our new discoveries, we conducted preliminary research and aimed at determining: (1) the critical mutations and recombination; (2) the influence of recombinant ORF8 on the phylogenetic analysis of betacoronaviruses and (3) the role of ORF8 in the modification of viral replication.

Results and Discussion

InDels were identified at six mutation sites, named M1 to M6 (**Table 1**), in ORF3a, membrane (M), ORF7a, ORF7b, ORF8 and nucleocapsid (N), respectively (**Figure 1**). Based on these InDels at six sites, betacoronavirus subgroup B (**Materials and Methods**) was classified into two classes: (1) the first class

includes SARS-CoV-2 (from humans) and SARS2-like CoV (from animals), and (2) the second class includes SARS-CoV (from humans) and SARS-like CoV (from animals). As a variable region, M1 has a length of 8 nt in betacoronaviruses of the second class and 11 nt in betacoronaviruses of the first class. M2, M3, M4 and M5 in betacoronaviruses of the first class have 3-nt deletions, which are complete codons, whereas M6 in betacoronaviruses of the first class has 6-nt deletion.

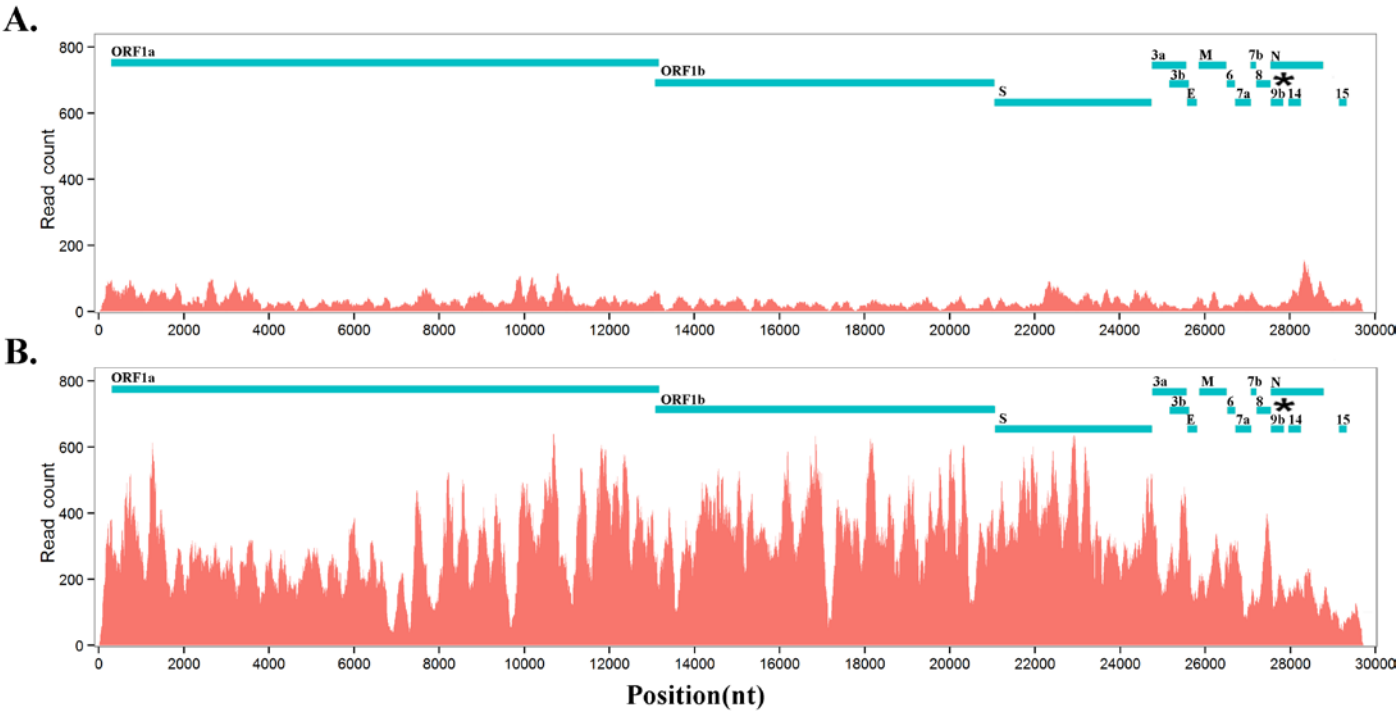


Figure 1 RNA abundances of RmYN01 and RmYN02 in a bat

RNA-seq data from a bat was aligned to two genomes of RmYN01 and RmYN02 (GISAID: EPI_ISL_412976 and EPI_ISL_412977). A RmYN01 was identified as belonging to the SARS-like CoV cluster and has a type 3 ORF8. B. RmYN02 was identified as belonging to the SARS-CoV-2 cluster and has a type 2 ORF8.

Recently, two betacoronavirus strains RmYN01 and RmYN02 (GISAID: EPI_ISL_412976 and EPI_ISL_412977) were detected from a bat (*Rhinolophus malayanus*) [4]. Since betacoronaviruses from subgroup B share many highly similar regions in their genome sequences (**Figure 1**), it is very difficult to assemble them correctly using high-throughput sequencing (HTS) data from one sample. Therefore, EPI_ISL_412976 was only assembled into a partial sequence in a previous study [4]. However, the exact identification of viruses requires the complete genomes or even the full-length genomes. Using paired-end sequencing data, we reassembled these two virus genomes and obtained two full-length sequences to update EPI_ISL_412976 and EPI_ISL_412977 (**Supplementary 1**). Using 5' UTR barcodes, the betacoronaviruses RmYN01 and RmYN02 were identified as belonging to subgroup B. Using the InDels at six sites, RmYN01 and RmYN02 were further identified as belonging to the second and first classes, respectively. In addition, RmYN02 was also identified as a recombinant SARS2-like CoV strain.

A recombination event occurred in a gene named open reading frame 8 (ORF8), existing only in the genomes of betacoronaviruses from subgroup B. Based our analysis of 1265 betacoronaviruses, most of identified recombination events occurred in genes S and ORF8. Although many recombination events in ORF8 of betacoronaviruses have been reported in sequence analysis results, it is difficult to determine whether they were recombination events or mutation accumulation as most of them only occurred over very small genomic regions, excepting a few events (e.g., the 382-nt deletion in ORF8 [7]). We reported—for the first time—a recombination event at the whole-gene level as substantial proof. Besides the recombination event in ORF8, other recombination events occurred on five sites, named RC1 to RC5 (**Table 1**) in the S1 region of the gene S. To initiate the coronavirus infection, the S protein encoded by the gene S need to be cleaved into the S1 and S2 subunits for receptor binding and membrane fusion.

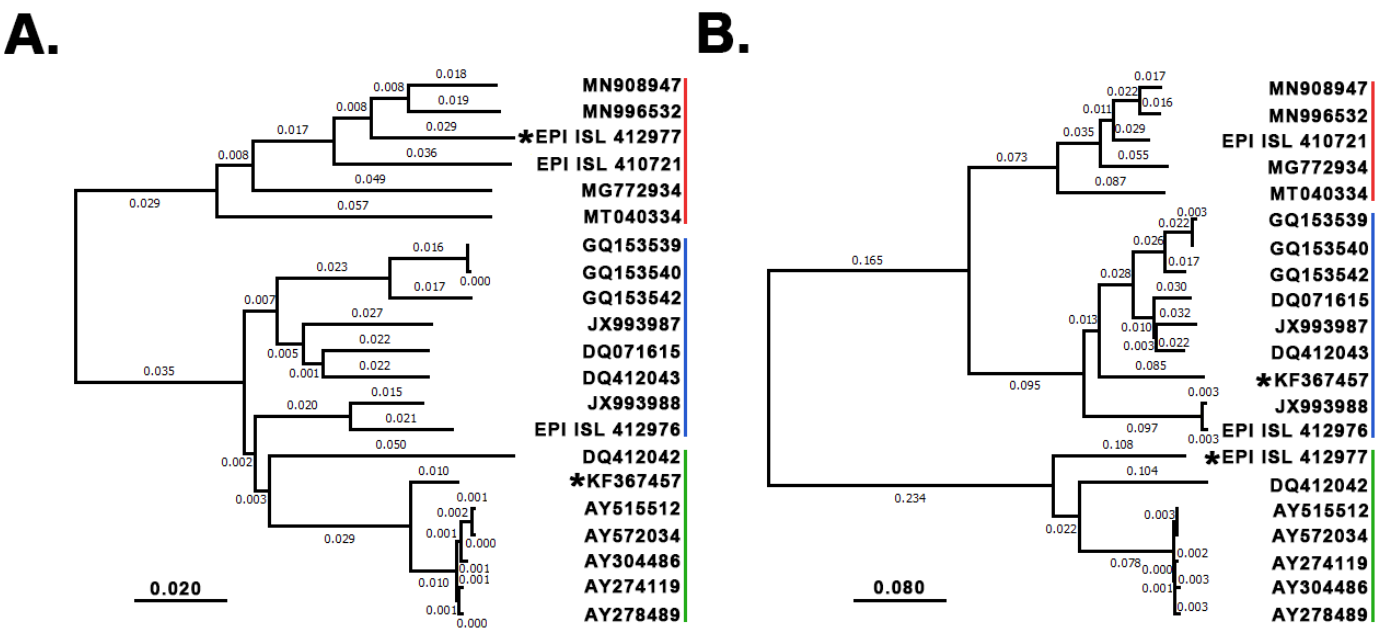


Figure 2 Phylogenetic analysis of the SARS-CoV-2, SARS-CoV and SARS-like CoV clusters

The accession numbers of the GenBank or GISAID databases were used to the viral genomes. The annotations of these viru genomes can be seen in Materials and Methods. The viral genomes of SARS-CoV (GenBank: AY278489) and SARS-CoV-2 (GenBank: MN908947) were used as reference genomes in the present study. Numbers above the branches are phylogenetic distances calculated using the NJ method. The SARS-CoV-2, SARS-like CoV and SARS-CoV clusters were marked by red, blue and green lines. A. Phylogenetic tree 1 was built using large segments spanning S2, ORF3a, ORF3b, envelope (E), M, ORF6, ORF7a, ORF7b, N, ORF9b, ORF14 and ORF15 (Table 1). B. Phylogenetic tree 2 was built using ORF8.

Using a large segment spanning S2, ORF3a, ORF3b, envelope (E), M, ORF6, ORF7a, ORF7b, N, ORF9b, ORF14 and ORF15 (**Table 1**), phylogenetic tree 1 (**Figure 2A**) showed that 21 betacoronaviruses from subgroup B (**Materials and Methods**) were classified into two major clades, corresponding to two classes classified using the InDels at six sites: (1) the first major clade, named the SARS-CoV-2 cluster, includes SARS-CoV-2 and SARS2-like CoV (from bats and pangolins), and (2) the second major clade includes two clusters—the SARS-CoV cluster including SARS-CoV and a few SARS-like CoVs (from bats

or civets) and the SARS-like CoV cluster including all other SARS-like CoVs (from bats). Therefore, these 21 betacoronaviruses were classified into the SARS-CoV-2, SARS-CoV and SARS-like CoV clusters named clusters 1, 2 and 3, respectively.

Identified as belonging to the SARS-CoV-2 cluster, RmYN02 was thought to have a 3-nt deletion at the V5 site; however, it did not (**Table 1**). This led us to identify three types of ORF8 genes in the betacoronaviruses from subgroup B. Using only ORF8, phylogenetic tree 2 (**Figure 2B**) also shows that the 21 betacoronaviruses were classified into the SARS-CoV-2, SARS-CoV, and SARS-like CoV clusters. However, this tree did not reflect the evolutionary relationship of the three clusters due to the recombination events of ORF8. Types 1, 2, and 3 ORF8 genes belong to the genomes of viruses from clusters 1, 2, and 3, respectively. Type 1 ORF8 genes possess low nucleotide identities (below 70%) to type 3 ORF8 genes, while type 2 ORF8 genes are so highly divergent from types 1 and 3 ORF8 genes, they cannot be aligned to calculate nucleotide identities between type 2 ORF8 genes and types 1 or 3 ORF8 genes. RmYN01 belongs to cluster 3 and has a type 3 ORF8, while RmYN02 belongs to cluster 1 but has a type 2 ORF8. Thus, RmYN02 was identified as a recombinant SARS2-like CoV strain.

Comparing phylogenetic tree 1 using large segments with tree 2 using only ORF8 genes, almost all viruses were consistently classified into the same clusters in both trees, except RmYN02 (GISAID: EPI_ISL_412977) and the SARS-like CoV strain WIV1 (GenBank: KF367457). WIV1 was classified into cluster 2 in tree 1 but cluster 3 in tree 2, as WIV1 has type 3 rather than type 2 ORF8 (**Figure 2B**). WIV1, isolated from Chinese horseshoe bats (*Rhinolophus sinicus*), was considered the most closely related to SARS-CoVs in humans and civets [8]. A previous study predicted the immediate ancestor of SARS-CoV based on the following hypothesis: the ancestor of civet SARS-CoV is a recombinant virus with ORF8 originating from greater horseshoe bats (*Rhinolophus ferrumequinum*) and other genomic regions originating from different horseshoe bats [5]. Although whether these recombination events occurred in bats or civets remains unknown [5], analysis of all recombination events in S1 of all betacoronaviruses may provide insights (See below).

As phylogenetic tree 1 was built without recombinant regions (i.e., ORF1a, S1 and ORF8), it revealed that SARS-CoV-2 is most closely related to the well-known strain RaTG13 (GenBank: MN996532) isolated from intermediate horseshoe bats (*Rhinolophus affinis*). Phylogenetic tree 2 shows that ORF8 of RaTG13 has the highest identity to that of SARS-CoV-2 (**Figure 2B**). RmYN02 was classified into cluster 1 in tree 1 but cluster 2 in tree 2. This suggested the recombination events happened across the SARS-CoV-2 and SARS-CoV clusters. In addition, all betacoronaviruses from pangolins (*Manis javanica*) used in the present study were identified as belonging to the SARS-CoV-2 cluster. However, pangolins are unlikely to be the intermediate host(s) of SARS-CoV-2 for two reasons: (1) betacoronaviruses from pangolins do not contain the FCS [2], and (2) the strains (e.g., GISAID: EPI_ISL_410721) are farther from SARS-CoV-2 than RaTG13 in the phylogenetic tree 1 (**Figure 2A**).

Guided by joint analysis of both molecular function and phylogeny, we conducted further research on the biological functions of ORF8. RmYN01 and RmYN02 were detected in a bat, providing a special opportunity to compare their copy numbers. As RmYN01 and RmYN02 have type 3 and type 2 ORF8 genes, respectively, the difference between copy numbers of RmYN01 and those of RmYN02 can be estimated by their relative RNA abundances to test a previous hypothesis that type 2 ORF8 genes enhance viral replication. Aligning RNA-seq data to the genomes of RmYN01 and RmYN02, our calculation showed that the RmYN01 genome was covered 99.85% of its length with an average depth of 32.89 (**Figure 1A**), while the RmYN02 genome was covered 99.89% with an average depth of 298.99 (**Figure 1B**). The RNA abundance of RmYN02 is 9 times that of RmYN01. Although this does not rule out other factors (e.g., differential gene expression) that contribute to the differences in RNA abundances of the two viruses, these results combined with other evidence [5] suggest that type 2 ORF8 genes enhance viral replication.

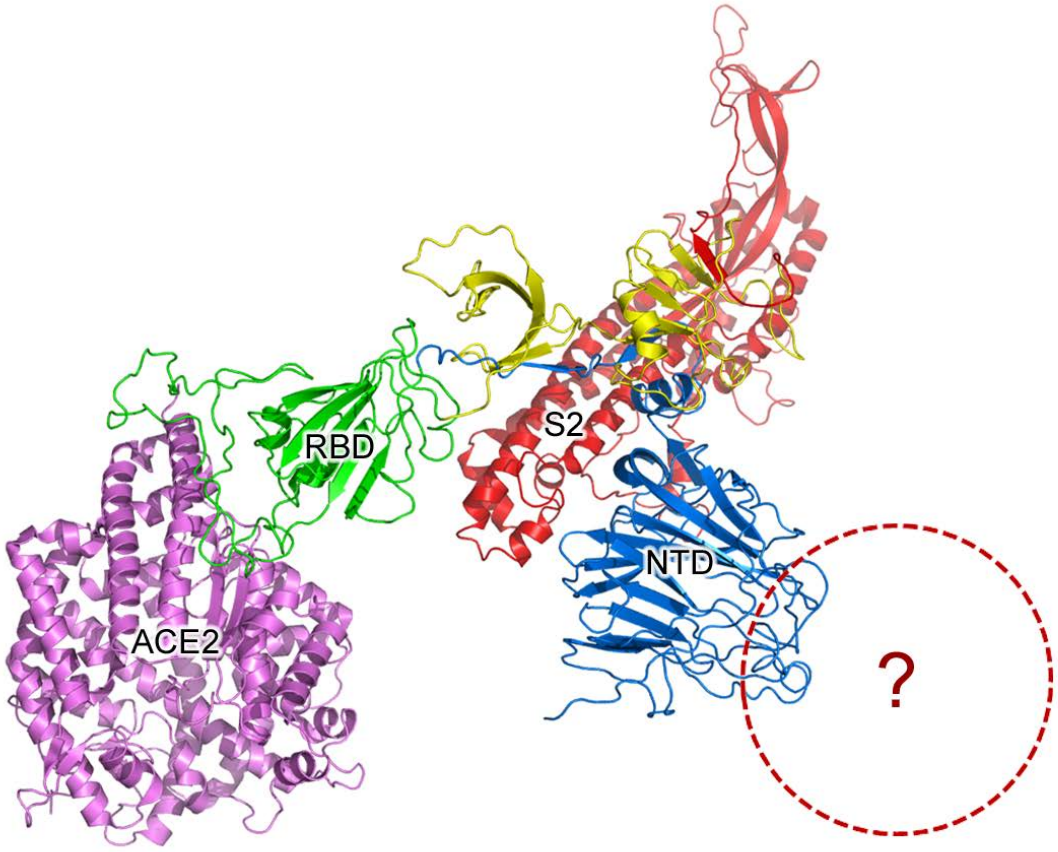


Figure 3 S protein of SARS-CoV-2 may have two specific receptors

The S protein is cleaved into two subunit S1 and S2 (in red color) for receptor binding and membrane fusion. S1 has two domain, RBD (in green color) and NTD (in blue color). It is well accepted that S1 binds to its specific receptor angiotensin-converting enzyme 2 (ACE2) by the interaction between RBD and ACE2 (in purple color). In the present study, we propose that the Spike protein of SARS-CoV-2 may have more than one specific receptor for its function as gp120 of HIV has CD4 and CCR5. The structure of S was predicted using trRosetta [18].

By analysis of all recombination events in S1 of all betacoronaviruses, we obtained the following results: (1) there were a few genotypes of recombinant segments at each recombination site; and (2) and betacoronaviruses within the SARS-CoV-2, SARS-CoV and SARS-like CoV clusters had the same genotypes. These results suggested that the occurrences of all recombination events in bats are prerequisite for speciation of SARS-CoV and SARS-CoV-2. Further analysis showed that two recombination regions reside in the receptor binding domain (RBD) of S1, while three other recombination regions reside in the N-terminal domain (NTD). Since positive selection of S was particularly strong [5] and both RBD and NTD have similar recombination events, we proposed that NTD is likely to have a specific receptor as RBD has angiotensin-converting enzyme 2 (ACE2). Thus, the S protein of SARS-CoV-2 may have more than one specific receptor (**Figure 3**) as gp120 of HIV has the cluster of differentiation 4 receptor (CD4) and the C-C chemokine receptor 5 (CCR5).

Conclusions

Betacoronaviruses have very high mutation and recombination rates, multiple hosts, and strong host adaptability. The recombinant regions (i.e., ORF1a, S1 and ORF8) must be removed for phylogenetic analysis. Based on phylogenetic analysis combined with the analysis of mutations and recombination, betacoronavirus subgroup B was classified into the SARS-CoV-2, SARS-CoV, and SARS-like CoV clusters. As all mutations and recombination are unlikely to undergo reversible changes together, we concluded the following: (1) the SARS-CoV-2 cluster separated from the other two clusters early and (2) later, the SARS-CoV cluster separated from the SARS-like CoV cluster. The FCS acquired by SARS-CoV-2 may increase the efficiency of viral entry into cells, while the type 2 ORF8 acquired by SARS-CoV may increase its replication efficiency. These two most critical events improved the adaptive ability of SARS-CoV and SARS-CoV-2 to new hosts and allowed for sustained transmission that can lead to significant outbreaks. The occurrences of all recombination events in bats are prerequisite for speciation of SARS-CoV and SARS-CoV-2. However, they are not the most critical reasons for the pandemics of SARS and COVID-2019.

Another gene, ORF3b, may also play a role in the host adaptability of betacoronaviruses. In our previous study, we discovered a special genomic region (AY274119: 25689-26153) encoding a complete ORF (named ORF3b) in SARS-CoV [1]. However, this region encodes several ORFs in other betacoronaviruses. Given that this region is not always an ORF, we used Nankai CDS (a 465- or 468-bp genomic region) to name ORF3b and its homologous sequences in all coronaviruses. We found that the alternative translation of Nankai CDS could produce more than 17 putative proteins that may be responsible for host adaptation. Nankai CDS encodes ORF3b in SARS-CoV, while it encodes shorter genes in many other SARS-like CoVs. ORF3b in SARS-CoV was predicted to be translated into a 155-aa (amino acid residue) protein—the largest one among all putative proteins predicted from Nankai CDS in SARS-CoV and SARS-like CoVs.

A previous study showed that the integrity of ORF8 facilitates the replication of SARS-CoV in cell lines [6]. However, another study suggested that the function of ORF8 may be more relevant for replication in cells from the actual animal reservoir than for replication in human or primate cell cultures [9]. Using data from an individual bat, we estimated that type 2 ORF8 genes could increase replication efficiency to 9 times that of type 3 ORF8 genes. Type 1 ORF8 genes are likely to exhibit similar replication efficiency to type 3 ORF8 genes, given the high identities between their nucleotide sequences. This also explains why SARS-CoV-2 (with type 1 ORF8 genes) caused mild or asymptomatic disease in most cases compared to SARS-CoV (with type 2 ORF8 genes). The significant differences in replication efficiency caused by the three types of ORF8 provide insight into SARS-CoV pandemics.

Materials and Methods

The software Fastq_clean [10] was used for sRNA data cleaning and quality control. The genomes of RmYN01 and RmYN02 (GISAID: EPI_ISL_412976 and EPI_ISL_412977) were reassembled by aligning RNA-seq data on two closest reference genomes JX993988 and MN908947. SVDetect v0.8b and SVFilter [11] were used to removed abnormal aligned reads. Several haploid contigs (**Supplementary 1**) highly similar to the complete RmYN01 genome were also assembled. This suggested that there exists more than one betacoronavirus strain belonging to the SARS-like CoV cluster in the same sample, from which RmYN01 and RmYN02 were detected.

1,265 complete genomes of betacoronavirus were downloaded from the NCBI Virus database (https://www.ncbi.nlm.nih.gov/labs/virus). Among these genomes, 292 belongs to the subgroup B of betacoronavirus. 35 complete genomes of betacoronavirus were also downloaded from the GISAID database. In our previous study, 10 complete genomes of betacoronavirus (GenBank: JX993987, JX993988, GQ153539, GQ153540, GQ153542, DQ071615, DQ412043, AY515512, AY572034 and DQ497008) were downloaded from the NCBI GenBank database and used for the analysis [12]. To trace the origin of SARS-CoV, five compelte genomes were added. They are DQ412042 (SARS-like CoV from *Rhinolophus ferrumequinum*), AY274119 (SARS-like CoV from human in Toronto, Tor2 [13]), AY278489 (SARS-like CoV from human in Guangdong, GD01 [14]), AY304486 (SARS-like CoV from civet [15]) and KF367457 (SARS-like CoV from bat [16]). DQ497008 was removed as a redundant sequence of AY274119 and AY278489. To trace the orgin of SARS-CoV-2, three compelte genomes were added. They are MN908947 (SARS-CoV-2), MN996532 (SARS2-like CoV hosted in Intermediate Horseshoe bats (*Rhinolophus affinis*) from Yunnan) and MG772934 (SARS2-like CoV hosted in Chinese horseshoe bats (*Rhinolophus sinicus*) from Zhejiang). A SARS2-like CoV (GISAID: EPI_ISL_410721) from pangolins (Collected in Guangdong, China) and a SARS2-like CoV (GenBank: MT040334) from pangolins (Collected in Guangxi, China) were also used. Plus RmYN01 and RmYN02 (GISAID: EPI_ISL_412976 and EPI_ISL_412977), totally 21 complete genomes were used for the phylogenetic analysis applying the neighbour joining (NJ) method. For

phylogenetic analysis, large segments spanning from S2 to ORF15 (**Supplementary 1**) were trimmed to only include CDS regions. Sequence alignment was performed using the Bowtie v0.12.7 software with paired-end alignment allowing 3 mismatches; mutation detection and other data processing were carried out using Perl scripts; the phylogenetic analysis was performed using MEGA v7.0.26; Statistics and plotting were conducted using the software R v2.15.3 the Bioconductor packages [17]. The structure of S (**Supplementary 2**) was predicted using trRosetta [18].

REFERENCES

1. C. Jiayuan, J. Shi, O. Yau Tung, C. Liu, X. Li, Q. Zhao, R. Jishou and G. Shan, *Bioinformatics Analysis of the 2019 Novel Coronavirus Genome.* Chinese Journal of Bioinformatics (In Chinese), 2020. **18**(2): p. 96-102.
2. X. Li, G. Duan, W. Zhang, J. Shi, J. Chen, S. Chen, S. Gao and J. Ruan, *A Furin Cleavage Site Was Discovered in the S Protein of the 2019 Novel Coronavirus.* Chinese Journal of Bioinformatics (In Chinese), 2020. **18**(2): p. 103-108.
3. G. Duan, J. Shi, Y. Xuan, J. Chen, C. Liu, J. Ruan, S. Gao and X. Li, *5' UTR Barcode of the 2019 Novel Coronavirus Leads to Insights into Its Virulence.* Chinese Journal of Virology (In Chinese), 2020. **36**(3): p. 365-369.
4. H. Zhou, X. Chen, T. Hu, J. Li, H. Song, Y. Liu, P. Wang, D. Liu, J. Yang and E.C. Holmes, *A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein.* Current Biology, 2020.
5. S.K.P. Lau, Y. Feng, H. Chen, H.K.H. Luk, W.H. Yang, K.S.M. Li, Y.Z. Zhang, Y. Huang, Z.Z. Song and W.N. Chow, *SARS coronavirus ORF8 protein is acquired from SARS-related coronavirus from greater horseshoe bats through recombination.* Journal of Virology, 2015: p. JVI.01048-15.
6. D. Muth, V.M. Corman, H. Roth, T. Binger, R. Dijkman, L.T. Gottula, F. Gloza-Rausch, A. Balboni, M. Battilani and D. Rihtarič, *Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission.* Scientific Reports, 2018. **8**(1).
7. Y.C. Su, D.E. Anderson, B.E. Young, F. Zhu, M. Linster, S. Kalimuddin, J.G. Low, Z. Yan, J. Jayakumar, L. Sun, G.Z. Yan, I.H. Mendenhall, Y.-S. Leo, D.C. Lye, L.-F. Wang and G.J. Smith, *Discovery of a 382-nt deletion during the early evolution of SARS-CoV-2.* bioRxiv, 2020. **1**(1): p. 1-23.
8. G. Xing-Yi, L. Jia-Lu, Y. Xing-Lou, C. Aleksei A, Z. Guangjian, E. Jonathan H, M. Jonna K, H. Ben, Z. Wei, P. Cheng, Z. Yu-Ji, L. Chu-Ming, T. Bing, W. Ning, Z. Yan, C. Gary, Z. Shu-Yi, W. Lin-Fa, D. Peter and S. Zheng-Li, *Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor*. Nature, 2013. **503**(7477): p. 535-538.
9. S. Perlman and J. Netland, *Coronaviruses post-SARS: update on replication and pathogenesis.* Nature Reviews Microbiology, 2009. **7**(6): p. 439.
10. M. Zhang, F. Zhan, H. Sun, X. Gong, Z. Fei and S. Gao. *Fastq_clean: An optimized pipeline to clean the Illumina sequencing data with quality control*. in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*. 2014. IEEE.
11. F. Zhang, T. Xu, L. Mao, S. Yan, X. Chen, Z. Wu, R. Chen, X. Luo, J. Xie and S. Gao, *Genome-wide analysis of Dongxiang wild rice (Oryza rufipogon Griff.) to investigate lost/acquired genes during rice domestication.* BMC plant biology, 2016. **16**(1): p. 1-11.
12. C. Liu, Z. Chen, Y. Hu, H. Ji, D. Yu, W. Shen, S. Li, J. Ruan, W. Bu and S. Gao, *Complemented Palindromic Small RNAs First Discovered from SARS Coronavirus.* Genes, 2018. **9**(9): p. 1-11.
13. R. He, F. Dobie, M. Ballantine, A. Leeson, Y. Li, N. Bastien, T. Cutts, A. Andonov, J. Cao and T.F. Booth, *Analysis of multimerization of the SARS coronavirus nucleocapsid protein.* Biochemical and biophysical research communications, 2004. **316**(2): p. 476-483.

14. Q. Wu, Y. Zhang, H. Lü, J. Wang, X. He, Y. Liu, C. Ye, W. Lin, J. Hu and J. Ji, *The E protein is a multifunctional membrane protein of SARS-CoV*. Genomics, proteomics & bioinformatics, 2003. **1**(2): p. 131-144.
15. Y. Guan, B. Zheng, Y. He, X. Liu, Z. Zhuang, C. Cheung, S. Luo, P. Li, L. Zhang and Y. Guan, *Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China*. Science, 2003. **302**(5643): p. 276-278.
16. X.-Y. Ge, J.-L. Li, X.-L. Yang, A.A. Chmura, G. Zhu, J.H. Epstein, J.K. Mazet, B. Hu, W. Zhang and C. Peng, *Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor*. Nature, 2013. **503**(7477): p. 535-538.
17. S. Gao, J. Ou and K. Xiao, *R language and Bioconductor in bioinformatics applications(Chinese Edition)*. 2014, Tianjin: Tianjin Science and Technology Translation Publishing Ltd.
18. J. Yang, I. Anishchenko, H. Park, Z. Peng and D. Baker, *Improved protein structure prediction using predicted interresidue orientations.* Proceedings of the National Academy of ences, 2020. **117**(3): p. 201914677.

Competing interests

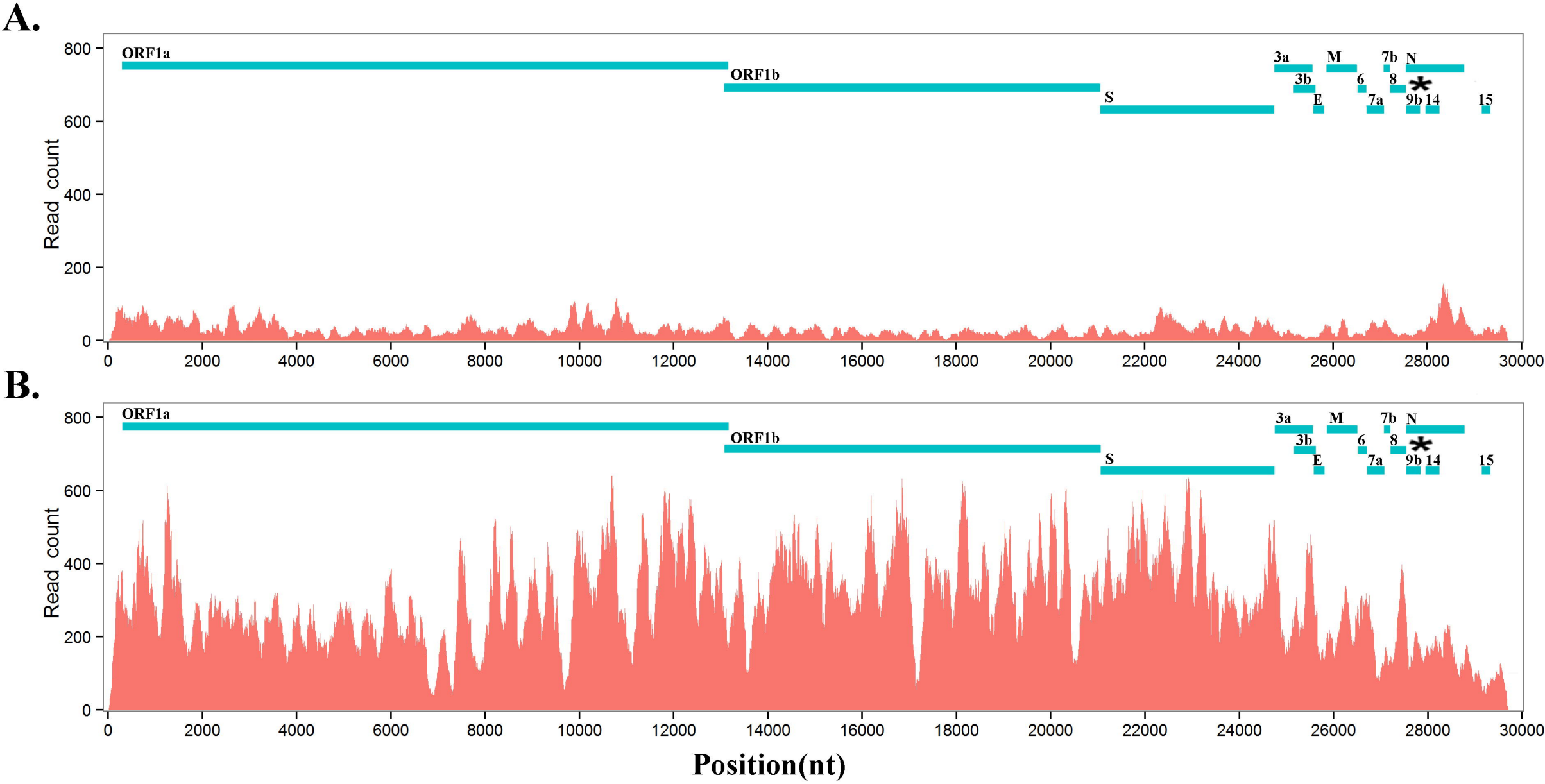
Non-financial competing interests

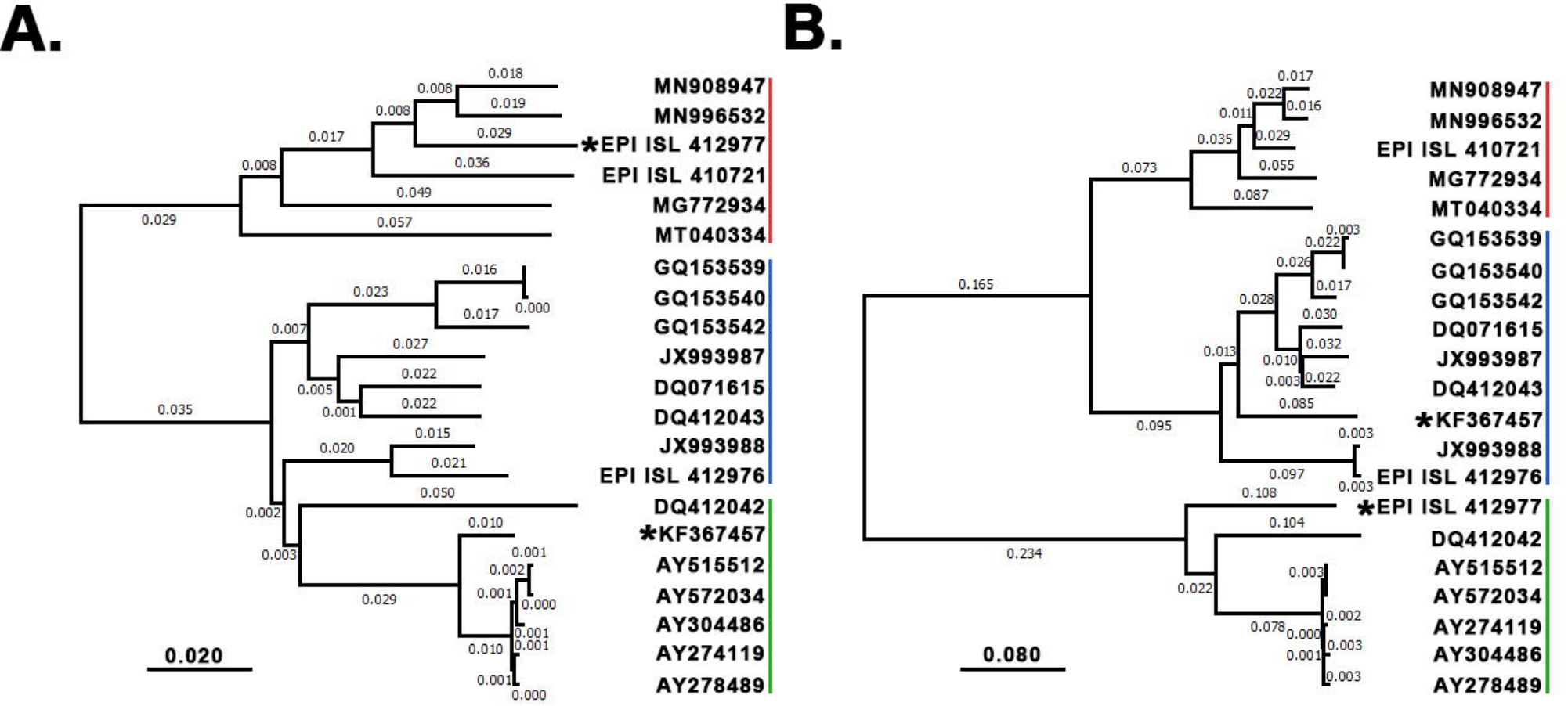
Acknowledgments

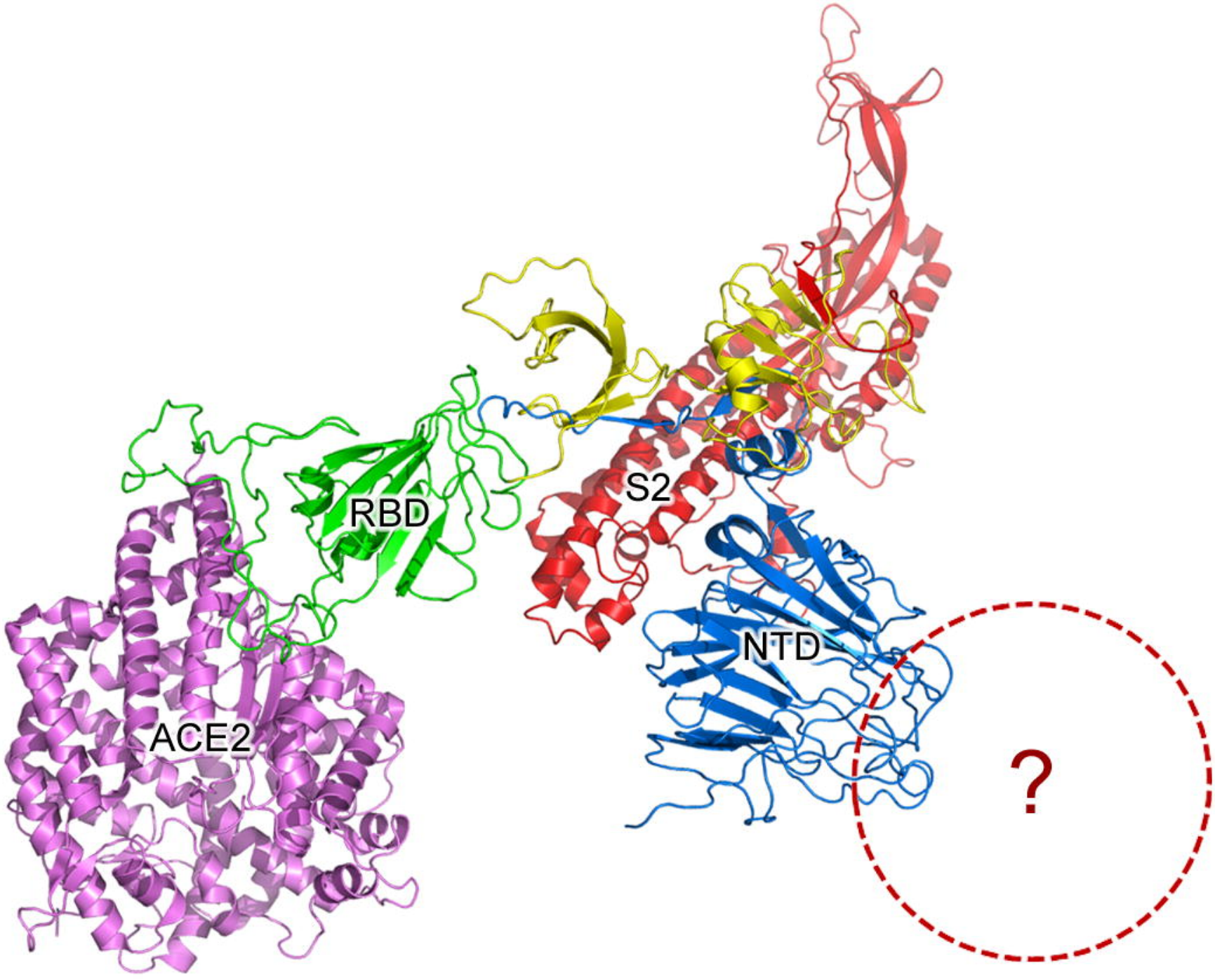
First, we thank Professor Weifeng Shi from Shandong First Medical University for his RNA-seq data sharing. We are grateful for the help from the following faculty members of College of Life Sciences at Nankai University: Deling Kong, Quan Chen, Wenjun Bu, Ting Ma, Tao Zhang, Dawei Huang, Mingqiang Qiao, Yanqiang Liu, Bingjun He and Zhen Ye. We also appreciate the cooperation and support from Professor Ze Chen and graduate student Yibo Xuan of Hebei Normal University. This project was supported by Yunnan Provincial Department of Education Science Research Fund Project Funding (No. 2018JS188) to Shunmei Chen and Tianjin Key Research and Development Program of China (19YFZCSY00500) to Shan Gao. We would like to thank Editage (www.editage.cn) for English language translation. This manuscript was online as a preprint on July 15nd, 2020 at https://www.researchgate.net/publication/343107038_The_discovery_of_a_recombinant_SARS2-like_CoV_strain_provides_insights_into_SARS_and_COVID-2019_pandemics

Author contributions statements

SG conceived the project. SG and GD supervised this study. XJ and SC conducted programming. XL, LW, and TY downloaded, managed and processed the data. JY predicted the structure of the S protein. JR analyzed the structure of S1. SG drafted the main manuscript text. SG and ZH revised the manuscript.


A.
B.
ORF1a
ORF1b
S
3a
3b
E
M
6
7a
7b
8
9b
N
14
15
Read count
800
600
400
200
0
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
24000
26000
28000
30000
Position(nt)


A.
MN908947
MN996532
*EPI ISL 412977
EPI ISL 410721
MG772934
MT040334
GQ153539
GQ153540
GQ153542
JX993987
DQ071615
DQ412043
JX993988
EPI ISL 412976
DQ412042
*KF367457
AY515512
AY572034
AY304486
AY274119
AY278489
0.020
B.
MN908947
MN996532
EPI ISL 410721
MG772934
MT040334
GQ153539
GQ153540
GQ153542
DQ071615
JX993987
DQ412043
*KF367457
JX993988
EPI ISL 412976
*EPI ISL 412977
DQ412042
AY515512
AY572034
AY274119
AY304486
AY278489
0.080


RBD
S2
NTD
ACE2
?