The Discovery of a Recombinant SARS2-like CoV Strain Provides Insight Into SARS and COVID-19 Pandemics

Xin Li

Nankai University

Xiufeng Jin

Nankai University

Shunmei Chen

Kunming Medical University

Liangge Wang

Nanjing University Medical School

Tung On Yau

Nottingham Trent University

Jianyi Yang

Nankai University

Zhangyong Hong

Nankai University

Jishou Ruan

Nankai University

Guangyou Duan

Qilu Normal University

Shan Gao (✉ gao_shan@mail.nankai.edu.cn)

Nankai University <https://orcid.org/0000-0002-8919-1338>

Research article

Keywords: SARS-CoV-2, furin cleavage, ORF8, recombination, alternative translation

Posted Date: August 20th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-60430/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: In December 2019, the world awoke to a new zoonotic strain of coronavirus named severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2).

Results: 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 we reported—for the first time—a recombination event of ORF8 at the whole-gene level in a bat, which had been co-infected by two betacoronavirus strains. 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. In the present study, we also found that the ancestor of betacoronavirus had a strong first Internal Ribosome Entry Site (IRES) and at least one furin cleavage site (FCS) in the junction region between S1 and S2 subunits.

Conclusions: We concluded that the junction FCS in SARS-CoV-2 may increase the efficiency of its 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-19 pandemics.

Background

A new zoonotic strain of coronavirus named severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) emerged in December 2019. Since SARS-CoV-2 is highly similar to SARS-CoV, most of studies focus on the investigation of the receptor binding domain (RBD) of Spike protein and its receptor ACE2 using the same strategies and methods as in SARS-CoV [1]. Different from these studies, we previously reported several other findings on SARS-CoV-2 for the first time, including the following in particular: (1) the alternative translation of Nankai coding sequence (CDS) that show the rapid mutation, multiple hosts and strong host adaptability of betacoronavirus at the molecular level [2]; (2) a furin cleavage site (FCS) in the junction region between S1 and S2 subunits (junction FCS) of SARS-CoV-2 that may increase the efficiency of viral entry into cells [3]; and (3) the use of 5' untranslated-region (UTR) barcoding for the detection, identification, classification and phylogenetic analysis of—though not limited to—coronaviruses [4]. 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 [4], respectively. The 3' end of each 5' UTR include the first Internal Ribosome Entry Site (IRES), which regulates the translational initiation of downstream genes. In particular, SARS-CoV-2 and SARS-CoV have strong first IRESs enhancing the translational initiation [4].

In the subsequent studies, InDels (insertions/deletions) at six mutation 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 [5] 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 was considered to have played a significant role in adaptation to human hosts following interspecies transmission [6] via the modification of viral replication [7]. 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 [7]. A 382-nt deletion during the early evolution of SARS-CoV-2 (GISAID: EPI_ISL_414378) was also reported [8]. In conjunction with our new discoveries, we conducted preliminary research and aimed at determining: (1) the critical recombination and mutations in SARS-CoV and SARS-CoV-2 genomes; (2) the influence of recombinant ORF8 on the phylogenetic analysis of betacoronaviruses; (3) the role of ORF8 in the modification of viral replication; and (4) the critical factors (e.g. a junction FCS, a strong first IRES and ORF8) which caused SARS and COVID-19 pandemics.

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 (Fig. 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 genomic site, 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.

Table 1
Annotations of recombinant regions and mutation sites

CDS	Start	End	Length(nt)	Start	End	Length(nt)
ORF1a	266	13483	13217	266	13477	13212
ORF1b	13768	21555	7788	13762	21549	7788
S	21563	25384	3822	21556	25239	3684
ORF3a	25393	26220	828	25248	26075	828
ORF3b/Nankai CDS	25814	26281	468	25669	26136	468
E	26245	26472	228	26100	26327	228
M	26523	27191	669	26378	27043	666
ORF6	27202	27387	186	27054	27239	186
ORF7a	27394	27759	366	27246	27611	366
ORF7b	27756	27887	132	27608	27739	132
ORF8	27894	28259	366	27746	28114	369
N/ORF9a	28274	29533	1260	28116	29375	1260
ORF9b	28284	28577	294	28126	28419	294
ORF14	28734	28955	222	28576	28797	222
ORF15	29558	29674	117	29400	29516	117
RC1	21761	21796	36	#AIHVSGTNGTKR		
RC2	21971	22054	84	#NDPFLGVYYHKNNKSWMESEFRVYSSAN		
RC3	22277	22348	72	#QTLALHRSYLT PGDSSSGWTAGA		
RC4	22874	22918	45	#SNNLDSKVGGNYNYL		
RC5	22964	23020	57	#ISTEIQAGSTPCNGVEGF		
M1	26109	26119	11	25964	25974	11
M2	26449	-3GAA		26303	-3GAA	

Six mutation sites named M1 to M6 in the viral genomes of SARS-CoV-2 (GenBank: MN908947) and RmYN02 (GISAID: EPI_ISL_412977). M1 has a 11-nt length in the genomes of viruses from the SARS-CoV-2 cluster, while M1 has a 8-nt length in the genomes of viruses from the SARS-CoV and SARS-like CoV clusters. As for M2 to M6, the deleted nucleotides belong to the viral genome of SARS-CoV (GenBank: AY278489). * Since RmYN02 has a type 2 ORF8, it has the same allele at the M5 site as SARS-CoV. Besides the recombination event in ORF8, other recombination events occurred in five regions, named RC1 to RC5 in the S1 subunit of the gene S. # These amino acid sequences are encoded by RC1 to RC5 from SARS-CoV-2 (GenBank: MN908947).

CDS	Start	End	Length(nt)	Start	End	Length(nt)
M3	27679	-3GAG		27530	-3GAG	
M4	27882	-3AAA		27733	-3AAA	
M5	27906	-3ATT		27757	ATT*	
M6	29512	-6AGCTTC		29353	-6AGCTTC	
<p>Six mutation sites named M1 to M6 in the viral genomes of SARS-CoV-2 (GenBank: MN908947) and RmYN02 (GISAID: EPI_ISL_412977). M1 has a 11-nt length in the genomes of viruses from the SARS-CoV-2 cluster, while M1 has a 8-nt length in the genomes of viruses from the SARS-CoV and SARS-like CoV clusters. As for M2 to M6, the deleted nucleotides belong to the viral genome of SARS-CoV (GenBank: AY278489). * Since RmYN02 has a type 2 ORF8, it has the same allele at the M5 site as SARS-CoV. Besides the recombination event in ORF8, other recombination events occurred in five regions, named RC1 to RC5 in the S1 subunit of the gene S. # These amino acid sequences are encoded by RC1 to RC5 from SARS-CoV-2 (GenBank: MN908947).</p>						

Recently, two betacoronavirus strains RmYN01 and RmYN02 (GISAID: EPI_ISL_412976 and EPI_ISL_412977) were detected from a bat (*Rhinolophus malayanus*) [5]. Since betacoronaviruses from subgroup B share many highly similar regions in their genome sequences (Fig. 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 [5]. 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, all of identified recombination events occurred in genes ORF1a, 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 small-size mutation (InDel & SNP) accumulation as most of them only occurred over very small genomic regions, excepting a few events (e.g., the 382-nt deletion in ORF8 [8]). We reported—for the first time—a recombination event at the whole-gene level in a bat, which had been co-infected by two betacoronavirus strains. Besides the recombination event in ORF8, other recombination events occurred over five regions, 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.

Using a large segment spanning S2, ORF3a, ORF3b, envelope (E), M, ORF6, ORF7a, ORF7b, N, ORF9b, ORF14 and ORF15 (Table 1), phylogenetic tree 1 (Fig. 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 M5 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 (Fig. 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 (Fig. 2B). WIV1, isolated from Chinese horseshoe bats (*Rhinolophus sinicus*), was considered the most closely related to SARS-CoVs in humans and civets [9]. 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 [6]. Although whether these recombination events occurred in bats or civets remains unknown [6], analysis of all recombination events in S1 subunits of all betacoronaviruses may help to find an answer (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 (Fig. 2B). RmYN02 was classified into cluster 1 in tree 1 but cluster 2 in tree 2. This suggested that 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 a junction FCS [3], and (2) the strains (e.g., GISAID: EPI_ISL_410721) are farther from SARS-CoV-2 than RaTG13 in the phylogenetic tree 1 (Fig. 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 simultaneously 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 increase replication efficiency of viruses. 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 (Fig. 1A), while the RmYN02 genome was covered 99.89% with an average depth of 298.99 (Fig. 1B). The RNA abundance of RmYN02 is 9 times that of RmYN01. Although this does not rule out other factors that contribute to the differences in RNA abundances of the two virus strains, these results combined with other evidence [6] suggest that type 2 ORF8 genes increase replication efficiency of RmYN02.

By analysis of all recombination events in S1 subunits of all betacoronaviruses, we obtained the following results: (1) there were a few genotypes of each recombinant region (RC1 to RC5); 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 were prerequisite for speciation of SARS-CoV and SARS-CoV-2. Further analysis showed that two recombination regions (RC4 to RC5) are localized in the receptor binding domain (RBD) of S1 (Fig. 3), while three other recombination regions (RC1, RC2 and RC3) are localized in the N-terminal domain (NTD). Almost all the secondary structures of five protein segments encoded by RC1 to RC5 are disordered, which are responsible for protein protein interaction (PPI). This suggested that the recombination of RC1 to RC5 improve the host adaptability of betacoronaviruses by enhancing interaction of RBD and NTD with their receptors, exhibiting that the positive selection of S was particularly strong [6]. Since both RBD and NTD have similar recombination events in their PPI regions, we proposed that NTD has a specific receptor as RBD has angiotensin-converting enzyme 2 (ACE2). Thus, the S1 subunit of SARS-CoV-2 may have more than one specific receptor (Fig. 3) as gp120 of HIV has the cluster of differentiation 4 receptor (CD4) and the C-C chemokine receptor 5 (CCR5).

Comprehensive analysis and reuse of data from different sources are necessary to determine the other receptor/s of SARS-CoV-2. A previous study identified two genetic susceptibility loci (rs11385942 at locus 3p21.31 and with rs657152 at locus 9q34.2) in COVID-19 patients with respiratory failure using genome-wide association analysis [10]. The locus 3p21.31 was associated to six genes SLC6A20, LZTFL1, CCR9, FYCO1, CXCR6 and XCR1. However, this study only focused on the further analysis of the locus 9q34.2 and confirmed a potential involvement of the ABO blood-group system. These researchers did not notice that three chemokine receptors CCR9, CXCR6 and XCR1 merit further investigation as candidates of SARS-CoV-2 receptors. The analysis of bulk RNA-seq data showed high expression of CCR9 and XCR1 in thymus and CXCR6 in T cells, compared to other tissues and cell types [10]. In particular, the thymic cells were consistently negative for ACE2 and many CoVs can infect thymus [11]. By investigating interaction of three protein segments encoded by RC1 to RC3 in NTD (Table 1) with CCR9, CXCR6 and XCR1, we found that CCR9 is the most possible candidate among three chemokine receptors. However, the final

determination of the other receptor/s of SARS-CoV-2 need more calculation and experiments on candidates at the whole-genome level. Our study did not rule out the possibilities of non-receptor proteins binding to NTD.

Rapid recombination and mutation, a strong first IRESs, a junction FCS and a type 2 ORF8 are four main factors contributing to transmission, virulence and host adaptability of coronavirus. By analysis of these four factors in 1,265 complete genomes of betacoronavirus, we concluded: (1) rapid recombination and mutation of viral genomes (particularly in the S gene) provide coronavirus the strong ability of cross-species transmission; (2) betacoronavirus jumps from bats to a new host species due to the host range expansion caused by recombination and mutations; (3) the ancestor of betacoronavirus (and even coronavirus) had a strong first IRES and at least one (most likely two) junction FCS, allowing for sustained transmission that can lead to significant outbreaks; (4) after jumping into a new host (e.g. hedgehog), betacoronavirus will be attenuated to spread widely in the host population by loss of junction FCSs or strong first IRESs and thus, it can not jump to another host species (e.g. human) if not triggered by new genetic events (e.g. acquiring a type 2 ORF8). The ancestor of betacoronavirus (and even coronavirus) had most likely two junction FCSs, which were close to N and C terminals. SARS-CoV-2 (GenBank: MN908947) has the junction FCS "RAAR" located at position 685 in the S protein, which is noted as R685. Although SARS-CoV-2 lost a junction FCS by substituting "KNTQ" (R779) for "RNTR", this FCS was inactivated in the alpha secondary structures. Due to the host range expansion, SARS-CoV-2 infected intermediate hosts (not necessary) and then humans before it attenuated. Originated from the same ancestor of SARS-CoV-2, SARS-like CoVs were attenuated by inactivation of the junction FCS "RNTR" (R761) and loss of the junction FCS (e.g. R667) close to N terminal. By acquiring a type 2 ORF8, SARS-CoV infected civets and then humans before being attenuated. After a period, a attenuated variant of SARS-CoV with a 29-nt deletion in the ORF8 gene was reported [7]. Originated from the same ancestor, the betacoronavirus strains RaTG13, RmYN02 and betacoronaviruses from pangolins were attenuated variants of SARS-CoV-2 by loss of junction FCSs [3]. In the betacoronavirus subgroup C, middle east respiratory syndrome coronavirus (MERS-CoV) have two junction FCSs (Fig. 4). The first one "RSTR" (R694) is likely to be dysfunctional, resulting in attenuation, as there is a disulfide bond cross the FCS R694. However, the second junction FCS "RSVR" (R751) is functional. Due to the host range expansion, MERS-CoV infected camels and then humans before it attenuated. Originated from the same ancestor of MERS-CoV, other MERS-like CoVs (e.g. hedgehog coronavirus) were attenuated by loss of junction FCSs. In the betacoronavirus subgroup A (Fig. 4), all strains were attenuated from the ancestor of betacoronavirus by loss of strong first IRESs. Since almost all betacoronaviruses of the subgroup A have junction FCSs, they (e.g. OC43 and HKU1) can still infect human with attenuated virulence. In addition, the average arginine (R) percentage of S proteins from betacoronaviruses of the subgroup A except mouse hepatitis virus (MHV) 2.63% is significantly lower than those from MHV, betacoronaviruses of the subgroup B, C and D (3.34%, 3.33%, 3.32% and 3.33%). This indicated that accumulated mutations caused attenuation by loss of arginine residues, since arginine residues usually reside at protease cleavage sites. In the betacoronavirus subgroup D, HKU9 does not have a junction FCS but have a strong first IRES. Therefore, HKU9 has the potential ability to cause an epidemic.

Conclusions

Betacoronaviruses have very high recombination and mutation 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 recombination and mutations, betacoronavirus subgroup B was classified into the SARS-CoV-2, SARS-CoV and SARS-like CoV clusters. As all recombination and mutations are unlikely to undergo reversible changes together, we concluded the following: (1) the SARS-CoV-2 cluster is more closely than the SARS-CoV and SARS-like CoV clusters related to the ancestor of betacoronavirus, which had at least one junction FCS; and (2) the SARS-CoV-2 cluster separated from the two other clusters and later, the SARS-CoV cluster separated from the SARS-like CoV cluster by acquiring a type 2 ORF8. The junction FCS in SARS-CoV-2 may increase the efficiency of its entry into cells, while the type 2 ORF8 acquired by SARS-CoV may increase its replication efficiency. These two most critical factors improved the adaptive ability of SARS-like CoV and SARS2-like CoV in bats to new hosts and allowed for sustained transmission that can lead to significant outbreaks. The occurrences of all recombination events in bats were prerequisite for speciation of SARS-CoV and SARS-CoV-2. However, they were not the most critical events to cause the pandemics of SARS and COVID-19. Based on our theories, two predictions can be made: (1) SARS2-like CoV containing at least one junction FCS will be detected in bats; and (2) the attenuated variant of SARS-CoV-2 without junction FCS will appear.

A previous study showed that the integrity of ORF8 facilitates the replication of SARS-CoV in cell lines [7]. 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 [12]. 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 three types of ORF8 provide insight into SARS pandemics.

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 [2]. However, this region encodes several ORFs in other betacoronaviruses. The alternative translation of Nankai CDS could produce more than 17 putative proteins. Given that this region is not always an ORF, we use Nankai CDS (a 465- or 468-bp genomic region) to name ORF3b and its homologous sequences in all coronaviruses. Nankai CDS encodes ORF3b in SARS-CoV, while it encodes shorter genes in many other SARS-like CoVs. ORF3b in SARS-CoV is predicted to be translated into a 155-aa (amino acid residue) protein—the largest one among all putative proteins predicted from Nankai CDS in all betacoronaviruses.

Materials And Methods

The software VirusDetect [13] was used to detect viruses in RNA-seq data from [5]. The software Fastq_clean [14] was used for RNA-seq 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 [15] 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 selected and used for the analysis [16]. To trace the origin of SARS-CoV, five complete genomes were added in the present study. They are DQ412042 (SARS-like CoV from *Rhinolophus ferrumequinum*), AY274119 (SARS-like CoV from human in Toronto, Tor2 [17]), AY278489 (SARS-like CoV from human in Guangdong, GD01 [18]), AY304486 (SARS-like CoV from civet [19]) and KF367457 (SARS-like CoV from bat [20]). DQ497008 was removed as a redundant sequence of AY274119 and AY278489. To trace the origin of SARS-CoV-2, three complete 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 [21]. The structure of S (**Supplementary 2**) was predicted using trRosetta [22].

Declarations

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. We would like to thank Editage (www.editage.cn) for English language translation. This manuscript was online as a preprint on July 22nd, 2020 at <https://biorxiv.org/cgi/content/short/2020.07.22.213926v1>.

Funding

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.

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.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Non-financial competing interests

References

1. Graham RL, Baric RS. Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission. *J Virol*. 2010;84(7):3134–46.
2. Jiayuan C, Shi J, Yau Tung O, Liu C, Li X, Zhao Q, Jishou R, Shan G. Bioinformatics Analysis of the 2019 Novel Coronavirus Genome. *Chinese Journal of Bioinformatics (In Chinese)*. 2020;18(2):96–102.
3. Li X, Duan G, Zhang W, Shi J, Chen J, Chen S, Gao S, Ruan J. A Furin Cleavage Site Was Discovered in the S Protein of the 2019 Novel Coronavirus. *Chinese Journal of Bioinformatics (In Chinese)*. 2020;18(2):103–8.
4. Duan G, Shi J, Xuan Y, Chen J, Liu C, Ruan J, Gao S, Li X, *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.
5. Zhou H, Chen X, Hu T, Li J, Song H, Liu Y, Wang P, Liu D, Yang J, Holmes EC. *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.

6. Lau SKP, Feng Y, Chen H, Luk HKH, Yang WH, Li KSM, Zhang YZ, Huang Y, Song ZZ, Chow WN. *SARS coronavirus ORF8 protein is acquired from SARS-related coronavirus from greater horseshoe bats through recombination*. Journal of Virology, 2015: p. JVI.01048-15.
7. Muth D, Corman VM, Roth H, Binger T, Dijkman R, Gottula LT, Gloza-Rausch F, Balboni A, Battilani M, Rihtarič D. *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).
8. Su YC, Anderson DE, Young BE, Zhu F, Linster M, Kalimuddin S, Low JG, Yan Z, Jayakumar J, Sun L, Yan GZ, Mendenhall IH, Leo Y-S, Lye DC, Wang L-F, Smith GJ. Discovery of a 382-nt deletion during the early evolution of SARS-CoV-2. bioRxiv. 2020;1(1):1–23.
9. Xing-Yi G, Jia-Lu L, Xing-Lou Y, Aleksei C, Guangjian AZ, Jonathan E, Jonna HM, Ben KH, Wei Z, Cheng P, Yu-Ji Z, Chu-Ming L, Bing T, Ning W, Yan Z, Gary C, Shu-Yi Z, Lin-Fa W, Peter D. and S. Zheng-Li, *Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor*. Nature. 2013;503(7477):535–8.
10. Ellinghaus D, Degenhardt F, Bujanda L, Buti M, Albillos A, Invernizzi P, Fernández J, Prati D, Baselli G, Asselta R, Grimsrud MM, Milani C, Aziz F, Kässens J, May S, Wendorff M, Wienbrandt L, Uellendahl-Werth F, Zheng T, Yi X, de Pablo R, Chercoles AG, Palom A, Garcia-Fernandez A-E, Rodriguez-Frias F, Zanella A, Bandera A, Protti A, Aghemo A, Lleo A, Biondi A, Caballero-Garralda A, Gori A, Tanck A, Carreras Nolla A, Latiano A, Fracanzani AL, Peschuck A, Julià A, Pesenti A, Voza A, Jiménez D, Mateos B, Nafria Jimenez B, Quereda C, Paccapelo C, Gassner C, Angelini C, Cea C, Solier A, Pestaña D, Muñiz-Diaz E, Sandoval E, Paraboschi EM, Navas E, García Sánchez F, Ceriotti F, Martinelli-Boneschi F, Peyvandi F, Blasi F, Téllez L, Blanco-Grau A, Hemmrich-Stanisak G, Grasselli G, Costantino G, Cardamone G, Foti G, Aneli S, Kurihara H, ElAbd H, My I, Galván-Femenia I, Martín J, Erdmann J, Ferrusquía-Acosta J, Garcia-Etxebarria K, Izquierdo-Sanchez L, Bettini LR, Sumoy L, Terranova L, Moreira L, Santoro L, Scudeller L, Mesonero F, Roade L, Rühlemann MC, Schaefer M, Carrabba M, Riveiro-Barciela M, Figuera Basso ME, Valsecchi MG, Hernandez-Tejero M, Acosta-Herrera M, D'Angiò M, Baldini M, Cazzaniga M, Schulzky M, Cecconi M, Wittig M, Ciccarelli M, Rodríguez-Gandía M, Boccione M, Miozzo M, Montano N, Braun N, Sacchi N, Martínez N, Özer O, Palmieri O, Faverio P, Preatoni P, Bonfanti P, Omodei P, Tentorio P, Castro P, Rodrigues PM, Blandino Ortiz A, de Cid R, Ferrer R, Gualtierotti R, Nieto R, Goerg S, Badalamenti S, Marsal S, Matullo G, Pelusi S, Juzenas S, Aliberti S, Monzani V, Moreno V, Wesse T, Lenz TL, Pumarola T, Rimoldi V, Bosari S, Albrecht W, Peter W, Romero-Gómez M, M. D'Amato, S. Duga, J.M. Banales, J.R. Hov, T. Folseraas, L. Valenti, A. Franke and T.H. Karlsen, *Genomewide Association Study of Severe Covid-19 with Respiratory Failure*. New England Journal of Medicine, 2020.
11. Lins MP, Smaniotto S. Potential impact of SARS-CoV-2 infection on the thymus. Can J Microbiol. 2020;66(8):1–20.
12. Perlman S, Netland J. Coronaviruses post-SARS: update on replication and pathogenesis. Nat Rev Microbiol. 2009;7(6):439.

13. Zheng Y, Gao S, Padmanabhan C, Li R, Galvez M, Gutierrez D, Fuentes S, Ling K, Kreuze JF, Fei Z. VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology*. 2016;500(2017):130–8.
14. Zhang M, Zhan F, Sun H, Gong X, Fei Z, Gao S. *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.
15. Zhang F, Xu T, Mao L, Yan S, Chen X, Wu Z, Chen R, Luo X, Xie J, Gao S. Genome-wide analysis of Dongxiang wild rice (*Oryza rufipogon* Griff.) to investigate lost/acquired genes during rice domestication. *BMC plant biology*. 2016;16(1):1–11.
16. Liu C, Chen Z, Hu Y, Ji H, Yu D, Shen W, Li S, Ruan J, Bu W, Gao S. Complemented Palindromic Small RNAs First Discovered from SARS Coronavirus. *Genes*. 2018;9(9):1–11.
17. He R, Dobie F, Ballantine M, Leeson A, Li Y, Bastien N, Cutts T, Andonov A, Cao J, Booth TF. Analysis of multimerization of the SARS coronavirus nucleocapsid protein. *Biochem Biophys Res Commun*. 2004;316(2):476–83.
18. Wu Q, Zhang Y, Lü H, Wang J, He X, Liu Y, Ye C, Lin W, Hu J, Ji J. The E protein is a multifunctional membrane protein of SARS-CoV. *Genom Proteom Bioinform*. 2003;1(2):131–44.
19. Guan Y, Zheng B, He Y, Liu X, Zhuang Z, Cheung C, Luo S, Li P, Zhang L, Guan Y. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science*. 2003;302(5643):276–8.
20. Ge X-Y, Li J-L, Yang X-L, Chmura AA, Zhu G, Epstein JH, Mazet JK, Hu B, Zhang W, Peng C. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature*. 2013;503(7477):535–8.
21. Gao S, Ou J, Xiao K. R language and Bioconductor in bioinformatics applications(Chinese Edition). Tianjin: Tianjin Science and Technology Translation Publishing Ltd.; 2014.
22. Yang J, Anishchenko I, Park H, Peng Z, Baker D, *Improved protein structure prediction using predicted interresidue orientations*. *Proceedings of the National Academy of Sciences*, 2020. **117**(3): p. 201914677.

Figures

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). The RNA abundance of RmYN02 is 9 times that of RmYN01. 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.

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 represent the viral genomes. 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.

Figure 3

S protein of SARS-CoV-2 may have more than one specific receptor The S protein is cleaved into two subunit S1 and S2 (in red color) for receptor binding and membrane fusion. S1 has two domains, 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 [22].

Figure 4

Furin cleavage sites of betacoronaviruses MERS-CoV have the furin cleavage sites (FCSs) “RSTR” and “RSVR”, while SARS-CoV and all other SARS-like CoVs do not have FCSs in the junction region between S1 and S2 subunits. However, there is a disulfide bond (in blue color) cross the FCS “RSTR”. MERS CoV (GenBank: JX869059) belongs to subgroup C; SARS-CoV (GenBank: AY278489) and SARS-CoV-2 (GenBank: MN908947) belong to subgroup B; Human OC43 (GenBank: KF530084), Human HKU1 (GenBank: KF686346), Yak CoV (GenBank: MH810163), PHEV (GenBank: KY419107), Waterbuck CoV (GenBank: FJ425186), Bovine CoV (GenBank: MH043954), Giraffe CoV (GenBank: EF424622), Sable antelope CoV (GenBank: EF424621), Water deer CoV (GenBank: MG518518), Camel CoV (GenBank: MN514963), Canine CoV (GenBank: JX860640), Horse CoV (GenBank: LC061274), Rat CoV (GenBank: JF792617) and MHV (GenBank: AF029248) belong to subgroup A. PHEV: porcine hemagglutinating encephalomyelitis virus; MHV: mouse hepatitis virus.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [s1.txt](#)