OXFORD

Briefings in Bioinformatics, 00(00), 2020, 1–12

doi: 10.1093/bib/bbaa262
Case Study

Compositional diversity and evolutionary pattern of coronavirus accessory proteins

Jingzhe Shang, Na Han, Ziyi Chen, Yousong Peng, Liang Li, Hangyu Zhou, Chengyang Ji, Jing Meng, Taijiao Jiang and Aiping Wu

Corresponding author: Aiping Wu, Suzhou Institute of Systems Medicine, Suzhou, Jiangsu 215123, China. Tel.: +86-512-62873529; Fax: +86-0512-62873779; E-mail: wap@ism. cams.cn

Abstract

Accessory proteins play important roles in the interaction between coronaviruses and their hosts. Accordingly, a comprehensive study of the compositional diversity and evolutionary patterns of accessory proteins is critical to understanding the host adaptation and epidemic variation of coronaviruses. Here, we developed a standardized genome annotation tool for coronavirus (CoroAnnoter) by combining open reading frame prediction, transcription regulatory sequence recognition and homologous alignment. Using CoroAnnoter, we annotated 39 representative coronavirus strains to form a compositional profile for all of the accessary proteins. Large variations were observed in the number of accessory proteins of 1–10 for different coronaviruses, with SARS-CoV-2 and SARS-CoV having the most (9 and 10, respectively). The variation between SARS-CoV and SARS-CoV-2 accessory proteins could be traced back to related coronaviruses in other hosts. The genomic distribution of accessory proteins had significant intra-genus conservation and inter-genus diversity and could be grouped into 1, 4, 2 and 1 types for alpha-, beta-, gamma-, and delta-coronaviruses, respectively. Evolutionary analysis suggested that accessory proteins are more conservative locating before the N-terminal of proteins E and M (E-M), while they are more diverse after these proteins. Furthermore, comparison of virus-host interaction networks of SARS-CoV-2 and SARS-CoV accessory proteins showed that they share multiple antiviral signaling pathways, those involved in the apoptotic process, viral life cycle and response to oxidative stress. In summary, our study provides a tool for coronavirus genome annotation and builds a comprehensive profile for coronavirus accessory proteins covering their composition, classification, evolutionary pattern and host interaction.

Key words: accessory proteins; coronavirus; evolution; compositional diversity

Jingzhe Shang is an Assistant Professor at the Suzhou Institute of Systems Medicine, Center for Systems Medicine, Chinese Academy of Medical Sciences and Peking Union Medical College, Jiangsu, Suzhou, China. He has been working in the field of virus-host interaction during virus infection.

Na Han is a Research Assistant at the Suzhou Institute of Systems Medicine, Center for Systems Medicine, Chinese Academy of Medical Sciences and Peking Union Medical College, Jiangsu, Suzhou, China. His research interests are multi-omics data integration and analysis.

Ziyi Chen is a PhD student at the Suzhou Institute of Systems Medicine, Center for Systems Medicine, Chinese Academy of Medical Sciences and Peking Union Medical College, Jiangsu, Suzhou, China. He is good at multi-omics data integration and analysis.

Yousong Peng is a Professor on the College of Biology at Hunan University, China. He is working in computational biology for infectious diseases.

Liang Li is a doctor-in-charge at the Linyi people’s hospital, Shandong, China. He is engaged in clinical work.

Hangyu Zhou is a Postdoc at the Suzhou Institute of Systems Medicine, Center for Systems Medicine, Chinese Academy of Medical Sciences and Peking Union Medical College, Jiangsu, Suzhou, China. He is good at vaccine development.

Chengyang Ji is a PhD student at the Suzhou Institute of Systems Medicine, Center for Systems Medicine, Chinese Academy of Medical Sciences and Peking Union Medical College, Jiangsu, Suzhou, China. He is good at evolution analysis.

Jing Meng is a Postdoc at the Suzhou Institute of Systems Medicine, Center for Systems Medicine, Chinese Academy of Medical Sciences and Peking Union Medical College, Jiangsu, Suzhou, China. Her research interests are in bioinformatics and deep learning.

Taijiao Jiang is a Professor at the Center for Systems Medicine, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. He is an expert in systems medicine and bioinformatics for infectious diseases.

Aiping Wu is a Professor at the Suzhou Institute of Systems Medicine, Center for Systems Medicine, Chinese Academy of Medical Sciences and Peking Union Medical College, Jiangsu, Suzhou, China. He is working in computational biology for infectious diseases.

Submitted: 24 July 2020; **Received (in revised form):** 29 August 2020

Introduction

Recently, a new coronavirus that causes the severe respiratory disease COVID-19, SARS-CoV-2, has become prevalent worldwide [1, 2]. The World Health Organization declared that the SARS-CoV-2 epidemic was an international emergency public health situation on 30 January 2020 [3]. As of 23 July 2020, the SARS-CoV-2 virus had expanded to 188 countries or regions, with more than 15.2 million people diagnosed. SARS-CoV-2 is the third coronavirus found to cause severe human diseases. As early as 2003, the SARS-CoV coronavirus spread worldwide, leading to infection of more than 8000 and a fatality rate of nearly 10% [4, 5]. Ten years later, another coronavirus with a lethality rate of more than 20%, MERS-CoV, spread from the Arabian Peninsula to 27 countries [6–9]. Although the fatality rate of SARS-CoV-2 (about 3.61%) is lower than that of the previous two viruses [10], this virus is more contagious with high R0 values estimated at 2–6.47 [11–13]. The frequent emergence of different types of coronavirus that cause serious diseases has raised important questions about the diversity and evolutionary associations of these viruses.

Coronaviruses are positive-stranded, single-stranded RNA enveloped viruses with the largest genome among RNA viruses [14]. Coronaviruses have a certain of replication fidelity, which may explain their relatively large genomes [15]. All coronaviruses have a similar genomic structure. At the 5′ end, two-thirds of the genome comprises two large open reading frames (ORFs) (ORF1a and ORF1b) encoding the coronavirus replicase, which is highly conserved among genera. At the 3′ end, the genome encodes four structural proteins (S, E, M and N) and a variable number of accessory proteins. Based on the highly conserved ORF1ab coding region, coronaviruses can be divided into four genera: alpha-, beta-, gamma- and delta-coronavirus. Alpha- and beta-coronaviruses mainly infect mammals, while gamma- and delta-coronaviruses primarily infect birds [16]. In addition to SARS-CoV, MERS-CoV and SARS-CoV-2, there are currently four other coronaviruses that can infect humans, HCoV-229E [17], HCoV-OC43 [18], HCoV-NL63 [19] and HCoV-HKU1 [20]. These viruses are all either alpha- or beta-coronaviruses. Studies have shown that the genome compositions of these seven human-infecting coronaviruses are significantly different, especially for accessory proteins [21, 22].

Coronaviruses have a unique discontinuous transcription mechanism. By recognizing specific transcription regulating sequences (TRSs), mature mRNA can be obtained by one-time transcription [23]. In theory, when naming encoded proteins, especially coronavirus variable accessory proteins, it is necessary to refer to the location of the TRS in the genome. However, because there is not a standardized genome annotation tool, some genome annotation studies do not have location information for the TRS. This makes the annotation and naming of coronaviruses a bit confusing, especially for accessory proteins that vary greatly in number and composition among different viral strains. Therefore, the need for a standardized method and tool for coronavirus genome annotation is urgent.

Coronavirus accessory proteins are normally located behind structural proteins, and each location may encode a different number of accessory proteins, resulting in a huge complexity of accessory proteins. Previous studies have shown that accessory proteins play an important role in virus–host interactions, especially in antagonizing or regulating host immunity and virus adaptation to the host [24]. However, there is still no normalized annotation and analysis for coronavirus accessory proteins, and no in-depth explorations of the evolutionary relationships of accessory proteins between different strains have been conducted to date (Figure 1A). To address these issues, we first developed a semi-automatic coronavirus annotation tool named CoroAnnoter for standardized annotation of all coronaviruses. We then obtained a comprehensive genome composition profile for all of the representative coronaviruses. The composition and evolutionary analysis revealed large variations in accessory proteins between strains as well as their inherent conservation of evolutionary patterns.

Materials and methods

Data sources

The representative sequences of the coronavirus genomes were obtained from the genome available in the National Center for Biotechnology Information (NCBI) database (Supplementary Table 1). The protein sequences of the self-built BLAST database were mainly obtained from the Virus Pathogen Resource (ViPR), including 86 747 protein sequences of all coronaviruses collected up to 16 February 2020. Some protein sequences were also from the annotated proteins in the genome in the NCBI database.

Genome annotation via CoroAnnoter

The coronavirus genome annotation process consists of the following six steps. (i) ORFfinder is used to predict the ORFs of the genome. The ORF starting sequence is ATG, the genetic code is set to the standard code and the minimum predicted length is 60 bp. (ii) The BLAST program is used to conduct sequence alignment to obtain credible ORFs with sequence similarities. The E value is set to 1 E-2, and the first three sequences with the best match are retained. The BLAST results are then manually checked, and the two redundant results are removed. (iii) The R script is used to obtain the 100 nucleotides before the 5′ end of the credible ORFs. (iv) The MEME kit is used to identify conservative transcription regulatory sequences (TRSs) [25]. The default classic method is used to select the zoos motif distribution mode and the motif length is between 6 and 8 nucleotides. The five best results are kept for manual selection of the most suitable motif as the viral TRS. In addition to the leader sequence, the TRSs of other ORFs are normally located at the 3′ end of 100 nucleotides. (v) Identify the position of TRS on the genome through the R script. (vi) The coronavirus genome is then annotated by integrating the information from the TRS and ORFs.

Sequence alignment

The similarity of accessory proteins is measured based on the score of the pairwise sequence alignment. We use the pairwise alignment function of the R package ‘Biostrings’ for sequence alignment [26]. The comparison type was global comparison based on the Needleman–Wunsch algorithm and the substitution Matrix was BLOSUM62.

Visualizing sequence similarity

The CDS sequences of the accessory proteins of each coronavirus were combined according to their orders in the genome to form a virtual accessory protein sequence. Coronavirus accessory protein sequences of the same genus were placed into a file and submitted to Circoletto (http://tools. bat.infspire.org/circoletto/), an online server that visualizes similar sequences [27]. The E value of cutoff is selected as 1 E-5, while the defaults were used for all parameters.

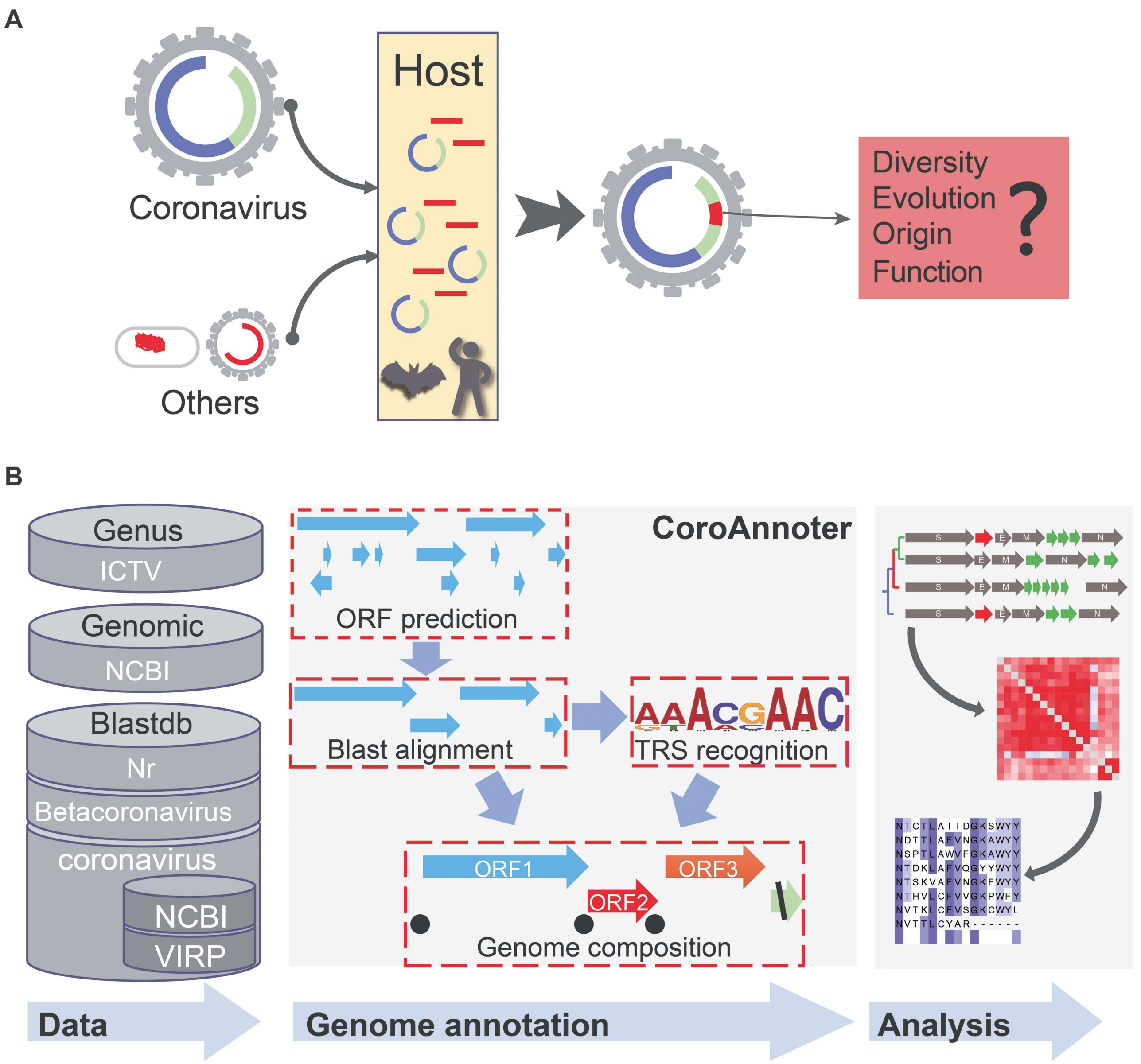


Figure 1. Methodology of the CoroAnnoter. (A) Unknown characteristics of the diversity, evolution, origin and function of coronavirus accessory protein analysis. Three modules are included. Data preparation included the selection of representative strains, the preparation of genomic sequences and the construction of BLAST database. Finally, the composition and evolutionary patterns of coronavirus accessory proteins were systematically annotated and analyzed.

Interaction network between accessory proteins and human proteins

We collected all interaction proteins between coronaviruses and humans up to 1 May 2020. We then used the R package 'clusterprofiler' to conduct GO and KEGG annotation of the proteins [28], after which we removed the redundant annotation results based on the kappa coefficient (Kappa similarity >0.3). The IPA software was used to obtain interaction information between human proteins. The above information was then integrated and visualized using the cytoscape software [29].

Code availability

CoroAnnoter is free-software and is available at https://github.com/wuaipinglab/CoroAnnoter.

Results

CoroAnnoter: a standardized genome annotation tool for coronavirus

The unique TRS recognition-based transcription mechanism can be used to guide the development of genome annotation tools for coronaviruses [30, 31]. By combining ORF prediction, TRS recognition and homologous alignment, we developed a semi-automatic and standardized genome annotation tool for coronaviruses named CoroAnnoter (Figure 1B). CoroAnnoter consists of the following steps (Supplementary Figure 1). First, the ORFs of the coronavirus genome sequence are predicted by ORFfinder [32], after which the potential ORFs are filtered based on sequence similarities via BLAST. After filtering, there were still some ORFs that should not be transcribed. To solve this problem, we introduced TRS position recognition to determine exactly which sub-genomes were transcribed. We extracted the 100 nucleotides before the 5′ terminal of each ORF to predict the conserved motifs using the MEME kit [33], and then, we manually determined the appropriate TRS core sequences (CSs) with 6–8 nucleotides. Finally, we annotated the coronavirus genome structure by integrating ORFs, blast similarities and TRS positions. The ORF following each TRS is an individual sub-genomic fragment. If more than one ORF is present at the same TRS position, then they can be named a, b, etc. (e.g. as 3a, 3b, etc.). The CoroAnnoter tool is freely available at https://github.com/wuaipinglab/CoroAnnoter.

One advantage of CoroAnnoter is to standardize the naming of identified ORFs based on the TRS information. For example, MERS-CoV and Bat-CoV-HKU4 were found to share a similar genome structure and have the same TRS CS as ACGAAY. The ORF names of some corresponding segments are different between these two coronaviruses. Four ORFs named as 3, 4a, 4b and 5 in MERS-CoV were named as 3a, 3b, 3c and 3d in Bat-CoV-HKU4, respectively [34]. However, these four ORFs could be named as the consistent 3, 4a, 4b and 5 with CoroAnnoter because their same TRS information locating before ORF3, ORF4a and ORF5, respectively. Therefore, with CoroAnnoter, the genome structure could be uniformly annotated in 39 representative species for all currently known coronaviruses.

Compositional diversity of coronavirus accessory proteins

Given the importance of TRS in the unique sub-genomic transcription mechanism of coronaviruses, a conserved TRS CS was identified for each type of coronavirus (Supplementary Table 2). Systematic comparison revealed that the TRS CSs of coronaviruses in the same genus were highly similar. The original TRS CSs of each genus were most likely to be CTAAAC (alpha), ACGAAC (beta), AACAA (gamma) and ACACCA (delta). Within each genus, the TRS sequence of every species may possess specific mutations introduced into the genus TRS (Figure 2). For example, the TRS of Human-Cov-NL63 strain of the alpha genus is CTMAAC, in which the third base M indicates A or C. In addition, some species-specific TRSs are inconsistent with their genus TRSs. For example, the TRS of the Human-CoV-OC43 strain of beta genus is TYYAAAC, which is very different from ACGAAC (beta) and more similar to CTAAAC (alpha).

Based on the comprehensive annotation profile of the coronavirus genome, we found that there are large variations in the number of accessory proteins (1–10) among coronaviruses (Figure 2). The number of accessory proteins of the alpha-coronaviruses is relatively lower, between 1 and 5, while beta-coronaviruses have 3–5 accessory proteins, except for SARS-CoV and SARS-CoV-2, which possess the largest number of accessory proteins among all coronaviruses (10 and 9, respectively). When compared with the evolutionarily similar strain, Bat-CoV-Hp, SARS-CoV and SARS-CoV-2 were found to have more complex compositions of accessory proteins. SARS-CoV and SARS-CoV-2 related viruses from non-human hosts showed similar complex accessory protein compositions (Supplementary Figure 2). The compositional variations between SARS-CoV and SARS-CoV-2 accessory proteins could be found from their evolutionarily related viruses. For example, protein 3b splits into a shorter 3c protein in some viruses, and even loses 3b, leaving only the 3c protein. SARS-CoV possesses the completed 3b protein, while SARS-CoV-2 contains only the shorter 3c protein. Protein 8 splits into 8a and 8b in some strains. SARS-CoV contains shorter 8a and 8b proteins, while SARS-CoV-2 has the complete protein 8. There are only two representative coronaviruses in the gamma genus, but they have relatively more accessory proteins (6 and 8, respectively). The number of accessory proteins in delta-coronaviruses is between 2 and 8.

To date, three of the seven human-infecting coronaviruses (SARS-CoV, MERS-CoV and SARS-CoV-2) have been shown to cause severe symptoms (Figure 2). We found that the number of accessory proteins of these three viruses is relatively high, between 5 and 10, while the other four viruses contained between 1 and 4 accessory proteins. Viruses from the alpha-genus (229E and NL63) were found to have the lowest number of accessory proteins, 2 and 1, respectively. When compared with other viruses of the alpha-genus, these viruses lacked the accessory proteins behind the N protein. Another two human-infecting coronaviruses, HKU1 and HKU24, both contain the HE protein and have 3 or 4 accessory proteins, respectively.

The distribution pattern of accessory proteins in the coronavirus genome

The distribution pattern of coronavirus accessory proteins was found to have intra-genus conservation and inter-genus diversity (Figures 2 and 3). Based on the distribution characteristics of coronavirus accessory proteins, we can divide the structural compositions of the accessory proteins into eight types, namely, Alpha, Beta-Lineage-A, Beta-Lineage-B, Beta-Lineage-C, Beta-Lineage-D, Gamma-Lineage-A, Gamma-Lineage-B and Delta (Figure 3A). All alpha-coronaviruses belong to the conservative Alpha type, in which 1–2 accessory proteins after the S protein and multiple accessory proteins after the N protein were observed (Figure 3B). However, beta-coronaviruses are highly diverse and consist of four compositional types of accessory proteins, lineages A, B, C and D. Beta-lineage-A strains have 2a and HE protein before the S protein, as well as 1–2 accessory proteins behind the S protein. Beta-lineage-B strains, including SARS-CoV and SARS-CoV-2, possess multiple accessory proteins located behind S and M, respectively. Members of Beta-lineage-C, which include MERS-CoV, possess four accessory proteins between S and E. Accessory proteins in Beta-Lineage-D have similar distributions as those of the alpha type (Figure 3C). The two-representative gamma-coronaviruses were found to have completely different accessory protein compositions. Gamma-Lineage-A contains all of the accessory proteins located between M and N proteins, while the distribution of Gamma-Lineage-B accessory proteins is similar to that of Beta-Lineage-B proteins, with accessory proteins located behind the S and M proteins, respectively (Figure 3D). In the Delta type, there is one accessory protein behind the M protein, as well as multiple accessory proteins behind the N proteins (Figure 3E). Interestingly, although the coronavirus accessory proteins have different compositions, the E and M proteins are always linked together with no accessory protein between them.

Sequence conservation and diversity among accessory proteins

We hypothesized that accessory proteins located at the same genomic position could have a close phylogenetic relationship and share similar sequences. To test this hypothesis, we merged coding sequences (cds) of all of the accessory proteins of a coronavirus genome to construct a virtual protein sequence, and then we analyzed the similarities among all of the virtual protein sequences by multiple sequence alignment. We found that sequence similarities of virtual accessory proteins had

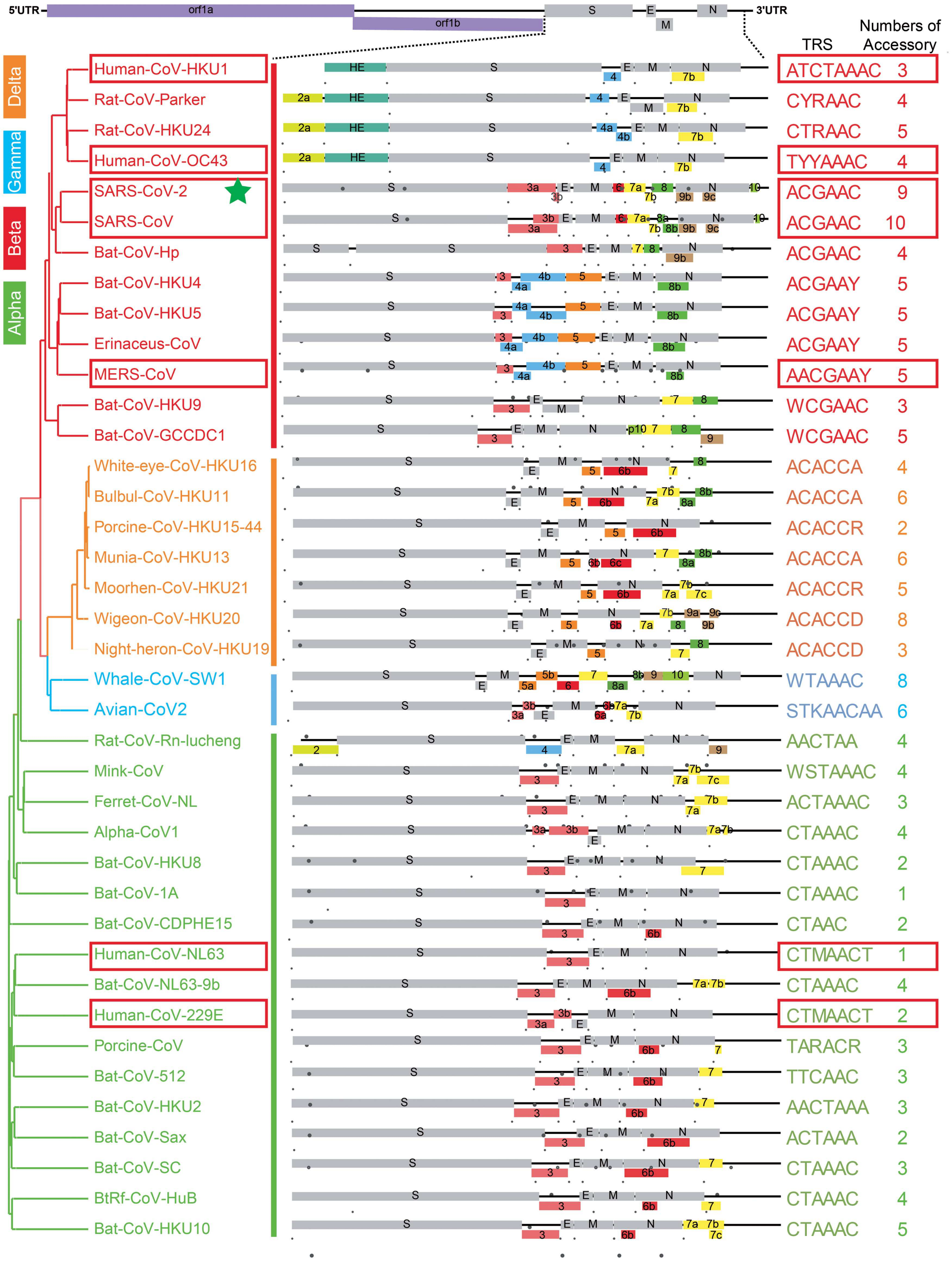


Figure 2. The comprehensive annotation of coronavirus accessory proteins. A phylogenetic tree of 39 representative reference coronaviruses was built based on the pp1b region. Four large branches, alpha, beta, gamma and delta coronaviruses, were clearly grouped as the classification from ICTV. Seven types of human-infecting coronaviruses are highlighted with rectangular boxes, while the recently pendemic SARS-COV-2 is indicated by a green star. For each reference strain, the 3'-terminal of genome encodes structural proteins (S, E, M and N) and accessary proteins are shown as a linear structure. Structural proteins are indicates in gray. The accessory proteins are indicates in different colors according to their names. The TRSs that regulate the discontinuous transcription of sub-genomes for coronaviruses are indicates as black points in the genome structure. The TRS sequence and numbers of accessory proteins are both listed.

significant intra-genus conservation and inter-genus diversity (Figure 4A), which was consistent with the distribution pattern among accessory protein sequences. Furthermore, we calculated the sequence similarities for all of the encoded accessory proteins. The results showed that only a small portion of the accessory proteins had similarity scores ≥40%, while most

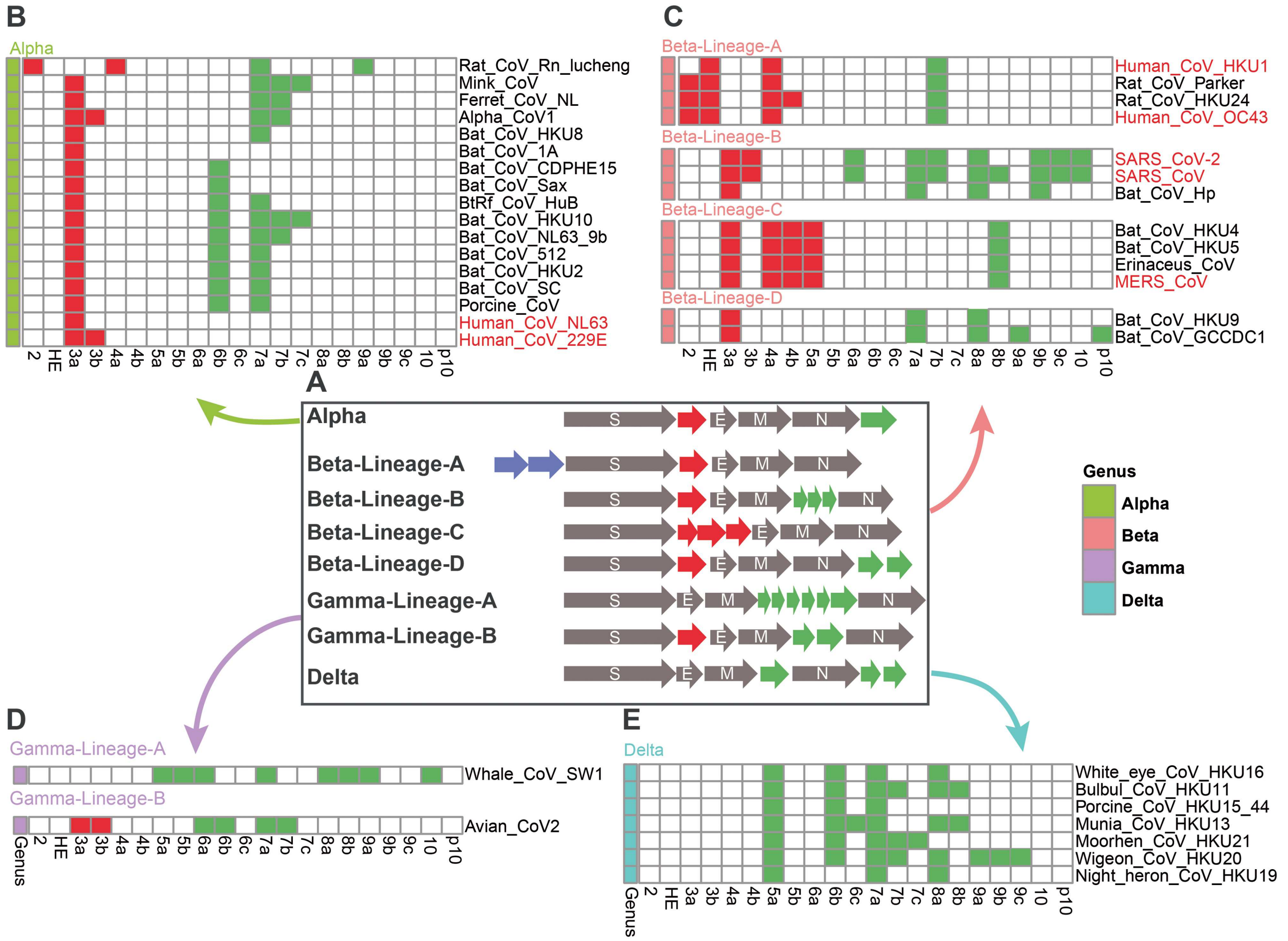


Figure 3. Distribution patterns of coronavirus accessory proteins on genome structure. (A) Eight compositional types of coronavirus accessory proteins are defined based on their genomic locations and compositions. There is no accessory protein between structural proteins E and M. The E-M proteins were used as the border, the accessory proteins in front are displayed in red and proteins behind E-M are green. (B–E) The distributions of accessory proteins for each strain in the alpha- (B), beta- (C), gamma- (D) and delta- (E) genus are shown. Seven human-infecting coronaviruses are highlighted in red.

accessory proteins had no sequence similarities with scores <20% (Supplementary Figure 3A). Accessory protein 3 in alpha-coronaviruses had a relatively higher consistency among different viral species. However, accessory protein 3 in beta-coronaviruses had significant sequence variation, forming three groups of consistent proteins (Supplementary Figure 3B).

By integrating the position distribution and sequence similarity of the accessory proteins, we found that they were relatively conserved before the E-M proteins, while they were more diverse behind the E-M proteins. Therefore, using the E-M proteins as a boundary, the accessory proteins could be divided into two parts: Pre-EM and Post-EM (Figure 4B). The Pre-EM accessory proteins could be distinguished and named according to the distribution pattern and TRS position of their accessory proteins as follows: Alpha-3, Beta-Lineage-A-4 (Beta-A4), Beta-Lineage-B-3 (Beta-B3), Beta-Lineage-C-3 (Beta-C3), Beta-Lineage-C-4a (Beta-C4a), Beta-Lineage-C-4b (Beta-C4b), Beta-Lineage-C-5 (Beta-C5) and Beta-Lineage-D-3 (Beta-D5). The 3b proteins of SARS-CoV and SARS-CoV-2 were found to have low consistency, which may have been because of truncation of the 3b protein of SARS-CoV-2 [22]. The accessory proteins of the two strains in Beta-A4 are more similar to that of Beta-C3 (Supplementary Figure 4A), while the accessory proteins of the two strains of Beta-D3 are more similar to that of Beta-C5. This implies that the Lineage C branch in the beta genus may be derived from Lineage A and Lineage D (Supplementary Figure 4B). Contrary to pre-EM, post-EM accessory proteins present high diversity, with only the Delta-5 and Delta-7 groups having sequence similarities ≥40%.

Conserved accessory proteins in beta-coronaviruses

The distribution and sequence characteristics of coronavirus accessory proteins suggest that there may be functional similarity within the genus. Because the three human-infecting coronaviruses that cause severe symptoms all belong to the beta-genus, we further investigated the characteristics of accessory proteins in beta-coronaviruses. The accessory proteins of beta-genus were found to have internal diversity and lineage conservation (Figure 5A). Three of the four lineages included coronaviruses that can infect humans. In Lineage-A, HKU1 and OC43 can cause mild symptoms in humans, and these contain a conserved accessory protein, Beta-A4 (Figure 5B). Studies have shown that OC43 Beta-A4 antagonizes type I IFN and NF-kb signaling [35].

Notably, Lineage-B includes SARS-CoV and SARS-CoV-2. The conserved protein in the pre-EM region of Lineage-B is Beta-B3, and the sequence similarity score of 3a between SARS-CoV and SARS-CoV-2 was 72% (Figure 5C), while it was only 20% for the

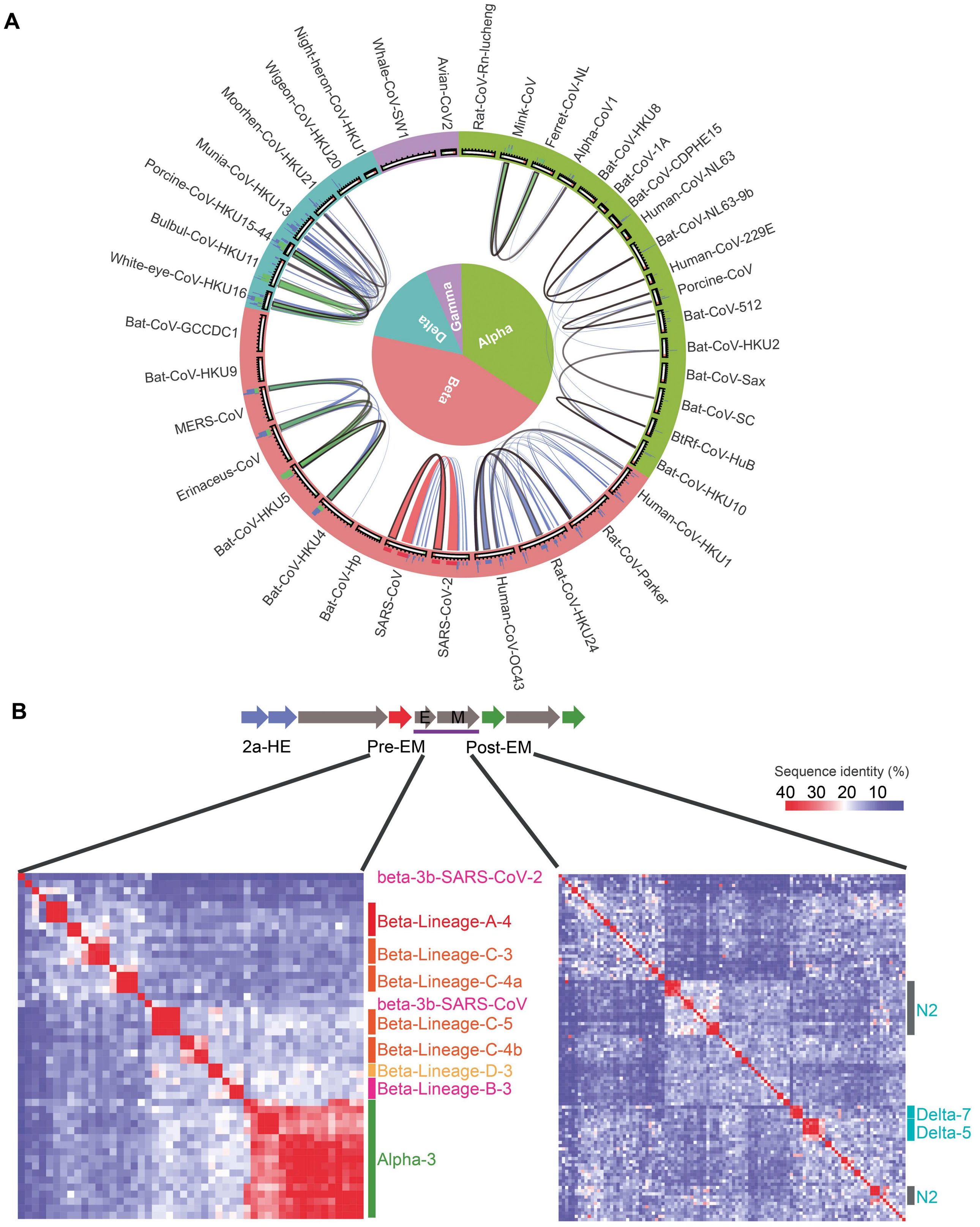
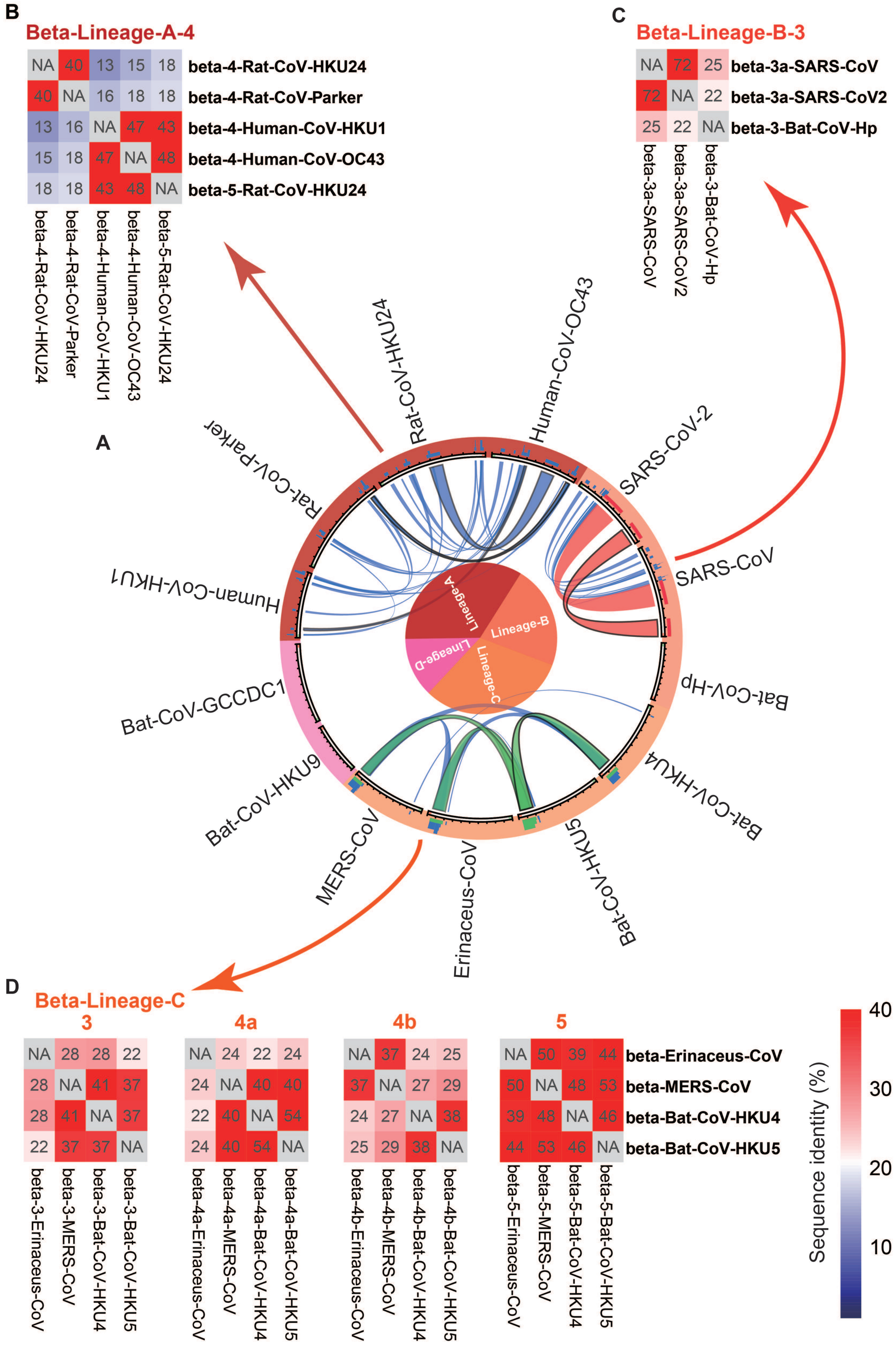


Figure 4. Sequence similarities of accessory proteins among four genera of coronaviruses. (A) Sequence similarities of accessory protein sequences between pairs of strains as determined with the Circos software. The E-value for the BLAST alignment is 1 E-5. Ribbon colors correspond to blast scores. Stains of alpha, beta, delta and gamma-coronaviruses are shown in green, pink, blue and purple, respectively. (B) Identification of conserved accessory proteins before and after E-M proteins. Pairwise alignment was used to measure the similarity between accessory protein sequences. Sequence identities greater than 40% are shown in red in the heatmap.



	beta-4-Rat-CoV-HKU24	beta-4-Rat-CoV-Parker	beta-4-Human-CoV-HKU1	beta-4-Human-CoV-OC43	beta-5-Rat-CoV-HKU24
beta-4-Rat-CoV-HKU24	NA	40	13	15	18
beta-4-Rat-CoV-Parker	40	NA	16	18	18
beta-4-Human-CoV-HKU1	13	16	NA	47	43
beta-4-Human-CoV-OC43	15	18	47	NA	48
beta-5-Rat-CoV-HKU24	18	18	43	48	NA

	beta-3a-SARS-CoV	beta-3a-SARS-CoV2	beta-3-Bat-CoV-Hp
beta-3a-SARS-CoV	NA	72	25
beta-3a-SARS-CoV2	72	NA	22
beta-3-Bat-CoV-Hp	25	22	NA

	beta-3-Erinaceus-CoV	beta-3-MERS-CoV	beta-3-Bat-CoV-HKU4	beta-3-Bat-CoV-HKU5
beta-Erinaceus-CoV	NA	28	28	22
beta-MERS-CoV	28	NA	41	37
beta-Bat-CoV-HKU4	28	41	NA	37
beta-Bat-CoV-HKU5	22	37	37	NA

	beta-4a-Erinaceus-CoV	beta-4a-MERS-CoV	beta-4a-Bat-CoV-HKU4	beta-4a-Bat-CoV-HKU5
beta-Erinaceus-CoV	NA	24	22	24
beta-MERS-CoV	24	NA	40	40
beta-Bat-CoV-HKU4	22	40	NA	54
beta-Bat-CoV-HKU5	24	40	54	NA

	beta-4b-Erinaceus-CoV	beta-4b-MERS-CoV	beta-4b-Bat-CoV-HKU4	beta-4b-Bat-CoV-HKU5
beta-Erinaceus-CoV	NA	37	24	25
beta-MERS-CoV	37	NA	27	29
beta-Bat-CoV-HKU4	24	27	NA	38
beta-Bat-CoV-HKU5	25	29	38	NA

	beta-5-Erinaceus-CoV	beta-5-MERS-CoV	beta-5-Bat-CoV-HKU4	beta-5-Bat-CoV-HKU5
beta-Erinaceus-CoV	NA	50	39	44
beta-MERS-CoV	50	NA	48	53
beta-Bat-CoV-HKU4	39	48	NA	46
beta-Bat-CoV-HKU5	44	53	46	NA

Figure 5. Conserved accessory proteins in beta-coronavirus. (A) Visualization of sequence similarity of accessory protein sequences in beta-coronavirus using the Circos software. The E-value for the blast alignment is 1 E-5. Ribbon colors correspond to BLAST scores. (B–D) Sequence identity of conserved accessory proteins in beta-coronavirus. Pairwise alignment was used to measure the similarity between accessory protein sequences. Sequence identities greater than 40% are shown in red in the heatmap.

Bat-CoV-Hp virus. Previous studies revealed that the 3a protein of SARS-CoV activates NF-Kb, has pro-inflammatory functions and can promote apoptosis and facilitate release of the virus [36–39]. Overexpression of the 3b protein of SARS-CoV in Vero E6 cells leads to apoptosis and necrosis [40], while 7a causes apoptosis by interfering with Bcl-XL [41], and ORF8a enhances virus replication and induces apoptosis through a mitochondrial-dependent pathway [42]. Finally, the 8b protein negatively regulates viral replication and activates the NLRP3 inflammasome [43, 44].

The similarities of the four accessory proteins before the E protein in Lineage-C were close to or higher than 40% among strains, including MERS-CoV (Figure 5D). Studies have shown that MERS-CoV 4a protein suppresses host immune response and antagonizes type I IFN production and NF-Kb activity [45] while also inhibiting the formation of stress granules and promoting viral protein translation [46, 47]. MERS-CoV 4b protein also suppresses the host immune response, preventing IFN-β production and NF-Kb signaling [48, 49].

Comparison of interaction networks of accessory proteins and host between SARS-CoV and SARS-CoV-2

SARS-CoV-2 has been shown to have a similar compositional structure of accessory proteins as SARS-CoV [22]. However, the similarity and differentiation of the functional interactions between the host and accessory proteins of SARS-CoV-2 and SARS-CoV are still unclear. Here, we collected all public virus-host protein–protein interactions of SARS-CoV-2 (229) and SARS-CoV (33) to construct networks of interactions between coronavirus accessory proteins and host proteins (Figure 6A and Supplementary Table 3). We found that they only share four interacting genes: SMOC1, MARK3, DCTN2 and BAG6. In addition, some human interaction genes of the two viruses were found to interact with each other. These genes are involved in multiple pathways of host resistance to viral infections, including apoptotic process, viral life cycle and response to oxidative stress. We found that SARS-CoV-2 and SARS-CoV share the BAG6 gene in the apoptosis signal. In addition, they were found to possess several proteins that can interact each other, BCL2, BCL2L1, ITGB2, PACK1 and HMOX1 (Figure 6B). Furthermore, evaluation of the interaction network of each accessory protein revealed that all of the accessory proteins participate in the host immune response, except for proteins 8a and 10 in SARS-CoV (Supplementary Figures 5 and 6).

Discussion

The key role of coronavirus accessory proteins in virus–host interactions prompted us to systematically study their compositional diversity and evolution pattern. Based on the unique discontinuous transcription of sub-genomes of coronavirus, we developed a standardized coronavirus genome annotation tool named CoroAnnoter. Using CoroAnnoter, we can correct the inaccurate naming of accessory proteins in some previous coronavirus annotation results, such as different naming of the same protein in different studies [7, 50]. Furthermore, we constructed a comprehensive profile for coronavirus accessory proteins and standardized the naming of accessory proteins by integrating ORFs sequence similarity and TRS positions. The normalized dataset enables the subsequent systematic studies of the composition and evolution pattern of different coronavirus accessory proteins.

We found that the compositions of coronavirus accessory proteins have significant intra-genus conservation and inter-genus diversity. We divided all of the representative coronaviruses into eight types based on the composition and location of accessory proteins in the viral genome. As expected, these eight types derived from accessory proteins are consistent with the evolutionary relationship of coronaviruses based on conserved proteins, such as pp1b or pp1ab. The consistent classifications suggest that the evolution of conserved proteins and relatively divergent accessory proteins occurs simultaneously with inherent association. We found that seven human-infecting coronaviruses distributed into four types in genus alpha and beta with significantly different compositions of accessory proteins, indicating the ability to infect humans is not closely associated with the composition of accessory proteins. However, different types of coronaviruses present variable pathogenicity. Types Alpha and Beta-Lineage-A coronaviruses cause mild cold symptoms, while coronaviruses from the Beta-Lineage-B and Beta-Lineage-C types cause potentially serious symptoms, including death. The compositional characteristics of the Alpha and Beta-Lineage-A types indicate that they both have few accessory proteins and simple compositional patterns; however, the accessory proteins of Beta-Lineage-B and Beta-Lineage-C types are more complex. Therefore, the pathogenicity of coronaviruses may be associated with the composition of accessory proteins.

Traditionally, proteins with the same name and genomic position among associated viral strains are considered to be homologous. However, coronavirus accessory proteins have more complex characteristics. Using the structural proteins E and M as the boundary, we found that the accessory proteins in the pre-EM region have intra-genus conservation and inter-genus diversity, while the accessory proteins in the post-EM region are not conserved within genera. The balance of conservation and variation of accessory proteins may play important roles in their functions and viral adaptation. For example, the inhibition of virus replication by conserved Alpha-3 protein may be beneficial to the long-term symbiosis of the virus in the host [7]. Additionally, the diverse accessory proteins of beta-coronaviruses have multiple functions, including antagonizing the host immune response and promoting viral protein translation [36, 39, 46, 47].

The interaction patterns between coronaviruses and their hosts are very important for the investigation of the pathogenic mechanism of the virus. Comparison of the virus–host interaction networks of SARS-CoV and SARS-CoV-2 accessory proteins revealed that they share multiple antiviral signaling pathways. This implies that although the two viruses induce differences in host–virus interaction proteins, these proteins function through similar pathways. Considering the large disease variations manifested by SARS-CoV-2 and SARS-CoV infection, the virus–host interaction mechanisms of the coronavirus accessory proteins need to be further investigated.

The divergent composition and evolutionary pattern of accessory proteins raises the question of their origins. Because the gain or loss of accessory proteins strongly affects the viral phenotypes, it is important to identify the sources of these coronavirus accessory proteins [21, 51, 52]. We identified all of the accessory proteins by BLAST and found a small number of similar sequences that may indicate their possible origins (Supplementary Table 4). Proteins 2a and HE from Beta-lineage-A are homologous with torovirus and influenza virus proteins. HE protein is known to be a glycoprotein that facilitates virus invasion [53]. However, protein 4b in Beta-lineage-C is similar to

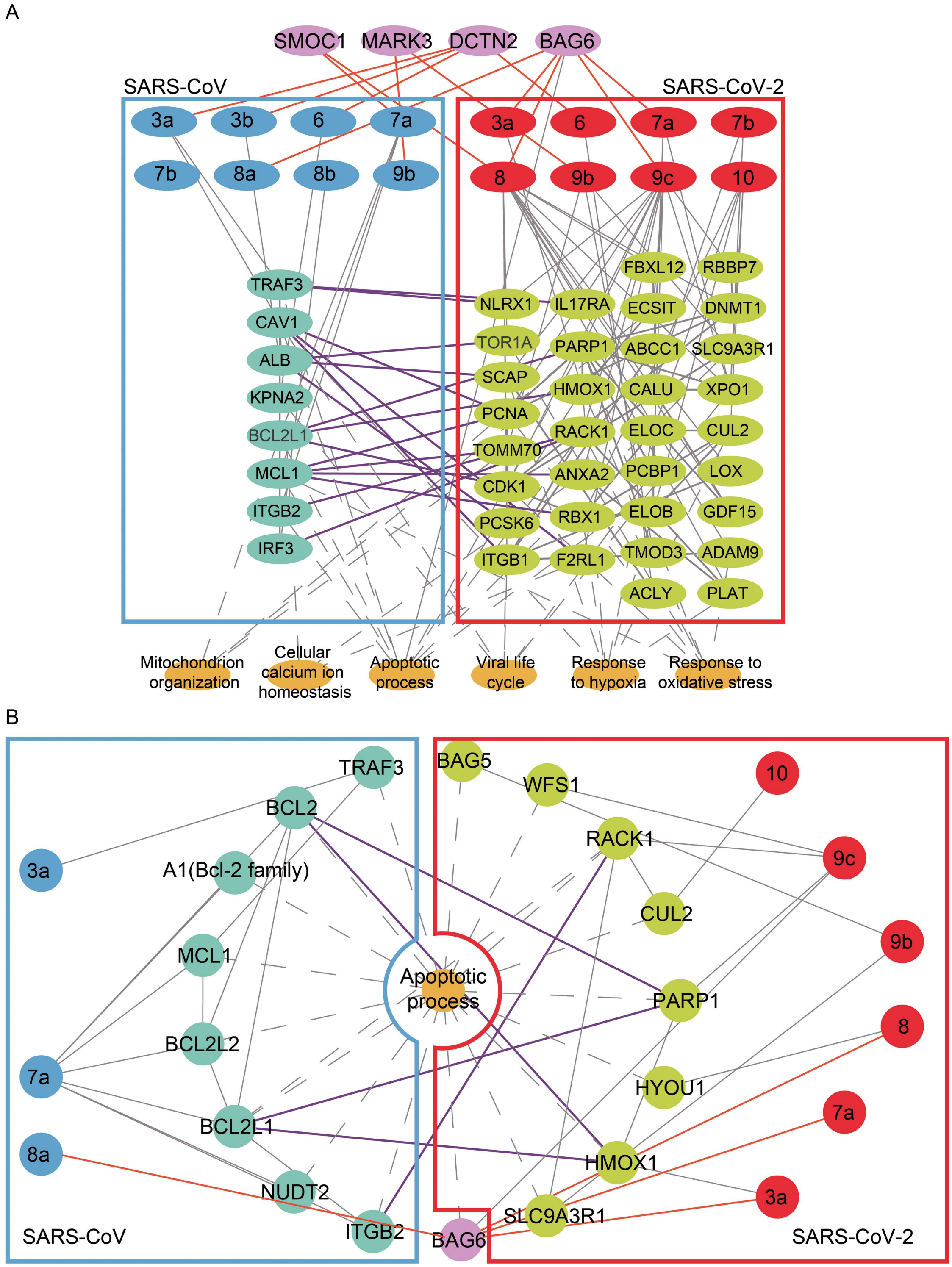


Figure 6. Interaction networks between accessory proteins and host for SARS-CoV and SARS-CoV-2. (A) Comparison of virus-host interaction networks of SARS-CoV and SARS-CoV-2 accessory proteins. Four common genes (SMOC1, MARK3, DCTN2 and BAG6) are shown in purple. Functions annotated via the Gene Ontology and IPA software are listed at the bottom. (B) Sub-networks of SARS-CoV and SARS-CoV-2 involved in the apoptotic process. Human proteins were selected from published papers. Interaction relationships were extracted by the IPA software.

that of an *Escherichia coli* protein with unclear function. Indeed, the sources of most accessory proteins are still unknown, which may be because there are still a large number of unknown sequences.

Three serious outbreaks caused by diverse coronaviruses in recent years indicate that more attention should be taken to some possibly human-susceptible coronaviruses in nature. Investigation in the genomic composition and evolution pattern of known coronaviruses may help with, when new viruses appear, understanding their molecular characteristics and evolutionary origin. We will further optimize the function of CoroAnnoter tool and develop a website for providing online genome annotation and comparison services especially for coronavirus. It may also play a role in promoting the standardized naming for coronavirus accessory proteins.

Key Points

- CoroAnnoter is a semi-automatic and standardized genome annotation tool for coronavirus proteins by combining open reading frame prediction, transcription regulatory sequence recognition and homologous alignment.
- We generated a comprehensive profile for the composition, homology, function and source of all coronavirus accessory proteins.
- The genomic distributions of accessory proteins have significant intra-genus conservation and inter-genus diversity.
- Evolutionary analysis suggested that the accessory proteins are more conservative in pre-EM group while significantly diverse in post-EM group.
- SARS-CoV-2 and SARS-CoV accessory proteins share multiple antiviral signaling pathways.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Authors' Contributions

Conceived and designed the experiments: J.Z.S. and A.P.W. Contributed to the data collection: N.H. and Z.Y.C. Data interpretation and discussion: A.P.W., Y.S.P., L.L., H.Y.Z., C.Y.J. and T.J.J. Wrote the paper: J.Z.S. and A.P.W. Reviewed the paper: A.P.W. and J.M.

Acknowledgements

We thank Peng Chao and Qiming Liang from Shanghai Jiaotong University for providing human protein information that interacts with coronavirus accessory proteins.

Conflict of Interest

No conflicts of interest.

Funding

National Key Plan for Scientific Research and Development of China (2016YFD0500301); CAMS Initiative for Innovative Medicine (CAMS-I2M, 2016-I2M-1-005); National Natural Science Foundation of China (31900472); Central Public-Interest Scientific Institution Basal Research Fund (2016ZX310195, 2017PT31026, 2018PT31016).

References

1. Huang C, Wang Y, Li X, *et al*. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020;**395**:497–506.
2. Zhu N, Zhang D, Wang W, *et al*. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* 2020;**382**:727–33.
3. WHO. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). 2020.
4. Drosten C, Günther S, Preiser W, *et al*. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N Engl J Med* 2003;**348**:1967–76.
5. Ksiazek TG, Erdman D, Goldsmith CS, *et al*. A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med* 2003;**348**:1953–66.
6. Haagmans BL, Al Dhahiry SH, Reusken CB, *et al*. Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation. *Lancet Infect Dis* 2014;**14**:140–5.
7. Cotten M, Watson SJ, Kellam P, *et al*. Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. *Lancet* 2013;**382**:1993–2002.
8. Hussain HY. Incidence and mortality rate of Middle East respiratory syndrome-corona virus (MERS-Cov), threatens and opportunities. *J Mycobac Dis* 2014;**4**:162.
9. Arabi YM, Arifi AA, Balkhy HH, *et al*. Clinical course and outcomes of critically ill patients with Middle East respiratory syndrome coronavirus infection. *Ann Intern Med* 2014;**160**:389–97.
10. Khafaie MA, Rahim F. Cross-country comparison of case fatality rates of COVID-19/SARS-COV-2. *Osong Public Health Res Perspect* 2020;**11**:74.
11. Wu D, Wu T, Liu Q, *et al*. The SARS-CoV-2 outbreak: what we know. *Int J Infect Dis* 2020;**94**:44–48.
12. Majumder M, Mandl KD. Early transmissibility assessment of a novel coronavirus in Wuhan, China, *Available at SSRN* 2020. https://media.ellinikahoaxes.gr/uploads/2020/02/SSRN-id3524675.pdf.
13. Tang B, Wang X, Li Q, *et al*. Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions. *J Clin Med* 2020;**9**:462.
14. van Regenmortel MH, Fauquet CM, Bishop DH, *et al*. *Virus taxonomy: classification and nomenclature of viruses. Seventh report of the International Committee on Taxonomy of Viruses*. Academic Press, 2000.
15. Eckerle LD, Lu X, Sperry SM, *et al*. High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. *J Virol* 2007;**81**:12135–44.
16. Woo PC, Lau SK, Lam CS, *et al*. Discovery of seven novel mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. *J Virol* 2012;**86**:3995–4008.
17. Thiel V, Herold J, Schelle B, *et al*. Infectious RNA transcribed in vitro from a cDNA copy of the human coronavirus genome cloned in vaccinia virus. *J Gen Virol* 2001;**82**:1273–81.
18. St-Jean JR, Jacomy H, Desforges M, *et al*. Human respiratory coronavirus OC43: genetic stability and neuroinvasion. *J Virol* 2004;**78**:8824–34.
19. van der Hoek L, Pyrc K, Jebbink MF, *et al*. Identification of a new human coronavirus. *Nat Med* 2004;**10**:368–73.

20. Woo PC, Lau SK, Chung-ming C, *et al*. Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. *J Virol* 2005;**79**:884–95.
21. Forni D, Cagliani R, Clerici M, *et al*. Molecular evolution of human coronavirus genomes. *Trends Microbiol* 2017;**25**:35–48.
22. Wu A, Peng Y, Huang B, *et al*. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host Microbe* 2020;**27**:325–28.
23. Zuniga S, Sola I, Alonso S, *et al*. Sequence motifs involved in the regulation of discontinuous coronavirus subgenomic RNA synthesis. *J Virol* 2004;**78**:980–94.
24. Liu DX, Fung TS, KK-L C, *et al*. Accessory proteins of SARS-CoV and other coronaviruses. *Antiviral Res* 2014;**109**:97–109.
25. Bailey TL, Boden M, Buske FA, *et al*. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 2009;**37**: W202–8.
26. Pages H, Aboyoun P, Gentleman R, *et al*. Biostrings: string objects representing biological sequences, and matching algorithms. *R package version 2.42* 2016;**1**:10–18129.
27. Darzentas N. Circoletto: visualizing sequence similarity with Circos. *Bioinformatics* 2010;**26**:2620–21.
28. Yu G, Wang L-G, Han Y, *et al*. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012;**16**:284–7.
29. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
30. Lai MM, Cavanagh D. The molecular biology of coronaviruses. *Advances in Virus Research* 1997;**48**:1–100.
31. Sawicki S, Sawicki D. A new model for coronavirus transcription. *Coronaviruses and Arteriviruses* 1998, 215–19.
32. Rombel IT, Sykes KF, Rayner S, *et al*. ORF-FINDER: a vector for high-throughput gene identification. *Gene* 2002;**282**: 33–41.
33. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *UCSD Technical Report* 1994;CS94–351.
34. van Boheemen S, de Graaf M, Lauber C, *et al*. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *MBio* 2012;**3**:e00473–12.
35. Beidas M, Chehadeh W. Effect of human coronavirus OC43 structural and accessory proteins on the transcriptional activation of antiviral response elements. *Intervirology* 2018;**61**:30–5.
36. Waye MM, Law PT, Wong C-H *et al*. The 3a Protein of SARS-coronavirus induces apoptosis in Vero E6 cells. In: *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. 2006, p. 7482–5. IEEE.
37. Freundt EC, Yu L, Goldsmith CS, *et al*. The open reading frame 3a protein of severe acute respiratory syndrome-associated coronavirus promotes membrane rearrangement and cell death. *J Virol* 2010;**84**:1097–109.
38. Siu K-L, Yuen K-S, Castaño-Rodriguez C, *et al*. Severe acute respiratory syndrome coronavirus ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC. *FASEB J* 2019;**33**:8865–77.
39. Lu W, Zheng B-J, Xu K, *et al*. Severe acute respiratory syndrome-associated coronavirus 3a protein forms an ion channel and modulates virus release. *Proc Natl Acad Sci* 2006;**103**:12540–5.
40. Khan S, Fielding BC, Tan TH, *et al*. Over-expression of severe acute respiratory syndrome coronavirus 3b protein induces both apoptosis and necrosis in Vero E6 cells. *Virus Res* 2006;**122**:20–7.
41. Tan Y-X, Tan TH, MJ-R L, *et al*. Induction of apoptosis by the severe acute respiratory syndrome coronavirus 7a protein is dependent on its interaction with the Bcl-XL protein. *J Virol* 2007;**81**:6346–55.
42. Chen C-Y, Ping Y-H, Lee H-C, *et al*. Open reading frame 8a of the human severe acute respiratory syndrome coronavirus not only promotes viral replication but also induces apoptosis. *J Infect Dis* 2007;**196**:405–15.
43. Keng C-T, Åkerström S, Leung CS-W, *et al*. SARS coronavirus 8b reduces viral replication by down-regulating E via an ubiquitin-independent proteasome pathway. *Microbes Infect* 2011;**13**:179–88.
44. Shi C-S, Nabar NR, Huang N-N, *et al*. SARS-coronavirus open reading frame-8b triggers intracellular stress pathways and activates NLRP3 inflammasomes. *Cell Death Dis* 2019;**5**:1–12.
45. Niemeyer D, Zillinger T, Muth D, *et al*. Middle East respiratory syndrome coronavirus accessory protein 4a is a type I interferon antagonist. *J Virol* 2013;**87**:12489–95.
46. Rabouw HH, Langereis MA, Knaap RC, *et al*. Middle East respiratory coronavirus accessory protein 4a inhibits PKR-mediated antiviral stress responses. *PLoS Pathog* 2016; **12**:e1005982.
47. Nakagawa K, Narayanan K, Wada M, *et al*. Inhibition of stress granule formation by Middle East respiratory syndrome coronavirus 4a accessory protein facilitates viral translation, leading to efficient virus replication. *J Virol* 2018;**92**: e00902–18.
48. Canton J, Fehr AR, Fernandez-Delgado R, *et al*. MERS-CoV 4b protein interferes with the NF-κB-dependent innate immune response during infection. *PLoS Pathog* 2018;**14**:e1006838.
49. Yang Y, Ye F, Zhu N, *et al*. Middle East respiratory syndrome coronavirus ORF4b protein inhibits type I interferon production through both cytoplasmic and nuclear targets. *Sci Rep* 2015;**5**:17554.
50. Woo PC, Wang M, Lau SK, *et al*. Comparative analysis of twelve genomes of three novel group 2c and group 2d coronaviruses reveals unique group and subgroup features. *J Virol* 2007;**81**:1574–85.
51. Chen L, Li F. Structural analysis of the evolutionary origins of influenza virus hemagglutinin and other viral lectins. *J Virol* 2013;**87**:4118–20.
52. Crossley BM, Mock RE, Callison SA, *et al*. Identification and characterization of a novel alpaca respiratory coronavirus most closely related to the human coronavirus 229E. *Viruses* 2012;**4**:3689–700.
53. Zeng Q, Langereis MA, van Vliet AL, *et al*. Structure of coronavirus hemagglutinin-esterase offers insight into corona and influenza virus evolution. *Proc Natl Acad Sci* 2008;**105**:9065–9.