Nonstructural proteins NS7b and NS8 are likely phylogenetically associated with evolution of 2019-nCoV

Muhamad Fahmi[a]†, Yukihiko Kubota[b]†, and Masahiro Ito[a,b,*,†]

[a]Advanced Life Sciences Program, Graduate School of Life Sciences, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan

[b]Department of Bioinformatics, College of Life Sciences, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan

† These authors contributed equally to this work.

[*]Corresponding Author: Masahiro Ito; maito@sk.ritsumei.ac.jp; 1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan

[1]

Abstract

Background: The seventh novel human infecting *Betacoronavirus* that causes pneumonia (2019 novel coronavirus, 2019-nCoV) originated in Wuhan, China. The evolutionary relationship between 2019-nCoV and the other human respiratory illness-causing coronavirus is not closely related. We sought to characterize the relationship of the translated proteins of 2019-nCoV with other species of *Orthocoronavirinae*.

Methods: A phylogenetic tree was constructed from the genome sequences. A cluster tree was developed from the profiles retrieved from the presence and absence of homologs of ten 2019-nCoV proteins. The combined data were used to characterize the relationship of the translated proteins of 2019-nCoV to other species of *Orthocoronavirinae*.

Results: Our analysis reliably suggests that 2019-nCoV is most closely related to BatCoV RaTG13 and belongs to subgenus *Sarbecovirus* of *Betacoronavirus*, together with SARS coronavirus and Bat-SARS-like coronavirus. The phylogenetic profiling cluster of homolog

[1] **Abbreviations:** 2019-nCoV, 2019 novel coronavirus; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; HCoV-NL63, human coronavirus NL63; HCoV-229E, human coronavirus 229E; HCoV-OC43, human coronavirus OC43; ECoV-HKU1, human coronavirus HKU1; MERS-CoV, Middle East respiratory syndrome coronavirus; NS, Nonstructural protein; ORF, open reading frame; NF-κB, nuclear factor kappa B

proteins of one annotated 2019-nCoV protein against other genome sequences revealed two clades of ten 2019-nCoV proteins. Clade 1 consisted of a group of conserved proteins in *Orthocoronavirinae* comprising Orf1ab polyprotein, Nucleocapsid protein, Spike glycoprotein, and Membrane protein. Clade 2 comprised six proteins exclusive to *Sarbecovirus* and *Hibecovirus*. Two of six Clade 2 nonstructural proteins, NS7b and NS8, were exclusively conserved among 2019-nCoV, BetaCoV_RaTG, and BatSARS-like Cov. NS7b and NS8 have previously been shown to affect immune response signaling in the SARS-CoV experimental model. Thus, we speculate that knowledge of the functional changes in the NS7b and NS8 proteins during evolution may provide important information to explore the human infective property of 2019-nCoV.

Keywords: 2019-nCoV, novel protein, phylogenetic tree, phylogenetic profile

1. Introduction

In December 2019, the seventh human coronavirus, termed 2019 novel coronavirus (2019-nCoV) or severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was found in Wuhan, China. As of February 8, 2020, the total number of infections and deaths due to 2019-nCoV globally was 34,439 and 720, respectively, according to the Johns Hopkins University Center for Systems Science and Engineering.

Coronaviruses are enveloped RNA viruses that infect many species, including humans, other mammals, and birds. After infection, the host may develop respiratory, bowel, liver, and neurological diseases (Weiss and Leibowitz, 2011; Cui et al., 2019). Coronaviruses are members of the order Nidovirales and subfamily *Orthocoronavirinae*. This subfamily is divided into four genera: *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus*, and *Deltacoronavirus*. Generally, *Alphacoronavirus* and *Betacoronavirus* tend to infect mammals, while *Gammacoronavirus* and *Deltacoronavirus* typically infect birds. However, some *Gammacoronavirus* and *Deltacoronavirus* can infect mammals under specific conditions (Woo et al., 2012).

In immunocompromised individuals, infection with one of the four human coronaviruses—human coronavirus NL63 (HCoV-NL63), human coronavirus 229E (HCoV-229E), human coronavirus OC43 (HCoV-OC43), or human coronavirus HKU1 (ECoV-HKU1)—usually results in cold-like symptoms. These viruses can cause severe infections in some infants and the elderly. SARS-CoV and Middle East respiratory syndrome coronavirus (MERS-CoV) are

zoonotic coronaviruses that can be associated with fatal diseases (Su et al., 2016; Forni et al., 2017; Cui et al., 2019). Due to the frequent interaction between wild animals and humans, wild animals are a common source of human zoonotic infections. SARS-CoV and MERS-CoV belong to *Betacoronavirus*. Both are zoonotic agents that can cause severe respiratory diseases in humans (Luk et al., 2019; Ramadan and Shaib, 2019). 2019-nCoV is the seventh coronavirus discovered that infects humans. It causes acute respiratory disease in respiratory infections. Immediately after its discovery, the complete genome sequence of 2019-nCoV was determined. The sequence (MN908947) was released by GenBank on 05 January 2020 (Lu et al., 2020). The sequence of 2019-nCoV is 96% identical, at the whole-genome level, to a bat coronavirus (Zhou et al., 2020).

The genomic characteristics and epidemiology of 2019-nCoV have been analyzed (Lu et al., 2020). Nine inpatient culture isolates were subjected to next-generation sequencing, and individual complete and partial 2019-nCoV genomic sequences were obtained. Phylogenetic analysis of these 2019-nCoV genomes and other coronaviruses was performed to determine the evolutionary history of the virus and to explore the origin of 2019-nCoV. At the first onset, homology modeling investigated the potential receptor-binding properties of the virus. However, SARS-CoV and MERS-CoV showed approximate similarities of 79% and 50% with 2019-nCoV, respectively. These findings indicated that there is not a close evolutionary relationship of 2019-nCoV with SARS-CoV and MERS-CoV. Thus, 2019-nCoV is considered the seventh novel human *Betacoronavirus* (Lu et al., 2020).

In this study, we comprehensively characterized the relationship of the translated proteins of 2019-nCoV to other species of *Orthocoronavirinae*. This was done using a combination of the phylogenetic tree constructed from the genome sequences and the cluster tree developed from the profiles retrieved from the presence and absence of homologs of ten 2019-nCoV proteins.

2. Methods

The genomes and the combination of genome and protein sequences were used to develop a phylogenetic tree and for phylogenetic profiling, respectively. The dataset of the genomes of the *Orthocoronavirinae* subfamily was collected from the Refseq database using the *Orthocoronavirinae* NCBI taxonomy ID (txid2501931). This dataset contains representative complete genomes from each species of the subfamily (Pruitt et al., 2007; Federhen, 2012) (Supplementary Table 1). Additionally, we collected genome sequences from Bat SARS-like coronavirus (MG772934 and MG772933) from NCBI and BetaCoV/bat/Yunnan/RaTG13/2013 (EPI_ISL_402131) from GISAID (http://www.GISAID.org). One species of the *Okanivirinae*

subfamily, the yellow head virus, was also collected as an outgroup (Supplementary Table 1). The genome sequences were aligned using the MAFFT multiple sequence alignment program provided at the XSEDE portal in the CIPRES Science Gateway with an automatic selection strategy (Miller et al., 2012; Katoh and Standley, 2013). A phylogenetic tree was constructed using the maximum likelihood method with RAxML-HPC BlackBox in the CIPRES Science Gateway (Stamatakis, 2006). The analysis used an automatic bootstrapping option with a general time-reversible substitution model with a gamma-shape parameter (GTR+ G). This was selected as the best-fit model under the Akaike information criterion using ModelTest-NG (Darriba et al., 2020). Phylogenetic trees were viewed using FigTree v1.4 (http://tree.bio.ed.ac.uk/software/figtree/).

The annotated protein sequences of 2019-nCoV were collected from the data of one representative genome from NCBI (MN996527). We built a BLAST database with the retrieved genome sequence data using BLAST+ version 2.2.30 (Camacho et al., 2009). We then determined the presence and absence of homolog proteins of one representative set of annotated 2019-nCoV proteins against other genome sequences in a database using tblastn with a threshold of >50 and >25 bits score for protein sequences >50 amino acids (aa) and <50 aa in length, respectively. The results of the presence and absence of homolog proteins were converted into a binary matrix and used to build a clustering tree using ward hierarchical clustering method (Ward, 1963). Nonstructural protein (NS) 7b and NS8 local alignments were only positive in the *Sarbecovirus* subgenus sample, excluding the SARS coronavirus. Additionally, we also predicted the structural properties of the 2019-nCoV NS7b protein, including the secondary structure and order-disorder propensity, using Jpred4 and DICHOT, respectively (Fukuchi et al., 2014; Drozdetskiy et al., 2015). We also predicted the structure using the contact assisted protein structure prediction (C-I TASSER) composite approach (Zhang et al., 2018). Additionally, we specifically collected the sequences that produced significant alignments of NS7b using the MEGA X software (Kumar et al., 2018).

3. Results and Discussion

3.1 Phylogenetic tree

The phylogenetic analysis using complete genome sequences showed that 2019-nCoV was the most closely related to BatCoV RaTG13 and belonged to the *Sarbecovirus* subgenus of *Betacoronavirus*, together with SARS coronavirus and Bat-SARS-like coronavirus with the full support of reliability (BAT-SL-CoVZXC21 and BAT-SL-CoVZC45) (Figure 1). Additionally, *Hibecovirus* with Bat Hp-betacoronavirus/Zhejiang2013, as the representative species, was the

Peer-reviewed version available at *Infect. Genet. Evol.* **2020**; doi:10.1016/j.meegid.2020.104272

most closely related subgenus of *Betacoronavirus* to *Sarbecovirus* as compared to other subgenera, including *Merbecovirus* (under which MERS-CoV has been classified), *Nobecovirus*, and *Embecovirus*. These findings agree with previous phylogenetic tree and similarity plot data (Paraskevis et al., 2020). 2019-nCoV was found to be more closely related to the bat-infecting Sarbecovirus species, Bat SARS-like coronavirus, and BetaCoV RaTG13 than to the SARS coronavirus that infects humans. This indicated that 2019-nCoV more likely originated from bats. However, the Wuhan outbreak was first detected in December, which is a time of year when most bat species hibernate. Moreover, the Huanan seafood market, which is considered as ground zero of the outbreak, does not sell bats. Instead, it has been suggested that there is an animal mediator for virus transmission from bats to humans, similar to the previous cases of SARS-CoV and MERS-CoV, wherein the masked palm civet (*Paguma larvata*) and dromedary camel (*Camelus dromedarius*) act as intermediate hosts, respectively (Lu et al., 2020). Although coronaviruses can exchange genetic material during coinfection, a recent report described the lack of a mosaic relationship of 2019-nCoV to the closely related *Sarbecovirus*, and the lack of a recombination event in the emergence of 2019-nCoV (Paraskevis et al., 2020). 2019-nCoV likely emerged from the accumulation of mutations responding to altered selective pressures or from the infidelity of RNA polymerase perpetuated as replication-neutral mutations. These speculations need to be studied further.

3.2 Phylogenetic profiling

A previously reported comprehensive similarity plot revealed notable mutational hotspots and conserved regions of the genome nucleotide positions of 2019-nCoV against closely related coronaviruses (Lu et al., 2020; Paraskevis et al., 2020). The present findings provide a different perspective of the similarity among *Orthocoronavirinae* species, using data from a cluster tree developed from the profiles retrieved from the presence and absence of homologs of ten 2019-nCoV proteins. This cluster was combined with the cladogram of a previously constructed phylogenetic tree (Figure 2). Both the trees were consistent in their heatmap distributions. The tree of 2019-nCoV proteins comprised two clades. The first, indicated by the blue bar in Figure 2, contained a group of conserved proteins in most *Orthocoronavirinae* species. These comprised Orf1ab polyprotein, Nucleocapsid protein, Spike glycoprotein, and Membrane protein. Spike and Orf1a regions of 2019-nCoV were previously shown to have the lowest sequence identity as compared to the closely related coronavirus species (Lu et al., 2020; Paraskevis et al., 2020). However, since the translated Spike glycoprotein and Orf1ab polyprotein from these regions are very long, the sequence similarity is still sufficient to classify them as homologs. In contrast, another clade, indicated by the green bar in Figure 2, comprised

proteins specific to *Sarbecovirus* for all proteins in this clade and *Hibecovirus* for the envelope protein only. This clade included proteins that were not completely conserved by all *Orthocoronavirus*. Two (NS7b and NS8) of the five nonstructural proteins were specific for 2019-nCoV and its closely related species, BatCoV RaTG13 and Bat-SARS-like coronavirus (BAT-SL-CoVZXC21 and BAT-SL-CoVZC45). The other three nonstructural proteins (NS3, NS6, and NS7a) were also detected in the SARS coronavirus. Based on these results, we propose that the comprehensive analysis of nonstructural proteins, especially NS7b and NS8, may provide new insights into the properties of 2019-nCoV.

As shown in Figure 2, NS7b and NS8 of 2019-nCoV, BatCoV RaTG13, and Bat-SARS-like coronavirus were distinct from other species of *Orthocoronavirus*. NS7b is an integral protein localized in the Golgi compartment. The protein is packaged into SARS-CoV particles (Schaecher et al., 2007). Interestingly, open reading frame (ORF) 7b, but not ORF 7b deletion, induces interferon (IFN)-dependent reporter gene expression as well as apoptosis and the type I IFN response (Pfefferle et al., 2009). Moreover, the deletion of ORF 7b can enhance virus growth (Pfefferle et al., 2009). Thus, we speculate that the property of the non-conserved NS7b in 2019-nCoV may affect the human infective property of the virus. Similarly, the existence of 29 nucleotide deletions in ORF 8b has been described in SARS-CoV (Oostra et al., 2007). A study involving MERS-CoV described that ORF 8b strongly antagonizes the INF-beta (β) promoter and ORF4b and 8b significantly suppress IFN induction (Lee et al., 2019b). Accessory proteins 8b and 8ab of SARS-CoV can suppress the IFN-β signaling pathway (and thus interferon production) by their participation in the ubiquitin-mediated rapid degradation of IFN regulatory factor 3 (Wong et al., 2018). In contrast, when we focused on MERS-CoV from bats and camels, ORF 8b antagonized melanoma differentiation-associated protein 5-mediated nuclear factor kappa B (NF-κB) activation. ORF 8b strongly inhibited TANK-binding kinase 1-mediated induction of NF-κB signaling, but not IκB kinase epsilon and interferon regulatory factor 3-mediated activations (Lee et al., 2019a). Thus, we speculate that the properties of the accessory proteins, NS7b and NS8, in 2019-nCoV may affect its ability to infect humans. Further studies are required to confirm this speculation.

NS7b is a short peptide of 43 residues. A three-dimensional structure is often difficult to obtain from such a short peptide. We predicted the three-dimensional structure of the queried NS7b amino acid sequence using DICHOT and C-I-TASSER (Supplementary Figure 1 and Figure 3) (Fukuchi et al., 2014; Zhang et al., 2018); a protein family (PF11395) was found, but no known three-dimensional structure was found. The secondary structure of this query was also predicted using Jpred4 (Supplementary Figure 2) (Drozdetskiy et al., 2015). The secondary structure was

predicted to be an α-helix. These predictions suggested that the sequence itself had the property of forming an α-helix, but that this very likely does not occur depending on the environment (Figure 3). The alignment of this protein revealed three polymorphism sites between 2019-nCoV, BatCoV RaTG13, and Bat-SARS-like coronavirus sequences (BAT-SL-CoVZXC21 and BAT-SL-CoVZC45) (Supplementary Figure 3).

In summary, some nonstructural proteins were conserved and others were not conserved between 2019-nCoV and SARS-CoV. By focusing on the 2019-nCoV-specific proteins, NS7b and NS8, we proposed a combination of phylogenetic profiling analysis and structural characterization of the genes that were specifically expressed in 2019-nCoV and the closely related bat coronavirus. The data provide insight for further characterization of the infective properties of this virus.

Funding

This work was supported by the MEXT-Supported Program for the Strategic Research Foundation at Private Universities [grant number S1511028 to T.I] and the Takeda Science Foundation.

Acknowledgments

We thank Dr. Motonori Ota, Dr. Satoshi Fukuchi, Dr. Kota Kasahara, and Dr. Takeshi Kikuchi for their support and helpful comments.

Conflicts of Interest

The authors declare no competing interests.

References

Camacho, C., Coulouris, G., Avagyan, V., et al., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10(1), 421.

Cui, J., Li, F., Shi, Z.L., 2019. Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. 17, 181–192.

Darriba, D., Posada, D., Kozlov, A.M., et al., 2020. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. Mol. Biol. Evol. 37(1), 291–294.

Drozdetskiy, A., Cole, C., Procter, J., et al., 2015. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 43, W389–W394.

Federhen, S., 2012. The NCBI taxonomy database. Nucleic Acids Res. 40(D1), D136–D143.

Forni, D., Cagliani, R., Clerici, M., et al., 2017. Molecular evolution of human coronavirus genomes. Trends Microbiol. 25, 35–48.

Fukuchi, S., Amemiya, T., Sakamoto, S., et al., 2014. IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. Nucleic Acids Res. 42, D320–D325.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30(4), 772–780.

Kumar, S., Stecher, G., Li, M., et al., 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35(6), 1547–1549.

Lee, J.Y., Bae, S., Myoung, J., 2019a. Middle East respiratory syndrome coronavirus-encoded accessory proteins impair MDA5-and TBK1-mediated activation of NF-κB. J. Microbiol. Biotechnol. 29(8), 1316–1323.

Lee, J.Y., Bae, S., Myoung, J., 2019b. Middle East respiratory syndrome coronavirus-encoded ORF8b strongly antagonizes IFN-β promoter activation: its implication for vaccine design. J. Microbiol. 57(9), 803–811.

Lu, R., Zhao, X., Li, J., et al., 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. S0140-6736(20), 30251–30258. doi: 10.1016/S0140-6736(20)30251-8.

Luk, H.K.H., Li, X., Fung, J., et al., 2019. Molecular epidemiology, evolution, and phylogeny of SARS coronavirus. Infect. Genet. Evol. 71, 21–30.

Miller, M.A., Pfeiffer, W., Schwartz, T., 2012. The CIPRES science gateway: enabling high-impact science for phylogenetics researchers with limited resources. In Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the extreme to the campus and beyond, pp. 1–8.

Oostra, M., de Haan, C.A., Rottier, P.J., 2007. The 29-nucleotide deletion present in human but not in animal severe acute respiratory syndrome coronaviruses disrupts the functional expression of open reading frame 8. J. Virol. 81(24), 13876–13888.

Paraskevis, D., Kostaki, E.G., Magiorkinis, G., et al., 2020. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. Infect. Genet. Evol. 104212.

Pfefferle, S., Krähling, V., Ditt, V., et al., 2009. Reverse genetic characterization of the natural genomic deletion in SARS-Coronavirus strain Frankfurt-1 open reading frame 7b reveals an attenuating function of the 7b protein in-vitro and in-vivo. Virol. J. 6(1), 131.

Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35(suppl_1), D61–D65.

Ramadan, N., Shaib, H., 2019. Middle East respiratory syndrome coronavirus (MERS-CoV): A review. Germs 9, 35–42.

Schaecher, S.R., Mackenzie, J.M., Pekosz, A., 2007. The ORF7b protein of severe acute respiratory syndrome coronavirus (SARS-CoV) is expressed in virus-infected cells and incorporated into SARS-CoV particles. J. Virol. 81(2), 718–731.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22(21), 2688–2690.

Su, S. et al., 2016. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. 24, 490–502.

Wang, N., Li, S.Y., Yang, X.L., et al., 2018. Serological evidence of Bat SARS-related coronavirus infection in humans, China. Virol. Sin. 33, 104–107.

Ward Jr, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58(301), 236–244.

Weiss, S.R., Leibowitz, J.L., 2011. Coronavirus pathogenesis. Adv. Virus Res. 81, 85–164.

Wong, H.H., Fung, T.S., Fang, S., et al., 2018. Accessory proteins 8b and 8ab of severe acute respiratory syndrome coronavirus suppress the interferon signaling pathway by mediating ubiquitin-dependent rapid degradation of interferon regulatory factor 3. Virology 515, 165–175.

Woo, P. C. et al., 2012. 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. 86, 3995–4008.

Zhang, C., Mortuza, S.M., He, B., Wang, Y., Zhang, Y., 2018. Template-based and free modeling of I‐TASSER and QUARK pipelines using predicted contact maps in CASP12. Proteins 86, 136-151.

Zhou, P., Yang, X.L., Wang, X.G., et al., 2020. A pneumonia outbreak associated with a new coronavirus of probable vat origin. Nature. doi: 10.1038/s41586-020-2012-7.

Figure legends

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2020 doi:10.20944/preprints202002.0310.v1

Peer-reviewed version available at *Infect. Genet. Evol.* **2020**; doi:10.1016/j.meegid.2020.104272

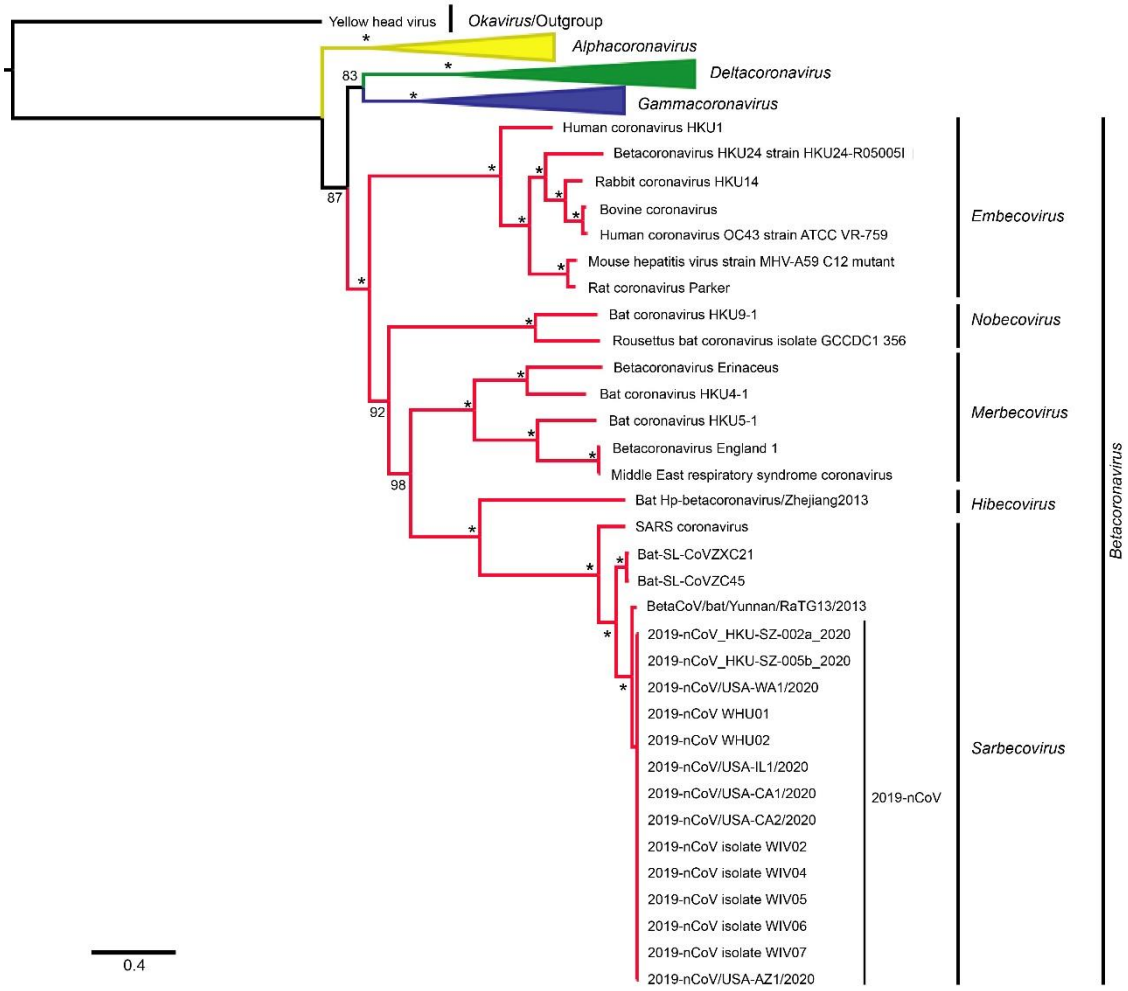

Figure 1. Phylogenetic tree of representative species from the *Orthocoronavirinae* subfamily with *Okavirus* as an outgroup, constructed using the maximum likelihood method. The asterisk indicates the fully supported reliability of maximum likelihood.

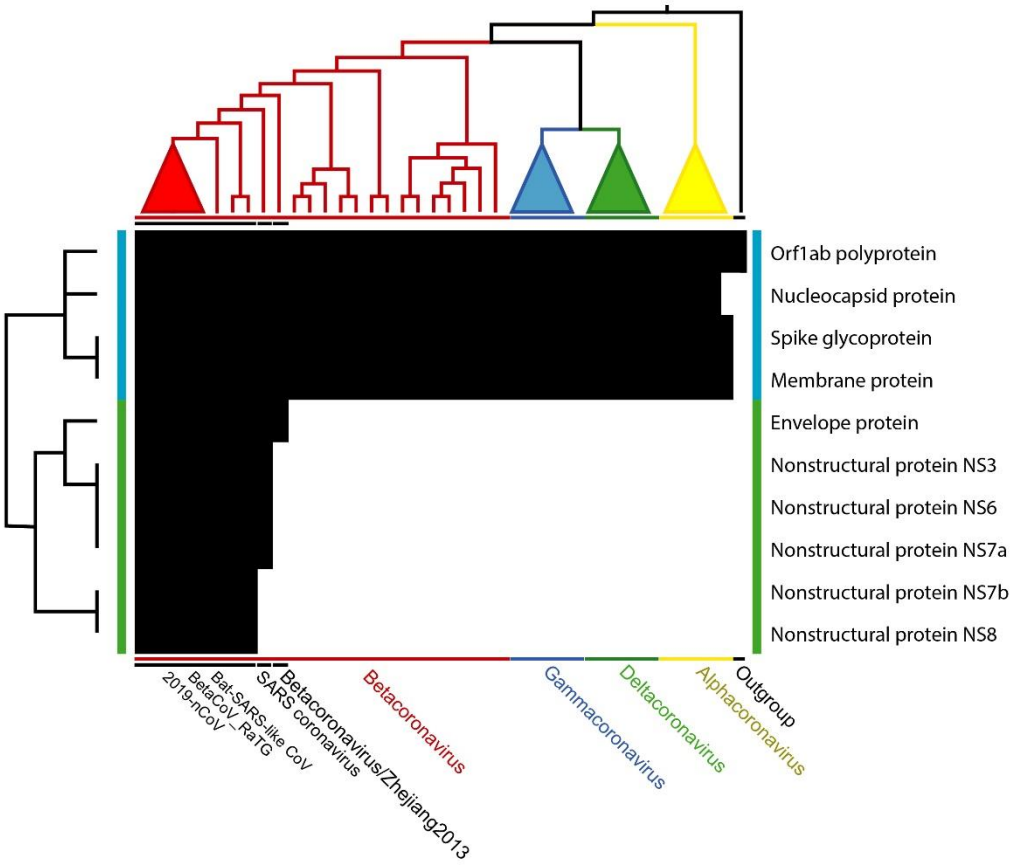

Figure 2. Phylogenetic profiling of the annotated 2019-nCoV proteins combined with the maximum likelihood cladogram of the *Orthocoronavirinae* subfamily with *Okavirus* as the outgroup. The heatmap indicates the binary matrix of the homolog proteins of 2019-nCoV against other species in the dataset, with black and white colors as presence and absence, respectively. The bit pattern was arranged following the vertical and horizontal trees. The vertical tree is a phylogenetic profiling tree constructed from a binary matrix of the presence and absence of homolog proteins. It has two clades, indicated by blue and green bars. The horizontal tree is the cladogram of the maximum likelihood tree, as shown in Figure 1, with a collapsed clade of 2019-nCoV.

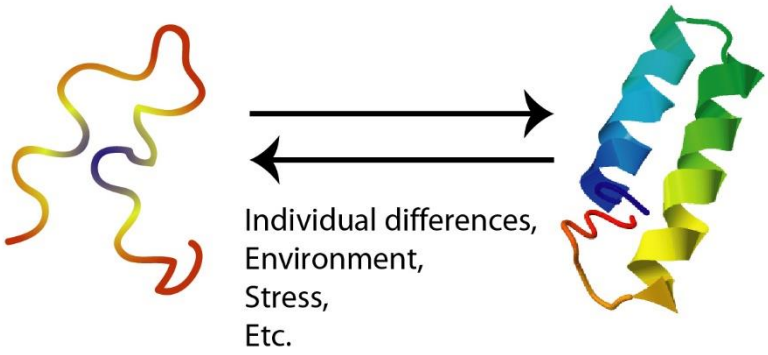

Figure 3. The model of nonstructural–structural transition of 2019-nCoV nonstructural protein 7b. The predicted protein structure of 2019-nCoV nonstructural protein 7b using C-I TASSER is shown as the helix structure protein.