Research article

**In silico** comparative analysis of SARS-CoV-2 nucleocapsid (N) protein using bioinformatics tools

Mehmet Emin Uras

1 Marmara University, Faculty of Science & Arts, Department of Biology, 34722, Goztepe, Istanbul, Turkey

**Abstract**

The world has been encountered to one of the biggest pandemics that causing by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). SARS-CoV-2 is placed in the Beta-CoV genus in the Coronaviridae family. N protein is one of the crucial structural proteins of SARS-CoV-2 that binds to the genome thereby generating helical ribonucleoprotein core. It is involved in viral transcription/replication, translation, and viral assembly after entering the host cell through interacting with host proteins. N protein sequences of SARS-CoV-2 and taxonomically related CoVs are examined using bioinformatics tools and approaches including sequence alignment, sequence and phylogenetic analyzes, and predicting of putative N-Glycosylation and phosphorylation positions and also predictions and comparative analyzes are performed on 3D structures of N proteins from SARS-CoV-2 related CoVs through using of some of applied bioinformatics analyzes. Results of mega BLAST search revealed that the most similar N protein sequence to SARS-CoV-2 is Bat-CoV RaTG13 N protein sequence in the taxonomically related CoVs. SARS-CoV-2 is grouped with SARS, pangolin, civet and bat CoVs (RATG13, SL ZC45 and SL ZXC21) in N protein, nucleotide and protein based ML phylogenetic trees. Some of SARS-CoV-2 N proteins were showed divergence from other SARS-CoV-2 N proteins analyzed due to amino acid substitutions detected in SARS-CoV-2 N proteins samples in phylogenetic trees. The highest amino acid substitutions were detected in Richmont/USA (QJA42209.1) and Greece (QIZ16579.1) samples, with 2 and 3 place substitutions, respectively. By domain analyzes, three domains were detected as Corona_nucleocora (Pfam), N terminal CoV RNA-binding domain (HAMAP) and C terminal N protein dimerization domain (HAMAP). Possible N-glycosylation positions of SARS-CoV-2 N protein were predicted at two positions. Assessments of possible serine, threonine and tyrosine phosphorylations were found to be at 100 positions, 34 of them were higher than 80% possibility. 3D structure analysis based on TM scores revealed that although the results of 3D structure analysis were shown consistency with the taxonomy of the CoVs, the 3D structures of SARS-CoV-2 N protein and taxonomically related CoVs were not at the same fold.

**Keywords:** 3D structure; bioinformatics; coronavirus; COVID-19; SARS-CoV-2; viral proteins

1. Introduction

By the end of 2019, the world has been encountered to one of the biggest pandemics that caused by severe acute respiratory syndrome coronavirus 2 (2019-nCoV or SARS-CoV-2). In December 2019, WHO authorities were informed by Chinese authorities for a new pneumonia infection, mainly resembling viral pneumonia, in Wuhan/China (Gene, 2020; Wu et al., 2020). After the first examinations, the cause of the infection was diagnosed as a novel CoV (SARS-CoV-2); thereafter, named as COVID-19. Meanwhile 282 cases and 6 deaths on January 21, 2020, and following 3,349,786 cases and 238,628 deaths on May 3, 2020 and 79,062,802 cases and 1,751,311 deaths on December 27, 2020 were reported by WHO in all around the world (WHO, 2020b, 2020a, 2020c).

SARS-CoV-2 is a member of Beta-CoV genera from
family Coronaviridae, like severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV) (Ye et al., 2020). Genome structure of SARS-CoV and MERS-CoV were extensively studied after the pandemics occurred during the periods of 2002-2003 and 2012-2015, respectively (De Wit, et al., 2016). The first genome sequence of SARS-CoV-2, sampled from a patient from Wuhan/China, were submitted to NCBI GenBank at the beginning of January 2020 with accession number MN908947 (Wu et al., 2020).

The genome analyzes of the Beta-CoVs revealed that they have a large single-stranded, approximately ~30 kb-bp long positive-sense RNA genome (+ssRNA). The strand has a poly-A tail at 3’ and cap at 5’. The two thirds of the genome encodes a non-structural protein named 1a/1b (ORF1a/b) polyprotein having a role in construction of replication/transcription complex (Chen, et al., 2020). The remaining part of genome encodes structural proteins including spike glycoprotein (S protein), envelope glycoprotein (E protein), and membrane (M protein) and nucleocapsid (N protein) proteins in a 5’ to 3’ order. Additionally, some proteins are also encoded by the structural proteins which are involved in immune response of host (De Wit, et al., 2016).

The S protein of CoVs is responsible for recognition and entry into the host. It binds to the receptor protein of the host, initiating membrane fusion. Angiotensin-converting enzyme 2 (ACE2) bound by SARS-CoV and SARS-CoV-2, and dipeptidyl peptidase 4 (DPP4 or CD26) bound by MERS-CoV are found as receptors on the hosts (De Wit, et al., 2016). E protein is an integral membrane protein and is crucial for virus assembly and budding. It enhances virus budding (Neuman, et al., 2011; Nieto-Torres et al., 2011). M protein is also an integral membrane protein and is vital in producing new viruses. It binds to the membrane allowing newly produced viruses to scatter by budding (Neuman, et al., 2011).

N protein of CoVs is 45-50 KDa phosphoprotein that is involved in: i) virus assembly and also viral genome +ssRNA replication and/or transcription, and ii) viral mRNA translation. N protein interacts with +ssRNA, M protein and N protein itself (Hogue & Machamer, 2008). It binds to viral genome and other N proteins in order to generate helical ribonucleoprotein core. During the process, it also acts as RNA chaperone. N protein enters into the host cell with the viral genome and interacts with host cell proteins. The structural organization analyzes of SARS-CoVs revealed that N protein has two non-interacting structural domains, of are as N-terminal RNA-binding domain between 45-181 residues and C-terminal dimerization domain between 248-365 residues (Chang et al., 2006). Additionally, SARS-CoV M protein binds to N protein between amino acids located at 211–254 and/or 168–208 positions and these regions play important roles in N protein-protein interactions (He et al., 2004; Fang et al., 2006). N protein may undergo some post-translational modifications, including sumoylation, proteolytic cleavage, ADP-ribosylation and phosphorylation (Fung & Liu, 2018). Phosphorylation can alter function of N proteins including serine-aromatic rich region or domain 3 in the protein. Phosphorylation has some effects on N proteins during viral life-cycle. Phosphorylation of N protein usually occurs in serine-arginine rich region or domain 3 in the protein. Phosphorylation has some effects on N protein including subcellular localization and antigenic specificity (Chang et al., 2006; Huang et al., 2015). Also, N protein has some additional functions such as nucleocyttoplasmic shuttling, inhibition of S phase progress of the host and development of viral infection (Satija & Lal, 2007; Fung & Liu, 2018).

Most of CoVs N protein have epitopic sites that can be used to diagnose the infection. For instance, in SARS-CoV, the site between 371–407th amino acids in C terminus is identified as most antigenic region (Li et al., 2003). The N protein is a major structural component of CoVs and one of the most abundant viral protein produced in the host cell; therefore, can be used as antigenic protein for diagnostic purposes, and efforts in developing medicine and/or vaccine and preventing the infection. Additionally, high level of antibodies against N protein is reported in SARS-CoV patients by several researchers (Satija & Lal, 2007).

This study is aimed to perform comparative bioinformatics analysis of N protein of CoVs in order to determine various properties of SARS-CoV-2 N protein. Additionally, 3D structures of SARS-CoV-2 N protein are also generated and analyzed.

2. Materials and methods

2.1. Sequence retrieving

40 nucleotide RNA and polypeptide sequences of SARS-CoV-2 N protein were retrieved from NCBI GenBank. Coding sequences (CDS) of N protein in the viral genome were retrieved from features of CDS option except only CDS of N protein from pangolin (Zhang et al., 2020), retrieved manually from viral genome MT084071. The accession numbers of protein sequences and viral genomes, the CDS positions of N proteins, the lengths of proteins and CDS, and the origins of the countries of the viral genomes were shown in Supplementary Table 1 (S-Table1). When selecting sequences, the countries considered as epicenters of the pandemic are preferred. The distribution of the sequences by countries is as of following: 13 from USA, 9 from China, 3 from Japan, 2 from Vietnam, 2 from Italy, 2 from Colombia, 1 from each of Israel, Iran, Greece, Brazil, Spain, India, Australia and Turkey.

Additionally, 42 coding nucleotide and protein sequences from the other CoVs were also retrieved from NCBI using the same method. The sources of the sequences, the accession numbers of the N protein sequences and the viral genomes, the CDS positions of N proteins, the lengths of the proteins and CDS were shown in S-Table 2. The selected other CoV sequences are as of following: 16 from bat CoVs, 6 from pangolin CoVs (except for the protein sequence of pangolin CoV isolate MP789), 2 from civet CoV, 1 from each of rabbit and camel, 2 from human beta CoV, 1 from human enteric CoV, 3 from human MERS-CoV, 7 from human SARS and finally, 3 from avian infectious bronchitis virus (Avian IBV) Gama-CoV.

2.2. Sequence alignment and analyses

All the 81 nucleotide and protein sequences retrieved were aligned as two data sets by using BioEdit software v7.2.5 (Hall, 1999) with Clustal W multiple alignment application (Thompson, Higgins, & Gibson, 1994). Separately, 19 Sarbecoviruses selected were analyzed for sequence variations. The SARS-CoV-2 N protein percentage identities of NCBI database (NCBI, 2020) were evaluated by using BLASTP suite (Basic Local Alignment Search Tool, protein-protein Blast) with blastp option. To compare SARS-CoV-2 N proteins, the percentage of identities and cover analysis were conducted by using SARS-CoV-2 isolate N protein sequences (YP_009724397.2 Wuhan-Hu-1 and QHD43423.2 Shanghai/
3. Results and discussion

3.1. Phylogeny and relative search for the N protein

The nucleocapsid protein (N protein) sequences were retrieved and compared using nucleotide mega BLAST for the estimations of the similarity levels existed in the taxonomically related CoV N proteins. The results of nucleotide blast search were shown in Table 1.

According to the results, the highest coverage and identity values were calculated between the SARS-CoV-2 samples from Kayseri/Turkey, Shanghai/China, Risaralda/ Colombia and Bat-CoV RaTG13 (100/99.05) whereas the lowest coverage and identity values were found to be between most of SARS-CoV-2 samples and Camel CoV (74/37.16). By the relevance of these data, the most similar N protein nucleotide sequence to SARS-CoV-2 N protein nucleotide sequence was found in Bat-CoV (Bat-CoV RaTG13) among all of the taxonomically related CoVs, in agreement with previous results (Cui et al., 2019; Zhang et al., 2020).

3.2. Phylogenetic analyses

Phylogenetic constructed based on the SARS-CoV-2 N protein was tested with nucleotide and protein sequences. Four main groups were identified in both of the nucleotide and protein based joining phylogenetic trees shown in Fig. 1. The main groups consist of individuals from Igacovirus (orange), Embecovirus (green), Merbecovirus (turquoise) and Sarbecovirus (yellow). All the SARS-CoV-2 N protein sequences were grouped in Sarbecovirus group with SARS, bat, pangolin and civet CoVs by 93% and 99% bootstrap values in the nucleotide sequence and protein sequence based phylogenetic trees, respectively. There were two Sarbecovirus subgroups in both phylogenetic trees: named as A and B in the nucleotide sequence tree, and C and D in the protein based phylogenetic tree. The A and C subgroups included bat, SARS-CoV-2 and pangolin CoVs by 93% and 99% bootstrap values in the nucleotide sequence and protein sequence based phylogenetic trees, respectively. The B and D subgroups consisted of SARS-CoV-2 and bat in the nucleotide sequence and protein sequence.
Fig. 1. Nucleotide blast results of N proteins of novel SARS-CoV-2 and taxonomically related CoVs. Cover/Identity values shown in percentage.

Fig. 2. ML phylogenetic tree of SARS-CoV-2 based on amino acid sequences of the N protein.

The amino acid sequences of SARS-CoV-2 N proteins were aligned with some of the other members of Sarbecovirus to reveal sequence variation and amino acid substitutions. The results of alignment were shown in Fig. 3. The results of domain analyzes conducted using Pfam and HAMAP databases showed that the sequence between 14 and 377 positions displayed matches with family Corona_nucleocora domain in Pfam database and the two sequences displayed matches in HAMAP database (the first match in the sequence was between 41 and 186 on N terminal with CoV RNA-binding domain and the second match in the sequence was between 258 and 361 on C terminal with N protein dimerization domain). Serine x” on position 176 was phosphorylated. Tajima’s D, segregating sites and nucleotide diversities (π) were calculated for the selected 19 Sarbecovirus members and they were also used for sequence variation analyzes as -0.337920, 0.053546 and 86, respectively.
Fig. 3. The amino acid sequences of SARS-CoV-2 and the other members of Sarbecovirus N proteins based on multiple sequence alignment.
According to the sequence variation analysis, there were some amino acid substitutions in both the samples of SARS-CoV-2, and between SARS-CoV-2 and the other Sarbecoviruses. The substitutions among the samples of SARS-CoV-2 were shown in Table 2 and Fig. 3. The substitutions detected were as following: serine (S) replaced with leucine (L) in position 198 (S198L, Valencia/Spain - QIM47474.2); arginine (R) replaced with lysine (K) in position 204; glycine (G) replaced with arginine (R) in position 205 (R204K and G204R, Richmont/USA - QJA42209.1 and Greece - QIZ16579.1); threonine (T) replaced with isoleucine (I) in position 206 (T206I, Greece - QIZ16579.1); glycine (G) replaced with cysteine (C) in position 239 (G239H, Colombia - QIB84680.1); glutamine (Q) replaced with histidine (H) in position 290 (Q290H, Kanagawa/Japan - BCB979908.1); proline (P) replaced with serin (S) in position 345 (P345S). The possible serine, threonine and tyrosine phosphorylation positions for 269 NNTA were in the N protein (with 50% possibility threshold) for all the samples. The possible N protein glycosylation positions for 47 VTQ and 47 NNTA were in the N protein at position (with 50% possibility threshold) for all the samples. The possible N protein glycosylation and phosphorylation positions may provide useful information for future studies especially related with vaccine and/or antiviral drug development. Two online servers used for the detection of putative N-glycosylation and phosphorylation positions were NetNGlyc 1.0 and NetPhos 3.1 servers. The results of the predictions were shown in Fig. 4.

The possible N-glycosylation positions of SARS-CoV-2 N protein were predicted as 47 NNTA (68% possibility) and 269 NVTQ (82% possibility) using NetNGlyc 1.0 server, in consistency with the results of Supekar et al., (2020). The two positions for N-glycosylation confirmed by the authors were with lower possibility for 47 NNTA (53%) and higher possibility for 269 NVTQ (94%) (Supekar et al. 2020). The N-glycosylation positions for 269 NVTQ and 47 NNTA were in the dimerization domain and RNA binding domain at the beginning, respectively. Additionally, three more N-glycosylation positions were also predicted with low possibility at 77 NNSR (21%), 192 NNSR (45%) and 196 NSTD (13%) positions. The possible serine, threonine and tyrosine phosphorylation of the N protein was predicted to be at the 100 position (with 50% possibility threshold) for all the samples. The extensive analysis of SARS-CoV-2 sample from Kayseri/Turkey revealed that the distribution of possibilities occurred as following: 41 at 50-60%, 14 at 60-70%, 11 at 70-80%, 10 at 80-90% and 24 at 90-100%.

3.5. Comparative analysis of 3D structure of CoV N protein

The analysis of the 3D structures of the N proteins gives us information for understanding of viral packaging, taxonomical relatedness and how it functions. For that, the putative 3D structures of N proteins from the two SARS-CoV-2 samples and 11 taxonomically related previous CoVs were generated using Phyre2 server and the assessments of topological similarities of the 3D structures of the N proteins were done by analyzing TM Scores. The results of the assessments were given in Table 3.

Xu and Zhang (2010) stated that protein pairs with a TM score >0.5 are usually not in the same fold and if TM score falls below 0.17, the protein pairs are assumed not in the same fold. In this study, the highest TM score was detected between Shanghai/China SARS-CoV-2 and Beta-CoV SX2013 (0.4067)
Table 3.
The results of the similarity assessments (TM score) between SARS-CoV-2 and taxonomically related previous CoVs N protein structures.

| SARS-CoV-2 | Taxonomically Related CoVs |
|-------------|---------------------------|
| Sample | Shanghai/China (QHD43423.2) | Bat-CHOV Rat-TG13 (QHR63308.1) | Pangolin-CoV (QIA48659.1) | SARS-CoV BJ01 (AAR86785.1) | Civet-CoV (AUA04642.1) | MERS-CoV (YP009047111.1) |
| Wuhan/China (YP_009724397.2) | 0.3946 | 0.3680 | 0.2513 | 0.3894 | 0.2904 | 0.2223 |
| Shanghai/China (QHD43423.2) | 1 | 0.3987 | 0.2953 | 0.3083 | 0.3246 | 0.1788 |

Fig. 4. The results of predicted putative N-glycosylation and phosphorylation positions on three SARS-CoV-2 N proteins.
whereas the lowest was detected between Wuhan/China SARS-CoV-2 and Camel CoV (0.1558). The TM scores between SARS-CoV-2 samples and Camel-CoV, Human Enteric CoV, Human Beta-CoV, Avian IBV and MERS-CoV samples were calculated as equal or below 1.7. The TM score between SARS-CoV samples and Pangolin-CoV/Civet-CoV was resulted as higher. The average TM scores between SARS-CoV-2, and Beta-CoV SX2013, Bat-CoV RaTG13, SARS-CoV BJ01 and SARS-CoV were resulted as 0.3660 but not exceed 0.5. Interestingly, The TM score between Wuhan/China and Shanghai/China SARS-CoV-2 samples was resulted as 0.3946.

In the light of this data, although the TM scores between SARS-CoV-2 and other taxonomically related CoVs were shown consistency with the taxonomical data, the 3D structures of the selected CoV samples were not at the same fold. Also, it can be said that the 3D structures of SARS-CoV-2 N proteins were shown significant divergences.

4. Conclusion

The novel coronavirus SARS-CoV-2, the cause of Covid 19 infection, created one of the biggest outbreaks in the world history. SARS-CoV-2 has structural proteins, including S, E, M and N proteins. N protein involves in viral assembly, replication, transcription and translation (Chen et al., 2020). The sequence analysis of SARS-CoV-2 N protein RaTG13 using cover and identity values revealed that the most similar N protein sequence to SARS-CoV-2 belongs to Bat-CoV. However, the lowest cover and identity values were detected for Camel CoV. The phylogenetic analysis of SARS-CoV-2 was conducted using both of the nucleotide and protein sequences of the N protein. Four main (Igacovirus, Embecovirus, Merbecovirus and Sarbecovirus) groups identified in both nucleotide and protein trees and SARS-CoV-2 were placed in Sarbecovirus with SARS-CoV, and bat, pangolin and civet CoVs. Bat-CoV RATG13, BAT-SL ZC45 and BAT-SL ZXC21 were grouped with SARS-CoV-2 in both trees in agreement with the related literature. Additionally, a third ML tree was constructed based on the protein sequences of 40 SARS-CoV-2 N protein samples. By the phylogenetic analysis, it was shown that some of the SARS-CoV-2 N protein samples showed divergences from the other selected samples. The diverged SARS-CoV-2 N protein samples were Tokyo/Japan (BCB15098), Alexandroupolis/ Greece (QIZ16579), Richmont/USA (QJA42209), Kanagawa/ Japan (BCB97908), Valencia/Spain (QIM47474) and Bogota/ Colombia (QIS30062).

The domain search was conducted using both of Pfam and HAMAP databases. Three domains were detected as Corona nucleocora (Pfam), N terminal CoV RNA-binding domain (HAMAP) and C terminal N protein dimerization domain (HAMAP). The sequence variation analysis revealed seven amino acid substitutions within the selected SARS-CoV-2 samples. The Richmont/USA (QJA42209.1) and Greece (QIZ16579.1) SARS-CoV-2 samples were shown as having higher substitution numbers, 2 and 3 substitutions, respectively. The possible N-glycosylation positions of SARS-CoV-2 N protein were predicted as 47 NNTA with 68% and 269 NVTQ with 82%. The possible serine, threonine and tyrosine phosphorylations were predicted for 100 positions with above 50% possibility (34 of them were having higher than 80% possibility). The TM score analysis revealed that the 3D structures of SARS-CoV-2 N protein and taxonomically related CoVs were not at the same fold. Also, the TM score of N protein pairs of SARS-CoV-2 samples was calculated as 0.3946. In the light of this data, although the analysis of 3D structure data was shown to have consistency with the taxonomy of the coronavirus, SARS-CoV-2 and taxonomically related CoV N proteins, they showed significant divergences.

In last three decades, the world has encountered with some severe outbreaks caused by CoV family, including SARS (2002/03), MERS (2012/15) and Covid19 (2019-still continues). The developments of vaccines and therapeutics for fighting with the current and potential future outbreaks are crucial. Analysis, identification and interpretation of all components of pathogenic CoVs will contribute in fighting with the disease. In hope, the information gained in this study will make contributions in fighting with the current and future CoV outbreaks.

Conflict of interest: The author declares that he has no conflict of interests.

Informed consent: This manuscript did not involve human or animal participants; therefore informed consent was not collected.

References

Blom, N., Sicherriz-Pontë, T., Gupta, R., Gammeltoft, S., & Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics, 4(6), 1633-1649. https://doi.org/10.1002/pmic.200300771

Caglioni, R., Forni, D., Clerici, M., & Sironi, M. (2020). Computational inference of selection underlying the evolution of the novel coronavirus, SARS-CoV-2. Journal of Virology, 94(12), 1-11. https://doi.org/10.1128/jvi.00411-20

Carlson, C. R., Asfaha, J. B., Ghent, C. M., Howard, C. J., Hartooni, N., Safari, M., … & Morgan, D. O. (2020). Phosphorephorylation of phase separation by the SARS-CoV-2 N protein suggests a biophysical basis for its dual functions. Molecular Cell, 80(6), 1092-1103.

Ceraolo, C., & Giorgi, F. M. (2020). Genomic variance of the 2019-nCoV coronavirus. Journal of Medical Virology, 92(5), 522-528. https://doi.org/10.1002/jmv.25700

Chang, C. K., Sue, S. C., Yu, T. H., Hsieh, C. M., Tsai, C. K., Chiang, Y. C., … Huang, T. H. (2006). Modular organization of SARS coronavirus nucleocapsid protein. Journal of Biomedical Science, 13(1), 59–72. https://doi.org/10.1007/s11373-005-9035-9

Chen, Y., Liu, Q., & Guo, D. (2020). Emerging coronaviruses: Genome structure, replication, and pathogenesis. Journal of Medical Virology, 92(4), 418-423. https://doi.org/10.1002/jmv.25681

Cui, J., Li, F., & Shi, Z. L. (2019). Origin and evolution of pathogenic coronaviruses. Nature Reviews Microbiology, 17(3), 181-192. https://doi.org/10.1038/s41579-018-0118-9

De Wit, E., Van Doremalen, N., Falzarano, D., & Munster, V. J. (2016). SARS and MERS: Recent insights into emerging coronaviruses. Nature Reviews Microbiology, 14(6), 523-534. https://doi.org/10.1038/nrmicro.2016.81

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., … Finn, R. D. (2019). The Pfam protein families database in 2019. Nucleic Acids Research, 47(D1), D427-D432. https://doi.org/10.1093/nar/gky995

Fang, X., Ye, L.-B., Zhang, Y., Li, B., Li, S., Kong, L., … Wu, Z. (2006). Nucleoside analog amino acids 211 to 254, in particular, tetrad glutamines, are essential for the interaction between the nucleocapsid and membrane proteins of SARS-associated coronavirus. Journal of Microbiology, 44(5), 577-580.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 39(4), 783-791. https://doi.org/10.2307/2408876

Fung, T. S., & Liu, D. X. (2018). Post-translational modifications of
Supplementary Table 1. NCBI GenBank accession numbers and some features of retrieved SARS-CoV-2 sequences.

| Seq No | Protein Accession Number | Source Viral Genome Accession Num. | Nucleotide Position in The Genome | City/Country |
|--------|--------------------------|------------------------------------|----------------------------------|--------------|
| 1      | YP_009724397.2           | NC_045512.2                        | 28274-29533                      | Wuhan/China  |
| 2      | QHR63298.1               | MN996531.1                         | 28261-29520                      | Wuhan/China  |
| 3      | QHZ60377.1               | MT0339874.1                        | 28260-29251                      | Beijing/China|
| 4      | QIC53221.1               | MT093631.2                         | 28261-29520                      | Beijing/China|
| 5      | QHD43423.2               | MN908947.3                         | 28268-29473                      | Shanghai/China|
| 6      | QIQ63701.1               | MT253710.1                         | 28220-29479                      | Zhejiang/China|
| 7      | QHN73817.1               | MN975262.1                         | 28274-29533                      | Shenzhen/China|
| 8      | QIE07488.1               | MT1239293.2                        | 28268-29527                      | Guangzhou/China|
| 9      | QIH55882.1               | MT114413.1                         | 28274-29533                      | Hong Kong/China|
| 10     | QIA98613.1               | MT066176.1                         | 28274-29533                      | Taipei/Taiwan |
| 11     | QIX14043.1               | MT322424.1                         | 28268-29527                      | Richmond/USA  |
| 12     | QIZ64727.1               | MT334560.1                         | 28235-29494                      | Utah/USA     |
| 13     | QIC19523.1               | MT358645.1                         | 28243-29502                      | Seattle/USA  |
| 14     | QIV65095.1               | MT08700.1                          | 28225-29484                      | Michigan/USA  |
| 15     | QIQ50199.1               | MT246667.1                         | 28270-29529                      | Maryland/USA  |
| 16     | QIA42209.1               | MT350254.1                         | 28273-29532                      | Richmond/USA  |
| 17     | QIU81918.1               | MT295465.1                         | 28267-29526                      | California/USA|
| 18     | QIK50435.1               | MT192765.1                         | 28267-29526                      | California/USA|
| 19     | QIV64997.1               | MT08700.1                          | 28249-29508                      | North Carolina/USA |
| 20     | QIK02971.1               | MT188341.1                         | 28223-29482                      | Minnesota/USA |
| 21     | QIT06983.1               | MT276331.1                         | 28274-29533                      | Atlanta/USA  |
| 22     | QIZ97070.1               | MT330941.1                         | 28274-29533                      | Arizona/USA  |
| 23     | QIZ13510.1               | MT326038.1                         | 27903-29162                      | Seattle/USA  |
| 24     | BCB97908.1               | LC534419.1                         | 28266-29521                      | Kanagawa/Japan|
| 25     | BCB15098.1               | LC529905.1                         | 28274-29533                      | Tokyo/Japan  |
| 26     | BCD87661.1               | LC542976.1                         | 28274-29533                      | Tokyo/Japan  |
| 27     | QIK55986.1               | MT127114.1                         | 143-1402                         | Hanoi/Vietnam|
| 28     | QIK50455.1               | MT192773.1                         | 28273-29532                      | Ho Chi Minh/Vietnam |
| 29     | QIC50505.1               | MT077125.1                         | 28218-29476                      | Rome/Italy   |
| 30     | QIA98856.1               | MT066156.1                         | 28274-29533                      | Rome/Italy   |
| 31     | QIT06985.1               | MT276597.1                         | 28254-29513                      | Nesa Ziona/Israel |
| 32     | QIS30602.1               | MT256924.2                         | 28220-29479                      | Bogota/Colombia |
| 33     | QIB84680.1               | MT072888.1                         | 28259-29518                      | Risaralda/Colombia |
| 34     | QIX16579.1               | MT328035.1                         | 28274-29533                      | Alexandroupoulos/Greece |
| 35     | QIG50001.1               | MT126808.1                         | 28274-29533                      | Sao Paulo/Brazil |
| 36     | QIM47447.2               | MT198362.2                         | 28280-29479                      | Valencia/Spain |
| 37     | QHS34553.1               | MT012998.1                         | 28258-29517                      | Maharashttra/India |
| 38     | QHR84456.1               | MT00544.1                          | 28274-29533                      | Melbourne/Australia |
| 39     | QIX17203.1               | MT320891.2                         | 28203-29489                      | Tehran/Iran  |

Protein length 419 aa Nucleotide length 1260 bp
**Suppl. Table 2.** NCBI GenBank accession numbers and some features of retrieved other CoV sequences.

| Seq No | Source | Protein Accession Number | Protein length | Source Viral Genome Accession Num. | Nucleotide length | Nucleotide Position in The Genome |
|--------|--------|--------------------------|----------------|-----------------------------------|------------------|---------------------------------|
| 1      | Bat-CoV RatG13 | QHR63308.1 | 419 | MN996532.1 | 1260 | 28240-29499 |
| 2      | Bat-CoV WIV1 | AGZ48841.1 | 422 | KF367457.1 | 1269 | 28686-29954 |
| 3      | Bat-CoV HKU3-13 | ADE34831.1 | 421 | GQ155148.1 | 1266 | 28073-29338 |
| 4      | Bat-CoV RsSHC014 | AGZ48815.1 | 422 | KC881005.1 | 1269 | 28162-29430 |
| 5      | Bat-CoV YNL1_31C | AKZ19084.1 | 421 | KPE88808.1 | 1268 | 28103-29368 |
| 6      | Bat-CoV RI/2004 | ABD75315.1 | 421 | DQ412042.1 | 1266 | 28084-29349 |
| 7      | Bat-CoV bat-SL-CoVZXC21 | AVP78049.1 | 419 | MG772934.1 | 1260 | 28110-29369 |
| 8      | Bat-CoV Longquan-140 | AID16719.1 | 421 | KF294457.1 | 1260 | 28072-29337 |
| 9      | Bat-CoV Rs7327 | ATO98228.1 | 422 | KY147151.1 | 1269 | 28684-29952 |
| 10     | Bat-CoV bat-SL-CoVZXC45 | AVP78038.1 | 419 | MG772933.1 | 1260 | 28179-29438 |
| 11     | Bat BtRs-Beta-CoV/HuB2013 | AIA62318.1 | 420 | KJ473814.1 | 1263 | 28049-29311 |
| 12     | Bat BtRs-Beta-CoV/GX2013 | AIA62328.1 | 422 | KJ473815.1 | 1269 | 27865-29133 |
| 13     | Bat BtRF-Beta-CoV/SX2013 | AIA62308.1 | 421 | KJ473813.1 | 1266 | 27849-29114 |
| 14     | Bat BtRI-Beta-CoV/SC2018 | QDF43818.1 | 421 | MK211374.1 | 1261 | 28076-29341 |
| 15     | Bat-CoV BM48-31/BGR/2008 | YP_003855891.1 | 417 | NC_014470.1 | 1254 | 27665-28918 |
| 16     | Bat –Beta-CoV/SC2013 | AHY61344.1 | 434 | KJ473821.1 | 1305 | 28819-30123 |
| 17     | Pangolin-CoV isolate MP789 | ---- | ---- | MT084071.1 | 1260 | 25752-27213 |
| 18     | Pangolin-CoV PCoV_GX-P5E | QIA48648.1 | 417 | MT040336.1 | 1254 | 28226-29479 |
| 19     | Pangolin-CoV PCoV_GX-P5L | QIA48639.1 | 417 | MT040335.1 | 1254 | 28230-29483 |
| 20     | Pangolin-CoV PCoV_GX-P1E | QIA48630.1 | 417 | MT040334.1 | 1254 | 28224-29477 |
| 21     | Pangolin-CoV PCoV_GX-P4L | QIA48621.1 | 417 | MT040333.1 | 1254 | 28229-29482 |
| 22     | Pangolin-CoV_GX-P2V | QIO54565.1 | 417 | MT072864.1 | 1254 | 28218-29471 |
| 23     | Civet-CoV 007/2004 | AUA04642.1 | 422 | AY572034.1 | 1269 | 28123-29391 |
| 24     | Civet-CoV civet010 | AUA04658.1 | 422 | AY572035.1 | 1269 | 28101-29369 |
| 25     | Rabbit-CoV HKU14 | YP_005452494.1 | 444 | NC_017083.1 | 1335 | 29462-30796 |
| 26     | Avian IBV Gama-CoV Beaudette | NP_040838.1 | 409 | NC_001451.1 | 1230 | 25873-27102 |
| 27     | Avian IBV Gama-CoV CK/CH/LHLJ/04V | ACO58535.1 | 409 | FJ821744.1 | 1230 | 1-1230 |
| 28     | Avian IBV Gama-CoV | QIO20949.1 | 409 | MN894514.1 | 1230 | 1-1230 |
| 29     | Camel-CoV HKU23 | ATIO9453.1 | 448 | MF593476.1 | 1346 | 29227-30573 |
| 30     | Human Beta-CoV England 1 | AFY13314.1 | 411 | KC164505.2 | 1236 | 28565-29800 |
| 31     | Human Beta-CoV 2c England-Qatar/2012 | AGG22549.1 | 411 | KC667074.1 | 1236 | 28566-29801 |
| 32     | Human Enteric-CoV 4408 | ACJ35489.1 | 449 | FJ415324.1 | 1347 | 29394-30740 |
| 33     | Human MERS-CoV HCoV-EMC | YP_009047211.1 | 413 | NC_019843.3 | 1242 | 28566-29807 |
| 34     | HCoV MERS Riyadh-KSA-19003852/2019 | QEJ82233.1 | 413 | MN365233.1 | 1242 | 28566-29807 |
| 35     | MERS related-CoV Qatar15 | AZU90738.1 | 413 | MK280984.2 | 1242 | 28539-29780 |
| 36     | Human SARS-CoV WH20 | AAX16200.1 | 422 | AY772062.1 | 1269 | 27853-29121 |
| 37     | Human SARS-CoV GZ02 | AAS00111.1 | 422 | AY390556.1 | 1269 | 28149-29417 |
| 38     | Human SARS-CoV BJ01 | AAS45456.1 | 422 | AY536760.3 | 1269 | 81-1349 |
| 39     | Human SARS-CoV BJ01 | AAP30037.1 | 422 | AY278488.2 | 1269 | 28101-29369 |
| 40     | Human SARS-CoV Rs. 672/2006 | ACU31039.1 | 422 | FJ588686.1 | 1269 | 27520-28788 |
| 41     | Human SARS-CoV ShanghaiQXC2 | AAR86785.1 | 422 | AY463060.1 | 1269 | 27462-28730 |
| 42     | Human SARS-CoV Taiwan TC3 | AAP97890.1 | 422 | AY348314.1 | 1269 | 28051-29319 |