*Baku State University Journal of Life Sciences & Biology **2024**, v.1 (1), p. 13-19*

journal homepage: http://bsuj.bsu.edu.az/en

SARS-COV-2: WHERE AND HOW DID IT EMERGE FROM?

Ilham A. Shahmuradov[a,b] *, Karim G. Gasimov[b]

[a]Institute of Molecular Biology and Biotechnologies, Baku, Azerbaijan
[b]Institute of Biophysics, Baku, Azerbaijan

Received 09 october 2023; accepted 06 november 2023

Abstract

The coronavirus disease discovered in 2019 (COVID-19) is caused by recently discovered human Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). An initial step of the SARS-CoV-2 entrance into the host cell depends on the spike (S) protein which interacts with the angiotensin-converting enzyme 2 (ACE2) as a receptor, via the receptor binding domain (RBD). We compared the S proteins from SARS-CoV-2, SARS, bat and pangolin CoVs. The most striking fact firstly discovered in this study is that the relative proportion of the synonymous substitution rates between SARS-CoV-2 and pangolin CoV are significantly higher than the corresponding characteristics for other CoVs studied. By comparing with SARS and bat CoVs, the significantly higher rate of the synonymous substitutions between the human SARS-CoV-2 and the pangolin CoV puts several intriguing questions on an emergence and the duration of divergence of the SARS-CoV-2.

Keywords: coronavirus; COVID-19; origin of the SARS-CoV-2; synonymous mutations.

1. Introduction

Coronaviruses (CoVs) are enveloped positive-sense RNA viruses with club-like spikes on their surface. CoVs cause diseases of wide range in mammals and birds [5]. The emergence and outbreak of a new acute respiratory syndrome is associated with recently discovered human Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and referred to as coronavirus disease discovered in 2019 (COVID-19) [15]. To date, there have been more than 31 million confirmed cases of COVID-19, including about one million deaths, reported by WHO (https://globalbiodefense.com/novel-coronavirus-covid-19-portal/). Prior to COVID-19, there were two other coronavirus-related syndromes reported worldwide, the highly pathogenic SARS-CoV (China, 2002-2003) and the Middle East respiratory syndrome Coronavirus (MERS-CoV; Kingdom of Saudi Arabia, 2012) [4]. Thus, only over the past two decades, the humanity has been plagued by 3 coronavirus diseases, while before widely spread (pandemic) diseases related to the human coronaviruses were discovered in the 1960s [11]. So, a question arises: is it a coincidence that coronaviruses have recently become very active in various forms and on a large scale with very dangerous complications? To answer this question, we have to clarify an origin and approximate time passed since a novel coronavirus emerged in the human organism. Moreover, understanding events of emergence of human coronaviruses might help to predict and prevent a new pandemic emergence in the future.

* Corresponding author. Tel.: +994 50 384 24 96
E-mail address: *ilhambaku@gmail.com*

At present, bats are supposed to be one of the natural reservoirs various viruses, including highly transmissible and pathogenic viruses as SARS-CoV and MERS-CoV (for a review see: [4]. Both experimental and theoretical studies indicate that many coronaviruses are capable of interspecies transmission [17]. In particular, various species of horseshoe bats in China harbor SARS-like coronaviruses, even some bat coronaviruses and SARS-CoV can use the same receptor for cell entry [7]. Furthermore, Menachery and colleagues [12] examined the disease potential of a SARS-like virus, SHC014-CoV from the Chinese horseshoe bat populations. For this purpose, they have designed a chimeric virus and introduced it to the mouse backbone. The results of the experiment indicated that the chimeric virus is able efficiently (i) to use orthologs of the SARS receptor for the ACE 2), (ii) replicate in the primary cells of the human respiratory tract and (iii) achieve *in vitro* titers close to the epidemic strains of SARS-CoV. Besides, it was *in vivo* demonstrated that the recombinant virus can replicate in mouse lung with significant pathogenesis using the novel spike protein. These findings suggested that the full-length recombinant viruses have a potential of emergence in human. However, it was later found that the S1 protein of the SARS-CoV-like virus from the pangolin lung samples is much more closely related to SARS-CoV-2 than to bat. Therefore, it was concluded that SARS-CoV-2-like CoVs are originating from pangolin species [9,10,13, 14, 21].

Coronaviruses contain a non-segmented, positive-sense RNA genome of ~30 kb, with a 5′-cap and a 3′ poly (A) tail structure. This structure allows for direct translation of RNA in the form of mRNA of viral polypeptides [5, 7]. The SARS-CoV-2 genome of 29,880 nt length encodes both structural (S, E, M, and N) and nonstructural (3-chymotrypsinlike protease, papain-like protease, and RNA-dependent RNA polymerase) proteins (for review see: [8].

An initial step of the SARS-CoV-2 entrance into the host cell depends on the spike (S) protein (~150 kDa) which uses the angiotensin-converting enzyme 2 (ACE2) as a receptor. The homotrimeric S protein gains access to the endoplasmic reticulum via the N-terminal signal sequence and is N-linked glycosylated. The protein is cleaved by a host cell furin-like protease into two subunits, the N-terminal S1 (13-685 aa) and the C-terminal S2 (686-1270 aa). The S1 subunit recognizes and binds to the host ACE2. The S2 plays a key role in a fusion of viral and host cell membranes. The S1 subunit was found to be much more variable in the inter-species manner, than the S2 subunit. This observation is probably related to a decisive role of the S1 subunit in recognition and interaction with the host cell receptor. Significant differences in S1 subunits of the spike proteins from different coronaviruses may partially determine their tissue tropism and host range. The crystal structure analysis of the S protein revealed that, comparing with the full S1 subunit, a 193 aa region (318-510 aa) binds the ACE2 receptor more efficiently and it was defined as the receptor binding domain (RBD) of SARS-CoV-2. Moreover, a loop subdomain (424-494 aa) was shown to contact directly with ACE2 and was defined as the receptor-binding motif, RBM [6].

According to resent reports, the neutralizing antibodies are generated in response to the penetration and fusion of surface-exposed S-protein (mainly the RBD domain), which is believed to be an important target for vaccine candidates [6, 19]. However, SARS-CoV-2 has remarkable properties such as glutamine-rich 42 aa long, exclusive molecular signature (*DSQQTVGQQDGSEDNQTTTIQTIVEVQPQLEMELTPVVQTIE*) in position 983-1024 of polyprotein 1ab (pp1ab) [2], diversified RB domain, unique furin cleavage site (PRRARain, unique furin 1ab (pp1ab) h as glutamine-rich 42 aa long, exclusviral pathogenesis, diagnosis and treatments [3].

In this study, we compared S proteins from SARS-CoV-2, SARS-Cov, bat CoV and pangolin CoV, both at the whole protein and distinguished subunits/domains levels. Here, we present some results of these studies.

2. Materials and methods

For analysis, both CDS and protein sequences of the S protein from the human SARS-CoV-2 (GenBank accession MN997409.1), SARS-CoV (NC_004718.3) and MERS CoV (NC_019843.3), as well as CoVs from bat (MG772934.1), pangolin (MT040335.1) and avian (M95169.1) were selected. Both pairwise and multiple alignments of the spike proteins were performed for whole proteins as well as RBDs (S1 subunit), RBMs (S1) and S2 subunits (Fig. 1).

A comparison CDS and protein sequences was done by BLAST tool [1]. The multiple alignment of CDS and protein sequence was performed by the Clustal Omega tool [16].

To explore statistical characteristics of variations, as identities, synonymous and non- synonymous substitutions, and insertion/deletions (Indels) we developed a new tool, MUTAN. As the source (input) data for the tool, an output of the pairwise alignment of the protein sequences by the Clustal Omega in the FASTA format, as well as corresponding query CDS sequences.

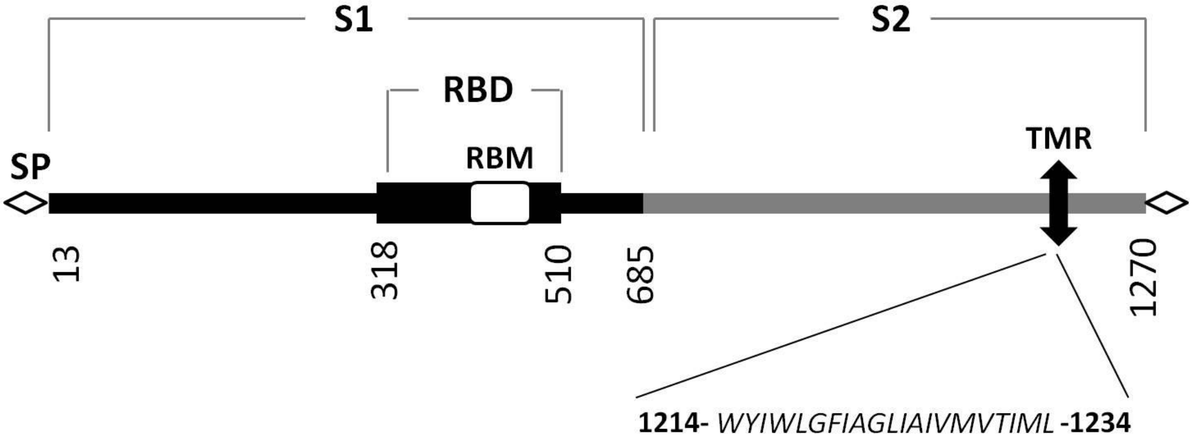

Fig 1. The schematic representation of the S protein structure (SARS-CoV-2) used in the study.

3. Results and discussion

Initially, using the BLAST tool, we compared whole S protein sequences from six different coronaviruses (see: Materials and Methods). As expected from literature data, S proteins of two coronaviruses (MERS and avian) have not significant similarity with SARS-CoV-2. Therefore, further we studied only two human coronaviruses (SARS-CoV-2, SARS-CoV), bat and pangolin CoVs.

First, for selected 4 coronaviruses, we aligned (i) the whole S proteins, (ii) the RBDs (S1), (iii) the RBMs (S1) and (iv) the S2-subunits of the spike proteins using the Clustal Omega tool. Then, the results of alignment were analyzed by the MUTAN program. Results of the analysis are summarized in the Table 1 and illustrated in Fig. 2.

The first noteworthy finding of this study is that pangolin CoV is significantly closer to the human SARs-CoV-2 virus in terms of S protein. This closeness is both on the scale of the whole protein and on the subunits S1 and S2. Thus, while the similarity between SARS-CoV-2 and pangolin virus on the whole protein is about 93%, these figures on SARS-CoV and bat CoV are 76% and 32%, respectively. Approximately the same tendency of difference is observed in RBDs, RBMs and S2 subunits: 87% against 72% and 26% in RBD, 80% against 55% and 23% in RBM, and 97% against 89% and 39% in S2 subunit. It should be noted that this result is fully consistent with the results recently obtained by some other authors [9, 10, 21].

Table 1. The similarity details of the CDS and amino acid sequences of the S proteins from human (SARS-COV-2, SARS-CoV), pangolin and bat CoVs

	Conserv. level: CoV-2 vs P-CoV	Conserv. level: CoV-2 vs SARS CoV	Conserv. level: CoV-2 vs B-CoV
For whole proteins:			
Identities (amino acids)	1176 (1273*), 92.4%	968 (1273), 76.0%	407 (1273), 32.0%
Identical codons (CDS)	739 (out of 1273), 58.1%	533 (1273), 41.9%	186 (1273), 5.3%
Syn. substitutions (CDS)	437 (528), 82.8%	435 (718), 60.6%	221 (1029),21.5%
Non-syn. substitutions (CDS)	91 (528), 17.0%	283 (718), 39.42%	808 (1029),78.5%

Indels (amino acids)	6 (1273), 0.5%	26 (1273), 2.04%	117 (1273), 9.2%
Glycosylation sites conserved	21 (22)	15/22	20/22
SS-bonds conserved	12/12	11/12	12/12
For RBD domains (S1 subunits):			
Identities (amino acids)	168 (193), 87.05%	138 (193), 71.50%	50 (193), 25.91%
Identical codons (CDS)	31 (193), 16.06%	5 (193), 2.59%	3 (193), 1.55%
Syn. substitutions (CDS)	137 (162), 84.57%	133 (187), 71.12%	47 (179), 26.26%
Non-syn. substitutions (CDS)	25 (162),15.43%	54 (187), 28.88%	132 (179),73.74%
Indels (CDS)	0 (193)	1 (193), 0.52%	11 (193), 5.70%
For RBM sites (S1 subunits):			
Identities (amino acids)	57 (71), 80.28%	39 (71), 54.93%	16 (71), 22.54%
Identical codons (CDS)	0 (71)	0 (71)	0 (62)
Syn. substitutions (CDS)	57 (71), 80.28%	39 (70), 55.71%	16 (62), 25.81%
Non-syn. substitutions (CDS)	14 (71), 19.72%	31 (70), 44.29%	46 (62), 74.19%
Indels (CDS)	0 (71)	1 (71), 1.41%	9 (71), 12.68%
For S2 subunits:			
Identities (amino acids)	590 (609), 96.88%	540 (609) 88.67%	236 (609), 38.75%
Identical codons (CDS)	16 (609), 2.63%	495 (609), 81.28%	6 (609), 0.99%
Syn. substitutions (CDS)	574 (589), 97.45%	45 (110), 40.91%	230 (586), 39.25%
Non-syn. substitutions (CDS)	15 (589), 2.55%	65 (110), 59.09%	356 (586), 60.75%
Indels (CDS)	4 (609), 0.66%	4 (609), 0.66%	17 (609), 2.79%

*Out of length of the comparison region. Syn. substitutions: Synonymous substitutions; Non-syn. substitutions: Non-synonymous substitutions; Conserv. level: Conservation level; P-CoV: Pangolin CoV; B-CoV: Bat CoV; SS-bonds: disulfide bridges. Most significant variations in the amino acid and codon compositions are marked in grey.

However, the most striking finding of this study is the following: for each of the 4 comparison regions mentioned above (i – iv), the relative proportion of the synonymous substitution rates between SARS-CoV-2 and pangolin CoV (83%, 85%, 80% and 98%) are significantly higher than the corresponding characteristics for both SARS-CoV-2/SARS-CoV and SARS-CoV/bat CoV (61%, 71%, 56% and 41%; 22%, 26%, 26% and 39%, respectively). In particular, comparing with the full S1 subunit of the S protein, it was previously found that the RBD of the S1 binds the ACE2 receptor more efficiently [8]. This observation agrees with our finding illustrated in Fig. 3.

In contrast, almost all known N-linked-glycosylation sites and SS-bonds were found to be conserved within the S protein from all 4 coronaviruses. Moreover, the transmembrane domain in the S proteins of all 4 coronaviruses analyzed is completely conserved (see: Fig. 1). This observation may indicate that the glycosylation sites, SS-bonds, and transmembrane domain are generally conserved features of the S proteins required for their total functional status, regardless the interaction specificity of the S proteins and the corresponding receptors.

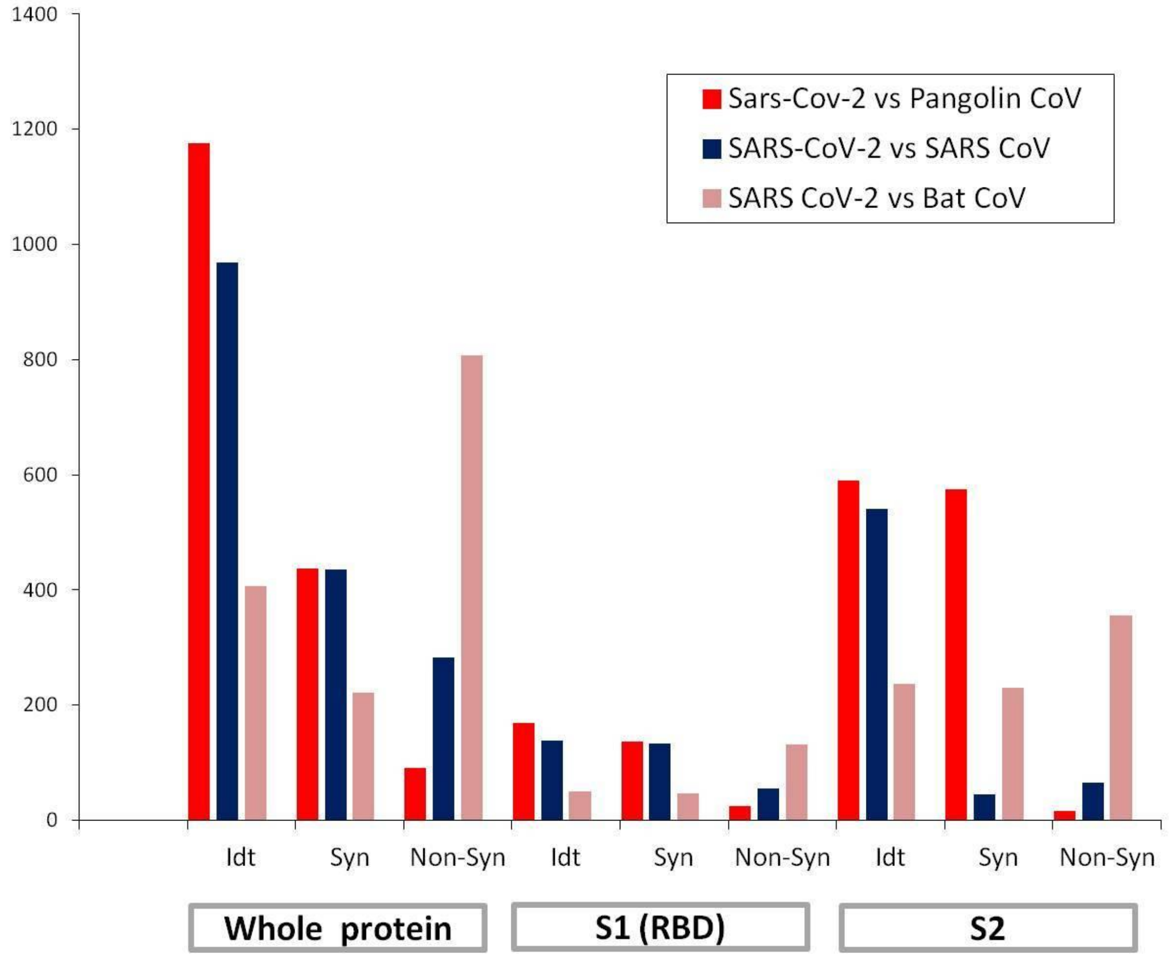

Fig 2. A graphical view of differences in identities, synonymous substitutions and non-synonymous substitutions within whole S proteins, RBDs (S1 subunit) and S2 subunits between SARS-CoV-2 and other 3 CoV-s analyzed.

```
SARS           FRVVPSGDVVRFPNITNLCPFGEVFNATKFPSVYAWERKKISNCVADYSVLYNSTFFSTF   344
Bat CoV        FRVQPTQSIVRFPNITNVCPFHKVFNATRFPSVYAWERTKISDCIADYTVFYNSTSFSTF   352
Pangolin CoV   FRVQPTISIVRFPNITNLCPFGEVFNASKFASVYAWNRKRISNCVADYSVLYNSTSFSTF   355
CoV2           FRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTF   377

SARS           KCYGVSATKLNDLCFSNVYADSFVVKGDDVRQIAPGQTGVIADYNYKLPDDFMGCVLAWN   404
Bat CoV        KCYGVSPSKLIDLCFTSVYADTFLIRFSEVRQVAPGQTGVIADYNYKLPDDFTGCVIAWN   412
Pangolin CoV   KCYGVSPTKLNDLCFTNVYADSFVVKGDEVRQIAPGQTGVIADYNYKLPDDFTGCVIAWN   415
CoV2           KCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWN   437

SARS           TRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCTP-PALNCYWPLNDYGF   464
Bat CoV        TAKQDT-----GHYFYRSHRSTKLKPFERDLSSDEN--------------GVRTLSTYDF   463
Pangolin CoV   SVKQDALTGGNYGYLYRLFRKSKLKPFERDISTEIYQAGSTPCNGQVGLNCYYPLERYGF   475
CoV2           SNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGF   497

SARS           YTTTGIGYQPYRV                                                  523
Bat CoV        NPNVPLEYQATRV                                                  513
Pangolin CoV   HPTTGVNYQPFRV                                                  535
CoV2           QPTNGVGYQPYRV                                                  510
```

Fig 3. Alignment of the RBDs in the S1 subunits of spike proteins from human (SARS-COV-2, SARS-CoV), pangolin and bat CoVs. The identical amino acids in all 4 COVs, as well as only in the SARS-CoV-2 and

pangolin CoV are given in red color. The identical amino acids in the SARS-COV-2, SARS-CoV and pangolin CoV are given in blue color. "-"indicates indels.

Our results is fully consistent with a recent suggestion on the pangolin origin of the SARS-CoV-2 [9, 10, 21]. Furthermore, by comparing with SARS and bat CoVs, the significantly higher rate of the synonymous substitutions between the human SARS-CoV-2 and the pangolin CoV puts several intriguing questions on an emergence and the duration of divergence of the SARS-CoV-2, including:

(1) Did the virus really pass from other animals (pangolins?) to humans recently (even in late 2019)? If so, how the virus acquired a lot of (more than 400!) synonymous mutations by a natural way in less than a year? Or there was a human intervention in emergence of a novel and very dangerous virus?

(2) If the virus entered the human body a few years ago, why and how did it not show its existence during these years?

Certain serious arguments can be made for and against each of these assumptions. However, the available facts do not preclude a definitive answer to these questions, and perhaps this will never be possible. In our personal opinion, we believe that scientists are capable of creating one or more terrible coronaviruses. The creation of the chimeric coronavirus a few years ago [12] is proof of this! And later, a certain researcher or technical staff, not knowingly, but out of simple irresponsibility, caused the new artificial virus to leave the laboratory.

References

[1] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997, 25 (17), p.3389-3402.

[2] Cardenas-Conejo Y, Linan-Rico A, Garcia-Rodruez DA, Centeno-Leija S, Serrano-Posada H. An exclusive 42 amino acid signature in pp1ab protein provides insights into the evolutive history of the 2019 novel human-pathogenic coronavirus (SARS-CoV2), *J. Med. Virol.* 2020, 92(6), p. 688-692.

[3] Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res*. 2020, 176, p. 104742; doi: 10.1016/j.antiviral.2020.104742.

[4] Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbial.* 2019, 17(3), p.181-192.

[5] Fehr AR, Perlman S. Coronaviruses: an overview of their replication and pathogenesis. *Methods. Mol. Biol.* 2015, 1282, p.1-23.

[6] Hoffmann M, Kleine-Weber H, Schroeder S, Kruger N et al. *Cell.* 2020, 181 (2), p. 271-280.

[7] Hu B, Ge X, Wang LF, Shi Z. Bat origin of human coronaviruses. *Virology J.* 2015, 12, p.221, DOI: 10.1186/s12985-015-0422-1.

[8] Huang Y, Yang C, Xu X-f, Xum W, Liu S-w. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacologica Sinica*. 2020, 41(9), p.1141-1149.

[9] Lopes LP, Cardillo GdeM, Paiva PB. Molecular evolution and phylogenetic analysis of SARS-CoV-2 and hosts ACE2 protein suggest Malayan pangolin as intermediary host. *Brazilian J. of Microbiology* (June 26, 2020), p.1-7; doi: 10.1007/s42770-020-00321-1.

[10] Malaiyan J, Arumugam S, Mohan K, Radhakrishnan GG. An update on the origin of SARS-CoV-2: Despite closest identity, bat (RaTG13) and pangolin derived coronaviruses varied in the critical binding site and O-linked glycan residues. *Med. Virol.* (July 7, 2020), p.1–7; doi:10.1002/jmv.26261.

[11] McIntosh K. Coronaviruses: A Comparative Review. In: *Arber W, Haas R, Henle W et al. (eds.). Current Topics in Microbiology and Immunology/Ergebnisse der Mikrobiologie und Immunitätsforschung*. Berlin, Heidelberg: Springer, 1974, p.85-129.

[12] Menachery V, Yount B, Debbink K, Agnihothram S, Gralinski LE et al. A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat Med*. 2015, 21(12), p.1508–1513.

[13] Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect. Dis.* (July 14, 2020); doi: 10.1016/S1473-3099(20) 30562-4.

[14] Munnink BB, Nieuwenhuijse DF, Stein M, O'Toole A, Haverkate M, et al. Dutch-Covid-19 response team. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nature Medicine*. 2020, 26(9), p.1405-1410.

[15] Shi Z, Hu Z. The genetic sequence, origin, and diagnosis of SARS-CoV-2. *Eur. J. Clin. Microbiol. Infect. Dis.* 2020, 39(9), p.1629-1635.

[16] Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* 2018, 27(1), p.135-145.

[17] Tang, Q, Song Y, Shi M. Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. *Scientific Reports*. 2015, 5, p.17155; DOI: 10.1038/srep17155.

[18] Tang X, Wu C, Li X, Song Y, Yao X et al. On the origin and continuing evolution of SARS-CoV-2. *National Science Review*. 2020, 7(6), p.1012-1023.

[19] Tian X, Li C, Huang A, Xia S, Lu S, et al. Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody, *Emerg. Microbes Infect,* 2020, 9(1), p. 382-385.

[20] Wang H, Li X, Li T, Zhang S, Wang L, Wu X, Liu J. The genetic sequence, origin, and diagnosis of SARS-CoV-2. *Eur. J. Clin. Microbiol. Infect. Dis.* 2020, 39(9), p.1629-1635.

[21] Zhang T, Wu Q, Zhang Z. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Current Biology*. 2020, 30(7), p.1346-1351.