IPAK RESEARCH REPORT COVID-19-2020_1

Motif Pathogenicity Typing: SARS-CoV-2 Coronavirus May Contain a Unique “Likely Pathogenic” Protein Motif Signature Also Found in a Natural Isolate from 2005

James Lyons-Weiler
The Institute for Pure and Applied Knowledge
Pittsburgh, PA

Abstract

Careful analysis of protein motif signatures and iterative phylogenetic analysis of amino acid sequences of coronaviruses reveal a characteristic functional motif signature of the coronavirus first detected in afflicted humans in China in December, 2019, currently responsible for the deaths of over 700 people so far and massive isolation efforts in China. New results show that original conjectures the modified protein signatures as indicative of supporting a likely laboratory origin via recombination technology do not, at this time, appear to be not supported. New data indicate that invoking a recent laboratory intentional modification is not necessary to understand the origins of the pathogenicity. This, however, does not rule out a laboratory escape of pathogenic virus for pathogenicity or vaccine development research. Continuous ongoing detailed sequence analysis is necessary. An evolutionary awareness of the inheritance and acquisition of characteristic functional motif signatures will aid in understanding limits of applicability such studies; studies of “SARS” and “SARS-Like” coronaviruses cannot be expected to be useful to generalize about the pathogenicity from sequence or structural insights alone. Determining, recording and publishing functional motif signatures could prove essential for communicating pathogenic-like coronavirus types with respect to expected pathogenicity. Laboratories with natural isolates, isolates from humans, and recombined coronaviruses that include spike protein sequences should analyze the spike protein for the putatively characteristic pathogenic functional motif signature: which appears at this time to include a shortened N-terminal spike domain, missing She-3 and KxDL motifs in the Spike 2 segment, and a C-terminal Gp41 (retroviral envelope) motif. Research into C-terminal motifs and elements may prove to potentially useful for tracking laboratory modified coronavirus types. The presence of the proposed pathogenicity signature and an understanding of the provenance of the sequence information involves laboratory origin of related sequences, but seems at this time to rule out recent laboratory origin of the SARS-CoV-2 lineage. A recently reported furin-like cleavage site that has a similar phylogenetic distribution in B-coronaviruses also seems like a promising lead for therapeutics.

Introduction

The purpose of the rapid communication is to present data that could test the hypothesis that SARS-CoV-2 necessarily had required a laboratory-mediated recombination origin. The world first gleaned knowledge of the genome of the SARS-CoV-2 from the genome sequence shared by Wu et al., (2020) from isolates from patients with SARS-like symptoms in Wuhan, China (Wu et al., 2020). A mystery segment that was difficult to match was detected (Ji et al., 2020) and referenced as a "middle segment" in a second publication but not explained (Paraskevis et al., 2020). The concept that a laboratory origin may be behind the outbreak was first published online via online accounts, and became popularized on social media due to a partial match found for the otherwise difficult-to-identify nucleotide segment in the SARS-CoV-2 genome sequences. At first glance, a link to recombination research might reasonably be considered plausible given that a successful match of the middle fragment via a search of non-viridae sequences, pShuttle-SN. The mere existence of part of a SARS-like coronavirus Spike protein S1 sequence in a pShuttle vector shows that it was evidently used at some time in China for recombination research on SARS spike protein function, or for vaccine research has fueled suspicion. The hypothesis of laboratory origin, originally represented among a list of hypotheses, seemed most likely given the apparent increased match between the vector sequence. The public's imagination was captured by the apparently close phylogenetic link between SARS-CoV-2 sequences deposited in Genbank by institutes in China: one from a military institute in Nanjian, China (AVP78042), and another deposited after the Dec 1, 2019 outbreak by the Wuhan Institute of Virology (QHR63300). The scenario has serious international policy implications and thus deserves closer analysis.

In spite of the concept being enumerated among other hypotheses, numerous outlets in the popular press picked up the potential link as a statement of certainty, and with regards to the pShuttle-link tended to emphasize the single "laboratory origin" hypothesis as an isolated, solitary possibility, confusing "most likely at this time" with "certainty". The exclusion of the other hypotheses in the press has been unfortunate, because they included, from the start, natural origin with, or without recombination. The present author also drew a similarly strong "most likely" assessment. The hypotheses enumerated also include laboratory accidental release of a natural sequence. The vector-technology link, however, required further investigation with an aim to falsify the hypothesis of laboratory origin.

To challenge and thereby test such hypotheses is essential and ideally would be based on available, not newly deposited evidence. If evidence that viruses isolated prior to 2019 from animals caught in the wild had been studied and existed in China, or elsewhere, resemble SARS-CoV-2 with respect especially to the function of the Spike protein elements, the hypothesis could be downgraded rather dramatically from "most likely from laboratory recombination" to "implausible" in this specific pandemic.

The complexities of such assessment is daunting; recombination and host-switching along with recombination - which is essential for coronavirus transcriptional biology - and both processes hamper understanding. In at least one case, natural sequences deposited in NCBI's nucleotide sequence database have been reconstructed as consensus sequences and made biologically active anew (from data to infective) and modified in the lab, with gain-of-function motifs added to enhance infectivity into human calls (Becker et al., 2008).

Pathogenicity in viruses is also known to be increased, or decreased via mutations within specific motifs (e.g., Liedmann et al., 2014). Analyzing such data alongside sequences from wild-caught animals is precarious because the mere presence of terms like “Military Institute” and “Synthetic Life Form” and “Recombinant virus” combined with knowledge of their geographic proximity to the epicenter of the outbreak, all enmeshed with complexities of interspecies transfer make teasing apart a sensible “most likely” position especially challenging. Tools available for phylogenetic analysis at the nucleotide or sequence are of limited use in the study of evolutionary histories for viruses that undergo recombination. Some tools exist for studying reticulation, and initial results have not favored evidence of recent natural recombination (in the wild) in the SARS-CoV-2 given sequences thought to be associated or related to the lineage. Motif analysis, however, allows rapid characterization of functional elements in proteins that may be useful for pathogenicity typing.

The goal of the present report is to share the discovery of a potentially characteristic signature Spike, S1 and S2 protein functional motif signature that may be characteristic of the pathogenicity associated with SARS-CoV-2, to report which other B-coronaviruses sequenced share the motif, and thereby test the hypothesis of a recent laboratory origin of SARS-CoV-2.

Results

Phylogenetic analysis of Spike protein amino acid sequences with Neighbor-Joining following alignment via G-INS-i led to a large phylogenetic tree revealing various subgroups within the available coronaviruses sequences. Motif analyses focused on the SARS-CoV-2 and closely related sequences and representative sequences from different parts of the tree showed variation in recognized motifs in a manner that mostly reflects a pattern of inheritance consistent with the phylogenetic tree estimated from amino acid sequences (Fig. 1).

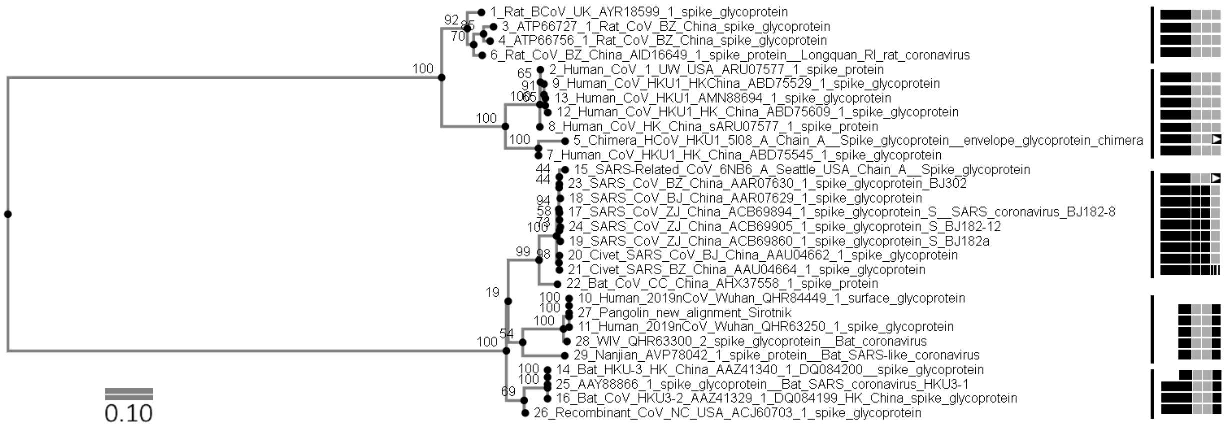
1_Rat_BCoV_UK_AYR18599_1_spike_glycoprotein
3_ATP66727_1_Rat_CoV_BZ_China_spike_glycoprotein
4_ATP66756_1_Rat_CoV_BZ_China_spike_glycoprotein
6_Rat_CoV_BZ_China_AID16649_1_spike_protein__Longquan_Rl_rat_coronavirus
2_Human_CoV_1_UW_USA_ARU07577_1_spike_protein
9_Human_CoV_HKU1_HKChina_ABD75529_1_spike_glycoprotein
13_Human_CoV_HKU1_AMN88694_1_spike_glycoprotein
12_Human_CoV_HKU1_HK_China_ABD75609_1_spike_glycoprotein
8_Human_CoV_HK_China_sARU07577_1_spike_protein
5_Chimera_HCoV_HKU1_5I08_A_Chain_A__Spike_glycoprotein__envelope_glycoprotein_chimera
7_Human_CoV_HKU1_HK_China_ABD75545_1_spike_glycoprotein
15_SARS-Related_CoV_6NB6_A_Seattle_USA_Chain_A__Spike_glycoprotein
23_SARS_CoV_BZ_China_AAR07630_1_spike_glycoprotein_BJ302
18_SARS_CoV_BJ_China_AAR07629_1_spike_glycoprotein
17_SARS_CoV_ZJ_China_ACB69894_1_spike_glycoprotein_S__SARS_coronavirus_BJ182-8
24_SARS_CoV_ZJ_China_ACB69905_1_spike_glycoprotein_S_BJ182-12
19_SARS_CoV_ZJ_China_ACB69860_1_spike_glycoprotein_S_BJ182a
20_Civet_SARS_CoV_BJ_China_AAU04662_1_spike_glycoprotein
21_Civet_SARS_BZ_China_AAU04664_1_spike_glycoprotein
22_Bat_CoV_CC_China_AHX37558_1_spike_protein
10_Human_2019nCoV_Wuhan_QHR84449_1_surface_glycoprotein
27_Pangolin_new_alignment_Sirotnik
11_Human_2019nCoV_Wuhan_QHR63250_1_spike_glycoprotein
28_WIV_QHR63300_2_spike_glycoprotein__Bat_coronavirus
29_Nanjian_AVP78042_1_spike_protein__Bat_SARS-like_coronavirus
14_Bat_HKU-3_HK_China_AAZ41340_1_DQ084200__spike_glycoprotein
25_AAY88866_1_spike_glycoprotein__Bat_SARS_coronavirus_HKU3-1
16_Bat_CoV_HKU3-2_AAZ41329_1_DQ084199_HK_China_spike_glycoprotein
26_Recombinant_CoV_NC_USA_ACJ60703_1_spike_glycoprotein
0.10

Figure 1. Phylogenetic tree (NJ) showing based on variable amino acid sites in the Spike protein sequences. Patterns on the right correspond to the presence (black), absence (grey), or length (width) of the discernable function motifs of S1 Spike, She3, KxDL and C-terminal motif (notably Gp41 envelope protein). For clarity the details are also presented in Table 1.

Pathogenicity Motif Fingerprint Search

The visual comparison of the motifs found in SARS-CoV-2 resulted in the discovery that a truncated N-terminal spike receptor binding domain seemed to distinguish sequences most closely related to SARS-CoV-2, compared to most other B-coronaviruses, combined with and missing She3, first recognized in an RNA-binding protein, and KxDL motifs in the middle of the S2 Spike protein segment and C-terminal region Gp41 motif. Using this arrangement as a “pathogenicity fingerprint”, comparisons were made across available sequences. These included three sequences from 2005 of bat-isolated “SARS” sequences (AAY88866, AAZ41340 and AAZ41329), reflected HKU-3-1, HKU-3-2, and HKU-3-3 sequences published by Lau et al. from bat samples taken in Hong Kong, China (Lau et al., 2005). Notably, various synthetic SARS-like coronaviruses were located in the protein database, including numerous sequences by Dr. Shi and colleagues (BJxxx in Table 1). A fully synthetic virus had been constructed from a consensus of sequences from HKU-3-1, HKU-3-2, and HKU-3-3 by researchers in North Carolina, USA (Becker et al., 2008). All synthetic types found were included in detailed motif analysis and in the final “representative” phylogenetic analysis. Any that might be found to match the motif would be included as candidates for a “laboratory” origin hypothesis, but the existence of a historical, natural bat coronavirus with the pathogenic motif fingerprint would make that hypothesis less implausible given the abundant opportunity for exposure to bat coronaviruses in the larger population via the handling, butchering and consumption of captured bats.

In the phylogenetic analysis, the group apparently most similar to 2019-npangolinCoV included the WIV bat coronavirus isolate (QHR63300), the Nanjian “Military Institute” bat coronavirus isolate (AVP78042), the de novo pangolin amino acid sequence (PRJNA573298)), derived from the viriome genome project PRJNA573298 (Liu et al. 2019), and of course published sequences from coronaviruses isolated from patients in Wuhan, China. The closest cluster to that group included the HKU-3-1, HKU-3-2, HKU-3 sequences, all isolated from bats in Hong Kong and published in 2005 (Lau et al., 2005), as well (of course) as the consensus-based synthetic coronavirus from North Carolina, USA (ACJ60703; Becker et al., 2008) that used those three sequences along with a “RP3” sequence (AAZ67052) to create the consensus.

None of the other coronaviruses analyzed had the pathogenic motif, making it potentially useful as a forensic fingerprint. The pathogenic motif was not found in any constructed or modified coronavirus associated with Dr. Shi; she is, in my scientific view, exonerated as far as available and trustworthy data can inform us.

In the larger phylogenetic analysis, pShuttle-SN was found to cluster with SARS sequenced by the Genome Center and is a product of research in the Cancer Center, Sun Yat-Sen University in Guangdong.

The pathogenic motif was partly found in HKU-3-1 and HKU-3-2 from natural isolates from bats from 2005 (Lau et al.,2005) . In contrast, and importantly, the pathogenic motif signature was found fully represented in HKU-3-3, from the same study, with a shortened N-terminal Spike 1 segment and the C-terminal Gp41 motif. Thus there was hidden variation not recognized by in

2005, nor in 2008 by Becker et al. (2008). This would tend to rule out at least a recent synthetic origin of SARS-CoV-2; at this time, such a hypothesis would be extraneous given available data. Like HKU-3-1 and HKU-3-2, the synthetic construct made from the consensus of the four coronaviruses isolated from one 3 HKU spike protein sequences (ACJ60703) has the GP41 C-terminal motif; however, its does not have a truncated N-terminal spike protein segment, and therefore this synthetic construct is also not a viable ancestor for SARS-CoV-2 from a functional perspective.

Table 1. Representative Sequences and Key Characteristic Functional Motifs Recognized

Accession	NTD, She3, KxDL, C-Terminal
AYR18599	Full, Missing, Missing, No
ATP66727	Full, Missing, Missing, No
ATP66756	Full, Missing, Missing, No
AID16649	Full, Missing, Missing, No
ABD75545	Full, Missing, Missing, No
ARU07577	Full, Missing, Missing, No
ABD75529	Full, Missing, Missing, No
ABD75609	Full, Missing, Missing, No
AMN88694	Full, Missing, Missing, No
5I08_A Chain A(Chimera)	Full, Missing, Missing, Fibritin_C
6NB6_A	Full, Missing, Missing, Fibritin_C
BJ182-8	Full, Present, Present, None
BJ182a	Full, Present, Present, None
BJ182-12	Full, Present, Present, None
BJ182-4	Full, Present, Present, None
BJ302 AAR07629.1	Full, Present, Present, None
ACJ60703	Full, Missing, Missing, GP41
AAR07630	Full, Present, Present, None
AH37558	Full, Present, Present, FbpA,Sly,h1A
AAU04664 civet	Full, Present, Present, None
AAU04662	Full, Present, Present, None
Pangolin NCoV-like	Short, Missing, Missing, Gp41
Nanjian_AVP78042	Short, Missing, Missing, Gp41
WUHAN_QHR84449	Short, Missing, Missing, GP41
WUHAN_QHR63250	Short, Missing, Missing, GP41
WIV_QHR63300	Short, Missing, Missing, GP41
Bat HKU3-3 AAZ41340	Short, Missing, Missing, GP41
Bat HKU3-2 AAZ41329	Full, Missing, Missing, GP41
Bat HKU-1 AAY88866	Full, Missing, Missing, GP41

Materials and Methods

Sequences from databases

All data except the Pangolin sequence were downloaded from NCBI's protein database using p-Blast. Various inquiries using Blast (NCBI), pBlast(NCBI) to identify sequences similar to the SARS-CoV-2 were conducted to catalog putatively related sequences. Details of de novo alignment analysis are provided in the Appendix. The Pangolin protein sequence was obtained as the genome project file (Liu et al., 2020; PRJNA573298/) with *de novo* alignment conducted against an SARS-CoV-2 sequence as a reference because the original protein sequence from the project has too many missing base positions in the Spike protein to be useful.

Software

Phylogenetic analysis was conducted using MAFFT, an online analysis service that integrates multiple sequence alignment and phylogenetic analysis with neighbor joining algorithm (Saitou and Nei, 1987) with bootstrapping (Katoh, 2019). MAFFT is a phylogenetic analysis web service that includes options for structure-optimized alignment and sequence inclusion/exclusion. Alignment of FASTA files were conducted using the G-INS-i method. Phylogenetic trees were rendered using MAFFT (NJ, bootstrapping; rate variation among sites alpha = 1.0; N=1000). Trees were visualized by sending the resulting phylogenetic tree directly from MAFFT to the Phylo.io service (Robinson et al., 2016). Motifs were recognized using the Motif Search web application (available at https://www.genome.jp/tools/motif/)

Discussion

There appears to be at least two functionally significant differences in SARS-CoV-2 and similar viruses with matching motif fingerprints. The first difference, reported here for the first time, is, a truncated recognized Spike protein S1. Importantly, the motif is truncated only in the ability of algorithms to detect a previously known motif. Other research supports the idea that this part of the B-coronavirus family tree has explored the human niche via adaptation to utilizing ACE2 as a receptor (Hoffmann et al., 2019) when the S1 protein attaches to the host cell membrane to initiate infection. The second potentially adaptive shift, first noted here, is variation in motifs in S2, which would alter how the virion fuses to cell membrane while acting as a class I viral fusion protein. Alternatively, the variation could also impact the ability of the the viral fusion peptide during cell fusion following S2 cleavage to effect virus endocytosis.

The results of this basic and fundamental analysis rule out all known synthetic SARS-like coronaviruses analyzed to date as plausible ancestors of SARS-CoV-2. It also provides strong support for the hypothesis that SARS-CoV-2, or at least the evidently pathogenic SARS-CoV-2-like Spike protein motif signature, has been isolated from nature and studied at least once in the past, as far back as 2005. The identified motif signature may be extremely useful for quickly "typing" coronaviruses in the lab for pathogenic potential.

An example of the utility of the motif signature can be seen in that in that Becker et al., (2008) been aware of the difference in functional motifs found between HK3-3, which matches nCoV2019, and HK3-1, HK-3-2 and RP3, they may have attempted a different consensus.

Future researchers raising retroviruses from the dead using sequence data from otherwise dead isolates might benefit via the motif fingerprint to enact Level 4 precautions in the event the sequences they are utilizing imply need for greater precaution.

It is reassuring to find that SARS-CoV-2 characteristic pathogenic signature was also found in a 2005 HK-3-3 sequence (AAZ41340). In spite of the existence laboratory generated SARS-CoV-2 retroviruses, the phylogenetic pattern of inheritance most strongly implies that the pathogenic signature and the overall amino acid sequence analysis suggests the motif is inherited via evolution, not recently derived in the laboratory. Low bootstrap values may hint at a more complex story of origin, but at this time, an extra hypothesis of lab-mediated transfer does not appear to be a plausible hypothesis. The transport of animals for research purposes from Hong Kong to Wuhan, or the transport of laboratory materials could either, or both, play a role. The independent confirmation of the proximity of the pangolin sequence provides assurance that the likely source was either bat, or pangolin, but does not by itself support nor immediately require an “intermediate host” hypothesis.

While these findings give a slight edge toward human exposure from an infected pangolin or bat by ruling out recent laboratory recombination, this absolutely does not rule out human infection from a live virus sample in the lab from infected cell lines or from handling field material. The pathogenic lineage may predate modern medicine, and the spike glycoprotein appears to contain a furin-like cleavage site (Coutard et al., 2020), that follows the same phylogenetic distribution as the pathogenic motif.

These results, however, do not rule out prior sensitization of major segments of the population of China via prior mass exposure to this or similar viruses leading to increased morbidity and mortality, as was seen in past studies of Spike-protein based coronavirus vaccine (Agrawal et al., 2016;Deming et al., 2006).

A large number of laboratories around the world are performing a variety of types of studies using recombinant coronaviruses, including research on vaccines. In the past this had led to controversy and to a temporary moratorium by the US on gain-of-function research out of concern of development of viruses in the laboratory that are more dangerous to humans and livestock than viruses found in nature (
https://www.the-scientist.com/news-opinion/lab-made-coronavirus-triggers-debate-34502

Evolution conserves function, but not necessarily sequence. Studies aimed at understanding gain-of-function evolution may do well to conduct analyses that map functional motifs on evolutionary trees or networks as we have. Others have made important discoveries in evolutionary patterns and processes using similar approaches in metazoans (e.g, Ogishima and Tanaka, 2007; Hadzhiev et al., 2007).

Motif analysis of a Fasta file is extremely rapid and requires less intensive analysis than sequence-based phylogenetic analysis. Thus, database annotation should be updated to include predicted motifs upon submission of new sequences or genomes, in particular of the spike protein of B-coronaviruses. Laboratories performing studies characterizing B-coronavirus spike protein tests, typing coronaviruses, and studying efficacy of treatment of B-coronavirus infections using viruses isolated from wild-caught animals, shared reagents, or derived de novo from published sequences would do well to publish the motif pattern present in the Spike protein so others can quickly surmise the applicability of the results of their studies to other B-coronaviruses. This may prove essential in search of origins of SARS-CoV-2 or detection of especially pathogenic B-coronaviruses in future outbreaks.

Literature Cited

Agrawal AS, Tao X, Algaissi A, et al. 2016. Immunization with inactivated Middle East Respiratory Syndrome coronavirus vaccine leads to lung immunopathology on challenge with live virus. Hum Vaccin Immunother. 12(9):2351-6. doi: 10.1080/21645515.2016.1177688.

Becker,M.M., Graham,R.L., Donaldson,E.F., et al. Synthetic recombinant bat SARS-like coronavirus is infectious in cultured cells and in mice. Proc. Natl. Acad. Sci. U.S.A. 105 (50), 19944-19949 (2008)

Coutard, B, C Valle, X de Lamballeri et al., 2020. The spike glycoprotein of the new coronavirus SARS-CoV-2 contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Research 10 Feb 104742

Deming, D. T. Sheahan, M Heise et al. 2006. Vaccine efficacy in senescent mice challenged with recombinant SARS-CoV bearing epidemic and zoonotic spike variants PLoS Med, 3 (12), e525.

Hadzhiev Y, Lang M, Ertzer R 2007. Functional diversification of sonic hedgehog paralog enhancers identified by phylogenomic reconstruction. Genome Biol. 8(6):R106.

Hoffmann M, Kleine-Weber H, Krueger N, Mueller MA, Drosten C, Poehlmann S. The novel coronavirus 2019 (SARS-CoV-2) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. bioRxiv. 2020 Jan 31 [Epub ahead of print]. Available from:
https://www.biorxiv.org/content/10.1101/2020.01.31.929042v1

Ji W, Wang W, Zhao X, Zai J, Li X.2020. Homologous recombination within the spike glycoprotein of the newly identified coronavirus may boost cross-species transmission from snake to human. J Med Virol. 2020 doi: 10.1002/jmv.25682.

Katoh, Rozewicki, Yamada 2019 (Briefings in Bioinformatics 20:1160-1166) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. (explains online service)

Lau,S.K., Woo,P.C., Li,K.S., Huang,Y., Tsoi,H.W., Wong,B.H., Wong,S.S., Leung,S.Y., Chan,K.H. and Yuen,K.Y. 2005. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats Proc. Natl. Acad. Sci. U.S.A. 102 (39), 14040-14045 (2005)

Liu P et al. 2019. Viral Metagenomics Revealed Sendai Virus and Coronavirus Infection of Malayan Pangolins (Manis javanica)., Viruses, 2019 Oct 24;11(11) https://www.ncbi.nlm.nih.gov/bioproject/PRJNA573298/

Liedmann S, Hrincius ER, Guy C. 2014. Viral suppressors of the RIG-I-mediated interferon response are pre-packaged in influenza virions. Nat Commun. Dec 9;5:5645. doi: 10.1038/ncomms6645.

Ogishima S, Tanaka H. 2007. Missing link in the evolution of Hox clusters. Gene. 387(1-2):21-30.

Paraskevis, D, EG Kostaki, G Magiorkinis et al., 2020. Full-genome evolutionary analysis of the novel corona virus (SARS-CoV-2) rejects the hypothesis of emergence as a result of a recent recombination event. Infection, Genetics and Evolution. 79:2020 104212

Robinson, O, D Dylus, C Dessimoz. 2016. Phylo.io : Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web. Molecular Biology and Evolution 33:2136-2166.

Saitou, N and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406-425.

Wu et al., 2020. Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome. NCBI Nucleotide Database, . https://www.ncbi.nlm.nih.gov/nuccore/MN908947