The codon usage pattern of the novel coronavirus is drastically different from those of other pathogenic viruses

*Xiaolong Wang**

College of Life Sciences, Ocean University of China, Qingdao, 266003, P. R. China

Abstract

The current outbreak of a novel coronavirus (COVID-19) has caused thousands of deaths and has been declared to be a worldwide pandemic by the World Health Organization. There have been various disputes but the origin of COVID-19 is not clear. Here we analyzed the similarities of codon usage patterns between humans and pathogenic viruses, such as human immunodeficiency virus (HIV), highly pathogenic avian influenza (HPAI), SARS, MERS, and COVID-19. In HIVs, HPAIs, SARS, and MERS, codon usages are highly similar to that of humans; in contrast, the codon usage pattern of COVID-19 is drastically different from those of humans and other pathogenic viruses. Besides, coronaviruses have been evolving in two opposite directions: human-preferred codons are adopted to substitute less-preferred ones in SARS and MERS but are substituted by less-preferred ones in COVID-19. The unique codon usage pattern suggesting that COVID-19 was evolved in an intermediate host, in which its codon usage pattern becomes drastically different from that of bats or humans, and its pathogenicity is weakened compared with SARS ad MERS COVs. Finally, we appeal to international cooperation to eliminate the epidemic by cutting off the transmission routes among humans and to search for the origin and intermediate hosts of the novel coronavirus to prevent future animal-to-human transmission.

*Xiaolong Wang, College of Life Sciences, Ocean University of China, No. 5 Yushan Road, Qingdao, 266003, Shandong, China, E-mail: Xiaolong@ouc.edu.cn.

1. Background

In the last two decades, three serious epidemics caused by pathogenic coronavirus have emerged, including Severe Acute Respiratory Syndrome (SARS) in 2002-2003 [1], the Middle East Respiratory Syndrome (MERS) in 2012-2015 [2], and the current outbreak of a novel coronavirus (COVID-19). COVID-19 has caused thousands of deaths and hundreds of thousands of hospitalized cases not only in China but at present more seriously in all over the world and has been declared to be a worldwide pandemic by the World Health Organization.

Most pathogenic viruses are of zoonotic origin. For example, human immunodeficiency virus (HIV) was originated from the chimpanzee simian immune deficiency virus (SIVcpz) [3], highly pathogenic avian influenza (HPAI) was originated from bird influenza [4], and pathogenic COVs are originated from a bat coronavirus [5-7]. Lentivirus like HIV inhabits in a host with no symptom for a long period; coronaviruses, such as SARS and MERS, cause severe acute immune responses and respiratory infections in a short period, may cause the death of the host if the viruses are not eliminated by the immune system or medical treatment.

Viral genomes are small in size and largely rely on the host to execute biological activities like replication, protein synthesis, and transmission. After the invasion of a human body, viruses adjust their growth rates and change their pathogenicity/immunogenicity to adapt for a short- or long-term inhabiting in humans. A common strategy for the evolution of viruses is to change the usage of codons, which has strong impacts on viral gene expression and the progress of the pathogenic virus. In 1996, Haas, Park, and Seed reported that the change of codon usage can lead to the inhibition of HIV protein synthesis and the limitation in the expression of HIV-1 envelop glycoprotein [8]. In 2017, Roy, Banerjee, and Basak demonstrated that the rate of substitution in the envelop gene is associated with disease progression [9]. Also, it was suggested that mutational pressure, rather than natural selection for specific coding triplets, is the main determinant of codon usage [10]. In 2013, Moratorio and his colleagues did a comprehensive analysis of the West Nile virus (WNV), which suggested that the genomic biases are the result of the evolution of genome composition, the need to escape the antiviral cell responses and to re-adapt its codon usage to different environments [11].

To analyze the evolutionary characteristics of the novel coronaviruses, we analyzed the codon

usages of SARS, MERS, and COVID-19, determined the changes of the codon usages by compared with those of their most recent common ancestors and those of other pathogenic viruses, including HIVs and HPAIs. The codon usages of the coronaviruses are associated with their high growth rates and severe acute inflammatory responses in humans, provides a theoretical basis for the prediction of possible changes of the pathogenic coronaviruses in the future.

2. Methods

2.1 Genomes sequences

The reference genome sequences and all available complete genome sequences of COVID-19, SARS, MERS, HIV, SIVcpz, and HPAI were downloaded from the NCBI Nucleotide Database during March 1st-16th, 2020. The accession numbers of the reference genome sequences are: NC_004718.3 (SARS), NC_019843.3 (MERS), NC_045512.2 (COVID-19), NC_001802.1 (HIV1), AF115393.1 (SIVcpz), NC_002022.1 (H1N1), NC_007361.1 (H5N1), NC_026422.1 (H7N9) and AF250131.1 (H7N2), respectively.

2.2 Phylogenetic trees

We constructed a multiple sequence alignment of 299 complete coronavirus genomes of using a phylogeny-aware alignment software, PRANK v170427. Maximum likelihood phylogenies were estimated using PhyML v3.115, utilizing the GTR+I+G model of nucleotide substitution with 1,000 bootstrap replicates. The phylogenetic tree was plotted using MEGA v7.0.26 [12].

2.3 Analyze of codon usages

Genes and genomes display a non-random usage of synonymous codons for specific amino acids. A measure of the extent of this non-randomness is given by the relative synonymous codon usage (RSCU), which is calculated as the ratio of the observed frequency of the codons divided by the expected frequency of the same codon if codon usage was uniform within a synonymous codon group [13]. An RSCU value greater than one indicates that the observed frequency of synonymous codons is more preferred compared to the expected frequency [14]. RSCU values of the 59 codons [excluding the single synonymous codons, AUG (Met) and UGG (Trp) and the termination codons, UGA, UAG, and UAA] of all coding gene sequences were calculated using CodonW v 1.4.2.

2.4 Assessment of the distance and the similarity index of codon usages

The relationship among the codon usages of humans and different viruses was calculated using

a squared Euclidean distance method as described by Wei Ji *et al* [15], which is computed as follows:

$$d(\mathrm{H},\mathrm{V}) = \sum_{i=1}^{59}(h_i - v_i)^2$$

where *d* (H, V) represents the distance between the overall codon usage pattern of human and a specific virus, h_i indicates the RSCU value for a particular codon in human, v_i signifies the RSCU value of the same codon for a certain viral gene or genome.

We also used a similarity index of the codon usages, as described by Roy, Banerjee, and Basak [9], to understand the influence of the host genome on the adaptability of the virus genome inside the host. The influence of the overall codon usage pattern of the host on the formation of the codon usage of the virus is defined as the similarity index, which is computed as follows:

$$R(\mathrm{H},\mathrm{V}) = \frac{\sum_{i=1}^{59} h_i \cdot v_i}{\sqrt{\sum_{i=1}^{59} h_i^2 \cdot \sum_{i=1}^{59} v_i^2}}$$

$$D(H,V) = \frac{1 - R(H,V)}{2}$$

where *R* (H, V) represents the degree of similarity between the overall codon usage pattern of human (H) and that of a specific viral gene/genome (V), h_i indicates the RSCU value for a particular codon in human, v_i signifies the RSCU value of the same codon for a certain viral gene/genome. *D* (H, V) represents the potential effect of the overall codon usage of humans on that of the virus. This value ranges from 0.0 to 1.0 and useful for cross-species comparison of codon usages.

2.5 Codon and aa unified sequence alignment

The packaging and the fusion of a virus into a cell rely on their surface/envelop proteins. The spike glycoprotein (S) of COVs and the envelop glycoproteins (GP120) of HIV have become the first choice of the targets in various studies. Here, the *gp120* gene of HIVs and the *s* gene of COVs were aligned by Codon-AA Unified Sequence Alignment (CAUSA v2.1.018) [16]. By comparing with their most recent common ancestors, synonymous and nonsynonymous codon substitutions were found and subject to the analyses of the change of codon preferences.

3. Results

3.1 Phylogenetic analyses

All available complete genome sequences that are related to SARS and COVID-19 viruses were aligned and a genome-wide maximum likelihood phylogenetic tree was established by phyML. As of March 1st, 2020, there were 45 COVID-19 viral genomes deposited in GenBank. As shown in Fig 1a, the overall phylogenetic tree is consistent with those reported earlier [15, 17-19]. COVID-19 share 79.5% identify to SARS-Cov and is 96% identical to a bat coronavirus (RaTG13) at the whole genome level [18], which is identified as the most recent common ancestor of COVID-19 and SARS COVs.

3.2 Codon usages of different viruses and their similarities to that of humans

The RSCU values of different viral genomes were compared with that of humans to assess the influence of the human host in shaping the patterns of codon usage among the viruses. It has been reported that the rate of codon substitution in the envelop gene is associate with disease progression, differs among the three different types of HIV, rapid progressor (RP), slow progressor (SP), and long-term non-progressor (LTNP) of HIV1 infected individuals [9]. Based on the RSCU values for different viruses given by CodonW, the relationship among codon usages of humans and different viruses was calculated using a squared Euclidean distance and a similarity index of codon usages. As shown in Table 1, the codon usage patterns of HIVs are all similar to that of human and that of HPAIs are even more similar to that of humans. The codon usage patterns of SARS and MERS are also highly similar to that of humans, however, COVID-19 has a very special codon usage pattern which is drastically different from that of humans, suggesting that COVID-19 was evolved in an intermediate host, in which its codon usage pattern becomes drastically different from that of bats or humans, and its pathogenicity is significantly weakened compared with SARS ad MERS COVs. Recently, it is reported that the intermediate hosts could be snakes [15] or pangolins [17], but further investigations are needed to validate these speculations.

3.3 The changes of codon usages in the protein-coding genes

Because the sizes of viral genomes are very small, the differences of codon usages could be obscured by noise when they were calculated by counting the number of codons used in the genome sequences. As shown in Fig 2, we performed codon alignments of their surface/envelop proteins, identified synonymous and nonsynonymous codon substitutions, calculate the codon preferences,

and investigated whether codon preferences have been changed in different viruses.

When compared with the *gp120* gene of SIVcpz, the HIV *gp120* gene has 226 synonymous and 279 nonsynonymous substitutions. The average RSCU of the substitutional codons is used as an index for human preference (HPI). As shown in Table 1, HPI decreased in the nonsynonymous substitutions significantly (paired t-test P=0.0301). In contrast, compared with the *s* gene of the bat coronavirus RaTG13, the *s* gene of SARS contains 450 synonymous and 262 nonsynonymous substitutions, while that of COVID-19 contains only 215 synonymous and 29 nonsynonymous substitutions. As shown in Table 1, compared with RaTG13, HPI increased in SARS but decreased in COVID-19.

Besides, compared with SARS, HPI decreased even further in COVID-19 in both synonymous and nonsynonymous substitutions. Although neither of the differences of preference among SARS, COVID-19 and bat COV is statistically significant, the difference of the preference of the synonymous substitutions between COVID-19 and SARS is close to statistically significant (paired t-test P=0.0523). It is clear that coronaviruses are evolving in two opposite directions: in SARS, human-preferred codons are adopted to substitute less-preferred ones; in COVID-19, however, human-preferred codons are abandoned and substituted by less-preferred ones, suggesting that the HPI of COVID-19 has been decreasing since it was isolated from bat COV.

2. Discussion & Conclusion

The above analysis concludes that codon usages have been changed in tested human pathogenic viruses comparing with their ancestors in wild animals. In HIV, HPAIs, SARS, and MERS, codon usages are highly similar to that of humans. In contrast, in COVID-19, hundreds of human-preferred codons were substituted by synonymous codons that are less preferred in humans, making its codon usage patterns drastically different from that of humans.

Moreover, SARS and MERS have an excessive number of highly human-preferred codons, the growth rate of them will be too fast and dysregulated in an infected human body, rob host cells of too many nutrients, energy, and resources. After infection, the fast growth of viruses is the cause of high mortality of the patients, as it may trigger a serve acute response, an inflammatory storm in the human body. Compared with the codon usages of SARS/MERS, the codon usage of COVID-19 is more different from that of humans. On one hand, COVID-19 infection is therefore not as

severe as SARS and MERS, on the other hand, however, as it is much milder than SARS and MERS, COVID-19 is indeed a more successful pathogenic coronavirus and perhaps has greater potential.

Like other pathogenic viruses, the coronaviruses evolve by optimizing the sequence, structure, and functionality of their proteins by changing their codons. If the current epidemic could not be eliminated in a short period, very likely, the coronavirus will develop a chronic disease eventually. It may evolve either into an HIV-like lentivirus or a flu-like self-limiting virus, or both, but they may keep their severe acute pathogenicity and remain to be dangerous for a long period. As a novel pathogenic coronavirus, they are in the early stage of their evolutionary journey in humans. . Finally, we appeal to international cooperation to eliminate the epidemic by cutting off the transmission routes among humans and to search for the origin and intermediate hosts of the novel coronavirus to prevent future animal-to-human transmission.

Data availability

This study conduct data analyses based on existing gene/genome sequences that are available in the NCBI Nucleotide Database and the Global Initiative on Sharing Avian Influenza Data (GISAID) database. The list of GenBank accession numbers of the genome sequences is available online as a text file (AllCoronaVirus.list.txt). The RSCU data for human and viruses are available online as a excel spreadsheet (RSCU-human-viruses.xlsx).

Acknowledgments

This research is funded by the National Natural Science Foundation of China (Grant 31571369). We acknowledge the doctors, nurses, and scientists from all over the world for battling against the epidemic and making the genomic sequences of coronavirus freely and publicly available.

Competing interests

We declare that we have no conflicts of interest.

References

1. Lau, S.K., et al., *Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats.* Proc Natl Acad Sci U S A, 2005. **102**(39): p. 14040-5.
2. Zaki, A.M., et al., *Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia.* N Engl J Med, 2012. **367**(19): p. 1814-20.

3. Rambaut, A., et al., *Human immunodeficiency virus. Phylogeny and the origin of HIV-1.* Nature, 2001. **410**(6832): p. 1047-8.
4. Lee, D.H., et al., *Evolution, global spread, and pathogenicity of highly pathogenic avian influenza H5Nx clade 2.3.4.4.* J Vet Sci, 2017. **18**(S1): p. 269-280.
5. Hon, C.C., et al., *Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus.* J Virol, 2008. **82**(4): p. 1819-26.
6. Mohd, H.A., J.A. Al-Tawfiq, and Z.A. Memish, *Middle East Respiratory Syndrome Coronavirus (MERS-CoV) origin and animal reservoir.* Virol J, 2016. **13**: p. 87.
7. Hu, B., et al., *Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus.* PLoS Pathog, 2017. **13**(11): p. e1006698.
8. Haas, J., E.C. Park, and B. Seed, *Codon usage limitation in the expression of HIV-1 envelope glycoprotein.* Curr Biol, 1996. **6**(3): p. 315-24.
9. Roy, A., R. Banerjee, and S. Basak, *HIV Progression Depends on Codon and Amino Acid Usage Profile of Envelope Protein and Associated Host-Genetic Influence.* Front Microbiol, 2017. **8**: p. 1083.
10. Shackelton, L.A., C.R. Parrish, and E.C. Holmes, *Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses.* J Mol Evol, 2006. **62**(5): p. 551-63.
11. Moratorio, G., et al., *A detailed comparative analysis of the overall codon usage patterns in the West Nile virus.* Infect Genet Evol, 2013. **14**: p. 396-400.
12. Kumar, S., G. Stecher, and K. Tamura, *MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.* Mol Biol Evol, 2016. **33**(7): p. 1870-4.
13. Sharp, P.M. and W.H. Li, *An evolutionary perspective on synonymous codon usage in unicellular organisms.* J Mol Evol, 1986. **24**(1-2): p. 28-38.
14. dos Reis, M., L. Wernisch, and R. Savva, *Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome.* Nucleic Acids Res, 2003. **31**(23): p. 6976-85.
15. Ji, W., et al., *Cross-species transmission of the newly identified coronavirus 2019-nCoV.* J Med Virol, 2020. **92**(4): p. 433-440.
16. Wang, X. and C. Yang, *CAUSA 2.0: accurate and consistent evolutionary analysis of proteins using codon and amino acid unified sequence alignments.* PeerJ PrePrints **3**.
17. Lam, T.T.-Y., et al., *Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China.* bioRxiv, 2020: p. 2020.02.13.945485.
18. Zhou, P., et al., *A pneumonia outbreak associated with a new coronavirus of probable bat origin.* Nature, 2020. **579**(7798): p. 270-273.
19. Matsuda, T., H. Suzuki, and N. Ogata, *Phylogenetic analyses of the severe acute respiratory syndrome coronavirus 2 reflected the several routes of introduction to Taiwan, the United States, and Japan*. 2020: arXiv:2002.08802 [q-bio.GN].

Table 1. Codon usages of different type of virus and their distance/similarity to that of humans

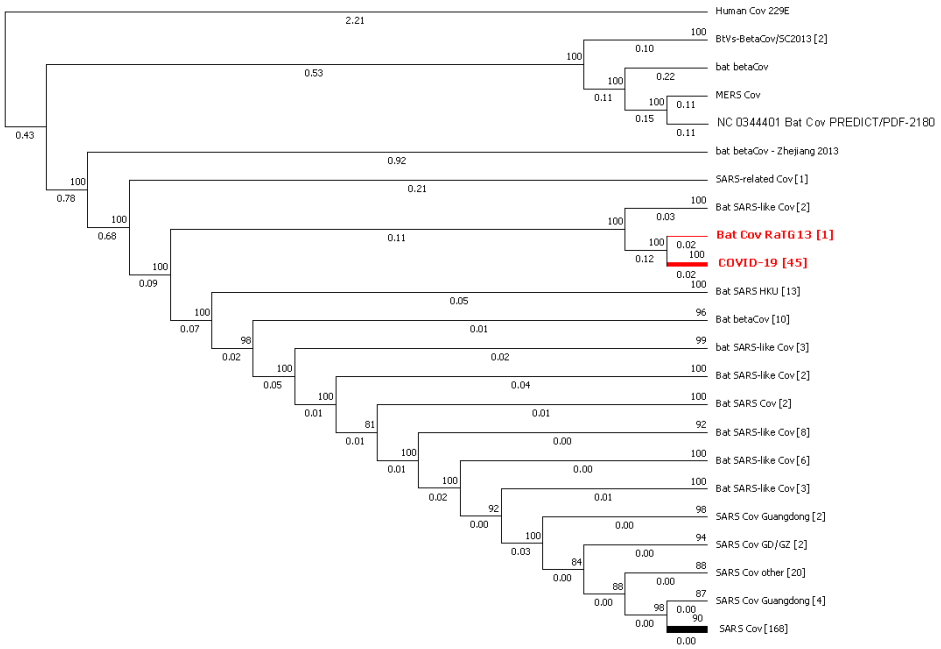
Virus	Strain / Type	*d (H, V)*	*R (H, V)*	*D (H, V)*
SIV	SIVcpz	27.1721	0.8263	0.0868
HIVs	RP	27.6765	0.8176	0.0912
	SP	28.2044	0.8143	0.0929
	LTNP	28.3472	0.8145	0.0927
HPAIs	H5N1	12.8378	0.9116	0.0442
	H7N9	13.0191	0.9069	0.0466
	H1N1	13.0251	0.9066	0.0467
	H7N2	14.4611	0.8972	0.0514
COVs	MERS	26.1475	0.8301	0.0850
	SARS	27.9605	0.8245	0.0878
	COVID-19	39.3928	0.7636	0.1182

1 Table 2. Codon substitutions and the average RSCU frequencies of codons of the envelop or surface protein
2

Compare	Type of codon Substitution	Number of Codon Substitutions	Human Preference Index (HPI) (Average RSCU Frequency Per Thousand)		
			SIVcpz	HV1	P-value
HIV1 vs SIV	Synonymous	226	17.7783	18.2407 ↑	0.2877
	Nonsynonymos	279	18.4039	17.1276 ↓	0.0301*
			Bat Cov	SARS	P-value
SARS vs Bat Cov	Synonymous	450	16.7153	17.0744 ↑	0.1856
	Nonsynonymos	262	16.8462	17.2905 ↑	0.2242
			Bat Cov	Cov-2019	P-value
Covid-19 vs Bat Cov	Synonymous	215	17.7167	17.2065 ↓	0.2052
	Nonsynonymos	29	16.9897	15.6690 ↓	0.2200
			SARS	COVID-9	P-value
Covid-19 vs SARS	Synonymous	435	17.2218	16.5579 ↓	0.0523
	Nonsynonymos	265	17.2400	16.6743 ↓	0.1710

3
4

A



B

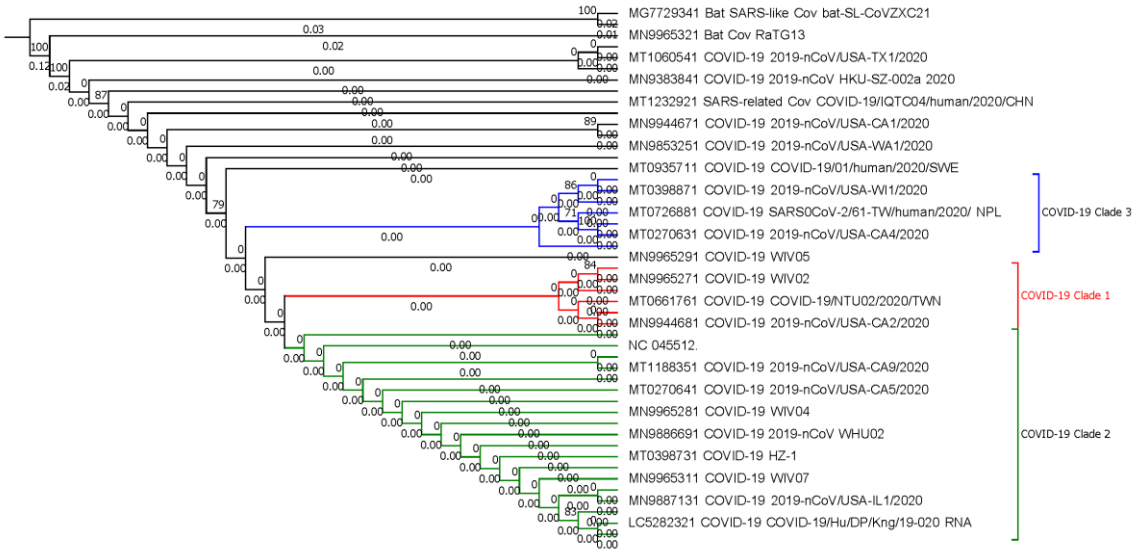
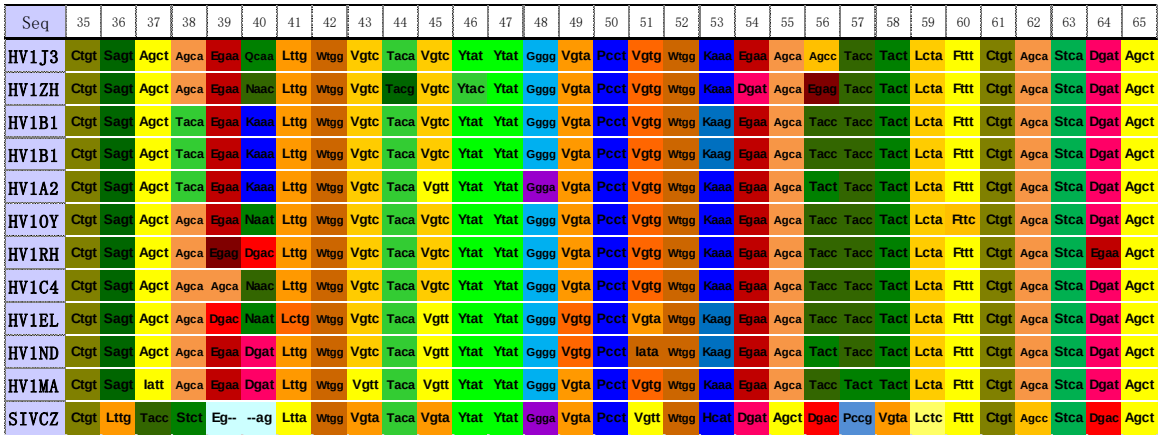


Fig 1. The molecular phylogenetic tree of coronavirus. (A) The phylogenetic tree of all SARS-related coronavirus; (B) the subtree of the novel coronavirus (COVID-19) in A. The analysis base on a PRANK alignment of 299 complete coronavirus genome sequences. All positions containing gaps and missing data were eliminated. The evolutionary history was inferred by using the Maximum Likelihood method by phyML. The tree with the highest log is shown. Trees were plotted in MEGA7. Root was placed on the most distant branch, human COV 229E. The bootstrap percentage of trees in which the associated taxa clustered together is shown next to the branches. The distances between the taxa are shown in the middle of the branches.

A



B

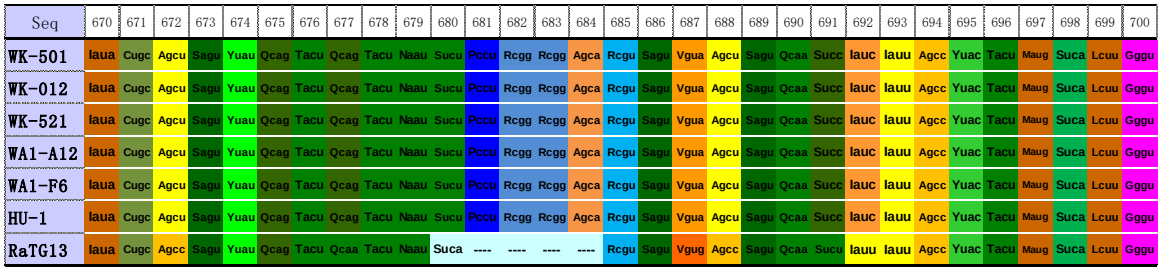


Fig 2. Codon and aa unified view of the codon alignments: (A) the codon alignment of HIV/SIV envelop protein gene; HV1J3-HV1MA: HIV strains, SIVCZ: chimpanzee SIV; (B) the codon alignment of the spike protein gene of coronaviruses. WK-501, WK-012, WK-521, WA1-A12, WA1-F6, HU-1: COVID-19 isolates; RaTG13: a bat coronavirus (MN996532.1) which is identified as the most recent common ancestor of COVID-19 and SARS COVs. Uppercase: amino acids; lowercase: nucleotides.