Wei Ji ORCID iD: 0000-0002-0818-5578

Homologous recombination within the spike glycoprotein of the newly identified coronavirus may boost cross-species transmission from snake to human.

Wei Ji[1, *, #], Wei Wang[2, *], Xiaofang Zhao[3, *], Junjie Zai [4, *], Xingguang Li[5, *]

1. Department of Microbiology, Peking University Health Science Center School of Basic Medical Sciences, Beijing, China; jiwei_yunlong@126.com

2. Department of Spleen and Stomach Diseases, The First affiliated Hospital of Guangxi university of Chinese Medicine, Nanning 530023, China.

3. Department of Science and Technology, Ruikang Hospital Affiliated to Guangxi University of Chinese Medicine, Nanning 530011, China

4. Immunology innovation team, School of Medicine, Ningbo University, Ningbo 315211, China.

5. Hubei Engineering Research Center of Viral Vector, Wuhan University of Bioengineering, Wuhan 430415, China.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jmv.25682.

*These authors contributed equally to this work.

Keywords

2019-nCoV; phylogenetic analysis; recombination; RSCU; cross-species transmission

Abstract

The current outbreak of viral pneumonia in the city of Wuhan, China, was caused by a novel coronavirus designated 2019-nCoV by the World Health Organization, as determined by sequencing the viral RNA genome. Many patients were potentially exposed to wildlife animals at the Huanan seafood wholesale market, where poultry, snake, bats, and other farm animals were also sold. To determine the possible virus reservoir, we have carried out comprehensive sequence analysis and comparison in conjunction with relative synonymous codon usage (RSCU) bias among different animal species based on existing sequences of the newly identified coronavirus 2019-nCoV. Results obtained from our analyses suggest that the 2019-nCoV appears to be a recombinant virus between the bat coronavirus and an origin-unknown coronavirus. The recombination occurred within the viral spike glycoprotein, which recognizes cell surface receptor. Additionally, our findings suggest that snake is the most probable wildlife animal reservoir for the 2019-nCoV based on its RSCU bias resembling snake compared to other animals. Taken together, our results suggest that homologous recombination within the spike glycoprotein may contribute to cross-species transmission from snake to humans.

Introduction

China has been the epicenter of emerging and reemerging viral infections that continue to stir a global concern. In the last 20 years, China has witnessed several emerging viral diseases, including an avian influenza in 1997[1], the severe acute respiratory syndrome (SARS) in 2003[2], and a severe fever with thrombocytopenia syndrome (SFTS) in 2010[3]. The most recent crisis was the outbreak of an ongoing viral pneumonia with unknown etiology in the city of Wuhan, China. On December 12, 2019, Wuhan Municipal Health Commission (WMHC) reported 27 cases of viral pneumonia with 7 of them being critically ill. All of them had history of exposure linked to the Huanan Seafood Wholesale Market where also sold poultry, bats, and snakes[4]. On January 3rd, 2020, WMHC updated the number of cases to a total of 44 with 11 of them in critical condition. On January 5th, the number of cases increased to 59 with 7 critically ill patients. The viral pneumonia outbreak was not caused by severe acute respiratory syndrome coronavirus (SARS-CoV), Middle East Respiratory Syndrome coronavirus (MERS-CoV), influenza virus, or adenovirus as determined by laboratory tests[4]. On January 10, it was reported that a novel coronavirus designated 2019-nCoV by the World Health Organization (WHO)[5] was identified by sequencing the viral RNA genome, which was released through virological.org. More significantly, the newly identified virus has also been isolated from one patient. The determination of viral RNA sequence has made it possible to develop reverse-transcription polymerase chain reaction (RT-PCR) methods for detection of viral RNA in samples from patients and potential hosts[6]. As a result, 217 patients were confirmed to be infected with the newly identified coronavirus 2019-nCoV, and three patients died as of January 20, 2020. Several patients travelling from Wuhan were also reported in Thailand, Singapore, Hong

Kong, South Korea and Japan. A total of 14 full-length sequences of the novel coronavirus were released to GISAID.

The *coronavirinae* family consists of four genera based on their genetic properties, including genus *Alphacoronavirus*, genus *Betacoronavirus*, genus *Gammacoronavirus*, and genus *Deltacoronavirus*[7]. The coronavirus RNA genome (ranging from 26 to 32 kb) is the largest among all RNA viruses[8]. Coronavirus can infect mammals, birds and reptile, including human, swine, cattle, horses, camels, cats, dogs, rodents, birds, bats, rabbits, ferrets, mink, snake, and various wildlife species[7,9]. Many coronavirus infections are subclinical[7,9]. The severe acute respiratory syndrome coronavirus (SARS-CoV) and the Middle East respiratory syndrome coronavirus(MERS-CoV) belong to the *Betacoronavirus* genus and are zoonotic pathogens that can cause severe respiratory diseases in humans[7].

The outbreak of viral pneumonia in Wuhan is associated with exposures at the Huanan seafood wholesale market, suggesting a possible zoonosis. The seafood market also sold live animals such as snakes, marmots, bats, birds, frogs, hedgehogs and rabbit. There is no evidence suggesting a specific wildlife host as virus reservoir. Studies of relative synonymous codon usage (RSCU) bias between viruses and their hosts suggested that viruses tends to evolve codon usage bias that are comparable to their hosts[10,11]. To search for potential virus reservoir, we have carried out a comprehensive sequence analysis and comparison. Results from our analysis suggest that snake is the most likely wildlife reservoir responsible for the current outbreak of 2019-nCoV infection. More interestingly, an origin-unknown homologous recombination was identified within the spike glycoprotein of the 2019-nCoV[5], which may explain its decreased pathogenesis, snake-to-human cross species transmission, and limited person-person spread.

Materials and Methods

Sequence data collection

Six newly sequenced Beta-coronavirus (Wuhan 2019-2020) genomes were downloaded from the Global Initiative on Sharing Avian Influenza Data (GISAID) database, including BetaCoV/Wuhan/IVDC-HB-01/2019, EPI_ISL_402119; BetaCoV /Wuhan/IVDC-HB-04/2020, EPI_ISL_402120; BetaCoV/Wuhan/IVDC-HB-05/2019, EPI_ISL_402121; BetaCoV/Wuhan/IPBCAMS-WH-01/2019, EPI_ISL_402123; Beta CoV/Wuhan-Hu-1/2019, EPI_ISL_402125, and BetaCoV/Wuhan/IVDC-HB-05/2019. It appeared that the BetaCoV/Wuhan/IVDC-HB-05/2019 was incorrect and therefore not included for the subsequent analysis. Five hundred closely related sequences were downloaded from GenBank. Out of them, 271 genome sequences (>19,000bp in length) were used in this study together with the above-described 5 Beta-coronavirus (2019-nCoV) genome sequences (Supplementary Table S1). The geographic origins of the sequences were from Bulgaria (*n*=1), Canada (*n*=2), China (n=71), Germany (n=1), Hong Kong (n=5), Italy (n=1), Kenya (n=1), Russia (n=1), Singapore (n=24), South Korea (n=1), Taiwan (n=11), United Kingdom (n=2), United States of America (*n*=67), and Unknown (n=88). The host gene origins include Bat (n=47), Civet (n=9), Human (n=27), Monkey (n=1), Mouse (n=13), and Unknown (n=179). Sequences were aligned using MAFFT v7.222[12], followed by manual adjustment using BioEdit v7.2.5[13].

Phylogenetic and recombination analysis

Phylogenetic trees were constructed using maximum-likelihood methods and general time-reversible model of nucleotide substitution with gamma-distributed

rates among sites (GTR+G substitution model) in RAxML v8.0.9[14]. Support for the inferred relationships was evaluated by a bootstrap analysis with 1000 replicates and trees were midpoint-rooted. Genetic distances between the 5 identified coronavirus (2019-nCoV) sequences were calculated with MEGA v7.1.18[15] using the maximum composite likelihood with 1000 bootstrap replicates.

To investigate the putative parents of the 2019-nCoV, we performed Similarity and Bootscanning plot analyses based on the Kimura two-parameter model with window size of 500 bp, step size of 30 bp using SimPlot v.3.5.1[16]. We divided our data set into 4 clades, the 5 newly discovered 2019-nCoV sequences were grouped as the query sequences. The closest relative coronaviruses (bat-SL-CoVZC45 and bat-SL-CoVZXC21) obtained from the city of Nanjing, China were grouped as 'Clade A'. The other two coronaviruses (BtCoV/BM48-31/BGR/2008 and BtKY72) from Bulgaria and Kenya were grouped as 'Clade B'. The rest sequences were grouped as 'Clade C' (Fig. 1).

Synonymous Codon Usage Analysis

To estimate the relative synonymous codon usage (RSCU) of the 2019-nCoV and its potential hosts, coding sequences of the 2019-nCoV-WIV04 genome (9711 codons) from GISAID, bat-SL-CoVZC45 genome (9680 codons), Bungarus multicinctus genes (32789 codons), Naja Atra genes (30270 codons), Erinaceus europaeus genes (16967352 codons), Marmota genes (21276356 codons), manis javanica genes (24183590 codons), Rhinolophus sinicus genes (28430 codons) and Gallus gallus genes (314483 codons) from GenBank were calculated with Codon W1.4.2 [17,18]. The RSCU of human genes (40662582 codons) was retrieved from the Codon Usage Database (http://www.kazusa.or.jp/codon/). The relationship among these sequences was calculated using a squared euclidean distance ($d_{ik} = \sum_{j=1}^{p}(X_{ij} - X_{kj})^2$), as we previously reported[19]. A heat map of

RSCU was drawn with MeV 4.9.0 software[20]. The Coronaviruses and their potential hosts were clustered using a Euclidean distance method.

Results

Phylogenetic classification

Phylogenetic analysis of 276 coronavirus genomes revealed that the 5 newly identified coronavirus (2019-nCoV) sequences were monophyletic with 100% bootstrap support. The Clade A (bat-SL-CoVZC45 and bat-SL-CoVZXC21) derived from bats in the city of Nanjing, China between 2015-2017 represents the sister lineage to 2019-nCoV. The Clade B (BtCoV/BM48-31/BGR/2008 and BtKY72) obtained from bats in Bulgaria and Kenya between 2005-2007 formed a distinct monophyletic cluster with 100% bootstrap support. The Clade C including 267 coronavirus strains was clustered together with 63% bootstrap support (Fig. 1).

The estimated genetic diversity of the 5 newly identified coronavirus genomes (2019-nCoV) was 0.000094 substitutions per site with an estimated evolutionary rate of 0.0038 substitutions per site per year based on 9 days of sampling interval and 29865bp in length for the 2019-nCoV. We estimated that the time of the most recent common ancestor (TMRCA) of the 2019-nCoV was 736.4 days (2.02 years). These results suggested that the newly identified 5 coronavirus (2019-nCoV) sequences were originated from the same isolate about 2 years ago.

Homologous recombination within the spike glycoprotein may boost Cross-species transmission.

Homologous recombination is an important evolutionary force and previous studies have found that homologous recombination occurred in many viruses, including Dengue virus[21], human immunodeficiency virus[22], hepatitis B virus[23], hepatitis C virus[24] and classical swine fever virus[19]. Similarity plot analysis of the 2019-nCoV revealed that homologous recombination occurred between Clade A strains (bat-coronaviruses) and the origin-unknown isolates in 21500-24000bp, located within the spike glycoprotein that recognizes cell surface receptor (Fig. 2). These characteristics indicates that homologous recombination within the spike glycoprotein may boost the 2019-nCoV cross-species spread to humans.

Snakes as the most likely wildlife reservoir for the 2019-nCoV.

As parasitic microorganism, virus codon usage pattern resembles its host to some extent. The RSCU shows that the 2019-nCoV, bat-SL-CoVZC45, and snakes from China have similar synonymous codon usage bias (Fig. 3A, Table1). The squared euclidean distance indicates that the 2019-nCoV and snakes from China have the highest similarity in synonymous codon usage bias compared to those of marmota, hedgehog, manis, bat, bird and human (Fig. 3B). Two types of snakes, containing Bungarus multicinctus (Many-banded krait) and Naja Atra (Chinese cobra) were used for RSCU analysis. Squared euclidean distance between the 2019-nCoV and Bungarus multicinctus is 12.77. The distance between the 2019-nCoV and another snake Naja Atra is 14.70. However, the distance between the 2019-nCoV and other animals is greater than 24, specifically 24.87 for marmota, 25.92 for hedgehog, 26.81 for manis, 27.47 for bat, 29.07 for bird, and 35.44 for human. These data suggest that the 2019-nCoV can more effectively use

snake's translation machinery than that of other animals, suggesting that the epidemic 2019-nCoV might originate from snakes.

Two types of snakes are common in the Southeastern China including the city of Wuhan (Fig. 4). Geographical distributions of Bungarus multicinctus include Taiwan, the Central and Southern China, Hong Kong, Myanmar (Burma), Laos, and Northern Vietnam[25]. Naja Atra is found in Southeastern China, Hong Kong, Northern Laos, Northern Vietnam, and Taiwan[26]. Snakes were also sold in the Huanan Seafood Wholesale Market where many patients worked or has exposed to wildlife or farm animals. Taken together, snakes could be the most likely wildlife animal reservoir for the 2019-nCoV.

Discussion

In this study, we have performed an evolutionary analysis using 276 genomic sequences of coronaviruses obtained from various geographic locations and host species. Our results show that the novel coronavirus sequences obtained from the viral pneumonia outbreak occurring in the city of Wuhan form a separate group that is highly distinctive to SARS-CoV. The SARS-CoV first emerged in China in 2002 and then spread to 37 countries/regions in 2003 and caused a travel-related global outbreak with 9.6% mortality rate [27]. More importantly, results from our analysis reveal a homologous recombination occurred between the bat coronavirus and an origin-unknown coronavirus in the nucleotides 21500-24000 within the spike glycoprotein gene. Sequence homology analysis of the partial spike glycoprotein genes (1-783bp) from the 2019-nCoV was done through BLAST at the NCBI website. Interestingly, no similar sequence was found with known sequence database, suggesting that a putative recombination parent virus was still unknown. Previous study suggested that recombination of SARS in the spike

glycoprotein genes might have mediated the initial cross-species transmission event from bats to other mammals[28]. Bootscanning plot analysis (data not shown) indicted that the major parents of the 2019-nCoV originated from Clade A (bat-SL-CoVZC45 and bat-SL-CoVZXC21) but formed a monophyletic cluster different from them. A possible explanation for the monophyly could be insertion of the spike glycoprotein regions of another parent lineage into Clade A. Overall, the ancestral origin of the 2019-nCoV was more likely from divergent host species rather than SARS-CoV.

The host range of some animal coronaviruses was promiscuous[7]. They caught our attention only when they caused human diseases such as SARS, MERS and 2019-nCoV pneumonia[4,9,29]. It is critical to determine the animal reservoir of the 2019-nCoV in order to understand the molecular mechanism of its cross-species spread. Homologous recombination within viral structural proteins between coronaviruses from different hosts may be responsible for "cross-species" transmission[28]. Information obtained from RSCU analysis provides some insights to the question of wildlife animal reservoir although it requires further validation by experimental studies in animal models. Currently, the 2019-nCoV has not been isolated from animal species although it was obtained from one patient. Identifying and characterizing the animal reservoirs for 2019-nCoV can help determine the recombination parent sources and for a better understand the potential spread of infection in human populations.

The 2019-nCoV has caused a total of 217 confirmed cases of pneumonia in China as of January 20, 2020 with new patients also reported in Hong Kong, Thailand, Singapore, South Korea and Japan. Unlike SARS-CoV, the 2019-nCoV appeared to cause mild form of viral pneumonia and have limited capability for person-person spread. This could be due to the recombination occurred within the

receptor-binding glycoprotein, virus originated from snake, or both. Snakes are cold-blooded reptiles with lower temperature than humans[30]. Accordingly, the 2019-nCoV will likely be attenuated upon infection to humans. However, there is a concern about its adaptation in humans that may acquire the capability to replicate more efficiently and spread more rapidly via close person-person contact.

In summary, results derived from our sequence analysis suggest for the first time that snake is the most likely wildlife animal reservoir for the 2019-nCoV based on their similar codon usage bias. Additionally, our analyses also identified a homologous recombination within the receptor-binding spike glycoprotein, which may determine cross-species transmission from snake to humans. These novel findings warrant future investigation to experimentally determine if snake serves as the 2019-nCoV reservoir and the homologous recombination within the spike glycoprotein determine the tropism of the 2019-nCoV in viral transmission and replication. New information obtained from our evolutionary analysis is highly significant for effective control of the outbreak caused by the 2019-nCoV-induced pneumonia.

References

1 Chan, P. K. Outbreak of avian influenza A (H5N1) virus infection in Hong Kong in 1997. *Clinical Infectious Diseases* **34**, S58-S64 (2002).

2 Guan, Y. *et al.* Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* **302**, 276-278 (2003).

3 Yu, X. J. *et al.* Fever with thrombocytopenia associated with a novel bunyavirus in China. *N Engl J Med* **364**, 1523-1532, doi:10.1056/NEJMoa1010095 (2011).

4 Lu, H., Stratton, C. W. & Tang, Y. W. Outbreak of Pneumonia of Unknown Etiology in Wuhan China: the Mystery and the Miracle. *J Med Virol*, doi:10.1002/jmv.25678 (2020).

5 WHO. *Coronavirus*, <https://www.who.int/health-topics/coronavirus> (2019).

6 Zhang, N. *et al.* Recent advances in the detection of respiratory virus infection in humans. *J Med Virol*, doi:10.1002/jmv.25674 (2020).

7 MacLachlan, N. J. & Dubovi, E. J. in *Fenner's Veterinary Virology (Fifth Edition)* (eds N. James MacLachlan & Edward J. Dubovi) 393-413 (Academic Press, 2017).

8 Howley, D. M. K. P. *Fields virology*. Vol. 1 830P (Lippincott Williams & Wilkins (LWW), 2013).

9 Organization, W. H. Consensus document on the epidemiology of severe acute respiratory syndrome (SARS). (Geneva: World Health Organization, 2003).

10 Wang, H., Liu, S., Zhang, B. & Wei, W. Analysis of synonymous codon usage bias of Zika virus and its adaption to the hosts. *PLoS One* **11**, e0166260 (2016).

11 Bahir, I., Fromer, M., Prat, Y. & Linial, M. Viral adaptation to host: a proteome based analysis of codon usage and amino acid preferences. *Molecular systems biology* **5** (2009).

12 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

13 Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**, 95-98, doi:citeulike-article-id:691774 (1999).

14 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).

15 Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular biology and evolution* **33**, 1870-1874, doi:10.1093/molbev/msw054 (2016).

16 Lole, K. S. *et al.* Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J Virol* **73**, 152-160 (1999).

17 Liu, X., Zhang, Y., Fang, Y. & Wang, Y. Patterns and influencing factor of synonymous codon usage in porcine circovirus. *Virology journal* **9**, 1-9 (2012).

18 Wright, F. The 'effective number of codons' used in a gene. *Gene* **87**, 23-29 (1990).

19 Ji, W., Niu, D.-D., Si, H.-L., Ding, N.-Z. & He, C.-Q. Vaccination influences the evolution of

classical swine fever virus. *Infection, Genetics and Evolution* **25**, 69-77 (2014).

20 Howe, E. *et al.* in *Biomedical Informatics for Cancer Research* (ed Michael F. OchsJohn T. CasagrandeRamana V. Davuluri) 267-277 (2010).

21 Tolou, H. *et al.* Evidence for recombination in natural populations of dengue virus type 1 based on the analysis of complete genome sequences. *Journal of General Virology* **82**, 1283-1290 (2001).

22 Clavel, F. *et al.* Genetic recombination of human immunodeficiency virus. *Journal of virology* **63**, 1455-1459 (1989).

23 Bollyky, P. L., Rambaut, A., Harvey, P. H. & Holmes, E. C. Recombination between sequences of hepatitis B virus from different genotypes. *Journal of Molecular Evolution* **42**, 97-102 (1996).

24 Colina, R. *et al.* Evidence of intratypic recombination in natural populations of hepatitis C virus. *Journal of general virology* **85**, 31-37 (2004).

25 wikipedia. *Many-banded krait*, <https://en.wikipedia.org/wiki/Many-banded_krait> (2020).

26 wikipedia. *Chinese cobra*, <https://en.wikipedia.org/wiki/Chinese_cobra> (2020).

27 Smith, R. D. Responding to global infectious disease outbreaks: lessons from SARS on the role of risk perception, communication and management. *Social science & medicine* **63**, 3113-3123 (2006).

28 Graham, R. L. & Baric, R. S. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. *Journal of virology* **84**, 3134-3146 (2010).

29 Cunha, C. B. & Opal, S. M. Middle East respiratory syndrome (MERS) A new zoonotic viral pneumonia. *Virulence* **5**, 650-654 (2014).

30 Hayes, R. O., Daniels, J. B., Maxfield, H. K. & Wheeler, R. E. Field and laboratory studies on eastern encephalitis in warm-and cold-blooded vertebrates. *The American journal of tropical medicine and hygiene* **13**, 595-606 (1964).

Acknowledgements

This work was supported by Project of Guangxi Health Committee (No. Z20191111) and Natural Science Foundation of Guangxi Province of China (No. 2017GXNSFAA198080) to Dr. Xiaofang Zhao. This study was sponsored by K.C. Wong Magna Fund in Ningbo University.

We gratefully acknowledge the Originating and Submitting Laboratories for sharing newly identified coronavirus sequences through GISAID, as follows:

1 EPI_ISL_402119, EPI_ISL_402120, EPI_ISL_402121:

Originating and submitting lab - National Institute for Viral Disease Control and Prevention, China CDC

Authors - Wenjie Tan, Xiang Zhao, Wenling Wang, Xuejun Ma, Yongzhong Jiang, Roujian Lu, Ji Wang, Weimin Zhou, Peihua Niu, Peipei Liu, Faxian Zhan, Weifeng Shi, Baoying Huang, Jun Liu, Li Zhao, Yao Meng, Xiaozhou He, Fei Ye, Na Zhu, Yang Li, Jing Chen, Wenbo Xu, George F. Gao, Guizhen Wu

2 EPI_ISL_402123:

Originating and submitting lab - Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College

Authors - Lili Ren, Jianwei Wang, Qi Jin, Zichun Xiang, Yongjun Li, Zhiqiang Wu, Chao Wu, Yiwei Liu

3 EPI_ISL_402124:

Originating lab - Wuhan Jinyintan Hospital

Submitting lab - Wuhan Institute of Virology, Chinese Academy of Sciences

Authors - Peng Zhou, Xing-Lou Yang, Ding-Yu Zhang, Lei Zhang, Yan Zhu, Hao-Rui Si, Zhengli Shi

4 EPI_ISL_402125:

Originating lab - Unknown

Submitting lab - National Institute for Communicable Disease Control and Prevention (ICDC) Chinese Center for Disease Control and Prevention (China CDC)

Authors - Zhang,Y.-Z., Wu,F., Chen,Y.-M., Pei,Y.-Y., Xu,L., Wang,W., Zhao,S., Yu,B., Hu,Y., Tao,Z.-W., Song,Z.-G., Tian,J.-H., Zhang,Y.-L., Liu,Y., Zheng,J.-J., Dai,F.-H., Wang,Q.-M., She,J.-L. and Zhu,T.-Y.

Author contributions

Conceptualization: Wei Ji

Writing: Wei Ji, Xingguang Li

Data collection: Junjie Zai, Wei Wang, Xiaofang Zhao

Data analysis: Wei Ji, Xingguang Li

Competing financial interests

The authors declare no competing financial interests.

Figure legends

Figure 1. Maximum likelihood phylogenetic tree of the 5 newly identified coronavirus genome sequences (2019-nCoV).

Phylogenetic tree inferred from 276 near-complete genome sequences of coronavirus was midpoint rooted and grouped into 4 clades (2019-nCoV, Clades A, B, and C). Tips are color labeled with sampling location.

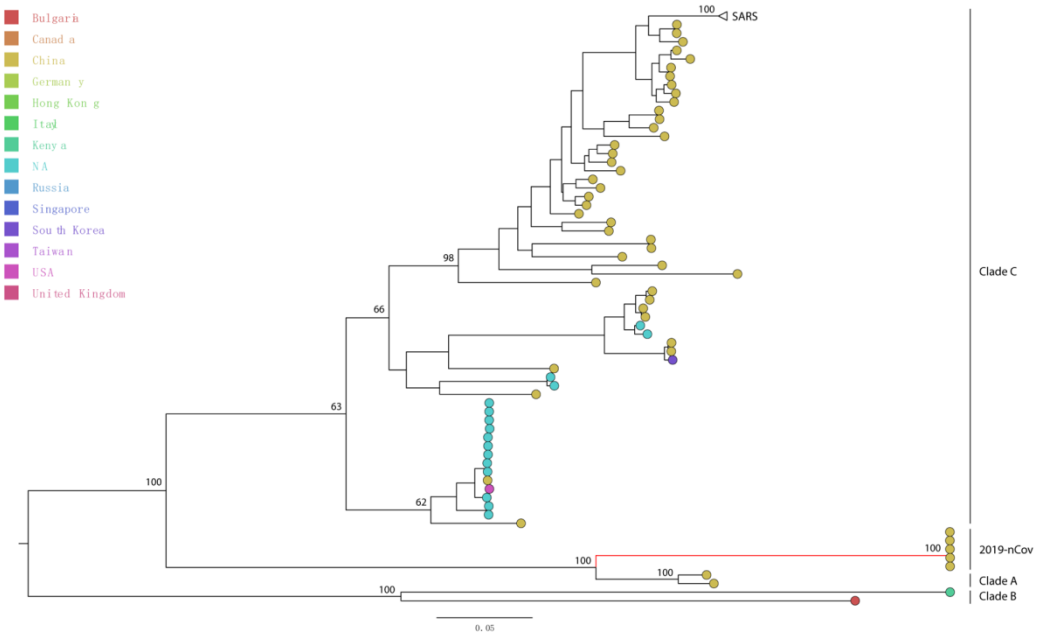

Figure 2. Recombinant analyses of the 5 newly identified coronavirus genome sequences (2019-nCoV).

Similarity plot analyses were performed with Clades A, B, and C. The recombinant analyses were performed with a sliding window of 500 bp and a step size of 30 bp. Recombination sites were located within spike glycoprotein genes, shown in Orange box at top.

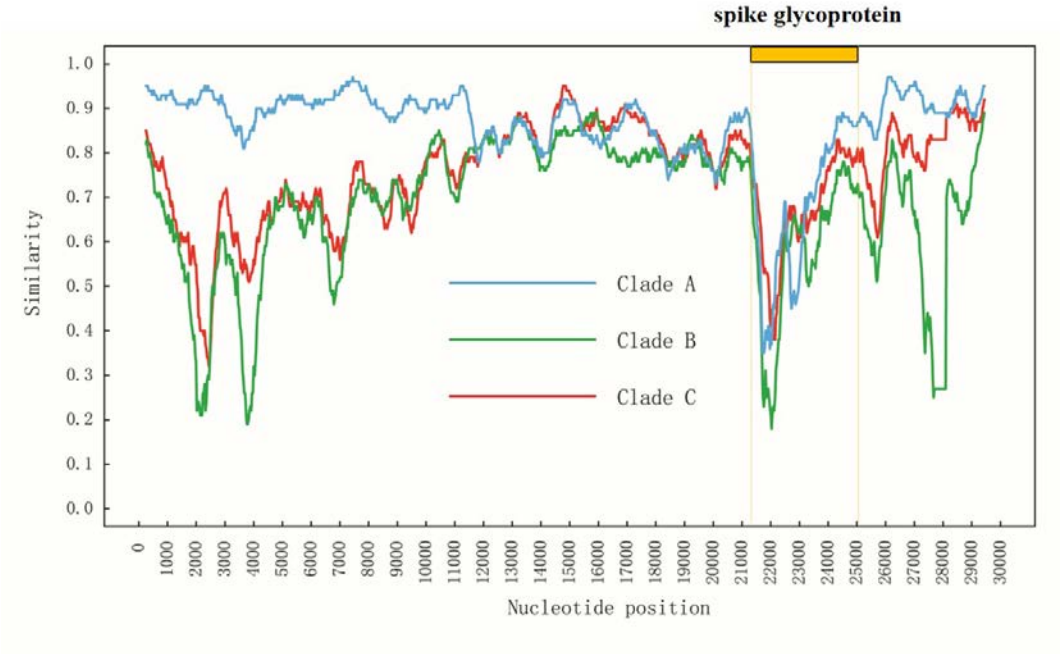

Figure 3 Comparison of relative synonymous codon usage between 2019-nCoV and its putative wildlife reservoir hosts. (A) Heat map shows Cluster analysis of RSCU from 2019-nCoV-WIV04, bat-SL-CoVZC45, Bungarus multicinctus, Naja atra, Marmota, Erinaceus europaeus, manis javanica, Rhinolophus sinicus, Gallus gallus, Homo sapiens. (B) Squared euclidean distance was used to compare Codon bias divergence between 2019-nCoV-WIV04 and its putative hosts.

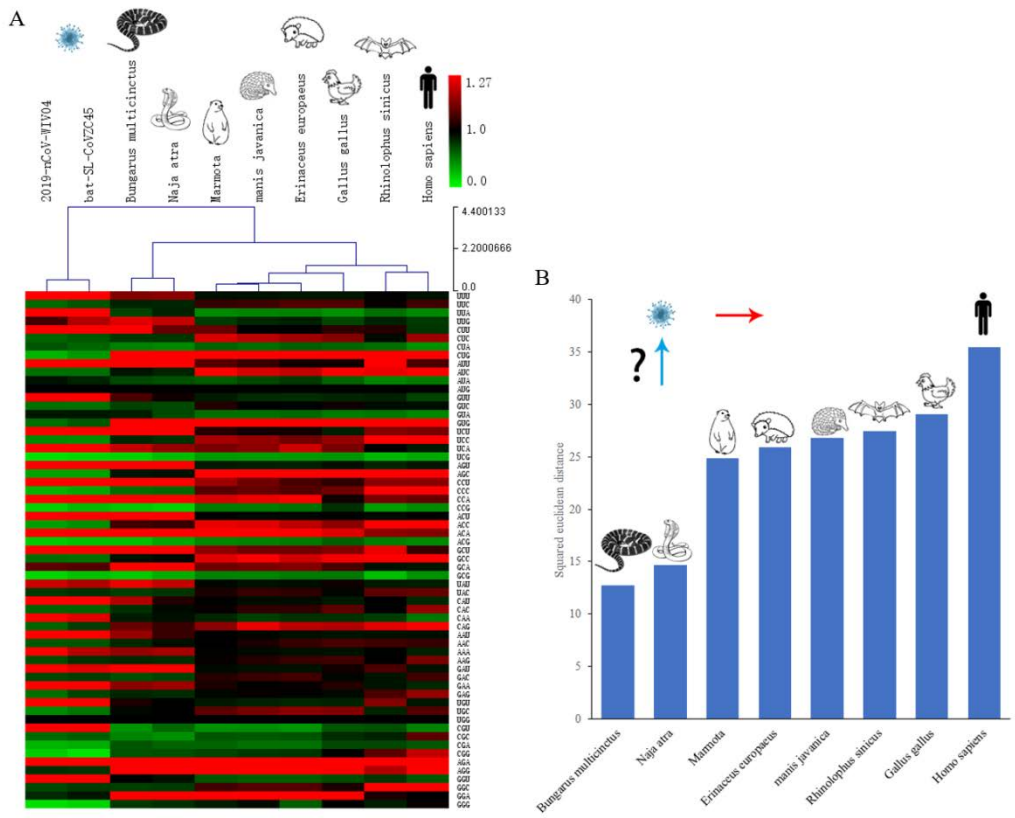

Figure 4. Geographic distribution of Bungarus multicinctus and Naja atra in China.

The geographic distribution of Bungarus multicinctus and Naja atra is shown at the provincial level. Yellow color represents the common geographic distribution of Bungarus multicinctus and Naja atra. Green color represents additional geographic distribution of Bungarus multicinctus. Maps were obtained from Craft MAP website (http://www.craftmap.box-i.net/).

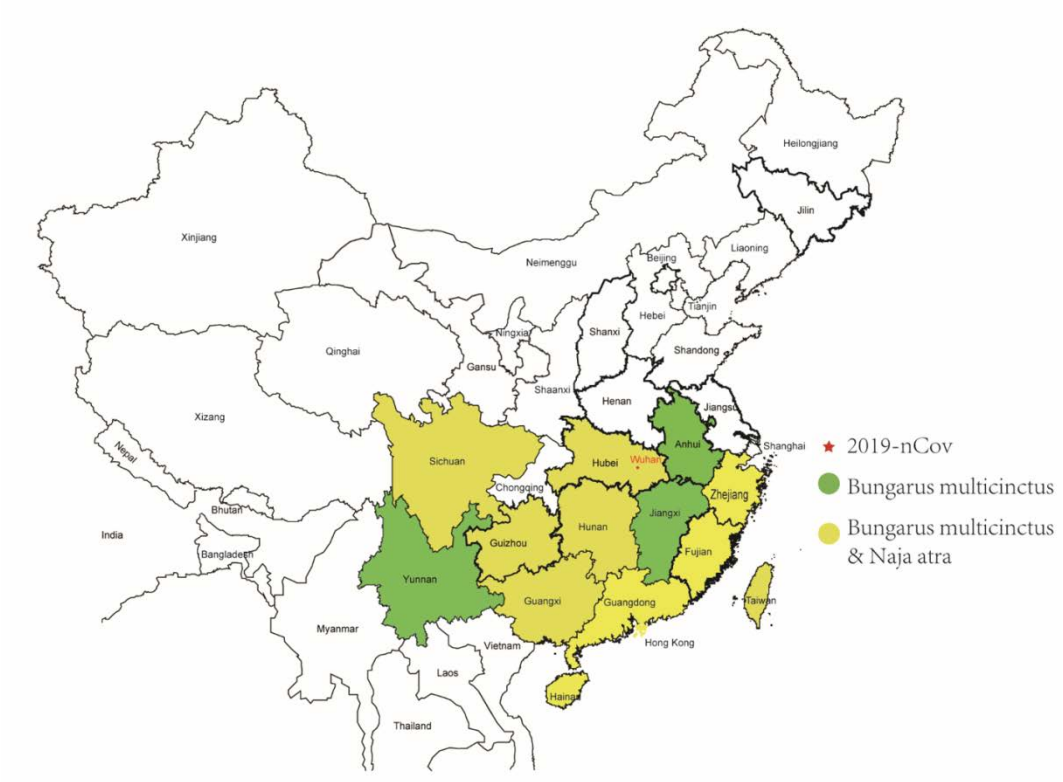

Table1 The RSCU analysis of the preferred codons (codons with RSCU > 1), the optimal codons and the rare codons for coronoviruses, snakes, hedgehog, bat Marmota, Manis, gallus, and human genome. The most preferred codons are in bold or red.

		bat-SL-CoVZC45	BetaCoV-Wuhan-WIV04	Bungarus multicinctus	Naja atra	Marmota	Erinaceus europaeus	manis javanica	Rhinolophus sinicus	Gallus gallus	Homo sapiens
Phe	UUU	**1.33**	**1.41**	**1.14**	**1.15**	0.94	0.91	0.93	0.99	0.91	0.93
	UUC	0.67	0.59	0.86	0.85	**1.06**	**1.09**	**1.07**	**1.01**	**1.09**	**1.07**
Leu	UU	1.38	1.64	0.73	0.88	0.4	0.49	0.4	0.56	0.	0.4

u	A				6	6		8		42	6
	U				1.		0.7		0.8		0.	
	U	1.19		1.07	1.24	2			5		72	0.7
	G					1		0.88		0.87		7
	C					1.						
	U	**1.78**		**1.75**	1.3	1	1.1		1		1.	0.7
	U					1	1	1.00		1.04	05	9
	C					0.						
	U	0.65		0.59	0.78	7	1.2		1.2		1.	1.1
	C					9	3	1.17		0.96	14	7
	C					0.						
	U	0.6		0.66	0.5	4	0.6		0.5		0.	0.4
	A					7	2	0.63	7	0.57	6	3
	C					**1.**						
	U	0.4		0.3	**1.45**	**5**	**1.8**	**1.83**	**1.8**	**2**	**2.**	**2.3**
	G					**5**	**1**		**9**		**06**	**7**

Ile	AUU	**1.58**	**1.53**	**1.37**	**1.46**	1.09	1.02	1.06	1.28	1.02	1.08
	AUC	0.57	0.56	0.92	0.86	**1.27**	**1.21**	**1.23**	**1.25**	**1.36**	**1.41**
	AUA	0.85	0.91	0.71	0.68	0.64	0.77	0.71	0.47	0.61	0.51
Met	AUG	1	1	1	1	1	1.00	1	1	1	1.00
Val	GUU	**1.89**	**1.95**	1.08	1.02	0.89	0.82	0.83	0.87	0.85	0.73
	GG	0.55	0.57	0.74	0.	1.0	**0.99**	1	**0.95**	1.	**0.9**

	UC				7 1	4			02	**5**
	GU A	0.9	0.9	0.88	0.7 1	0.56	0.59	0.56	0.51	0.47
	GU G	0.66	0.58	**1.29**	**1.56**	**1.51**	**1.60**	**1.61**	**1.62**	**1.85**
Ser	UC U	**2.04**	**1.96**	**1.5**	**1.43**	1.04	1.02	1.02	0.88	1.13
	UC C	0.44	0.47	0.85	0.81	1.18	1.12	1.16	1.31	1.31
	UC	1.66	1.66	1.21	1.1	1.13	1.22	1.15	1	1.13	0.90

	A				6					
	U				0.					
	C	0.15	0.11	0.17	2	0.39	0.40	0.38	0.28	0.41	0.33
	G				6					
	A				1.					
	G	1.36	1.43	1.35	3	0.89	0.86	0.87	0.85	0.93	0.90
	U				9					
	A				0.					
	G	0.36	0.37	0.91	9	**1.37**	**1.38**	**1.41**	**1.42**	**1.5**	**1.44**
	C				3					
P	C				**1.**					
r	C	**1.83**	**1.94**	**1.7**	**6**	1.17	1.08	1.12	1.15	1.05	1.15
o	U				**6**					
	C				0.					
	C	0.34	0.3	0.57	5	1.09	1.09	1.11	**1.35**	**1.12**	**1.29**
	C				9					

	CCA	1.59	1.6	1.51	1.58	**1.23**	**1.27**	**1.25**	1.13	1.102	1.11
	CCG	0.25	0.16	0.21	0.17	0.51	0.56	0.51	0.37	0.81	0.45
Thr	ACU	**1.75**	**1.78**	1.3	1.28	1	0.97	0.97	0.96	0.999	0.99
	ACC	0.44	0.38	1.07	1.06	1.25	1.20	1.24	**1.39**	1.152	**1.42**
	ACA	1.58	1.64	**1.38**	**1.33**	**1.29**	**1.31**	**1.31**	1.15	**1.24**	1.14
	AAA	0.24	0.2	0.24	0.03	0.4	0.52	0.4	0.5	0.01	0.4

	C				3	7		9		62	6
	G				4						
	G				1.						
	C				3	1.1		1.1		1.	
A	U	**2.13**	**2.18**	**1.54**	4	5	1.12	2	1.25	14	1.0
la											6
	G				0.						
	C	0.55	0.57	0.93	9	**1.2**		**1.2**		**1.**	**1.6**
	C				7	**4**	**1.22**	**7**	**1.59**	**26**	**0**
	G				**1.**						
	C	1.09	1.09	1.33	**3**	1.1		1.1		1.	0.9
	A				**8**	2	1.14	3	0.93	08	1
	G				0.						
	C	0.24	0.15	0.2	3	0.4		0.4		0.	0.4
	G				1	8	0.52	8	0.24	52	2
T	U	**1.19**	**1.22**	**1.25**	**1.**	0.9		0.9		**1.**	0.8
y	A				**2**	7	0.95	4	0.9	**02**	9

r	U				**1**						
	U				0.						
	A	0.81	0.78	0.75	7	**1.03**	**1.05**	**1.06**	**1.1**	0.98	**1.11**
	C				9						
	C				**1.**						
H	A	**1.39**	**1.39**	**1.19**	**0**	0.99	0.95	0.95	1	0.9	0.84
is	U				**4**						
	C				0.						
	A	0.61	0.61	0.81	9	**1.01**	**1.05**	**1.05**	1	**1.1**	**1.16**
	C				6						
	C				0.						
G	A	**1.24**	**1.39**	0.89	9	0.85	0.80	0.75	0.76	0.81	0.53
ln	A				4						
	C				**1.**						
	A	0.76	0.61	**1.11**	**0**	**1.15**	**1.20**	**1.25**	**1.24**	**1.19**	**1.47**
	G				**6**						

A	A				**1.**			0.9		0.	
s	A	**1.35**	**1.35**	**1.16**	**0**	1		7		92	0.9
n	U				**6**		0.94		0.93		4
	A				0.			**1.0**		**1.**	**1.0**
	A	0.65	0.65	0.84	9	1	**1.06**	**3**	**1.07**	**08**	**6**
	C				4						
L	A				**1.**						
						0.9		0.9		0.	
y	A	**1.2**	**1.31**	**1.16**	**1**	9		6		93	0.8
s	A				**8**		0.94		1		7
	A				0.						
						1.0		**1.0**		**1.**	**1.1**
	A	0.8	0.69	0.84	8		**1.06**				
	G				2	**1**		**4**	1	**07**	**3**
A	G				**1.**						
						0.9		0.9		**1.**	
s	A	**1.24**	**1.28**	**1.32**	**2**	6		4	**1.08**	**02**	0.9
p	U				**7**		0.93				3
	G	0.76	0.72	0.68	0.	**1.0**	**1.07**	**1.0**	0.92	0.	**1.0**

	A C				7 3	**4**		**6**		98	**7**
G lu	G A A	**1.27**	**1.44**	**1.15**	**1. 1 7**	**1.0 3**	**1.01**	0.9 9	0.92	**1. 04**	0.8 4
	G A G	0.73	0.56	0.85	0. 8 3	0.9 7	0.99	**1.0 1**	**1.08**	0. 96	**1.1 6**
C y s	U G U	**1.48**	**1.56**	**1.03**	1	0.9	0.85	0.8 7	**1.14**	0. 86	0.9 1
	U G C	0.52	0.44	0.97	1	**1.1**	**1.15**	**1.1 3**	0.86	**1. 14**	**1.0 9**
T r	U G	1	1	1	1	1	1.00	1	1	1	1.0 0

p	G									
A	C				0.					
						0.4		0.4		0.	
r	G	1.51	1.45	0.42	5						0.4
						3		4		66	
g	U				9		0.41		0.69		8
	C				0.					
						0.6		0.6		0.	
	G	0.64	0.59	0.48	4						1.1
						6		5		82	
	C				3		0.67		0.87		0
	C										
	G	0.33	0.29	0.66	0.			0.5		0.	
						0.6					0.6
	A				6			5		71	
							0.55		0.72		5
	C				0.					
						0.7		0.7		0.	
	G	0.1	0.19	0.67	6						1.2
						3		8		99	
	G				9		0.79		1.1		1
	A				**2.**						
						2.0		**1.9**		**1.**	**1.2**
	G	**2.63**	**2.67**	**2.29**	**0**		**1.97**		**1.42**		
						7		**4**		**46**	**9**
	A				**7**						

	A				1.					
						1.5		1.6	1.	
	G	0.79	0.81	1.47	6					1.2
						1		3	36	
	G				1		1.61		1.19	7
	G				0.					
						0.6		0.6	0.	
G	G	**2.16**	**2.34**	0.99	9					0.6
						6		3	7	
ly	U				3		0.63		0.79	5
	G				0.					
						1.0		1.0	1.	**1.3**
	G	0.76	0.71	0.75	7				**1.39**	
						7		9	03	**5**
	C				7		1.09			
	G				**1.**					
						1.3		**1.3**	**1.**	
	G	0.92	0.83	**1.59**	**6**		**1.33**			1.0
						9		**1**	**29**	
	A				**4**				1.03	0
	G				0.					
						0.8		0.9	0.	
	G	0.16	0.12	0.77	8					1.0
						8		5	99	
	G				1		0.62		0.88	1