Check for updates

RESEARCH ARTICLE Identification *Sus scrofa* and *Mus musculus* as potential hosts of SARS-CoV-2 via phylogenetic and homologous recombination analysis [version 1; peer review: awaiting peer review]

Xiaopeng Hu ^{1,2}, Weixin Li¹, Zhendan He², Fengxue Zhang³

¹Shenzhen Samii Medical Center, Shenzhen, China
²Department of Pharmacology, Shenzhen University, Shenzhen, China
³School of Basic Medical Sciences, Guangzhou University of Chinese Medicine, Guangzhou, China

V1 First published: 16 Mar 2020, 9:190 (https://doi.org/10.12688/f1000research.22627.1)

Latest published: 16 Mar 2020, 9:190 (https://doi.org/10.12688/f1000research.22627.1)

Abstract

Background: Wuhan Huanan seafood wholesale market is highly suspected as the original place for the outbreak of SARS-CoV-2 previously. Most studies focus on the livestock being sold in the market, neglecting that the livestock living around the city of Wuhan could also be the original hosts.

Methods: First, relative synonymous codon usage was utilized to analyze the potential hosts of SARS-CoV-2; Then cluster SARS-CoV-2 and related coronavirus through the phylogenetic tree. Next, we used Recombination Detection Program to identify the possible recombination region, as well as verifying via Simplot.

Results: Related coronavirus from porcine or murine sources may faciliatate the evolution and reorganization of SARS-CoV-2. **Conclusions:** Overall, to our knowledge, this is the first paper to illustrate that swine and mice could be probable reservoirs for the SARS-CoV-2.

Keywords

SARS-CoV-2, Sus scrofa, Mus musculus, phylogenetic analysis, homologous recombination analysis



This article is included in the Disease Outbreaks

gateway.

Open Peer Review

Reviewer Status AWAITING PEER REVIEW

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Fengxue Zhang (zhangfengxue@gzucm.edu.cn)

Author roles: Hu X: Data Curation, Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing; Li W: Project Administration, Resources, Software; He Z: Funding Acquisition, Investigation; Zhang F: Supervision, Validation, Visualization

Competing interests: No competing interests were disclosed.

Grant information: This study was funded by National Natural Science Foundation of China (U1702286, 31670360). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Hu X *et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Hu X, Li W, He Z and Zhang F. Identification *Sus scrofa* and *Mus musculus* as potential hosts of SARS-CoV-2 via phylogenetic and homologous recombination analysis [version 1; peer review: awaiting peer review] F1000Research 2020, 9:190 (https://doi.org/10.12688/f1000research.22627.1)

First published: 16 Mar 2020, 9:190 (https://doi.org/10.12688/f1000research.22627.1)

Introduction

A novel coronavirus, SARS-CoV-2, was recently reported in the city of Wuhan, Hubei province, China, causing severe respiratory diseases as well as epidemic all around the China. The inflection point of confirmed cases didn't occur until February 2020, with the first patient hospitalized on the 12th of December 2019¹. On 29th February, 573 new confirmed cases of novel coronavirus infection were reported on the Chinese mainland, bringing the total to 78,630. There were 35 new fatalities reported daily, with the cumulative fatalities were up to 2761. Meanwhile, outside the Chinese mainland, more than 5000 cases have been confirmed in Asia, in places such as Japan, Singapore, Thailand and South Korea, in Europe, in places like Germany and France, and in the Americas. Consensus that the SARS-CoV-2 originated from bats has been reached². However, intermediate hosts are deemed as having mediated human infection via gradually adapting the mechanism of transcription and translation in the human body.

The exact putative parent of SARS-CoV-2 remains uncertain. As a matter of fact, either wild animal or rear livestock deserve the suspicion. In the past ten years, there were more than 300 strains of porcine coronavirus had been reported in China (Data collected from NCBI Virus). Frequent contact between humans and swine could lead to a higher risk of cross-species transmission or virus recombination. For the sake of identifying intermediate host, relative synonymous codon usage (RSCU) analysis was applied to evaluate the potential diversity of species acting as reservoirs. Phylogenetic and homologous recombination analysis were also used to illuminate the correlated coronavirus hosting in the most possible host.

Methods

Sequence data collection

More than 300 genome sequences were obtained and analyzed from GenBank (a list of downloaded sequences is provided with Extended data A)³. Porcine and murine coronavirus from China were refined and downloaded from NCBI virus database (Table S1, Extended data B)⁴. Here, the mitochondrial genes represent the whole genome of potential host (Table S2, Extended data B)⁴. ClustalX 1.83 was applied to align the sequences. Coronavirus hosting in Sus scrofa from different regions of China were calculated (Extended data A)⁵. To be more specific, results from NCBI Virus database were refined using terms "coronaviridae", "China" and "sus scrofa" from Jan 1, 2010 to Mar 2, 2020. More than 300 nucleotide outputs were identified. Eventually, confirm the exact geographic region according with every GenBank accession. The porcine coronavirus belonging to diverse regions was presented as a heatmap.

Phylogenetic analysis

MEGA X (v1.0.3) was used to construct the phylogenetic trees using the Neighbor-Joining method⁶ (The input GenBank accessions are shown in Table S1, Extended data B)⁴. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches⁷. The tree is drawn to scale, with branch

lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All genome sequences were aligned before implementing of phylogenetic analysis.

Synonymous codon usage analysis

In order to identify the relative synonymous codon usage (RSCU) bias of the SARS-CoV-2 and its potential host, the coding sequences of suspect species were calculated using CodonW 1.4.2. (Table S4, Extended data C)⁵. Whole coding sequences of the genomes, downloaded from NCBI, were utilized for identifying the host among different viruses⁸ (the input GenBank accessions are shown in Table S1, Extended data B)⁴. The RSCU were calculated with CodonW 1.4.2. The heat map clustering of RSCU was realized via applying MeV 4.9.0. The homologous analysis was executed via analyzing the pairwise distance of mitochondrial genes from the potential host animal. The pairwise distances were computed with MEGA X via the bootstrap test (1000 replicates) to evaluate potential hosts⁹.

Genome homologous recombination analysis¹

The genome sequences of bat RaTG13 (MN996532.1), murine JHM (AC_000192.1), Human HKU1 (NC_006577.2), PEDV H11-SD2017 (MH708243.1) and PEDV YN15 (KT021228.1) were obtained from the NCBI Virus database. Potential recombination events of SARS-CoV-2 were first implemented with Recombination Detection Program v4 (RDP4), then Simplot (version 3.5.1) was utilized to verify the possible breakpoints. The BootScan method were taken into the RDP4 analysis. Potential recombination events were characterized with similarity plots, with possible regions of recombination indicated.

Results

Geographical hot map

The distribution of PEDV with Geographical heat map indicated the risk of pandemic of porcine epidemic diarrhea in different regions¹⁰ (Figure 1). In the past 10 years, more than 300 strains of PEDV have been reported in China¹¹. Guangdong province ranked top with 81 strains, followed by Jiangsu with 27 strains. The distribution of PEDV in Hubei province ranked third with the amount of 21 strains. Higher amounts of PEDV means higher infection exposure. Therefore, we believe that the outbreak of SARS-CoV-2 is associated with porcine coronavirus to some extent.

The intermediate animal reservoirs of the SARS-CoV-2 point to swine and mice.

RSCU has been widely utilized to analyze the association between virus and potential host⁸. The output from RSCU heat map, based on the Euclidean distance, inferred that both *Sus scrofa* and *Mus musculus* have similar synonymous codon usage bias with SARS-CoV-2 (Figure 2A). The Euclidean distance between *Sus scrofa* and SARS-CoV-2 is minimal, hinting that SARS-CoV-2 could effectively use porcine translation machinery better than that of other animals, suggesting that the epidemic SARS-CoV-2 might originate from swine. The pairwise distance between the SARS-CoV-2 complete genome and the potential host mitochondrial genome also supports

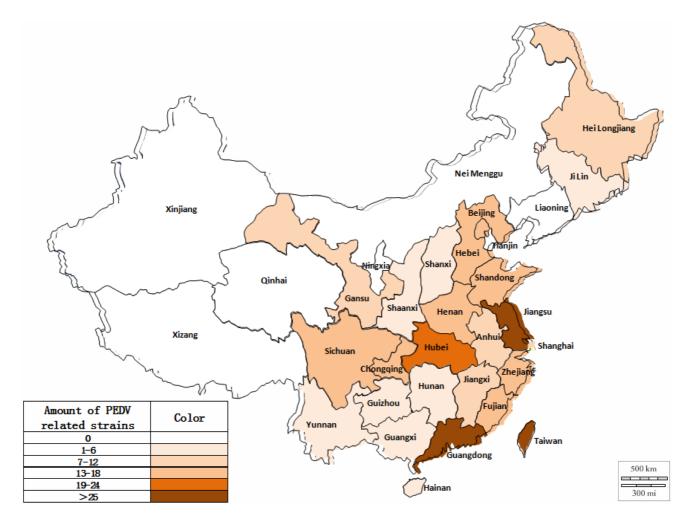


Figure 1. The hot map of 329 PEDV genomes distribution in China. Note that the Porcine Epidemic Diarrhea Virus (PEDV) strains locating accurately in different provinces were collected from NCBI virus database between 2010~2020(Table S3). The depth of the color represents the amount of PEDV related strains in distinctive regions. Note: Chinese map has been reproduced with permission from www.d-maps.com.

the above judgments. Regarding the pairwise distances of SARS-CoV-2, both swine and mice presented shorter than snakes, marmots, mink, bats and humans and (Figure 2B). Pairwise distance between the SARS-CoV-2 and *Sus scrofa* is 14.89, and for SARS-CoV-2 and *Mus musculus* is 14.94. This result indicates that SARS-CoV-2 may originate from both swine and mice.

Phylogenetic analysis

The 21 PEDV strains were filtered from 2010~2020 in Hubei based on the NCBI virus database (The GenBank accessions were shown in Figure 3). Phylogenetic analysis of the coronavirus (HKU15, PHEV, NL63, H11-SD2017, 229E, HKU2, TGEV, etc.) derived from varied species (bats, swine, humans), representing the sister lineage to SARS-CoV-2 with 99% bootstrap support (Figure 3). RSCU revealed that related coronaviruses have similar synonymous codon usage bias with

SARS-CoV-2 (Figure 4). The other PEDV strains obtained from Hubei also showed the closely phylogenetic correlation with SARS-CoV-2. Overall, the close phylogenetic relationship to *Sus scrofa* provides evidence for bat-swine axis, being one of the origins of SARS-CoV-2.

Meanwhile, the result of RSCU also presented that both SARS-CoV-2 and related coronavirus lean towards to having similar synonymous codon usage bias (Figure 4 and Table S5, Extended data C)⁵. Therefore, the origination of SARS-CoV-2 could be further focus on the coronavirus isolating from swine and mice. Particularly, the relationship with PEDV H11-SD2017 and PEDV YN15 needed to be further studied.

To our knowledge, this is the first study to report that porcine and murine coronavirus may attend the reorganization of SARS-CoV-2. The RDP4 estimated the possible reorganization

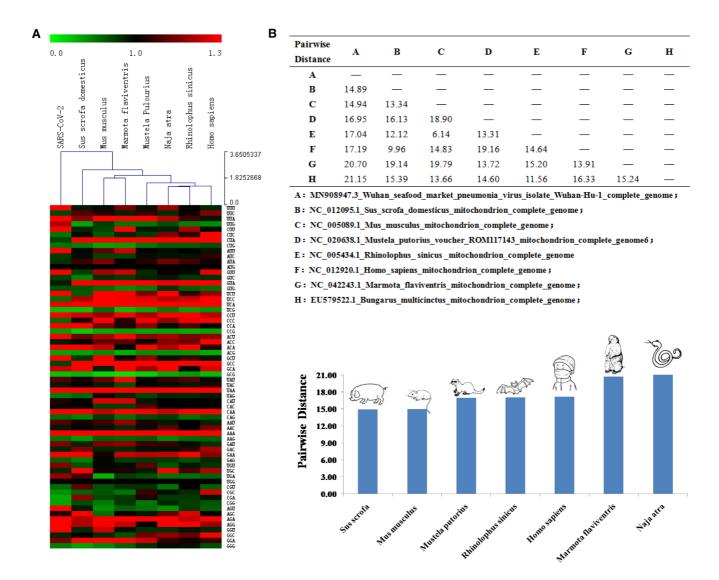


Figure 2. SARS-CoV-2 phylogeny analysis based on the codon usage. (A)Heat map of relative synonymous codon usage (RSCU) derived from the complete genome of SARS-CoV-2, as well as the mitochondrial genome of diverse animals (*Sus scrofa, Mus musculus, Najaatra, Mustela Pulourius, Marmota flaviventris, Rhinolophus sinicus, Homo sapiens*). (B)Homology analysis between the SARS-CoV-2 and different animal species (Table S4), Pairwise Distance was applied to evaluate the homology as compared with SARS-CoV-2 via MEGA-X.

regions for SARS-CoV-2 (Table 1, Figure 5A). Furthermore, SimPlot analysis confirmed the homologous recombination of sequence similarity between SARS-CoV-2 and coronavirus from potential hosts. The potential recombination breakpoints (16205-16358nt) are shown in red dashed lines (Figure 5B), indicating recombination between PEDV YN15 and the murine hepatitis virus JHM when SARS-CoV-2 was queried. In addition, the region between 20923 and 21181 nt also indicated a recombination event taking place between PEDV H11-SD2017 and Human coronavirus HKU1 when SARS-CoV-2 was queried (Figure 5C).

Discussion

In China, the cumulative number of patients diagnosed as infected with SARS-CoV-2 is thought to over 100,000, with more than 2000 deaths. It is the most severe public health emergency since the outbreak of SARS 17 years ago¹². Local residents were reined with anxiety and confused during the virus outbreak, although the Chinese government has taken substantial action to prevent and control the spread of the virus. However, multiple different messages could disrupt the attention of medical workers and science researchers. Current research about the origin of SARS-CoV-2 mostly focuses on wildlife, since the first case was thought to be highly associated with wild animals in Wuhan sea food market.

The origin of SARS-CoV-2 had caused great concern to the public. Natural variation was the dominant view holding by most scholars, believing that bats and other wild animals provide reservoirs for the virus. However, coronavirus from bats

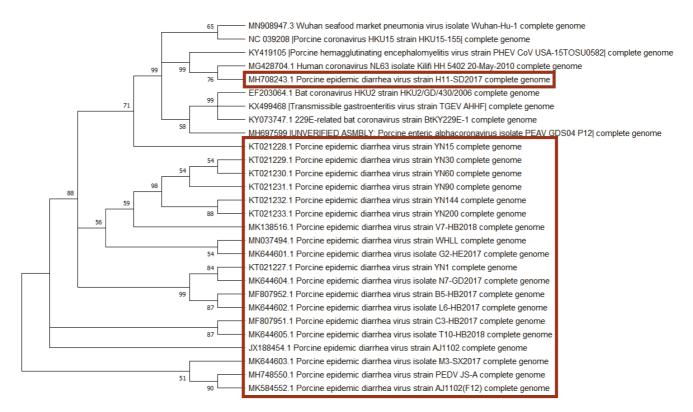


Figure 3. Phylogenetic tree of SARS-CoV-2 and related coronavirus strains. The neighbor-joining tree (bootstrap n=1,000; p-distance) was exerted to represent the evolutionary history of the taxa analyzed. Phylogeny-based geographical dissection of 21 PEDV strains (Brown box) derived from Hubei. Information about the 21 PEDV strains is shown in Table S1 Extended data B⁴.

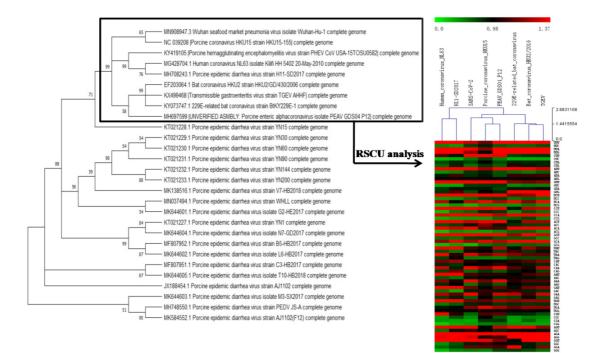


Figure 4. Relative synonymous codon usage (RSCU) analysis of SARS-CoV-2 and related coronavirus. The heat map of RSCU derived from the complete genome of SARS-CoV-2 and other polyphyletic coronaviruses. The Euclidean distance was calculated to cluster the related coronavirus.

Reorganization sequence	Major and minor parent sequence	P value	Reorganization region (nt)
MN908947.3	KT021228.1—AC000192	4.764×10 ⁻³	16205–16358
	MH708243.1-NC006577.2	2.579×10 ⁻²	20923-21181

MN908947.3: Wuhan_seafood_market_pneumonia_virus_isolate_Wuhan-Hu-1_complete_ genome; KT021228.1: Porcine_epidemic_diarrhea_virus_strain_YN15_complete_genome; AC_000192.1: Murine_hepatitis_virus_strain_JHM_complete_genome; NC_006577.2: Human_coronavirus_HKU1_complete_genome; MH708243.1: Porcine_epidemic_ diarrhea_virus_strain_H11-SD2017_complete_genome.

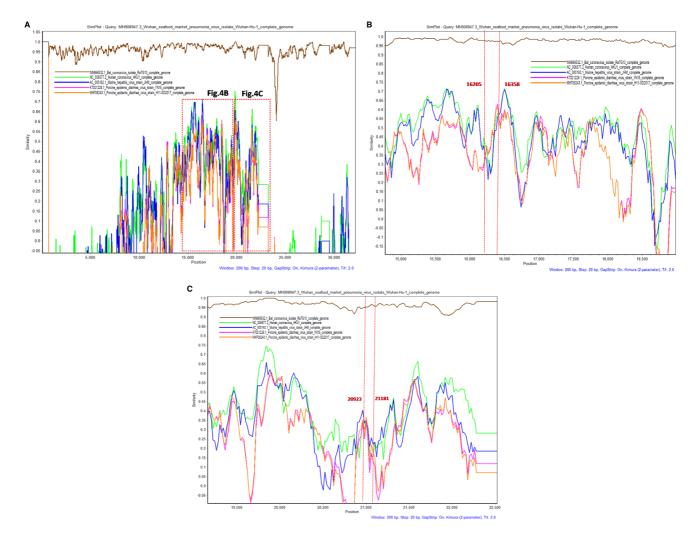


Figure 5. Simplot analysis of varies coronavirus from potential host. The similarities to different reference sequences are indicated by different colors shown in the legend box at the top. (A) The sequence of SARS-CoV-2(query), Bat coronavirus RaTG13, Human coronavirus HKU1, Murine hepatitis virus JHM, PEDV YN15 and PEDV H11-SD2017 were assessed with similarity plot. (B) The enlarged figure identifies the homologous recombination region (16205–16358 nt) of PEDV YN15 and Murine hepatitis strain JHM. (C) The enlarged figure identifies the homologous recombination region(20923–21181nt) of PEDV H11-SD2017 and Human coronavirus HKU1.

is unlikely to infect humans directly; one or two intermediate hosts may facilitate the homologous recombination, enabling the coronavirus to gradually adjust to the human genetic code, then survival and breeding successfully.

From our perspective, natural variation would be a more reasonable explanation for the origin of SARS-CoV-2. However, the wild animals in Wuhan seafood market should not accept all the liability, since there is evidence that patients with early infection didn't have any contact with the market¹². Some researchers¹³ have also pointed out that the Wuhan seafood market is not the only area of origin, indicating that another creature may act as intermediate host apart from those animals being sold at the market. In our study, the captivity animal, *Sus scrofa* (swine) and *Mus musculus* (mice) were suspected to be critical hosts of SARS-CoV-2.

Our finding supports the theory of natural variation. Natural variation believes that people get infected because they eat or come into contact with intermediate hosts. SARS-CoV-2 was found to be 96% identical at the whole-genome level to the bat coronavirus². Other reports state that snakes, mink, and pangolins could be potential hosts for SARS-CoV-2¹⁴.

Either wild animals or reared livestock could serve as hosts. It has been reported that Wuhan seafood market may not be the only source of novel virus spreading globally because the earliest patient became ill on 1 December 2019 and had no epidemiological link to the seafood market or later cases. The official details about the first 41 hospitalized patients showed 13 of the 41 patients had no link to the marketplace at all¹². One possible hypothesis is that the cross-species transmission has occurred in other places before the outbreak of Wuhan Huanan Seafood market.

Previous study into fatal swine acute diarrhea syndrome (SADS) revealed that SADS-related coronavirus was responsible for a large-scale outbreak of fatal disease in pigs in China¹⁰. Here we discovered Porcine epidemic diarrhea virus (PEDV) periodicity burst in China from 2010~2020¹¹ (Figure 1, Extended data A)³. Among one of the strains, H11-SD2017 showed closely affiliation with SARS-CoV-2 via implementing relative synonymous codon usage (RSCU) and phylogenetic analysis. Swine-to-human cross species transmission may explain why many patients with coronavirus disease-19 (COVID-19) not only suffer from severe respiratory diseases, but also diarrhea¹².

In the past 10 years, PEDV has spread into most provinces with the swine industry in China. Hubei was the first region to be infected with PEDV strain CH/HBQX/10 in 2010; from then on, more and more provinces reported the presence of PEDV¹¹. The spread of PEDV has gone beyond its initial geographical limitation. There were more than 300 PEDV strains clustered into pandemic, meaning the significant natural variation took place in the spread of PEDV in different regions. According to the statistics in China from 2010–2020, up to 21 strains of PEDV had been founded in Hubei, ranking as the third highest in China (Figure 1). Further analysis among 21 strains of PEDV and coronavirus from varied species indicate that swine and mice could be other hosts of SARS-CoV-2 (Figure 2A, Figure 3 and Figure 4).

RDP4 and Simplot analysis helped us better understand the homologous recombination of SARS-CoV-2 (Table 1, Figure 5). It verified that not only porcine coronavirus, but also murine coronavirus, experienced recombination events. Therefore, we speculate that SARS-CoV-2 may originate from the bat firstly, undergoing a series of recombination events, with swine and mice playing critical role in mediating cross species transmission.

Pairwise analysis of distance also indicated that Mus musculus could be a possible host of SARS-CoV-2 (Figure 2B). Previously, the Chinese Centers for Disease Control said that 33 of the samples were positive for the novel coronavirus nucleic acid. The positive samples were distributed among 22 stalls and a garbage truck in the Wuhan Huanan seafood market, so the outbreak is highly suspected to be related to the wildlife trade. But how could wild animal movement around the market lead to cross-infection with disparate species in different regions? One plausible phenomenon is that mice could be infected with SARS-CoV-2 firstly, transferring the infection to other wild species in the market. Thus, the role of mice in the market deserves more attention. In order to verify the exact intermediate host, mice and swine living around Wuhan should be collected for further proof test basing on Koch's postulates

Eventually, the rear livestock should be deserved more notice apart from wild vertebrate creatures. To our knowledge, this is the first study to illustrate that swine and mice are the probable livestock reservoir for the SARS-CoV-2. Furthermore, the mice around the seafood market may also be involved in the cross transmission of the virus to some extent. All of those output based on the bioinformatics analysis, further identification of host should be verified from isolation and other experiments.

Data availability Source data

The genomic data assessed in this study are described below in the *Extended data*.

Extended data

Figshare: Identification Sus scrofaand Mus musculus as potential parasitifersof SARS-CoV-2 via phylogenetic and homologous recombination analysis. https://doi.org/10.6084/m9.figshare. 11925588³.

This project contains Extended data A: PEDV distribution in China from 2010–2020. (The 329 PEDV genomes assessed in this study.)

Figshare: Identification Sus scrofaand Mus musculus as potential parasitifersof SARS-CoV-2 via phylogenetic and homologous recombination analysis. https://doi.org/10.6084/m9.figshare. 11925693⁴.

This project contains Extended data B: The GenBank accession numbers used for analysis. (Contains accession numbers, strain name and species used for phylogenetic analysis.)

Figshare: Identification Sus scrofaand Mus musculus as potential parasitifersof SARS-CoV-2 via phylogenetic and homologous recombination analysis. https://doi.org/10.6084/m9.figshare. 11925654⁵.

This project contains Extended data C: RSCU analysis of diverse genome. (Contains RSCU analysis of codons derived from SARS-CoV-2.)

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Acknowledgements

We are appreciated with the sharing of SARS-CoV-2 complete genome (GenBank accession MN908947) from Prof. Yongzhen Zhang research team, as well as the related genome GenBank accessions from other researchers.

References

- Wu F, Zhao S, Yu B, et al.: A new coronavirus associated with human respiratory disease in China. Nature. 2020; 579: 265–269.
 PubMed Abstract | Publisher Full Text
- Zhou P, Yang XL, Wang XG, et al.: A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020; 579: 270–273. PubMed Abstract | Publisher Full Text
- Hu X: Identification Sus scrofaand Mus musculus as potential parasitifersof SARS-CoV-2 via phylogenetic and homologous recombination analysis. *figshare*. Dataset. 2020. http://www.doi.org/10.6084/m9.figshare.11925588.v2
- Hu X: Identification Sus scrofaand Mus musculus as potential parasitifersof SARS-CoV-2 via phylogenetic and homologous recombination analysis. *figshare*. Dataset. 2020. http://www.doi.org/10.6084/m9.figshare.11925693.v1
- Hu X: Identification Sus scrofaand Mus musculus as potential parasitifersof SARS-CoV-2 via phylogenetic and homologous recombination analysis. *figshare*. Dataset. 2020. http://www.doi.org/10.6084/m9.figshare.11925654.v1
- Saitou N, Nei M: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987; 4(4): 406–425. PubMed Abstract | Publisher Full Text
- Felsenstein J: Confidence limits on phylogenies: An approach using the bootstrap. Evolution. 1985; 39(4): 783–791.
 PubMed Abstract | Publisher Full Text

- Chen LL, Ou HY, Zhang R, et al.: ZCURVE_CoV: a new system to recognize protein coding genes in coronavirus genomes, and its applications in analyzing SARS-CoV genomes. Biochem Biophys Res Commun. 2003; 307(2): 382–388.
 PubMed Abstract | Publisher Full Text
 - Tamura K. Stochar G. Poterson D. et al.: MEGA6: Ma
- Tamura K, Stecher G, Peterson D, et al.: MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013; 30(12): 2725–9. PubMed Abstract | Publisher Full Text | Free Full Text
- Zhou P, Fan H, Lan T, et al.: Fatal swine acute diarrhoea syndrome caused by an HKU2-related coronavirus of bat origin. Nature. 2018; 556(7700): 255–258. PubMed Abstract | Publisher Full Text
- Sun M, Ma J, Wang Y, et al.: Genomic and epidemiological characteristics provide new insights into the phylogeographical and spatiotemporal spread of porcine epidemic diarrhea virus in Asia. J Clin Microbiol. 2015; 53(5): 1484–92. PubMed Abstract | Publisher Full Text | Free Full Text
- Wang C, Horby PW, Hayden FG, *et al.*: A novel coronavirus outbreak of global health concern. *Lancet*. 2020; 395(10223): 470–473.
 PubMed Abstract | Publisher Full Text
- Yu WB, Zhang L, Tang GD, et al.: Decoding evolution and transmissions of novel pneumonia coronavirus using the whole genomic data. [ChinaXiv: 202002.00033], 2020.
 Reference Source
- Ji W, Wang W, Zhao X, et al.: Cross-species transmission of the newly identified coronavirus 2019-nCoV. J Med Virol. 2020; 92(4): 433–440.
 PubMed Abstract | Publisher Full Text

Comments on this article

Version 1

Reader Comment 20 Mar 2020

Billy Brazelton, University of Utah, Salt Lake City, USA

Figure 2 of this article is remarkably similar to Figure 3 of cited Ref. 14 (Ji et al.) https://onlinelibrary.wiley.com/doi/10.1002/jmv.25682. This is not necessarily wrong - maybe this is a standard visualization in this field? But perhaps the figure legend should acknowledge the inspiration?

A separate comment is that the authors' use of "natural variation" as the proposed mechanism for transmission is not a very precise description of what they mean, which is that the virus "naturally" transmits among people and animals frequently encountered in every day life and does not require unusual encounters of "wild" animals in meat markets. There must be a better term for this mechanism. Perhaps "routine contact"? Epidemiology must have a term for this, but that is not my area of expertise.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

