RESEARCH ARTICLE

REVISED Genomic and protein structure modelling analysis depicts the origin and pathogenicity of 2019-nCoV, a new coronavirus which caused a pneumonia outbreak in Wuhan, China [version 2; peer review: 2 not approved]

Ning Dong^{1*}, Xuemei Yang^{1*}, Lianwei Ye^{1*}, Kaichao Chen^{1*}, Edward Wai-Chi Chan², Sheng Chen¹

¹Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Kowloon, Hong Kong

²State Key Lab of Chirosciences, Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

* Equal contributors

V2 First published: 18 Feb 2020, 9:121
<https://doi.org/10.12688/f1000research.22357.1>
 Latest published: 09 Mar 2020, 9:121
<https://doi.org/10.12688/f1000research.22357.2>

Abstract

Background: A pandemic outbreak caused by a novel coronavirus, 2019-nCoV, has originated from Wuhan, China and spread to many countries around the world. The outbreak has led to around 45 thousand cases and over one thousand death so far.

Methods: Phylogenetic analysis and sequence alignment were used to align the whole genome sequence of 2019-nCoV with other over 200 sequences of coronaviruses to predict the origin of this novel virus. In addition, protein modeling and analysis were performed to access the potential binding of the spike protein of 2019-nCoV with human cell receptor, angiotensin-converting enzyme 2 (ACE2).

Results: Detailed genomic and structure-based analysis of a new coronavirus, namely 2019-nCoV, showed that the new virus is a new type of bat coronavirus and is genetically fairly distant from the human SARS coronavirus. Structure analysis of the spike (S) protein of this new virus showed that its S protein only binds much weaker to the ACE2 receptor on human cells whereas the human SARS coronavirus exhibits strongly affinity to the ACE receptor.

Conclusions: These findings suggest that the new virus should theoretically not be able to cause very serious human infection when compared to human SARS virus. However, the lower pathogenicity of this new virus may lead to longer incubation time and better adaption to human, which may favor its efficient transmission in human. These data are important to guide design of infection control policy and

Open Peer Review

Approval Status **XX**

	1	2
version 2		
(revision)	X	X
09 Mar 2020	view	view
version 1		
18 Feb 2020		
1. Fikret Aydemir , Cornell University, Ithaca, USA		
Hector Aguilar , Cornell University, Ithaca, USA		
2. Noam Auslander , National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, USA		

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

inform the public on the nature of threat imposed by 2019-nCoV. Most importantly, using the analysis platform that we have developed, we should be able to predict whether the new mutations could lead to the increase of infectivity of the mutated virus in a very short time.

Keywords

2019-nCoV, Genomics, Protein modelling, Origin, Pathogenicity, Wuhan

This article is included in the **Emerging Diseases and Outbreaks** gateway.

This article is included in the **Coronavirus** collection.

Corresponding author: Sheng Chen (shechen@cityu.edu.hk)

Author roles: **Dong N:** Formal Analysis, Methodology; **Yang X:** Formal Analysis, Methodology; **Ye L:** Formal Analysis, Methodology; **Chen K:** Formal Analysis, Methodology; **Chan EWC:** Formal Analysis, Writing – Review & Editing; **Chen S:** Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2020 Dong N *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: Dong N, Yang X, Ye L *et al.* **Genomic and protein structure modelling analysis depicts the origin and pathogenicity of 2019-nCoV, a new coronavirus which caused a pneumonia outbreak in Wuhan, China [version 2; peer review: 2 not approved]** F1000Research 2020, 9:121 <https://doi.org/10.12688/f1000research.22357.2>

First published: 18 Feb 2020, 9:121 <https://doi.org/10.12688/f1000research.22357.1>

REVISED Amendments from Version 1

Upon request by Prof. Mengsu Yang, we would like to remove him from the author list to the acknowledgment section.

Any further responses from the reviewers can be found at the end of the article

Introduction

A cluster of pneumonia cases of unknown cause were reported in Wuhan, the capital City of Hubei Province of China in December 2019. On 18th January 2020, a total of 44 such cases were documented, among whom two patients have died, five in critical condition, and six have been discharged from hospital. Most patients had visited or worked in a seafood wholesale market in Wuhan. In the early stage of this outbreak, there is no strong evidence which suggests that the unknown agent is highly infectious as no health care personnel in the hospitals where the patients were admitted were infected. However, the situation changed dramatically with newly infected cases rapidly increased and human to human transmission has been confirmed with several health care personnel being confirmed to be infected. As of Jan 28, 2020, the infection cases sharply increased in the past few days reaching a total of 4500 cases with over 100 confirmed deaths. Sporadic new cases were reported in more and more provinces in China mainly due to the traveler from Wuhan. Many confirmed cases were reported in many other countries. These patients were known to originate from or have been to Wuhan. These epidemiological data were quite different from the data reported in the beginning and may suggest that the new virus could undergo human host adaption/evolution and become more adaptive to human host leading to more efficient human to human transmission.

As of February 11, 2020, a total of 42,714 confirmed infections and 21,675 suspected cases were reported worldwide with over 31,728 confirmed cases in Hubei Province according to [data from the National Health Commission of China](#). A total of 1,017 deaths were reported so far with 974 deaths being from Hubei Province. The mortality rate is very high within Hubei Province reaching 3.06% (974/31,728), while the mortality rate outside Hubei Province was about 0.39% (43/10986). These data suggested that the 2019-nCoV is very infectious, while the pathogenicity seemed to be lower than SARS virus. It is urgent to find scientific evidence to support the epidemiological data to guide further control measure development.

Methods

Phylogenetic analysis and sequence alignment

A total of 211 genome sequences of viruses in the *Coronaviridae* family including the 2019-nCoV Wuhan virus were downloaded from [GenBank](#) (last accessed on 11 January 2020; list of downloaded sequences is provided as extended data²²). Circular proteomic trees were computed using [ViPTreeGen v1.1.2](#)¹. The sequence of a Breda virus (accession: [NC_007447](#)) was used as an outgroup. Alignment of sequences in different viral genomes was conducted using the alignment function of [ViPTree](#)¹. The phylogenetic tree and sequence

alignment products were manually edited using [Inkscape v0.91](#)². Ten spike protein sequences which were similar to that of 2019-nCoV were downloaded from NCBI. SNP analysis was performed using [Mega X](#)³, and the alignment was carried out by using [ClustalW](#)⁴. The aligned sequences were edited and viewed in [CLC Genomics Workbench 20](#).

Protein structure prediction and contacts between the human ACE2 and spike receptor-binding domains

Spike receptor-binding domain (RBD) of coronavirus 2019-nCoV and four bat-originated coronavirus were predicted by aligning their spike protein sequences to spike RBD of SARS coronavirus⁵ using [Clustal Omega](#)⁶. Homology modeling of spike proteins and the related RBDs of 2019-nCoV and four bat-originated coronavirus was performed using SARS-CoV spike glycoprotein as template (PDB ID: [5X58](#))⁷ on the [Swiss-Model workspace](#)⁸. The structure assessment results are presented below. The models were visualized with [PyMol](#). The contacts between human ACE2 and spike RBD were predicted by aligning to structure of the RBD in complex with the human receptor ACE2 (PDB ID: [2AJF](#))⁹.

Open source alternatives

Sequence alignment performed with CLC Genomics Workbench can be replicated using open source alternatives such as [ClustalW](#).

An open source version of PyMol has been made freely available by the developers, available from GitHub (<https://github.com/schrodinger/pymol-open-source>).

An earlier version of this article can be found on bioRxiv (<https://doi.org/10.1101/2020.01.20.913368>).

Results and discussion

On 6th January 2020, the Chinese authority released the sequence (accession#: [MN908947](#)) of a novel coronavirus, designated as 2019-nCoV, which was isolated from one of the pneumonia patients and confirmed to be the causative agent for this outbreak⁹. Coronaviruses are a large family of viruses, most of which cause mild infections such as the common cold, but some such as the SARS and MERS (Middle East Respiratory Syndrome) viruses cause severe and potential fatal respiratory tract infections^{10,11}. Some coronaviruses are known to be transmitted easily between humans, while others do not. Based on currently available information, the 2019-nCoV virus belongs to a category that can cause severe illness in some patients but does not transmit readily between people¹²⁻¹⁴. It is necessary to investigate the genetic and functional data of this new virus and compare to other coronaviruses so as to guide future research and design of appropriate infection control policy to prevent widespread dissemination of another potentially deadly coronavirus since the emergence of the SARS and MERS viruses. In this study, we performed in-depth genetic analysis of 2019-nCoV and generated data which provide timely and valuable insight into the potential origin of this virus, its ability to cause human infection, and its genetic relatedness with SARS and MERS.

Phylogenetic analysis of genomic sequences of coronaviruses deposited in the GenBank revealed that 2019-nCoV belonged to betacoronavirus and exhibited the closest linkage with two SARS-like coronavirus from bat (bat-SL-CoVZXX45 and bat-SL-CoVZXX21) (Figure 1a). According to the phylogenetic tree, human SARS viruses were closest to bat SARS-like viruses but with a lesser degree to bat coronaviruses, and was least related to other coronaviruses. The 2019-nCoV stands in a position between bat SARS-like viruses and bat coronaviruses, suggesting it is less related to the human SARS virus than other bat SARS-like viruses, and is likely a new type of bat coronavirus. Nevertheless, all coronaviruses that exhibit close linkage with 2019-nCoV originated from bat, strongly suggesting that this new coronavirus originated from bat (Figure 1a). Coronaviruses of other species including the murine coronavirus are genetically distant from this new coronavirus, indicating that 2019-nCoV did not originate from other animal hosts. As bats are not sold in the Wuhan market, animals that serve as the transmission vehicle remains to be identified^{15,16}.

The sequence of 2019-nCoV was annotated and aligned with several representative coronaviruses selected according to the degree of genetic relatedness depicted by the phylogenetic tree (Figure 1b). These included two highly homologous human SARS coronavirus: SARS CoV P2 (FJ882963) and SARS CoV ZJ02 (EU371559), one bat SARS virus that exhibits high homology with human SARS virus and similar potential to infect human as human SARS coronavirus: bat SARS CoV W1V1 (KF367457)¹⁷, two bat SARS-like viruses that were not able to infect human: (bat-SL-CoVZXX45 and Rp Shaanxi2011), and two un-related coronaviruses, the MERS virus MERS CoV (NC019843) and the Avian Infectious Bronchitis (IBV) virus IBV CoV (AY646283). The new 2019-nCoV was annotated slightly different from the human SARS virus and other coronaviruses, but the functionally important ORFs, ORF1a and ORF1b, and major structural proteins including the spike (S), membrane (M) and envelop (E) and nucleic capsid (N) proteins are well annotated (Figure 1b). Consistent with the phylogenetic tree data, 2019-nCoV did not align well with the MERS and IBV virus (Figure S1, Extended data²²). Among the SARS viruses, bat coronaviruses and 2019-nCoV, the non-structural proteins generally aligned well but variations were observable in major structural proteins and some small ORFs (Figure 1b). Detailed sequence alignment showed that 2019-nCoV exhibited significant sequence variation at several regions with the human SARS coronavirus including the N-terminal region of ORF1a and S protein, ORF3, E, ORF6, 7 and 8, and the middle part of N protein. Bat SARS-like virus W1V1 aligned well with human SARS virus P2, with some variations at ORF8 and an insertion between ORF6 and 7. Coronavirus 2019-nCoV aligned best with the bat SARS-like virus bat-SL-CoVZXX45, with the majority of genetic variations being seen at the N-terminal part of S protein. The bat-SL-CoVZXX45 virus itself exhibited a high degree of variation with another bat SARS-like CoV Rp Shaanxi2011 at ORF1a, the N-terminus of S protein and other structural proteins. However, bat SARS-like CoV Rp Shaanxi2011 exhibited high homology with human SARS virus ZJ02, with variation being seen at the N-terminus of S protein and the middle part of N

protein (Figure 1b). To check if 2019-nCoV is more close to bat SARS like virus or bat coronavirus, we selected two adjacent viruses from each group, bat SARSCoV Rf1/2004 and bat coronavirus BM48-31/BGR/2008, to perform alignment with the new virus with result showing that 2019-nCoV showed similar diversity from these two strains (Figure S1b, Extended data²²). When 2019-nCoV was aligned with bat CoV HKU9-1, another bat coronavirus with further distance, it showed that 2019-nCoV was very different from this virus (Figure S1c, Extended data²²). These sequence alignment data were consistent with results of phylogenetic tree analysis and indicated that 2019-nCoV exhibited in between bat SARS like viruses and bat coronaviruses, but is genetically distant from the human SARS virus. It should be considered a new type of bat coronavirus.

Since the S protein is the protein that exhibits the highest degree of genetic variations among different coronaviruses, we performed phylogenetic analysis of the S protein of different coronaviruses (Figure S2, Extended data²²). Our data showed that the S protein of 2019-nCoV exhibited high homology with bat SARS-like coronaviruses such as bat-SL-CoVZXX45 and bat-SL-CoVZXX21, human SARS virus and bat coronaviruses. Homology of S protein of 2019-nCoV with representative human SARS virus, bat SARS like viruses and bat coronaviruses was determined as shown in Table 2. S protein of 2019-nCoV showed about 76% homology to human SARS virus P2 and high homology to bat SARS like viruses, while it showed 72% homology to closest bat coronavirus, bat coronavirus BM48-31, even lower homology with other bat coronaviruses. These data further suggested that 2019-nCoV is more likely a new type of bat coronavirus with only loose linkage with the SARS virus. Interestingly, different regions of the S proteins exhibited different levels of homology among the known coronaviruses (Figure 1b). Amino acid sequence alignment showed consistently that the N-terminal regions were far more diverse than the C-terminus, which seemed to be highly conservative (Figure 2). Aligning these regions to the structure of S protein indicated that the structurally conserved C-terminal aligned well to the transmembrane domain which consists of a double helix, whereas the most variable region aligned to the N-terminal domain; on the other hand, the receptor binding domain exhibited intermediate level of sequence variation (Figure 2).

Due to the high amino acid sequence homology of the S protein in 2019-nCoV and the SARS virus which can cause severe human infection, we analyzed the structural similarity of this protein in various viruses. Protein structure modeling was performed to obtain high quality structure of S proteins from different coronaviruses (Table 1). The high level similarity observable between the structures of S protein from different viruses implied that the S protein of 2019-nCoV and bat coronaviruses would most likely use the same human cell receptor as SARS virus (Table 2). It was shown that angiotensin-converting enzyme 2 (ACE2) was the cellular receptor of the SARS virus S protein¹⁸. Complex structure of ACE2 with the receptor binding domain (RBD) of S protein of SARS virus has been resolved and demonstrated tight interaction

Figure 1. Phylogenetic analysis and sequence alignment of coronaviruses of different species. (a) Phylogenetic tree of coronaviruses from different species. The type of coronavirus and the host were labelled. Virus labeled with red is the newly discovered coronavirus 2019-nCoV. (b) Sequence alignment of representative the new 2019-nCoV, bat SARS like coronaviruses and bat coronaviruses. These included two highly homologous human SARS coronaviruses: SARS CoV P2 (FJ882963) and SARS CoV ZJ02 (EU371559), one bat SARS virus showing high homology with human SARS virus and similar potential to infect human as human SARS coronavirus: bat SARS CoV WIV1 (KF367457), two bat SARS-like viruses that are not able to infect human: (bat-SL-CoVZC45 and Rp Shaanxi2011) and the newly discovered 2019-nCoV from Wuhan.

Figure 2. Amino acid sequence alignment of coronaviruses from bat and human. Upper panel is the amino acid sequence alignment and the lower panel is the structure of the S protein. Three domains of S protein were labelled as red, blue and green, which aligned with the different colors for the aligned sequence. Sequences that exhibits the highest degree of variation are in the N-terminal domain, followed by those in the RBD. The C-terminal green domain is the most conservative among the test viruses.

between these two proteins at the interaction interface⁵. Current studies have confirmed that 2019-nCoV indeed showed binding to ACE2. These data prompted us to determine the level of interaction between the S protein of 2019-nCoV with its potential

cellular receptor ACE2. Using modeled S protein structure and by further performing structure-based alignment, we obtained the complex structure of RBD of the S protein of 2019-nCoV and several bat coronaviruses with human ACE2, using the complex

Table 1. Assessment of the quality of modeled structures of spike protein and its receptor binding domain (RBD) of different coronavirus using protein structures of the human SARS virus as templates.

S protein	GenBank	residues	MolProbity Score ¹	Ramachandran Favoured ¹	GMQE ²	QMEAN ²
2019-nCoV	MN908947	1273	1.64	89.4%	0.73	-4.46
BatSARS-likeCoV-WIV1	KC881007	1256	1.36	90.02%	0.74	-3.45
BatSARS-likeCoV	MG772934	1245	1.64	88.96%	0.73	-3.91
BatSARSr-CoV	DQ022305	1242	1.34	89.49%	0.73	-3.96
BatCoV	JX993987	1240	1.4	90.02%	0.73	-3.90

¹, Structure assessment. MolProbity is all-atom contact analysis based only on properties of the predicted model. Lower numbers indicate better models. Ramachandran favoured indicates energetically favoured regions for backbone dihedral angles against amino acid residues in protein structure. Larger numbers indicate better models.

², Model evaluation. GMQE (Global Model Quality Estimation) is a quality estimation which combines properties of the target-template alignment and the template search method. The resulting GMQE score is expressed as a number between 0 and 1, larger numbers indicate higher reliability. The QMEAN Z-score provides an estimate of the "degree of nativeness" of the structural features observed in the model on a global scale. QMEAN Z-scores around zero indicate good agreement between the model structure and experimental structures of similar size.

Table 2. Homology of S protein between different coronaviruses.

	2019-nCoV	SL-CoVZXC21	CoV-HKU3-1	CoV-WIV1	SARS_P2	Shaanxi-2011	Bat-BM48-31	Bat-zj2013	CoV-HKU5-1	CoV-HKU4-1
2019-nCoV	100%	81.35%	77.93%	77.7%	76.58%	76.56%	72.5%	41.06%	31.21%	30.58%
SLCoVZXC21	81.35%	100%	83.09%	78.15%	77.5%	82.05%	71.8%	40.59%	31.62%	31.46%
CoV-HKU3-1	77.93%	83.09%	100%	80.11%	79.55%	88.38%	75.1%	40.79%	32.27%	31.87%
CoV-WIV1	77.7%	78.15%	80.11%	100%	92.11%	81.06%	75.66%	41.35%	31.31%	30.8%
SARS_P2	76.58%	77.5%	79.55%	92.11%	100%	81.14%	75.64%	41.14%	31.42%	30.99%
Shaanxi-2011	76.56%	82.05%	88.38%	81.06%	81.14%	100%	75.06%	41.6%	31.91%	31.67%
Bat-BM48-31	72.5%	71.8%	75.1%	75.66%	75.64%	75.06%	100%	40.21%	32.12%	32.05%
Bat-zj2013	41.06%	40.59%	40.79%	41.35%	41.14%	41.6%	40.21%	100%	28.77%	28.9%
CoV-HKU5-1	31.21%	31.62%	32.27%	31.31%	31.42%	31.91%	32.12%	28.77%	100%	69.23%
CoV-HKU4-1	30.58%	31.46%	31.87%	30.8%	30.99%	31.67%	32.05%	28.9%	69.23%	100%

structure of RBS of S protein from human SARS virus and ACE2 (2ajf) as reference ⁵. Structural analysis of potential interactions between RBD of S protein from human SARS virus and ACE2 protein depicted several interaction points including four hydrophobic interactions: ACE2(Y⁴¹)/RBD(Y⁴⁸⁴), ACE2(L⁴⁵)/RBD(Y⁴⁸⁴), ACE2(L⁷⁹, M⁸²)/RBD(L⁴⁷²), ACE2(Y⁸³)/RBD(Y⁴⁷⁵), one salt-bridge: ACE2(E³²⁹)/RBD(R⁴²⁶) and one cation- π interaction: ACE2(K³⁵³)/RBD(Y⁴⁹¹) (Figure 3a–e). However, examination of interaction between RBD of 2019-nCoV and human ACE2 depicted only one potential hydrophobic interaction between ACE2(L⁷⁹, M⁸²) and RBD(F⁴⁸⁶) and one cation- π interaction ACE2(K³⁵³)/RBD(Y⁴⁹²). Further examination of interactions between RBD from bat SARS like coronaviruses, bat-SL-CoVZXC45 and bat coronavirus HKU3-1 that do not infect human, showed only one cation- π interaction interaction, ACE2(K³⁵³)/RBD(Y⁴⁸¹)^{20,21}. Another bat-originated coronavirus, bat SARS CoV WIV1 that displayed strong binding

to ACE2 and exhibited potential to cause human infection was also included for analysis¹⁷. Binding affinity of RBD from this virus to ACE2 was as tight as that of the human SARS virus, involving four hydrophobic interactions: ACE2(Y⁴¹)/RBD(Y⁴⁸⁵), ACE2(L⁴⁵)/RBD(Y⁴⁸⁵), ACE2(L⁷⁹, M⁸²)/RBD(F⁴⁷³), ACE2(Y⁸³)/RBD(Y⁴⁷⁶), one salt-bridge: ACE2(E³²⁹)/RBD(R⁴²⁷) and one cation- π interaction site ACE2(K³⁵³)/RBD(Y⁴⁹²). These data suggested that the higher binding affinity of RBD of coronavirus to ACE2 will confer the virus higher infectivity and pathogenicity. The fact that the RBD of 2019-nCoV exhibited much lower affinity to ACE2 implies that the virulence potential of 2019-nCoV should be much lower than that of human SARS virus, but is nevertheless stronger than viruses that do not cause human infection^{20,21}; such finding is also consistent with the current epidemiological data in that 2019-nCoV only caused severe pneumonia in patients with weaker immune system such as the elderly and people with underlying diseases. The weaker binding affinity of

Figure 3. Potential interactions between receptor binding domain (RBD) of S proteins from different coronaviruses and the human cell receptor ACE2. Interactions between ACE2 with RBD of human SARS virus (a), highly similar bat SARS like coronavirus, CoV-WIV1 that can infect human (b), new type of coronavirus, 2019-nCoV (c), bat SARS like coronavirus CoV SL-CoVZXC21 and bat coronavirus HKU3-1 (d) are shown. The detailed amino acid interaction sites between these two proteins are shown in (e). Arrows showed the areas with interacted residues from both proteins. Amino acid highlighted with different colors indicated the potential interaction residues between different proteins, which was highlighted with different colors.

2019-nCoV to human cell might also explain the limited human to human transmission potential of this virus observed to date.

Discussion

In this study, we utilized the whole genome sequence of the newly discovered coronavirus, 2019-nCoV, that caused an outbreak of pneumonia in Wuhan, China to perform comparative genetic and functional analysis with the human SARS virus and coronaviruses recovered from different animals. Phylogenetic analysis of coronavirus of different species indicated that 2019-nCoV might have originated from bat, but the intermediate transmission vehicle is not known at this stage. Genetic linkage analysis showed that 2019-nCoV lied at the interface between bat SARS like coronavirus and bat coronavirus and should belong to a novel type of bat coronavirus owing to high degree of variation from the human SARS virus. Analysis of the potential interaction of RBD of 2019-nCoV with human ACE2 receptor protein indicated that its affinity to human cell is much lower than that of human SARS virus due to the loss of several important interaction sites, implying that the infectivity and pathogenicity of this new virus should be much lower than the human SARS virus. These data were the most comprehensive scientific data that support the origin of this new virus, which is important for the following research to identify the intermediate transmission vehicles most likely wild animals. Most importantly, our data supported that the pathogenicity of 2019-nCoV

is much lower when compared to SARS virus. These data are very important for the current prevention and treatment of infections caused by this new virus. First, the lower pathogenicity of this new virus may suggest that the infectious dose should be large and the incubation time should be longer, which imply that 1) human to human transmission should mainly happen with close contact within a certain period of time; 2) the longer incubation time should favor the transmission of this virus in human, which suggest that patients with no or mild symptom should also be isolated to prevent further transmission; 3) supportive treatment should be enough to save lives of severe cases. Second, the lower pathogenicity of the virus may suggest that most of the patients should have mild symptom, but they are infectious. Appropriate measures should be designed to isolate these patients without occupying large amount of hospital resources. Thirdly, the lower pathogenicity of the new virus may suggest that it adapts to human much better, which imply that the virus may not disappear after few generations of transmission and long term battle with this virus should be prepared. Lastly, the epidemiology of the outbreak in the early stage is very different from the current dramatic situation. This may not exclude that the virus may have undergone human adaption and mutation in the past month. Using the analysis platform that we have developed above, we should be able to predict whether the new mutations could lead to the increase of infectivity of the mutated virus in a very short time.

Data availability

Underlying data

All data underlying the results are available as part of the article and no additional source data are required.

Extended data

Figshare: Extended data for 2019-nCoV. <https://doi.org/10.6084/m9.figshare.11848224>²²

This project contains the following extended data:

- Extended+data_F1000Research.docx (Document containing supplementary table and figures)
- all-virus-accessions.xlsx (Excel spreadsheet of virus accession numbers used for analysis)

- spike-protein-accessions.xlsx (Excel spreadsheet of spike protein accession numbers used for analysis)

Data are available under the terms of the [Creative Commons Zero “No rights reserved” data waiver](#) (CC0 1.0 Public domain dedication).

Acknowledgments

We are grateful for the discussion and comments from Prof. Mengsu Yang from CityU of Hong Kong. We acknowledge the use of the genome sequence of 2019-nCoV (deposited in GenBank with the accession number MN908947) and other coronavirus sequences that we have used in this study in GenBank. An earlier version of this article can be found on bioRxiv (doi: <https://doi.org/10.1101/2020.01.20.913368>).

References

- Nishimura Y, Yoshida T, Kuronishi M, *et al.*: **ViPTree: the viral proteomic tree server.** *Bioinformatics.* 2017; **33**(15): 2379–2380.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Bah T: **Inkscape: guide to a vector drawing program**[M]. Upper Saddle River, NJ USA: Prentice Hall. 2011.
[Reference Source](#)
- Kumar S, Stecher G, Li M, *et al.*: **MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.** *Mol Biol Evol.* 2018; **35**(6): 1547–1549.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Larkin MA, Blackshields G, Brown NP, *et al.*: **Clustal W and Clustal X version 2.0.** *Bioinformatics.* 2007; **23**(21): 2947–2948.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Li F, Li W, Farzan M, *et al.*: **Structure of SARS coronavirus spike receptor-binding domain complexed with receptor.** *Science.* 2005; **309**(5742): 1864–1868.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Madeira F, Park YM, Lee J, *et al.*: **The EMBL-EBI search and sequence analysis tools APIs in 2019.** *Nucleic Acids Res.* 2019; **47**(W1): W636–W641.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Yuan Y, Cao D, Zhang Y, *et al.*: **Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains.** *Nat Commun.* 2017; **8**: 15092.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Waterhouse A, Bertoni M, Bienert S, *et al.*: **SWISS-MODEL: homology modelling of protein structures and complexes.** *Nucleic Acids Res.* 2018; **46**(W1): W296–W303.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lu R, Zhao X, Li J, *et al.*: **Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding.** *Lancet.* 2020; pii: S0140-6736(20)30251-8.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Ji JS: **Origins of MERS-CoV, and lessons for 2019-nCoV.** *Lancet Planet Health.* 2020; pii: S2542-5196(20)30032-2.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Lee PI, Hsueh PR: **Emerging threats from zoonotic coronaviruses-from SARS and MERS to 2019-nCoV.** *J Microbiol Immunol Infect.* 2020; pii: S1684-1182(20)30011-6.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Li Q, Guan X, Wu P, *et al.*: **Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia.** *N Engl J Med.* 2020.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Rothe C, Schunk M, Sothmann P, *et al.*: **Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany.** *N Engl J Med.* 2020; **382**(10): 970–971.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Zhu N, Zhang D, Wang W, *et al.*: **A Novel Coronavirus from Patients with Pneumonia in China, 2019.** *N Engl J Med.* 2020.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Lau SK, Woo PC, Li KS, *et al.*: **Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats.** *Proc Natl Acad Sci U S A.* 2005; **102**(39): 14040–14045.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- To KK, Hung IF, Chan JF, *et al.*: **From SARS coronavirus to novel animal and human coronaviruses.** *J Thorac Dis.* 2013; **5** Suppl 2: S103–108.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ge XY, Li JL, Yang XL, *et al.*: **Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor.** *Nature.* 2013; **503**(7477): 535–538.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Li W, Moore MJ, Vasilieva N, *et al.*: **Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus.** *Nature.* 2003; **426**(6965): 450–454.
[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.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Dominguez SR, Shrivastava S, Berglund A, *et al.*: **Isolation, propagation, genome analysis and epidemiology of HKU1 betacoronaviruses.** *J Gen Virol.* 2014; **95**(Pt 4): 836–848.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hu D, Zhu C, Ai L, *et al.*: **Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats.** *Emerg Microbes Infect.* 2018; **7**(1): 154.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Chen S: **Extended data for 2019-nCoV.** *figshare.* Dataset. 2020.
<http://www.doi.org/10.6084/m9.figshare.11848224.v1>

Open Peer Review

Current Peer Review Status:

Version 2

Reviewer Report 07 July 2020

<https://doi.org/10.5256/f1000research.25120.r65534>

© 2020 Auslander N. This is an open access peer review report 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. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

Noam Auslander

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

In the study by Dong *et al.* the authors perform a genomic, phylogenetic and structural analysis of SARS-CoV-2, the causal agent of COVID-19. While the goal of the study is worthy and timely, the conclusions drawn are not justified or supported, possibly due to incomplete data or suboptimal methodology that was used throughout this study. The following concerns should be addressed:

1. Data: the 211 complete genomes of *Coronaviridae* family that has been collected seem to be only a subset of the available data. There are thousands of complete genomes of viruses from the *Coronaviridae* family, and substantially more than 211 complete genomes of human CoVs. The selection of these specific 211 should be explained and justified.
2. Alignment and phylogenetic tree software: The software used to construct alignments and phylogenetic trees is not standard. The authors draw critical conclusions from the results of these applications. However, it is not clear how accurate are the alignments and trees built using the ViPTree tool for closely related viruses within one particular viral family. To draw conclusions from this analysis, the authors should use standard multiple sequence alignment software, such as MUSCLE or MAFFT, and standard phylogenetic tree construction, such as phyML.
3. It is unclear how the *"Ten spike protein sequences which were similar to that of 2019-nCoV"* were selected. What is the criterion used for similarity? Why these particular ten?
4. The conclusions drawn appear to be in contradiction to recent literature. The authors write *"Based on currently available information, the 2019-nCoV virus belongs to a category that can cause severe illness in some patients but does not transmit readily between people"*. However, the references that they cite do not support this notion.
5. The authors write *"Phylogenetic analysis of genomic sequences of coronaviruses deposited in*

the GenBank revealed that 2019-nCoV belonged to betacoronavirus and exhibited the closest linkage with two SARS-like coronavirus from bat (bat-SL-CoVZXC45 and bat-SL-CoVZXC21)". First, these genomes could not be found in NCBI. I assume that there was a typo, and the authors meant "bat-SL-CoVZXC21" and "bat-SL-CoVZXC45". However, these are not the current closest relatives of SARS-CoV-2, but MN996532 (RaTG13). This may be due to missing sequences or issues with the alignments/tree constructions that are mentioned before. As this is critical to the analysis, it should be revised and addressed.

6. The authors conclude that the new coronavirus originated from bat because all coronaviruses that exhibit close linkage with 2019-nCoV originated from bat. However, this may be due to the incomplete data that is used, or other issues, and does not suffice to conclude that SARS-CoV-2 originated from bats. Moreover, many recent studies explored the origin of SARS-CoV-2 in depth. The authors should cite these and compare against their results and conclusions.
7. The authors write *"Since the S protein is the protein that exhibits the highest degree of genetic variations among different coronaviruses"* Is that true? It needs to be cited or supported with data.
8. The authors conclude that the infectivity and pathogenicity of SARS-CoV-2 should be much lower than that of SARS-CoV, and that its incubation time should be longer. However, this too seems to contradict most recent studies, which show that the infectivity is not lower (and may be higher) than that of SARS-CoV, and most studies suggest a similar incubation period for SARS-CoV and SARS-CoV-2. These discrepancies should be addressed, and the conclusions supported by recent literature.
9. The authors further write that *"Analysis of the potential interaction of RBD of 2019-nCoV with human ACE2 receptor protein indicated that its affinity to human cell is much lower than that of human SARS virus"*, again, there is a lot of recent work exploring the interaction of RBD of SARS-CoV-2 with human receptors, which seem to report contradicting evidence. This authors should discuss the differences and compare to recent literature.

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational Biology, Bioinformatics, Computer science, Systems Biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 07 April 2020

<https://doi.org/10.5256/f1000research.25120.r61587>

© 2020 Aguilar H et al. This is an open access peer review report 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.

Fikret Aydemir

Cornell University, Ithaca, New York, USA

Hector Aguilar

Department of Microbiology and Immunology, College of Veterinary Medicine, Cornell University, Ithaca, NY, USA

Dong *et al.* examined the available sequences of coronaviruses to depict the origin and the pathogenicity of 2019-nCoV that caused the current global pandemic. While the data presented in the manuscript could be of interest, several concerns need to be addressed. These concerns presented below.

Major:

1. Main findings in this study suggest that the new virus should theoretically not be able to cause very serious human disease as compared to other coronaviruses. The mechanism(s) of lowering the pathogenicity, increase human adaptation, or increase incubation times as compared to other strains are not explained, addressed, or cited.
2. The epidemiological data are not explained, addressed, or cited.
3. The reference for accession # MN908947, the very main isolate obtained from Wuhan, is repeatedly not cited.
4. The center that obtained the isolate is not mentioned or cited.
5. What is the exact sequence of the whole genome? Where is it referenced?
6. Clear English writing/editing is needed throughout the paper.

7. The limited human transmission potential needs to be explained and cited.
8. The intermediate transmission vehicle for other coronaviruses, if exists, needs to be explained and cited.
9. The relationship between sequence information and lower pathogenicity needs to be explained and cited.
10. The relationship between sequence information, receptor binding, human adaptation, and mutations needs to be explained and cited.

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and is the work technically sound?

No

Are sufficient details of methods and analysis provided to allow replication by others?

No

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

No

Are the conclusions drawn adequately supported by the results?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Virology, Biochemistry, Molecular Biology, Vaccine Development

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

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

F1000Research