Briefings in Bioinformatics, 00(00), 2020, 1–9

doi: 10.1093/bib/bbaa154
Problem Solving Protocol

Application of Bayesian phylogenetic inference modelling for evolutionary genetic analysis and dynamic changes in 2019-nCoV

Tong Shao†, Wenfang Wang†, Meiyu Duan, Jiahui Pan, Zhuoyuan Xin, Baoyue Liu, Fengfeng Zhou and Guoqing Wang

Corresponding authors: Guoqing Wang, College of Basic Medical Science, Jilin University, Changchun, China. Tel.: +86 0431-85167458. E-mail: qing@jlu.edu.cn; Fengfeng Zhou, College of Computer Science and Technology, Jilin University, Changchun, China. E-mail: FengfengZhou@gmail.com
†These authors contributed equally to this work.

Abstract

The novel coronavirus (2019-nCoV) has recently caused a large-scale outbreak of viral pneumonia both in China and worldwide. In this study, we obtained the entire genome sequence of 777 new coronavirus strains as of 29 February 2020 from a public gene bank. Bioinformatics analysis of these strains indicated that the mutation rate of these new coronaviruses is not high at present, similar to the mutation rate of the severe acute respiratory syndrome (SARS) virus. The similarities of 2019-nCoV and SARS virus suggested that the S and ORF6 proteins shared a low similarity, while the E protein shared the higher similarity. The 2019-nCoV sequence has similar potential phosphorylation sites and glycosylation sites on the surface protein and the ORF1ab polyprotein as the SARS virus; however, there are differences in potential modification sites between the Chinese strain and some American strains. At the same time, we proposed two possible recombination sites for 2019-nCoV. Based on the results of the skyline, we speculate that the activity of the gene population of 2019-nCoV may be before the end of 2019. As the scope of the 2019-nCoV infection further expands, it may produce different adaptive evolutions due to different environments. Finally, evolutionary genetic analysis can be a useful resource for studying the spread and virulence of 2019-nCoV, which are essential aspects of preventive and precise medicine.

Key words: 2019-nCoV; COVID-19; Bayesian model; genetic evolutionary analysis; pneumonia

Tong Shao is a PhD student in College of Basic Medical Science, Jilin University. Her research interests include bioinformatics and virological analysis.
Wenfang Wang is a Graduate student in College of Basic Medical Science, Jilin University. Her research interests include bioinformatics and integration analysis of virus.
Meiyu Duan is a PhD student in College of Computer Science and Technology, Jilin University. Her research interests include bioinformatics.
Jiahui Pan is a PhD student in College of College of Basic Medical Science, Jilin University. Her research interests include bioinformatics.
Zhuoyuan Xin is a lecturer in College of College of Basic Medical Science, Jilin University. His research interests include bioinformatics and comparative genomics analysis.
Baoyue Liu is a bachelor student in College of Basic Medical Science, Jilin University. Her research interests include virus and immune system.
Fengfeng Zhou is a professor at College of Computer Science and Technology, Jilin University, Changchun, Jilin, China. His research interests include bioinformatics and systems biology.
Guoqing Wang is a professor at the Department of Pathogenobiology, College of Basic Medicine, Jilin University, Changchun, Jilin, China. His research interests include bioinformatics and systems biology.
Submitted: 11 March 2020; **Received (in revised form):** 26 May 2020

Introduction

In December 2019, cases of a new coronavirus infection were reported in Wuhan, Hubei Province, China. The main clinical manifestations are a series of symptoms, such as fever and pneumonia, which can lead to severe breathing difficulties [1, 2]. In January 2020, the World Health Organization (WHO) named the new coronavirus '2019-nCoV.' In February 2020, the WHO designated 2019-nCoV–caused pneumonia 'COVID-19.' By 4 April 2020, there were 2 329 539 confirmed cases, of which 160 717 patients died; a fatality rate of 2–4% [3].

2019-nCoV is an enveloped positive-strand single-stranded RNA virus belonging to subfamily *Coronavirinae*, family *Coronaviridae*, order *Nidovirales*. There are four genera of coronavirus (*CoV*): *Alphacoronavirus* (α*CoV*), *Betacoronavirus* (β*CoV*), *Deltacoronavirus* (δ*CoV*) and *Gammacoronavirus* (γ*CoV*) [1]. The main infection targets of α*CoV* and β*CoV* are mammals; δ*CoV* and γ*CoV* mostly infect birds [4]. Within the last two decades, *CoV* has given rise to two large-scale pandemics: severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) [5]. These disease outbreaks prompt people to attach great importance to *CoV*. According to the results of phylogenetic analysis, 2019-nCoV, like the SARS, belongs to β*CoV* genus and is most similar to the SARS-like coronavirus from bats, with a nucleotide homology of 84%. It has a homology with the human SARS virus of 78% and with the MERS virus of ~50% [6, 7].

At present, there are few studies investigating the recombination events of new coronaviruses. Among them, studies have found that the whole genome of 2019-nCoV is very similar to the rat coronavirus RaTG13, but at the same time, the receptor-binding domain (RBD) of its S protein is closer to the Pangolin-CoV isolated in Guangdong, China [8]. A research group also found that the S protein of 2019-nCoV has a multifunctional cleavage site at the S1-S2 boundary [9], and another research found that S protein has two possible recombination points [10].

In this study, the differences between the amino acid composition of the new coronavirus and that of the SARS virus, their recombination with other coronaviruses and their molecular evolutionary rules were studied to identify the origin of the strong infectivity of 2019-nCoV and provide a basis for the prevention of this disease.

Materials and methods

Sequence collection

For Bayesian analysis, 777 total 2019-nCoV gene sequences as of 29 February 2020 were downloaded from GISAID [11]. A phylogenetic tree does not allow a sequence with significant recombination events. This study used the Recombination Detection Program version 4 (RDP4) method [12] to detect and remove those strains with recombined signals. Finally, 746 strains were kept for constructing the phylogenetic tree, in a similar protocol [13, 14]. These gene sequences were the complete 2019-nCoV sequences. The detailed sequence information and the acknowledgments of their original contributors are listed in Supplementary Table 1. Other nucleotide sequences were retrieved from GenBank [15].

Sequence analysis and comparison of 2019-nCoV

DNASTAR Lasergene 8.0 was used to compare the homology of the nucleotide sequences of 2019-nCoV. The similarity of amino acid levels between different viruses and the 2019-nCoV (MN908947) was also compared. Three other coronaviruses were selected: Bat coronavirus RaTG13 (MN996532), Bat-SLCovZXC21 (MG772934) and Pangolin coronavirus PCoV_GX-P5E (MT040336).

Analysis of potential protein modification sites

The NetOGlyc 4.0 Server (http://www.cbs.dtu.dk/services/NetOGlyc/) was used to estimate the O-linked glycosylation of the surface protein and ORF1ab polyprotein. The server is a support vector machine used to produce a predictor of O-GalNAc glycosylation. GlycoMine [16] was used to predict the C-linked glycosylation (http://glycomine.erc.monash.edu/Lab/GlycoMine/). GPS 5.0 [17] was used to predict phosphorylation. A total of 150 ORF1ab polyprotein amino acids sequences were predicted, using samples from 14 different countries. A total of 121 S protein from were predicted, using samples from 12 different countries. The location information of the strains is listed in Supplementary Table 2.

Analysis of the obtained viral genome data using the Bayesian method

The Bayesian analysis method was used to study the evolution rate and evolution model of the recent outbreaks of the 2019-nCoV strain. The complete gene sequence alignment of the 2019-nCoV was carefully performed using the ClustalW program in MEGA 5.0 [18]. All sequences were analysed using RDP4 [12] to detect recombination using a variety of methods, including RDP, GENECONV, BootScan, maximum chi square, Chimera, SISCAN and 3SEQ. We ensured that no intragroup recombination occurred in the strains for evolutionary tree analysis. Putative recombination events were investigated using two methods: RDP4 [12] and Simplot v3.5.1 [19].Then, to test the saturation monitoring, sequences were screened using Date Analysis and Molecular Biology and Evolution (DAMBE) [20]. If the result is ISS < ISS.c, then the sequence substitution was not saturated, and it meets the requirements for building a phylogenetic tree using Bayesian methods. Finally, the best evolution model is selected using the IQ-TREE Web Server (http://iqtree.cibiv.univie.ac.at/). It selects the most suitable model for constructing the evolutionary tree of these strains. We will choose this model in the next analysis. Through BEAST v2.6 [21] under the GTR (General Time-Reversible) + I model of nucleotide substitutions and a Strict clock. For the strict clock model, the rate of evolution is the same for all lineages of the phylogenetic tree. Then, we used 100 million Markov chain Monte Carlo (MCMC) runs to construct a maximum clade credibility (MCC) tree (effective sampling size > 200). The analysis was sampled every 10 000 states. Posterior probabilities were calculated with a burn-in of 10 million states. These parameters are the optimal choices. The analysis of sampling data was output by Tracer v1.6 [22] and then the Tree Annotator program was employed to output the results of the MCC tree model. In the end, the Fig Tree v1.4.2 program (http://tree.bio.ed.ac.uk/software/figtree/) can illustrate the MCC molecular evolutionary tree. The significance level is 0.05. The default values of all the other parameters were used, assuming that the popular programs used in this study have already evaluated the rationale of these parameters. The limitation is that the extreme values of some parameters may cause large changes in the experimental results.

Results

Homologous comparison of the 2019-nCoV sequences

The phylogenetic analysis of the major coding regions of representative members of the *Sarbecovirus* subgenus indicated that 2019-nCoV and SARS virus and human coronaviruses 229E, NL63, OC43 and HKU1 are within the same evolutionary branch (Figure 1A) and indicated that the new virus belongs to

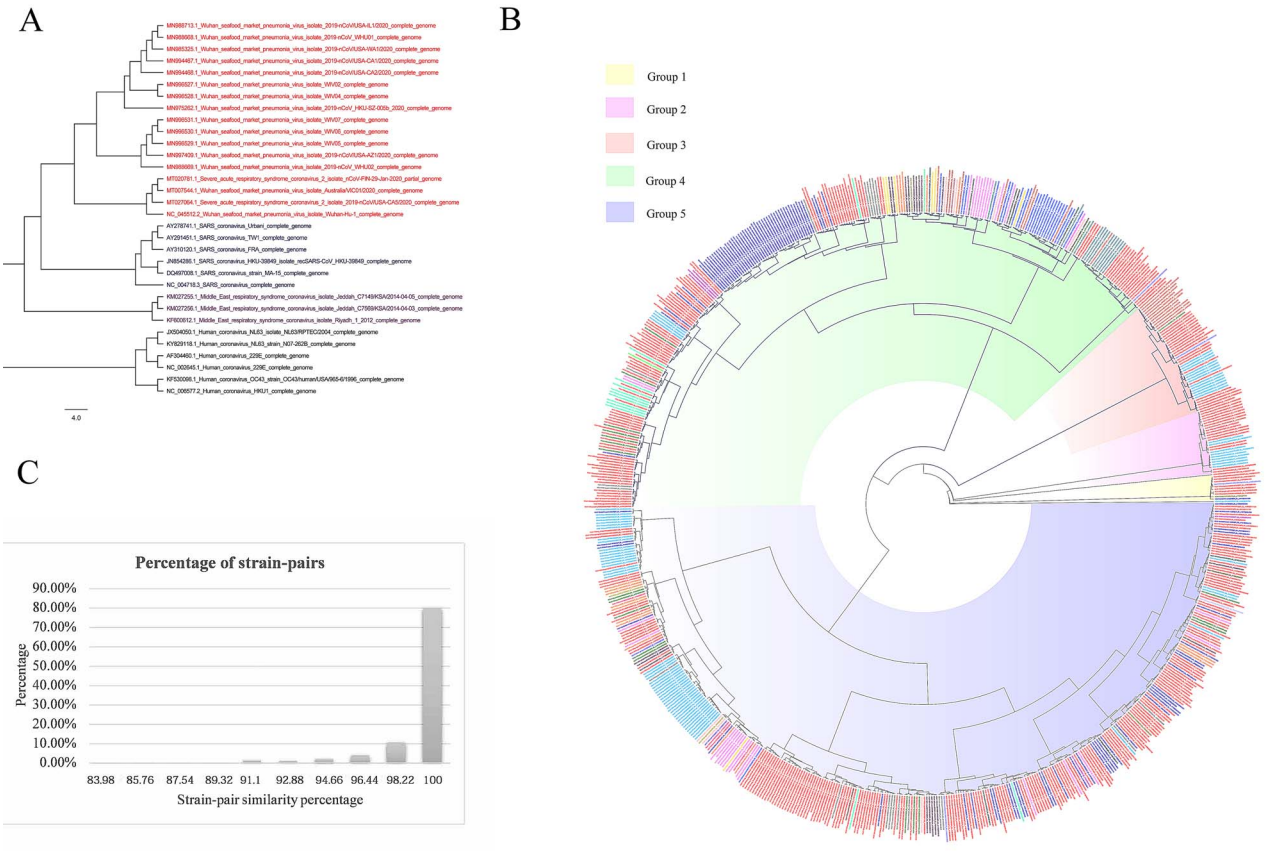

Figure 1. Genetic analysis of novel coronavirus (2019-nCoV), severe acute respiratory syndrome (SARS) virus and human coronaviruses 229E, NL63, OC43 and HKU1. Phylogenetic analysis of the 2019-nCoV gene: (A) coronavirus phylogenetic tree and (B) phylogenetic tree of 746 strains of 2019-nCoV. Strains from different regions are shown in different colours. Those marked red are from China. (C) Histogram of similarity distributions between all the strain pairs. The minimum and maximum pair-wise similarities were 82.2% and 100%, respectively.

Sarbecovirus, which is the SARS virus subgenus. The sequence similarity of the 15 strains of 2019-nCoV reached 99.5%. Although the new coronavirus belongs to the SARS virus evolutionary branch, it is a unique evolutionary branch, which indicates that 2019-nCoV is a new beta coronavirus from the *Sarbecovirus* subgenus. All strains collected in this study were subjected to homologous sequence alignment. From the results (Supplementary Table 3) we observed that nucleotide similarity was 82.2–100.0%. This suggests that 2019-nCoV has developed a low level of change during its expansion in the population. Phylogenetic analysis was conducted on all 746 isolated strains. These strains were analysed through the phylogenetic tree using the Bayesian method (Figure 1B). The Chinese strains were mainly in Group 5 and the US strains were mainly in Group 4. Most of the European strains were in Group 4. The fifth group was dominated by strains from Asia. A significant portion of the South Korean strain came from the fourth group. The virus strains in each region are intermingled with each other and do not individually form a cluster, which means that 2019-nCoV has been in a state of widespread human mobility.

Amino acid sequence alignment

After comparing the amino acid similarities between different proteins within the 2019-nCoV and SARS viruses (Figure 2A), we observed the similarity of ORF1ab is 86%, that of S is 76%, that of ORF3a is 72%, that of E is 95%, that of ORF6 is 69%, that of ORF7a is 85% and that of N is 91%. In the above results, E has the highest amino acid identity, and ORF6 has the lowest. S (76% identity) in the coronavirus genome encodes spike proteins on the surface of the virus. Whether these differences in genetic similarity affect the infectivity of the virus deserves further investigation. At the same time, we also compared the amino acid similarity of 2019-nCoV with two kinds of bat coronaviruses and one kind of pangolin coronavirus (Figure 2B–D). From the results, we can see that the similarity of E protein is still the highest among the three strains, all of which are 100%. By comparison with the two bat strains, we can see that the amino acid sequence similarity between the new coronavirus and RaTG13 is very high, the S protein similarity is 98%. For strain SLCoZXC21, the lowest similarity was also observed for the S protein (80%). Compared to the strain from pangolin, the 2019-nCoV has the lowest similarity of 88% ORF7a.

Prediction of glycosylation and phosphorylation sites

We conducted phosphorylation and two types of glycosylation, (O-linked and C-linked) modifications on 150 ORF1ab polyproteins (Figure 3A) and 121 surface proteins (Figure 3B), and displayed the results on a heat map. These three types of post-translational modifications (PTMs) demonstrated different

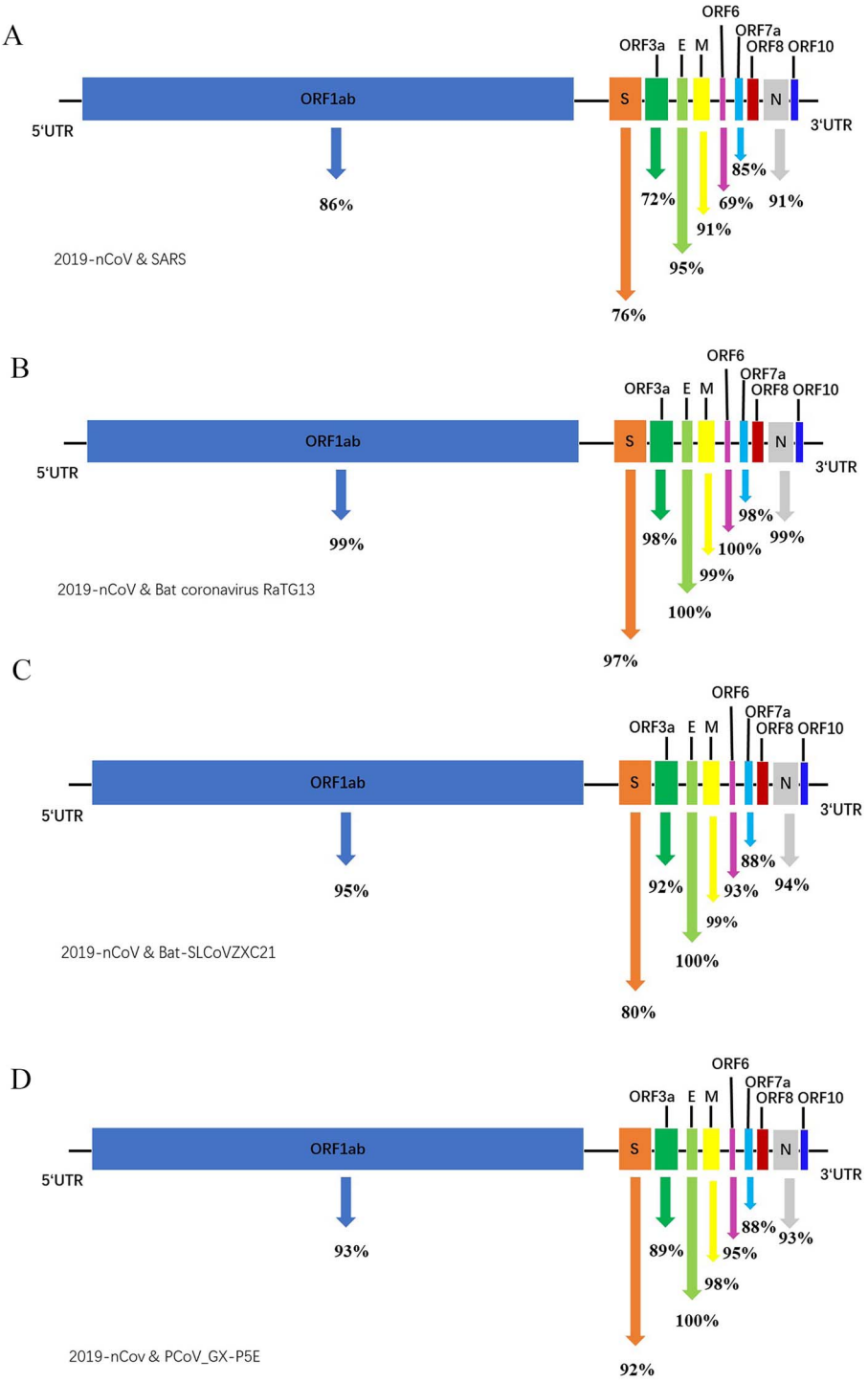

Figure 2. Coronavirus genome comparison. (A) A comparison between severe acute respiratory syndrome (SARS) virus (FJ882963) and novel coronavirus (2019-nCoV) (MN908947). (B–C) Comparison of coronavirus 2019-nCoV and two bat coronaviruses (RaTG13 MN996532 and Bat-SLCoVZX21 MG772934). (D) Comparison of 2019-nCoV and pangolin coronavirus (MT040336).

distributions of prediction scores, as shown in Figure 3. The result of the modification is predicted by three approaches, and the scores of different sites are presented using heat maps. The results showed that there was no significant difference in the potential modification sites of the surface proteins of these 121 strains. However, it is worth noting that the surface protein prediction results of seven strains (QHS34546 from India; QIM47476, QIM47457, QIQ08800 from Spain; QIQ49922, QIQ49902.1, QIQ50102 from the USA) are different from those of the other strains. As for the prediction of open reading frame ORF1ab modification sites, there were some differences in the potential sites of ORF1ab proteins in 150 strains. Some of the strains that are quite different from other strains come from the USA.

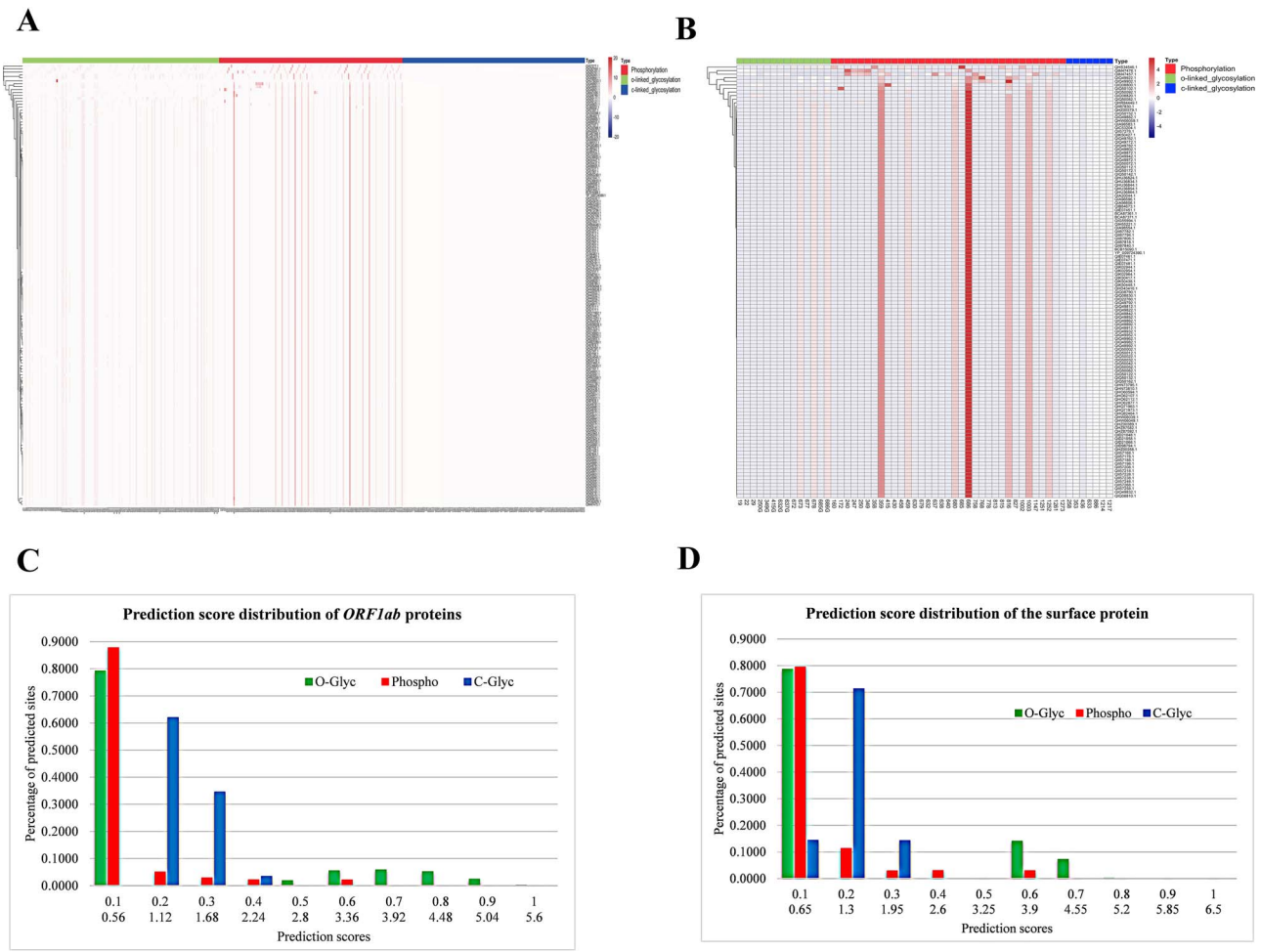

Figure 3. Heat map of predicted phosphorylation sites and N- and C-linked glycosylation sites. The ordinate of the heat map represents the different strains, and the abscissa represents the potential modification sites after the prediction. C-linked glycosylation in blue, O-linked glycosylation in green and phosphorylation in red. Different sites have different scores. The sites with high scores are shown in red. The higher the score, the darker the red. (A) Prediction scores of ORF1ab. (B) Prediction scores of the surface protein. (C) Prediction score distribution of the three PTM types on ORF1ab. (D) Prediction score distribution of the three PTM types on the surface protein. The two scores for each horizontal axis point were the upper bounds of the prediction scores of glycosylation and phosphorylation sites, respectively. The terms O-Glyc, C-Glyc and Phospho refer to the O-linked and C-linked glycosylation sites and phosphorylation sites, respectively.

The prediction scores of O-linked glycosylation and phosphorylation sites were relatively larger than those of the C-linked glycosylation scores for the ORF1ab and surface proteins, as shown in Figure 3C and D. Most candidate O-linked glycosylation sites and phosphorylation sites of the ORF1ab protein were predicted to have scores of zero, so that the first bins of these two PTM types were very high. The predicted positive PTM sites of the ORF1ab protein had large scores ranging between 0.5 and 0.9. The majority of the predicted positive C-linked glycosylation sites have scores ranging between 0.1 and 0.4, as shown in Figure 3C. Similar patterns were observed for these three PTM types on the surface proteins, as shown in Figure 3D. The biochemical properties of the amino acids flanking the candidate PTM sites play an essential role in determining the prediction scores, which were positively correlated with the probabilities of real PTM modifications [23, 24]. However, the prediction scores are only meaningful for the same PTM type and are not comparable between different PTM types.

Additional comparisons with SARS and MERS viruses indicated that the modification sites on these viruses were completely different from those on 2019-nCoV.

Recombinant analysis of the virus strains

After recombination testing of SARS virus, MERS virus and 2019-nCoV, and of human coronaviruses HKU1, 229E and OC43 (Figure 4), we observed that the recombinant sequence *human coronavirus HKU1* (NC_006577.2) was more closely related to the main parental sequence 2019-nCoV (NC_0445512.2) with a probability of 41.2%. The probability of a closer relationship with the secondary parental sequence was 66.3%. We found two possible recombination sites throughout the genome. They are 13 208 and 21 743, respectively. Further experiments are needed to verify whether recombination occurs. The recombination site contains mainly *ORF1b*. The proteins identified in this framework are NSP12, NSP14, NSP15 and NSP16, which play

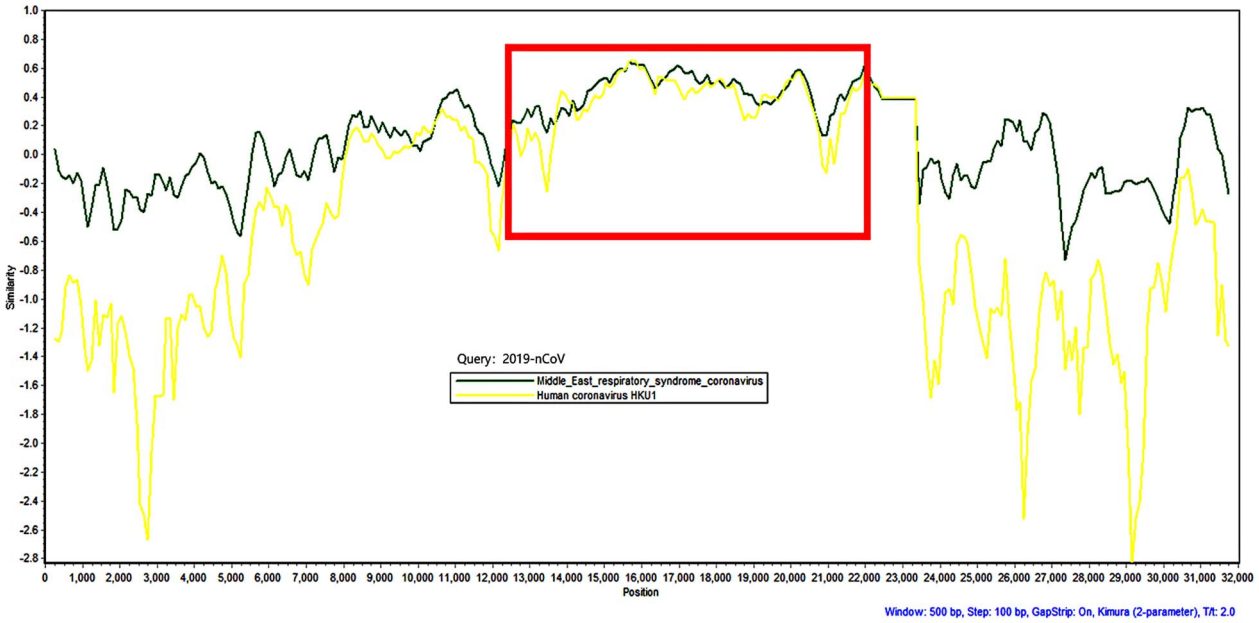

Figure 4. Recombination analysis of 2019-nCoV. Plots of similarity (generated by SimPlot) among human coronavirus HKU1, novel coronavirus (2019-nCoV) and Middle East respiratory syndrome (MERS) virus. Different colours correspond to the nucleotide similarity between the 2019-nCoV and different groups. The regions with discordant phylogenetic clustering of the 2019-nCoV with human coronavirus HKU1 sequences are shown in different colours.

important roles in the life cycle of the virus and are involved in virus replication and transcription. The other recombinant site is located on the S protein. The S protein is responsible for the interaction between the virus and the host.

Evolutionary tree construction based on the Bayesian–Markov chain method

From the results, all three codon positions had different relative substitution rates: the mean values of the first, second and third codon positions were 0.9759, 1.0354 and 0.9887, respectively (Figure 5A–B). Among these codon positions, the relative substitution rate of the second codons was the highest. From the results of the skyline, we speculate that the activity of the gene population of 2019-nCoV may be before the end of 2019, and then we estimated the time to a most recent common ancestor (TMRCA) dates for 2019-nCoV, that is 29 July 2019, 95% HPD interval (5 February 2019 and 28 November 2019) (Figure 5A). We constructed the SARS virus skyline in the same way, selecting 267 strains of SARS virus (Supplementary Figure 1). From the figure, we can see that the population of SARS virus increased in a small area in 2002, and then decreased in 2003 after reaching a high point. This is consistent with the actual situation we have known, and it also indicates that the prediction of the skyline for species and population is persuasive to some extent.

Discussion

The mutation rate was not high for the 2019-nCoV nucleotide sequences collected, which indicated that the virus had not extensively mutated and was still in a stage of stable transmission. In this study, we compared the amino acid sequences of the proteins encoded by *ORF1ab, S, ORF3a, E, ORF6, ORF7a* and *N* of 2019-nCoV and SARS virus, respectively. The similarity with E was highest and the similarity with ORF6 was the lowest. The E protein is not necessary for CoV genomic replication or subgenomic mRNA synthesis, but it can affect virus morphogenesis, budding, assembly, intracellular transport and virulence, and is one of the causes of acute respiratory distress syndrome (ARDS) [25]. Compared with SARS virus, the E protein of 2019-nCoV has the highest similarity, and both have ARDS symptoms, indicating that the E protein of 2019-nCoV may have the same function, but further research is needed. The ORF6 protein of SARS virus, which encodes a 7 kDa protein with a hydrophobic N-terminus, is considered to have an N-endo-C-endo conformation [26]. Some researchers have performed functional studies on the ORF6 protein and found that it can interact with the non-structural protein 8 (NSP8) in the SARS virus replicase complex [27], which can increase the infection titre during early infection with a lower multiplicity of infection and by interacting with the nuclear transporter $\alpha 2$ to inhibit the rate of cellular gene synthesis [28, 29], interferon production [30] and nuclear translocation of signal transducer and activator of transcription 1 (STAT1) [31]. Whether the new ORF6 of coronavirus has different functions needs further study.

Another study has shown [9] that 2019-nCoV uses hACE2 as the receptor for entering the host, which has a similar affinity to SARS isolates from 2002 to 2003. At the same time, the article pointed out that research [32] has shown that the increase in the strength of the binding force between SARS S and hACE2 is related to the increase in human viral transmission capacity and the severity of the disease. For the current status, the number of 2019-nCoV infections is higher than that of SARS. Whether this is related to the difference between S genes needs further study. Notably, the similarity to ORF6 was only 69%. Purnima Kumar *et al*. [33] reported for the first time [33] on the interaction between ORF6 and nsp8 in SARS virus and found that ORF6 protein may play a role in virus replication. At the same time, a previous study [34] showed that the expression of ORF6 during infection may play an important role in the pathogenesis of the virus. Therefore, the similarity between the 2019-nCoV and SARS virus ORF6 is not high, and whether it will affect the

A

Summary statistic	CP1.mu	CP2.mu	CP3.mu	TMRCA
Mean	0.9759	1.0354	0.9887	29/7/19
95% HPD lower	0.8795	0.9429	0.8987	5/2/19
95% HPD upper	1.0620	1.1241	1.0806	28/11/19

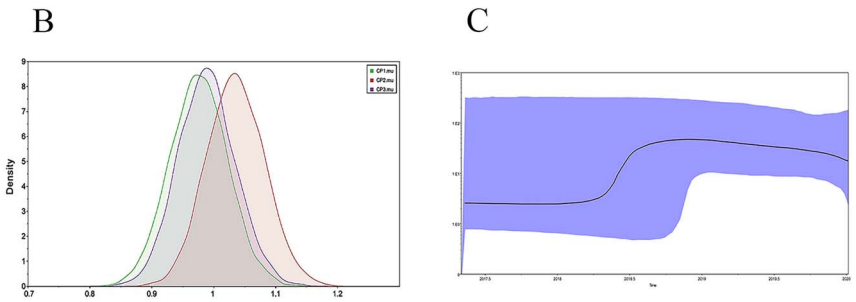

Figure 5. Codon mutation rate and TMRCA of 2019-nCoV and skyline plot. (A) The codon substitution rate and TMRCA of novel coronavirus (2019-nCoV) were estimated using BEAST. (B) The codon substitution rate of 2019-nCoV was estimated using the Bayes-Markov chain method and is the result of BEAST run using Tracer analysis. (C) Dynamic study of the 2019-nCoV genetic diversity using the Bayesian skyline plot. The thick solid line is the median estimate and the dotted line shows the 95% confidence interval. The abscissa is time and the ordinate is the effective population size.

pathogenesis and pathogenicity of the 2019-nCoV is worthy of further study.

E protein, with 95% similarity, is involved in virus assembly, budding and envelope formation. It plays an important role in the generation and maturation of viruses [35]. Coronaviruses lacking E are promising candidates for vaccine development [35]. Some teams have found in their research on SARS [36] and MERS [37] viruses that the mutation of the E protein or the use of coronaviruses lacking E protein may be possibilities for a live attenuated vaccine. Therefore, it is worth continuing to explore whether the E protein, with its high similarity between novel coronavirus and SARS virus, can also be a research direction for a live attenuated vaccine. The main function of nucleocapsid proteins is to package viral genome RNA molecules into ribonucleoprotein (RNP) complexes, which are responsible for the replication of the virus [38]. Both E and S proteins play an essential role in the viral life cycle. One study [8] has shown that the similarity between the whole genome of 2019-nCoV and RaTG13 is very high, while the RBD region of its S protein is closer to pangolin-CoV from Guangdong, China. In the different coronaviruses we compared, the similarity of each protein of 2019-nCoV to RaTG13 was very high; the lowest was 97%. In our results, the full-length amino acids of S protein were 97% similar when compared to RaTG13 and 92% similar when compared to pCoV-GX-P5E from pangolin. However, looking solely at the RDB region of the S protein [8], this region is more closely related to the pangolin virus.

We analysed the recombination of 2019-nCoV. It was found that there was no recombination relationship with the SARS virus. Furthermore, after recombination testing on SARS virus, MERS virus and 2019-nCoV, and on human coronaviruses HKU1, 229E and OC43, we observed that the recombinant sequence *human coronavirus HKU1* (NC_006577.2) was most closely related (probability 41.2%) to the main parental sequence of 2019-nCoV (NC_0445512.2). The probability of a closer relationship with the secondary parental sequence was 66.3%.

At the same time, we found two potential recombinant sites of 2019-nCoV, 13 208 and 21 473, which encode ORF1b and S proteins. Previous studies [10] focused on the recombination of S proteins and found that S protein may have two potential recombination sites. This fact indicates that the nucleotides in this part are largely obtained by recombination, which is worthy of further analysis and verification.

By predicting viral protein modification sites, we observed that some virus strains have different potential sites. Proteins can be phosphorylated during enzyme action and mediate protein activity. With glycosylation, proteins can resist digestive enzymes, giving them the ability to transmit signals, and some fold properly only after glycosylation. Whether the potential modification sites of different strains have a certain influence on the life cycle and activity of the virus deserves further study. The sequences of the Chinese strain and the US strain are not significantly different, but there are differences in potential modification sites, which may cause changes in the glycosylation and phosphorylation sites of surface proteins and ORF1ab polyproteins, which may affect the viral genome, and the stability and function of the protein structure may have adaptive value in the process of virus transmission. These predicted possible modification sites also need to be studied experimentally to determine whether they can be used for new coronaviruses.

Based on the Bayesian evolution analysis method, we constructed a phylogenetic tree for the 746 strains collected and analysed the codons and skylines. From the results of skyline analyses, we speculated that 2019-nCoV may have been active as early as the end of 2019. The mutation rate of the code may prove that certain mutations occurred during early viral transmission [39], which led to the current large-scale outbreak of human infection. At present, because of the 2019-nCoV sequences we selected are in the early stage of the outbreak, the results have certain limitations. With the emergence of more sequences and studies, the understanding of this virus will be further developed.

Conclusion

At present, the 2019-nCoV collected in this research has not undergone a large number of mutations and is in the stage of stable transmission. However, by analysing the data up to the

end of February, we found that the virus may have shown signs of activity before the end of 2019, and some mutations may have occurred during the early transmission process that led to the outbreak of COVID-19. In addition, we speculate that there may be recombination between 2019-nCoV and human coronavirus HKU1. We also found that 2019-nCoV has the lowest similarity to ORF6 of the SARS virus. Whether the large ORF6 difference is one of the reasons for the different infectivity intensities of 2019-nCoV and other coronaviruses deserves further analysis. By predicting the modification sites of the virus surface protein and ORF1ab polyprotein, we found some potential modification sites for 2019-nCoV. The above findings will provide new ideas for further study of 2019-nCoV. Currently, 2019-nCoV continues to cause infection worldwide. In the future, new directions of evolution may appear due to environmental impacts. This work is basically a proof-of-principle study, and further functional analysis will be needed in future investigations.

Key Points

- We found that during the evolution of 2019-nCoV, ORF6 has a large variation and a high evolutionary dynamic efficiency.
- Based on the results of the skyline and TMRCA, we speculate that 2019-nCoV may have been active in early and mid-2019.
- We analysed the key role of the 2019-nCoV surface, ORF1ab and other protein modification sites in the 2019-nCoV evolutionary dynamics. Differences in modification sites between Chinese and American strains may suggest the virus's adaptability during transmission.

Supplementary data

Supplementary data mentioned in the text are available to subscribers in *Briefings in Bioinformatics*.

Acknowledgements

We would like to thank Editage (www.editage.cn) for English language editing. The insightful comments from the three anonymous reviewers were also greatly appreciated.

Funding

The epidemiology, early warning and response techniques of major infectious diseases in the Belt and Road Initiative [#2018ZX10101002]; National Natural Science Foundation of China [#81871699]; Foundation of Jilin Province Science and Technology Department [#172408GH010234983].

References

1. Chan JF, Yuan S, Kok KH, *et al*. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* 2020;**395**:514–23.
2. Huang C, Wang Y, Li X, *et al*. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020;**395**:497–506.
3. Rao H, Shi X, Rodrigue AK, *et al*. Feature selection based on artificial bee colony and gradient boosting decision tree. *Appl Soft Comput* 2019;**74**:634–42.
4. Tang Q, Song Y, Shi M, *et al*. Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. *Sci Rep* 2015;**5**:17155.
5. Corman VM, Muth D, Niemeyer D, *et al*. Hosts and sources of endemic human coronaviruses. *Adv Virus Res* 2018;**100**:163–88.
6. 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;**395**:565–74.
7. Zhang L, Shen FM, Chen F, *et al*. Origin and evolution of the 2019 novel coronavirus. *Clin Infect Dis* 2020; ciaa112.
8. Wu A, Niu P, Wang L *et al*. Mutations, recombination and insertion in the evolution of 2019-nCoV, *bioRxiv* 2020; 2020.02.29.971101.
9. Walls AC, Park YJ, Tortorici MA, *et al*. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 2020;**181**:281–292 e286.
10. Wu F, Zhao S, Yu B, *et al*. A new coronavirus associated with human respiratory disease in China. *Nature* 2020;**579**:265–9.
11. Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* 2017;**22**:30494.
12. Martin DP, Murrell B, Golden M, *et al*. RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol* 2015;**1**:vev003.
13. Xu W, Weese JS, Ojkic D, *et al*. Phylogenetic inference of H3N2 canine influenza a outbreak in Ontario, Canada in 2018. *Sci Rep* 2020;**10**(1):6309.
14. Aiewsakun P, Richard L, Gessain A, *et al*. Modular nature of simian foamy virus genomes and their evolutionary history. *Virus Evol* 2019;**5**:vez032.
15. Sayers EW, Cavanaugh M, Clark K, *et al*. GenBank. *Nucleic Acids Res* 2020;**48**:D84–6.
16. Li F, Li C, Wang M, *et al*. GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. *Bioinformatics* 2015;**31**: 1411–9.
17. Wang C, Xu H, Lin S, *et al*. GPS 5.0: an update on the prediction of kinase-specific phosphorylation sites in proteins. *Genomics Proteomics Bioinformatics* 2020;S1672-0229(20) 30027-9
18. Tamura K, Peterson D, Peterson N, *et al*. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011;**28**:2731–9.
19. Lole KS, Bollinger RC, Paranjape RS, *et al*. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J Virol* 1999;**73**:152–60.
20. Xia X. DAMBE7: new and improved tools for data analysis in molecular biology and evolution. *Mol Biol Evol* 2018;**35**:1550–2.
21. Bouckaert R, Vaughan TG, Barido-Sottani J, *et al*. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 2019;**15**(4):e1006650.
22. Rambaut A, Drummond AJ, Xie D, *et al*. Posterior summarization in Bayesian Phylogenetics using tracer 1.7. *Syst Biol* 2018;**67**:901–4.
23. Zhou F, Xue Y, Yao X, *et al*. A general user interface for prediction servers of proteins' post-translational modification sites. *Nat Protoc* 2006;**1**:1318–21.

24. Zhou FF, Xue Y, Chen GL, *et al*. GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem Biophys Res Commun* 2004;**325**:1443–8.
25. DeDiego ML, Nieto-Torres JL, Jimenez-Guardeno JM, *et al*. Coronavirus virulence genes with main focus on SARS-CoV envelope gene. *Virus Res* 2014;**194**:124–37.
26. Zhou H, Ferraro D, Zhao J, *et al*. The N-terminal region of severe acute respiratory syndrome coronavirus protein 6 induces membrane rearrangement and enhances virus replication. *J Virol* 2010;**84**:3542–51.
27. Kirchdoerfer RN, Ward AB. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nat Commun* 2019;**10**:2342.
28. Zhao J, Falcon A, Zhou H, *et al*. Severe acute respiratory syndrome coronavirus protein 6 is required for optimal replication. *J Virol* 2009;**83**:2368–73.
29. McBride R, Fielding BC. The role of severe acute respiratory syndrome (SARS)-coronavirus accessory proteins in virus pathogenesis. *Viruses* 2012;**4**:2902–23.
30. Kindler E, Thiel V, Weber F. Interaction of SARS and MERS coronaviruses with the antiviral interferon response. *Adv Virus Res* 2016;**96**:219–43.
31. Huang SH, Lee TY, Lin YJ, *et al*. Phage display technique identifies the interaction of severe acute respiratory syndrome coronavirus open reading frame 6 protein with nuclear pore complex interacting protein NPIPB3 in modulating type I interferon antagonism. *J Microbiol Immunol Infect* 2017;**50**:277–85.
32. Li W, Zhang C, Sui J, *et al*. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *EMBO J* 2005;**24**:1634–43.
33. Kumar P, Gunalan V, Liu B, *et al*. The nonstructural protein 8 (nsp8) of the SARS coronavirus interacts with its ORF6 accessory protein. *Virology* 2007;**366**:293–303.
34. Pewe L, Zhou H, Netland J, *et al*. A severe acute respiratory syndrome-associated coronavirus-specific protein enhances virulence of an attenuated murine coronavirus. *J Virol* 2005;**79**:11335–42.
35. Schoeman D, Fielding BC. Coronavirus envelope protein: current knowledge. *Virol J* 2019;**16**(1):69.
36. Regla-Nava JA, Nieto-Torres JL, Jimenez-Guardeno JM, *et al*. Severe acute respiratory syndrome coronaviruses with mutations in the E protein are attenuated and promising vaccine candidates. *J Virol* 2015;**89**:3870–87.
37. Almazan F, DeDiego ML, Sola I, *et al*. Engineering a replication-competent, propagation-defective Middle East respiratory syndrome coronavirus as a vaccine candidate. *MBio* 2013;**4**:e00650–13.
38. Chang CK, Hou MH, Chang CF, *et al*. The SARS coronavirus nucleocapsid protein–forms and functions. *Antiviral Res* 2014;**103**:39–50.
39. Jiang L, Zhang M, Sang M, *et al*. Evo-Devo-EpiR: a genome-wide search platform for epistatic control on the evolution of development. *Brief Bioinform* 2017;**18**:754–60.