Briefings in Bioinformatics, 00(00), 2021, 1–9

doi: 10.1093/bib/bbab042
Case Study

Functional alterations caused by mutations reflect evolutionary trends of SARS-CoV-2

Liang Cheng †, Xudong Han†, Zijun Zhu, Changlu Qi, Ping Wang and Xue Zhang

Corresponding authors: Liang Cheng, NHC and CAMS Key Laboratory of Molecular Probe and Targeted Theranostics, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, PR China. Tel.: +15303614540; E-mail: liangcheng@hrbmu.edu.cn; Xue Zhang, McKusick-Zhang Center for Genetic Medicine, Peking Union Medical College, Beijing 100005, China. Tel.: +18245196868; E-mail: xuezhang@hrbmu.edu.cn
†These authors contributed equally to this work.

Abstract

Since the first report of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in December 2019, the COVID-19 pandemic has spread rapidly worldwide. Due to the limited virus strains, few key mutations that would be very important with the evolutionary trends of virus genome were observed in early studies. Here, we downloaded 1809 sequence data of SARS-CoV-2 strains from GISAID before April 2020 to identify mutations and functional alterations caused by these mutations. Totally, we identified 1017 nonsynonymous and 512 synonymous mutations with alignment to reference genome NC_045512, none of which were observed in the receptor-binding domain (RBD) of the spike protein. On average, each of the strains could have about 1.75 new mutations each month. The current mutations may have few impacts on antibodies. Although it shows the purifying selection in whole-genome, ORF3a, ORF8 and ORF10 were under positive selection. Only 36 mutations occurred in 1% and more virus strains were further analyzed to reveal linkage disequilibrium (LD) variants and dominant mutations. As a result, we observed five dominant mutations involving three nonsynonymous mutations C28144T, C14408T and A23403G and two synonymous mutations T8782C, and C3037T. These five mutations occurred in almost all strains in April 2020. Besides, we also observed two potential dominant nonsynonymous mutations C1059T and G25563T, which occurred in most of the strains in April 2020. Further functional analysis shows that these mutations decreased protein stability largely, which could lead to a significant reduction of virus virulence. In addition, the A23403G mutation increases the spike-ACE2 interaction and finally leads to the enhancement of its infectivity. All of these proved that the evolution of SARS-CoV-2 is toward the enhancement of infectivity and reduction of virulence.

Key words: SARS-CoV-2; dominant mutation; virus virulence; interaction; evolutionary trend

Liang Cheng is a professor at the College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China. His research interests include genetics, disease system biology and metagenome.
Xudong Han is a PhD candidate at the College of Bioinformatics Science and Technology, Harbin Medical University. His research interests include bioinformatics and transcriptomics analysis.
Zijun Zhu is a master candidate at the College of Bioinformatics Science and Technology, Harbin Medical University. His research interests include bioinformatics and epigenetics.
Changlu Qi is a master candidate at the College of Bioinformatics Science and Technology, Harbin Medical University. His research interests include bioinformatics and metagenomics.
Ping Wang is a master candidate at the College of Bioinformatics Science and Technology, Harbin Medical University. His research interests include bioinformatics and deep learning.
Xue Zhang, professor of human genetics. NHC and CAMS Key Laboratory of Molecular Probe and Targeted Theranostics, Harbin Medical University. McKusick-Zhang Center for Genetic Medicine, Peking Union Medical College.
Submitted: 28 October 2020; **Received (in revised form):** 4 January 2021

Introduction

In December 2019, the respiratory disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first reported in Wuhan, China [1, 2]. Since then, it has rapidly spread across the world, leading to an unprecedented global public health emergency. As of 19 August 2020, SARS-CoV-2 has infected over 20 million individuals, and caused over 700 thousand individuals death worldwide. Like the other two coronaviridae family known to infect humans, Middle East respiratory syndrome coronavirus (MERS-CoV) and severe acute respiratory syndrome (SARS), SARS-CoV-2 is also associated with high case fatality rates (CFR) [3, 4].

According to the reference genome of SARS-CoV-2 (NC_045512), the virus genome contains 29 903 nucleotides and consists of 12 major open-reading frames (ORFs) involving ORF1a, ORF1b, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, N and ORF10. Analysis of the nucleotide and protein sequence of these ORFs can help to expose derivation and high CFR of SARS-CoV-2 [4–6]. In January 2020, Zhou *et al*. identified that SARS-CoV-2 share 79.6% sequence identity to SARS-CoV and 96% sequence identical to a bat coronavirus RaTG13 at the whole-genome level, suggesting that the virus is probable bat origin [5]. Furthermore, Zhou *et al*. found that SARS-CoV-2 and SARS-CoV share 94.4% identical at CoV species classification domains in ORF1ab, which shows that the two viruses belong to the same species [5]. In May 2020, Ayal *et al*. conducted an in-depth molecular analysis of 3001 coronavirus genomes to differentiating high CFR strains including SARS-CoV-2, SARS and MERS-CoV from low CFR strains [4]. And they identified 11 regions of nucleotide alignments in four ORFs ORF1ab, S, M and E for predicting high CFR of coronaviruses, of which GAAL insertion in the spike protein of coronavirus strains appears to be associated with high CFR [4].

Since the first report of SARS-CoV-2 strain in December 2019, the virus evolves constantly through mutation in genome, which were identified in recent researches [7, 8]. In March 2020, Tang *et al*. analyzed 103 SARS-CoV-2 genomes and identified two complete linkage SNPs T8782C and C28144T [7]. This indicates that the virus was evolved into L and S types. In the same time, Peter *et al*. analyzed 160 complete SARS-CoV-2 genomes sequenced in 28 February 2020 and before [8]. They divided SARS-CoV-2 into three types according to the three central variants. Type B is derived from type A with T8782C and C28144T, and type C is derived from type B with one nonsynonymous mutation G26144T. In total, both of these two researches support that T8782C and C28144T play important roles in the evolution of SARS-CoV-2.

Though current discoveries about high CFR associated GAAL insertion in SARS-CoV-2 and evolution associated SNPs T8782C and C28144T, researchers did not investigate the functional alterations caused by these mutations, which could explain high CFR and reflect the evolutionary trends of SARS-CoV-2. Since early researches are limited by the small number of SARS-CoV-2 strains, more mutations need to be investigated further with the increase of SARS-CoV-2 strains. Herein, we analyzed SNPs and functional alterations caused by mutations in 1809 sequences of SARS-CoV-2 strains. It provides new insights into understanding evolutionary trends of SARS-CoV-2.

Materials and methods

Here the sequence data of SARS-CoV-2 strains were downloaded from GISAID (https://www.gisaid.org/). As shown in Table 1, it contains 648, 703 and 458 sequence data isolated from America, Europe and Asia, respectively. All of these viral genomes were aligned to the reference genome of SARS-CoV-2 (NC_045512) using MAFFT [9]. Figure 1 shows the flow chart of our work.

Table 1. The distribution of SARS-CoV-2 strains

District	January	February	March	April
America	10	73	290	328
Europe	12	47	433	158
Asia	144	149	153	12

Totally, 1809 SARS-CoV-2 strains were downloaded from GISAID.

Analysis of mutations in 1809 SARS-CoV-2 strains

We analyzed all the SNPs in 1809 SARS-CoV-2 strains to evaluate the tendency of mutations, and the significance of synonymous and nonsynonymous mutation rates in each ORF. Here the significance of mutations in ORF was evaluated using fisher's exact test based on 2×2 tables. For example, the following table was used to evaluate the significance of the number of mutations in each ORF: the number of mutations in ORF, the number of nucleobase in ORF and the number of mutations in CDS, the number of nucleobase in CDS. To calculate synonymous and nonsynonymous mutation rates, we calculated the number of synonymous and nonsynonymous nucleotide substitutions based on a classical method [10].

Identification of dominant mutations and analysis of their potential functions

Mutations with high frequency were analyzed to detect dominant mutations. First, Haploview was used to detect the patterns of linkage disequilibrium (LD) between SNPs [11]. Then, the SARS-CoV-2 genomes were divided into four groups by month to detect the change of the mutation frequency. We further evaluated functional alteration of genes caused by those key mutations through bioinformatics tools ProtScale [12], I-Mutant [13] and PPA-Pred [14]. ProtScale and PPA-Pred are used for evaluating the hydrophobicity and binding affinity of protein [12, 14]. I-Mutant is used for prediction of protein stability [13].

Investigation of associations between the SARS-CoV-2 strains

To analysis of associations between the 1809 SARS-CoV-2 strains, we performed hierarchical clustering using R package factoextra (https://cran.r-project.org/web/packages/factoextra/index.html). Then, we constructed a maximum likelihood phylogenetic tree for SARS-CoV-2 strains and a bat coronavirus (BatCov RaTG13), which is a probable origin of SARS-CoV-2 based on sequence similarity [5, 7]. Here jModelTest (version 2.1.10) [15] and PhyML (version 3.1) [16] was used to complete the construction of maximum likelihood phylogenetic tree.

Results

Mutations in 1809 SARS-CoV-2 strains

Based on the ORF alignments of reference genome NC_045512, we identified 1529 SNPs involving 1017 nonsynonymous and 512 synonymous mutations in 1809 SARS-CoV-2 strains. None of these mutations were located in the receptor-binding domain

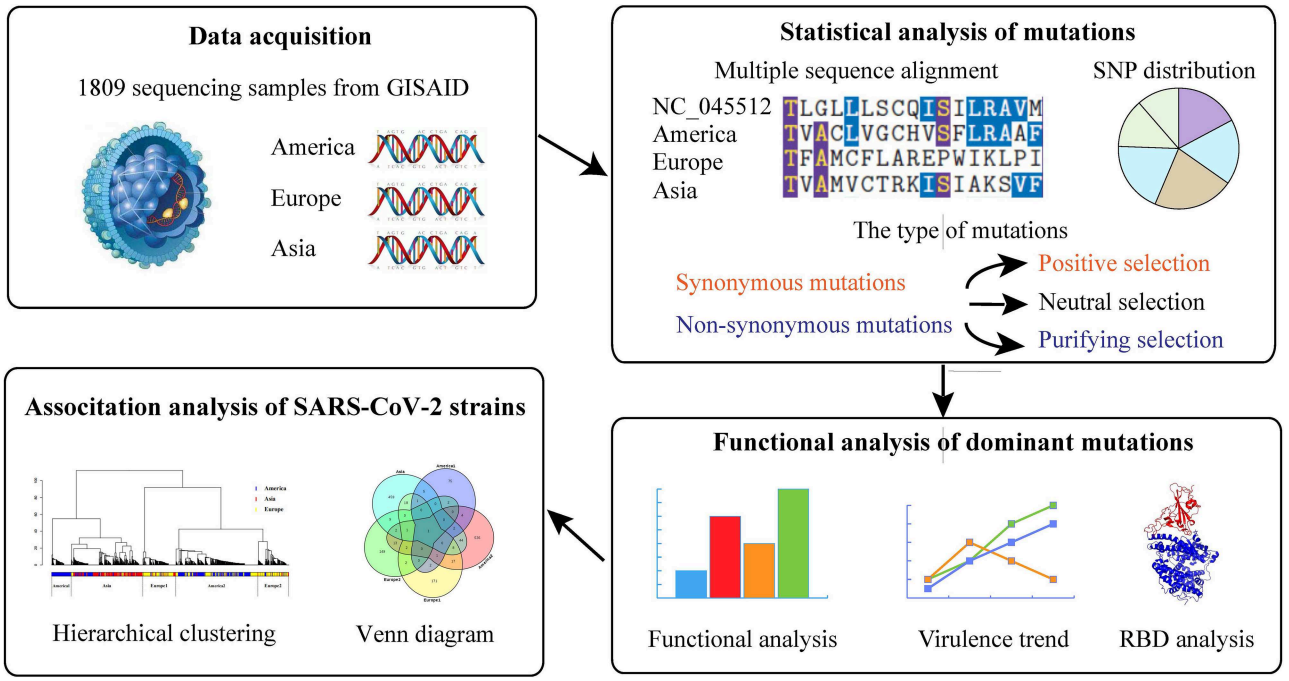

Figure 1. The workflow of our analysis on SARS-CoV-2 strains.

(RBD) of Spike protein. Although the number of nonsynonymous mutations is more than synonymous mutations, the nonsynonymous substitution rate (0.0444) is lower than synonymous substation rate (0.0796). According to the frequency of derived mutations in these virus strains in Figure 2A, the proportion of singleton nonsynonymous mutations (726/1017 = 0.7139) is higher than that of synonymous mutations (325/512 = 0.6348). For those mutations that occurred in 1% and more virus strains, the proportion of nonsynonymous mutations (21/1017 = 0.0206) is also lower than that of synonymous mutation (15/512 = 0.0293). All of these provide the evidence of purifying selection. Whereas, more nonsynonymous substitution rate (11/1017 = 0.0108) than synonymous substitution rate (3/512 = 0.0059) derived over 100 virus strains. These mean the derived nonsynonymous mutations are expected to spread more widely.

Figure 2B shows the average of accumulative mutations grows correspondingly with the time. In general, the number of nonsynonymous mutation is more than the number of synonymous mutations in each month. Although it has only 166 virus strains (Table 1) in January, it has the largest number of average mutation (2.48). Each of virus strains in the fourth month contains 6.99 mutations. It indicates that SARS-CoV-2 has about 1.75 new mutations each month on average. We further calculated the average of accumulative mutations in different locations by month. Figure 2C–E shows the number of mutations in America, Asia and Europe, respectively. Asian synonymous mutations in February and European nonsynonymous mutations in April decrease a little. Overall, mutation rate is almost same in America, Asia and Europe.

Distribution of mutations in each of ORFs

In order to determine whether the distribution of these mutations has a tendency in the ORFs, the fisher's exact test is used to evaluate the significance of the number of individual mutation sites in each of the ORFs (section 'Materials and methods'). As shown in Figure 3A, the number of individual mutation sites has a significant tendency in ORF1b, ORF3a and N (P value < 0.01). We then calculated the ratio of mutation sites based on the number of individual mutation sites and the sequence length of each ORF in Figure 3D. It shows that the ratio of mutation sites in ORF1b is smaller than that in other ORFs and the ratio of mutation sites in ORF3a and N is larger than that in other ORFs. All of these indicate that ORF1b is a conserved region and ORF3a and N is the divergent region.

We then investigated the diversity of synonymous and nonsynonymous mutations in these ORFs. Figure 3B shows that the number of synonymous mutations are evenly distributed in different ORFs. This means the mutations in synonymous sites are random in ORFs, which may be because synonymous sites are affected by small pressure of natural selection. In comparison with the synonymous site, nonsynonymous sites are under the greater pressure of natural selection, thus the distribution of their mutations should be different for each of ORFs. In fact, the number of nonsynonymous mutation sites has a significant tendency in ORF1b, ORF3a, N and ORF8 (P value < 0.01) according to Figure 3C, which is almost consistent with Figure 3A. And Figure 3F shows that the ratio of nonsynonymous mutation sites in ORF1b is smaller than that in other ORFs and the ratio of nonsynonymous mutation sites in ORF3a, ORF8 and N is larger than that in other ORFs.

In order to determine the tendency of natural selection, we calculated the nonsynonymous substation rate and synonymous substation rate in each of these ORFs in Figure 3E and F. Result shows that ORF3a, ORF8 and ORF10 were under positive selection and other ORFs were under purifying selection.

Dominant mutations derived in SARS-CoV-2 strains

Thirty-six mutations occurred in 1% and more virus strains were analyzed to reveal LD variants and dominant mutations. As shown in Figure 4A, r2 and LOD values for each pair-wise

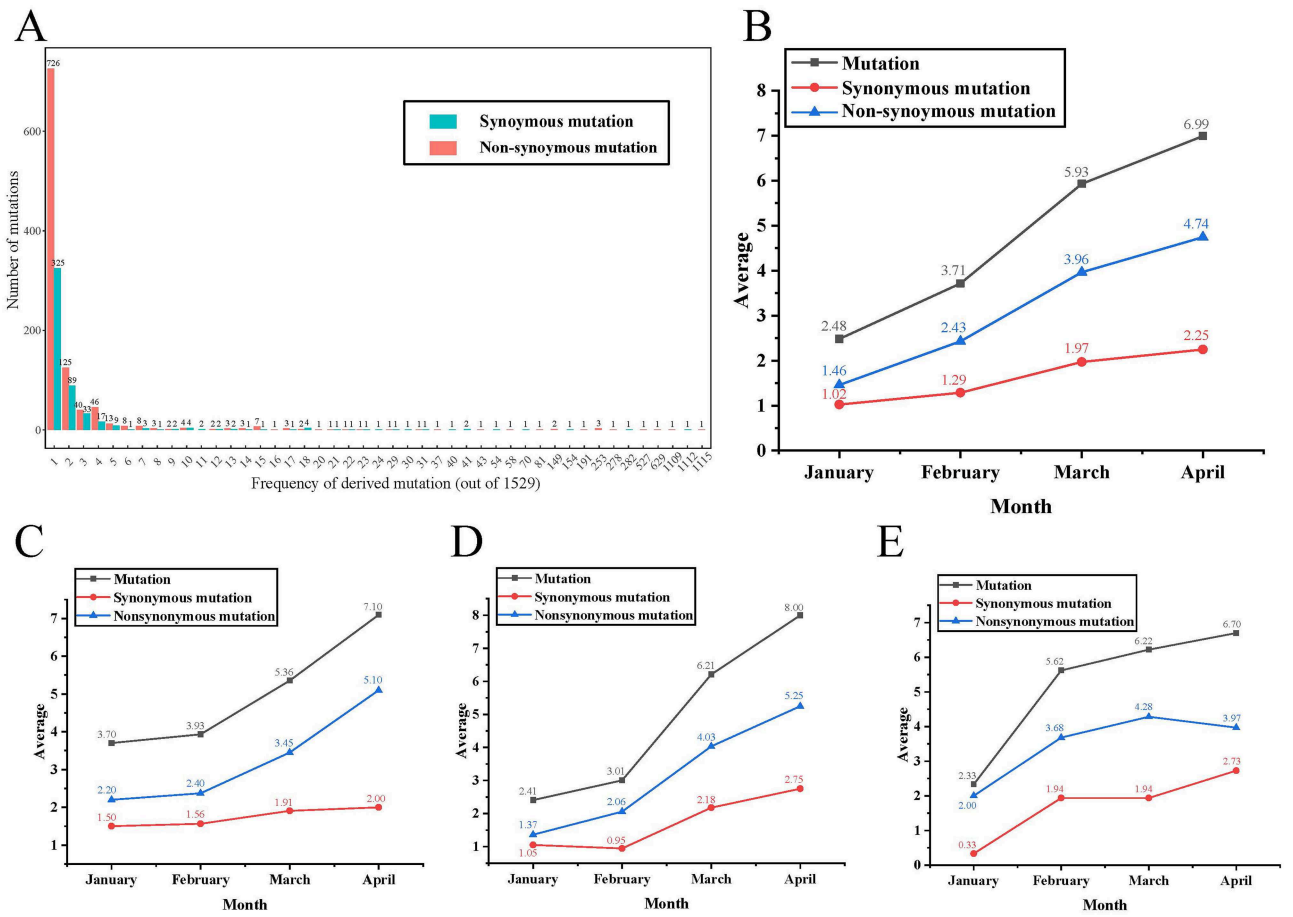

Figure 2. The distribution of mutations in 1809 SARS-CoV-2 strains. (A) The number of derived strains with individual mutations in 1809 SARS-CoV-2 virus. (B) The average of accumulative mutations by month. (C) The average of accumulative mutations in America by month. (D) The average of accumulative mutations in Asia by month. (E) The average of accumulative mutations in Europe by month.

Table 2. The significant LD variants

Location 1	Location 2	LOD	r^2
379	2244	129.1	1
17 747	17 858	447.04	1
28 881	28 882	635.9	1
28 881	28 883	635.9	1
28 882	28 883	635.9	1
3037	23 403	1029.39	0.993
3037	14 408	1019.04	0.988
14 408	23 403	1005.18	0.981
8782	28 144	645.08	0.975
1397	28 688	128.82	0.967
17 747	18 060	427.89	0.965
17 858	18 060	427.89	0.965

variants were calculated. As shown in Figure 4A, 12 pairs with significant LD were detected ($r^2 > 0.95$, LOD > 100). In the previous study, linkage in locations 8782 and 28 144 was also identified by Forster *et al.* [8] and Tang *et al.* [7] for dividing subtype of SARS-CoV-2 strains. All of the other LD variants (Table 2) are reported for the first time. For example, locations 3037, 23 403 and 14 408 have a very high LOD value (>1000).

The ratio of these 36 mutations occurs in each month were further analyzed to detect advantage mutations (section 'Materials and methods'), which was shown in Figure 4B. As a result, five locations 8782 (orf1a, synonymous), 28 144 (orf8, nonsynonymous), 3037 (orf1a, synonymous), 14 408 (orf1b, nonsynonymous) and 23 403 (s, nonsynonymous) were identified as dominant mutations in Table 3. In January, about 38% strains had nucleotides of C and T at LD locations 8782 and 28 144, respectively. Then the number of strains with C (8782) and T (28144) reduces gradually with the growth of the month. In April, almost all the strains had T (8782) and C (28 144). In January, almost all of the strains had nucleotides of C, C and A at LD locations 3037, 14 408 and 23 403, respectively. Whereas in April, most of the strains (93%) had T (3037), T (14408) and G (23403). The details of these five advantage mutations were described in Table 3. In total, we detected five dominant mutations T8782C, C28144T, C3037T, C14408T and A23403G, the origin nucleotides of which were almost substituted by the mutation in the latest virus strains. Besides, there are two important nonsynonymous mutations C1059T (ORF1a), G25563T (ORF3a). Although no strains in January and few strains in February occurred these two mutations, about 50% strains in April have the mutations. Thus, C1059T and G25563T could be potential dominant mutations. To explore the trend of mutations in different locations, we calculated the ratio of these 36 mutations for virus strains in America, Asia and Europe. Figure 4C–E shows that the virus strains in different locations have the same dominant mutations.

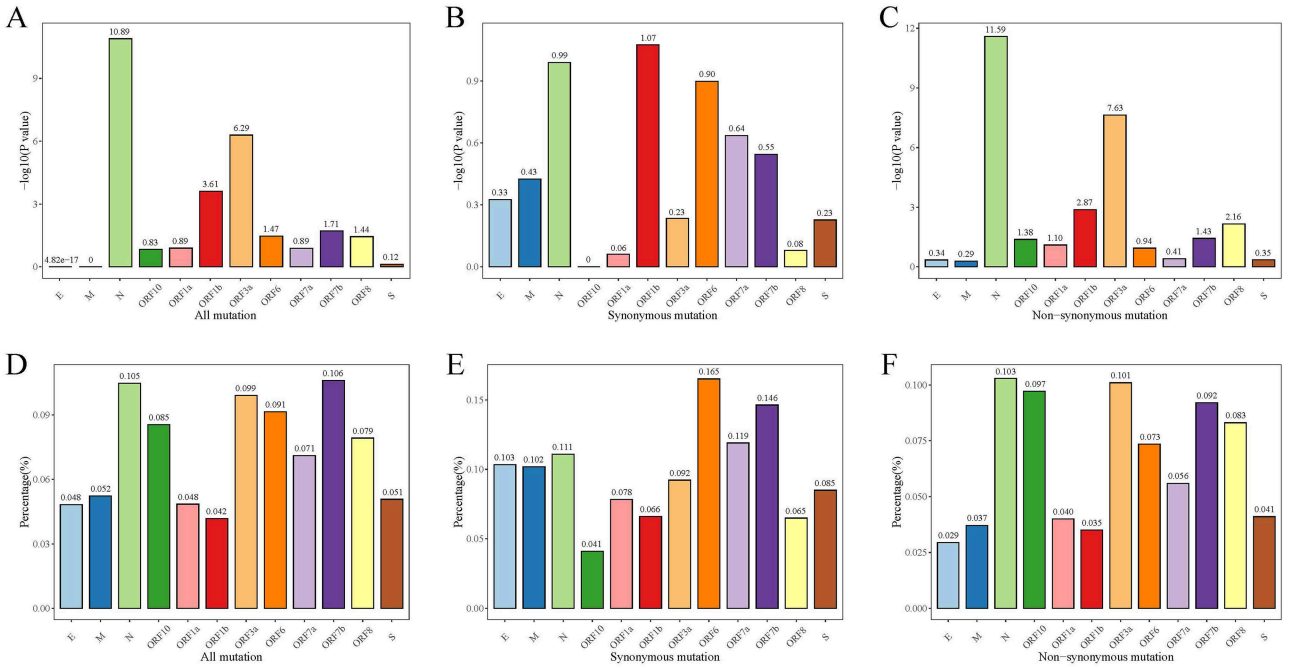

Figure 3. The distribution of mutations in each of ORFs. (A) Significant score of the number of mutation locations in each of ORFs. (B) Significant score of the number of synonymous mutation locations in each of ORFs. (C) Significant score of the number of nonsynonymous mutation locations in each of ORFs. (D) Mutation rate in each of ORFs. (E) Synonymous substitution rate in each of ORFs. (F) Nonsynonymous substitution rate in each of ORFs.

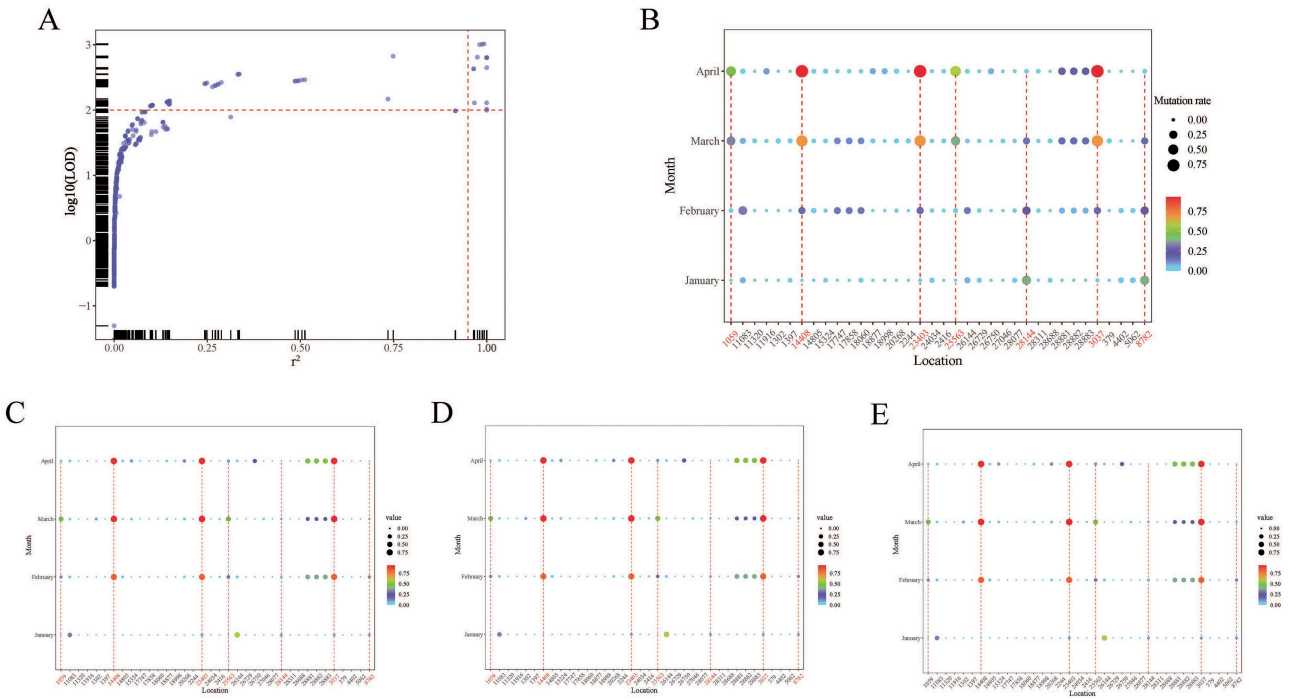

Figure 4. Linkage and tendency of 36 mutations occurred in 1% and more virus strains. (A) Scatter diagram of linkage disequilibrium between 36 SNPs. Horizontal axis and vertical axis represent r2 and LOD of pair-wise SNPs, respectively. (B) The ratio of 36 mutations occurs by month. (C) The ratio of 36 mutations in America occurs by month. (D) The ratio of 36 mutations in Asia occurs by month. (E) The ratio of 36 mutations in Europe occurs by month.

Functional analysis of dominant mutations revealed trends of evolution

Three dominant nonsynonymous mutations C28144T, C14408T and A23403G and two potential dominant mutations C1059T and G25563T were evaluated by bioinformatics tools for investigating the functional alterations caused by these mutations. We predicted the potential changes of protein stability due to the nonsynonymous mutations using I-Mutant [13], which is a widely used online tool based on support vector machine. I-Mutant directly estimates the relative stability changes upon

Table 3. Substitution rate of dominant mutations in each month

Location	RaTG13 sequence	Reference sequence	Mutation sequence	ORF	Mutation type	Substitution rate in January	Substitution rate in February	Substitution rate in March	Substitution rate in April
3037	T	C	T	ORF1a	Synonymy	0.006	0.16	0.69	0.93
8782	T	C	T	ORF1a	Synonymy	0.38	0.26	0.16	0.02
14 408	C	C	T	ORF1b	Nonsynonymy	0	0.16	0.69	0.93
23 403	A	A	G	S	Nonsynonymy	0.006	0.16	0.69	0.93
28 144	C	T	C	ORF8	Nonsynonymy	0.38	0.26	0.15	0.02
1059	C	C	T	ORF1a	Nonsynonymy	0	0.022	0.33	0.46
25 563	A	G	T	ORF3a	Nonsynonymy	0	0.03	0.38	0.58

Table 4. Prediction results of A23403G binding affinity using PPA-Pred

Nucleotide	ΔG (kcal/mol)	Kd (M)
A	−14.36	2.96e−11
G	−14.37	2.90e−11

ΔG is dissociation free energy and K_d is dissociation constant.

Table 5. The distribution of SARS-CoV-2 strains on different clusters

District	January	February	March	April
America1	0	37	106	6
America2	0	8	331	288
Asia	165	188	163	30
Europe1	1	15	138	138
Europe2	0	21	138	94

protein mutation through ΔΔG values [13]. Here we got ΔΔG values with −0.67, −0.83, −0.93, −0.9 for C1059T, C14408T, A23403G and G25563T, respectively. The very low ΔΔG values (<−0.5) show these mutations decreased protein stability largely, which could lead to the significant reduction of virus virulence [17]. By comparison, C28144T could reduce virus virulence a little, since the ΔΔG values for the mutation is near zero.

Since spike-ACE2 interaction can affect virus infectivity [18, 19], we analyzed the alteration of the interaction caused by spike-ACE2 binding affinity due to A23403G using PPA-Pred [14]. The tool can evaluate the alterations of binding affinity through two aspects: dissociation free energy ΔG and dissociation constant Kd. Both of these two aspects are inversely proportional to protein–protein binding affinity and interactions [14]. In Table 4, ΔG and Kd in SARS-CoV-2 spike is decreased by the A23403G, which means that the mutation in SARS-CoV-2 increases the spike-ACE2 interaction, and finally leads to the enhancement of its infectivity [18, 19].

Associations between SARS-CoV-2 strains among different continents

According to the continent where the patients with SARS-CoV-2 are located, SARS-CoV-2 genomes are marked as America, Asia and Europe. Because the patient's area does not represent the difference of SARS-CoV-2 strains, we encoded the virus genome to perform hierarchical clustering (section 'Materials and methods'). As shown in Figure 5A, the results of hierarchical clustering show five distinct groups. According to the continent where the main sample of each group is located, the five groups of SARS-CoV-2 strains were named America1, Asia, Europe1, America2 and Europe2 in turn. The sample sizes of these five regional groups are shown in Table 5. We then performed hierarchical clustering of virus genomes based on nonsynonymous mutation, the results of which (Figure 5B) is consistent with that base on all mutations.

As shown in Figure 5C, we identified each regional group SARS-CoV-2 CDS mutations sites and the five groups SARS-CoV-2 strains have few intersections in these mutations, which indicates that our grouping method can effectively distinguish SARS-CoV-2 strains. Further, we used the complete genomes of Bat RaTG13 from the GeneBank and five regional groups SARS-CoV-2 strains specific mutation sites that totaled 147 to construct a maximum likelihood phylogenetic tree (Figure 5D).

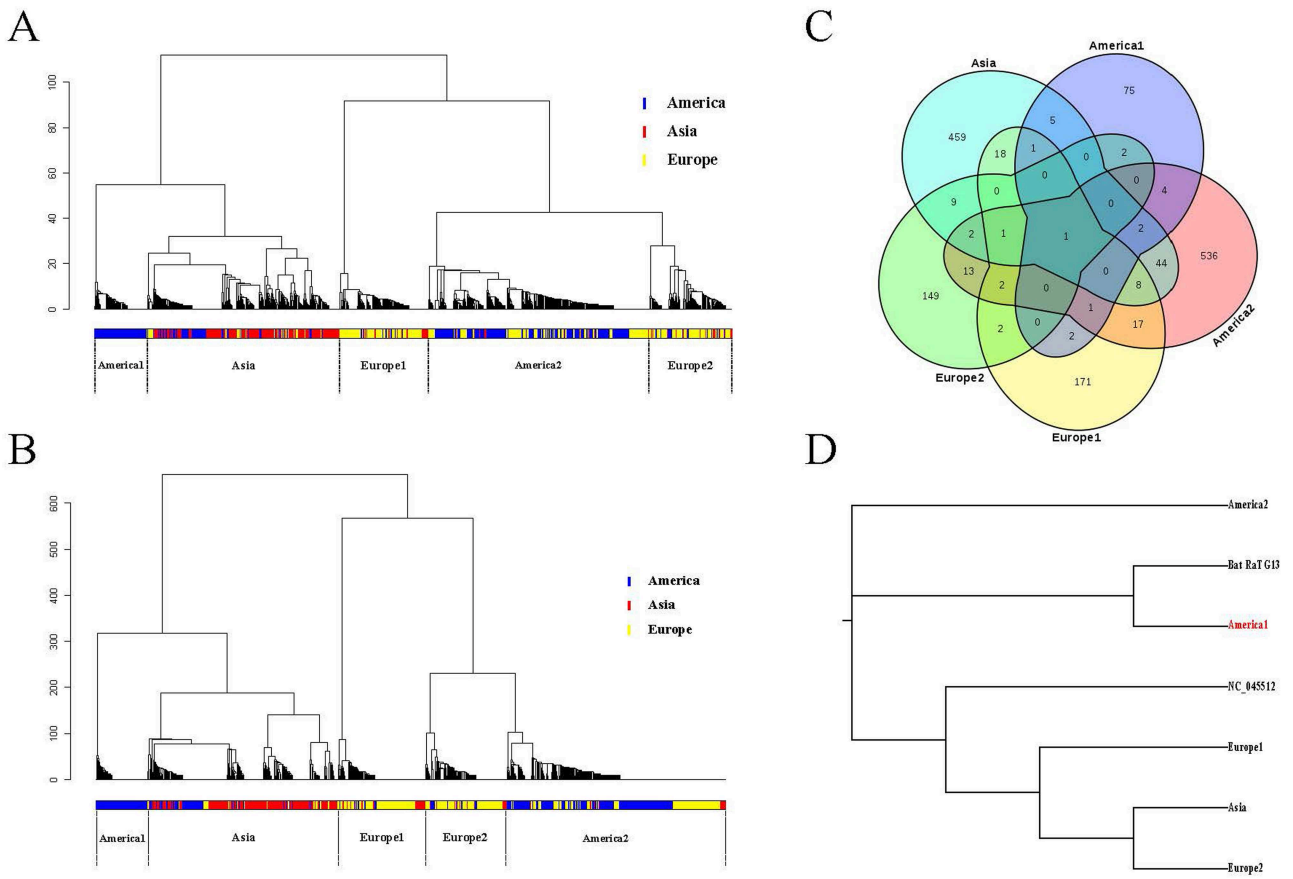

Figure 5. Associations between SARS-CoV-2 strains among different continents. (A) Hierarchical clustering of virus genomes from different continents. (B) Hierarchical clustering of virus genomes based on nonsynonymous mutation. (C) Venn diagram of mutation sites of the five regional grouping SARS-CoV-2 strains. *D. maximum* likelihood phylogenetic tree of the five regional groups SARS-CoV-2 strains specific mutations.

In the phylogenetic tree, the five regional groups SARS-CoV-2 strains are also clearly separated. Compared to Asia, America1 and America2 SARS-CoV-2 strains are closer to bat-derived coronavirus.

Availability and implementation

All the codes for conducting this study could be downloaded from the website: https://github.com/liangcheng-hrbmu/FE-SARS-CoV-2. In addition, we will download and update the latest data each month.

Discussion

Totally, there were 1529 SNPs in 1809 virus strains, none of which were located in RBD of the spike protein. In addition, each of the strains could has about 1.75 new mutations each month in average. Since RBD are targeted by many known neutralizing antibodies and the number of mutation in each strains is very few, the accumulated mutations may have few impacts on antibodies. The nonsynonymous substation rate is lower than synonymous substation rate. It provides evidence of purifying selection. The further analysis in each of ORFs shows that ORF3a, ORF8 and ORF10 were under positive selection, ORF1b is a conserved region and ORF3a, and N is the divergent region.

Like ORF1b, even if not significant, mutations in ORF1a also tend to be less. ORF1a and ORF1b can encode two nonstructural proteins of the SARS-CoV-2. The two nonstructural proteins are essential for the basic function (like viral replication, viral assembly) of the SARS-CoV-2 [20]. The stability of ORF1a and ORF1b ensures the basic needs of SARS-CoV-2 survival. The gene region encoding the N protein has the highest tendency to mutation. The 168-208 amino acid region of N protein can directly bind to M protein through ionic interaction [21]. The M protein plays an important role in the assembly, germination and release of the SARS-CoV-2 and O-Glycosylation of M protein is related to the interaction between coronavirus and host [22, 23]. ORF3a, ORF8 and ORF10 are all accessory proteins of the SARS-CoV-2. Some accessory proteins can regulate interferon signaling pathways and the production of proinflammatory cytokines, which makes it play an important role in the host response to coronavirus infection and thereby [24, 25]. A significant body of evidence has found SARS-CoV-2 ORF3a could coordinate attack the heme on the 1-beta chain of hemoglobin and could efficiently induce apoptosis in cells [26, 27]. SARS-CoV-2 ORF8 stands out by structural plasticity and high diversity and its gene transcripts are expressed in higher amounts [28, 29]. Furthermore, SARS-CoV-2 ORF8 protein may inhibit the type I interferon signaling pathway, an important role of antiviral infection [30]. Although the function of ORF10 remains to be elucidated, we infer that ORF10 with positive selection may have an important role in SARS-CoV-2 infection and spread.

We further identified dominant mutations in 1809 virus strains, and analyzed the functional alterations caused by these dominant mutations. Totally, we identified five dominant mutations T8782C, C28144T, C3037T, C14408T and A23403G

and two potential dominant mutations C1059T and G25563T. There T8782C, and C28144T were also identified by Peter *et al.* for distinguishing the subtype of SARS-CoV-2 [8]. Viruses with 3037T-14408T-23403G have a fitness gain, which was reported by Yang *et al.* in their latest discovery [31]. A23403G were deemed as important mutations in spike protein. And this mutation has become the most prevalent form in the global pandemic [32]. Mutations C1059T and G25563T are first highlighted here. We analyzed the alteration of protein stability due to the dominant mutations and using I-Mutant and the alteration of spike-ACE2 binding affinity due to A23403G using PPA-Pred. Results show that mutations decreased protein stability largely, which could lead to a significant reduction of virus virulence. The A23403G mutation increases the spike-ACE2 interaction, and finally leads to the enhancement of its infectivity [18, 19]. This were further validated recently in clinical trials by Plante *et al.* [33]. All of these proved that the evolution of SARS-CoV-2 is toward enhancement of infectivity and reduction of virulence as other viruses [34, 35].

Up to now, seven types of coronavirus have been known to infect humans, which includes low CFR and high CFR named SARS-CoV-2, SARS and Middle East respiratory syndrome (MERS). In previous studies, Bethany *et al.* has highlighted an important insert from location 32 029 to 32 040 that encodes GAAL of spike protein [4]. Whereas, the significance of these positions was not pointed out. Here we analyzed changes in binding affinity due to GAAL insertion in SARS-CoV-2 reference genome NC_045512 using PPA-Pred [14]. Dissociation free energy ΔG and dissociation constant Kd were discussed since both of them are inversely proportional to protein–protein binding affinity [14]. Due to the GAAL insertion, ΔG is decreased from -19.19 to -20.08, Kd is decreased from 8.40e-15 to 1.89e-15. It means that the insertion increases the spike-ACE2 binding affinity, and finally leads to the enhancement of its infectivity and virulence [18, 19].

Key Points

- Sequence analysis of 1809 SARS-CoV-2 strains.
- Identification of positive selection in ORF3a, ORF8 and ORF10.
- Identification of five dominant mutations and two potential dominant mutations.
- Discovery of significant reduction of virulence and enhancement of infectivity on current mutations in SARS-CoV-2.
- Association analysis of SARS-CoV-2 strains in different continents.

Authors' contributions

LC and XZ conceived and designed the experiments. XH and ZZ analyzed data. CQ and PW wrote and revised this manuscript. All authors read and approved the final manuscript.

Acknowledgments

We thank researchers for sharing their sequencing data from GISAID (https://www.gisaid.org/), and thank GISAID for providing instructions and advice on the database. We also thank researchers and NCBI for providing sequencing data and instructions for SARS-CoV-2 reference sequence (https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2). This work was supported by the Tou-Yan Innovation Team Program of the Heilongjiang Province (2019-15), and National Natural Science Foundation of China [61871160].

Funding

Tou-Yan Innovation Team Program of the Heilongjiang Province (2019–15), National Natural Science Foundation of China [61871160], Heilongjiang Province Postdoctoral Fund [LBH-TZ20], and Young Innovative Talents in Colleges and Universities of Heilongjiang Province (2018–69).

References

1. Lu H, Stratton CW, Tang YW. Outbreak of pneumonia of unknown etiology in Wuhan, China: the mystery and the miracle. *J Med Virol* 2020;**92**:401–2.
2. Jiang S, Du L, Shi Z. An emerging coronavirus causing pneumonia outbreak in Wuhan, China: calling for developing therapeutic and prophylactic strategies. *Emerg Microbes Infect* 2020;**9**:275–7.
3. Su S, Wong G, Shi W, *et al.* Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol* 2016;**24**:490–502.
4. Gussow AB, Auslander N, Faure G, *et al.* Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proc Natl Acad Sci USA* 2020;**117**:15193–9.
5. Zhou P, Yang XL, Wang XG, *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020;**579**:270–3.
6. Wong MC, Javornik Cregeen SJ, Ajami NJ, *et al.* Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv* 2020. doi: https://doi.org/10.1101/2020.02.07.939207.
7. Tang X, Wu C, Li X, *et al.* On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev* 2020;**7**:1012–23.
8. Forster P, Forster L, Renfrew C, *et al.* Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci USA* 2020;**117**:9241–3.
9. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;**30**:772–80.
10. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 1986;**3**:418–26.
11. Barrett JC, Fry B, Maller J, *et al.* Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005;**21**:263–5.
12. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;**157**:105–32.
13. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res* 2005;**33**:W306–10.
14. Yugandhar K, Michael Gromiha M. Protein–protein binding affinity prediction from amino acid sequence. *Bioinformatics* 2014;**30**:3583–9.
15. Darriba D, Taboada GL, Doallo R, *et al.* jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 2012;**9**:772.
16. Guindon S, Dufayard J, Lefort V, *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 2010;**59**:307–21.
17. Xu K, Klenk C, Liu B, *et al.* Modification of nonstructural protein 1 of influenza A virus by SUMO1. *J Virol* 2011;**85**:1086–98.

18. Brielle ES, Schneidman-Duhovny D, Linial M. The SARS-CoV-2 exerts a distinctive strategy for interacting with the ACE2 human receptor. *Viruses* 2020;**12**:497.
19. Lan J, Ge J, Yu J, *et al*. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* 2020;**581**:215–20.
20. Wu F, Zhao S, Yu B, *et al*. A new coronavirus associated with human respiratory disease in China. *Nature* 2020;**579**:265–9.
21. He R, Leeson A, Ballantine M, *et al*. Characterization of protein-protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus. *Virus Res* 2004;**105**:121–5.
22. de Haan CAM, Vennema H, Rottier PJM. Assembly of the coronavirus envelope: homotypic interactions between the M proteins. *J Virol* 2000;**74**:4967–78.
23. de Haan CAM, de Wit M, Kuo L, *et al*. O-glycosylation of the mouse hepatitis coronavirus membrane protein. *Virus Res* 2002;**82**:77–81.
24. Liu DX, Fung TS, Chong KK, *et al*. Accessory proteins of SARS-CoV and other coronaviruses. *Antiviral Res* 2014;**109**:97–109.
25. Narayanan K, Huang C, Makino S. SARS coronavirus accessory proteins. *Virus Res* 2008;**133**:113–21.
26. Liu W, Li H. COVID-19: Attacks the 1-Beta Chain of Hemoglobin and Captures the Porphyrin to Inhibit Human Heme Metabolism. *ChemRxiv* 2020:v7. Preimpresión. doi: https://doi.org/10.26434/chemrxiv11938173.
27. Ren Y, Shu T, Wu D, *et al*. The ORF3a protein of SARS-CoV-2 induces apoptosis in cells. *Cell Mol Immunol* 2020;**17**:881–3.
28. Nyayanit DA, Sarkale P, Baradkar S, *et al*. Transcriptome & viral growth analysis of SARS-CoV-2-infected Vero CCL-81 cells. *Indian J Med Res* 2020;**152**:70–6.
29. Pereira F. Evolutionary dynamics of the SARS-CoV-2 ORF8 accessory gene. *Infect Genet Evol* 2020;**85**:104525.
30. Li JY, Liao CH, Wang Q, *et al*. The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway. *Virus Res* 2020;**286**:198074.
31. Yang HC, Chen CH, Wang JH, *et al*. Analysis of genomic distributions of SARS-CoV-2 reveals a dominant strain type with strong allelic associations. *Proc Natl Acad Sci USA* 2020;**117**:30679–86.
32. Korber B, Fischer WM, Gnanakaran S, *et al*. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 2020;**182**:812–27 e819.
33. Plante JA, Liu Y, Liu J, *et al*. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* 2020. doi: 10.1038/s41586-020-2895-3.
34. Liu Y, Liu J, Du S, *et al*. Evolutionary enhancement of Zika virus infectivity in Aedes aegypti mosquitoes. *Nature* 2017;**545**:482–6.
35. Arien KK, Vanham G, Arts EJ. Is HIV-1 evolving to a less virulent form in humans? *Nat Rev Microbiol* 2007;**5**:141–51.