Received: 5 July 2021 | Accepted: 13 July 2021

DOI: 10.1002/jmv.27225

RESEARCH ARTICLE

JOURNAL OF MEDICAL VIROLOGY WILEY

Evolution trace of SARS-CoV-2 from January 19 to March 12, 2020, in the United States

Ziying Lin[1] | Hua Qing[1] | Rui Li[1] | Lei Zheng[2] | Huipeng Yao[1]

[1]College of Life Science, Sichuan Agriculture University, Ya'an, China

[2]Sichuan University, Chengdu, China

Correspondence
Huipeng Yao, College of Life Science, Sichuan Agriculture University, Ya'an, 625014
Sichuan, China.
Email: yaohuipeng0921@163.com

Abstract

As a kind of human betacoronavirus, SARS-CoV-2 has endangered globally public health. As of January 2021, the virus had resulted in 2,209,195 deaths. By studying the evolution trend and characteristics of 265 SARS-CoV-2 strains in the United States from January to March, it is found that the strains can be divided into six clades, USA clade-1, USA clade-2, USA clade-3, USA clade-4, USA clade-5, and USA clade-6, in which US clade-1 may be the most ancestral clade, USA clade-2 is an interim clade of USA clade-1 and USA clade-3, the other three clades have similar codon usage pattern, while USA clade-6 is the newest and most adaptable clade. Mismatch analysis and protein alignment showed that the evolution of the clades arises from some special mutations in viral proteins, which may help the strain to invade, replicate, transcribe and so on. Compared with previous research and classifications, we suggest that clade O in GISAID should not be an independent clade and Wuhan-Hu-1 (EPI_ISL_402125) should not be an ancestral reference sequence. Our study decoded the evolutionary dynamic of SARS-CoV-2 in the early stage from the United States, which give some clues to infer the current evolution trend of SARS-CoV-2 and study the function of viral mutational protein.

KEYWORDS
clade, coronavirus, evolution, SARS-CoV-2

1 | INTRODUCTION

SARS-CoV-2, a member of the genus *betacoronavirus* that infects humans,[1] which can not only cause a series of mild symptom such as high inflammation and microangiopathy, but also develop into severe symptoms like extensive thrombosis and severe acute respiratory syndrome, endangering patient's life.[2] Since it first identified in December 2019 in Wuhan, China, the virus has spread globally, resulting in the currently ongoing COVID-19 pandemic. As of January 31, 2021, there have been 102,083,344 confirmed cases of COVID-19 with 2,209,195 deaths globally, and 25,676,612 confirmed cases with 433,173 deaths in the United States, reported to WHO (https://covid19.who.int/).

SARS-CoV-2 genome is an about 30 kb positive-sense single-stranded RNA and firstly published in January 2020,[3] which is made up of 14 open reading frames (ORFs), ORF1a, ORF1b, spike (S), ORF3a, envelope (E), membrane (M), ORF6, ORF7a, ORF7b, ORF8, nucleocapsid (N), ORF10, ORF9, and ORF14, encoding corresponding proteins.[4]

The 14 proteins are classified into three kinds, polyproteins, main structural proteins and accessory proteins. SARS-CoV-2 has two kinds of polyproteins, polyprotein 1a and polyprotein 1b which are encoded by two overlapping ORFs, ORF1a and ORF1b, occupying 2/3 of its genome.[5,6] Polyprotein 1a is proceed into 10 non-structural proteins (Nsps), Nsp1 to Nsp10 by viral protease Nsp5 and the papain-like protease domain from Nsp3.[7] Nsp1 can interfere with the host innate immune response, by lowing the expression of some immune protein factors after binding to the 40S ribosomal subunit and destroying its messenger RNA (mRNA).[8] Nsp3 hydrolyze the sequence LXGG↓X, resulting in producing Nsp1, Nsp2, Nsp3, from

ORF1a[9] and blocking the activation of innate immune by deubiquitinating and deISG15ylating.[10–13] Through the capacity, after binding to ADP-ribose, Nsp3 removes the chemical group from the ADP-ribosylated proteins in immune response.[14–16] In addition, Nsp3 can induce double-membrane vesicles outside viral replication-transcriptional complexes (RTC) with Nsp4 and Nsp6.[17] Nsp5, as the main protease, can hydrolyze the sequence (L/V/F)Q↓(S/A/G), catalyzing polyprotein 1ab to produce Nsp4 to Nsp10 and Nsp12 to Nsp16, 12 proteins.[18] Nsp7 and Nsp8 form a heterodimer which binds to RNA dependent RNA polymerase, stabilizing the enzyme polymerase domain, increasing the affinity of RNA binding to the enzyme significantly and enhancing the catalytic activity.[4,19,20] Nsp9 is a kind of single stranded RNA-binding protein, which may influence the virulence of the virus.[21] Nsp10 assists methyltransferase to form a methylation complex which caps the virus mRNA and enhances the enzyme activity.[22] Nsp12 is a kind of RNA dependent RNA polymerase (RdRp), which catalyze the replication and transcription of virus genome after forming RTC with other Nsps. Beside, with Nsp8, it can regulate the activity of helicase.[23] Nsp13 is a kind of helicase, participating in virus replication and preservation in the life activity of virus.[24] Nsp14 has a N-terminal exonuclease domain and a C terminal guanine-N7 methyl transferase domain, in which, the former is responsible for proofreading new error nucleotides and the latter is involved in capping the virus RNA from the degradation of the host immune.[25–27] Nsp15 is a kind of ribonucleic acid endonuclease,[28] which can prevent the host from detecting the virus double stranded RNA to escape the attack of the immune system. Nsp16 is a specific methyl transferase, which catalyzes 2'-*O*-methylation of the first nucleotide in the viral capped RNA to be protected from the degradation of host innate immune response.[29,30]

SARS-CoV-2 has spike (S), membrane (M), envelope (E), nucleocapsid (N), four kinds of Nsps. S protein helps the virus enter host cells by binding to human ACE2 receptors,[31–33] and induces inflammatory response after recognized by TLR4.[33] N protein is responsible for packaging viral RNA into helical ribonucleocapsid, forming viral nucleocapsid structure with M protein.[34–36] E may play an important role in virus maturation, transmission and reproduction.[4] Besides, the virus has several accessory proteins, such as: ORF3a, ORF6, ORF7a, ORF7b, ORF8, ORF9, and ORF10.[4] As a transmembrane protein, ORF3a is related to virion release and viral pathogenicity.[37,38] ORF7a may play a role in protein transport mediated by endoplasmic reticulum and Golgi complex.[4] The exogenous overexpression of ORF8 in cells can destroy IFN-I signal from the host.[39] ORF9b, as a part of N protein, inhibits host immune response.[40]

SARS-CoV-2 mutates rapidly, producing a large number of lineages or clades by different methods. As of February 12, 2020, Yu et al.[41] found that SARS-CoV-2 is classified into five groups including 58 haplotypes, in which, H13 and H35 were ancestral haplotypes and H1 (which from the Hua Nan market) was derived from the H3 haplotype. Two months later, Peter et al.[42] proposed an interesting viewpoint that the virus should be divided into three major variants, Types A, B, and C, in which Type B appear in East Asia while other types are mainly distributed outside of East Asia but Type A was the ancestor type. By using 64 whole genome sequences from December 30, 2019, to March 9, 2020, in Europe, North America, South America, and Southeast Asia, Maías et al.[43] suggests that SARS-CoV-2 is divided into three genetic clades cocirculating in all over the world and the most recent common ancestor may appear in around November 1, 2019. On March 31, 2020, Jennifer et al. firstly classified 10,959 viral isolates into lineages A and B, in which the genomes of lineage A is characterized by 8782T and 28144C, while lineage B by 8782C and 28144T. Each sublineages is designated or defined by an additional unique mutations and can further diversify into sublineages.[44] On May 2, Priyanka et al.[45] found that SARS-CoV-2 can be classified into two groups, G1 and G2, in which the former is restricted to moderate warm climate and the later start to spread neighboring cold climate and hot climate of the tropics. Houriiyah et al.[46] had identified 16 new lineages, including B.1.1.54, B.1.1.56, and C.1 spreading widely in South Africa after analyzing 1365 whole genomes of SARS-CoV-2 isolate between March 6 and August 26, 2020. BII/GIS from A*STAR Singapore analyzed all SARS-CoV-2 sequences as of January 22, 2021 and found that 373,805 full genomes excluding 21,358 low coverage entries can be classified into S, L, V, G, GH, GR, GV, 7 clades, in which clade S is ancestral and clade O is lacking.[47]

Changtai et al.[48] identified 13 variation sites in ORF1a, ORF1b, S, ORF3a, M, ORF8, and N regions, on February 14, 2020. Maria et al.[49] found eight new mutations at positions 1397, 2891, 14,408, 17,746, 17,857, 18,060, 23,403, and 28,881 in the viral genomic sequences worldwide from December 2019 to mid-March 2020. As of April 19, 2020, Van Dorpan Dorp et al.[50] found that many recurrent mutations occurred in S protein, Nsp6, Nsp11, and Nsp3 and nonsynonymous mutations account for nearly 80%. Takahiko et al.[51] found that C3037T synonymous sites in genome, P4715L in ORF1ab and D614G in S protein were the most frequent mutations, after aligning 10,022 SARS CoV-2 genomes between February 1, and May 1, 2020. On May 7, 2020, Yujiro et al.[52] found that P4715L of ORF1ab and D614G of S protein are linkage and lethality. In July 2020, Domenico et al.[53] found two mutations at Nsp6 position 37 and ORF 10 position 3 or 4, reducing the two proteins structure stability. In July 2020, Phan[54] found three mutations (N354D, D364Y and V367F) on the surface of S protein, which may change its conformation, resulting in changes in antigenicity of the virus. On October 2, 2020, Sarmilah et al.[55] found that for S protein, the N501Y mutation was more infectious than the D614G mutation. On December 18, 2020, Yixuan et al.[56] found that after the mutation D614G in S protein, the variant has more effective infection, replication and competitive adaptability in human primary airway epithelial cells.

In this study, we firstly try to use all ORFs combined sequence to analyze the phylogenetical relation of the strains in the United States from January 19 to March 12, 2020. Then, to test our phylogenetic tree we ran a series of follow-up analysis. We calculate each ORF nucleotide substitution rates, analyze its codon usage pattern and population expansion and align each corresponding protein sequences. Finally, we compare our phylogenetical result with others, to reveal the evolution dynamics of SARS-CoV-2 in early stage of US epidemic outbreak, assess the characteristics of its classification and understand the trend of the epidemic.

2 | MATERIALS AND METHODS

2.1 | Data acquisition

On May 12, 2020, flagged as "complete (>29,000 bp)" and "high coverage," 683 SARS-CoV-2 genomes from January 19 to March 12 in the United States were downloaded from the GISAID Initiative EpiCoV platform. Filtering any sequences with N, W, and other missing site produces a final data set of 265 genomes. In addition, two reference ancestral sequences, hCoV-19/bat/Yunnan/RaTG13/2013 and hCoV-19/Wuhan-Hu-1/2019, are also downloaded from the GISAID. All voucher information of 267 strains can be found in Table S1.

2.2 | ORFs finding and combining

With the aid of the reference sequence, hCoV-19/Wuhan-Hu-1/2019, the tool of ORFfinder in linux x64 is used to look for the meaning ORFs from each genome. Because ORF9 gene is a part of the nucleocapsid gene, and the sequence and function of ORF14 gene can not been found in NCBI and GISAID, 12 ORFs or genes from all genomes are screened. According to the order of the genes on the genome, 12 ORFs sequences were spliced into a large assembly sequence with end to end mode, producing 267 assemblies.

2.3 | Maximum likelihood (ML) tree and Bayesian phylogenetic tree

Taking the assembly sequence of hCoV-19/bat/Yunnan/RaTG13/2013 as the reference sequence, the other 266 assemblies is used to build a phylogenetic tree by the software of MEGA-X with the ML statistical method (Model, Tamura-Nei; Bootstrap, 500). Bayesian phylogenetic analysis can be used to infer the divergency time of each clade, further determining the virus evolutionary trend. The 266 assemblies are used to build a Bayesian phylogenetic tree by using BEAST 2 software with GTR model. The clock model chooses Relaxed Clock Log Normal and the chain length of MCMC is 100,000,000. The priors model use the Coalescent Bayesian Skyline, and statistical uncertainty in the data was reflected by the 95% highest posterior density values. Using the TRACER program v1.7.1, it is found that one hundred million generations were produced after a burn-in of 10 million steps, assessed by effective sample sizes (ESS) over 200. The tree was processed by TreeAnnotator v1.10.4 and viewed in program FigTree v1.4.4.

2.4 | Mismatch distribution

To characterize of the virus gene evolution, DNAsp is used to analyze mismatch distribution of each gene and assemblies of all SARS-CoV-2 strains in the United States. Mismatch distribution is a way to visually reflect the historical dynamics of the population. If the mismatch curve shows a unimodal Poisson distribution, it is generally accepted that the population size has experienced expansion or continuous growth. On the other side, if the curve coincides with the expectation curve, the population size remains stable in the past.

2.5 | Protein alignment and nucleotide substitution rates calculation

Twelve viral proteins of all strains are aligned to find out the common and different characteristics of each clade produced by the phylogenetic tree. For the same purpose, we calculate Ka (synonymous substitution rate), Ks (nonsynonymous substitution rate) and ω (the ratio of nonsynonymous substitution rate to synonymous substitution rate) of different genes in each clade of 265 virus strains in United States by MEGA-X (model, Nei-Gojobori).

2.6 | Principal component analysis (PCA)

To determine the codon usage pattern of SARS-CoV-2 in different clades, CodonW 1.4.2 was used to calculate the values of relative synonymous codon usage (RSCU) of 265 assemblies. Basing on the RSCU, by the software of SPSSv21.0, we calculate axis 1 and axis 2 infecting the codon usage pattern which are plotted by Excel 2010 according to axis 1 as horizontal axis, the second axis as vertical axis.

3 | RESULTS

3.1 | Establishment of clades and estimation of evolutionary time

ML tree and the Bayesian phylogenetic tree were used to classify 265 SARS-CoV-2 strains in the United States and estimate the evolutionary time of each clade. According to Figure 1, it is seen that the strains are divided into six main independent clades (USA clade-1 to USA clade-6), among which Wuhan-Hu-1 belongs to USA clade-3, indicating it is not appropriate to be as the reference ancestor sequence in many papers.[42,43,49,49,50,52] In the phylogenetic tree, it is found that USA clade-1 is the closest to the ancestors (RaTG13) in kinship, followed by USA clade-2, USA clade-3, USA clade-4, USA clade-5, and USA clade-6. In addition, as the largest clade, USA clade-6 includes 76 virus strains, followed by USA clade-1 (53), USA clade-3 (46), USA clade-5 (40), USA clade-4 (30), and USA clade-2 (20). It seem to indicate that compared with USA clade-2, the youngest clade (USA clade-6) and the oldest clade (USA clade-1) are more suitable to United States at the time. The virus large population expansion may occur in the initial and later stage of the its evolution and the virus mutated and evolved very rapidly.

Bayesian phylogenetic tree is showed as Figure 2, in which it is found that about 80 days ago, the common ancestor of all virus clades may begin to appear on December 24, 2019. About 11 days later, the ancestors of both USA clade-1 and USA clade-2

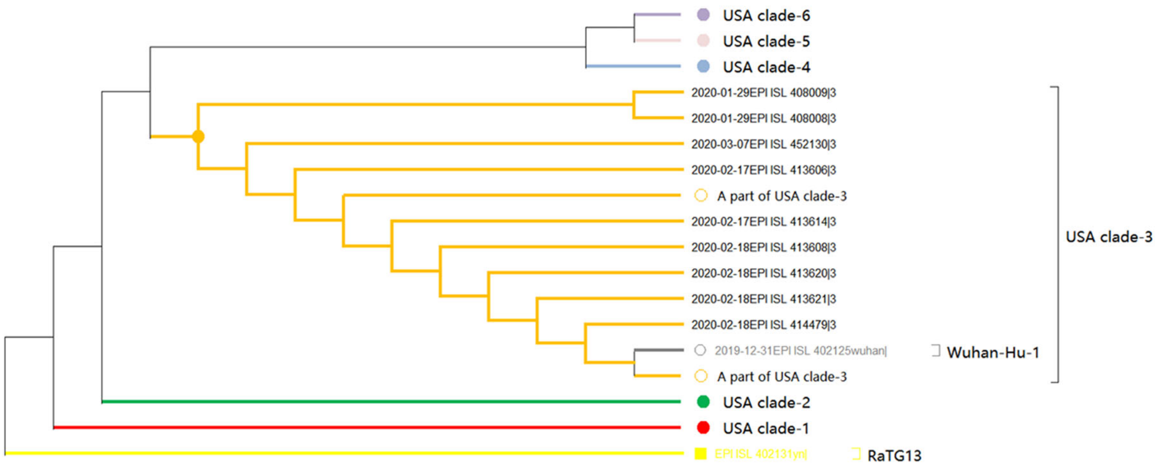
USA clade-6
USA clade-5
USA clade-4
2020-01-29EPI ISL 408009|3
2020-01-29EPI ISL 408008|3
2020-03-07EPI ISL 452130|3
2020-02-17EPI ISL 413606|3
A part of USA clade-3
2020-02-17EPI ISL 413614|3
2020-02-18EPI ISL 413608|3
2020-02-18EPI ISL 413620|3
2020-02-18EPI ISL 413621|3
2020-02-18EPI ISL 414479|3
2019-12-31EPI ISL 402125wuhan|
Wuhan-Hu-1
A part of USA clade-3
USA clade-3
USA clade-2
USA clade-1
EPI ISL 402131yn|
RaTG13

FIGURE 1 Phylogenetic trees of 265 SARS-CoV-2 strains in the United States from January 19 to March 12, 2020, which is rooted by RaTG13 and compared with Wuhan-Hu-1. The yellow colour represent ancestor RaTG13, red represent USA clade-1, green represent USA clade-2, orange represent USA clade-3, blue represent USA clade-4, pink represent USA clade-5, purple represent USA clade-6, and grey represent Wuhan-Hu-1. The details of each clade are shown in Figure S1

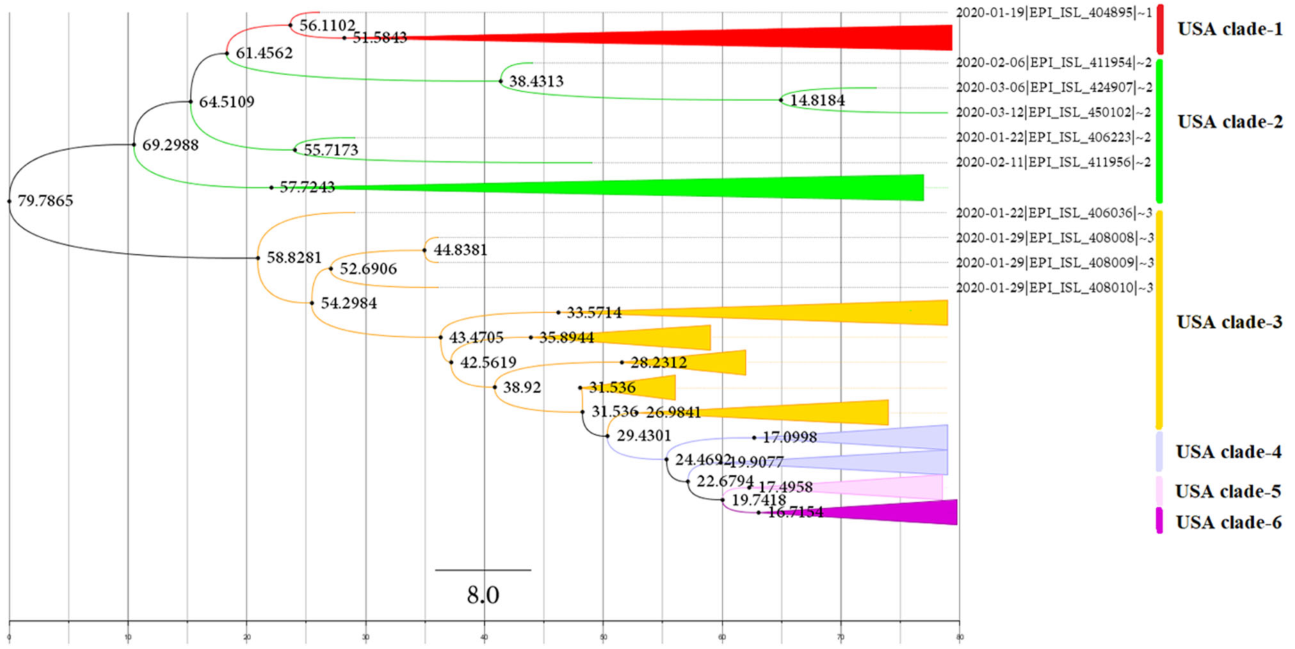

FIGURE 2 Bayesian phylogenetic tree of 265 SARS-CoV-2 strains in the United States from January 19 to March 12, 2020, showing the divergency time of six clades. The red colour represent USA clade-1, green represent USA clade-2, orange represent USA clade-3, blue represent USA clade-4, pale pink represent USA clade-5 and USA clade-6 use purple to represent. The details of each clade are shown in Figure S2

start to appear, but USA clade-1 seem to diverge from USA clade-2 on January 16, 2020. In the other aspect, USA clade-3 emerges around January 13, 2020 and it is also the appearance date of the common ancestor of USA clade-4 to USA clade-6, which is diverged into USA clade-4 on February 22, USA clade-5 on January 25 and USA clade-6 on February 26. The details of each clade are in Figure S2.

3.2 | Population expansion analysis

Mismatch distribution can be used to seek for the trace of virus population expansion.[41] The mismatch distribution of the assemblies and different genes are shown as Figures 3 and S4. From Figure 3, it is known the mismatch distribution curve shows a multipeak state, suggesting that the assemblies used to have experienced several dominant mutations in SARS-CoV-2 evolution, which may help itself better adapt to the environment at that time. Combining Figures S4A, S4B, S4C, and S4D in Figure S4, it is seen that the mismatch distribution curves of ORF1a, ORF1b, S and ORF3a show a pattern of unimodal Poisson distribution, which means that the four genes had experienced a rapid variation expansion in the past and the genes mutants have been significantly increasing. From protein analysis in Tables 1 and S2, it is known that the ORF1a mutation T265I may play an important role in the evolution from USA clade-5 to USA clade-6, the same as the ORF1b mutation P4715L from USA clade-3 to USA clade-4, S mutation D614G from USA clade-3 to USA clade-4 and the ORF3a mutation Q57H in from USA clade-4 to USA clade-5. Linked to the function of the proteins, it is speculated that the four mutations proteins may help SARS-CoV-2 easily bind to the host and beneficially proliferate themselves. On the contrary, from the other figures, E to K in Figure S4, it is known that mismatch distribution curves of other genes are similar to their expectation curves, indicating that these genes are relatively conservative in the evolution of SARS-CoV-2.

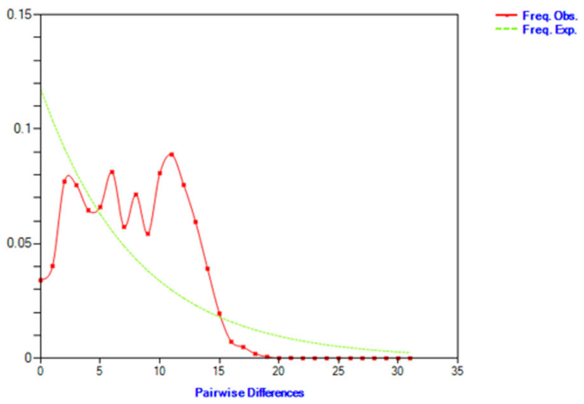

FIGURE 3 Mismatch distribution analyses of assemblies sequences alignment of 265 SARS-CoV-2 strains

3.3 | Viral proteins sequences alignment

Proteins sequences are aligned to find out the common and different characteristics of each clade. Twelve viral proteins alignments of 267 strains are seen as Tables S2 and 1, in which it is found that compared with RaTG13, both E and ORF6 of 265 strain in United States have no mutation site, but the second amino acid, S mutates into I in ORF7b and the 35th amino acid, L mutates into F in ORF10, which may be related to the choice of the host. Each clade from phylogenetic tree has special mutation sites characteristics. Compared with the other clades and RaTG13, USA clade-2 have some unique mutation sites (most of members), such as the 75th amino acid (D75E), the 265 amino acid (P971L) in ORF1a and the 62 amino acid (V62L) in ORF8. However, from Tables 1 and S2, it is easily seen that USA clade-1 is same as the ancestral protein sequence in these mutation sites, seeming to indicate USA clade-1 is the oldest clade, just as Figure 1. Except for both USA clade-1 and USA clade-2, it is found that the number of mutation sites gradually increases with the appearance of USA clade-3 to USA clade-6 and the 84th amino acid, S mutates into L in ORF8 in the clades. In addition, the

TABLE 1 Characteristics of amino acids mutations of different genes among six clades of 265 SARS-CoV-2 strains in the United States from January 19 to March 12, 2020

	Mutation sites										
Gene	ORF1a	ORF1a	ORF1a	ORF1b	S	ORF3a	ORF3a	ORF8	ORF8	ORF8	N	N
Position	75	265	971	4715	614	57	251	24	62	84	203	204
USA clade-1(53)	D53	T53	P53	P53	D53	Q53	G53	S53	V53	S53	R53	G53
USA clade-2(20)	**E14D6**	T20	**L14P6**	P20	D20	Q20	G20	S20	**L15V5**	S20	R20	G20
USA clade-3(46)	D46	T46	P46	P46	D46	Q46	**G31V15**	S46	V46	**L46**	R46	G46
USA clade-4(30)	D30	T30	P30	**L30**	**G30**	Q30	G30	S30	V30	**L30**	**R21K9**	**G21R9**
USA clade-5(40)	D40	T40	P40	**L40**	**G40**	**H40**	G40	S40	V40	**L40**	R40	G40
USA clade-6(76)	D76	**I76**	P76	**L76**	**G76**	**H76**	G76	**S69L7**	V76	**L76**	R76	G76
RATG13	D	T	P	P	D	Q	G	S	V	S	R	G
Wuhan-Hu-1	D	T	P	P	D	Q	G	S	V	**L**	R	G

Note: Bold font represent mutation sites of six USA clades. The capital letters represents an amino acids. The numbers after clades and amino acids represent the number of virus strains. "Position" means the site of an amino acid in gene shown in the preceding row. The details can be found in Table S2.

mutation site (G215V, G31V15) in ORF3a is the unique clade characteristics of USA clade-3. Compared with ancestral clades, in USA clade-4 to USA clade-6, the 4,715th amino acid, P mutates into L in ORF1b and the 614th amino acid, D mutates into G in protein S, which form a pair of coupling mutation sites. USA clade-4 diverges into USA clade-5 and USA clade-6, represented by the mutation site, Q57H in ORF3a. USA clade-6 is diverged from USA clade-5, with both the sites of T265I site in ORF1a and S24L(S69L7) in ORF8 as a symbol. Moreover, a pair of coupling mutations, R203K and G204R in protein N are found in USA clade-4.

3.4 | Nucleotide substitution rates analysis

The synonymous substitution rate (Ks), nonsynonymous substitution rate (Ka) and the ω ratio (ω = Ka/Ks) can be used to infer the pressure of evolution in genes of SARS-CoV-2. According to the theory of Michael et al.,[57] Ks < 0.05 may indicate that genes experience a phase of accelerated proteins evolution in their early evolution history, followed by the period of a gradual increase in selective constraint, with the progressive decline of ω.

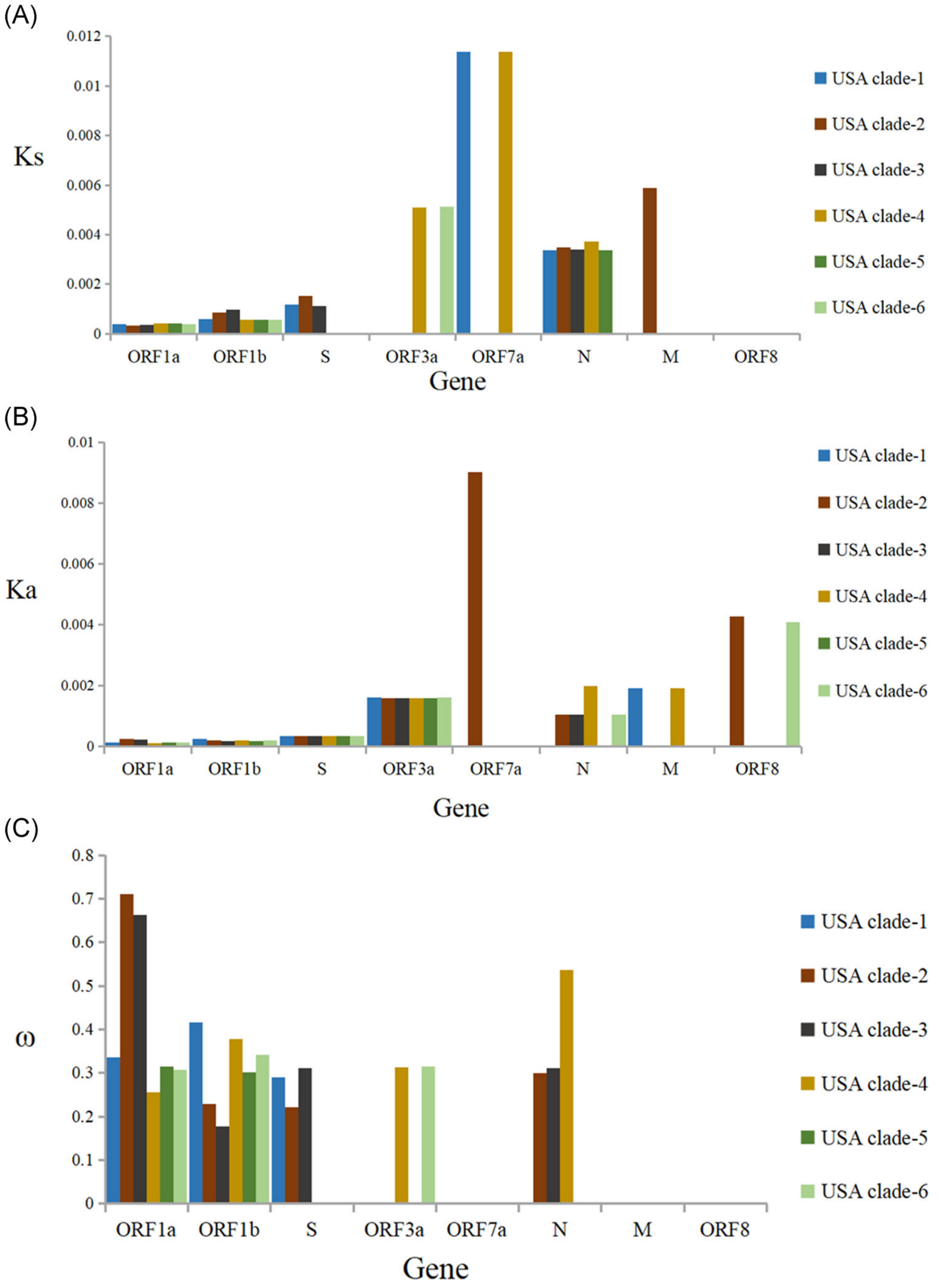

FIGURE 4 Ks: Nucleotide synonymous substitution rates (A), Ka: nonsynonymous substitution rates (B); ω: the ratio of nonsynonymous substitution rate to synonymous substitution rate (C) of different clades in each gene and the assemblies

Ks, Ka, and ω of viral gene from each clade can be seen as Figure 4 and Table 2, from which it is found that Ks of each gene is far <0.05 in all clades, indicating the viral genes are in the early stage of the evolution. For example, Ks of ORF8 is zero in all clades and Ka is also zero in five clades, suggesting ORF8 gene may be in the earliest period of evolutionary process. It is the same as other genes, such as ORF3a, ORF7a, M in most of clades, and S protein in clade-4 to clade-6. In the period, SARS-CoV-2 is experiencing accelerated evolution of some proteins in several clades, such as S in clade-4 to clade-6 ($Ks = 0$, $Ka > 0$). In addition, for each viral gene, the value of Ks is steady among different clades except for ORF1b in clade-2 and clade-3, partly verifying the assumption that synonymous substitutions are largely immune from selection and accumulate at a stochastic rate that is proportional to time. On the other aspect, similar to Ks, the value of Ka is also steady among all clades except for ORF1a in clade-2 and clade-3. Except for the genes of $Ks = 0$ or $Ka = 0$, the value of ω is smaller than 1, meaning that most of genes from SARS-CoV-2 were mainly affected by purifying selection. From Figure 4, it is observed that for ORF1b in USA clade-2, USA clade-3 and N in USA clade-4, the values of ω are higher, indicating both of genes are under greater purifying pressure. For ORF1a, ω progressively decline from USA clade-2 to USA clade-6, which may mean that the strain gradually adapts to the environment with evolutionary process.

3.5 | PCA analysis

PCA analysis is usually used to analyze the potential evolutionary trends of genes codon usage patterns. PCA analysis plot of 265 assemblies is shown in Figure 5, from which it is seen that USA clade-1 distributes in the first quadrant, USA clade-2 distributes near the horizontal positive

TABLE 2 Ks: Nucleotide synonymous substitution rates (A), Ka: nonsynonymous substitution rates (B); ω: the ratio of nonsynonymous substitution rate to synonymous substitution rate (C) of different clades in each gene and the assemblies

A

	Synonymous substitution rate							
Clade	ORF1a	ORF1b	S	ORF3a	ORF7a	N	M	ORF8
USA clade-1	0.000393400	0.000587098	0.001191506	0	0.011363854	0.003365121	0	0
USA clade-2	0.000334869	0.000863252	0.001528091	0	0	0.003492679	0.005893940	0
USA clade-3	0.000356284	0.000981839	0.001108461	0	0	0.003403186	0	0
USA clade-4	0.000411639	0.000560392	0	0.005102210	0.011363854	0.003724469	0	0
USA clade-5	0.000406571	0.000560495	0	0	0	0.003365121	0	0
USA clade-6	0.000397937	0.000567960	0	0.005110869	0	0	0	0

B

	Nonsynonymous substitution rate							
Clade	ORF1a	ORF1b	S	ORF3a	ORF7a	N	M	ORF8
USA clade-1	0.000132088	0.000244478	0.000346517	0.001608595	0	0	0.001926165	0
USA clade-2	0.000237824	0.000197298	0.000338142	0.001593414	0.009016407	0.001043297	0	0.004261928
USA clade-3	0.000236377	0.000174444	0.000344039	0.001593203	0	0.001059123	0	0
USA clade-4	0.000105171	0.000212060	0.000338238	0.001592780	0	0.001996637	0.001928022	0
USA clade-5	0.000128122	0.000168839	0.000338162	0.001593203	0	0	0	0
USA clade-6	0.000122386	0.000194281	0.000342772	0.001610483	0	0.001043025	0	0.004077137

C

	The ratio of nonsynonymous substitution rate to synonymous substitution rate							
Clade	ORF1a	ORF1b	S	ORF3a	ORF7a	N	M	ORF8
USA clade-1	0.335761345	0.416418618	0.290823082	-	0	0	-	-
USA clade-2	0.710201161	0.228551726	0.221284247	-	-	0.298709700	0	-
USA clade-3	0.663450957	0.177670197	0.310375695	-	-	0.311215042	-	-
USA clade-4	0.255493386	0.378414538	-	0.312174514	0	0.536086453	-	-
USA clade-5	0.315128706	0.301230902	-	-	-	0	-	-
USA clade-6	0.307551500	0.342068812	-	0.315109399	-	-	-	-

axis, USA clade-3 in the fourth quadrant, and most points of USA clade-4 to clade-6 in the second and third quadrant. The distribution zones of USA clade-1 and USA clade-3 are relatively independent, but USA clade-2 seem to partly overlap with the area of USA clade-1 or USA clade-3, suggesting that USA clade-2 may be a interim clade of USA clade-1 to USA clade-3 in the virus evolution. USA clade-4, USA clade-5 and USA clade-6 concentrate together, indicating the clades have many common mutation sites or amino acid sequences. There is no overlap between USA clade-6 and three ancestral clades (USA clade-1 to USA clade-3), implying compared to other clades, USA clade-6 is obviously different from its ancestors and it was also the newest product of SARS-CoV-2 evolution.

4 | DISCUSSION

In the study, we analyzed the evolutionary characteristics of SARS-CoV-2 in United States from January 19 to March 12, 2020 and found the strains are classified into six clades (USA clade-1 to USA clade-6). The common ancestor of the clades may appear on December 24, 2019.

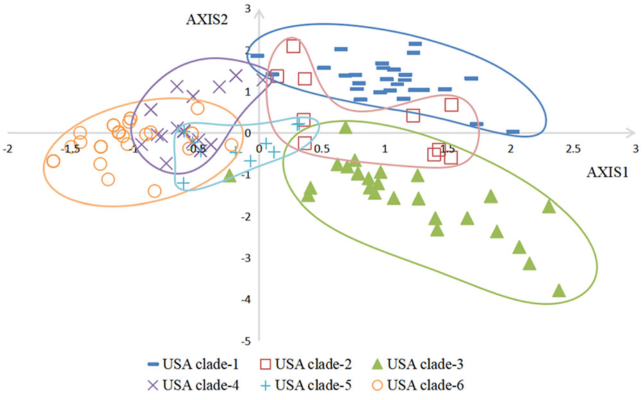

FIGURE 5 PCA in each clade of the assemblies. A plot with Axis 1 against Axis 2 was plotted based on RSCU values of all genes. Different clades are represented by different colors and shapes. PCA, principal component analysis; RSCU, relative synonymous codon usage

Though Bayesian analysis seems to indicate that USA clade-1 diverged from USA clade-2, other more arguments suggest that the former is closer to the ancestors, such as phylogenetic tree, mutation sites similarity and codon usage pattern from PCA analysis. In other words, USA clade-1 should be the oldest clade and the ancestor of SARS-CoV-2 in United States. The number of strains in USA clade-1 also indicates the clade used to experience a scale of expansion and be more suitable to the environment at the time. USA clade-2 is the closest to USA clade-1 and USA clade-3 in kinship (Figure 1), partly overlaps with both clades in codon usage pattern (Figure 5), and has the fewest members, which suggests that USA clade-2 is a interim clade of the evolution process from clade-1 to clade-3.

Compared with traditional classification, it is easily found that USA clade-1 and USA clade-2 can merge into S clade in GISAID website (Figure 6),[47] in which the 8782nd nucleotide is T and the 28,144th is C, in accordance with Lineage A in the study of Jennifer et al.[44] Compared with both the clades, the 84th amino acid, S mutates into L in ORF8 and the number of mutation sites gradually increases in USA clade-3 to USA clade-6, characterized by lineage B in Jennifer et al.[44] According to the classification in GISAID website, USA clade-3 is composed of the strains of L clade and V clade. USA clade-4 is composed of G clade and GR clade. USA clade-5 and USA clade-6 are composed of GH clade. However, clade O is scattered among the four clades (Table S3 and Figure 6). The study of BII/GIS from A*STAR Singapore as of January 22, 2021[47] demonstrates that clade S is the ancestor, from which the clades of L and V are diverged, followed by clade G, clade GH, clade GR, clade GV and so on. Yu et al.[41] also found that H13 and H35 (belonging to S clade) were ancestral haplotypes, which is simlar to our conclusion.

Wuhan-Hu-1 (EPI_ISL_402125) is always regarded as the reference sequence or the ancestor sequence in phylogenetic analysis of SARS-CoV-2 strains[42,43,49,49,50,52] but the strain belongs to USA clade-3 in our research, the clade of L in GISAID website, Lineage B in Jennifer et al., which suggest that it is necessary to seek for a new strain as the source of ancestral or reference sequence. Combining 24 genome sequences in December 2019, we think the candidate should be EPI_ISL_529213. In other words, SARS-CoV-2 may have spread widely in some unknown manners before outbreak.

The characteristic mutation sites in different clades may take part in the viral replication and transmission. For example, during the

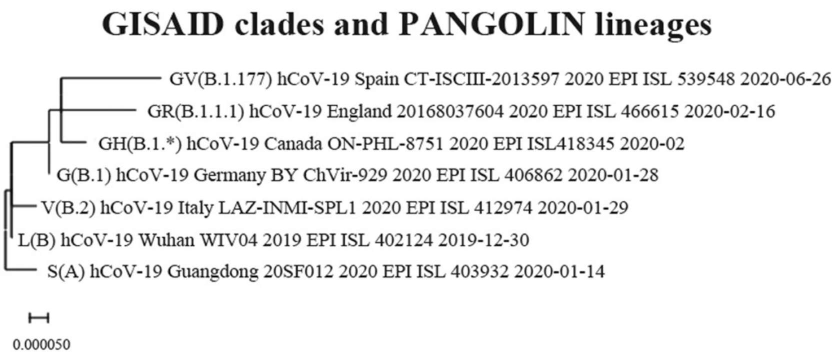
GISAID clades and PANGOLIN lineages
GV(B.1.177) hCoV-19 Spain CT-ISCIII-2013597 2020 EPI ISL 539548 2020-06-26
GR(B.1.1.1) hCoV-19 England 20168037604 2020 EPI ISL 466615 2020-02-16
GH(B.1.*) hCoV-19 Canada ON-PHL-8751 2020 EPI ISL418345 2020-02
G(B.1) hCoV-19 Germany BY ChVir-929 2020 EPI ISL 406862 2020-01-28
V(B.2) hCoV-19 Italy LAZ-INMI-SPL1 2020 EPI ISL 412974 2020-01-29
L(B) hCoV-19 Wuhan WIV04 2019 EPI ISL 402124 2019-12-30
S(A) hCoV-19 Guangdong 20SF012 2020 EPI ISL 403932 2020-01-14
0.000050

FIGURE 6 Phylogenetic tree and clades of the global viruses as of January 22, 2021, acquired in GISAID[46]

evolution from the first three clades (USA clade-1, USA clade-2, and USA clade-3) to the last three clades (USA clade-4, USA clade-5, and USA clade-6), the mutation D614G in S protein has been proved to strengthen viral infection and viral replication in human epithelial cells[56] and the strong linkage with P4715L of ORF1ab showed significant positive correlations with fatality rates.[52] Similarly, during the evolution from the first four clades to the last two clades, the mutation Q57H in ORF3a protein may enhance the capacity of virion release and viral pathogenicity.[37,38] During the evolution from the first two clades to the last four clades, the mutation S84L in ORF8 protein may help virus escape from immune response.[39] During the evolution from other clades to the USA clade-6, the mutation T265I in ORF1a protein may help SARS-CoV-2 to form an advantage group. Moreover, for USA clade-4, newly strong linkage of R203K and G204R in N protein may help virus to for helical ribonucleocapsid[36]

All in all, we do not only describe the evolutionary process in six clades of SARS-CoV-2 in the early stage from the United States, but also identify their evolutionary direction and characteristics, which supplements the strain classification and helps to infer the current evolution trend of SARS-CoV-2. In addition, we newly find many special mutations in viral proteins in different clades, which lay foundation to study the function of viral mutational protein.

ACKNOWLEDGEMENTS

This study was supported by Research Interest Training Scheme from Sichuan Agricultural University (Project No.: 2021582).

AUTHOR CONTRIBUTION STATEMENT

Conceptualization, methodology, software, formal analysis, investigation, writing—original draft, visualization: Ziying Lin. *Data curation, formal analysis, and visualization*: Hua Qing. *Software and data curation*: Rui Li. *Software and data curation*: Lei Zheng. *Validation, supervision, and writing—original draft*: Huipeng Yao.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article.

ORCID

Huipeng Yao http://orcid.org/0000-0001-5078-6492

REFERENCES

1. Malik YA. Properties of coronavirus and SARS-CoV-2. *Malays J Pathol*. 2020;42(1):3-11.
2. Stasi C, Fallani S, Voller F, Silvestri C. Treatment for COVID-19: an overview. *Eur J Pharmacol*. 2020;889:173644.
3. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265-269.
4. Aryaya R, Kumari S, Pandey B, et al. Structural insights into SARS-CoV-2 proteins. *J Mol Biol*. 2021;433(2):166725.
5. Alanagreh L, Alzoughool F, Atoum M. The human coronavirus disease COVID-19: its origin, characteristics, and insights into potential drugs and its mechanisms. *Pathogens*. 2020;9:331.
6. Helmy YA, Fawzy M, Elaswad A, Sobieh A, Kenney SP, Shehata AA. The COVID-19 pandemic: a comprehensive review of taxonomy, genetics, epidemiology, diagnosis, treatment, and control. *J Clin Med*. 2020;9:1225.
7. Chen Y, Liu Q, Guo D. Emerging coronaviruses: genome structure, replication, and pathogenesis. *J Med Virol*. 2020;92:418-423.
8. Kamitani W, Huang C, Narayanan K, Lokugamage KG, Makino S. A two-pronged strategy to suppress host protein synthesis by SARS coronavirus Nsp1 protein. *Nature Struct. Mol. Biol*. 2009;16:1134-1140.
9. Harcourt BH, Jukneliene D, Kanjanahaluethai A, et al. Identification of severe acute respiratory syndrome coronavirus replicase products and characterization of papain-like protease activity. *J Virol*. 2004;78:13600-13612.
10. Barretto N, Jukneliene D, Ratia K, Chen Z, Mesecar AD, Baker SC. The papain-like protease of severe acute respiratory syndrome coronavirus has deubiquitinating activity. *J Virol*. 2005;79:15189-15198.
11. Barretto N, Jukneliene D, Ratia K, Chen Z, Mesecar AD, Baker SC. Deubiquitinating activity of the SARS-CoV papain-like protease. *Adv Exp Med Biol*. 2006;581:37-41.
12. Lindner HA, Fotouhi-Ardakani N, Lytvyn V, Lachance P, Sulea T, Me´nard R. The papain-like protease from the severe acute respiratory syndrome coronavirus is a deubiquitinating enzyme. *J Virol* . 2005;79:15199-15208.
13. Lindner HA, Lytvyn V, Qi H, Lachance P, Ziomek E, Me´nard R. Selectivity in ISG15 and ubiquitin recognition by the SARS coronavirus papain-like protease. *Arch Biochem Biophys*. 2007;466:8-14.
14. Frick DN, Virdi RS, Vuksanovic N, Dahal N, Silvaggi NR. Molecular basis for ADP-ribose binding to the Mac1 domain of SARS-CoV-2 nsp3. *Biochemistry*. 2020;59:2608-2615.
15. Michalska K, Kim Y, Jedrzejczak R, et al. Crystal structures of SARS-CoV-2 ADP-ribose phosphatase: from the apo form to ligand complexes. *IUCrJ*. 2020;7:814-824.
16. Alhammad YMO, Kashipathy MM, Roy A, et al. The SARS-CoV-2 conserved macrodomain is a highly effificient ADP-ribosylhydrolase. *BioRxiv*. 2020
17. Angelini MM, Akhlaghpour M, Neuman BW, Buchmeier MJ. Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles. *mBio*. 2013;4:e00524-13.
18. Ziebuhr J, Snijder EJ, Gorbalenya AE. Virusencoded proteinases and proteolytic processing in the Nidovirales. *J Gen Virol*. 2000;81:853-879.
19. Gao Y, Yan L, Huang Y, et al. Structure of the RNAdependent RNA polymerase from COVID-19 virus. *Science*. 2020;368:779-782.
20. Yin W, Mao C, Luan X, et al. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science*. 2020;368:1499-1504.
21. Miknis ZJ, Donaldson EF, Umland TC, Rimmer RA, Baric RS, Schultz LW. Severe acute respiratory syndrome coronavirus nsp9 dimerization is essential for efficient viral growth. *J Virol*. 2009;83:3007-3018.
22. Romano M, Ruggiero A, Squeglia F, Maga G, Berisio R. A structural view of SARS-CoV-2 RNA replication machinery: RNA synthesis, proofreading and final capping. *Cells*. 2020;9:1267.
23. Chen J, Malone B, Llewellyn E, et al. Structural basis for helicase-polymerase coupling in the SARS-CoV-2 replication-transcription complex. *Cell*. 2020;182:1560-1573.e13.
24. Jia Z, Yan L, Ren Z, et al. Delicate structural coordination of the severe acute respiratory syndrome coronavirus Nsp13 upon ATP hydrolysis. *Nucleic Acids Res*. 2019;47:6538-6550.
25. Ma Y, Wu L, Shaw, et al. Structural basis and functional analysis of the SARS coronavirus nsp14–nsp10 complex. *PNAS*. 2015;112:9436-9441.
26. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ. Nidovirales: evolving the largest RNA virus genome. *Virus Res*. 2006;117:17-37.
27. Koonin EV, Moss B. Viruses know more than one way to don a cap. *Proc Natl Acad Sci USA*. 2010;107:3283-3284.

28. Kim Y, Jedrzejczak R, Maltseva NI, et al. Crystal structure of Nsp15 endoribonuclease NendoU from SARS-CoV-2. *Protein Sci*. 2020; 29:1596-1605.
29. Bouvet M, Debarnot C, Imbert I, et al. In Vitro Reconstitution of SARS-Coronavirus mRNA Cap Methylation. *PLOS Pathog*. 2010;6: e1000863.
30. Decroly E, Debarnot C, Ferron F, et al. Crystal structure and functional analysis of the SARS-coronavirus RNA Cap 20-O-methyltransferase nsp10/nsp16 complex. *PLOS Pathog*. 2011;7: e1002059.
31. Wang Q, Zhang Y, Wu L, et al. Structural and functional basis of SARS-CoV-2 Entry by using human ACE2. *Cell*. 2020;181:894-904.e9.
32. Zhou P, Yang XL, X.G., Hu, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579: 270-273.
33. Choudhury A, Mukherjee S. In silico studies on the comparative characterization of the interactions of SARS-CoV-2 spike glycoprotein with ACE-2 receptor homologs and human TLRs. *J Med Virol*. 2020;92(10):2105-2113. https://doi.org/10.1002/jmv.25987
34. Chang C, Chen C-MM, Chiang M, Hsu Y, Huang T. Transient oligomerization of the SARS-CoV N protein–implication for virus ribonucleoprotein packaging. *PLOS One*. 2013;8:e65045.
35. He R, Leeson A, Ballantine, et al. Characterization of protein–protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus. *Virus Res*. 2004;105:121-125.
36. Zhao X, Nicholls JM, Chen Y-G. Severe acute respiratory syndrome-associated coronavirus nucleocapsid protein interacts with smad3 and modulates transforming growth factor-b signaling. *J Biol Chem*. 2008;283:3272-3280.
37. Castan˜o-Rodriguez C, Honrubia JM, Gutiérrez-Álvarez IJ, et al. Role of severe acute respiratory syndrome coronavirus viroporins E, 3a, and 8a in replication and pathogenesis. *mBio*. 2018;9(3):e02325-17.
38. Ren Y, Shu T, Wu D, et al. The ORF3a protein of SARS-CoV-2 induces apoptosis in cells. *Cell Mol Immunol*. 2020;17:881-883.
39. Li J-Y, Liao C-H, 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.
40. Jiang HW, Zhang HN, Meng QF, et al. SARS-CoV-2 Orf9b suppresses type I interferon responses by targeting TOM70. *Cell Mol Immunol*. 2020;17:998-1000.
41. Yu WB, Tang GD, Zhang L, Zhang L, Corlett RT. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2) using whole genomic data. *Zool Res*. 2020;41(3): 247-257.
42. Forster P, Forster L, Renfrew C, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci USA*. 2020;117(17):9241-9243.
43. Castells M, Lopez-Tort F, Colina R, Colina R, Cristina J. Evidence of increasing diversification of emerging SARS-CoV-2 strains. *J Med Virol*. 2020;92(10):2165-2172.
44. Giandharii J, Pillay S, Wilkinson E, et al. Early transmission of SARS-CoV-2 in South Africa: an epidemiological and phylogenetic report. *Int J Infect Dis*. 2021;103:234-241.
45. Bajaj P & Arya PC Evolution and spread of SARS-CoV-2 likely to be affected by climate. In press. https://www.biorxiv.org/content/10.1101/2020.06.18.147074v3
46. Houriiyah T, Eduan W, Lessells RJ, et al. Sixteen novel lineages of SARS-CoV-2 in South Africa. *Nat Med*. 2021;27:440-446.
47. BII/GIS. Phylogenetic tree and clades of global area as of January 22, 2021 acquired in platform GISAID. January 22, 2021. https://www.epicov.org
48. Wang C, Liu Z, Chen Z, et al. The establishment of reference sequence for SARS-CoV-2 and variation analysis. *J Med Virol*. 2020; 92(6):667-674.
49. Pachetti M, Marini B, Benedetti F, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J Transl Med*. 2020;18(1):179.
50. Van Dorpan Dorp L, Acman M, Richard D, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol*. 2020;83:104351.
51. Koyama T, Platt D, Parida L. Variant analysis of SARS-CoV-2 genomes. *Bull World Health Organ*. 2020;98(7):495-504.
52. Toyoshima Y, Nemoto K, Matsumoto S, Nakamura Y, Kiyotani K. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. *J Hum Genet*. 2020;65(12):1075-1082.
53. Benvenuto D, Angeletti S, Giovanetti M, et al. Evolutionary analysis of SARS-CoV-2: how mutation of non-structural protein 6 (Nsp6) could affect viral autophagy. *J Infect*. 2020;81(1):e24-e27.
54. Phan T. Genetic diversity and evolution of SARS-CoV-2. *Infect Genet Evol*. 2020;81:104260.
55. Mathavan S & Kumar S Evaluation of the effect of D614G, N501Y and S477N mutation in SARS-CoV-2 through computational approach. In press. https://doi.org/10.20944/preprints202012.0710.v1
56. Hou YJ, Chiba S, Halfmann P, et al. SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science*. 2020;370(6523):1464-1468.
57. lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science*. 2000;290(5494):1151-1155.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Lin Z, Qing H, Li R, Zheng L, Yao H. Evolution trace of SARS-CoV-2 from January 19 to March 12, 2020, in the United States. *J Med Virol*. 2021;93:6595-6604. https://doi.org/10.1002/jmv.27225