Received: 11 July 2021 | Accepted: 8 September 2021

DOI: 10.1002/jmv.27331

RESEARCH ARTICLE

JOURNAL OF MEDICAL VIROLOGY WILEY

Molecular evolutionary characteristics of SARS-CoV-2 emerging in the United States

Shihang Wang[1,2] | Xuanyu Xu[1,2] | Cai Wei[1,2] | Sicong Li[1,2] | Jingying Zhao[2,3] | Yin Zheng[1,2] | Xiaoyu Liu[1,2] | Xiaomin Zeng[4] | Wenliang Yuan[5] | Sihua Peng[1,2]

[1]Department of Virology, National Pathogen Collection Center for Aquatic Animals, Ministry of Agriculture of China, Shanghai, China

[2]Department of Developmental Biology, College of Fisheries and Life Science, Shanghai Ocean University, Shanghai, China

[3]Department of Health Care, School of Physical Education & Health Care, East China Normal University, Shanghai, China

[4]Department of Biostatistics, Central South University, Xiangya Public Health School, Changsha, China

[5]Department of Mathematics, College of Mathematics and Information Engineering, Jiaxing University, Jiaxing, China

Correspondence

Wenliang Yuan, Department of Mathematics, College of Mathematics and Information Engineering, Jiaxing University, Jiaxing 314033, China.
Email: yuanwl@zjxu.edu.cn

Sihua Peng, Department of Virology, National Pathogen Collection Center for Aquatic Animals, Ministry of Agriculture of China, Shanghai 201306, China.
Email: shpeng@shou.edu.cn

Funding information

Shanghai Natural Science Foundation, Grant/Award Number: 15zr1420800

Abstract

SARS-CoV-2 is a newly discovered beta coronavirus at the end of 2019, which is highly pathogenic and poses a serious threat to human health. In this paper, 1875 SARS-CoV-2 whole genome sequences and the sequence coding spike protein (S gene) sampled from the United States were used for bioinformatics analysis to study the molecular evolutionary characteristics of its genome and spike protein. The MCMC method was used to calculate the evolution rate of the whole genome sequence and the nucleotide mutation rate of the S gene. The results showed that the nucleotide mutation rate of the whole genome was 6.677×10^{-4} substitution per site per year, and the nucleotide mutation rate of the S gene was 8.066×10^{-4} substitution per site per year, which was at a medium level compared with other RNA viruses. Our findings confirmed the scientific hypothesis that the rate of evolution of the virus gradually decreases over time. We also found 13 statistically significant positive selection sites in the SARS-CoV-2 genome. In addition, the results showed that there were 101 nonsynonymous mutation sites in the amino acid sequence of S protein, including seven putative harmful mutation sites. This paper has preliminarily clarified the evolutionary characteristics of SARS-CoV-2 in the United States, providing a scientific basis for future surveillance and prevention of virus variants.

KEYWORDS

bioinformatics, molecular evolution, SARS-CoV-2, spike protein

1 | INTRODUCTION

Coronaviruses (CoV) is an enveloped RNA virus, widely distributed in humans and other mammals, and can cause a variety of diseases, e.g., respiratory intestinal, liver, and nervous system diseases.[1–6] There have been seven types of coronaviruses that infect humans, including SARS-CoV, MERS-CoV, HCoV-229E, HCoV-HKU1, HCoV-NL63, HCoV-OC43, and SARS-CoV-2,[7] and the genome sequence and spike protein structure of SARS-CoV-2 are similar to that of SARS-CoV.[8–10] The receptor binding and membrane fusion are the initial and critical steps in the SARS-CoV-2 infection cycle, during which the SARS-CoV-2 spike protein (S protein) plays a key role. Therefore, it is very important to study the SARS-CoV-2 S protein.[11,12]

As of July 2021, the number of SARS-CoV-2 infections has reached 180 million people worldwide. As for the United States, more than 33 million COVID-19 cases have been diagnosed, which is close to a quarter of the total number of confirmed cases worldwide. The death toll in the United States exceeds 60 million.[13]

In previous studies, Li et al.[12] analyzed the global evolution rate of SARS-CoV-2 in the first month of the outbreak, with an estimated mean nucleotide mutation rate ranging from 1.7926×10^{-3} to 1.8266×10^{-3} substitution per site per year. Four months after the outbreak, the mutation rate became 3.95×10^{-4} per nucleotide per year,[14] which was almost seven times lower than the mutation rate of SARS-CoV and two times lower than that of MERS-CoV.[15–17] Motayo et al.[18] reported that the evolution rate of SARS-CoV-2 in Africa from February 24 to April 24 was 4.133×10^{-4} substitution per nucleotide per year. The Nextstrain website estimates that the annual nucleotide evolution rate is 8×10^{-4} substitution per nucleotide per year based on current statistics.[19] As for the evolution rate of the gene encoding the S protein, Pereson et al.[20] reported that the global evolution rate was 2.19×10^{-3} nucleotide substitutions per site per year as of April 2020, while the evolution rate reduced to 1.08×10^{-3} as of September 2020.[21] However, so far, no relevant research has been reported on the US cases, so we initiate this study.

2 | MATERIALS AND METHODS

2.1 | Sequence data collection

2.1.1 | Various coronavirus genome sequence data for the phylogenetic analysis

A total of 15 whole genome sequences and the corresponding S gene sequences were employed in this study from the NCBI database (https://www.ncbi.nlm.nih.gov/) (Table S1), including SARS-CoV-2, SARS-CoV, MERS-CoV, HCoV-229E, HCoV-OC43, HCoV-NL63, HCoV-HKU1, Bat SARS-like, Civet SARS-CoV, Pangolin CoV, Murine hepatitis virus, Rat CoV, Erinaceus CoV, Camel MERS-CoV, and Bovine CoV.

2.1.2 | SARS-CoV-2 genome sequence and the corresponding S gene sequence data collected from US cases to study molecular evolution

A total of 2241 whole-genome sequences of SARS-CoV-2 and the corresponding S gene sequences (from 2020/2/27 to 2021/4/8) were obtained from the severe acute respiratory syndrome coronavirus 2 data hub of NCBI Virus (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) (Table S2). After removing the low-quality sequences (with more than 10 ambiguous nucleotides), 1875 SARS-CoV-2 genome sequences were retained for further analysis, using MN908947 (isolate from Wuhan-Hu-1) as a reference sequence.

In addition, another 69 098 S protein amino acid sequences obtained from the NCBI database were used for the mutation analysis of amino acids, which were also retrieved from the SARS-CoV-2 cases in the United States.

2.2 | Sequence alignment and phylodynamics analysis

A multiple sequence alignment of the 1875 SARS-CoV-2 genome sequences was performed using MAFFT v7.464.[22] ModelFinder was used to analyze the optimal substitution model for the genome sequences based on the results of Bayesian Information Criterion,[23] with a result of the best nucleotide substitution model to be GTR+F+G4.

Mutation rate can be defined as the number of mutations per cell division, per generation or per unit time.[24] Markov Chain Monte Carlo (MCMC) method was used to reconstruct the maximum clade credibility (MCC) tree and calculate the mutation rate by BEAST V2.6.2.[25] To set the time scale prior for the dataset, we used a constrained evolution rate with a Log-normal prior averaged at 10^{-3} by substitution per site per year. We performed a phylogenetic Bayesian analysis using the Relaxed Clock Log-Normal molecular clock model and selected the Coalescent Bayesian Skyline as the model of population size and growth according to the relevant studies.[26–28] The whole-genome sequence and the sequence coding spike protein sequence (S gene) were separately analyzed by MCMC to calculate the mutation rate, with a length of 4×10^{8} steps, sampling every 4×10^{4} steps. The convergence of all the parameters (ESS >200, burn-in 10%) was verified with Tracer v1.7.1.[29] The final MCC tree was generated by TreeAnnotator (a software package in BEAST2) and displayed in Figtree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

2.3 | Adaptive evolution analysis and recombination analysis

The SARS-CoV-2 MN908947 was used as the reference sequence, which was aligned against the genome sequences of the other types of coronavirus by MAFFT. IQ-Tree 2.0.3 was used to establish the ML tree.[30] The CODEML program in PAML was used for the selection pressure analysis.[31] In the branch-site model, the SARS-CoV-2 was set as the foreground and the other viruses as the background. Full genomes and S gene sequences were analyzed separately to detect recombination events using RDP4 and SimPlot.[32,33]

2.4 | The S protein nonsynonymous or indel variant biological function analysis

PROVEAN was used to predict whether nonsynonymous mutations of S protein would affect its function,[34] which can detect harmful substitution of amino acid and predict whether the substitution will affect its phenotype. A PROVEAN score of −2.5 or lower indicates that amino acid substitution is harmful, and a score higher than −2.5 is considered neutral.

3 | RESULTS AND DISCUSSION

3.1 | Mutation rate in SARS-CoV-2

According to the results of BEAST2, we obtained the mutation rate of 6.677×10^{-4} per site per year (95% highest posterior density [HPD]: 6.117×10^{-4} to 7.270×10^{-4}) for the whole genome and of 8.066×10^{-4} per site per year (95% HPD: 5.969×10^{-4} to 1.038×10^{-3}) for the sequence coding S protein. The first more infectious mutation discovered by the scientists is D614G, which is caused by the change of base A to G at position 23403.[35] We identified 12 high-frequency mutations in the S gene among the 69 098 amino acid sequences, with the highest mutation frequency to be 98.47% at the position of amino acid (AA) 614 (Figure 1).

Figure 2 shows the MCC tree with Bayesian phylogeographic reconstruction of SARS-CoV-2 isolates. To make the image visual effect clearer without losing the representativeness of the genomes, we only show 380 genome sequences in Figure 2 to display the MCC tree, and the MCC tree including the complete list of 1875 sequences is shown in Figure S1. The detailed information obtained from BEAST2 results is shown in Table S3. We estimate that November 5, 2019, is the time of the most recent common ancestor of SARS-CoV-2 emerging in the United States, with the 95% HPD ranging from September 21, 2019, to December 16, 2019. This conclusion is also consistent with the first case reported in the literature on December 1, 2019.[36]

3.2 | Phylogenetic analysis of SARS-CoV-2

To explore the evolutionary selection pressure of SARS-CoV-2, a phylogenetic analysis was carried out using MN908947 as the reference sequence, and the whole genome of MN908947 was aligned with the other 14 coronaviruses to establish the phylogenetic tree. MN908947 was used as the foreground branch with the other 14 coronaviruses as the background branch for the branch-site model. The analysis results are shown in Table 1 and Table S4. Thirteen and two statistically significant positive sites were detected in the whole genome and S gene respectively ($p < 0.005$).

RDP4 and SimPlot were used to detect the recombination event of the 15 coronaviruses. However, the results showed that no statistically significant recombination event was found.

3.3 | Harmful mutation detection in S protein amino acids

We found that there were 101 nonsynonymous mutations in all the 69 098 protein sequences. The results showed that there were seven harmful mutations (Table 2), while the remaining 94 were neutral.

4 | DISCUSSION

SARS-CoV-2 has evolved into thousands of variants worldwide. Although the virus might mutate a lot, only a few could cause serious harm to humans.[37] Compared with other coronaviruses and RNA viruses, the mutation rate of SARS-CoV-2 (6.677×10^{-4}) is at a medium level. Moreover, the mutation rate gradually decreased during the subsequent infection process.[12,14,18] The reason for this may be that SARS-CoV-2 has its own copy "proofreading mechanism", which can correct some errors that may occur during the copying process, leading to the decrease of the mutation rate of SARS-CoV-2.[38] On the other hand, harmful mutations may lead to a certain degree of protein structure and functional changes, ultimately affecting the reproduction of SARS-CoV-2. Under the pressure of natural selection and with the accumulation of harmful mutations, the number of viruses will gradually decrease and even mutational meltdown may occur, leading to population extinction,[39,40] with a result of the relatively low evolution rate of SARS-CoV-2 (Table 3).

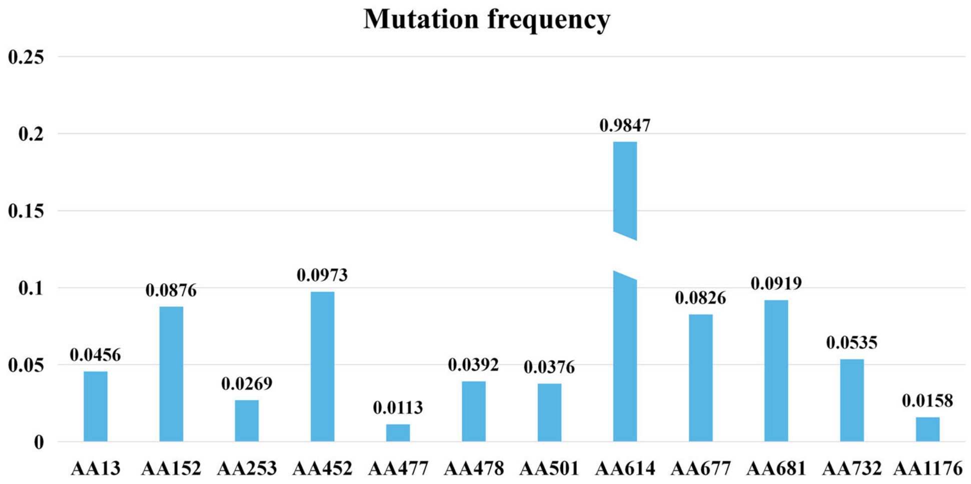

FIGURE 1 Distribution of mutations in the S protein. All mutations in the S protein are nonsynonymous, with position AA614 having the highest frequency

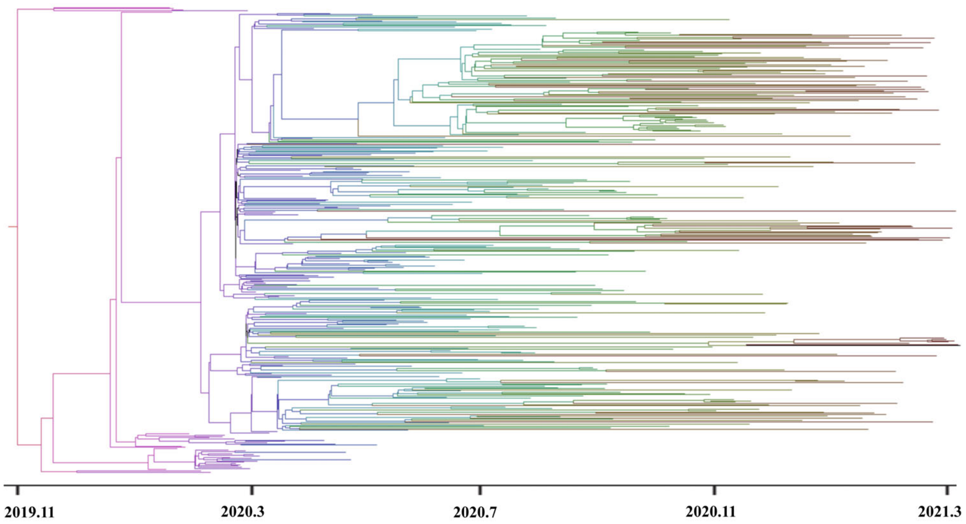

FIGURE 2 Bayesian maximum clade credibility tree

TABLE 1 Results of branch-site model for SARS-CoV-2

Model	Ln L	Parameter estimation					model comparison	LRT *p* values	Positive selection site
Model A	−157,747.79203	Site type	0	1	2a	2b	Model A vs. Model A null	0.00000	498V* 1,039M* 1,125V* 1,183 E* 1,388N* 1,592R* 1,968P* 2,001S* 2,020S* 2,169S* 2,233S* 2,257S* 2,360S*
		Site Ratio f	0.41494	0.35320	0.12525	0.10661			
		Background branch ω0	0.31558	1.00000	0.31558	1.00000			
		Detection of branch ω1	0.31558	1.00000	65.38486	65.38486			
Model A null	−157,831.26325	1							

TABLE 2 Seven harmful mutations detected in the S protein

Mutation	Score	Prediction (critical = −2.5)
P589S	−3.966	Deleterious
T716I	−3.293	Deleterious
D936Y	−2.602	Deleterious
S939F	−3.094	Deleterious
P1162S	−2.722	Deleterious
C1236F	−4.061	Deleterious
C1250F	−5.057	Deleterious

However, given that SARS-CoV-2 has not yet been effectively controlled in the United States, we cannot rule out the possibility that more new mutations will appear in the United States. Compared with other coronaviruses, 13 positive selection sites were detected in the SARS-CoV-2 genome, indicating that SARS-CoV-2 has a special evolution pattern.

In addition, nonsynonymous mutations were found in many samples of the S protein, with a harmful mutation proportion of 6.93%. These harmful mutations may affect the structure and function of the S proteins. Once the amino acid of RBD of S protein is mutated, the binding affinity with the human ACE2 receptor may change, which may cause the increase of the ability to infect humans and also make the existing vaccines ineffective.[54] Global mutation data collected from the GISAID database revealed that mutations occurred at almost every site of the S gene.[55] However, further experiments are needed to verify whether these mutations would indeed affect the function of the S protein.

Viruses are the masters of evolution, which create new variants by mutating and recombining in an unpredictable way during each

TABLE 3 The nucleotide mutation rate (substitutions per site per year) of different RNA virus

Group	Family	Virus	Mutation rate	Reference
ss(+)RNA	Coronaviridae	SARS-CoV-2	8×10^{-4}	19
ss(+)RNA	Coronaviridae	SARS	3.01×10^{-3}	17
ss(+)RNA	Coronaviridae	MERS-CoV	1.12×10^{-3}	15
ss(+)RNA	Coronaviridae	HCoV-OC43	1.06×10^{-4}	41
ss(+)RNA	Coronaviridae	HCoV-229E	3.28×10^{-4}	42
ss(+)RNA	Coronaviridae	Avian coronavirus	2.40×10^{-4}	43
ss(+)RNA	Coronaviridae	Bovine coronavirus	5.37×10^{-4}	44
ss(+)RNA	Filoviridae	EBOV	1.23×10^{-3}	45
ss(+)RNA	Picornaviridae	Hepatitis A virus	9.76×10^{-4}	46
ss(+)RNA	Flaviviridae	Hepatitis C virus	1.39×10^{-3}	47
ss(−)RNA	Orthomyxoviridae	Influenza A virus	3.15×10^{-3}	48
ss(+)RNA	Flaviviridae	Dengue virus	6.50×10^{-4}	49
ss(+)RNA	Picornaviridae	Human enterovirus A	5.53×10^{-3}	50
ss(+)RNA	Picornaviridae	Human enterovirus B	5.27×10^{-3}	50
ss(+)RNA	Picornaviridae	Poliovirus 1	1.17×10^{-2}	50
ss(−)RNA	Paramyxoviridae	Measles virus	6.02×10^{-4}	51
ss(−)RNA	Rhabdoviridae	Rabies virus	3.32×10^{-4}	52
dsRNA	Reoviridae	Human rotavirus A	1.87×10^{-3}	53

replication cycle. From the perspective of biological evolution, the existence of viruses is a natural selection pressure for human beings. The human body's immune response will produce a certain degree of adaptability, which will promote the development of human beings in a direction that is more conducive to human survival. However, the evolution of the human immune system and various interventions may also promote the adaptive mutations of the virus. Usually, viruses become "mild" as they circulate in the same host because high pathogenicity will lead to the death of the host, resulting in loss of transmission and reproduction, which is not conducive to the survival and reproduction of the virus itself. As mentioned above, the current mutation rate of SARS-CoV-2 in the United States is lower than that at the beginning of the outbreak. Obviously, this view is in line with our expectations.

Experimental results showed that some variants of SARS-CoV-2 strains increased the infectivity, such as variants D614G, S477N, and N439K, and these variants had higher transmission ability and faster replication speed than that of the original viruses, however, their pathogenicity did not increase.[35,56–58]

At present, more infectious variants of SARS-CoV-2 appearing in the world include the first D614G variant B.1.5-B.1.72, British variant B.1.1.7, South African variant B.1.351, Brazilian variant P.1, and Philippine variant P.3, as well as the two native varieties of the United States, the California variety B.1.429/B.1.427 and New York variety B.1.526. The spreading power, virulence, and immune evasion ability of new variants have been increasing.[59] Among them, the British variant B.1.1.7 is a virus that has a relatively large impact on the population. In December 2020, the new strain B.1.1.7 was discovered in the UK for the first time, and its S gene sequence accumulated 16 nucleotide mutations, resulting in 10 amino acid site changes (H69del, V70del, Y144del, N501Y, A570D, D614G, P681H, T716I, S982A, and D1118H).[60] According to the data from PANGO lineages, B.1.1.7 mutant strain has been found in more than 90 countries.[61] Studies have shown that the infectiousness of this mutant strain is more than 50% higher than that of the current prevailing strains.[62] As of April 10, the strain has caused 20,915 confirmed infections in the United States and has become the main type of transmission in the United States.[59] The California variant B.1.429/B.1.427 discovered in July 2020 has a mutation site of L452R, which leads to a 20% increase in its transmission power, and exhibits moderate immune evasion leading to a fairly fast transmission rate. The New York variant B.1.526 discovered in November 2020, whose mutation site is E484K or S477N, also exhibits moderate immune evasion.

At present, the main variants concerned in the United States include B1.1.7, B1.351, P.1, B.1.429/B.1.427. There is evidence that these variants could lead to increased infectivity, more severe disease, reduced effectiveness of treatments or vaccines, or diagnostic detection failures.[63–65] At present, the proportion of B1.1.7 has reached 44.1% and shows a significant trend of continuing to increase. Taking New York as an example, B1.1.7 accounted for 11.9% in February, 26.2% in mid-March, and 28.2% at the end of March. At the same time, the proportion of the New York native variant B.1.526 (E484K) is still as high as 27.9%. These different variants have greatly affected the effectiveness of the vaccine.

It is worth noting that the B.1.617 variant strain first appeared in India and has now become one of the most popular strains in the United States, and B.1.617 has a 110% higher affinity for human ACE2 receptor, making it highly infectious.[66,67]

In conclusion, we clarify the evolutionary characteristics of SARS-CoV-2 in the United States, providing a scientific basis for future surveillance and prevention of virus variants. To control SARS-CoV-2 and to restore people's normal life activities as soon as possible, it is necessary to continue to monitor specific mutations, which is still of great significance for further in-depth study of SARS-CoV-2 and the evaluation of the effectiveness of existing vaccines.

ACKNOWLEDGMENTS

This research was supported by the Shanghai Natural Science Foundation (SNSF, Grant No. 15zr1420800) to Sihua Peng. The authors would like to thank Dr. Fangluan Gao from Fujian Agriculture and Forestry University for providing key guidance assistance. The two reviewers provided very valuable comments on this article, and we are grateful for their contributions.

CONFLICT OF INTERESTS

The authors declare that there are no conflicts of interest.

AUTHOR CONTRIBUTIONS

Sihua Peng conceived and designed the study. Shihang Wang, Jingying Zhao, Yin Zheng, and Xiaoyu Liu collected data and prepared the datasets. Shihang Wang, Xuanyu Xu, and Cai Wei participated in phylogenetic analyses. Shihang Wang and Xuanyu Xu drafted the manuscript. Xiaomin Zeng, Wenliang Yuan, and Sihua Peng provided critical comments. All authors contributed to manuscript revision, read, and approved the submitted version.

DATA AVAILABILITY STATEMENT

Data derived from public domain resources.

ORCID

Sihua Peng http://orcid.org/0000-0001-7231-666X

REFERENCES

1. Peiris JS, Lai ST, Poon LL, et al. Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet*. 2003;361(9366):1319-1325. https://doi.org/10.1016/s0140-6736(03)13077-2
2. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet*. 2020;395(10223):507-513. https://doi.org/10.1016/s0140-6736(20)30211-7
3. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020;395(10223):497-506. https://doi.org/10.1016/s0140-6736(20)30183-5
4. Ksiazek TG, Erdman D, Goldsmith CS, et al. A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med*. 2003;348(20):1953-1966. https://doi.org/10.1056/NEJMoa030781
5. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265-269. https://doi.org/10.1038/s41586-020-2008-3
6. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet*. 2020;395(10229):1054-1062. https://doi.org/10.1016/s0140-6736(20)30566-3
7. Lu L, Liu Q, Du L, Jiang S. Middle East respiratory syndrome coronavirus (MERS-CoV): challenges in identifying its source and controlling its spread. *Microb Infect*. 2013;15(8-9):625-629. https://doi.org/10.1016/j.micinf.2013.06.003
8. Chan JF-W, Kok K-H, Zhu Z, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg Microbes Infect*. 2020;9(1):221-236. https://doi.org/10.1080/22221751.2020.1719902
9. Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *J Virol*. 2020;94(7):e00127-00120. https://doi.org/10.1128/jvi.00127-20
10. Xu X, Chen P, Wang J, et al. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *Science China-Life Sciences*. 2020;63(3):457-460. https://doi.org/10.1007/s11427-020-1637-5
11. Benvenuto D, Giovanetti M, Salemi M, et al. The global spread of 2019-nCoV: a molecular evolutionary analysis. *Pathog Glob Health*. 2020;114(2):64-67. https://doi.org/10.1080/20477724.2020.1725339
12. Li X, Wang W, Zhao X, et al. Transmission dynamics and evolutionary history of 2019-nCoV. *J Med Virol*. 2020;92(5):501-511. https://doi.org/10.1002/jmv.25701
13. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis*. 2020;20(5):533-534. https://doi.org/10.1016/s1473-3099(20)30120-1
14. Shen S, Zhang Z, He F. The phylogenetic relationship within SARS-CoV-2s: an expanding basal Glade. *Mol Phylogenet Evol*. 2021;157:107017. https://doi.org/10.1016/j.ympev.2020.107017
15. Cotten M, Watson SJ, Zumla AI, et al. Spread, circulation, and evolution of the Middle East Respiratory Syndrome Coronavirus. *mBio*. 2014;5(1):e01062-01013. https://doi.org/10.1128/mBio.01062-13
16. Zhang Z, Shen L, Gu X. Evolutionary dynamics of MERS-CoV: potential recombination, positive selection and transmission. *Sci Rep*. 2016;6:25049. https://doi.org/10.1038/srep25049
17. Pavlovic-Lazetic GM, Mitic NS, Tomovic AM, Pavlovic MD, Beljanski MV. SARS-CoV genome polymorphism: a bioinformatics study. *Genomics Insights*. 2005;3(1):18-35.
18. Motayo BO, Oluwasemowo OO, Olusola BA, et al. Evolution and genetic diversity of SARS-CoV-2 in Africa using whole genome sequences. *Int J Infect Dis*. 2021;103:282-287. https://doi.org/10.1016/j.ijid.2020.11.190
19. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018;34(23):4121-4123. https://doi.org/10.1093/bioinformatics/bty407
20. Pereson MJ, Mojsiejczuk L, Martinez AP, Flichman DM, Garcia GH, Di Lello FA. Phylogenetic analysis of SARS-CoV-2 in the first few months since its emergence. *J Med Virol*. 2021;93(3):1722-1731. https://doi.org/10.1002/jmv.26545
21. Pereson MJ, Flichman DM, Martinez AP, Bare P, Garcia GH, Di Lello FA. Evolutionary analysis of SARS-CoV-2 spike protein for its different clades. *J Med Virol*. 2021;93:3000-3006. https://doi.org/10.1002/jmv.26834
22. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772-780. https://doi.org/10.1093/molbev/mst010
23. Kalyaanamoorthy S, Bui Quang M, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate

phylogenetic estimates. *Nature Methods*. 2017;14(6):587-589. https://doi.org/10.1038/nmeth.4285

24. Baer CF, Miyamoto MM, Denver DR. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat Rev Genet*. 2007;8(8):619-631. https://doi.org/10.1038/nrg2158
25. Bouckaert R, Heled J, Kuehnert D, et al. BEAST 2: A Software Platform for Bayesian evolutionary analysis. *PLoS Comput Biol*. 2014; 10(4):e1003537. https://doi.org/10.1371/journal.pcbi.1003537
26. Farah S, Atkulwar A, Praharaj MR, Khan R, Gandham R, Baig M. Phylogenomics and phylodynamics of SARS-CoV-2 genomes retrieved from India. *Future Virol*. 2020;15(11):8. https://doi.org/10.2217/fvl-2020-0243
27. He W-T, Ji X, He W, et al. Genomic epidemiology, evolution, and transmission dynamics of porcine deltacoronavirus. *Mol Biol Evol*. 2020;37(9):2641-2654. https://doi.org/10.1093/molbev/msaa117
28. Nabil B, Sabrina B, Abdelhakim B. Transmission route and introduction of pandemic SARS-CoV-2 between China, Italy, and Spain. *J Med Virol*. 2021;93(1):564-568. https://doi.org/10.1002/jmv.26333
29. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol*. 2018;67(5):901-904. https://doi.org/10.1093/sysbio/syy032
30. Lam-Tung N, Schmidt HA, von Haeseler A, Bui Quang M. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015;32(1):268-274. https://doi.org/10.1093/molbev/msu300
31. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24(8):1586-1591. https://doi.org/10.1093/molbev/msm088
32. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol*. 2015;1(1):vev003. https://doi.org/10.1093/ve/vev003
33. 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(1):152-160. https://doi.org/10.1128/jvi.73.1.152-160.1999
34. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*. 2015;31(16):2745-2747. https://doi.org/10.1093/bioinformatics/btv195
35. Plante JA, Liu Y, Liu J, et al. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*. 2020;592:116-121. https://doi.org/10.1038/s41586-020-2895-3
36. Chan JF, Yuan S, Kok K-H, 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(10223):514-523. https://doi.org/10.1016/s0140-6736(20)30154-9
37. Li Q, Wu J, Nie J, et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell*. 2020;182(5): 1284-1294.e1289. https://doi.org/10.1016/j.cell.2020.07.012
38. Minskaia E, Hertzig T, Gorbalenya AE, et al. Discovery of an RNA virus 3 '-> 5 ' exoribonuclease that is critically involved in coronavirus RNA synthesis. *Proc Natl Acad Sci USA*. 2006;103(13): 5108-5113. https://doi.org/10.1073/pnas.0508200103
39. Jensen JD, Lynch M. Considering mutational meltdown as a potential SARS-CoV-2 treatment strategy. *Heredity*. 2020;124(5):619-620. https://doi.org/10.1038/s41437-020-0314-z
40. Lynch M, Burger R, Butcher D, Gabriel W. The mutational meltdown in asexual populations. *J Hered*. 1993;84(5):339-344. https://doi.org/10.1093/oxfordjournals.jhered.a111354
41. Motayo BO, Oluwasemowo OO, Akinduti PA. Evolutionary dynamics and geographic dispersal of beta coronaviruses in African bats. *PeerJ*. 2020;8:e10434. https://doi.org/10.7717/peerj.10434
42. Pyrc K, Dijkman R, Deng L, et al. Mosaic structure of human coronavirus NL63, one thousand years of evolution. *J Mol Biol*. 2006; 364(5):964-973. https://doi.org/10.1016/j.jmb.2006.09.074
43. McKinley ET, Jackwood MW, Hilt DA, et al. Attenuated live vaccine usage affects accurate measures of virus diversity and mutation rates in avian coronavirus infectious bronchitis virus. *Virus Res*. 2011;158(1-2):225-234. https://doi.org/10.1016/j.virusres.2011.04.006
44. Vijgen L, Keyaerts E, Moes E, et al. Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. *J Virol*. 2005;79(3): 1595-1604. https://doi.org/10.1128/jvi.79.3.1595-1604.2005
45. Tong YG, Shi WF, Liu D, et al. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature*. 2015;524(7563): 93-96. https://doi.org/10.1038/nature14490
46. Moratorio G, Costa-Mattioli M, Piovani R, Romero H, Musto H, Cristina J. Bayesian coalescent inference of hepatitis A virus populations: evolutionary rates and patterns. *J Gen Virol*. 2007;88: 3039-3042. https://doi.org/10.1099/vir.0.83038-0
47. Gray RR, Parker J, Lemey P, Salemi M, Katzourakis A, Pybus OG. The mode and tempo of hepatitis C virus evolution within and among hosts. *BMC Evol Biol*. 2011;11:131. https://doi.org/10.1186/1471-2148-11-131
48. Goni N, Fajardo A, Moratorio G, Colina R, Cristina J. Modeling gene sequences over time in 2009 H1N1 Influenza A Virus populations. *Virol J*. 2009;6:215. https://doi.org/10.1186/1743-422x-6-215
49. Patil JA, Cherian S, Walimbe AM, et al. Evolutionary dynamics of the American African genotype of dengue type 1 virus in India (1962-2005). *Infect Genet Evol*. 2011;11(6):1443-1448. https://doi.org/10.1016/j.meegid.2011.05.011
50. Hicks AL, Duffy S. Genus-specific substitution rate variability among picornaviruses. *J Virol*. 2011;85(15):7942-7947. https://doi.org/10.1128/jvi.02535-10
51. Furuse Y, Suzuki A, Oshitani H. Origin of measles virus: divergence from rinderpest virus between the 11(th) and 12(th) centuries. *Virol J*. 2010;7:52. https://doi.org/10.1186/1743-422x-7-52
52. Davis PL, Bourhy H, Holmes EC. The evolutionary history and dynamics of bat rabies virus. *Infect Genet Evol*. 2006;6(6):464-473. https://doi.org/10.1016/j.meegid.2006.02.007
53. Matthijnssens J, Heylen E, Zeller M, Rahman M, Lemey P, Van Ranst M. Phylodynamic analyses of rotavirus genotypes G9 and G12 underscore their potential for swift global spread. *Mol Biol Evol*. 2010;27(10):2431-2436. https://doi.org/10.1093/molbev/msq137
54. Mao Y, Bian D. Bioinformatics analysis of SARS-CoV-2 strain S protein mutation and its effect. *Chin J Virol*. 2020;36(06):44-51.
55. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*. 2017; 1(1):33-46. https://doi.org/10.1002/gch2.1018
56. Singh A, Steinkellner G, Kochl K, Gruber K, Gruber CC. Serine 477 plays a crucial role in the interaction of the SARS-CoV-2 spike protein with the human receptor ACE2. *Sci Rep*. 2021;11(1):4320-4320. https://doi.org/10.1038/s41598-021-83761-5
57. Thomson EC, Rosen LE, Shepherd JG, et al. Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. *Cell*. 2021;184(5):1171-1187.e20. https://doi.org/10.1016/j.cell.2021.01.037
58. van Dorp L, Richard D, Tan CCS, Shaw LP, Acman M, Balloux F. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat Commun*. 2020;11(1):5986. https://doi.org/10.1038/s41467-020-19818-2
59. Control CfD. US COVID-19 Cases Caused by Variants. 2021.
60. Feng Y, Chen Z, Meng Y, et al. Global early transmission of the new coronavirus variant strain VOC 202012/01 and analysis of the evolutionary characteristics of the spike protein. *Chin J Virol*. 2021: 1-7. https://doi.org/10.13242/j.cnki.bingduxuebao.003866

61. Rambaut A, Holmes EC, O'Toole A, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*. 2020;5(11):1403-1407. https://doi.org/10.1038/s41564-020-0770-5

62. Davies NG, Jarvis CI, Edmunds WJ, et al.Increased hazard of death in community-tested cases of SARS-CoV-2 Variant of Concern 202012/01. *medRxiv*. 2021. 1–31. https://doi.org/10.1101/2021.1102.1101.21250959. 10.1101/2021.02.01.21250959

63. Davies NG, Abbott S, Barnard RC, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*. 2021;372(6538):eabg3055. https://doi.org/10.1126/science.abg3055

64. Deng X, Garcia-Knight MA, Khalid MM, et al. Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant. *Cell*. 2021;184(13):3426-3437.e3428. https://doi.org/10.1016/j.cell.2021.04.025

65. Wang P, Casner RG, Nair MS, et al. Increased resistance of SARS-CoV-2 variant P.1 to antibody neutralization. *Cell Host Microbe*. 2021;29(5):747-751. https://doi.org/10.1016/j.chom.2021.04.007

66. Choudhary OP, Priyanka, Singh I, Rodriguez-Morales AJ. Second wave of COVID-19 in India: Dissection of the causes and lessons learnt. *Travel Med Infect Dis*. 2021;43:102126. https://doi.org/10.1016/j.tmaid.2021.102126

67. Quinonez E, Vahed M, Hashemi Shahraki A, Mirsaeidi M. Structural analysis of the novel variants of SARS-CoV-2 and Forecasting in North America. *Viruses*. 2021;13(5):930. https://doi.org/10.3390/v13050930

SUPPORTING INFORMATION

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

How to cite this article: Wang S, Xu X, Wei C, et al. Molecular evolutionary characteristics of SARS-CoV-2 emerging in the United States. *J Med Virol*. 2022;94:310-317.
https://doi.org/10.1002/jmv.27331