


RESEARCH ARTICLE

Establishing reference sequences for each clade of SARS-CoV-2 to provide a basis for virus variation and function research

Jian Yu^{1,2,3} | Shanshan Sun^{1,2} | Qianqian Tang^{1,2} | Chengzhuo Wang^{1,2} |
Liangchen Yu⁴ | Lulu Ren⁴ | Jun Li³ | Zhenhua Zhang^{1,2} 

¹Department of Infectious Diseases, The Second Hospital of Anhui Medical University, Hefei, China

²Institute of Clinical Virology, The Second Hospital of Anhui Medical University, Hefei, China

³Inflammation and Immune Mediated Diseases Laboratory of Anhui Province, Anhui Institute of Innovative Drugs, The School of Pharmacy, Anhui Medical University, Hefei, China

⁴The Second Clinical Medical School, Anhui Medical University, Hefei, China

Correspondence

Zhenhua Zhang, Department of Infectious Diseases, The Second Hospital of Anhui Medical University, Hefei, China.
Email: zzh1974cn@163.com

Jun Li, Inflammation and Immune Mediated Diseases Laboratory of Anhui Province, Anhui Institute of Innovative Drugs, School of Pharmacy, Anhui Medical University, Hefei, China.
Email: lj@ahmu.edu.cn

Funding information

Anhui Provincial Natural Science Foundation, Grant/Award Number: 1608085MH162

Abstract

Coronavirus disease 2019 (COVID-19) is a severe respiratory disease caused by the highly infectious severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). As the COVID-19 pandemic continues, mutations of SARS-CoV-2 accumulate. These mutations may not only make the virus spread faster, but also render current vaccines less effective. In this study, we established a reference sequence for each clade defined using the GISAID typing method. Homology analysis of each reference sequence confirmed a low mutation rate for SARS-CoV-2, with the latest clade GRY having the lowest homology with other clades (99.89%–99.93%), and the homology between other clade being greater than or equal to 99.95%. Variation analyses showed that the earliest genotypes S, V, and G had 2, 3, and 3 characterizing mutations in the genome respectively. The G-derived clades GR, GH, and GV had 5, 6, and 13 characterizing mutations in the genome respectively. A total of 28 characterizing mutations existed in the genome of the latest clades GRY. In addition, we found differences in the geographic distribution of different clades. G, GH, and GR are popular in the USA, while GV and GRY are common in the UK. Our work may facilitate the custom design of antiviral strategies depending on the molecular characteristics of SARS-CoV-2.

KEYWORDS

characterizing mutation, high-frequency mutation, reference sequence, SARS-CoV-2, variation analyses

1 | INTRODUCTION

Coronavirus disease 2019 (COVID-19) is a severe respiratory disease caused by the highly infectious severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a beta coronavirus with a nearly 30 kb positive-sense, single-strand RNA genome that encodes 29 proteins.¹ It has caused more than 120 million infections and more than 2.7 million deaths worldwide. Efficient prevention and control of

SARS-CoV-2 require a deeper understanding of different aspects of the virus, including sequence variations and evolution. To date, more than 800 000 SARS-CoV-2 sequences have been published, and it has been reported that the mutation rate of the viral genome was constantly increasing from 0.1338% on January 1, 2020 to 1.2373% on February 28, 2021.² In fact, more than 27 000 mutations have been identified so far. Mutations within the SARS-CoV-2 genome can affect biological consequences from an infection: mutations in

structural proteins targeted by vaccines may impair vaccine efficacy; mutations in nonstructural proteins may result in antiviral-resistant strains.³

In the early stage of the COVID-19 outbreak, some scientists were controversial over virus genotyping, thinking that it was difficult to prove the relationship between virus mutation and its function, and it was recommended not to over-interpret genome mutations during the pandemic.⁴ However, as the SARS-CoV-2 pandemic became prolonged, continuous accumulation of genomic data and in-depth study of the pathogenic and immune characteristics of the virus resulted in development of a few different methods to genotype and classify the virus. Current commonly used typing methods include the Chinese typing method, Pangolin typing method, GISAID typing method, and Nextstrain typing method.

In this study, we established a reference sequence for each clade based on the GISAID typing method and further analyzed characterizing mutations and frequent mutations for each clade. In addition, the evolution of characterizing mutations of all genotypes was analyzed in key regions (UK, South Africa, USA, India, and Brazil).

2 | METHODS

2.1 | Genomic sequence collection

SARS-CoV-2 genomic sequences were downloaded from GISAID (<http://www.gisaid.org>) with the restrictive conditions: 1. The virus only infected human hosts; 2. The sequence was submitted before February 28; 3. The sequence was in full length.

2.2 | Establishment of the reference sequence

Sequence analyses were performed by a previously reported method.⁵ Homology analysis and sequence alignment were conducted for downloaded sequences by using Primer 7.0 and Mega (7.0.14). The reference sequence was established by selecting the most common nucleotide in each position.

2.3 | Phylogenetic analysis

The ClustalW program of the MEGA software (7.0.14) was used to conduct multiple sequence alignment and the phylogenetic tree was constructed by using a maximum likelihood approach based on reference sequences.

2.4 | Analyses of nucleotide and amino acid sequence variation

Primer 7.0 was used to compare the reference nucleotide sequence to those of related human isolates and analyze the variation at different locations. Sequence comparison and variation analysis were also conducted at the amino acid level. The changes in the rate of mutations over time were conducted on GISAID.

3 | RESULTS

3.1 | Establishment of the reference sequence for each clade

With the ever-increasing number of SARS-CoV-2 sequences being uploaded onto GISAID, the GISAID team merged smaller evolutionary branches into main branches based on the shared marker strains and divided SARS-CoV-2 into 8 distinct clades, L, S, V, G, GH, GR, GV, and GRY, named by a collection of distinguishingly characterizing mutations. Here, we established reference sequences for all these clades to provide suitable reference standards for studies on the molecular biology and virology of SARS-CoV-2. To that end, we randomly downloaded 100 sequences for each clade from the GISAID website deposited during January 1, 2020, and January 31, 2021, established the reference sequence by selecting the most common nucleotide in each position. And the reference sequence of clade L is consistent with the reference sequence we established based on earlier sequences (accession number: EPI_ISL_412026).

TABLE 1 Comparison of homology between different clade of SARS-COV-2

Homology (%)	L	S	V	G	GH	GR	GV	GRY
L	100	99.99	99.99	99.99	99.98	99.98	99.96	99.91
S		100	99.98	99.98	99.97	99.97	99.95	99.90
V			100	99.98	99.97	99.97	99.95	99.90
G				100	99.99	99.99	99.97	99.92
GH					100	99.98	99.96	99.91
GR						100	99.96	99.93
GV							100	99.89
GRY								100

Abbreviation: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

TABLE 2 Characterizing mutations at nucleotide level of SARS-CoV-2 based on reference sequences

Region	Position	L	S	V	G	GH	GR	GV	GRY	
5'UTR	204	G						T		
	241	C			T	T	T	T	T	
1a	445	T						C		
	913	C							T	
	1059	C				T				
	3037	C			T	T	T	T	T	
	3267	C							T	
	5388	C							A	
	5986	C							T	
	6286	C						T		
	6954	T							C	
	8782	C		T						
	11083	G			T					
	11288-11296	TCTGGTTTT							del	
	1b	14408	C			T	T	T	T	T
		14676	C							T
14805		C		T						
15279		C							T	
16176		T							C	
21255		G						C		
S	21766-21771	ACATGT							del	
	21994-21996	TTA							del	
	22227	C						T		
	23063	A							T	
	23271	C							A	
	23403	A			G	G	G	G	G	
	23604	C							A	
	23709	C							T	
	24506	T							G	
	24914	G							C	
3a	25563	G				T				
	26144	G		T						
M	26801	C					G			
8	27944	C						T		
	27972	C							T	
	28048	G							T	
	28111	A							G	

TABLE 2 (Continued)

Region	Position	L	S	V	G	GH	GR	GV	GRY
	28144	T	C						
N	28274	A							del
	28280-28282	GAT							CTA
	28881-28883	GGG					AAC		AAC
	28932	C						T	
	28977	C							T
10	29645	G						T	

Abbreviation: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

3.2 | Phylogenetic and homology analyses

The MEGA6.0 software was used to construct a phylogenetic tree with the established reference sequences (Figure S2). Here, the phylogenetic tree simply shows the magnitude of differences between each clade, does not represent the evolutionary relationship between each clade. Then we analyzed the homology between genotypes (Table 1), and the results showed that the homology between genotypes was high. The latest clade GRY had the lowest homology with other clades (99.89%–99.93%), and the homology between other clades was greater than or equal to 99.95%.

3.3 | Variation analyses at the nucleotide and amino acids levels

By comparing to the established reference sequences, we identified characterizing mutations for each clade at the nucleotide and amino acid levels (Tables 2 and 3). Two characterizing mutations at the nucleotide level (C8782T and T28144C) and one characterizing mutation at the amino acid level (ORF8_L84S) were identified for clade S; three characterizing mutations at the nucleotide level (G11083T, C14805T, and G26144T) and two characterizing mutations at the amino acid level (NSP6_L37F and ORF3a_G251V) were found in clade V; Clade G and its four derivative GH, GR, GV and GRY had four characterizing mutations at the nucleotide level (C241T, C3037T, C14408T, and A23403G) and two characterizing mutations at the amino acid level (NSP12_P323L and S_D614G). In addition to mutations mentioned above, other characterizing mutations were also observed in certain clades: two mutations at the nucleotide level (C1059T and G25563T) and two mutations at the amino acid level (NSP2_T85I and ORF3a_Q57H) were found in clade GH; three mutations at the nucleotide level (GGG28881AAC) and two mutations at the amino acid level (N_R203K and N_G204R) in clade GR; nine mutations at the nucleotide level (G204T, T445C, C6286T, G21255C,

TABLE 3 Characterizing mutations at the amino acid level of SARS-CoV-2 based on reference sequences

Region	position	L	S	V	G	GH	GR	GV	GRY
1a	265 (NSP2_85)	T				I			
	1001 (NSP3_183)	T							I
	1708 (NSP3_890)	A							D
	2230 (NSP3_1412I)	I							T
	3606 (NSP6_37)	L		F					
	3675-3677 (NSP6_106-108)	SGF							del
1b	314	P			L	L	L	L	L
S	69	H							del
	70	V							del
	144	Y							del
	222	A						V	
	501	N							Y
	570	A							D
	614	D			G	G	G	G	G
	681	P							H
	716	T							I
	982	S							A
1118	D							H	
3a	57	Q				H			
	251	G		V					
8	27	Q							stop
	84	L	S						
N	3	D							L
	203	R					K		K
	204	G					R		R
	220	A						V	
	235	S							F
10	30	V						L	

Abbreviation: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

C22227T, C26801G, C27944T, C28932T, and G29645T) and three mutations at the amino acid level (S_A222V, N_A220V, and ORF10_V30L) in clade GV. On top of all mutations observed in clade GR, clade GRY had 23 more mutations at the nucleotide level (C913T, C3267T, C5388A, C5986T, T6954C, TCTGGTTTT11288-11296-del, C14676T, C15279T, T16176C, ACATGT21766-21771del, TTA21994-21996del, A23063T, C23271A, C23604A, C23709T, T24506G, G24914C, C27972T, G28048T, A28111G, A28274del, GAT28280CTA, and C28977T) and 16 more mutations at the amino acid level (NSP3_T183I, NSP3_A890D, NSP3_I1412T,

NSP6_SGF1-6-108del, S_H69del, S_V70del, S_Y144del, S_N501Y, S_A570D, S_P681H, S_T706I, S_S982A, S_D1118L, ORF8_Q27stop, N_D3L, and N_S235F).

We further analyzed the changes in the rate of characterizing mutations in each clade over time (Figure 1), and the results showed that almost all mutations were kept over time except for NSP2_T85I in clade GH. These analyses also demonstrated that NS8_L84S was specific for clade S; NSP6_L37F and ORF8_G251V were specific for clade V; NSP2_T85I and ORF3a_Q57H were specific for clade GH; N_A220V, S_A222V, and ORF10_V30L were specific for clade GV; Clade GRY had 16 specific mutations, including NSP3_T183I, NSP3_A890D, NSP3_I1412T, NSP6_SDF106-108del, S_H69del, S_V70del, S_Y144del, S_N501Y, S_A570D, S_P681H, S_T706I, S_S982A, S_D1118H, ORF8_Q27stop, N_D3L, and N_S235F; No unique mutation was found in clades G and GR.

We then analyzed high-frequency mutations (the mutation rate is between 20% and 50%) in each clade, which could potentially become new characterizing mutations with the spread and evolution of the virus. By comparing 100 sequences from each clade to the corresponding reference sequence established, we identified high-frequency mutations for different clades (Figure 2A). Clade S had three high-frequency mutations at the amino acid level: NSP13_P504L (30 in 100), NSP13_Y541C (33 in 100), and N_S202N (26 in 100); clade V had two high-frequency mutations at the amino acid level: NSP2_I559V (35 in 100) and NSP2_P585S (27 in 100); clade G had one high-frequency mutation N_S194L (23 in 100) at the amino acid level; clade GH has three high-frequency mutations at the amino acid level: NSP5_L89F (24 in 100), ORF8_S24L (28 in 100) and N_P199L (20 in 100); Clade GR had no high-frequency mutations at the amino acid level; Clade GV had one high-frequency mutation S_L18F (30 in 100) at the amino acid level; Clade GRY had one high-frequency mutation NSP13_K460R (42 in 100) at the amino acid level. Then we analyzed the mutation frequencies of these sites on GISIAD (Figure 2B), and the mutation rates were basically consistent with those calculated with the 100 sequences downloaded. Finally, we analyzed the changes of these mutations over time (Figure 2C-H). It is worth noting that the mutation rate of N_S202 was close to 100% after September 2020, which can be considered as a characterizing mutation of clade S; and the frequency of high-frequency mutations of GH showed an upward trend over time, with the mutation frequency reaching about 50% for both by January 2021.

3.4 | Characterizing mutations in key regions

Finally, we analyzed the occurrence of characterizing mutations in key regions (Figure 3). Mutation NSP6_L37F (V-specific mutation) was prevalent in the UK and USA in the early days, then appeared in India, but is currently about to disappear; ORF3a_G251V is another specific mutation for V, but it was only prevalent in the early months in the UK and USA. These observations suggested that although both NSP6_L37F and ORF3a_G251V were characterizing mutations of clade V, they did not necessarily occur at the same time.

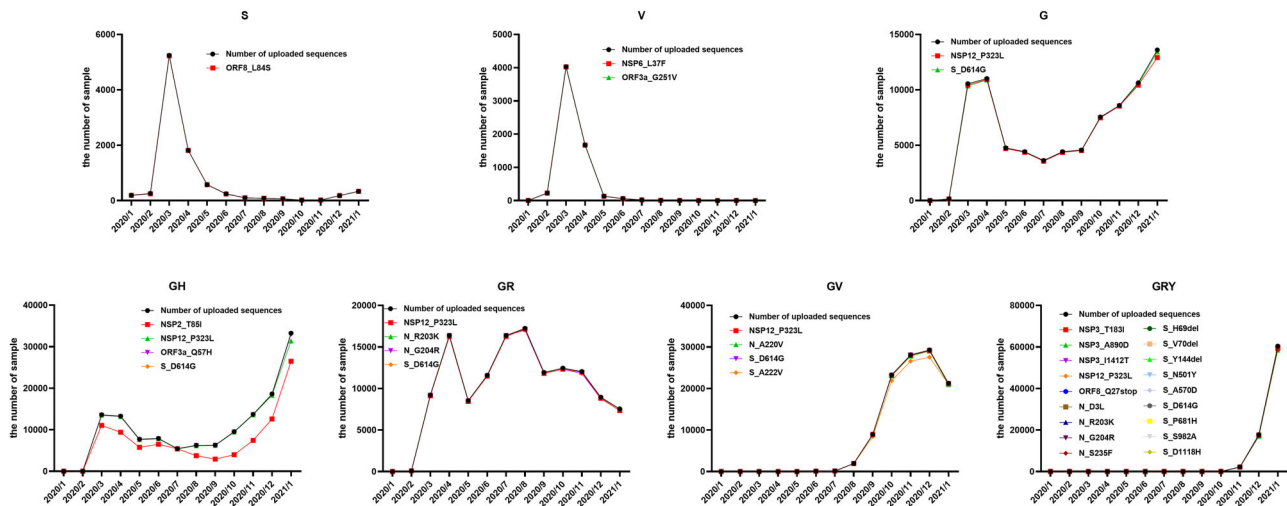


FIGURE 1 The changes in the rate of characterizing mutations in each clade over time

The mutation rate of NSP12_P323L and S_D614G in each key region from June 2020 is close to 100%, demonstrating the global prevalence of clade G and its offspring. NSP2_T85I and ORF3a_Q57 are GH-specific mutations that first appeared in the USA, then occurred simultaneously in several key regions except for India. Analysis for mutations N_A220V and S_A222V (GV-specific mutations) found that clade GV mainly occurred in the UK. Currently, GRY-specific mutations exist simultaneously, but we noted that NSP3_T183I was the earliest popular mutation that had occurred in India and then became popular in the UK; The calculated mutation rate of S_144del in January was as high as 50%, but it was mostly due to the small number of early uploaded sequences from India (The number of sequence in January is 2). G, GH, and GR are currently mainly popular in the USA, and GV and GRY are currently mainly common in the UK (Figure S3).

4 | DISCUSSION

Previously we have established the reference sequence for SARS-CoV-2 from 95 full-length genomic sequences of SARS-CoV-2 strains retrieved from the National Center for Biotechnology Information and GISAID databases up to February 14, 2020.⁵ Although the mutation rate of SARS-CoV-2 is extremely low, as the epidemic continues, different genotypes of SARS-CoV-2 have evolved, and scientists have found that mutations of SARS-CoV-2 have a profound impact on the pathogenic and immune characteristics of the virus. In this study, we systematically established the reference sequence for each viral clade, analyzed and compared the genetic mutations in the SARS-CoV-2 genome, and further analyzed the evolution of characterizing mutations of all genotypes in key regions (UK, South Africa, USA, India, and Brazil). It is noteworthy that the reference sequence of clade L is consistent with the reference sequence we established based on earlier sequences (accession number: EPI_ISL_412026), confirming that clade L be the earliest type of virus. This study also

confirmed a low mutation rate for SARS-CoV-2, with homology between different genotypes being between 99.89% and 99.99%, which is much higher than other viruses. For example, the homology between different genotypes of HBV is less than 92%.⁶ However, our study found that the mutations of SARS-CoV-2 are constantly increasing. The earliest genotypes S, V, and G have 2, 3, and 3 mutations in the genome respectively. G-derived clades GR, GH, and GV have 5, 6, and 13 mutations in the genome respectively.⁷ The latest clades GRY have 28 mutations in the genome. These characterizing mutations can be used as targets for primer design to distinguish different genotypes. As the number of mutations continues to increase, these clades can be further characterized by including the most recent mutations and will likely be split even further in the future. We also analyzed high-frequency mutations in different clades and found that GH contained the highest number of high-frequency mutations at the amino acid level. Based on our homology alignment and mutation analyses, we speculate that clade L, S, and V belong to a virus group, clade G, GH, GR, and GV belong to another virus group, and GRY is the latest virus group evolved from GR.

SARS-CoV-2 belongs to the coronavirus family and is a positive-sense virus without segmentation. It contains two large overlapping open reading frames (ORF1a and ORF1b) and encodes four structural proteins, including spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins, and nine accessory proteins.⁸ The ORF1ab is closely related to viral replication. The S protein is responsible for the binding of the virus to the host cell membrane receptor and membrane fusion,⁹ and a critical site for host neutralizing antibodies and a key target for vaccine design. The N protein is the central component of virions. It binds to viral genomic RNA to package the RNA into a ribonucleoprotein (RNP) complex and is capable of inducing both humoral and cellular immune responses after infection.¹⁰ Characterizing mutations and high-frequency mutations occur mainly in regions within ORF1ab, S, and N. ORF1ab and N genes are often selected as the genes targeted for primer design,¹¹ we must pay attention to avoid these mutations

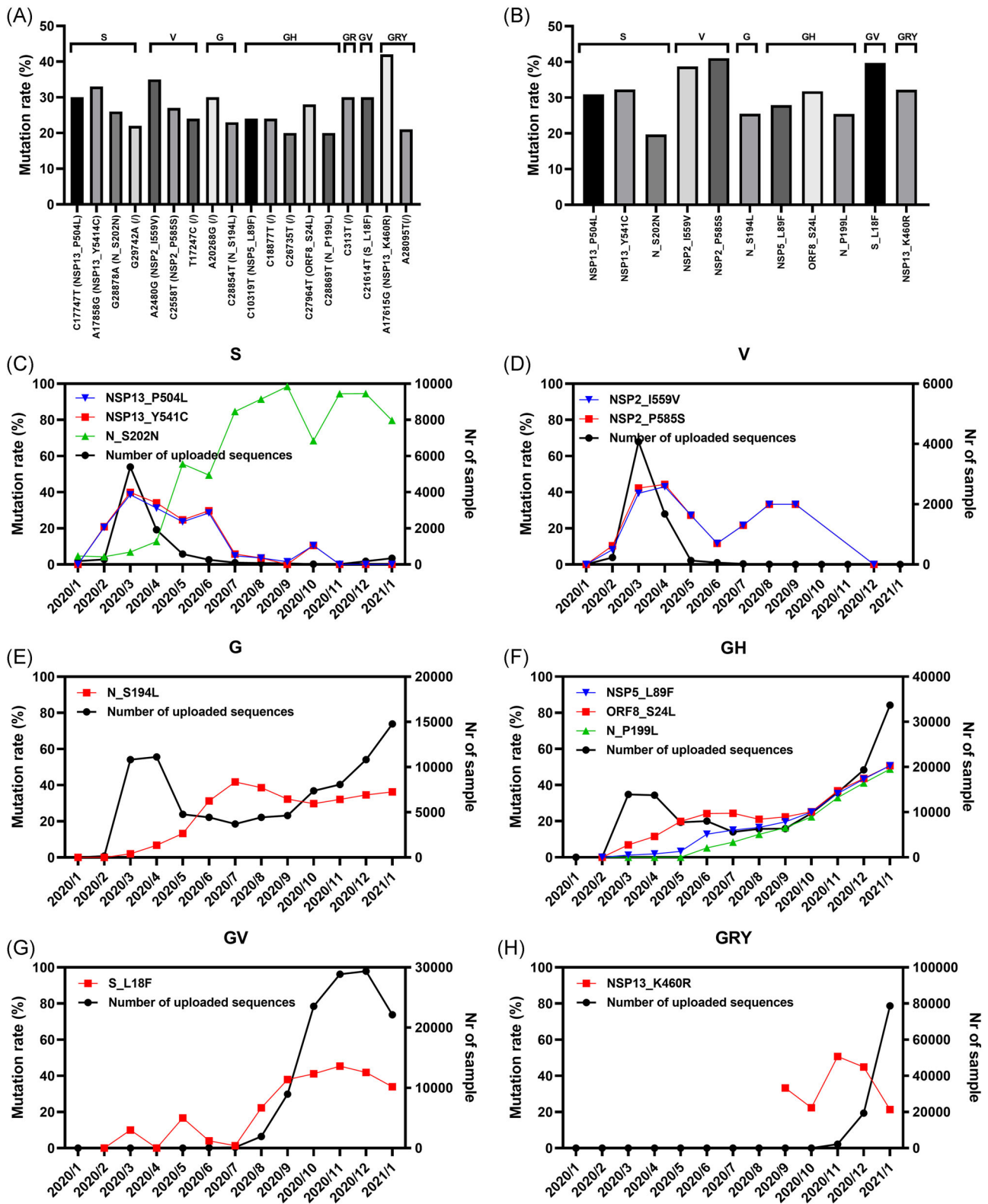


FIGURE 2 High-frequency mutations (the mutation rate is between 20% and 50%) of each clade. (A) High-frequency mutations through comparison of 100 downloaded sequences with corresponding reference sequences; (B) The mutation rate of high-frequency mutations were conducted on GISAID; (C)–(H) represent the changes in the rate of high-frequency mutations over time in clade S, V, G, GH, GV, and GRY, respectively

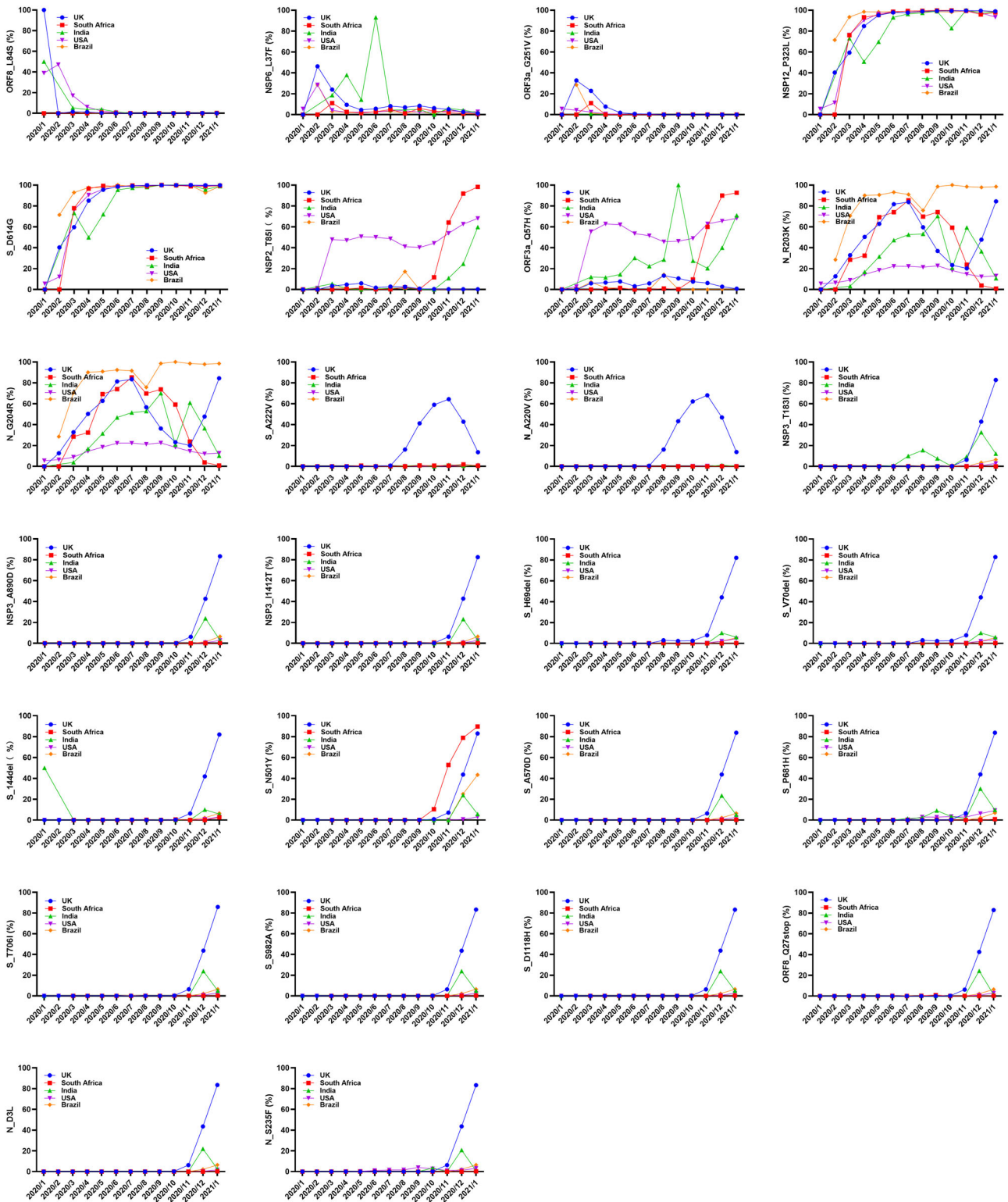


FIGURE 3 The changes in the rate of characterizing mutations in key regions over time

when designing primers. Mutations can result in the change of structure, function, and pathogenicity of the SARS-CoV-2. Analyses of these mutations can help us predict the trend of virus mutation and changes in virulence and infectivity in advance, providing a basis for vaccine development. For example, Shinya's study showed

that the NS8_L84S variant exhibited greater solubility than wild-type under acidic conditions¹²; this mutation may be to allow the virus to survive in a harsher environment. The study showed that the S_D614G variant increases SARS-CoV-2 transmission and replication,¹³ and Clade G and its offspring containing this

substitution have become the predominant circulating clades after March 2021. Clade GRY adds the S_N501Y mutation on the basis of the S_D614G mutation, a series of studies have shown that the N501Y mutation strengthens its binding to human receptor angiotensin-converting enzyme 2 (ACE2) and further enhances the ability of the virus to enter host cells.^{14–16} Therefore, GRY has gradually become the main epidemic Clade from 2021.

In general, results from this study will facilitate viral detection, functional analysis, vaccine design, epidemic investigation, and evaluation of drug efficacy, among others.

5 | CONCLUSION

In this study, we established a reference sequence for each clade classified using the GISAID typing method. By comparing with the established reference sequences, we found that the earliest genotypes S, V, and G have 2, 3, and 3 characterizing mutations in the genome respectively. G-derived clades GR, GH, and GV have 5, 6, and 13 characterizing mutations respectively. The latest clade GRY has 28 characterizing mutations in the genome. GH contains the highest number of high-frequency mutations at the nucleotide and amino acid levels and the number of high-frequency mutations of GH has shown an upward trend over time. This study provides suitable reference standards for studies on the molecular biology and virology of SARS-CoV-2, and our study may facilitate custom-designed antiviral strategies.

ACKNOWLEDGEMENT

This study was supported by the Anhui Provincial Natural Science Foundation of China (grant number: 1608085MH162).

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

AUTHOR CONTRIBUTIONS

Jian Yu and Shanshan Sun contributed equally to this study. Jian Yu and Shanshan Sun analyzed the data and wrote the manuscript; Qianqian Tang, Chengzhuo Wang, and Liangchen Yu were responsible for collecting, collating, and checking data; Lulu Ren contributed to drawing; Zhenhua Zhang and Jun Li conceptualized and designed the study and critically revised the manuscript.

DATA AVAILABILITY STATEMENT

Data for this study can be accessed through <http://www.gisaid.org>. The information of data is listed at Table S2.

ORCID

Zhenhua Zhang  <http://orcid.org/0000-0002-8480-9004>

REFERENCES

1. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265-269.
2. Gong Z, Zhu JW, Li CP, et al. An online coronavirus analysis platform from the National Genomics Data Center. *Zool Res*. 2020;41(6):705-708.
3. Kadam SB, Sukhrmani GS, Bishnoi P, Pable AA, Barvkar VT. SARS-CoV-2, the pandemic coronavirus: Molecular and structural insights. *J Basic Microbiol*. 2021;61(3):180-202.
4. MacLean OA, Orton RJ, Singer JB, Robertson DL. No evidence for distinct types in the evolution of SARS-CoV-2. *Virus Evol*. 2020;6:034. doi:10.1093/ve/veaa034
5. 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.
6. Cai Q, Zhu H, Zhang Y, Li X, Zhang Z. Hepatitis B virus genotype A: design of reference sequences for sub-genotypes. *Virus Genes*. 2016;52(3):325-333.
7. Mercatelli D, Giorgi FM. Geographic and genomic distribution of SARS-CoV-2 mutations. *Front Microbiol*. 2020;11:1800.
8. Weiss SR, Navas-Martin S. Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus. *Microbiol Mol Biol Rev*. 2005;69(4):635-664.
9. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*. 2020;181(2):281-292.
10. Peng Y, Du N, Lei Y, et al. Structures of the SARS-CoV-2 nucleocapsid and their perspectives for drug design. *EMBO J*. 2020;39(20):e105938.
11. Corman VM, Landt O, Kaiser M, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill*. 2020;25(3):23-30.
12. Ohki S, Imamura T, Higashimura Y, Matsumoto K, Mori M. Similarities and differences in the conformational stability and reversibility of ORF8, an accessory protein of SARS-CoV-2, and its L84S variant. *Biochem Biophys Res Commun*. 2021;563:92-97.
13. Zhou B, Thao TTN, Hoffmann D, et al. SARS-CoV-2 spike D614G change enhances replication and transmission. *Nature*. 2021;592(7852):122-127.
14. Ali F, Kasry A, Amin M. The new SARS-CoV-2 strain shows a stronger binding affinity to ACE2 due to N501Y mutant. *Med Drug Discov*. 2021;10:100086.
15. Tian F, Tong B, Sun L, et al. N501Y mutation of spike protein in SARS-CoV-2 strengthens its binding to receptor ACE2. *eLife*. 2021;10:10.
16. Khan A, Zia T, Suleman M, et al. Higher infectivity of the SARS-CoV-2 new variants is associated with K417N/T, E484K, and N501Y mutants: an insight from structural data. *J Cell Physiol*. 2021;236(10):7045-7057.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Yu J, Sun S, Tang Q, et al. Establishing reference sequences for each clade of SARS-CoV-2 to provide a basis for virus variation and function research. *J Med Virol*. 2022;94:1494-1501. doi:10.1002/jmv.27476