Heliyon 9 (2023) e18980

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

SARS-CoV-2 genomic characterization and evolution in China

Peng Zhang [a,1], Dongzi Liu [b,1], Lei Ji [a], Fenfen Dong [a,*]

[a] Huzhou Center for Disease Control and Prevention, 999 Changxing Road, Huzhou, Zhejiang, 313000, China
[b] State Key Laboratory of Virology, College of Life Sciences, Wuhan University, Wuhan, 430072, China

ARTICLE INFO

Keywords:
SARS-CoV-2
Evolution
Genomic characterization
China
Nonpharmaceutical intervention

ABSTRACT

The pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) affected global health worldwide due to its high contagiousness. During the viral spread, many mutations occurred within the virus genome. China has adopted nonpharmaceutical intervention (NPI) to contain COVID-19 outbreaks. In order to understand the evolution and genomic variation of SARS-CoV-2 in China under this policy, a total of 524 sequences downloaded from Global Initiative on Sharing All Influenza Data (GISAID) between 2019 and 2022 were included in this study. The time-scaled evolutionary analysis showed that these sequences clustered in three groups (Group A-C). Group B and C accounted for the majority of the sequences whose divergence times were around 2020 and distributed in multiple regions. Group A was mainly composed of G variants, which were mainly isolated from several regions. Moreover, we found that 191 sites had mutations with no less than 3 times, including 30 amino acids that were deleted. Finally, we found that spike and nucleocapsid genes underwent positive selection evolution, indicating that the mutations within spike and nucleocapsid genes increased the SARS-CoV-2 contagiousness. Hence, this study preliminarily elucidates the evolutionary characteristics and genomic mutations of SARS-CoV-2 under the implementation of the NPI policy in China, providing scientific basis for further understanding the control effect of the NPI policy on the epidemic.

1. Introduction

A new type of infectious respiratory syndrome was first discovered in Wuhan on December 2019, and it was later confirmed that severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was the etiological agent [1]. The virus can be transmitted from person to person through aerosols or small respiratory droplets [2]. Since its first detection, the virus spread worldwide. According to the World Health Organization, the most common symptoms showed by patients affected by coronavirus disease of 2019 (COVID-19) are mostly fever, cough, tiredness, and loss of taste or smell. Less commonly symptoms are sore throat, headache, general aches and chest pain. Difficulty in breathing or shortness of breath as well as loss of speech or mobility, and/or confusion can be found in the most severe cases (https://www.who.int/health-topics/coronavirus#tab=tab_3). Through the 2020, the evolution rate of SARS-COV-2 was relatively slow, with a high effectiveness of the vaccine [3]. However, by the end of 2020, the SARS-CoV-2 variant Alpha which carried 23 nucleotide mutations with 17 amino acids changes was identified for the first time in the UK [4]. Later, the variants Beta, Gamma, Delta, and Omicron emersed, accompanied by a higher mutations rate. To this regard, the evolution of the virus in immunocompromised patients was considered as an important factor in terms of occurrence of many mutations in the virus genome [5].

* Corresponding author.
E-mail address: dongfenf@mail.ustc.edu.cn (F. Dong).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2023.e18980
Received 27 January 2023; Received in revised form 25 July 2023; Accepted 3 August 2023
Available online 12 August 2023

SARS-CoV-2 is a single stranded RNA virus, and it belongs to the same genus beta coronavirus as SARS-CoV, sharing both a size of the genome of 30 kb, including many open reading frames (ORFs) [1]. The size of the first ORF (ORF1ab) accounts for about 2/3 of the whole genome length, and it encodes 16 non-structural proteins (nsps) involved in many processes after translation splicing [6]. For example, Nsp1 is responsible for the expression of the viral genome by inhibiting translation of the host cells proteins [7]. Nsp2 binds to the nucleic acids to regulate important intracellular signal pathways [8]. The complex composed of Nsp12-Nsp7-Nsp8 mediates the viral genome synthesis [9]. The remaining nsps genes encode accessory and structural proteins. On the other side, structural proteins are encoded by spike (S), membrane (M), nucleocapsid (N) and envelope (E)genes [6]. The S protein wraps on the surface of the virus and mediates the binding to the angiotensin-converting enzyme 2 (ACE2) receptor on the host cell surface [10]. The M protein plays a role in the formation of the viral envelope, promoting the genome assembly as well [11]. The N protein binds to the RNA genome and it is involved not only in the viral assembly, but also it antagonizes the host immunity [11]. The E protein is involved in the viral assembly and maturation through interaction with host cell membrane proteins [12]. Accessory proteins are encoded by ORF3a, ORF3b, ORF6, ORF7a, ORF7b, ORF8b, ORF9b, ORF9c and ORF10, and they play an important role in helping the virus in evasion of the immune system and enhancement of virulence by interacting with host cells [13].

According to Global Initiative on Sharing All Influenza Data (GISAID) nomenclature system, most of the sequenced SARS-CoV-2 genomes were classified into major clades including S, O, L, V, G, GH, GR, GV, and GRY. These clades were bases on shared marker mutations. These mutations include: NS8-L84S for clade S; coexisting NSP6-L37F and NS3-G251V for clade V; S-D614G for clade G. Clades GK, GH, GR, GV not only have S-D614G, but have S-T478K, NS3-Q57H, N-G204R and S-A222V respectively. S–H69del, S–V70del, S–Y144del, S–N501Y, S-D614G, N-G204R characterize the clade GRY. Sequences that don't belong to any of these major clades were grouped into clade O. The S and L clades were around at the beginning of the outbreak. L split into V and G. GR, GH, GV, and GK are the descendants of G. GR evolved into GRY and later GRA.

After the first detection in China in 2019, nonpharmaceutical intervention (NPI) to contain COVID-19 outbreaks were put in place [14]. However, the impact of this NPI on the spread and evolution of the virus in China was not yet known. So far, according to the GISAID (https://gisaid.org/), viruses from China were classified in 10 clades. However, still little information is available on SARS-CoV-2 mutations, evolution and selectively pressure in China. Therefore, in this study, we used sequences downloaded from GISAID to (i) collect the information of clinical cases; (ii) perform Bayesian Evolutionary Analysis Sampling Trees (BEAST) to investigate the evolution time of the virus; (iii) analyze the virus genome mutation characteristics; (iv) analyze the selection pressure of virus genes. We aimed to unravel the evolutionary characteristics and variations in SARS-CoV-2 during these years when China implemented NPI.

2. Materials and methods

2.1. Collection of SARS-CoV-2 whole-genome sequences

We retrieved 524 nucleotide sequences GISAID (www.gisaid.org) using the following parameters: virus name: hCov-19; host: human; collection date: from 2019 to 12-01 to 2022-10-13; location: China; complete genome; high coverage. The metadata include collection date, location, gender, patient age, lineage and clade.

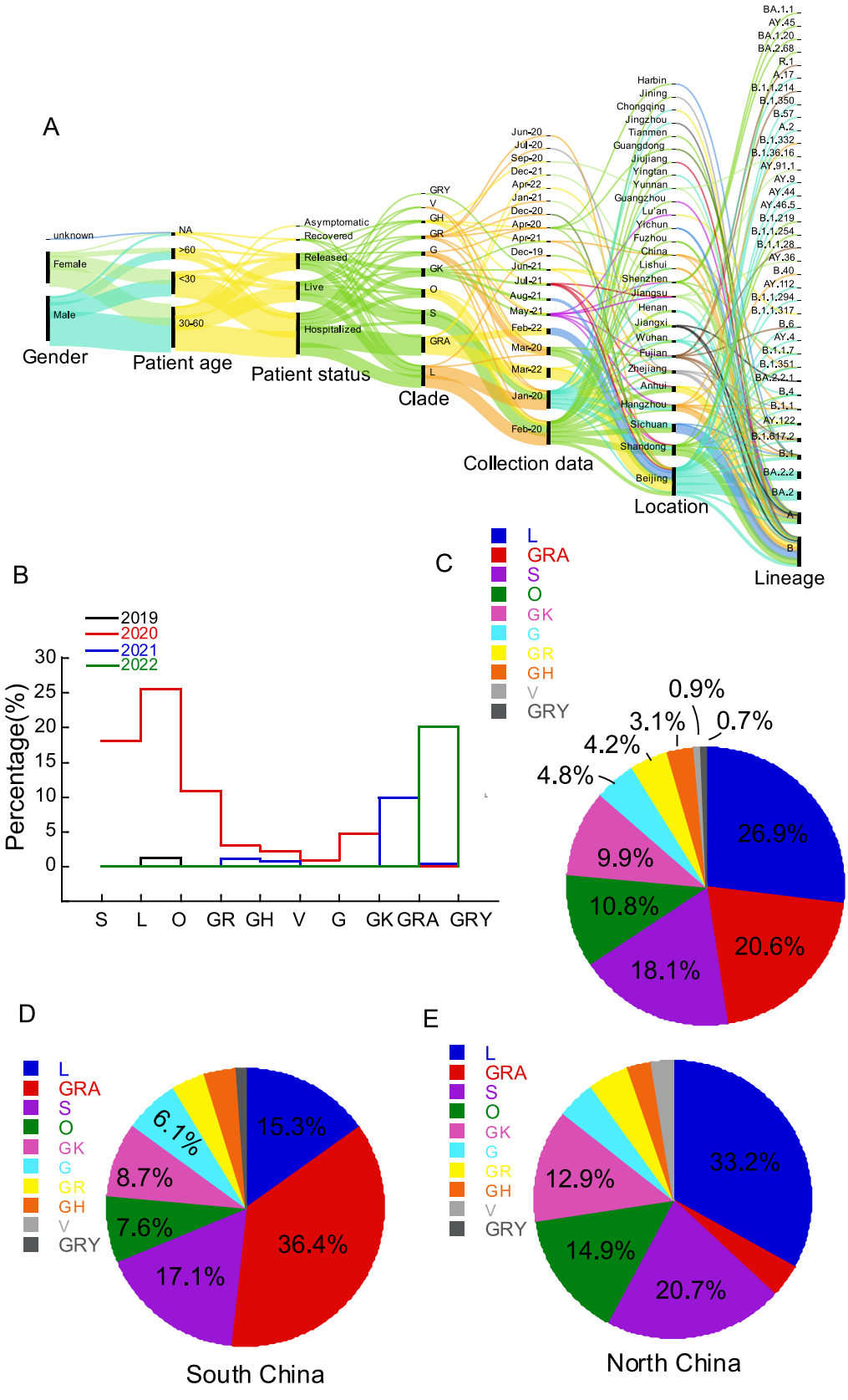
2.2. Genomic analysis

All the sequences were mapped to the reference SARS-CoV-2 genome (acc. No. NC_045512.2) using the online software Coronavirus Typing Tool (Version 1.25) [15]. The amino acids changes were obtained in this tool as well by using a novel dynamic aligner to allow tracking new viral mutations [16]. The amino acid mutation frequency including deletions were calculated artificially. The mutation frequency was determined by dividing the number of mutations of each site in all samples by the total number of samples. Graphs and plots were generated using Kaleida Graph 4.5 (Synergy Software, Reading, PA, USA). Multi-categorical alluvial diagrams were generated using RawGraphs 2.0 [17].

2.3. Reconstruction of time-scaled phylogenies

Sequence alignments were performed using Multiple Alignment using Fast Fourier Transform (MAFFT) [18]. A total of 365 sequences were obtained after removing low-quality sequences and UTR regions and further analyzed. To generate the files required by the Bayesian analysis, the sequences were used as input into the Bayesian Evolutionary Analysis Utility (BEAUti) v1.10.0 tool [19] and the tip dates with the corresponding collection data were downloaded from GISAID as mentioned above. GTR (generalized time reversible) +F + G4 was the fit model used accordingly with PhyloSuite v1.2.2 [20], followed by relaxed clock log normal and coalescent exponential growth trees previously selected. Finally, a chain length of one hundred million every ten thousand iterations was selected for further analyses. Data formatted in XML files were generated and used as input into BEAST 1.10.0. Runs were repeated twice, and the results were combined using the LogCombiner 1.10.0.

Data were examined by Tracer (v1.7.2) [21] with a 10% burn in. Effective sampling size (ESS) of all parameters was greater than 200. The final MCC tree was generated by TreeAnnotator (v1.10.0) after a 10% burn in. The tree was viewed by FigTree (v1.4.4) and each lineage and location were highlighted in distinct colors.



(caption on next page)

Fig. 1. Characteristics of the research objectives.
(A) The age, sex, virus strain, and sampling time of the patients enrolled are represented in a multi-categorical alluvial diagram. **(B)** Plot of the trends of variants from 2019 to 2022. (C) Pie chart of the proportion of each virus clades. The percentage has been marked on the pie chart. (D) Pie chart of the proportion of each virus clades in South China. The percentage has been marked on the pie chart. (E) Pie chart of the proportion of each virus clades in North China. The percentage has been marked on the pie chart.

2.4. Nucleotide substitution rates calculation

Ka and Ks of distinct genes from 365 sequences belonged to each lineage were calculated by MEGA-11 (model, Nei-Gojobori). Ka/Ks values were calculated using Excel 2019.

3. Results

3.1. Research aims and characteristics

Metadata from a total of 524 COVID-19 positive samples selected for this study were collected. In these cases, 29% of them were younger than 30 years old (yo), 13% were older than 60 yo, and 53% were between 30 and 60 yo. The patient status mainly includes released, live and hospitalized. More than half of the patients were hospitalized. The patients came from 26 different regions, with the most coming from Beijing, followed by Shandong and Sichuan (Fig. 1A). In the first two months of 2020, infected patients were distributed in many provinces and regions. After that, patients in each month were mainly distributed in a few provinces and regions (Fig. 1A).

In January–February 2020, clades L and S were predominant, followed by clade GRA two years later in China (Fig. 1A). More specifically, in 2019, all COVID-19 cases belonged to clade L, and clades S, O, GR, GH, V, and G appeared in the following year, 2020. By 2021, COVID-19 evolved to GK, GRA, and GRY clades, with GK accounting for the most detections. In 2022, the virus kept evolving, and all patients were infected with clade GRA in China (Fig. 1A and B). Among the clades mentioned, clade L (26.9%) was the most predominant, followed by GRA (20.6%), S (18.1%), O (10.8%), GK (9.9%), G (4.8%), GR (4.2%), GH (3.1%), V (0.9%), and GRY (0.7%) (Fig. 1C). In terms of geographical distribution, clades G, GH and GRA were mainly distributed in northern China, while clades L, O, S and V were mostly distributed in southern China (Fig. 1D and E).

3.2. Phylogenetic tree identifying an evolutionary pattern

To estimate the evolutionary time of each clade, we performed a Bayesian phylogenetic tree. A total of 365 out of 524 sequences were selected after alignment, quality check and UTR regions removal. BEAST analysis further investigated the evolutionary relationship within sequences considering the sampling time and location. The time-scaled maximum clade credibility (MCC) tree of 365 sequences was shown in Fig. 2. The tree represented ten GISAID clades (L, S, O, V, G, GK, GR, GH, GRY, and GRA). Group B and C were

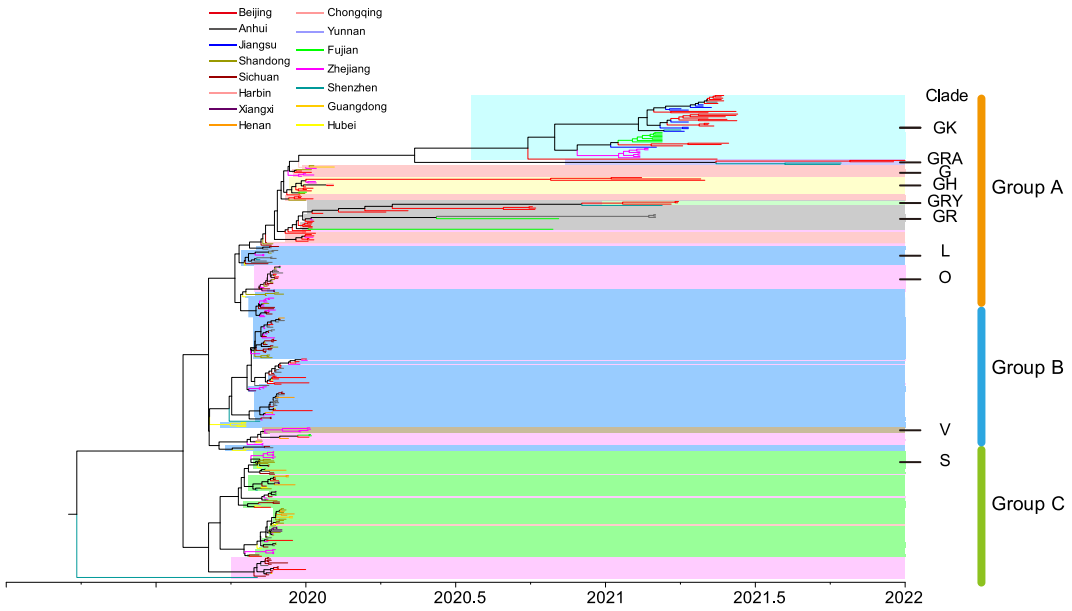
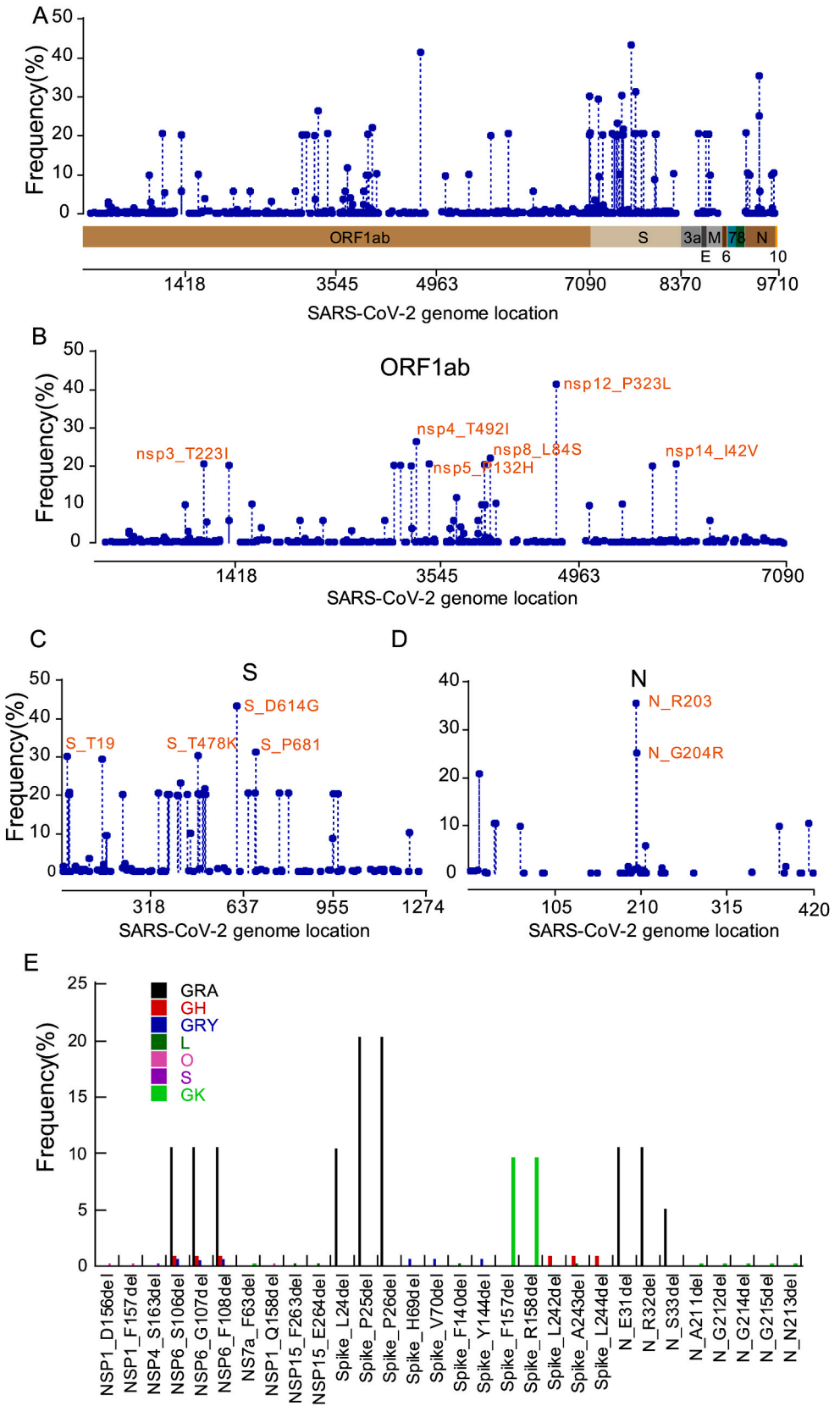


Fig. 2. Time-scaled phylogenetic analysis to identify the evolutionary relationship of each clade. The colors of the branches represent different regions. The colors after branches represent different clades. The tree was divided into three groups, namely group A, B, and C.



(caption on next page)

Fig. 3. The characteristics of SARS-CoV-2 genome mutation.
Dot plot of the amino acid mutation frequency in the whole genome **(A)**, ORF1ab **(B)**
S protein **(C)**, N protein **(D)**. Y-coordinates indicate mutation frequencies. X-coordinates indicate the SARS-CoV-2 genome location. **(E)** Histogram of the frequency of deletions at different sites of each viral clade.

composed of clades L, S, O and V, which were the major clades in the early stage of the outbreak. Group A was mainly composed of clades GK, GRA, G, GH, GRY and GR, which were evolved on the basic of clade L (Fig. 2). From 2019 to the beginning of 2020, the virus evolved from the original clade L to O, G and V. In 2021, new clades arose, including GR, GH, GRY, GK, and GRA, which were shown in Group A. Since clade L was first detected in Wuhan, it spread to the whole country quickly, including Anhui, Shandong, Sichuan, and Beijing. At the same time, clade S was transmitted in many cities like Zhejiang and Shandong. Clade O was also detected in the early stages of the outbreak (Fig. 2). Clade G was the ancestor of GR, GH, GK. The Beijing isolates belonging to the GR clade clustered with isolates from Fujian and Anhui. However, the Beijing isolates belonging to the GH clade clustered with isolates from Zhejiang. It is also important to note that the Beijing isolates within the GK clade clustered with Jiangsu, Fujian and Zhejiang isolates belonging to this clade (Fig. 2). These were great hints for traceability. Finally, the Beijing isolates evolved into clades GRY and GRA.

3.3. Different SARS-CoV-2 mutation profiles

Among the sequences analyzed, a total of 603 amino acid variations were identified, including 30 deletions (Fig. 3A, E). About half of mutations (393) were found only in ORF1ab gene, while Nsp12_P323L showed the highest frequency, reaching 41.4%. In addition, the mutations Nsp4_T492I, Nsp8_L84S, Nsp3_T223I, Nsp5_P132H, Nsp14_I42V, Nsp8_S24P, Nsp4_L264F, Nsp3_G489S, Nsp4_T327I, Nsp13_R392C, and Nsp4_L438F were all found at a rate of 20% (Fig. 3B). The S gene reported a total of 120 gene mutations, of which S_D614G showed the highest frequency, reaching 43.3%. Secondly, the mutation frequencies of S_P681, S_T478K and S_T19 were more than 30% (Fig. 3C). A total of 44 mutations were found in the N gene, with the mutation N_R203 found at 35.5% (Fig. 3D). Besides N_R203K, N_R203 M was also identified. The more details were shown in Table 1. In addition, 2 amino acids mutated into the stop codon (Table 1), and the effect of this mutation on the virus needs further study. Additionally, Nsp3, S gene, and N gene were hotspots for mutations (Table 1).

It was demonstrated that gene deletions can facilitate the immune evasion of SARS-COV-2 and adaptation to the host [22]. The S gene and N gene were hotspots for deletion (Fig. 3E). S_ P25del and S_ P26del had the highest frequency (20.2%), followed by NSP6_ S106del, NSP6_ G107del, NSP6_ F108del, N_ E31del, and N_ R32del (up to 10.5%) (Fig. 3E). Gene deletions were found to be more frequent in clade GRA followed by GK, GH. These results showed that variants carried their unique deletions.

In addition to the maker mutations contained in each clade, there were many other high-frequency mutations. For example, the mutation frequencies of NSP6–F108del, NSP6-G107del, and NSP6–S106del in GRY were also higher (11.8%). Table 1 showed that the same position on the genome can mutate into two different amino acids, including stop codon.

3.4. Evolutionary characteristics of SARS-COV-2 genes

In order to investigate the evolutionary selection pressure of SARS-COV-2, Ka (synonymous substitution rate) (Fig. 4A), Ks (nonsynonymous substitution rate) (Fig. 4B), and the Ka/Ks ratio (Fig. 4C) for each gene sequence from December 2019 to October 2022 were calculated and they were reported in Fig. 4. The Ka/Ks ratios of S genes were all higher than 1 in clades GH, GRY, GRA, and GK, indicating that S gene underwent a positive selection. Similarly, the N genes of clades GRY, GK, and GRA were higher than 1, suggesting that N genes evolved toward positive selection as well. For ORF1ab and M genes, the Ka/Ks values were less than 1, suggesting that a purifying selection occurred in recent years. These results may suggest that SARS-COV-2 can infect and escape the immune host defense more with a positive selection of S and N genes in the continuous evolution process.

4. Discussion

In this study, SARS-CoV-2 sequences downloaded from GISAID database from 2019 to April 12, 2022 in China, were included in the study. These sequences were divided into 10 clades, named S, L, O, V, G, GK, GH, GR, GRY, and GRA. The China epidemic was characterized by an early clades L, O, V and S that was subsequently replaced by clade G variants (Figs. 1B and 2). According to the Bayesian evolutionary analysis, GR originated from G and then evolved into GRY, GRA (Fig. 2), which was consistent with previous report [23]. There were no reported cases in some months of these years (Fig. 1). Containment and suppression strategies have been implemented in China to control the prevalence of COVID-19 (14). Case finding and management, with identification and quarantine of close contacts limit the wide spread of the virus. This may be the reason for the relatively concentrated distribution of cases in a certain period of time. The absence of cases in some months in these three years may be due to this policy leading to the clearing of the virus. In these years, L, S and GRA were the major clades which accounted for a large proportion (Fig. 1). Between the end of 2019 and the beginning of 2020, clades S and L accounted for the majority, while the major clade was replaced by GRA in 2022 (Figs. 1 and 2). The proportions of clades G, GR and GH were relatively low, which is different from the global trend (Fig. 1) [24]. This may be due to the implementation of the NPI policy controlling the spread of GR and GH, resulting in fewer sequences. However, with the clade GRA continuous strengthening of the virus transmission ability, this strategy may be difficult to implement all the time, and the wide spread of the virus was inevitable.

Table 1
Mutation frequency of sites with no less than three mutations.

Frequency of mutation sites (mutate no less than 3 times)					
genome region	amino acid position	reference	allele	frequency	clades which the mutation occurred
nsp1	73	R	C	0.00381679	S
nsp2	85	T	I	0.030534	GH, O
nsp2	81	K	N	0.024809	GK
nsp2	129	P	L	0.017176	GK
nsp2	447	V	F	0.015267	GK
nsp2	198	V	I	0.009542	O, L
nsp2	475	C	Y	0.0076336	GR
nsp2	174	A	S	0.0057252	GRA
nsp2	326	C	Y	0.0057252	GR
nsp2	370	R	H	0.0057252	S
nsp2	429	T	I	0.0057252	GR
nsp3	223	T	I	0.20611	GR, GRA
nsp3	489	G	S	0.20229	GRA
nsp3	725	T	I	0.10115	GRA
nsp3	26	S	L	0.09923664	GK
nsp3	488	A	S	0.05916	GK
nsp3	1228	P	L	0.05916	GK
nsp3	1469	P	S	0.05916	GK
nsp3	251	G	V	0.055344	V, O
nsp3	822	P	L	0.040076	GK
nsp3	1768	V	G	0.03244275	GH, L, G, GR, O
nsp3	57	Q	H	0.03053435	GH
nsp3	72	L	F	0.013359	GR
nsp3	1278	M	R	0.01145	GR, G, L
nsp3	185	Q	H	0.0076336	S
nsp3	171	S	L	0.0076336	GH
nsp3	1711	A	V	0.0076336	GK
nsp3	890	A	D	0.0076336	GRY
nsp3	1274	H	Y	0.0076336	GK
nsp3	789	I	V	0.0076336	L
nsp3	837	K	N	0.0076336	GH
nsp3	389	P	H	0.0076336	L
nsp3	1275	T	I	0.0076336	S
nsp3	1412	I	T	0.00763359	GRY
nsp3	183	T	I	0.00763359	GRY
nsp3	218	D	E	0.0057252	GR
nsp3	1881	I	V	0.0057252	GR
nsp3	1198	T	K	0.0057252	O
nsp3	55	V	F	0.00572519	GK
nsp3	333	A	V	0.00572519	L
nsp3	768	M	I	0.00572519	GK
nsp3	1103	P	Q, S	0.0057	GK
nsp4	492	T	I	0.2652	GK, GRA
nsp4	264	L	F	0.20229008	GRA
nsp4	327	T	I	0.20229	GRA
nsp4	438	L	F	0.20038	GRA
nsp4	167	V	L	0.05916	GK
nsp4	446	A	V	0.03816794	GK
nsp5	132	P	H	0.20611	GRA
nsp5	184	P	S	0.009542	O, L
nsp5	90	K	R	0.00763359	GH
nsp6	108	F	delete	0.11832061	GH, GRY, GRA
nsp6	106	S	delete	0.11832061	GH, GRY, GRA
nsp6	107	G	delete	0.11832	GH, GRY, GRA
nsp6	77	T	A	0.05916031	GK
nsp6	149	V	A	0.042	GK
nsp6	37	L	F	0.03816794	O, S, V, GH, GK
nsp6	181	T	I	0.024809	GK
nsp6	153	Y	C	0.009542	O, S
nsp7	82	V	A	0.099237	GK
nsp7	120	T	I	0.09923664	GK
nsp7	40	T	I	0.05916	GK
nsp7	45	P	L	0.024809	GK
nsp7	116	L	F	0.01526718	GK
nsp8	84	L	S	0.22137405	O, G, S
nsp8	145	T	I	0.10305344	GRA
nsp8	62	V	L	0.013359	S, O

(*continued on next page*)

Table 1 (*continued*)

Frequency of mutation sites (mutate no less than 3 times)					
genome region	amino acid position	reference	allele	frequency	clades which the mutation occurred
nsp8	76	I	V	0.01145038	GRA
nsp8	27	Q	stop	0.00954198	GR, GRA, GRY
nsp8	52	R	I	0.00763359	GR, GRY
nsp8	73	Y	C	0.00763359	GR, GRY
nsp9	95	N	S	0.0076336	S
nsp10	51	T	A	0.0076336	L
nsp12	323	P	L	0.41412	GR, GRA, GK, GH, G, GRY, O
nsp12	671	G	S	0.097328	GK
nsp12	736	D	G	0.0076336	L
nsp12	829	L	F	0.0057252	GH
nsp12	629	M	I	0.0057252	L
nsp12	97	A	V	0.00572519	O
nsp13	392	R	C	0.20038168	GRA
nsp13	77	P	L	0.10114504	GK
nsp13	598	A	V	0.0076336	O, L
nsp13	581	L	F	0.0076336	L
nsp13	481	T	M	0.0076336	S
nsp13	529	P	L	0.0057252	GR
nsp13	100	S	G	0.00572519	GR
nsp14	42	I	V	0.20611	GRA
nsp14	394	A	V	0.05916	GK
nsp14	31	T	I	0.0076336	S
nsp14	420	Y	stop	0.0076336	GH, L
nsp14	482	A	V	0.0057252	GR
nsp14	177	L	F	0.00572519	G
nsp15	234	H	Y	0.01526718	GK
nsp15	33	T	I	0.01145038	S
nsp15	276	K	R	0.0076336	GR, S
nsp15	127	V	F	0.00763359	GH, GR
nsp16	78	V	G	0.01335878	GH, L, GR, G
nsp16	120	T	I	0.00572519	GRA
S	614	D	G	0.43321	G, GR, GH, GRA, GRY, GK
S	681	P	H, R	0.31298	GRY, GR, GRA
S	478	T	K	0.30344	GK, GRA
S	19	T	I, R	0.30153	GK, GRA
S	142	G	D	0.29389	GK, GRA
S	417	K	N	0.23282	GH, GRA
S	501	N	Y	0.21756	GH, GR, GRY, GRA
S	27	A	S	0.20802	GH, GRA
S	339	G	D	0.20611	GRA
S	655	H	Y	0.20611	GRA
S	679	N	K	0.20611	GRA
S	764	N	K	0.20611	GRA
S	796	D	Y	0.20611	GRA
S	477	S	N	0.2042	GRA
S	484	E	A, K	0.2042	GH, GR, GRA
S	954	Q	H	0.2042	GRA
S	969	N	K	0.2042	GRA
S	24	L	delete	0.20229	GRA
S	25	P	delete	0.20229	GRA
S	26	P	delete	0.20229	GRA
S	213	V	G	0.20229	GRA
S	371	S	F	0.20229	GRA
S	373	S	P	0.20229	GRA
S	375	S	F	0.20229	GRA
S	376	T	A	0.20229	GRA
S	440	N	K	0.20229	GRA
S	493	Q	R	0.20229	GRA
S	498	Q	R	0.20229	GRA
S	505	Y	H	0.20229	GRA
S	405	D	N	0.20038	GRA
S	408	R	S	0.19847	GRA
S	1221	I	T	0.10305	GRA
S	452	L	R	0.10115	GK, GRA
S	156	E	G	0.09542	GK
S	157	F	delete	0.09542	GK
S	158	R	delete	0.09542	GK
S	950	D	N	0.087786	GK

(*continued on next page*)

Table 1 (*continued*)

Frequency of mutation sites (mutate no less than 3 times)					
genome region	amino acid position	reference	allele	frequency	clades which the mutation occurred
S	95	T	I	0.03626	GR, GRA, GK
S	222	A	V	0.022901	GK
S	146	H	Y	0.020992	GK
S	5	L	F	0.015267	GRA, GH, S, O
S	49	H	Y	0.015267	L
S	215	D	G, H	0.01145	GH, L, S
S	570	A	D, V	0.01145	GR, S, GRY
S	243	A	delete	0.009542	GH, L
S	244	L	delete	0.009542	GH, L
S	484	E	A	0.20419847	GRA
S	547	T	I,K	0.009542	GRA, GK
S	69	H	Y, delete	0.0076336	S, GRY
S	80	D	A	0.0076336	GH
S	144	Y	delete	0.0076336	S, GRY
S	242	L	delete	0.0076336	GH
S	701	A	V	0.0076336	GH
S	982	S	A	0.0076336	GR, GRY
S	1084	D	Y	0.0076336	L
S	1118	D	H	0.0076336	GR, GRY
S	1130	I	L	0.0076336	GR
S	1162	P	S	0.0076336	S, GRY
S	50	S	L	0.0057252	L, S
S	70	V	delete	0.0057252	GRY
S	153	M	T	0.0057252	L
S	688	A	S	0.0057252	GR
S	938	L	F	0.0057252	GK
S	943	S	P	0.0057252	G,GR
S	1040	V	F	0.0057252	L
S	1114	I	V	0.0057252	GK
E	9	T	I	0.20611	GRA
E	71	P	L	0.0076336	GH
M	19	Q	E	0.2042	GRA
M	63	A	T	0.2042	GRA
M	82	I	T	0.099237	GK
M	3	D	G	0.0057252	GRA,G
M	64	C	F	0.0057252	GK
N	203	R	K, M	0.35496	GR, GK, O, G, GRY, GRA
N	204	G	R	0.25191	GR, O, GRY, GRA
N	13	P	S, L	0.21374	O, GK, GRA
N	31	E	delete	0.10496	GRA
N	32	R	delete	0.10496	GRA
N	33	S	delete	0.10496	GRA
N	413	S	R	0.10496	GRA
N	63	D	G	0.099237	GK
N	377	D	Y	0.099237	GK
N	215	G	C, delete	0.05916	GK
N	194	S	L	0.015267	GH, L
N	385	R	K	0.015267	GK
N	205	T	I	0.01145	GH, S
N	235	S	F	0.01145	GR, GH, L, GRY
N	202	S	N	0.009542	S, O
N	11	N	T	0.0076336	L
N	211	A	V, delete	0.0076336	GR, GK
N	3	D	L	0.0057252	GRY
N	8	N	T	0.0057252	GR
N	213	N	Y, delete	0.0057252	GK, S

Between 2020 and 2021, we observed a subsequent switch in dominance from clade L to clade G-variants (G, GK, GH, GR) (Figs. 1B and 2). The mutation S_D614G was found to enhance viral replication and infectiousness [25]. Indeed, G614 can enhance virus replication in the upper respiratory tract without entering the lower respiratory tract, including the lungs, and with no severe disease in patients infected with G614 [26,27]. S_D614G is the marker mutant of clades G, GR, GK, GH, GRA, and GRY. In addition to S_D614G, clade GH has the other marker mutation Nsp3_Q57H which was found in this study (Table 1). Q57 is responsible for the formation of hydrophilic constrictions in the transmembrane helix 1 (TM-1). Patients who were positive for the Nsp3_Q57H mutation had better outcomes according to previous study [28]. Nsp_P323L was associated with the highest death risk [28]. Nsp_P323L was found in clades G, GH, GK, GR, GRA, GRY, and O (Table 1). Further studies will be required to explore what kind of impact on the severity of COVID-19 if the combination of nsp_P323L and Nsp3_Q57H in one genome. The clade GRA have shown notable changes regarding S protein

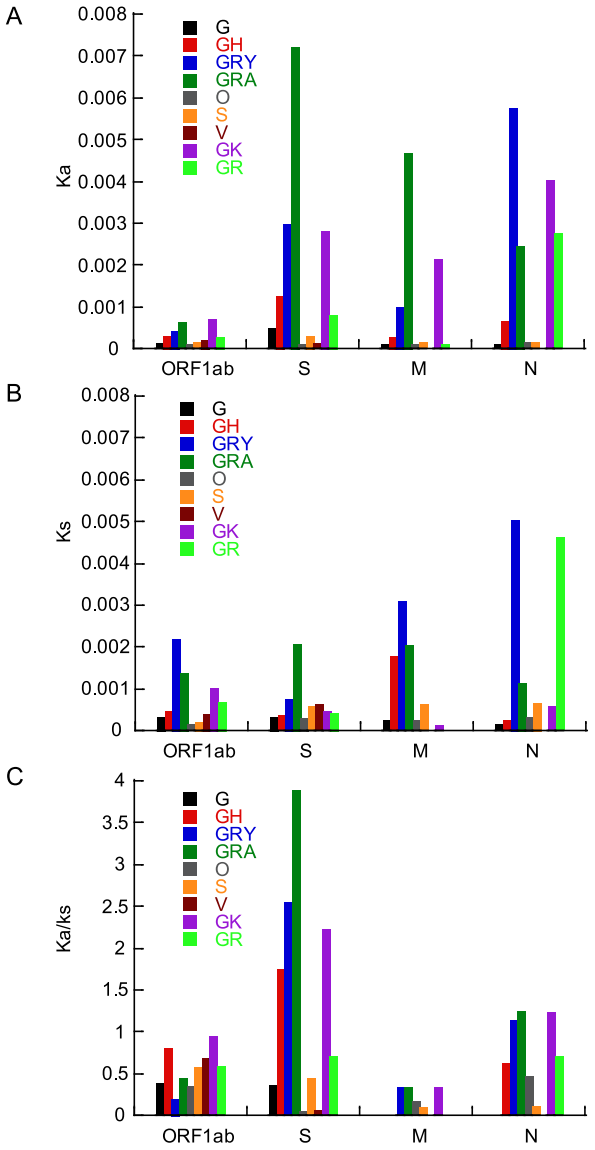


Fig. 4. Ka value (A), Ks value (B), and the value of Ka/Ks (C) of each gene in different viral clade.

mutants which are known to contain important neutralizing antibody epitopes [29]. In our study, the spike protein has more than 36 mutations including S375F, K417 N, T547K (Table 1). These mutations make the clade GRA less sensitive to vaccine-induced antibody neutralization [30,31]. In addition, N_ R203 K/G204G, which mostly occurred in clade GRA (Table 1), was associated with a more rapid virus replication in primary human upper airway tissues [32]. The increases in the infectivity and virulence of R203 K/G204R variants could contribute to the increased transmission and mortality of GRA [33,34]. Previous studies have shown that deletions can influence the replication and transmission of SARS-CoV-2 [35,36]. In this study, 30 amino acid deletions were found, with clade GRA reporting the highest number and frequency. These deletions showed a clear regional preference, which mainly occurred in the Nsp6, S, N proteins (Table 1). This result is consistent with previous study [22]. In details, spike_P25 and P26 showed the highest frequency of deletions (20.2%), followed by Nsp6_S106, Nsp6_G107, Nsp6_F108, N_E31, N_R32, S_L24, S_F157, and S_R158 (Fig. 3). Deletions in the S protein contribute to the viral adaption, including the viral transmissibility and immune escape [37]. However, the function of other deletions was still unknown. Furthermore, the cooperation of these deletions with some SNPs may play a certain role in the SARS-CoV-2 adaption and evolution. Therefore, further studies needed to understand their role in viral evolution and transmission. The first discovery of nucleotide mutated into stop codon in SARS-CoV-2 genome was in Italian patient samples [38]. This mutation was in ORF6 protein and did not affect the replication of the virus or the neutralization ability of the patient's antibody [38]. This study found that there were 2 amino acids mutated into the stop codons in Nsp8 and Nsp14 (Table 1). Most stop codons are highly deleterious [39]. Further studies needed to understand their exact role in viral evolution and transmission.

In addition, the evolutionary pattern and characteristics of the proteins were also investigated. Mutations refer to the virus to

undergo certain changes which can lead to develop some new isolates after replications. Non-synonymous substitutions play a very significant role as this type of mutation makes change in amino acid. In this study, we found that S protein was under an evolutionary trend of strong positive selection, together with N protein (Fig. 4) which was consistent with precious study [40]. This suggested that SARS-CoV-2 may enhance its infectivity and transmission through the positive selection evolution of S and N proteins.

Not all clades in this study had a sufficient number of genome sequences. In order to obtain more precise insight regarding mutation frequency and genetic variation characteristics, we need further studies with more genome sequences of each clade. This larger amount of samples will allow us to evaluate the virus's evolution better.

In conclusion, this study identified the mutation sites and evolutionary characteristics of SARS-CoV-2 genomes downloaded from GISAID in China when implementing NPI. In the early stages of the epidemic, the virus spread to many regions and was very dispersed (Figs. 1 and 2, Group B, C). After 2021, the distribution of the virus was relatively concentrated and the sequence number was relatively small (Figs. 1 and 2, Group A), indicating the control effect of NPI on the epidemic. Our analysis provided a perspective to believe the idea that outbreaks can be contained successfully through NPI.

Author contribution statement

Peng Zhang: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data. Dongzi Liu: Analyzed and interpreted the data. Lei Ji: Conceived and designed the experiments; Performed the experiments. Fenfen Dong: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

Data availability statement

Data associated with this study has been deposited at Global Initiative on Sharing All Influenza Data (GISAID), under the accession number EPI_ISL_1040031; EPI_ISL_1040033; EPI_ISL_1040034; EPI_ISL_1040036; EPI_ISL_1040037; EPI_ISL_1040040; EPI_ISL_1040041; EPI_ISL_1040042; EPI_ISL_1040043; EPI_ISL_1040045; EPI_ISL_1040046; EPI_ISL_1040047; EPI_ISL_1040048; EPI_ISL_1068593; EPI_ISL_1068682; EPI_ISL_1069173; EPI_ISL_1069192; EPI_ISL_1069193; EPI_ISL_1069198; EPI_ISL_1069199; EPI_ISL_1069200; EPI_ISL_1069201; EPI_ISL_1069203; EPI_ISL_1069205; EPI_ISL_1069206; EPI_ISL_1069207; EPI_ISL_1069208; EPI_ISL_1069210; EPI_ISL_1069211; EPI_ISL_1069213; EPI_ISL_1069215; EPI_ISL_1069216; EPI_ISL_1069219; EPI_ISL_1081193; EPI_ISL_1081194; EPI_ISL_1081195; EPI_ISL_1081196; EPI_ISL_1081197; EPI_ISL_1081198; EPI_ISL_1081199; EPI_ISL_1081260; EPI_ISL_1081291; EPI_ISL_1081310; EPI_ISL_1081311; EPI_ISL_1081324; EPI_ISL_1081325; EPI_ISL_1081326; EPI_ISL_1081331; EPI_ISL_1081340; EPI_ISL_1117376; EPI_ISL_1117377; EPI_ISL_1117378; EPI_ISL_11799970; EPI_ISL_11799982; EPI_ISL_11799983; EPI_ISL_11799984; EPI_ISL_12040149; EPI_ISL_12241077; EPI_ISL_12241078; EPI_ISL_12241079; EPI_ISL_12241080; EPI_ISL_12241081; EPI_ISL_12241082; EPI_ISL_1259400; EPI_ISL_1260834; EPI_ISL_1260836; EPI_ISL_12901913; EPI_ISL_13499639; EPI_ISL_13499640; EPI_ISL_13499641; EPI_ISL_13858925; EPI_ISL_13858927; EPI_ISL_13858931; EPI_ISL_13858932; EPI_ISL_13858933; EPI_ISL_13858934; EPI_ISL_13858935; EPI_ISL_13858938; EPI_ISL_13858939; EPI_ISL_13858940; EPI_ISL_13858941; EPI_ISL_13858942; EPI_ISL_13858943; EPI_ISL_13858944; EPI_ISL_13858945; EPI_ISL_13858947; EPI_ISL_13858949; EPI_ISL_13858950; EPI_ISL_13858951; EPI_ISL_13858952; EPI_ISL_13858953; EPI_ISL_13858954; EPI_ISL_13858955; EPI_ISL_13858956; EPI_ISL_13858957; EPI_ISL_13858958; EPI_ISL_13858959; EPI_ISL_13858960; EPI_ISL_13858961; EPI_ISL_13858963; EPI_ISL_13858964; EPI_ISL_13858965; EPI_ISL_13858966; EPI_ISL_13858967; EPI_ISL_13858968; EPI_ISL_13858969; EPI_ISL_13858970; EPI_ISL_13858971; EPI_ISL_13858972; EPI_ISL_13858973; EPI_ISL_13858974; EPI_ISL_13858975; EPI_ISL_13858976; EPI_ISL_13858977; EPI_ISL_13858978; EPI_ISL_13858979; EPI_ISL_13858980; EPI_ISL_13858981; EPI_ISL_13858982; EPI_ISL_13858983; EPI_ISL_13858984; EPI_ISL_13858985; EPI_ISL_13858986; EPI_ISL_13858987; EPI_ISL_13858988; EPI_ISL_13858989; EPI_ISL_13858990; EPI_ISL_13858991; EPI_ISL_13858992; EPI_ISL_13858993; EPI_ISL_13858994; EPI_ISL_13858995; EPI_ISL_13858996; EPI_ISL_13858997; EPI_ISL_13858998; EPI_ISL_13858999; EPI_ISL_13859000; EPI_ISL_13859001; EPI_ISL_13859002; EPI_ISL_13859003; EPI_ISL_13859004; EPI_ISL_13859005; EPI_ISL_13859006; EPI_ISL_13859007; EPI_ISL_13859008; EPI_ISL_13859009; EPI_ISL_13859010; EPI_ISL_13859011; EPI_ISL_13859013; EPI_ISL_13859014; EPI_ISL_13859015; EPI_ISL_13859016; EPI_ISL_13859017; EPI_ISL_13859018; EPI_ISL_13859019; EPI_ISL_13859020; EPI_ISL_13859021; EPI_ISL_13859022; EPI_ISL_13859023; EPI_ISL_13859024; EPI_ISL_13859025; EPI_ISL_13859026; EPI_ISL_13859027; EPI_ISL_13859028; EPI_ISL_13859030; EPI_ISL_13859031; EPI_ISL_13859032; EPI_ISL_13859033; EPI_ISL_13859034; EPI_ISL_14675318; EPI_ISL_1655075; EPI_ISL_1655937; EPI_ISL_1663017; EPI_ISL_1911195; EPI_ISL_1911196; EPI_ISL_1911197; EPI_ISL_1911250; EPI_ISL_2131805; EPI_ISL_2131843; EPI_ISL_2132221; EPI_ISL_2433280; EPI_ISL_2433281; EPI_ISL_2433282; EPI_ISL_2433283; EPI_ISL_2433284; EPI_ISL_2779638; EPI_ISL_2779639; EPI_ISL_2811805; EPI_ISL_2965600; EPI_ISL_2965601; EPI_ISL_2965602; EPI_ISL_2965603; EPI_ISL_2965604; EPI_ISL_3006797; EPI_ISL_3006798; EPI_ISL_3006799; EPI_ISL_3127444; EPI_ISL_3127445; EPI_ISL_3154875; EPI_ISL_3501737; EPI_ISL_3501738; EPI_ISL_3501739; EPI_ISL_3501740; EPI_ISL_3501741; EPI_ISL_3611059; EPI_ISL_3611060; EPI_ISL_402124; EPI_ISL_402127; EPI_ISL_402128; EPI_ISL_402129; EPI_ISL_402130; EPI_ISL_402132; EPI_ISL_404227; EPI_ISL_404228; EPI_ISL_406592; EPI_ISL_406593; EPI_ISL_406594; EPI_ISL_406595; EPI_ISL_406970; EPI_ISL_407313; EPI_ISL_408480; EPI_ISL_408481; EPI_ISL_408482; EPI_ISL_408484; EPI_ISL_408485; EPI_ISL_408486; EPI_ISL_408488; EPI_ISL_412386; EPI_ISL_412459; EPI_ISL_412978; EPI_ISL_412979; EPI_ISL_412980; EPI_ISL_412981; EPI_ISL_412982; EPI_ISL_412983; EPI_ISL_414663; EPI_ISL_414691; EPI_ISL_415709; EPI_ISL_415711; EPI_ISL_416042; EPI_ISL_416044; EPI_ISL_416046; EPI_ISL_416047; EPI_ISL_416425; EPI_ISL_416473; EPI_ISL_416474; EPI_ISL_418990; EPI_ISL_418991; EPI_ISL_421221; EPI_ISL_421222; EPI_ISL_421224; EPI_ISL_421225; EPI_ISL_421226; EPI_ISL_421227; EPI_ISL_421228; EPI_ISL_421229; EPI_ISL_421230; EPI_ISL_421231; EPI_ISL_421232; EPI_ISL_421233; EPI_ISL_421234; EPI_ISL_421235; EPI_ISL_421236; EPI_ISL_429852; EPI_ISL_429853; EPI_ISL_429854; EPI_ISL_430722;

EPI_ISL_430724; EPI_ISL_430725; EPI_ISL_430728; EPI_ISL_430729; EPI_ISL_430730; EPI_ISL_430731; EPI_ISL_430733; EPI_ISL_430734; EPI_ISL_430735; EPI_ISL_430736; EPI_ISL_430737; EPI_ISL_430738; EPI_ISL_430740; EPI_ISL_430741; EPI_ISL_430742; EPI_ISL_431118; EPI_ISL_431180; EPI_ISL_431240; EPI_ISL_431782; EPI_ISL_431783; EPI_ISL_431784; EPI_ISL_431785; EPI_ISL_434534; EPI_ISL_444969; EPI_ISL_451076; EPI_ISL_451313; EPI_ISL_451314; EPI_ISL_451315; EPI_ISL_451316; EPI_ISL_451318; EPI_ISL_451319; EPI_ISL_451320; EPI_ISL_451321; EPI_ISL_451322; EPI_ISL_451325; EPI_ISL_451326; EPI_ISL_451327; EPI_ISL_451328; EPI_ISL_451329; EPI_ISL_451330; EPI_ISL_451331; EPI_ISL_451334; EPI_ISL_451337; EPI_ISL_451338; EPI_ISL_451344; EPI_ISL_451345; EPI_ISL_451346; EPI_ISL_451348; EPI_ISL_451353; EPI_ISL_451354; EPI_ISL_451356; EPI_ISL_451357; EPI_ISL_451359; EPI_ISL_451360; EPI_ISL_451365; EPI_ISL_451369; EPI_ISL_451370; EPI_ISL_451371.

EPI_ISL_451374; EPI_ISL_451376; EPI_ISL_451377; EPI_ISL_451378; EPI_ISL_451379; EPI_ISL_451380; EPI_ISL_451381; EPI_ISL_451382; EPI_ISL_451383; EPI_ISL_451384; EPI_ISL_451385; EPI_ISL_451386; EPI_ISL_451387; EPI_ISL_451388; EPI_ISL_451389; EPI_ISL_451390; EPI_ISL_451391; EPI_ISL_451392; EPI_ISL_451393; EPI_ISL_451394; EPI_ISL_451395; EPI_ISL_451398; EPI_ISL_4515846; EPI_ISL_4515902; EPI_ISL_452327; EPI_ISL_452328; EPI_ISL_452329; EPI_ISL_452330; EPI_ISL_452331; EPI_ISL_452332; EPI_ISL_452333; EPI_ISL_452334; EPI_ISL_452335; EPI_ISL_452336; EPI_ISL_452337; EPI_ISL_452338; EPI_ISL_452339; EPI_ISL_452340; EPI_ISL_452341; EPI_ISL_452342; EPI_ISL_452343; EPI_ISL_452344; EPI_ISL_452345; EPI_ISL_452346; EPI_ISL_452347; EPI_ISL_452348; EPI_ISL_452349; EPI_ISL_452350; EPI_ISL_452351; EPI_ISL_452352; EPI_ISL_452353; EPI_ISL_452354; EPI_ISL_452355; EPI_ISL_452356; EPI_ISL_452357; EPI_ISL_452358; EPI_ISL_452359; EPI_ISL_452360; EPI_ISL_452361; EPI_ISL_452362; EPI_ISL_452363; EPI_ISL_452364; EPI_ISL_453779; EPI_ISL_453780; EPI_ISL_453781; EPI_ISL_453782; EPI_ISL_453783; EPI_ISL_455460; EPI_ISL_455461; EPI_ISL_455462; EPI_ISL_455463; EPI_ISL_455464; EPI_ISL_455465; EPI_ISL_455466; EPI_ISL_455467; EPI_ISL_455502; EPI_ISL_455503; EPI_ISL_455504; EPI_ISL_455505; EPI_ISL_455506; EPI_ISL_455507; EPI_ISL_455508; EPI_ISL_455509; EPI_ISL_482575; EPI_ISL_482576; EPI_ISL_482577; EPI_ISL_482578; EPI_ISL_482579; EPI_ISL_482580; EPI_ISL_482581; EPI_ISL_482582; EPI_ISL_482583; EPI_ISL_482584; EPI_ISL_482585; EPI_ISL_482586; EPI_ISL_495459; EPI_ISL_497950; EPI_ISL_575330; EPI_ISL_5903004; EPI_ISL_632934; EPI_ISL_7876604; EPI_ISL_7876605; EPI_ISL_7876606; EPI_ISL_7876607; PI_ISL_7876608; EPI_ISL_7876609; EPI_ISL_7876610; EPI_ISL_8582058; EPI_ISL_8582059; EPI_ISL_8582060; EPI_ISL_8582061; EPI_ISL_8582062; EPI_ISL_8582063; EPI_ISL_8582064; EPI_ISL_8582065; EPI_ISL_8582066; EPI_ISL_8582067; EPI_ISL_8582068; EPI_ISL_8582069; EPI_ISL_8582070; EPI_ISL_8582071; EPI_ISL_8582072; EPI_ISL_8582073; EPI_ISL_8582074; EPI_ISL_8582075; EPI_ISL_8582076; EPI_ISL_8582077; EPI_ISL_8582078; EPI_ISL_8582079; EPI_ISL_8582080; EPI_ISL_8582083; EPI_ISL_8582085; EPI_ISL_8582086; EPI_ISL_8582087; EPI_ISL_8582088; EPI_ISL_962528; EPI_ISL_962529; EPI_ISL_962530; EPI_ISL_962531; EPI_ISL_962532; EPI_ISL_962533; EPI_ISL_962534; EPI_ISL_962535; EPI_ISL_962536; EPI_ISL_962537; EPI_ISL_962538; EPI_ISL_962539; EPI_ISL_962540; EPI_ISL_962541; EPI_ISL_962542; EPI_ISL_962543; EPI_ISL_962544; EPI_ISL_962545; EPI_ISL_962546; EPI_ISL_962547; EPI_ISL_962548; EPI_ISL_962549; EPI_ISL_962550; EPI_ISL_962551; EPI_ISL_962552; EPI_ISL_962553; EPI_ISL_962554; EPI_ISL_962555; EPI_ISL_962556; EPI_ISL_962557; EPI_ISL_962558; EPI_ISL_962559; EPI_ISL_962560; EPI_ISL_962561; EPI_ISL_962562; EPI_ISL_962563; EPI_ISL_962564; EPI_ISL_962565; EPI_ISL_962566; EPI_ISL_962567; EPI_ISL_962568; EPI_ISL_962569; EPI_ISL_962570; EPI_ISL_962571; EPI_ISL_962572; EPI_ISL_962573; EPI_ISL_962574; EPI_ISL_962575; EPI_ISL_962576; EPI_ISL_962577; EPI_ISL_962578; EPI_ISL_962579; EPI_ISL_962580; EPI_ISL_962581; EPI_ISL_962582; EPI_ISL_962583; EPI_ISL_962584; EPI_ISL_962868; EPI_ISL_962869; EPI_ISL_962870; EPI_ISL_962871; EPI_ISL_962873; EPI_ISL_962874; EPI_ISL_962875.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P. Zhou, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (2020) 270–273, https://doi.org/10.1038/s41586-020-2012-7.
[2] B. Salzberger, et al., Epidemiology of SARS-CoV-2, Infection 49 (2021) 233–239, https://doi.org/10.1007/s15010-020-01531-3.
[3] B. Dearlove, et al., A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants, Proc. Natl. Acad. Sci. U.S.A. 117 (2020) 23652–23662, https://doi.org/10.1073/pnas.2008281117.
[4] Correction to: the origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK, Virus Evol 8 (2022) veac119, https://doi.org/10.1093/ve/veac119.
[5] L. Corey, et al., SARS-CoV-2 variants in patients with immunosuppression, N. Engl. J. Med. 385 (2021) 562–566, https://doi.org/10.1056/NEJMsb2104756.
[6] R.A. Khailany, et al., Genomic characterization of a novel SARS-CoV-2, Gene Rep 19 (2020), 100682, https://doi.org/10.1016/j.genrep.2020.100682.
[7] K. Schubert, et al., SARS-CoV-2 Nsp1 binds the ribosomal mRNA channel to inhibit translation, Nat. Struct. Mol. Biol. 27 (2020) 959–966, https://doi.org/10.1038/s41594-020-0511-8.
[8] J. Ma, et al., Structure and function of N-terminal zinc finger domain of SARS-CoV-2 NSP2, Virol. Sin. 36 (2021) 1104–1112, https://doi.org/10.1007/s12250-021-00431-6.
[9] Q. Peng, et al., Structural and biochemical characterization of the nsp12-nsp7-nsp8 core polymerase complex from SARS-CoV-2, Cell Rep. 31 (2020), 107774, https://doi.org/10.1016/j.celrep.2020.107774.
[10] Q. Wang, et al., Structural and functional basis of SARS-CoV-2 entry by using human ACE2, Cell 181 (2020) 894–904 e899, https://doi.org/10.1016/j.cell.2020.03.045.
[11] S. Lu, et al., The SARS-CoV-2 Nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein, bioRxiv (2020), https://doi.org/10.1101/2020.07.30.228023.
[12] E.A.J. Alsaadi, et al., Identification of a membrane binding peptide in the envelope protein of MHV coronavirus, Viruses (2020), https://doi.org/10.3390/v12091054, 12.
[13] C. Bai, et al., Overview of SARS-CoV-2 genome-encoded proteins, Sci. China Life Sci. 65 (2022) 280–294, https://doi.org/10.1007/s11427-021-1964-4.
[14] Z. Li, et al., Active case finding with case management: the key to tackling the COVID-19 pandemic, Lancet 396 (2020) 63–70, https://doi.org/10.1016/S0140-6736(20)31278-2.

[15] S. Cleemput, et al., Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes, Bioinformatics 36 (2020) 3552–3555, https://doi.org/10.1093/bioinformatics/btaa145.
[16] S. Cleemput, et al., Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes, bioRxiv (2020), https://doi.org/10.1101/2020.01.31.928796.
[17] M. Mauri, et al., RAWGraphs: A Visualisation Platform to Create Open Outputs 28 (2017) 21–28, https://doi.org/10.1145/3125571.3125585, 25.
[18] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, Mol. Biol. Evol. 30 (2013) 772–780, https://doi.org/10.1093/molbev/mst010.
[19] M.A. Suchard, et al., Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10, Virus Evol 4 (2018) vey016, https://doi.org/10.1093/ve/vey016.
[20] D. Zhang, et al., PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies, Mol Ecol Resour 20 (2020) 348–355, https://doi.org/10.1111/1755-0998.13096.
[21] A. Rambaut, et al., Posterior summarization in bayesian phylogenetics using tracer 1.7, Syst. Biol. 67 (2018) 901–904, https://doi.org/10.1093/sysbio/syy032.
[22] S. Weng, et al., Conserved pattern and potential role of recurrent deletions in SARS-CoV-2 evolution, Microbiol. Spectr. 10 (2022), e0219121, https://doi.org/10.1128/spectrum.02191-21.
[23] L. Zoppi, Viral clades of SARS-CoV-2.2023-01-13. https://www.news-medical.net/health/Viral-Clades-of-SARS-CoV-2.aspx.
[24] Nextstrain, Genomic epidemiology of SARS-CoV-2 with subsampling focused globally since pandemic start. https://nextstrain.org/ncov/gisaid/global/all-time?c=GISAID_clade.
[25] J.A. Plante, et al., Spike mutation D614G alters SARS-CoV-2 fitness, Nature 592 (2021) 116–121, https://doi.org/10.1038/s41586-020-2895-3.
[26] B. Korber, et al., Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus, Cell 182 (2020) 812, https://doi.org/10.1016/j.cell.2020.06.043, 827 e819.
[27] L. Yurkovetskiy, et al., Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant, Cell 183 (2020) 739, https://doi.org/10.1016/j.cell.2020.09.032, 751 e738.
[28] D.A. Obeid, et al., SARS-CoV-2 genetic diversity and variants of concern in Saudi Arabia, J Infect Dev Ctries 15 (2021) 1782–1791, https://doi.org/10.3855/jidc.15350.
[29] P. Arora, et al., Comparable neutralisation evasion of SARS-CoV-2 omicron subvariants BA.1, BA.2, and BA.3, Lancet Infect. Dis. 22 (2022) 766–767, https://doi.org/10.1016/S1473-3099(22)00224-9.
[30] B. Hu, et al., Spike mutations contributing to the altered entry preference of SARS-CoV-2 omicron BA.1 and BA.2, Emerg. Microb. Infect. 11 (2022) 2275–2287, https://doi.org/10.1080/22221751.2022.2117098.
[31] J. Zou, et al., Cross-neutralization of omicron BA.1 against BA.2 and BA.3 SARS-CoV-2, Nat. Commun. 13 (2022) 2956, https://doi.org/10.1038/s41467-022-30580-5.
[32] H. Wu, et al., Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2, Cell Host Microbe 29 (2021) 1788, https://doi.org/10.1016/j.chom.2021.11.005, 1801 e1786.
[33] N.G. Davies, et al., Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7, Nature 593 (2021) 270–274, https://doi.org/10.1038/s41586-021-03426-1.
[34] N.L. Washington, et al., Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States, Cell 184 (2021) 2587, https://doi.org/10.1016/j.cell.2021.03.052, 2594 e2587.
[35] B.E. Young, et al., Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study, Lancet 396 (2020) 603–611, https://doi.org/10.1016/S0140-6736(20)31757-8.
[36] F. Benedetti, et al., Emerging of a SARS-CoV-2 viral strain with a deletion in nsp1, J. Transl. Med. 18 (2020) 329, https://doi.org/10.1186/s12967-020-02507-5.
[37] W.T. Harvey, et al., SARS-CoV-2 variants, spike mutations and immune escape, Nat. Rev. Microbiol. 19 (2021) 409–424, https://doi.org/10.1038/s41579-021-00573-0.
[38] S. Delbue, et al., Isolation of SARS-CoV-2 strains carrying a nucleotide mutation, leading to a stop codon in the ORF 6 protein, Emerg. Microb. Infect. 10 (2021) 252–255, https://doi.org/10.1080/22221751.2021.1884003.
[39] J.D. Bloom, R.A. Neher, Fitness effects of mutations to SARS-CoV-2 proteins, *bioRxiv*, https://doi.org/10.1101/2023.01.30.526314, 2023.
[40] B. Xi, et al., Analyses of Long-Term Epidemic Trends and Evolution Characteristics of Haplotype Subtypes Reveal the Dynamic Selection on SARS-CoV-2, 2022, https://doi.org/10.3390/v14030454. Viruses **14**.