Computational and Structural Biotechnology Journal 19 (2021) 1976–1985

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

AutoVEM: An automated tool to real-time monitor epidemic trends and key mutations in SARS-CoV-2 evolution

Binbin Xi [1], Dawei Jiang [1], Shuhua Li, Jerome R. Lon, Yunmeng Bai, Shudai Lin, Meiling Hu, Yuhuan Meng, Yimo Qu, Yuting Huang, Wei Liu, Lizhen Huang, Hongli Du *

School of Biology and Biological Engineering, South China University of Technology, Guangzhou 510006, China

ARTICLE INFO

Article history:
Received 7 January 2021
Received in revised form 31 March 2021
Accepted 2 April 2021
Available online 5 April 2021

Keywords:
SARS-CoV-2
Automated tool
Epidemic trends
Key mutations

ABSTRACT

With the global epidemic of SARS-CoV-2, it is important to effectively monitor the variation, haplotype subgroup epidemic trends and key mutations of SARS-CoV-2 over time. This is of great significance to the development of new vaccines, the update of therapeutic drugs, and the improvement of detection methods. The AutoVEM tool developed in the present study could complete all mutations detections, haplotypes classification, haplotype subgroup epidemic trends and candidate key mutations analysis for 131,576 SARS-CoV-2 genome sequences in 18 h on a 1 core CPU and 2 GB RAM computer. Through haplotype subgroup epidemic trends analysis of 131,576 genome sequences, the great significance of the previous 4 specific sites (C241T, C3037T, C14408T and A23403G) was further revealed, and 6 new mutation sites of highly linked (T445C, C6286T, C22227T, G25563T, C26801G and G29645T) were discovered for the first time that might be related to the infectivity, pathogenicity or host adaptability of SARS-CoV-2. In brief, we proposed an integrative method and developed an efficient automated tool to monitor haplotype subgroup epidemic trends and screen for the candidate key mutations in the evolution of SARS-CoV-2 over time for the first time, and all data could be updated quickly to track the prevalence of previous key mutations and new candidate key mutations because of high efficiency of the tool. In addition, the idea of combinatorial analysis in the present study can also provide a reference for the mutation monitoring of other viruses.

1. Introduction

SARS-CoV-2 had infected over 61.8 million (61,866,635) people and caused over 1.4 million (1,448,990) deaths in 216 countries or regions by November 29, 2020 [1], and the ongoing epidemic trend of COVID-19 had posed a great threat to global public health [2]. Previous researches on the origin or evolution of SARS-CoV-2 were mostly restricted to a limited number of virus genomes [3–5], and the results were rather controversial [4]. Recently, a study had built a single nucleotide variants (SNVs) database of 42,461 SARS-CoV-2 genomes (GESS), which could track the epidemic trend of single SNV over time in the genomes they had analyzed [6]. Xing et al. have developed a Python package, MicroGMT, which focuses on an upstream process of sequence mapping or SNV calling [7]. While CoV-GLUE is an online web application for the interpretation and analysis of SARS-CoV-2 virus genome sequences, with a focus on amino acid sequence variation [8]. All of these tools can analyze mutations or provide some useful information of these mutations. However, few studies focused on identification of candidate key mutations in the evolution of SARS-CoV-2 over time through linkage analysis and haplotype subgroup epidemic trends, which could reduce SARS-CoV-2 subgroup complexity greatly. In a previous study, we tracked the evolution trends of SARS-CoV-2 through linkage analysis and haplotype subgroup epidemic trends at three time points (March 22, 2020, April 6, 2020 and May 10, 2020), and found that the frequency of H1 haplotype with the 4 specific mutations (C241T, C3037T, C14408T and A23403G) increased over time, which indicated that they might be related to infectivity, pathogenicity or host adaptability of SARS-CoV-2 [9]. Thereinto, the A23403G mutation, which resulted in amino acid change of D614G in the spike protein, had been proved to be related to infectivity by several in vitro experiments subsequently [10–15].

Phylogenetic trees have been used in most studies on the evolution of SARS-CoV-2 [3–5,9,16,17], but a reliable phylogenetic tree is relatively time-consuming and requires a huge amount of computer resources because of bootstrapping, especially in the case of

* Corresponding author.
E-mail address: hldu@scut.edu.cn (H. Du).
[1] Equal contribution.

https://doi.org/10.1016/j.csbj.2021.04.002

a large number of genome sequence analyses [18]. According to the characteristics of virus transmission and epidemic, if the frequency of mutant allele at a certain locus gradually increases over time, it indicates that the mutant locus is likely to be related to viral infectivity, pathogenicity or host adaptability [9]. With the global epidemic of SARS-CoV-2, it is important to monitor the variation, haplotype subgroup epidemic trend and candidate key mutations of SARS-CoV-2 effectively in real-time, which may help in the development of new vaccines, update therapeutic drugs, and improve detection methods. Here we presented an innovative and integrative method and an automated tool to monitor haplotype subgroup epidemic trends and screen for the candidate key mutations in the evolution of SARS-CoV-2 over time efficiently. This tool skips the process of using a large number of genome sequences to construct the phylogenetic tree, and it can complete all mutations detection, haplotypes classification, haplotype subgroup epidemic trends and candidate key mutations analysis for 131,576 SARS-CoV-2 genome sequences in 18 h on a computer with a single core CPU and 2 GB RAM, which will play an important role in monitoring the epidemic trend of SARS-CoV-2 and finding candidate key mutation sites over time (Fig. 1).

2. Materials and methods

2.1. AutoVEM

AutoVEM is a highly specialized pipeline for quick monitoring the mutations and haplotype subgroup epidemic trends of SARS-CoV-2 by using virus genome sequences from GISAID. AutoVEM is written in Python language (Python 3.8.6) and runs on Linux machines with centos. Bowtie 2 [19], SAMtools [20], BCFtools [21], VCFtools [22] and Haploview [23] were applied in this automated tool.

2.2. Preparing genome sequences

The genome sequences of SARS-CoV-2 were downloaded from GISAID (https://www.epicov.org/) by 30 November 2020. All genome sequences should be placed in a folder, which would be as input of AutoVEM.

2.3. Operating system and hardware requirements

The software was tested on a computer with a single core CPU and 2 GB RAM running CentOS 7. For faster analysis, we recommend that you do not use computers with lower hardware resources.

2.4. Workflow of AutoVEM

Step 1: Quality control of genome sequences

The genome sequences were filtered out according to the following criteria: (1) sequences less than 29,000 in length; (2) low-quality sequences with > 15 unknown bases and > 50 degenerate bases; (3) sequences with unclear collection time information or country information.

Step2: Alignment and SNVs calling

Each genome sequence passed the quality control was aligned to the reference genome (NC_045512.2) using Bowtie 2 v2.4.2 (bowtie2-build –f, bowtie2 -f -x -U -S) [19]. SNVs and INDELs were

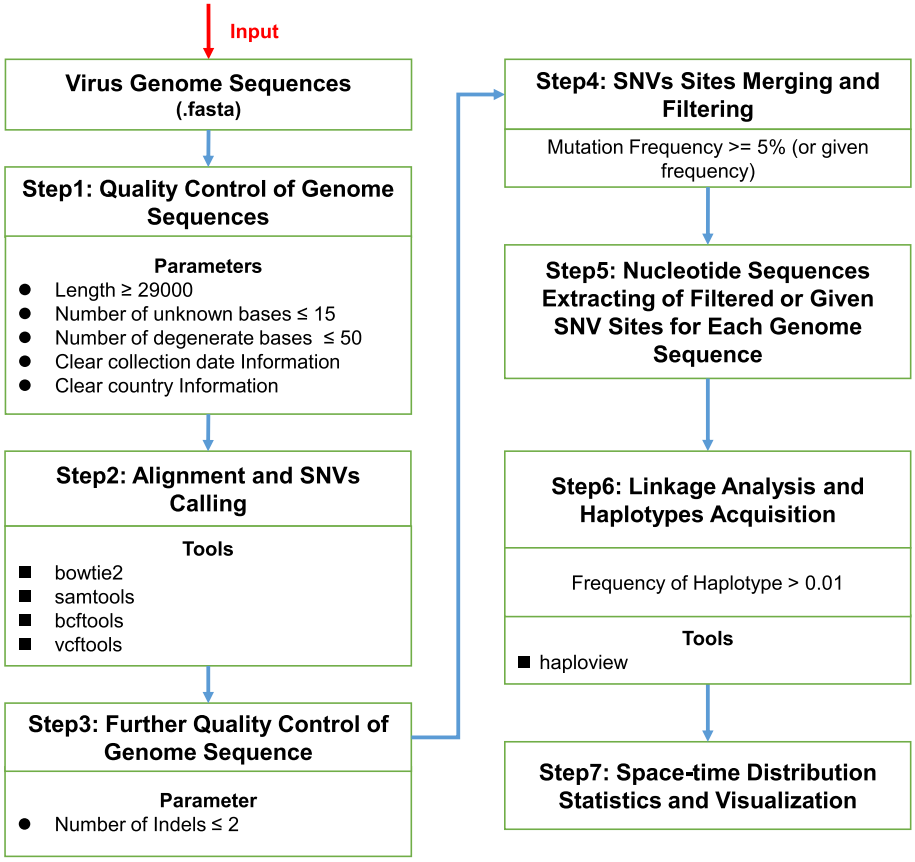

Fig. 1. Workflow chart of AutoVEM tool.

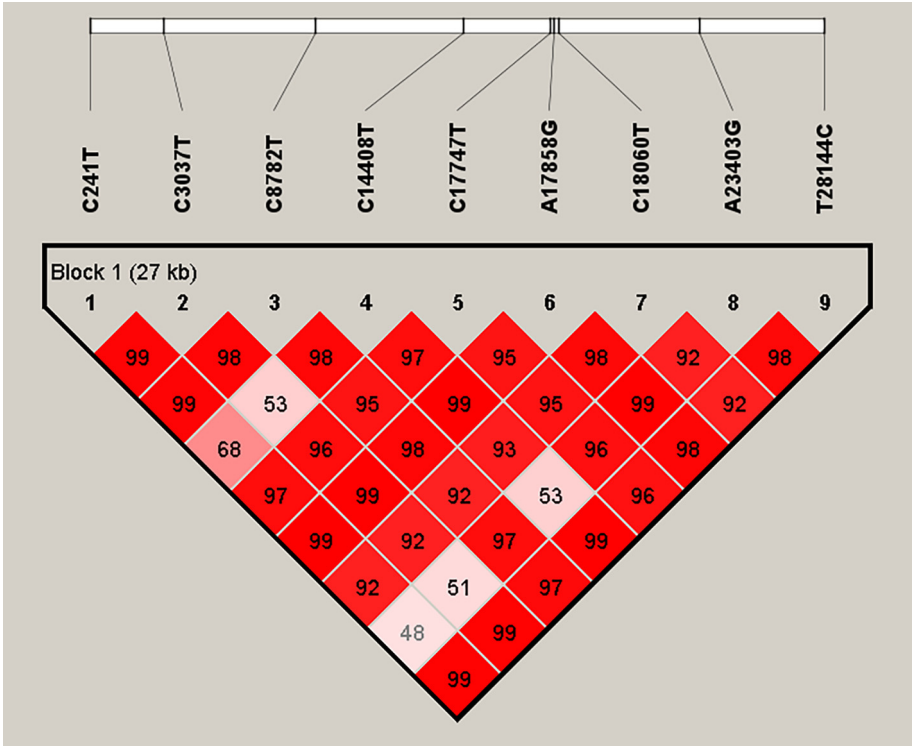

Fig. 2. Linkage analysis of the previous 9 specific sites.

called by SAMtools v1.10 (samtools sort) [20] and BCFtools v1.10.2 (bcftools mpileup --fasta-ref, bcftools call --ploidy-file -vm –o) [21], resulting in a Variant Call Format (VCF) file that contains both SNVs and INDELs information.

Step3: Further quality control of genome sequences.

INDELs of each sequence were extracted using VCFtools v0.1.16 (vcftools --vcf --recode --keep-only-indels –stdout, vcftools --vcf --recode --remove-indels --stdout) [22] and the sequences would be kept with the counts of INDELs ≤ 2. The SNVs were extracted from the kept sequences in VCF file.

Step4: SNV sites merging and filtering

SNVs of each sequence were merged and the mutation allele frequency of each SNV was calculated. The SNVs with mutation allele frequency less than 5% (or defined frequency) were filtered out.

Step5: Nucleotide Sequences Extracting of Filtered or Given SNV Sites for Each Genome Sequence

Nucleotides at specific sites with mutation allele frequency $\geq 5\%$ (or defined frequency) or given SNVs were extracted and organized in the order of genome position.

Step6: Linkage analysis of filtered or given mutation sites and haplotypes acquisition

Linkage analysis was performed and haplotypes with a frequency $\geq$ 1% were obtained using Haploview v4.2 (java –jar Haploview.jar -n -skipcheck -pedfile -info -blocks -png -out) [23]. The haplotype of each genome sequence was defined according to the haplotype sequence, and it was defined as 'other' if it had a frequency less than 1%.

Table 1
Haplotypes and frequencies of the previous 9 specific sites.

Name	Sequence	Frequency	Previous[1]
H1	TTCTCACGT	0.7184	H1
H2	CCCCCACAT	0.0454	H2
H3	CCTCTGTAC	0.0134	H3
H4	CCTCCACAC	0.0145	H4
H5	TTCCCACGT	0.0322	H5
H7	CTCTCACGT	0.0306	H7
H9 (new)	CCCTCACGT	0.0630	NA
*H*10 (new)	TTCTCACAT	0.0555	NA
*H*11 (new)	CTCCCACGT	0.0184	NA
other	NA	0.0084	NA

[1]Bai et al. [9].

Step7: Space-time distribution statistics of haplotypes and visualization

Sample information was captured from the annotation line in fasta format file of each genome sequence, then the haplotype subgroups were organized according to the country and collection date and the final results were visualized.

2.5. Variation annotation

The specific mutation sites were annotated by an online tool of *China National Center for Bioinformation* (https://bigd.big.ac.cn/ncov/online/tool/annotation?lang=en).

3. Results

3.1. AutoVEM development

The developed AutoVEM software has been shared on the website (https://github.com/Dulab2020/AutoVEM), which is freely available and applied in analyzing any public or local genome

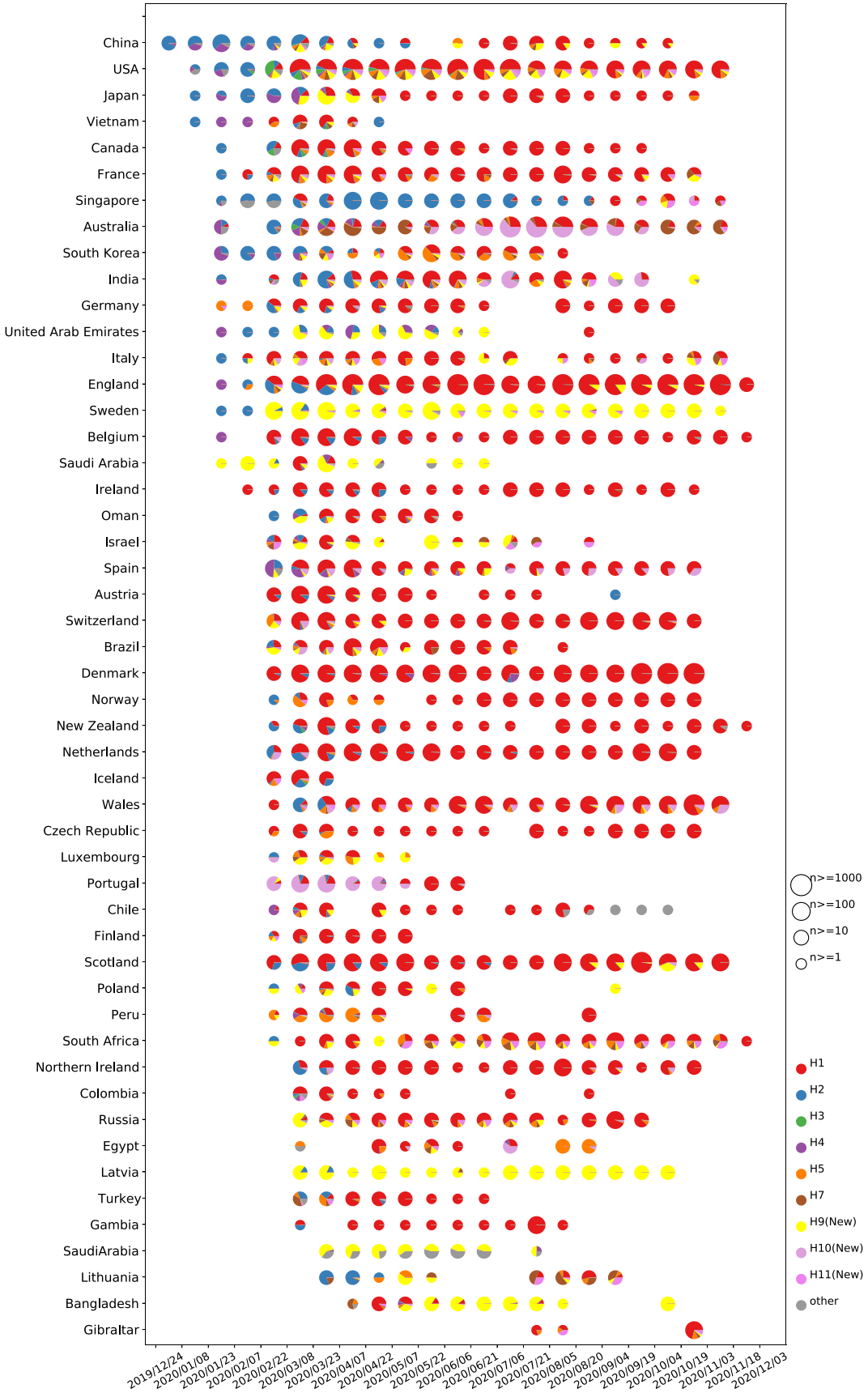

Fig. 3. Haplotype subgroup prevalence trends of the previous 9 specific sites. The numbers of haplotype subgroups of the previous 9 specific sites for 131,576 genomes with clear collection data detected in each country or region in chronological order. Countries or regions with a total number of genomes $\leq$ 100 were not shown in the figures.

Table 2
The information of the 23 sites with a mutation frequency greater than 5%.

Position	Ref	Alt	Frequency	Gene region	Mutation type	Protein changed	Coden changed	Impact
204	G	T	0.1303	5′-UTR	upstream	NA	NA	Modifier
241	C	T	0.8089	5′-UTR	upstream	NA	NA	Modifier
445	T	C	0.2103	gene-ORF1ab	synonymous	60 V	180gtT > gtC	Low
1059	C	T	0.1380	gene-ORF1ab	missense	265 T > I	794aCc > aTc	Moderate
1163	A	T	0.0609	gene-ORF1ab	missense	300I > F	898Att > Ttt	Moderate
3037	C	T	0.8574	gene-ORF1ab	synonymous	924F	2772ttC > ttT	Low
6286	C	T	0.2114	gene-ORF1ab	synonymous	2007 T	6021acC > acT	Low
7540	T	C	0.0558	gene-ORF1ab	synonymous	2425 T	7275acT > acC	Low
11,083	G	T	0.0607	gene-ORF1ab	missense	3606L > F	10818ttG > ttT	Moderate
14,408	C	T	0.8711	gene-ORF1ab	missense	4715P > L	14144cCt > cTt	Moderate
16,647	G	T	0.0555	gene-ORF1ab	synonymous	5461 T	16383acG > acT	Low
18,555	C	T	0.0583	gene-ORF1ab	synonymous	6097D	18291gaC > gaT	Low
18,877	C	T	0.0546	gene-ORF1ab	synonymous	6205L	18613Cta > Tta	Low
20,268	A	G	0.0587	gene-ORF1ab	synonymous	6668L	20004ttA > ttG	Low
21,614	C	T	0.1064	gene-S	missense	18L > F	52Ctt > Ttt	Moderate
22,227	C	T	0.2141	gene-S	missense	222A > V	665gCt > gTt	Moderate
22,992	G	A	0.0761	gene-S	missense	477S > N	1430aGc > aAc	Moderate
23,403	A	G	0.8664	gene-S	missense	614D > G	1841gAt > gGt	Moderate
24,334	C	T	0.0514	gene-S	synonymous	924A	2772gcC > gcT	Low
25,563	G	T	0.2164	gene-ORF3a	missense	57Q > H	171caG > caT	Moderate
26,801	C	G	0.2034	gene-M	synonymous	93L	279ctC > ctG	Low
27,944	C	T	0.1552	gene-ORF8	synonymous	17H	51caC > caT	Low
29,645	G	T	0.2110	gene-ORF10	missense	30 V > L	88Gta > Tta	Moderate

sequences of SARS-CoV-2 conveniently. The AutoVEM software package contains AutoVEM software script, installation and operation instructions, and examples of the input directory or file and the output files. The software can automatically output all SNVs information of the sequences passed the quality control, the information of the filtered or given sites, the linkage map of the filtered or given sites, the information of haplotypes, and the haplotype subgroup epidemic trends in various countries and regions over time.

3.2. Genome sequences

In total, 169,207 genome sequences of SARS-CoV-2 were downloaded from GISAID (https://www.epicov.org/) by November 30, 2020, thereinto a total of 131,576 genome sequences were passed at two steps of quality control, which were used for all subsequent analysis.

3.3. Linkage and haplotype analysis of the previous 9 specific sites

Linkage and haplotype analysis of the previous 9 specific sites showed that the 9 sites were still highly linked (Fig. 2), and 9 haplotypes with a frequency $\geq$ 1% were found and accounted for 99.16% of the total population (Table 1). Thereinto, 6 of them were found before January 23, 2020 and all of them were found before February 23, 2020 in different countries (Fig. 3), which indicated the complexity of SARS-CoV-2 evolution and spread at the early stage. The frequency and epidemic trend of haplotype subgroups of the 9 specific sites showed that H1, H5, H7, H9, H10 and H11 were prevalent in the world before November 30, 2020, and H1 with the frequency of 0.7184 is the most epidemic haplotype subgroup (Table 1, Fig. 3). However, H2 with a larger proportion and H3 and H4 with a smaller proportion at the early stage have almost disappeared at the present stage (Fig. 3). By carefully comparing the base composition of the 9 specific sites of H1, H5, H7, H9, H10 and H11 with those of H2, H3 and H4 (Table 1, Fig. 3), we still find that the 4 specific sites (C241T、C3037T、C14408T and A23403G) in Europe have an important influence on the viral infectivity, pathogenicity or host adaptability. Among the prevalent haplotype subgroups, H5, H7, H9 and H11 all had A23403G mutations, which indicated that the single A23403G mutation was related to infectivity, pathogenicity or host adaptability of SARS-CoV-2, while H10 had the other three specific mutations, including C241T, C3037T and C14408T, indicating that the combined mutations of these 3 sites also had a certain impact on infectivity, pathogenicity or host adaptability of SARS-CoV-2. However, H1 had these four mutations at the same time and showed an absolute epidemic advantage, which indicated that the simultaneous mutation of these four sites had a cumulative effect on infectivity, pathogenicity or host adaptability of SARS-CoV-2.

3.4. Linkage and haplotype analysis of the 23 sites with a frequency greater than 5%

A total of 23 SNVs with a frequency $\geq$ 5% were filtered from 131,576 SARS-CoV-2 genomes (Table 2). According to the linkage analysis of the 23 sites, it showed that not all of them were highly linked (Fig. S1). The 23 haplotypes with a frequency $\geq$ 1% were found and accounted for only 87.07% of the total population (Table S1). The frequency distribution of 23 haplotypes was relatively dispersed and the haplotype subgroup epidemic trends were complex (Table S1, Fig. S2), which was difficult to find the regular pattern.

Among the 23 sites, we found that 6 mutation sites (T445C, C6286T, C22227T, G25563T, C26801G and G29645T) with a frequency greater than 20% might be highly linked (Table 2, Fig. S1). Therefore, we performed linkage analysis of the 6 sites separately, and only 4 haplotypes with a frequency greater than 1% were found and accounted for 99.50% of the total population (Fig. 4), which suggested that the 6 sites were indeed highly linked. Among these 23 sites, except for the 4 specific sites (C241T、C3037T、C14408T and A23403G) of the previous H1 haplotype with mutation frequency greater than 0.8, only the above 6 mutation sites of highly linked had higher frequencies, which indicated that the mutations of these 10 sites were more significant at the present stage. Therefore, we constructed haplotypes by combining the previous 4 specific sites and the 6 new sites to reveal the landscape of virus continuous evolution (from the early to the present stage).

3.5. Linkage and haplotype analysis of the 10 sites

According to the linkage analysis and haplotype frequencies of the 10 sites (including the 4 specific sites of previous haplotype H1 and the 6 new sites), 11 haplotypes with a frequency $\geq$ 1% were

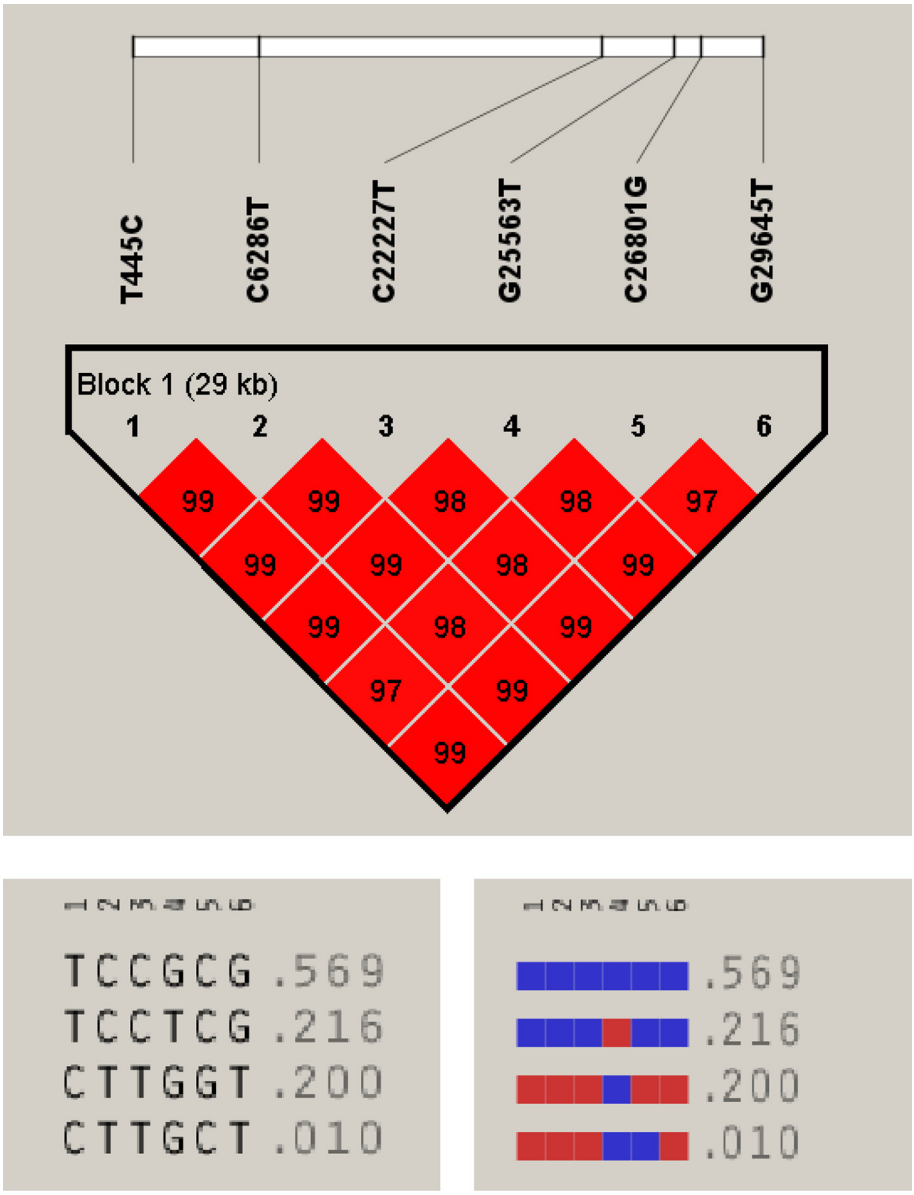

Fig. 4. Linkage analysis and haplotypes of the 6 new sites.

found and accounted for 95.14% of the total population (Fig. 5, Table 3). Among them, the three haplotypes (H1-1, H1-2 and H1-3) derived from the previous H1 accounted for 71.23% of the population (Table 3). Haplotype H1-2 with 5 new mutations (T445C, C6286T, C22227T, C26801G and G29645T) accounted for 19.36% of the population and appeared in the later stage (July 21, 2020), which showed a trend of increasing gradually (Table 3, Fig. 6). While haplotype H1-3 with only one new mutation (G25563T) accounted for 15.46% of the population and appeared in the early stage (February 7, 2020) (Table 3, Fig. 6). Besides, the G25563T mutation was also found in several other haplotypes H9-2, H5-2 and H7-2 (Table 3). The above haplotype subgroup epidemic trends showed that the mutation of 5 sites (T445C, C6286T, C22227T, C26801G and G29645T) or the single G25563T mutation may have some influence on infectivity, pathogenicity or host adaptability of SARS-CoV-2.

In general, the haplotype subgroups at the later stage are more complex and diverse than those at the earlier stage (Fig. 3, Fig. 6, Fig. S2), which may be related to the larger population and the more complex genome diversity. In addition, we observed that there are more complex haplotypes in the United States.

4. Discussion

According to linkage analysis and haplotype subgroup epidemic trends of the previous 9 specific sites for 131,576 genome sequences, we found that the frequency of H1 increased from 0.2880 (March 22, 2020), 0.4540 (April 6, 2020) and 0.6083 (May 10, 2020) at the early stage [9] to 0.7184 (November 30, 2020) at the present stage (Table 1). Moreover, both the single mutation of A23403G or the combined mutations of C241T, C3037T and C14408T could influence the infectivity, pathogenicity or host adaptability of SARS-CoV-2, which further confirmed the 4 specific sites (C241T, C3037T, C14408T and A23403G) were important in the previous H1 haplotype [9]. The mutation of A23403G located in S genes resulted in amino acid change of 614D > G (Table 2), which has been proved to be related to infectivity by several in vitro experiments [10–15]. Therefore, it is strongly

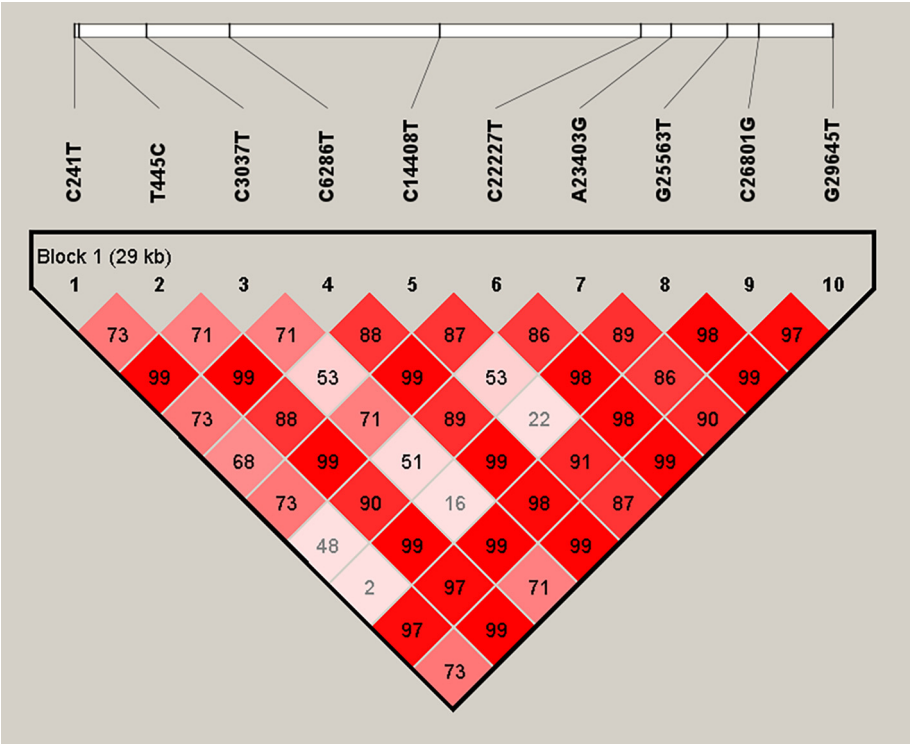

Fig. 5. Linkage analysis of the 10 sites.

Table 3
Haplotypes and frequencies of the 10 sites.

Name	Sequence	Frequency	Previous[1]
H1-1	TTTCTCGGCG	0.3641	H1
H1-2	TCTTTTGGGT	0.1936	NA
H1-3	TTTCTCGTCG	0.1546	NA
H2 (or H3 or H4)-1	CTCCCCAGCG	0.0759	H2 (or H3 or H4)
H5-1	TTTCCCGGCG	0.0165	H5
H5-2	TTTCCCGTCG	0.0134	NA
H7-1	CTTCTCGGCG	0.0181	H7
H7-2	CTTCTCGTCG	0.0113	NA
H9-1	CTCCTCGGCG	0.0300	H9
H9-2	CTCCTCGTCG	0.0245	NA
*H*10-1	TTTCTCAGCG	0.0494	*H*10
other	NA	0.0486	NA

[1]Table 1.

recommended that the A23403G mutation should be taken into account in the development of SARS-CoV-2 vaccines, especially RNA vaccines.

Since mutations in the virus genome occur randomly, the frequencies of most mutations in the population could not reach a certain level, and they should have no impact on infectivity, pathogenicity or host adaptability of virus. In the present study, we screened the SNVs with the mutation frequency of more than 5% and only 23 SNVs were screened in 131,576 SARS-CoV-2 genomes, which indicated the random mutation phenomena of SARS-CoV-2 genome. In the linkage and haplotype analysis of the 23 sites, we found the haplotypes of the 23 sites were complicated and could not find a specific haplotype subgroup trend (Table S1, Fig. S2), which suggested that the selection of 5% mutation frequency might not be appropriate at the present stage. Then we focused on the 6 mutations (T445C, C6286T, C22227T, G25563T, C26801G and G29645T) with a frequency greater than 20% among 23 sites. It was found that the 6 sites were highly linked and only 4 haplotypes with a frequency greater than 1% were found and accounted for 99.50% of the total population (Fig. 4). A few haplotype subgroups represent almost all of the population, indicating that the linkage analysis of appropriate sites can reduce haplotype subgroup complexity. Moreover, the frequencies of two mutated haplotypes were 0.2160 and 0.2000 (Fig. 4), suggesting that these 6 sites might be valuable. Therefore, in the practice of using our tool to screen the candidate key sites, we can adjust the setting frequency according to the situation.

Since the 4 specific sites and 6 new sites are more valuable at the early and present stage, the linkage analysis and haplotype subgroup epidemic trends of the 10 sites would reveal the landscape of virus continuous evolution (from the early to present stage). According to the haplotype subgroup epidemic trends and frequencies of the 10 sites, the previous H1 haplotype derived H1-2 and H1-3 haplotypes with new mutations and increasing trends, which indicated that the combined mutations of T445C, C6286T, C22227T, C26801G and G29645T at later stage or the single mutation of G25563T at earlier stage may have some influence on infectivity, pathogenicity or host adaptability of SARS-CoV-2. These 6 sites (T445C, C6286T, C22227T, G25563T, C26801G and G29645T) were located in ORF1ab, ORF1ab, S, ORF3a, M and ORF10 genes, respectively. Among them, only C22227T, G25563T and G29645T caused amino acid changes of S, ORF3a and ORF10 proteins (Table 2), which indicated that these 3 sites might have more important contributions to infectivity, pathogenicity or host adaptability of SARS-CoV-2. Thereby, their epidemic trend should be tracked in the future. Among the haplotypes of the 10 sites, H1-2 haplotype subgroup had 614D > G and 222A > V double mutation in S protein, which might be related to its rapid spread, should be further confirmed and verified.

In the latest report, a lineage B1.177 with high proportion was obtained by using 126,219 genomes generated by the COG-UK consortium [24]. This lineage had 222A > V (C22227T) mutation on the basis of 614D > G (A23403G) mutation, which was consistent with the results of the present study. Our results showed that the early

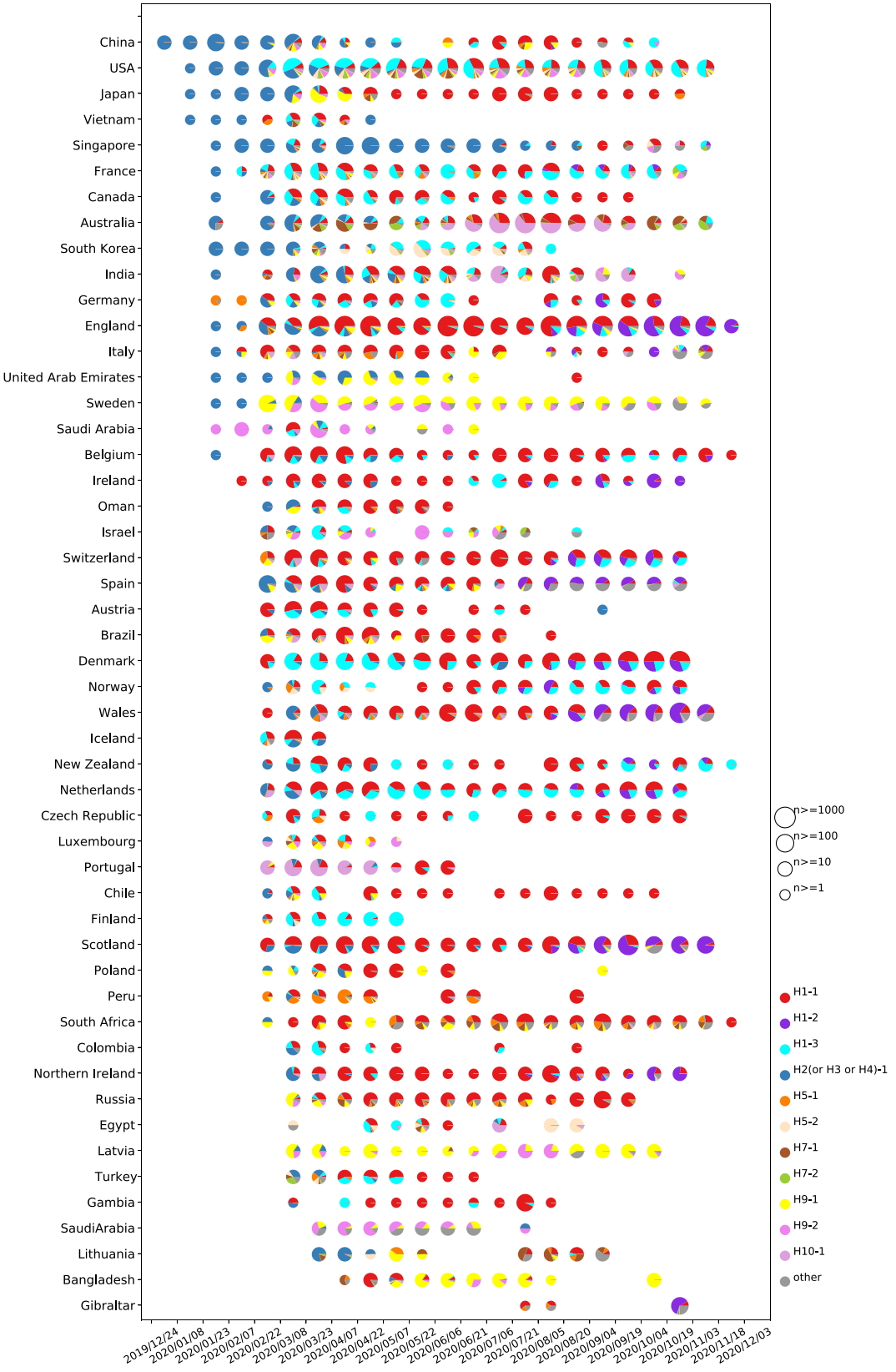

Fig. 6. Haplotype subgroup prevalence trends of the 10 sites. The numbers of haplotype subgroups of the 10 sites for 131,576 genomes with clear collection data detected in each country or region in chronological order. Countries or regions with a total number of genomes ≤ 100 were not shown in the figures.

H1 haplotype subgroup with the 4 highly linked sites (C241T, C3037T, C14408T and A23403G) derived H1-2 subgroup had another 5 highly linked sites (T445C, C6286T, C22227T, C26801G and G29645T). Therefore, the use of linkage analysis in this study can provide some highly linked co-mutation information, which combined with haplotype subgroup epidemic trends over time, can provide a more comprehensive assessment of which mutations may contribute significantly to infectivity, pathogenicity or host

adaptability of SARS-CoV-2. In addition, previous researchers identified several other mutations of N439K, Y453F and N501Y in the S protein, but we did not [24,25]. The possible reason is that the genome data analyzed was different. Besides, these mutations had a frequency of less than 5% according to their data [24,25], while our analysis filtered out mutation sites with a frequency of less than 5%. Among them, the frequency of N501Y mutation seems to reach 10% in the latest 28 days (13/11/2020–10/12/2020), whether the frequency of the mutation will increase over time should be monitored to further infer the significance of the mutation. Our tool can be utilized to analyze any local or public SARS-CoV-2 genomes for real-time monitoring virus mutations and epidemic trends.

In conclusion, the AutoVEM tool, which integrated screening SNVs of relatively high mutation frequency, linkage analysis and haplotype subgroup epidemic trends over time, could automatically complete the analysis of 169,207 initial genome sequences and 131,576 filtered genome sequences on a computer with a single core CPU and 2 GB RAM within 18 h. Through haplotype subgroup epidemic trends of 131,576 genome sequences, the significance of the previous 4 specific sites was further addressed, and 6 new highly linked mutation sites, which might be related to infectivity, pathogenicity or host adaptability of SARS-CoV-2, were found for the first time and should be further monitored in the future. We provide a new idea of combinatorial analysis and an automated tool to monitor haplotype subgroup epidemic trends and candidate key mutations in virus evolution in real-time and efficiently for the first time, which is of great significance for the development of new SARS-CoV-2 vaccine, the update of therapeutic drugs and detection methods in advance. At the same time, the idea of combinatorial analysis could also provide a reference for mutation monitoring of other viruses.

CRediT authorship contribution statement

Binbin Xi: Software, Validation, Data curation, Visualization, Investigation, Writing - original draft, Writing - review & editing. **Dawei Jiang:** Software, Writing - review & editing. **Shuhua Li:** Data curation, Writing - original draft, Writing - review & editing. **Jerome R. Lon:** Data curation, Writing - original draft. **Yunmeng Bai:** Data curation, Writing - original draft, Writing - review & editing. **Shudai Lin:** Writing - review & editing. **Meiling Hu:** Data curation, Writing - review & editing. **Yuhuan Meng:** Software. **Yimo Qu:** Data curation, Writing - review & editing. **Yuting Huang:** Data curation, Writing - review & editing. **Wei Liu:** Data curation, Writing - review & editing. **Lizhen Huang:** Writing - review & editing. **Hongli Du:** Conceptualization, Funding acquisition, Supervision, Writing - review & editing.

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.

Acknowledgments

We are very grateful to the GISAID initiative, database maintenance engineers and all their data contributors.

Authors' contributions

BX, DJ and YM developed the tool, BX, SL, YB and JL carried out the data analysis and wrote the manuscript. SL, MH, YQ, YH and WL collected data and revised the manuscript. LH revised the manuscript. HD conceived and supervised the study and revised the manuscript.

Availability

The developed AutoVEM software has been shared on the website (https://github.com/Dulab2020/AutoVEM) and can be available freely.

Funding

This work was supported by the National Key R&D Program of China (2018YFC0910201), the Key R&D Program of Guangdong Province (2019B020226001), the Science and the Technology Planning Project of Guangzhou (201704020176) and the Science and Technology Innovation Project of Foshan Municipality, China (2020001000431).

Data availability statement

All data relevant to the study are included in the article or uploaded as supplementary information.

Ethical Approval

Not required.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.04.002.

References

[1] COVID-19. Weekly Epidemiological Update. 2020.
[2] Hu B, Guo H, Zhou P, Shi Z. Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol 2020.
[3] Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. PNAS 2020;117(17):9241–3.
[4] Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev 2020;7(6):1012–23.
[5] van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 2020;83(104351).
[6] Fang S, Li K, Shen J, Liu S, Liu J, Yang L, Hu C, Wan J. GESS: a database of global evaluation of SARS-CoV-2/hCoV-19 sequences. Nucleic Acids Res 2020.
[7] Xing Y, Li X, Gao X, Dong Q. MicroGMT: a mutation tracker for SARS-CoV-2 and other microbial genome sequences. Front Microbiol 2020;11.
[8] Joshua B. Singer RJGM: CoV-GLUE: a web application for tracking SARS- CoV-2 genomic variation. 10.20944/preprints202006.0225.v1 2020.
[9] Bai Yunmeng, Jiang Dawei, Lon Jerome R, Chen Xiaoshi, Hu Meiling, Lin Shudai, et al. Comprehensive evolution and molecular characteristics of a large number of SARS-CoV-2 genomes reveal its epidemic trends. Int J Infect Dis 2020;100:164–73.
[10] Daniloski Z, Jordan TX, Ilmain JK, Guo X, Bhabha G, TenOeve BR, et al. The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. BioRxiv 2020.
[11] Jiang X, Zhang Z, Wang C, Ren H, Gao L, Peng H, Niu Z, Ren H, Huang H, Sun Q. Bimodular effects of D614G mutation on the spike glycoprotein of SARS-CoV-2 enhance protein processing, membrane fusion, and viral infectivity. Signal Transduct Target Ther 2020;5(2681).
[12] Fernández Ariel. Structural impact of mutation D614G in SARS-CoV-2 spike protein: enhanced infectivity and therapeutic opportunity. ACS Med Chem Lett 2020;11(9):1667–70.
[13] Li Qianqian, Wu Jiajing, Nie Jianhui, Zhang Li, Hao Huan, Liu Shuo, et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell 2020;182(5):1284–1294.e9.
[14] Zhang L, Jackson CB, Mou H, Ojha A, Rangarajan ES, Izard T, et al. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. BioRxiv 2020.
[15] Yurkovetskiy Leonid, Wang Xue, Pascal Kristen E, Tomkins-Tinch Christopher, Nyalile Thomas P, Wang Yetao, et al. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. Cell 2020;183(3):739–751.e8.

[16] Dearlove Bethany, Lewitus Eric, Bai Hongjun, Li Yifan, Reeves Daniel B, Joyce M Gordon, et al. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. PNAS 2020;117(38):23652–62.
[17] Ling Jiaxin, Hickman Rachel A, Li Jinlin, Lu Xi, Lindahl Johanna F, Lundkvist Åke, et al. Spatio-temporal mutational profile appearances of Swedish SARS-CoV-2 during the early pandemic. Viruses 2020;12(9):1026. https://doi.org/10.3390/v12091026.
[18] Aberer Andre J, Pattengale Nicholas D, Stamatakis Alexandros. Parallel computation of phylogenetic consensus trees. Procedia Comput Sci 2010;1(1):1065–73.
[19] Langmead Ben, Salzberg Steven L. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9(4):357–9.
[20] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAM tools. Bioinformatics 2009;25(16):2078–9.
[21] Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011;27(21):2987–93.
[22] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics 2011;27(15):2156–8.
[23] Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21(2):263–5.
[24] COG-UK update on SARS-CoV-2 Spike mutations of special interest. 2020.
[25] Sa K, Wt H, Rp D, Da C, Iatm F, Am C, et al. Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion H69/V70. BioRxiv 2020.