

AutoVEM2: a flexible automated tool to analyze candidate key mutations and epidemic trends for virus

Binbin Xi^{1#}, Shuhua Li^{1#}, Wei Liu¹, Dawei Jiang¹, Yunmeng Bai¹, Yimo Qu¹, Jerome Rumdon Lon¹, Lizhen Huang¹, Hongli Du^{1*}

¹School of Biology and Biological Engineering, South China University of Technology, Guangzhou 510006, China

Equal contribution.

* To whom correspondence should be addressed. Tel: 862039380667; Fax: 862039380667; Email: hldu@scut.edu.cn

Abstract

In our previous work, we developed an automated tool, AutoVEM, for real-time monitoring the candidate key mutations and epidemic trends of SARS-CoV-2. In this research, we further developed AutoVEM into AutoVEM2. AutoVEM2 is composed of three modules, including call module, analysis module, and plot module, which can be used modularly or as a whole for any virus, as long as the corresponding reference genome is provided. Therefore, it's much more flexible than AutoVEM. Here, we analyzed three existing viruses by AutoVEM2, including SARS-CoV-2, HBV and HPV-16, to show the functions, effectiveness and flexibility of AutoVEM2. We found that the N501Y locus was almost completely linked to the other 16 loci in SARS-CoV-2 genomes from the UK and Europe. Among the 17 loci, 5 loci were on the S protein and all of the five mutations cause amino acid changes, which may influence the epidemic traits of SARS-CoV-2. And some candidate key mutations of HBV and HPV-16, including T350G of HPV-16 and C659T of HBV, were detected. In

brief, we developed a flexible automated tool to analyze candidate key mutations and epidemic trends for any virus, which would become a standard process for virus analysis based on genome sequences in the future.

Keywords

virus; SARS-CoV-2; HBV; HPV-16; automated tool; candidate key mutations; haplotypes; epidemic trends

Highlights

1. An automatic tool to quickly analyze candidate key mutations and epidemic trends for any virus was developed.
2. Our integrated analysis method and tool could become a standard process for virus mutation and epidemic trend analysis based on genome sequences in the future.
3. N501Y with the other 16 highly linked mutation sites of SARS-CoV-2 in the UK and Europe were further confirmed, and some valuable mutation sites of HBV and HPV-16 were detected.

1. Introduction

SARS-CoV-2 has infected over 151,812,556 people and caused 3,186,817 deaths by 2 May 2021 [1]. At present, a variety of vaccines against SARS-CoV-2 are being used over the world, including mRNA-1273 [2], BNT162b2 [3], CoronaVac [4] and so on, hoping to form the effect of herd immunity. However, it is reported that N501Y mutation in the spike protein may reduce the neutralization sensitivity of antibodies, and finally influence the effectiveness of some vaccines [5]. Therefore, real-time monitoring the epidemic trend of SARS-CoV-2 mutations is of great significance to the update of detection reagents and vaccines. In our previous work, we found 9 candidate key mutations [6], including A23403G causing D614G amino acid change on the S protein, which has been proved to increase the infectivity of SARS-CoV-2 by several in vitro experiences [7-11]. With the further global spread of SARS-CoV-2, it

is difficult to prevent its mutation. Therefore, we proposed an innovative and integrative method that combines high-frequency mutation site screening, linkage analysis, haplotype typing and haplotype epidemic trend analysis to monitor the evolution of SARS-CoV-2 in real time. And we developed the whole process into an automated tool: AutoVEM [12]. We further found that the 4 highly linked sites (C241T, C3037T, C14408T and A23403G) of the previous 9 candidate key mutations have been almost fixed in the virus population, and the other 5 mutations disappeared gradually [12]. In addition, we found another 6 candidate key mutations with increased frequencies over time [12].

Our research on the trend of haplotype prevalence and other studies on the trend of single site prevalence both show that SARS-CoV-2 is constantly emerging new mutations, and the frequency of some mutations is increasing over time, while the frequency of some mutations is decreasing or even completely disappearing over time [6, 12, 13]. The consistent findings indicated that the integrative method we proposed is reliable. Moreover, the haplotype prevalence trend we used makes the new epidemic mutants less complicated. However, AutoVEM we developed is only for SARS-CoV-2 analysis. With the changes in the global natural environment, new and sudden infectious diseases are continuously emerging, such as the outbreak of SARS in Feb 2003 [14], MERS in 2012 [15], Ebola in 2014 [16], and the ZIKV in 2015 [17]. Therefore, we need a more flexible automated tool to identify and monitor the key mutation sites and evolution of various viruses.

In this research, we further developed AutoVEM into AutoVEM2. AutoVEM2 is composed of three different modules, including call module, analysis module and plot module. The call module can carry out quality control of genomes and find all SNVs for any virus genome sequences with various optional parameters. The analysis module can carry out candidate key mutations screening, linkage analysis, haplotype typing with optional parameters of mutation frequency and mutation sites. And the plot module can visualize the epidemic trends of haplotypes. The three modules can be used modularly or as a whole for any virus, as long as the corresponding reference genome is provided. Therefore, AutoVEM2 is much more flexible than AutoVEM.

Here, we analyzed 3 existing viruses by AutoVEM2, including SARS-CoV-2, HBV and HPV-16, to show its functions, effectiveness and flexibility. The SARS-CoV-2 genomes from the UK, Europe, and the USA were analyzed separately due to their large number of SARS-CoV-2 genomes in the GISAID. In addition to existing viruses, AutoVEM2 can also be used to analyze any virus that may appear in the future. We think our integrated analysis method and tool could become a standard process for virus mutation and epidemic trend analysis based on genome sequences in the future.

2. Materials and Methods

2.1. Functions of Three Modules of AutoVEM2

AutoVEM2 is a highly specialized, flexible, and modular pipeline for quickly monitoring the candidate key mutations, haplotype subgroups, and epidemic trends of different viruses by using virus whole genome sequences. It is written in Python language, in which Bowtie 2 [18], SAMtools [19], BCFtools [20], VCFtools [21] and Haploview [22] are used. AutoVEM2 consists of three modules, including call module, analysis module, and plot module, which can be used modularly or as a whole, and each module performs specific function(s) (Fig 1).

2.1.1. Call Module

The call module performs the function of finding all SNVs for all genome sequences. The input of the call module is a folder that stores formatted fasta format genome sequences. The call module processes are as follows:

1. Quality control of genome sequence according to four optional parameters: `--length`, `--number_n`, `--number_db`, and `--region_date_filter`.
2. Align the genome sequence to the corresponding reference sequence by Bowtie 2 v2.4.2 [18].
3. Call SNVs and INDELs by SAMtools v1.10 [19] and BCFtools v1.10.2 [20], resulting in a file named Variant Call Format (VCF) containing both SNVs and

INDELs information.

4. Further quality control of genome sequence according to the `--number_indels` optional parameter. Remove all INDELs from the sequence that has passed further quality control by VCFtools v0.1.16 [21], resulting in a VCF file that only contains SNVs information.
5. Merge SNVs for all genome sequences, resulting in a Tab-Separated Values (tsv) file named `snp_merged.tsv`.

2.1.2. Analysis Module

The analysis module performs three functions: screening out candidate key mutations, linkage analysis of these candidate key mutations, and acquisition haplotype of each genome sequence according to the result of linkage analysis. The input is the `snp_merged.tsv` file produced by the call module. The analysis module processes are as follows:

1. Count the mutation frequency of all mutation sites.
2. Screen out candidate key mutation sites according to the `--frequency` (default 0.05) optional parameter, and candidate key mutation sites can also be specified by the `--sites` optional parameter.
3. Nucleotides at these specific sites of each genome are extracted and organized according to the order of genome position.
4. Linkage analysis of these specific sites by Haploview v4.2 [22].
5. Acquire haplotypes using Haploview v4.2 [22]. Define the haplotype of each genome sequence according to the haplotype sequence, and if frequency of one haplotype <1%, it will be defined as “other”. This finally results in a tsv file named `data_plot.tsv`.

2.1.3. Plot Module

The plot module performs the function of visualizing epidemic trends of each haplotype in different countries or regions. The input of the plot module is the

data_plot.tsv file produced by the analysis module. The plot module processes are as follows:

1. Divide the whole time into different time periods according to the --days parameter.
2. Count the number of different haplotypes in each time period of different countries or regions.
3. Visualize the statistical results.

2.2. Genome Sequences Acquisition, Pre-processing, and Analyzing

SARS-CoV-2 whole genome sequences of the United Kingdom, Europe, and the United States were downloaded from GISAID between 01 Dec 2020 and 28 Feb 2021, resulting in 93,262, 161,703, and 40,405 genome sequences, respectively (Table 1). All HBV and HPV-16 nucleotide sequences, including whole genome sequences and fragments of whole genome, were downloaded from NCBI, resulting in 119,721 and 10,269 sequences, respectively (Table 1). Reference genome sequences of the three viruses were downloaded from NCBI (Table 1). The genome sequences were processed by in-house python script to make them meet the input format of AutoVEM2. Each formatted sequence consisted of two sections, the head section and the body section. The head section started with a greater than sign, followed by the virus name, sequence unique identifier, sequence collection time, and country or region where the sequence was collected, which were separated by vertical lines. And the body section was the nucleotide sequence.

For SARS-CoV-2, sequences with length < 29000, number of unknown bases > 15, number of degenerate bases > 50, number of indels > 2, or unclear collection time information or country information were filtered out [6, 12]. Finally, there were 79,269 sequences of the UK, 139,703 sequences of Europe, and 30,142 sequences of the USA (Table 1). All SNVs of these genomes were found by the call module. Mutation sites with mutation frequency ≥ 0.15 of the UK and Europe (in order to

include the five high linkage sites we found before [12]), and 0.25 of the USA would be as their candidate key mutation sites. Linkage analysis of these specific sites was performed and haplotype of each genome sequence was obtained by the analysis module. Epidemic trends of each haplotype were visualized by the plot module. (Table 1)

For HBV and HPV-16, sequences with length <90% and the number of unknown bases >1% the length of reference genomes were filtered out, resulting in 11,088 HBV genome sequences and 1,637 HPV-16 genome sequences. All SNVs of HBV and HPV-16 were found using the call module. Mutation sites with mutation frequency ≥ 0.25 of HBV and HPV-16 would be as the candidate key mutations. Linkage analysis of these specific sites was performed and haplotype of each genome sequence was obtained by the analysis module. (Table 1)

2.3. Variation Annotation

The candidate key mutation sites of SARS-CoV-2 in the UK, Europe, and the USA were annotated by an online tool of China National Center for Bioinformation (<https://bigd.big.ac.cn/ncov/online/tool/annotation?lang=en>), respectively. The candidate key mutation sites of HBV and HPV-16 were annotated manually.

3. Results

3.1. Overview of the analysis results of SARS-CoV-2, HBV, and HPV-16

The same 27 candidate key mutation sites were screened from the 79,269 SARS-CoV-2 (UK) and 139,703 SARS-CoV-2(Europe) genomes. Through linkage analysis of the 27 sites, it can be divided into 6 and 5 haplotypes with a proportion $\geq 1\%$ for the UK and Europe, respectively. 13 candidate key mutation sites were screened

from the 30,142 SARS-CoV-2(USA) genomes. Through linkage analysis of the 13 sites, the SARS-CoV-2 in the USA can be divided into 21 haplotypes with a proportion $\geq 1\%$. (Table 2)

7 of HBV and 12 of HPV-16 candidate key mutation sites were found from the 11,088 HBV genomes and 1,637 HPV-16 genomes, respectively. HBV and HPV-16 can be divided into 24 and 18 haplotypes with a proportion $\geq 1\%$ by the 7 sites and 12 sites, respectively. (Table 2)

3.2. Linkage and haplotype analysis of the 27 sites with a frequency ≥ 0.15 of SARS-CoV-2 in the United Kingdom and Europe

The detailed information for the 27 candidate key mutation sites screened from the UK and Europe was showed in Table 3. According to the linkage analysis, only 6 and 5 haplotypes with a frequency $\geq 1\%$ were found and accounted for 93.47% and 85.77% of SARS-CoV-2 population in the UK and Europe, respectively (Table 4), which showed highly linked among the 27 candidate key mutation sites (Fig 2A, Fig 2B).

For the UK, the 5 of 6 haplotypes (including H1-1-1, H1-2-1, H1-4-1, H1-4-2, and H1-4-3), which derived from H1 with previous 4 specific mutation sites (C241T, C3037T, C14408T, and A23403G) [6], accounted for 91.95% of the population (Table 4). H1-1-1 with only previous 4 specific mutation sites had almost disappeared in the UK by early 2021 (Fig 3). H1-2-1 with previous 4 specific mutation sites and the other 5 specific mutation sites (T445C, C6286T, C22227T, C26801G, and G29645T) appeared around July 21, 2020, became one of the major haplotypes circulating in the UK in early December 2020 [12], and gradually decreased, and there was only a very small population still circulating by late Feb, 2021 (Fig 3). While H1-4-1 with previous 4 specific mutation sites and another 17 specific mutation sites (C913T, C3267T, C5388A, C5986T, T6954C, C14676T, C15279T, T16176C, A23063T,

C23604A, C23709T, T24506G, G24914C, C27972T, G28048T, A28111G, and C28977T) with mutation frequencies around 0.78, and H1-4-2 with one more mutation site (A17615G) compared with H1-4-1 showed a trend of increasing gradually since early December, 2020. And H1-4-1 and H1-4-2 had become the dominant epidemic haplotypes in the UK by early February, 2021 (Fig 3). Notably, the H1-4-1 and H1-4-2 haplotypes both had A23063T mutation causing the N501Y mutation on the S protein, and the N501Y mutation was almost completely linked with the other 16 mutation sites (C913T, C3267T, C5388A, C5986T, T6954C, C14676T, C15279T, T16176C, C23604A, C23709T, T24506G, G24914C, C27972T, G28048T, A28111G, and C28977T). Among the 17 sites, 11 caused amino acid changes, of which 5 mutation sites were located on the S protein (including N501Y, P681H, T716I, S982A, and D1118H) (Table 3). This may influence the epidemic traits of SARS-CoV-2 and the effectiveness of vaccines, especially mRNA vaccines.

For Europe, the 5 haplotypes were the same as the 5 of 6 haplotypes of the UK (Table 4). Among the 5 haplotypes, 4 haplotypes (including H1-1-1, H1-2-1, H1-4-1, and H1-4-2) derived from H1 with previous 4 specific sites accounted for 84.67% of the population. And the epidemic trends of H1-1-1, H1-2-1, H1-4-1, and H1-4-2 were similar to those in the UK (Fig 4). That is, the H1-1-1 and H1-2-1 were gradually decreased, while the H1-4-1 and H1-4-2 were gradually increased.

3.3. Linkage and haplotype analysis of the 13 sites with a frequency ≥ 0.25 of SARS-CoV-2 in the USA

The detailed information for the 13 candidate key mutation sites screened from the USA was showed in Table 5. According to the linkage analysis, 21 haplotypes with a frequency $\geq 1\%$ were found and accounted for 87.94% of SARS-CoV-2 population in the USA (Table 6), which showed some degree linked among the 13 candidate key mutation sites (Fig 2C). Among the 21 haplotypes, H1-1-1, H1-3-2, and H1-3-3, with a frequency $>5\%$, all derived from H1 with previous 4 specific sites [6] (Table 6). H1-1-1 with previous 4 specific sites had a stable proportion (about 18%)

between December 1, 2020 and February 28, 2021 in the USA (Fig 5). H1-3-2 and H1-3-3 were derived from H1-3 directly, and H1-3 derived from H1 directly with one more mutation site (G25563T) compared with H1 [6, 12]. H1-3-2 had previous 5 specific sites (C241T, C3037T, C14408T, A23403G, and G25563T) [12] and C1059T (Table 5, Table 6), which had a stable prevalent trend between December 01, 2020 and February 02, 2021 in the USA (Fig 5). H1-3-3 had previous 5 specific sites and 8 new missense mutation sites (C1059T, C10319T, A18424G, C21304T, G25907T, C27964T, C28472T, and C28869T) (Table 5, Table 6), which increased gradually between December 01, 2020 and February 02, 2021 in the USA (Fig 5). In general, the haplotype subgroup diversity in the USA is much more complicated than those of in the UK and Europe.

3.4. Linkage and haplotype analysis of 7 sites with a frequency ≥ 0.25 of HBV

The detailed information for the 7 candidate key mutation sites screened from HBV genomes was showed in Table 7. 5 of the 7 sites were missense mutations, including 356S>A (T192G), 444S>P (T456C), 807D>V (A1546T), 10R>K (G2337A) on P gene, and 331A>V (C659T) on the S gene (Table 7). These 5 mutations were all on the P gene or the overlapping part of the P gene and other genes. Linkage analysis and haplotype analysis were performed and found 24 haplotypes with a proportion $\geq 1\%$, of which there was not a major haplotype, indicating that the 7 sites of HBV had a low degree of linkage(Fig S1A, Table S1).

3.5. Linkage and haplotype analysis of 12 sites with a frequency ≥ 0.25 of HPV-16

The detailed information for the 12 candidate key mutation sites screened from HPV-16 genomes was showed in Table 7. Among them, 8 specific mutations were missense mutation, including 83L>V (T350G) on the E6 gene, 219P>S (C3409T) on the E2 gene, 39I>L (A3977C) and 60I>V (A4040G) on the E5 gene, 43E>D (A4363T)

and 330L>F (A5224C) on the L2 gene, 228H>D (C6240G) and 292T>A (A6432G) on the L1 gene. Linkage analysis and haplotype analysis were performed on the 12 specific mutation sites and screened out 18 haplotypes with a proportion $\geq 1\%$ (Table S2), and the 12 specific sites showed a low degree of linkage (Fig S1B). Among the 18 haplotypes, there were 5 major haplotypes with a frequency $\geq 4\%$, including H1, H2, H3, H4, and H5. The haplotype H2 had 5 specific mutation sites (A2925G, T4226C, A4363T, G4936A, and A5224C). H4 has 9 specific mutation sites (A2925G, C3409T, A3977C, A4040G, A4363T, G4936A, A5224C, A6432G, and G7191T), and H3 had one more mutation site (T350G) compared with H4, while H1 had two more mutation sites (T350G and T4226C) compared with H4. (Table 7, Table S2)

Discussion

In this study, we developed a flexible tool to quickly monitor the candidate key mutations, haplotype subgroups, and epidemic trends for different viruses by using virus whole genome sequences, and analyzed a large number of SARS-CoV-2, HBV and HPV-16 genomes to show its functions, effectiveness and flexibility.

For the UK and Europe, we obtained the same 27 candidate key mutation sites, which could divide the SARS-CoV-2 population into 6 and 5 haplotypes, respectively. From the epidemic trend analysis, it showed that H1-4-1 and H1-4-2 with N501Y mutation on the S protein, which almost completely linked with the other 16 loci, had continued increasing from early December 2020 and became the dominant epidemic haplotypes in the United Kingdom and Europe by late February 2021. The B.1.1.7 mutant [23], corresponding to H1-4-1 and H1-4-2, has been reported that it has a more substantial transmission advantage based on several epidemiology researches [24, 25] and is greater in infectivity and adaptability [26]. Several studies have reported that the N501Y mutant reduces the neutralizing effect of the convalescent serum [27, 28], indicating that the N501Y variants may change neutralization sensitivity to reduce the effectiveness of the vaccine. Besides, the N501Y variants may reduce the

effectiveness of antibodies [29]. Therefore, we should pay continuous attention to the N501Y mutant, which is almost completely linked with the other 16 loci.

For HPV-16 and HBV, we also screened out multiple specific sites which may be related to infectivity. For HPV-16, the T350G (83L>V) mutation we detected is the most common mutation on the E6 gene of HPV-16 [30-32]. Several studies have shown that the T350G mutant may cause persistent virus infection and further increase cancer risk [31-34]. It is reported that T350G variants can down-regulate the expression of E-cadherin, which is an adhesion protein that acts cell-cell adhesion. E-cadherin down-regulation can reduce the adhesion between cells, allowing infected cells to escape the host's immune surveillance, and increase the risk of continued virus infection and the risk of cancer [33]. The C3410T mutation we detected, on the E2 gene of HPV-16, is also one of the common mutations of HPV-16 [35, 36]. Furthermore, The A2926G mutation we detected has been reported due to a reference genome sequencing error [37, 38]. For HBV, the C659T mutation, which causes A331V mutation on S gene, is reported to be associated with increasing the efficiency of HBV replication [39].

Due to the continuous mutations and evolution of viruses, it should be carefully considered whether the new mutations have an influence on developing and updating vaccines. AutoVEM2 provides a fast and reliable process of continuously monitoring candidate key mutations and epidemic trends of these mutations. Through AutoVEM2, we have analyzed a large number of SARS-CoV-2, HBV, and HPV-16 genomes and obtained some candidate key mutation sites fast and effectively. Among them, some mutations, such as D614G and N501Y of SARS-CoV-2, T350G of HBV, and C659T of HVP-16, have been proved to play an important role in the viruses, indicating the reliability and effectiveness of AutoVEM2. In total, we developed a flexible automatic tool for monitoring candidate key mutations and epidemic trends for any virus. It can be used in the study of mutations and epidemic trends analysis of existing viruses, and can be also used in analyzing the virus that may appear in the future. Our integrated analysis method and tool could become a standard process for virus mutation and epidemic trend analysis based on genome sequences in the future.

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.

Authors' contributions

BX developed the tool, carried out the data analysis, and wrote the manuscript. SL collected the data and wrote the manuscript. WL collected the data. DW, YB, YQ, RL, and LH revised the manuscript. HD conceived and supervised the study and revised the manuscript.

Availability

The developed AutoVEM2 software has been shared on the website (<https://github.com/Dulab2020/AutoVEM2>) and can be freely available.

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

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

Ethical Approval

Not required.

References:

1. WHO: **COVID-19 Weekly Epidemiological Update.**
<https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---4-may-2021>.
2. Jackson LA, Anderson EJ, Roupael NG, Roberts PC, Makhene M, Coler RN, McCullough MP, Chappell JD, Denison MR, Stevens LJ *et al*: **An mRNA Vaccine against SARS-CoV-2 — Preliminary Report.** *The New England journal of medicine* 2020, **383**(20):1920-1931.
3. Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, Perez JL, Perez Marc G, Moreira ED, Zerbini C *et al*: **Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine.** *NEW ENGL J MED* 2020, **383**(27):2603-2615.
4. Zhang Y, Zeng G, Pan H, Li C, Hu Y, Chu K, Han W, Chen Z, Tang R, Yin W *et al*: **Safety, tolerability, and immunogenicity of an inactivated SARS-CoV-2 vaccine in healthy adults aged 18–59 years: a randomised, double-blind, placebo-controlled, phase 1/2 clinical trial.** *The Lancet infectious diseases* 2021, **21**(2):181-192.
5. Kai Wu APWJ, Stewart-Jones HBSB, Andrea Carfi KSCR: **mRNA-1273 vaccine induces neutralizing antibodies against spike mutants from global SARS-CoV-2 variants.** *bioRxiv preprint doi: 10.1101/2021.01.25.427948*. 2021.
6. Bai Y, Jiang D, Lon JR, Chen X, Hu M, Lin S, Chen Z, Wang X, Meng Y, Du H: **Comprehensive evolution and molecular characteristics of a large number of SARS-CoV-2 genomes reveal its epidemic trends.** *INT J INFECT DIS* 2020, **100**:164-173.
7. Daniloski Z, Jordan TX, Ilmain JK, Guo X, Bhabha G, TenOever BR, Sanjana NE: **The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types.** *ELIFE* 2021, **10**.
8. Fernández A: **Structural Impact of Mutation D614G in SARS-CoV-2 Spike Protein: Enhanced Infectivity and Therapeutic Opportunity.** *ACS MED CHEM LETT* 2020, **11**(9):1667-1670.
9. 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 TRANSDUCTION AND TARGETED THERAPY* 2020, **5**(2681).
10. Zhang L, Jackson CB, Mou H, Ojha A, Rangarajan ES, Izard T, Farzan M, Choe H: **The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity.** *bioRxiv preprint doi: <https://doi.org/10.1101/2020.06.12.148726>*. 2020.
11. Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, Zhao C, Zhang Q, Liu H, Nie L *et al*: **The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity.** *CELL* 2020, **182**(5):1284-1294.
12. Xi B, Jiang D, Li S, Lon JR, Bai Y, Lin S, Hu M, Meng Y, Qu Y, Huang Y *et al*: **AutoVEM: An automated tool to real-time monitor epidemic trends and key mutations in SARS-CoV-2 evolution.** *COMPUT STRUCT BIOTEC* 2021, **19**:1976-1985.
13. 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* 2021, **49**(D1):D706-D714.
14. Zhong NS, Zheng BJ, Li YM, Poon, Xie ZH, Chan KH, Li PH, Tan SY, Chang Q, Xie JP *et al*: **Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's**

Republic of China, in February, 2003. *LANCET* 2003, **362**(9393):1353-1358.

15. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA: **Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia.** *N Engl J Med* 2012, **367**(19):1814-1820.
16. Coltart CE, Lindsey B, Ghinai I, Johnson AM, Heymann DL: **The Ebola outbreak, 2013-2016: old lessons for new epidemics.** *Philos Trans R Soc Lond B Biol Sci* 2017, **372**(1721).
17. Heukelbach J, Alencar CH, Kelvin AA, De Oliveira WK, Pamplona De Góes Cavalcanti L: **Zika virus outbreak in Brazil.** *The Journal of Infection in Developing Countries* 2016, **10**(02):116-120.
18. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *NAT METHODS* 2012, **9**(4):357-359.
19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *BIOINFORMATICS* 2009, **25**(16):2078-2079.
20. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.** *BIOINFORMATICS* 2011, **27**(21):2987-2993.
21. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST *et al*: **The variant call format and VCFtools.** *BIOINFORMATICS* 2011, **27**(15):2156-2158.
22. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *BIOINFORMATICS* 2005, **21**(2):263-265.
23. COG-UK: **COG-UK update on SARS-CoV-2 Spike mutations of special interest.** https://www.attogene.com/wp-content/uploads/2020/12/Report-1_COG-UK_19-December-2020_SARS-CoV-2-Mutations.pdf.
24. Zhao S, Lou J, Cao L, Zheng H, Chong MKC, Chen Z, Chan RWY, Zee BCY, Chan PKS, Wang MH: **Quantifying the transmission advantage associated with N501Y substitution of SARS-CoV-2 in the UK: an early data-driven analysis.** *J TRAVEL MED* 2021, **28**(2).
25. Leung K, Shum MH, Leung GM, Lam TT, Wu JT: **Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020.** *Euro surveillance : bulletin européen sur les maladies transmissibles* 2021, **26**(1):1.
26. Hu J, Peng P, Wang K, Fang L, Luo F, Jin A, Liu B, Tang N, Huang A: **Emerging SARS-CoV-2 variants reduce neutralization sensitivity to convalescent sera and monoclonal antibodies.** *CELL MOL IMMUNOL* 2021, **18**(4):1061-1063.
27. Xie X, Liu Y, Liu J, Zhang X, Zou J, Fontes-Garfias CR, Xia H, Swanson KA, Cutler M, Cooper D *et al*: **Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K and N501Y variants by BNT162b2 vaccine-elicited sera.** *NAT MED* 2021, **27**(4):620-621.
28. Rees-Spear C, Muir L, Griffith SA, Heaney J, Aldon Y, Snitselaar JL, Thomas P, Graham C, Seow J, Lee N *et al*: **The effect of spike mutations on SARS-CoV-2 neutralization.** *CELL REP* 2021, **34**(12):108890.
29. Wang Z, Schmidt F, Weisblum Y, Muecksch F, Barnes CO, Finkin S, Schaefer-Babajew D, Cipolla M, Gaebler C, Lieberman JA *et al*: **mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants.** *NATURE* 2021, **592**(7855):616-622.
30. Hang D, Yin Y, Han J, Jiang J, Ma H, Xie S, Feng X, Zhang K, Hu Z, Shen H *et al*: **Analysis of human papillomavirus 16 variants and risk for cervical cancer in Chinese population.** *VIROLOGY* 2016, **488**:156-161.
31. Escobar-Escamilla N, González-Martínez BE, Araiza-Rodríguez A, Fragosó-Fonseca DE,

- Pedroza-Torres A, Landa-Flores MG, Garcés-Ayala F, Mendieta-Condado E, Díaz-Quinonez JA, Castro-Escarpulli G *et al*: **Mutational landscape and intra-host diversity of human papillomavirus type 16 long control region and E6 variants in cervical samples.** *ARCH VIROL* 2019, **164**(12):2953-2961.
32. Tan G, Duan M, Li YE, Zhang N, Zhang W, Li B, Qu P: **Distribution of HPV 16 E6 gene variants in screening women and its associations with cervical lesions progression.** *VIRUS RES* 2019, **273**:197740.
33. Togtema M, Jackson R, Richard C, Niccoli S, Zehbe I: **The human papillomavirus 16 European-T350G E6 variant can immortalize but not transform keratinocytes in the absence of E7.** *VIROLOGY* 2015, **485**:274-282.
34. Zhang L, Liao H, Yang B, Geffre CP, Zhang A, Zhou A, Cao H, Wang J, Zhang Z, Zheng W: **Variants of human papillomavirus type 16 predispose toward persistent infection.** *INT J CLIN EXP PATHO* 2015, **8**(7):8453-8459.
35. Kahla S, Hammami S, Kochbati L, Chanoufi MB, Oueslati R: **HPV16 E2 variants correlated with radiotherapy treatment and biological significance in cervical cell carcinoma.** *Infection, Genetics and Evolution* 2018, **65**:238-243.
36. Lee K, Magalhaes I, Clavel C, Briolat J, Birembaut P, Tommasino M, Zehbe I: **Human papillomavirus 16 E6, L1, L2 and E2 gene variants in cervical lesion progression.** *VIRUS RES* 2008, **131**(1):106-110.
37. Arroyo-Mühr LS, Lagheden C, Hultin E, Eklund C, Adami H, Dillner J, Sundström K: **Human papillomavirus type 16 genomic variation in women with subsequent in situ or invasive cervical cancer: prospective population-based study.** *BRIT J CANCER* 2018, **119**(9):1163-1168.
38. Meissner, J.1997. Sequencing errors in reference HPV clones, p. III-110–III-123. In G. Myers, C. Baker, K. Munger, F. Sverdup, A. McBride, H.-U. Bernard, and J. Meissner (ed.), Human papillomaviruses 1997: a compilation and analysis of nucleic acid and amino acid sequences. Theoretical biology and biophysics. Los Alamos National Laboratory, Los Alamos, N.M.
39. Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou K: **An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation.** *J THEOR BIOL* 2005, **235**(4):555-565.

Tables and Figures Legends

Figure 1: Functions and optional parameters of three modules of AutoVEM2

Figure 2: Linkage analysis results of SARS-CoV-2.

A. Linkage analysis of 27 candidate key mutation sites of SARS-CoV-2 (UK);

B. Linkage analysis of 27 candidate key mutation sites of SARS-CoV-2 (Europe);

C. Linkage analysis of 13 candidate key mutation sites of SARS-CoV-2 (USA)

Figure 3: Epidemic trends of 6 haplotypes of 93,262 SARS-CoV-2 genomes from the UK

Figure 4: Epidemic trends of 5 haplotypes of 139,703 SARS-CoV-2 genomes from Europe. Countries or regions with a total number of genomes ≤ 100 were not shown in the figure.

Figure 5: Epidemic trends of 21 haplotypes of 30,142 SARS-CoV-2 genomes from the USA

Figure S1: Linkage analysis of candidate key mutation sites of HBV and HPV-16.

- A. Linkage analysis of 7 candidate key mutation sites of HBV;
- B. Linkage analysis of 12 candidate key mutation sites of HPV-16

Table 1: Information of SARS-CoV-2, HBV, and HPV-16 genomes and the analysis process of the three viruses

Table 2: Candidate key mutation sites and haplotypes results of SARS-CoV-2, HBV, and HPV-16

Table 3: The annotation of the 27 sites of SARS-CoV-2(UK and Europe) with a mutation frequency $\geq 15\%$

Table 4: Haplotypes and their frequencies of the 27 sites of SARS-CoV-2(UK and Europe)

Table 5: The annotation of the 13 sites of SARS-CoV-2(USA) with a mutation frequency $\geq 25\%$

Table 6: Haplotypes and their frequencies of the 13 sites of SARS-CoV-2(USA)

Table 7: The annotation of the 7 sites of HBV and 12 sites of HPV-16 with mutation frequency $\geq 25\%$

Table S1: Haplotypes and their frequencies of the 7 sites of HBV

Table S2: Haplotypes and their frequencies of the 12 sites of HPV-16

AutoVEM2

Formatted fasta format whole genome sequences



Call module

Functions

- Quality control of genomes
- Find all SNVs of all genomes

Optional parameters

- Length of genome
- Number of unknown bases
- Number of degenerate bases
- Number of INDELs
- Clear country or collection date information

snp_merged.tsv

Analysis module

Functions

- Screen out candidate key mutations
- Linkage analysis
- Acquisition haplotype of each genome sequence

Optional parameters

- Mutation frequency
- Mutation sites

data_plot.tsv

Plot module

Function

- Visualize epidemic trends of each haplotypes in different countries or regions

Optional parameters

- The time interval

Table 1. Information of SARS-CoV-2, HBV, and HPV-16 genomes and the analysis process of the three viruses

Virus	Sequences Collection Date	Database	Number of Downloaded Genomes	Number of Filtered Genomes ¹	Reference Sequence	Find all SNVs	Screen Out Candidate Key Mutation Sites	Linkage Analysis and Acquire Haplotypes	Epidemic Trends of Haplotypes
SARS-CoV-2 (UK)	2020.12.01 - 2021.02.28	GISAID	93,262	79,269	NC_045512.2	yes	yes	yes	yes
SARS-CoV-2 (Europe)	2020.12.01 - 2021.02.28	GISAID	161,703	139,703	NC_045512.2	yes	yes	yes	yes
SARS-CoV-2 (USA)	2020.12.01 - 2021.02.28	GISAID	40,405	30,142	NC_045512.2	yes	yes	yes	yes
HBV	- 2021.01.25	NCBI	119,721	11,088	NC_003977.2	yes	yes	yes	no
HPV-16	- 2021.01.25	NCBI	10,269	1,637	K02718	yes	yes	yes	no

¹ Filtered criteria: the genomic sequences with more than 90% full length and less than 1% N were retained for HBV and HPV-6, while the filtered criteria for SARS-CoV-2 genome was referred to AutoVEM^[12]

Table 2. Candidate key mutation sites and haplotypes results of SARS-CoV-2, HBV, and HPV-16

Virus	Number of Candidate Key Mutation Sites	Candidate Key Mutation Sites	Number of Haplotypes ¹
SARS-CoV-2 (UK)	27	C241T T445C C913T C3037T C3267T C5388A C5986T C6286T T6954C C14408T C14676T C15279T T16176C A17615G C22227T A23063T A23403G C23604A C23709T T24506G G24914C C26801G C27972T G28048T A28111G C28977T G29645T	6
SARS-CoV-2 (Europe)			5
SARS-CoV-2 (USA)	13	C241T C1059T C3037T C10319T C14408T A18424G C21304T A23403G G25563T G25907T C27964T C28472T C28869T	21
HBV	7	T192G T456C C659T C669T A1546T G2337A G2479A	24
HPV-16	12	T350G A2925G C3409T A3977C A4040G T4226C A4363T G4936A A5224C C6240G A6432G G7191T	18

¹Haplotypes with a proportion $\geq 1\%$

Table 3. The annotation of the 27 sites of SARS-CoV-2(UK and Europe) with a mutation frequency $\geq 15\%$

Position	Ref	Alt	Frequency UK ¹	Frequency Europe ²	Gene Region	Mutation Type	Protein Changed	Codon Changed	Impact
241	C	T	0.9659	0.9552	5'UTR	upstream	NA	NA	MODIFIER
445	T	C	0.1801	0.2394	gene-orf1ab	synonymous	60V	180gtT>gtC	LOW
913	C	T	0.7831	0.5692	gene-orf1ab	synonymous	216S	648tcC>tcT	LOW
3,037	C	T	0.9779	0.9737	gene-orf1ab	synonymous	924F	2772ttC>ttT	LOW
3,267	C	T	0.7892	0.5794	gene-orf1ab	missense	1001T>I	3002aCt>aTt	MODERATE
5,388	C	A	0.7881	0.5738	gene-orf1ab	missense	1708A>D	5123gCt>gAt	MODERATE
5,986	C	T	0.7891	0.5827	gene-orf1ab	synonymous	1907F	5721ttC>ttT	LOW
6,286	C	T	0.1813	0.2421	gene-orf1ab	synonymous	2007T	6021acC>acT	LOW
6,954	T	C	0.7896	0.5799	gene-orf1ab	missense	2230L>T	6689aTa>aCa	MODERATE
14,408	C	T	0.9718	0.9679	gene-orf1ab	missense	4715P>L	14144cCt>cTt	MODERATE
14,676	C	T	0.7862	0.5747	gene-orf1ab	synonymous	4804P	14412ccC>ccT	LOW
15,279	C	T	0.7904	0.5801	gene-orf1ab	synonymous	5005H	15015caC>caT	LOW
16,176	T	C	0.7862	0.5745	gene-orf1ab	synonymous	5304T	15912acT>acC	LOW
17,615	A	G	0.2579	0.1790	gene-orf1ab	missense	5784K>R	17351aAg>aGg	MODERATE
22,227	C	T	0.1810	0.2439	gene-S	missense	222A>V	665gCt>gTt	MODERATE
23,063	A	T	0.7860	0.5777	gene-S	missense	501N>Y	1501Aat>Tat	MODERATE
23,403	A	G	0.9914	0.9770	gene-S	missense	614D>G	1841gAt>gGt	MODERATE
23,604	C	A	0.7913	0.5829	gene-S	missense	681P>H	2042cCt>cAt	MODERATE
23,709	C	T	0.7854	0.5748	gene-S	missense	716T>I	2147aCa>aTa	MODERATE
24,506	T	G	0.7858	0.5740	gene-S	missense	982S>A	2944Tca>Gca	MODERATE
24,914	G	C	0.7847	0.5739	gene-S	missense	1118D>H	3352Gac>Cac	MODERATE
26,801	C	G	0.1739	0.2365	gene-M	synonymous	93L	279ctC>ctG	LOW
27,972	C	T	0.7725	0.5630	gene-ORF8	stop	27Q>*	79Caa>Taa	HIGH
28,048	G	T	0.7862	0.5695	gene-ORF8	missense	52R>I	155aGa>aTa	MODERATE
28,111	A	G	0.7834	0.5716	gene-ORF8	missense	73Y>C	218tAc>tGc	MODERATE
28,977	C	T	0.7746	0.5690	gene-N	missense	235S>F	704tCt>tTt	MODERATE
29,645	G	T	0.1807	0.2370	gene-ORF10	missense	30V>L	88Gta>Tta	MODERATE

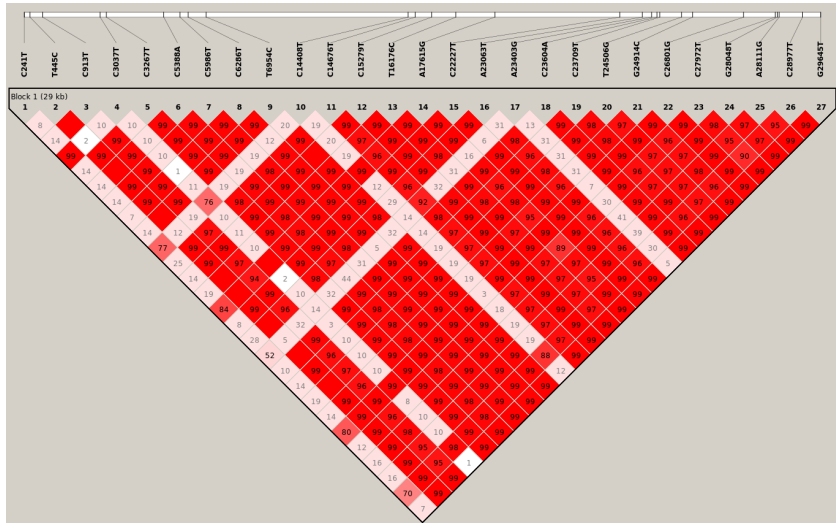
¹Mutation frequency of the 27 sites of 79,269 SARS-CoV-2 genomes from the UK

²Mutation frequency of the 27 sites of 139,703 SARS-CoV-2 genomes from Europe

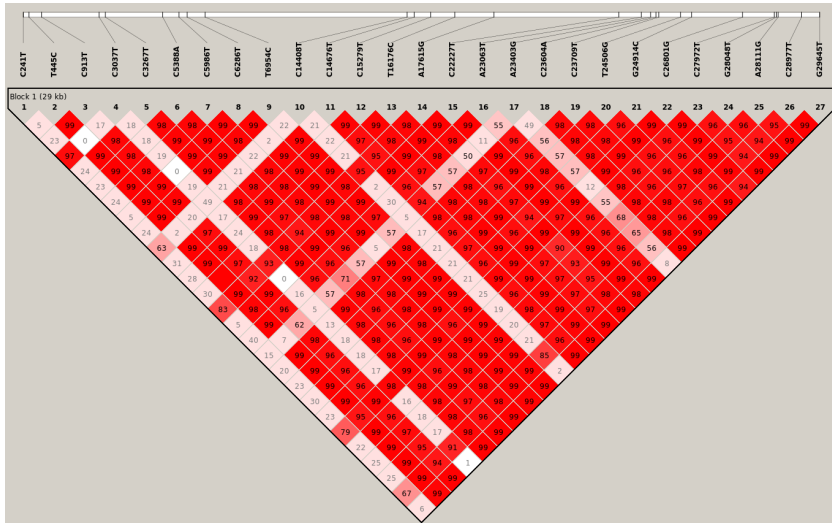
Table 4. Haplotypes and their frequencies of the 27 sites of SARS-CoV-2(UK and Europe)

Country or Region	Name	Sequence	Frequency	Corresponding to the UK ^[23]
	reference	CTCCCCCTCCCTACAACCTGCCGACG	NA	NA
The UK	H1-1-1	TTCTCCCCTTCCTACAGCCTGCCGACG	0.0219	B.1
	H1-2-1	TCCTCCCTTTCCTATAGCCTGGCGACT	0.1621	B.1.177
	H1-4-1	TTTTTATCCTTTCCTGATGCCTTGTG	0.4834	B.1.1.7
	H1-4-2	TTTTTATCCTTTCGCTGATGCCTTGTG	0.2421	B.1.1.7
	H1-4-3	TTTTTATCCTTTCCTGATGCCTTGTG	0.0100	B.1.1.7
	H2(or H3 or H4)-1-2	CTTCTATCCCTTCACTGATGCCTTGCG	0.0152	NA
	other	NA	0.0653	NA
Europe	H1-1-1	TTCTCCCCTTCCTACAGCCTGCCGACG	0.1302	B.1
	H1-2-1	TCCTCCCTTTCCTATAGCCTGGCGACT	0.2093	B.1.177
	H1-4-1	TTTTTATCCTTTCCTGATGCCTTGTG	0.3475	B.1.1.7
	H1-4-2	TTTTTATCCTTTCGCTGATGCCTTGTG	0.1597	B.1.1.7
	H2(or H3 or H4)-1-2	CTTCTATCCCTTCACTGATGCCTTGCG	0.0110	NA
	other	NA	0.1423	NA

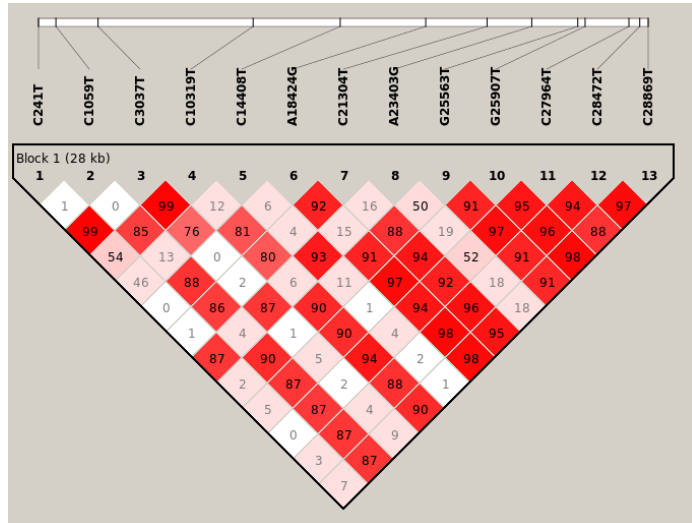
A

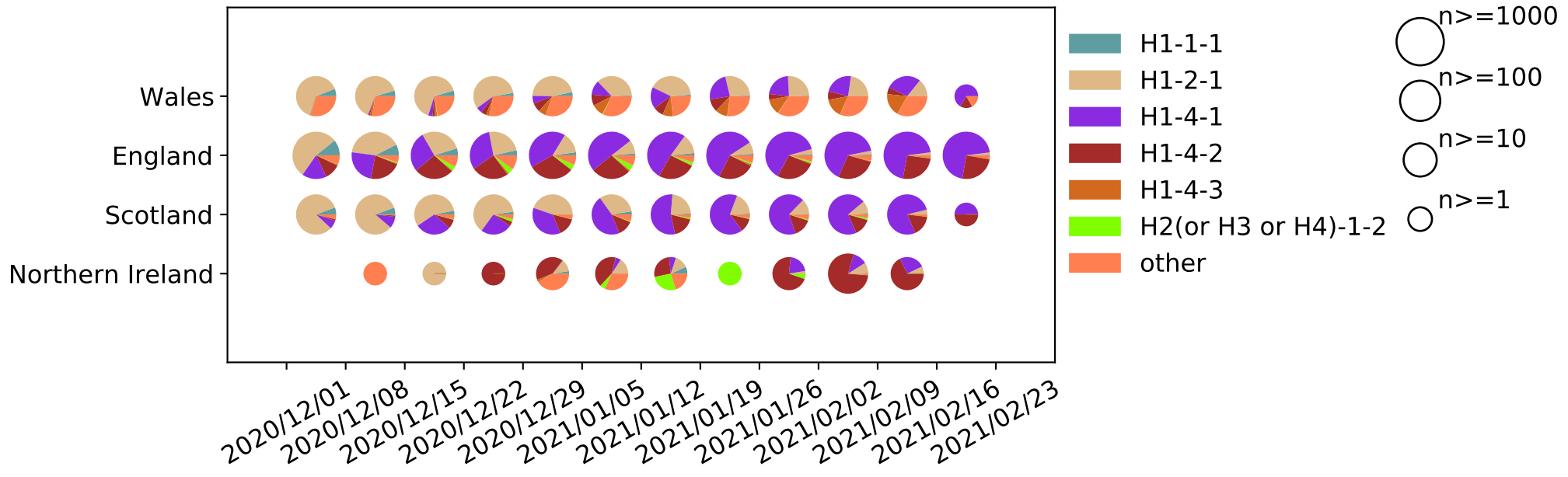


B



C





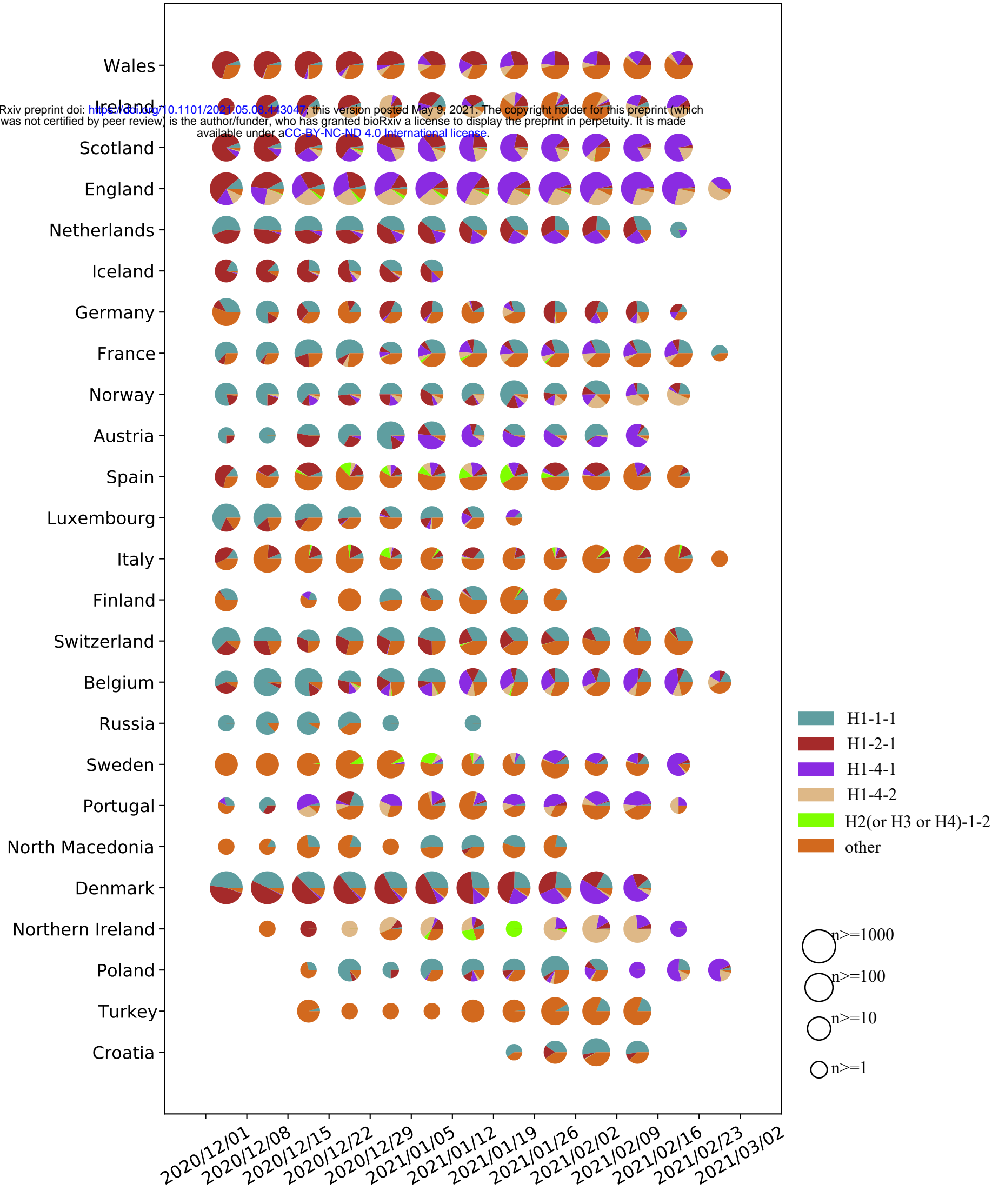


Table 5. The annotation of the 13 sites of SARS-CoV-2(USA) with a mutation frequency $\geq 25\%$

Position	Ref	Alt	Frequency	Gene Region	Mutation Type	Protein Changed	Codon Changed	Impact
241	C	T	0.7720	5'UTR	upstream	NA	NA	MODIFIER
1,059	C	T	0.6667	orf1ab	missense	265T>I	794aCc>aTc	MODERATE
3,037	C	T	0.9117	orf1ab	synonymous	924F	2772ttC>ttT	LOW
10,319	C	T	0.4435	orf1ab	missense	3352L>F	10054Ctt>Ttt	MODERATE
14,408	C	T	0.8505	orf1ab	missense	4715P>L	14144cCt>cTt	MODERATE
18,424	A	G	0.4625	orf1ab	missense	6054N>D	18160Aat>Gat	MODERATE
21,304	C	T	0.4593	orf1ab	missense	7014R>C	21040Cgc>Tgc	MODERATE
23,403	A	G	0.9454	S	missense	614D>G	1841gAt>gGt	MODERATE
25,563	G	T	0.6850	ORF3a	missense	57Q>H	171caG>caT	MODERATE
25,907	G	T	0.4846	ORF3a	missense	172G>V	515gGt>gTt	MODERATE
27,964	C	T	0.5072	ORF8	missense	24S>L	71tCa>tTa	MODERATE
28,472	C	T	0.4827	N	missense	67P>S	199Cct>Tct	MODERATE
28,869	C	T	0.5029	N	missense	199P>L	596cCa>cTa	MODERATE

Table 6. Haplotypes and their frequencies of the 13 sites of SARS-CoV-2(USA)

Name	Sequence	Frequency
reference	CCCCACAGGCC	NA
H1-1-1	TCTCTACGGGCC	0.1820
H1-3-1	TCTCTACGTGCCC	0.0141
H1-3-2	TTTCTACGTGCCC	0.0920
H1-3-3	TTTTTGTGTTTT	0.2724
H1-3-4	TCTTTGTGTTTT	0.0135
H1-3-5	TTTCTACGTGCCT	0.0134
H1-3-6	TTTTTATGTTTT	0.0103
H1-3-7	TTTTTACGTGTCC	0.0102
H5-1-1	TCTCCACGGGCC	0.0149
H5-2-1	TTTCCACGTGCCC	0.0138
H5-2-2	TTTTCGTGTTTT	0.0279
H7-1-1	CCTCTACGGGCC	0.0223
H7-2-2	CTTCTGTGTTTT	0.0197
H9-1-1	CCCCTACGGGCC	0.0213
H9-2-2	CTCCTACGTGCCC	0.0122
H9-2-3	CTCCTGTGTTTT	0.0347
H10-1-1	TCTCTACAGGCC	0.0172
H10-1-2	TTTTTGTATTTTT	0.0102
H11-1-1	CCTCCACGGGCC	0.0227
H11-2-1	CTTCCACGTGCCC	0.0137
H11-2-2	CTTTCGTGTTTT	0.0409
other	NA	0.1206

USA

2020/12/01
2020/12/08
2020/12/15
2020/12/22
2020/12/29
2021/01/05
2021/01/12
2021/01/19
2021/01/26
2021/02/02
2021/02/09
2021/02/16
2021/02/23
2021/03/02

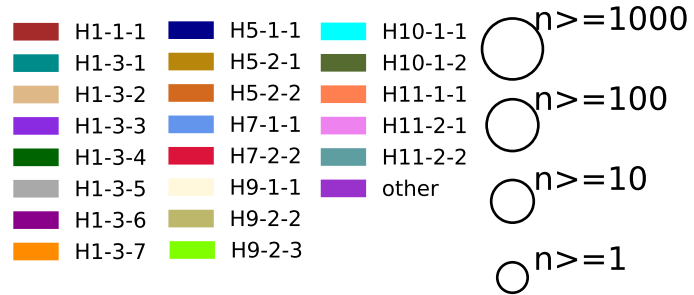


Table 7. The annotation of the 7 sites of HBV and 12 sites of HPV-16 with a mutation frequency $\geq 25\%$

Virus	Position	Ref	Alt	Frequency	Gene Region ¹	Mutation Type	Protein Changed	Codon Changed
HBV	192	T	G	0.2804	P, S	P: missense S: synonymous	P: 356S>A S: 175L	P: 356Tct>Gct S: 175ctT>ctG
	456	T	C	0.2750	P, S	P: missense S: synonymous	P: 444S>P S: 263Y	P: 444Tca>Cca S: 263taT>taC
	659	C	T	0.5515	P, S	P: synonymous S: missense	P: 511S S: 331A>V	P: 511agC>agT S: 331gCc>gTc
	669	C	T	0.3205	P, S	P: synonymous S: synonymous	P: 515L S: 334S	P: 515Ctg>Ttg S: 334tcC>tcT
	1,546	A	T	0.4638	P, X	P: missense X: synonymous	P: 807D>V X: 57G	P: 807gAc>gTc X: 57ggA>ggT
	2,337	G	A	0.4863	P, C	P: missense C: synonymous	P: 10R>K C: 174E	P: 10aGa>aAa C: 174gaG>gaA
	2,479	G	A	0.4016	P	P: synonymous	P: 57G	P: 57ggG>ggA
HPV-16	350	T	G	0.4508	E6	missense	83L>V	83Ttg>Gtg
	2,925	A	G	0.9157	E2	synonymous	57Q	57caA>caG
	3,409	C	T	0.5125	E2, E4	E2: missense E4: synonymous	E2: 219P>S E4: 26T	E2: 219Ccc>Tcc E4: 26acC>acT
	3,977	A	C	0.5596	E5	missense	39I>L	39Ata>Cta
	4,040	A	G	0.6121	E5	missense	60I>V	60Ata>Gta
	4,226	T	C	0.4844	Non-coding Region	NA	NA	NA
	4,363	A	T	0.9157	L2	missense	43E>D	43gaA>gaT
	4,936	G	A	0.8607	L2	synonymous	234Q	234caG>caA
	5,224	A	C	0.8192	L2	missense	330L>F	330ttA>ttC
	6,240	C	G	0.3415	L1	missense	228H>D	228Cat>Gat
	6,432	A	G	0.7019	L1	missense	292T>A	292Act>Gct
	7,191	G	T	0.6115	Non-coding Region	NA	NA	NA

¹The HBV genome contains four genes: P gene, S gene, X gene, and C gene, some of which overlap partially