Evolution and molecular characteristics of SARS-CoV-2 genome

Yunmeng Bai[1#], Dawei Jiang[1#], Jerome R Lon[1#], Xiaoshi Chen[1], Meiling Hu[1], Shudai Lin[1], Zixi Chen[1], Yuhuan Meng[2*], Hongli Du[1*]

1. School of Biology and Biological Engineering, South China University of Technology, Guangzhou, China
2. Guangzhou KingMed Transformative Medicine Institute Co., Ltd, Guangzhou, China

# Contributed equally as co-first authors

*Correspondence address should be addressed to Yuhuan Meng (zb-mengyuhuan@kingmed.com.cn) and Hongli Du (hldu@scut.edu.cn)

Abstract

In the evolution analysis of 622 complete human severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomes with high quality, the estimated Ka/Ks ratio of SARS-CoV-2 is 1.008, which is significantly higher than that of SARS-CoV and MERS-CoV, and the time to the most recent common ancestor (tMRCA) of SARS-CoV-2 is inferred in late September 2019 (95% CI: 2019/08/28-2019/10/26), which indicating that SARS-CoV-2 may have completed a positive selection pressure of the cross-host evolution in the early stage and be going through a neutral evolution at present. In addition, no-root phylogenetic tree of the 622 SARS-CoV-2 genomes were constructed by maximum likelihood (ML) with the bootstrap value of 100. According to the phylogenetic trees, all genomes were divided into Cluster 1 to 3, in which genomes were mainly from North America, global and Europe respectively. Further we find 9 key specific sites of highly linkage which play a decisive role in the classification of each cluster. Among them, 3 and 4 sites of almost complete linkage are the specific sites for Cluster 1 and Cluster 3 respectively. Notably the frequencies of haplotype TTTG and H1 are generally high in European countries and correlated to death rate (r>0.4) based on more than 3500 SARS-CoV-2 genomes, which indicated that the haplotypes might be related to pathogenicity of SARS-CoV-2 and need to be addressed. According to haplotype changes in chronological order, the H3 haplotype subgroup disappeared soon after detection, while H1 haplotype subgroup was globally increasing with time. The evolution and molecular characteristics of more than 3500 genomic sequences provided a new perspective for

revealing the epidemiology mechanism of SARS-CoV-2 and coping with SARS-CoV-2 effectively.

1. Introduction

The global outbreak of SARS-CoV-2 is currently and increasingly recognized as a serious, public health concern worldwide. To date, a total of seven types of coronaviruses that can infect humans have been found, among which the more typical ones are SARS-CoV, MERS-CoV and SARS-CoV-2. They spread across species and usually cause mild to severe respiratory diseases. The total number of SARS-CoV and MERS-CoV infections are only 8,069 and 2,494 with reproduction number (R0) fluctuates 2.5-3.9 and 0.3-0.8, respectively. However, SARS-CoV-2 has infected 2471136 people in 212 countries up to April 22th[1], with the basic R0 ranging from 1.4 to 6.49[2]. Among these three typical coronavirus, MERS-CoV has the highest death rate of 34.40%, SARS-CoV has the modest death rate of 9.59%, SARS-CoV-2 is about 6.99% and 7.69% death rate of the global and Wuhan, but is quite high death rate in some countries of Europe, such as Belgium, Italy, United Kingdom, Netherlands, Spain and France, which even reaches 14.95%, 13.39%, 13.48%, 11.61%, 10.42% and 13.60% respectively according to the data of April 22th, 2020. Except for the shortage of medical supplies, aging and other factors, it is not clear whether there is virus mutation effect in these countries with such a significantly increased death rate.

It has been reported that SARS-CoV-2 belongs to beta-coronavirus and is mainly transmitted by the respiratory tract, which belongs to the same subgenus (SarbeCoVirus) as SARS-CoV[3]. Through analyzing the genome and structure of SARS-CoV-2, its receptor-binding domain (RBD) was found to bind with angiotensin-converting enzyme 2 (ACE2), which is also one of the receptors for binding SARS-CoV[4]. Some early genomic studies have shown that SARS-CoV-2 is similar to certain bat viruses (RaTG13, with the whole genome homology of 96.2%[5]) and Malayan pangolins coronaviruses (GD/P1L and GDP2S, with the whole genome homology of 92.4%[6]). They have speculated several possible origins of SARS-CoV-2 based on its spike protein characteristics (cleavage sites or the RBD)[5-7], in particular, SARS-CoV-2 exhibits 97.4% amino acid similarity to the Guangdong pangolin coronaviruses in RBD, even though it is most closely related to bat coronavirus RaTG13 at the whole genome level. However, it is not enough to present genome-wide evolution by a single gene or local evolution of RBD, whether bats or pangolins play an important role in the zoonotic origin of SARS-CoV-2 remains uncertain[8]. Furthermore, since there is little molecular characteristics analysis of SARS-CoV-2 based on a large number of genomes, it is

difficult to determine whether the virus has significant variation that affects its phenotype. Thereby, it is necessary to further reveal the phylogenetic evolution and molecular characteristics of the whole genome of SARS-CoV-2 in order to develop a comprehensive understanding of the virus and provide a basis for designing vaccines and antiviral treatment strategies. Fortunately, with the increase of SARS-CoV-2 genome data, we have an opportunity to reveal phylogenetic evolution and molecular characteristics of SARS-CoV-2 more comprehensively. In this study, the Ka/Ks ratio, substitution rate and the time to the most recent common ancestor (tMRCA) of SARS-CoV-2, SARS-CoV and MERS-CoV were analyzed using KaKs_Calculator and BEAST, and then 622 SARS-CoV-2 genome sequences with high quality were clustered using maximum likelihood (ML) with bootstrap of 100 to ensure the reliability of evolutionary trees. According to the clusters of phylogenetic trees, the characteristic mutations of SARS-CoV-2 were screened and their potential significance were explored. The results of this study will provide a landscape for us to further understand SARS-CoV-2, and supply a possible basis for the prevention and treatment of SARS-CoV-2 in different areas.

2. Materials and methods

2.1 Genome Sequences

The complete genome sequences of SARS-CoV-2 were downloaded from China National Center for Bioinformation (https://bigd.big.ac.cn/ncov/release_genome/) and GISAID (https://www.gisaid.org/) up to March 22th, 2020. The sequences were filtered out according to the following criteria: (1) sequences with ambiguous time; (2) sequences with low quality which contained the counts of unknown bases > 15 and degenerate bases > 50 (https://bigd.big.ac.cn/ncov/release_genome); (3) sequences with similarity of 100% were removed to unique one. Finally, 624 high quality genomes with precise collection time were selected and aligned using MAFFT v7 with automatic parameters. Besides, the genome sequences of 7 SARS-CoV and 475 MERS-CoV were also downloaded from NCBI (https://www.ncbi.nlm.nih.gov/), and the MERS-CoV dataset including samples collected from both human and camel.

2.2 Estimate of evolution rate and the time to the most recent common ancestor for SARS-CoV, MERS-CoV, and SARS-CoV-2

The average Ka, Ks and Ka/Ks for all coding sequences were calculated using KaKs_Calculator v1.2[9], and the substitution rate and tMRCA were estimated using BEAST v2.6.2[10]. The temporal

signal with root-to-tip divergence was visualized in TempEst v1.5.3[11] using a ML whole genome tree as input. For SARS-CoV and SARS-CoV-2, we selected a strict molecular clock and Coalescent Exponential Population Model. For MERS-CoV, we selected a relaxed molecular clock and Birth Death Skyline Serial Cond Root Model. We used the tip dates and chose the HKY as the site substitution model in all these analyses. Markov Chain Monte Carlo (MCMC) chain length was set to 10,000,000 steps sampling after every 1000 steps. The output was examined in Tracer v1.6 (http://tree.bio.ed.ac.uk/software/tracer/).

2.3 Variants calling of SARS-CoV-2 genome sequences

Each genome sequence was aligned to the Wuhan-Hu-1 reference genome (NC_045512.2) using bowtie2 with default parameters[12], and variants were called by samtools (sort; mpileup -gf) and bcftoots (call -vm). The merge VCF files were created by bgzip and bcftools (merge --missing-to-ref)[13,14].

2.4 Phylogenetic tree construction and virus isolates clustering for SARS-CoV-2

After alignment of 624 SARS-CoV-2 genomes with high quality and manually deleted 2 highly divergent genomes (EPI_ISL_415710, EPI_ISL_414690) according to the firstly constructed phylogenetic tree, the aligned dataset was phylogenetically analyzed. The SMS method was used to select GTR+G as the base substitution model[15], and the MEGA[16] and PhyML 3.1[17] were used to construct the phylogenetic tree by the maximum likelihood method with the bootstrap value of 100. The online tool iTOL[18] was used to visually beautify the phylogenetic tree. The clusters were defined by the shape of phylogenetic tree.

2.5 Detection of specific sites from each Cluster

Information (ID, countries/regions and collection time) and variants (NC_045512.2 as reference genome) of each genome from each Cluster were extracted. The allele frequency and nucleotide divergency (pi) for each site in the virus population of each Cluster were measured by vcftools[19]. The *Fst* were also calculated by vcftools[19] to assess the diversity between the Clusters. Sites with high level of *F*st together with different major allele in each Cluster were filtered as the specific sites. PCA were analyzed by the GCTA v1.93.1beta[20] with the specific sites and all SNV dataset respectively.

2.6 Linkage analysis of specific sites and characteristics of main haplotype subgroups

The linkage disequilibrium of the specific sites for 622 genomes and the following redownloaded

genomes were analyzed by haploview[21], and the statistics of the haplotype of the specific sites for each Cluster or country were used in-house perl script.

2.7 Frequencies of specific sites or haplotypes and correlation with death rate

We redownloaded the sequences from GISAID up to April 6th, 2020 to further explore possible roles of the specific sites, and obtained 3624 genome sequences from the countries with more than 10 complete genome sequences. The frequencies of specific sites for each country were calculated. The death rate was estimated with Total Deaths /Confirmed Cases based on the data from Johns Hopkins resources on April 22th, 2020 (https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6). The correlation coefficient between death rate and frequencies of specific site or haplotype in different countries was calculated using Pearson method.

3. Results

3.1 Genome Sequences

We got a total of 1053 sequences up to March 22th, 2020. According to the filter criteria, 37 sequences with ambiguous time, 314 with low quality and 78 with similarity of 100% were removed. A total of 624 sequences were obtained to perform multiple sequences alignment. Two highly divergent sequences (EPI_ ISL_414690, EPI_ISL_415710) according to the firstly constructed phylogenetic tree were also filtered out (**Extended Data Table 1**). The remaining 622 sequences were used to reconstruct a phylogenetic tree.

3.2 Estimate of evolution rate and the time to the most recent common ancestor for SARS-CoV, MERS-CoV, and SARS-CoV-2

In this study, date for SARS-CoV-2 ranged from 2019/12/26 to 2020/03/18 was collected. The average Ka/Ks of all the coding sequences was closer to 1 (1.008), which indicating the genome was going through a neutral evolution. We also reevaluated the Ka/Ks of SARS-CoV and MERS-CoV through the whole period, and found the ratio were significantly smaller than SARS-CoV-2 (**Table 1**). To estimate more credible Ka/Ks for SARS-CoV-2, we recalculated it using redownloaded 3624 genome sequences ranged from 2019/12/26 to 2020/04/06, and the average Ka/Ks of it was 1.009, which was almost same with the above result.

We assessed the temporal signal using TempEst v1.5.3. All three data sets exhibit a positive

correlation between root-to-tip divergence and sample collecting time (**Extended Data Fig 1**), hence they are suitable for molecular clock analysis in BEAST[11]. The substitution rate of SARS-CoV-2 genome was estimated to be 1.60×10^{-3} (95% CI: 1.42×10^{-3} - 1.80×10^{-3}, **Table 2**, **Extended Data Fig 2a**) substitution/site/year, which is as in the same order of magnitude as SARS-CoV and MERS-CoV. The tMRCA was inferred on the late September, 2019 (95% CI: 2019/08/28-2019/10/26,**Table 2**, **Extended Data Fig 2b**), about 2 months before the early cases of SARS-CoV-2[22].

3.3 Phylogenetic tree and clusters of SARS-CoV-2

We firstly constructed the phylogenetic tree using PhyML 3.1 with fast likelihood-based method (aLRT SH-like), while it failed when set the bootstrap value as 100 due to no response for a long time, so that MEGA was used with GTR+G model and the bootstrap value was set to 100. According to the shape of phylogenetic trees (**Fig 1, Extended Data Fig 3a and b)**, we divided 622 sequences into three clusters: Cluster 1 including 76 sequences mainly from North America, Cluster 2 including 367 sequences from all regions of the world, and Cluster 3 including 179 sequences mainly from Europe (**Extended Data Table 2**).

3.4 The specific sites of each Cluster

The Fst and population frequency of a total of 9 sites (NC_045512.2 as reference genome) were detected (**Table 3**, **Extended Data Table 3**). Thereinto, three (C17747T, A17858G and C18060T) are the specific sites of Cluster 1, and four (C241T, C3037T, C14408T and A23403G) are the specific sites of Cluster 3. Notably, C241T was located in the 5'-UTR region and the others were located in coding regions (6 in *ofr1ab* gene, 1 in *S* gene and 1 in *ORF8* gene). Five of them were missense variant, including C14408T, C17747T and A17858G in *ofr1ab* gene, A23403G in *S* gene, and C28144T in *ORF8* gene. The PCA results showed that these 9 specific sites could clearly separate the three Clusters, while all SNV dataset could not clearly separate Cluster 1 and Cluster 2 (**Extended Data Fig 4**), which further suggested that these 9 specific sites are the key sites for separating the three Clusters.

3.5 Linkage of specific sites and characteristics of main haplotype subgroups

We found the 9 specific sites are highly linkage based on 622 genome sequences (**Fig 2a**), then we carried out a further linkage analysis using the redownloaded 3624 genome sequences (**Fig 2b**). As a result, for the 3624 genome sequences, 3 specific sites in the Cluster 1 were almost complete

linkage, and haplotype CAC and TGT accounted for 98.65% of all the 3 site haplotypes. The same phenomenon was also found in 4 specific sites of Cluster 3, and haplotype CCCA and TTTG accounted for 97.68% of all the 4 site haplotypes. Intriguingly, the 9 specific sites were still highly linkage, and four haplotypes, including TTCTCACGT (H1), CCCCCACAT (H2), CCTCTGTAC (H3) and CCTCCACAC (H4), accounted for 95.89% of all the 9 site haplotypes. The frequencies of each site and main haplotype for each country were showed in **Fig 3** and **Extended Data Table 4**. The data of 3624 genomes showed that the haplotype TTTG of the 4 specific sites in Cluster 3 had existed globally at present, and still exhibited high frequencies in most European countries but quite low in Asian countries. While the haplotype TGT of the 3 specific sites in Cluster 1 existed almost only in North America and Australia. For the 9 specific sites, most countries had only two or three main haplotypes, except America, Australia and United Kingdom had all of four main haplotypes.

All haplotypes of 9 specific sites and the numbers of them were shown in **Extended Data Table 5,** and the numbers of these haplotypes detected in each country in chronological order were shown in **Fig 4a.** From these results, H2 and H4 haplotype subgroups had existed for a long time (2019-12-31 to 2020-04-14), and the H3 haplotype subgroup disappeared soon after detection (2020-02-18 to 2020-04-07), while H1 haplotype subgroup was globally increasing with time, which indicated H1 subgroup had adapted to the human hosts, and was under an adaptive growth period worldwide. However, due to the nonrandom sampling on early phase (only patients having a recent travel to Wuhan were detected), some earlier cases of H3 may be lost, the high proportion of H3 subgroup during Feb 18 to 25 could confirm this point.

From the whole genome mutations in each haplotype subgroup (**Fig 4b, Extended Data Table 6**), we found that except the 9 specific sites, there was no common mutations with frequencies more than 0.01 in these haplotype subgroups but between H2 and H4 subgroups. H3 and H4 subgroups had the lowest Ka/Ks ratio (0.795 and 0.796 respectively) among the 4 main subgroups (**Table 4**), suggesting that H3 and H4 subgroup might be going through a purifying evolution and have existed for a long time, while H2 (1.268) and H1 subgroup (1.099) might be going through a positive selection and a neutral evolution according to their Ka/Ks ratios, respectively, which were consistent with the above phenomenon that only H1 subgroup was expanding with time around the world.

3.6 Correlation analysis of specific sites with death rate

To explore the relationship between death rate and the 9 specific sites, we used Pearson method to calculate the correlation coefficient between death rate and frequency of each specific site or main haplotype in 17 countries. As a result, all r values of 241T, 3037T, 14408T, 23403G and haplotype TTTG and H1 in Cluster 3 were more than 0.4 (**Fig 5**, **Extended Data Table 4**). This finding integrating their high frequencies in most European countries indicated that the 4 sites and haplotype TTTG and H1 might be related to the pathogenicity of SARS-CoV-2.

4. Discussion

SARS-CoV-2 poses a great threat to the production, living and even survival of human beings[23]. With the further outbreak of SARS-CoV-2 in the world, further understanding on the evolution and molecular characteristics of SARS-CoV-2 based on a large number of genome sequences will help us cope better with the challenges brought by SARS-CoV-2. In the present study, we estimated the Ka/Ks ratio and tMRCA of SARS-CoV-2, SARS-CoV and MERS-CoV, and constructed the phylogenetic tree of SARS-CoV-2 genomes. As a result, we found that 9 specific sites of highly linkage could play a decisive role in the classification of SARS-CoV-2, and the haplotype TTTG might be related to pathogenicity of SARS-CoV-2 based on more than 3500 genomes. All these results could accelerate our further understanding of SARS-CoV-2.

Exploring evolution rate, tMRCA and phylogenetic tree of SARS-CoV-2 would help us better understand the virus[24]**.** The average Ka/Ks for all the coding sequences of 622 and 3624 SARS-CoV-2 genomes was 1.008 and 1.009, which was significantly higher than those of SARS-CoV and MERS-CoV, indicating that the SARS-CoV-2 was going through a neutral evolution. In general, SARS-CoV would show purifying selection pressures and the Ka/Ks would go down at the middle or end of the epidemic[25]. Therefore, it seemed to be reasonable considering that SARS-CoV-2 was under an exponential growth period around the world. Interestingly, we also found that the subgroups of different haplotypes (H1, H2, H3 and H4) seemed to experience different evolutionary patterns according to their Ka/Ks. The H3 haplotype subgroup disappeared soon after detection (2020-02-18 to 2020-04-07) (2020-02-18 to 2020-04-07, Fig 4a), while H1 haplotype subgroup was globally increasing with time, these evolution and changes should be considered in developing therapeutic drugs and vaccines as soon as possible. The tMRCA of SARS-CoV-2 was inferred on the late September, 2019 (95% CI: 2019/08/28- 2019/10/26), about 2 months before the early cases

of SARS-CoV-2[22]. We also estimated tMRCA of SARS-CoV and MERS-CoV with the same methods, both were about 3 months later than the corresponding tMRCAs estimated by previous studied[26,27], which indicated the tMRCA of SARS-CoV-2 we estimated was reasonable.

A recent study clustered 160 SARS-CoV-2 whole-genome sequences into A, B and C groups by a phylogenetic network analysis[28]. The cluster result was similar to our study: both the samples in Cluster A and our Cluster 1 were mainly from the United States; the samples in Cluster C and our Cluster 3 were mainly from European countries, while Cluster B and our Cluster 2 were mainly from China and the other regions. It was interesting that the markers in Cluster B (T8782C and C28144T) were also found in our study, but the other markers in Cluster A (T29095C) and Cluster C (G26144T) were not significantly in our study. That may be caused by different sample sizes and different constructing methods of phylogenetic tree. Based on the base substitution model, the ML method avoids the possible "long-branch attraction" problem in the maximum parsimony method and is faster than the Bayesian method[29], hence it could be used as a reliable method for phylogenetic analysis. Nowadays, there is still a shortage of studies on phylogenetic analysis of SARS-CoV-2. Some studies used the genome of bat SARS-like-CoV[30], RaTG13[30] or MT019529 from Wuhan (https://bigd.big.ac.cn/ncov/tree) as the root of phylogenetic tree. Unfortunately, there was no obvious evidence showing that SARS-CoV-2 was from the bat coronavirus even the identity between SARS-CoV-2 and RaTG13 was up to 96.2%[5]. In our study, the tMRCA of SARS-CoV-2 was inferred on the late September 2019, which indicated there might exist an earlier SARS-CoV-2 strain we didn't find. Then in the case of unclear source of SARS-CoV-2 and high homology of its genomes (> 99.9% homology), it may be inappropriate to identify the evolutionary characteristics inside the genomes by taking bat SARS-like-CoV, RaTG13 or MT019529 as root. Therefore, the maximum likelihood method was used in current study to construct a no-root tree to obtain the reliable clusters with different characteristics. Based on the no-root tree, we identified 9 specific sites of highly linkage that play a decisive role in the classification of clusters successfully. Among the four main haplotypes, H1 and H3 were in Cluster 3 and Cluster 1, respectively, while there were two haplotypes H2 and H4 in Cluster 2 (**Fig 1, Extended Data Fig 3b)**.

Among the 9 specific sites, 8 of them were located in coding regions (6 in *ofr1ab* gene, 1 in *S* gene and 1 in *ORF8* gene), and 5 of them were missense variant, including C14408T, C17747T and A17858G in *ofr1ab* gene, A23403G in *S* gene, and C28144T in *ORF8* gene. *Orf1ab* is composed of two partially

overlapping open reading frames (orf), orf1a and 1b. It is proteolytic cleaved into 16 non-structural proteins (nsp), including nsp1 (suppress antiviral host response), nsp9 (RNA/DNA binding activity), nsp12 (RNA-dependent RNA polymerase), nsp13 (helicase) and others[31], indicating the vital role of it in transcription, replication, innate immune response and virulence[32]. C14408T with high frequencies of T in European countries were located at nsp12 region, indicating that this missense variant might influence the role of RNA polymerase. Spike glycoprotein, the largest structural protein on the surface of coronaviruses, comprises of S1 and S2 subunits which mediating binding the receptor on the host cell surface and fusing membranes, respectively[33]. It has been reported that S protein of SARS-CoV-2 can bind angiotensin-converting enzyme 2 (ACE2) with higher affinity than that of SARS[4,5]. Whether the missense variant of A23403G in *S* gene, which with high frequencies of G in European countries, will change the affinity between S protein and ACE2 remains to be further investigated.

Previous studies showed that the death rate of SARS-CoV-2 is affected by many factors, such as medical supplies, aging, life style and other factors[34,35]. However, our study found that the 4 specific sites (C241T, C3037T, C14408T and A23403G) in the Cluster 3 were almost complete linkage, and the frequency of haplotype TTTG was generally high in European countries and correlated to death rate ($r>0.4$), which provides a new perspective to the reasons of relatively high death rate in Europe. In order to make frequency relatively credible, we selected countries with more than 10 genomes (except South Korea with 16 genomes, all the other countries with more than 20 genomes) to analyze the correlation between frequencies of haplotypes and death rate. Therefore, the correlation between the frequencies of haplotypes and death rate obtained in this study is quite reliable, and these specific sites should be considered in designing new vaccine and drug development of SARS-CoV-2.

5. Conclusion and prospective

The Ka/Ks ratio of SARS-CoV-2 (1.008) and tMRCA of SARS-CoV-2 (95% CI: 2019/08/28-2019/10/26) indicated that SARS-CoV-2 might have completed the positive selection pressure of cross-host evolution in the early stage and be going through a neutral evolution at present. The 9 specific sites with highly linkage were found to play a decisive role in the classification of clusters. Thereinto, 3 of them are the specific sites of Cluster 1 mainly from North America, 4 of them are the specific sites of Cluster 3 mainly from Europe, both of these 3 and 4 specific sites are almost

complete linkage. The frequencies of haplotype TTTG for the 4 specific sites in Cluster 3 were generally high in European countries and correlated to death rate (r>0.4) based on more than 3500 SARS-CoV-2 genomes, which suggested this haplotype might relate to pathogenicity of SARS-CoV-2 and provided a new perspective to the reasons of relatively high death rates in Europe. The relationship between the haplotype TTTG and the pathogenicity of SARS-CoV-2 needs to be further verified by clinical samples or virus virulence test experiment, and the 9 specific sites with highly linkage may provide a starting point for the traceability research of SARS-CoV-2. Given that the different evolution patterns of different haplotypes subgroups, we should consider these evolution and changes in the development of therapeutic drugs and vaccines as soon as possible.

Acknowledgements

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 Program of Guangzhou (2020Q-P013).

Author contributions

H.D. conceived the study. Y.M., Y.B., D.J., J.L and X.C. carried out the data analysis and wrote the manuscript. M.H., S.L., and Z.C. collected data and revised the manuscript. H.D. and Y.M. supervised the whole work, attended the discussions and revised the manuscript.

Competing interests

The authors declare no competing interests.

Data availability

Sequence data that support the findings of this study could be downloaded in China National Center for Bioinformation(https://bigd.big.ac.cn/ncov/release_genome/), GISAID (https://www.gisaid.org/) and NCBI (https://www.ncbi.nlm.nih.gov/).

Code availability

The Perl script is available on Github (https://github.com/yunmengbai/SARS-CoV-2-genome.git).

Reference

1 Coronavirus disease 2019 (COVID-19) Situation Report – 93. *WHO* (2020).

2 Liu, Y., Gayle, A. A., Wilder-Smith, A. & Rocklov, J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J Travel Med* **27**, doi:10.1093/jtm/taaa021 (2020).

3 Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565-574, doi:10.1016/S0140-6736(20)30251-8 (2020).

4 Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260-1263, doi:10.1126/science.abb2507 (2020).

5 Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273, doi:10.1038/s41586-020-2012-7 (2020).

6 Lam, T. T. *et al.* Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature*, doi:10.1038/s41586-020-2169-0 (2020).

7 Zhang, Y. Z. & Holmes, E. C. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. *Cell*, doi:10.1016/j.cell.2020.03.035 (2020).

8 Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature Medicine*, doi:10.1038/s41591-020-0820-9 (2020).

9 Zhang, Z. *et al.* KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**, 259-263, doi:10.1016/S1672-0229(07)60007-2 (2006).

10 Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **15**, e1006650, doi:10.1371/journal.pcbi.1006650 (2019).

11 Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* **2**, vew007 (2016).

12 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

13 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

14 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).

15 Lefort, V., Longueville, J. E. & Gascuel, O. SMS: Smart Model Selection in PhyML. *Mol Biol Evol* **34**, 2422-2424, doi:10.1093/molbev/msx149 (2017).

16 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547-1549, doi:10.1093/molbev/msy096 (2018).

17 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321, doi:10.1093/sysbio/syq010 (2010).

18 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **47**, W256-W259, doi:10.1093/nar/gkz239 (2019).

19 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics (Oxford, England)* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).

20 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82, doi:10.1016/j.ajhg.2010.11.011 (2011).

21 Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-265 (2005).

22 Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497-506, doi:10.1016/S0140-6736(20)30183-5 (2020).

23 Guo, Y. R. *et al.* The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak - an update on the status. *Mil Med Res* **7**, 11, doi:10.1186/s40779-020-00240-0 (2020).

24 Yuen, K.-S., Ye, Z.-W., Fung, S.-Y., Chan, C.-P. & Jin, D.-Y. SARS-CoV-2 and COVID-19: The most important research questions. *Cell Biosci* **10**, 40, doi:10.1186/s13578-020-00404-4 (2020).

25 Chinese, S. M. E. C. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* **303**, 1666-1669, doi:10.1126/science.1092002 (2004).

26 Zhao, Z. *et al.* Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evol Biol* **4**, 21, doi:10.1186/1471-2148-4-21 (2004).

27 Cotten, M. *et al.* Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. *mBio* **5**, doi:10.1128/mBio.01062-13 (2014).

28 Forster, P., Forster, L., Renfrew, C. & Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences*, 202004999, doi:10.1073/pnas.2004999117 (2020).

29 Holder, M. & Lewis, P. O. Phylogeny estimation: traditional and Bayesian approaches. *Nat Rev Genet* **4**, 275-284, doi:10.1038/nrg1044 (2003).

30 Zhang, L., Shen, F.-m., Chen, F. & Lin, Z. Origin and Evolution of the 2019 Novel Coronavirus. *Clinical Infectious Diseases*, doi:10.1093/cid/ciaa112 (2020).

31 Chan, J. F. *et al.* Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg Microbes Infect* **9**, 221-236, doi:10.1080/22221751.2020.1719902 (2020).

32 Graham, R. L., Sparks, J. S., Eckerle, L. D., Sims, A. C. & Denison, M. R. SARS coronavirus replicase proteins in pathogenesis. *Virus Res* **133**, 88-100, doi:10.1016/j.virusres.2007.02.017 (2008).

33 Li, F. Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annu Rev Virol* **3**, 237-261, doi:10.1146/annurev-virology-110615-042301 (2016).

34 Malavolta, M., Giacconi, R., Brunetti, D., Provinciali, M. & Maggi, F. Exploring the Relevance of Senotherapeutics for the Current SARS-CoV-2 Emergency and Similar Future Global Health Threats. *Cells* **9**, doi:10.3390/cells9040909 (2020).

35 Liu, Y. *et al.* Neutrophil-to-lymphocyte ratio as an independent risk factor for mortality in hospitalized patients with COVID-19. *J Infect*, doi:10.1016/j.jinf.2020.04.002 (2020).

Figure Legends

Figure 1. Phylogenetic tree and clusters of 622 SARS-CoV-2 genomes. The 622 sequences were clustered into three clusters: Cluster 1 were mainly from North America, Cluster 2 were from regions all over the world, and Cluster 3 were mainly from Europe.

Figure 2. Linkage disequilibrium plot of haplotypes of the 9 specific sites. **a**, The plot for 622

genome sequences; **b**, The plot for 3624 genome sequences.

Figure 3. The frequencies of both the 9 specific sites and haplotypes. The frequencies of the 9 specific sites (**a**) and haplotypes (**b**) in each country which with more than 10 genome sequences up to April 6th, 2020.

Figure 4. The characteristics of haplotype subgroups. **a,** The numbers of haplotypes of the 9 specific sites detected in each country in chronological order (In order to get the latest trend of haplotype subgroup in chronological order, we had updated the genome data to April 20th, 2020); **b,** The whole genome mutations in each main haplotype subgroup.

Figure 5. The correlation between death rate and frequencies of both the 9 specific sites and haplotypes.

Tables

Table 1. Statistics of Ka, Ks and Ka/Ks ratios for all coding regions of the SARS-CoV, MERS-CoV and SARS-CoV-2 genome sequences

Table 2. Substitution rate and tMRCA estimated by BEAST v2.6.2

Table 3. The information of the 9 specific sites in each Cluster.

Table 4. Statistics of Ka, Ks and Ka/Ks ratios for all coding regions of each 4 main haplotype genomes

Extended Data

Extended Data Figure 1. Regression analyses of the root-to-tip divergence with the maximum-likelihood phylogenetic tree of SARS-CoV (**a**), MERS-CoV (**b**) and SARS-CoV-2 (**c**). The root-to-tip genetic distances were inferred in TempEst v1.5.3. Three data sets exhibit a positive correlation between root-to-tip divergence and sample collecting time, and they appear to be suitable for molecular clock analysis.

Extended Data Figure 2. Substitution rate and tMRCA estimate of SARS-CoV-2

a, Estimated substitution rate of SARS-CoV-2 using BEAST v2.6.2 under a strict molecular clock. The mean rate was 1.60×10^{-3} (substitutions per site per year). **b**, The estimated tMRCA of SARS-CoV-2 using BEAST v2.6.2 under a strict molecular clock. The mean tMRACA was 2019.74 (date: 2019-09-27).

Extended Data Figure 3. Phylogenetic tree of 622 SARS-CoV-2 genomes. **a**, Phylogenetic tree constructed by MEGA with bootstrap value of 100. **b**, Phylogenetic tree constructed by PhyML 3.1.

Extended Data Figure 4. The PCAs analysis result based on the 9 specific sites (**a**) and all SNVs (**b**).

Extended Data Table 1. The information of the first downloaded complete genome sequences.

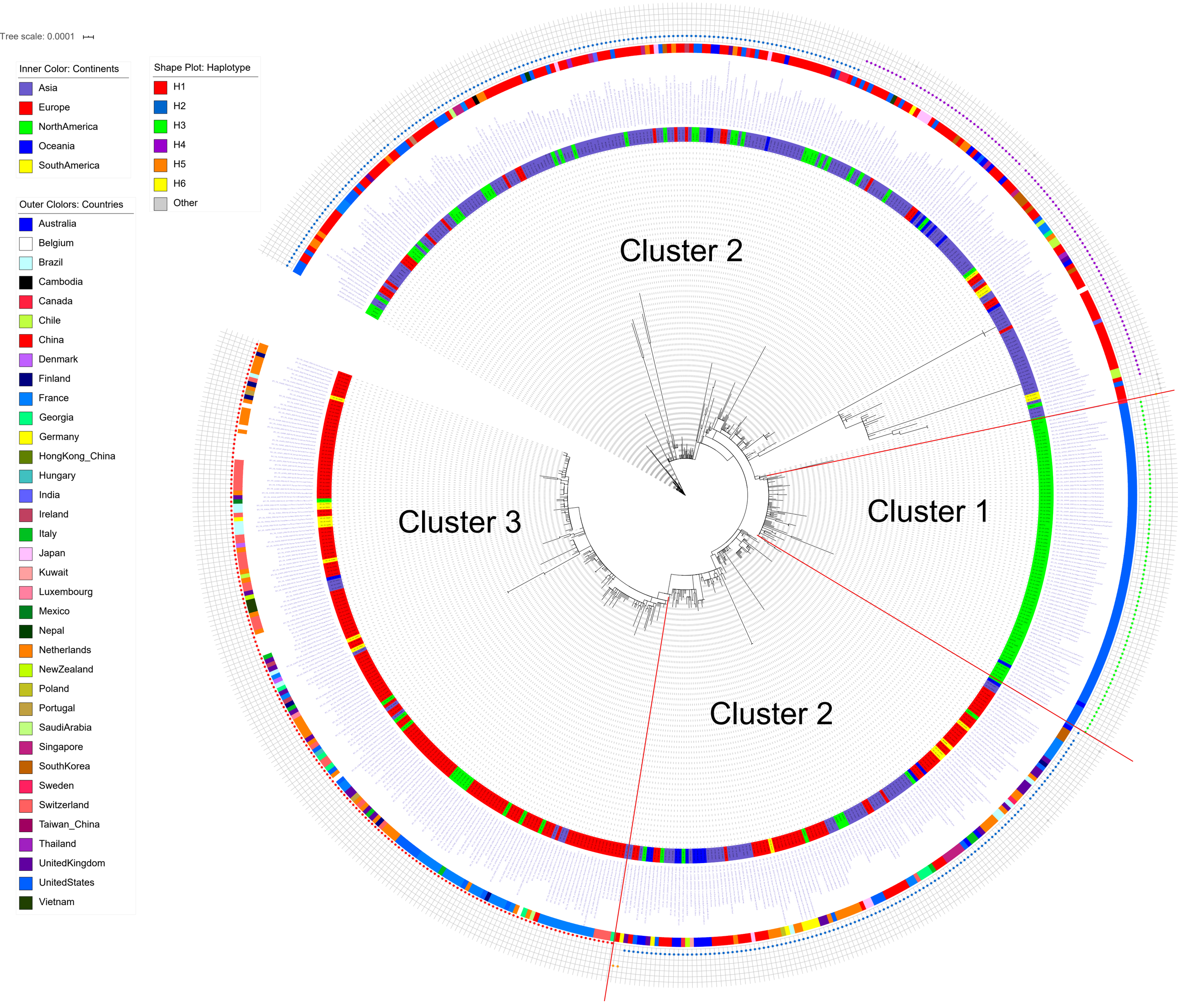
Extended Data Table 2. The sequence IDs in each Cluster.

Extended Data Table 3. The details information of the SNVs in all Clusters.

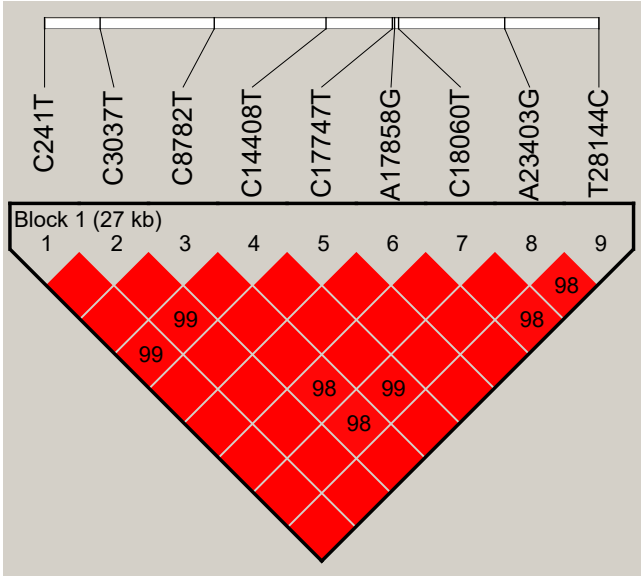
Extended Data Table 4. The information of death rate, frequencies of the 9 specific sites and main haplotypes in each country for 3624 genomes.

Extended Data Table 5. All haplotypes of the 9 specific sites and numbers of them for 3624 genomes.

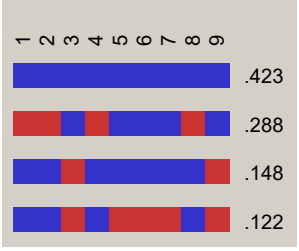
Extended Data Table 6. Whole genome mutations and frequencies in each main haplotype subgroup (NC_045512.2 as reference genome)


Tree scale: 0.0001
Inner Color: Continents
Asia
Europe
NorthAmerica
Oceania
SouthAmerica
Shape Plot: Haplotype
H1
H2
H3
H4
H5
H6
Other
Outer Clolors: Countries
Australia
Belgium
Brazil
Cambodia
Canada
Chile
China
Denmark
Finland
France
Georgia
Germany
HongKong_China
Hungary
India
Ireland
Italy
Japan
Kuwait
Luxembourg
Mexico
Nepal
Netherlands
NewZealand
Poland
Portugal
SaudiArabia
Singapore
SouthKorea
Sweden
Switzerland
Taiwan_China
Thailand
UnitedKingdom
UnitedStates
Vietnam
Cluster 2
Cluster 3
Cluster 1
Cluster 2

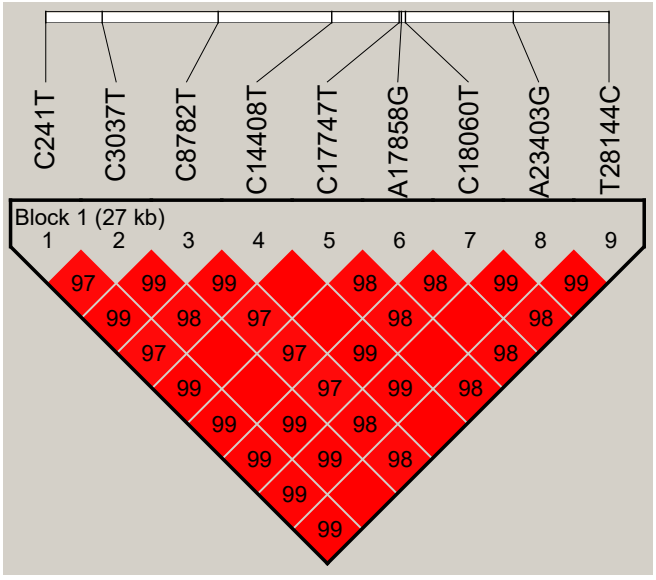
a

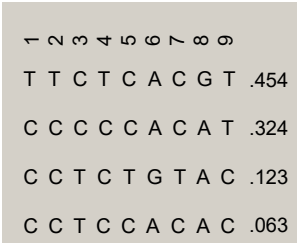

C241T
C3037T
C8782T
C14408T
C17747T
A17858G
C18060T
A23403G
T28144C
Block 1 (27 kb)
1 2 3 4 5 6 7 8 9
98
99
98
99
98
99
98

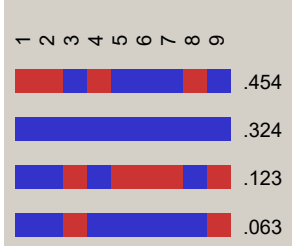
1 2 3 4 5 6 7 8 9
C C C C C A C A T .423
T T C T C A C G T .288
C C T C C A C A C .148
C C T C T G T A C .122

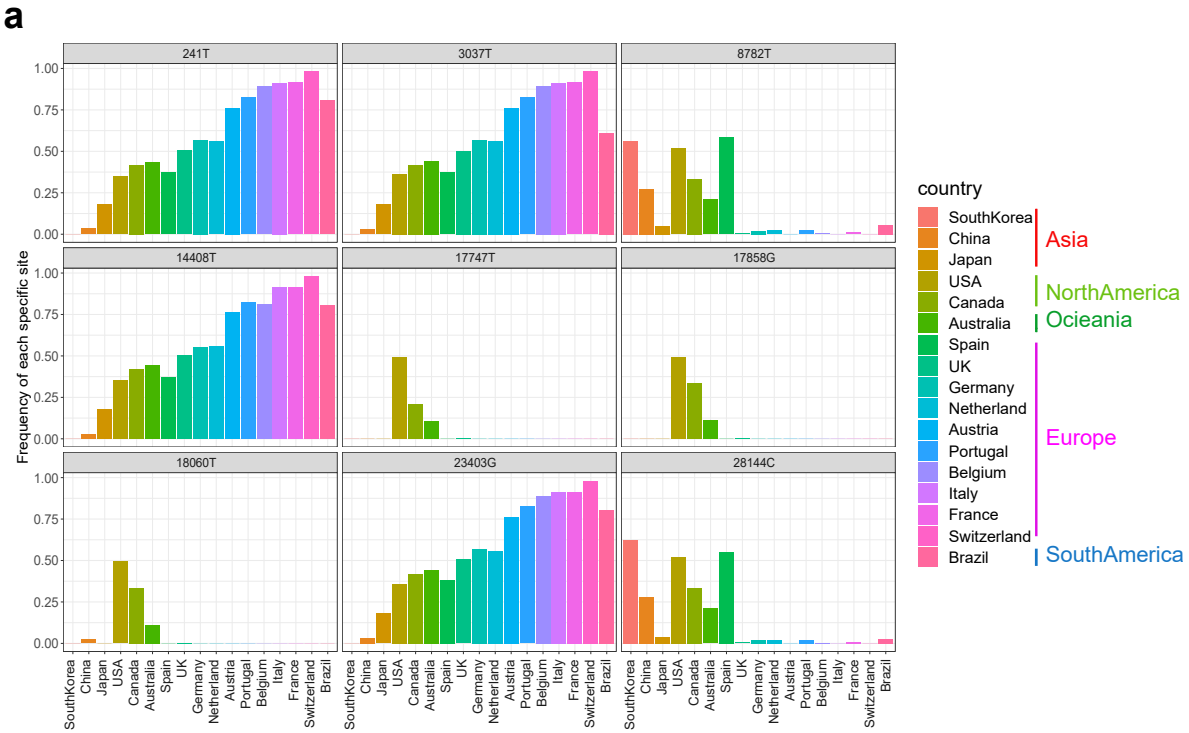

1 2 3 4 5 6 7 8 9
.423
.288
.148
.122

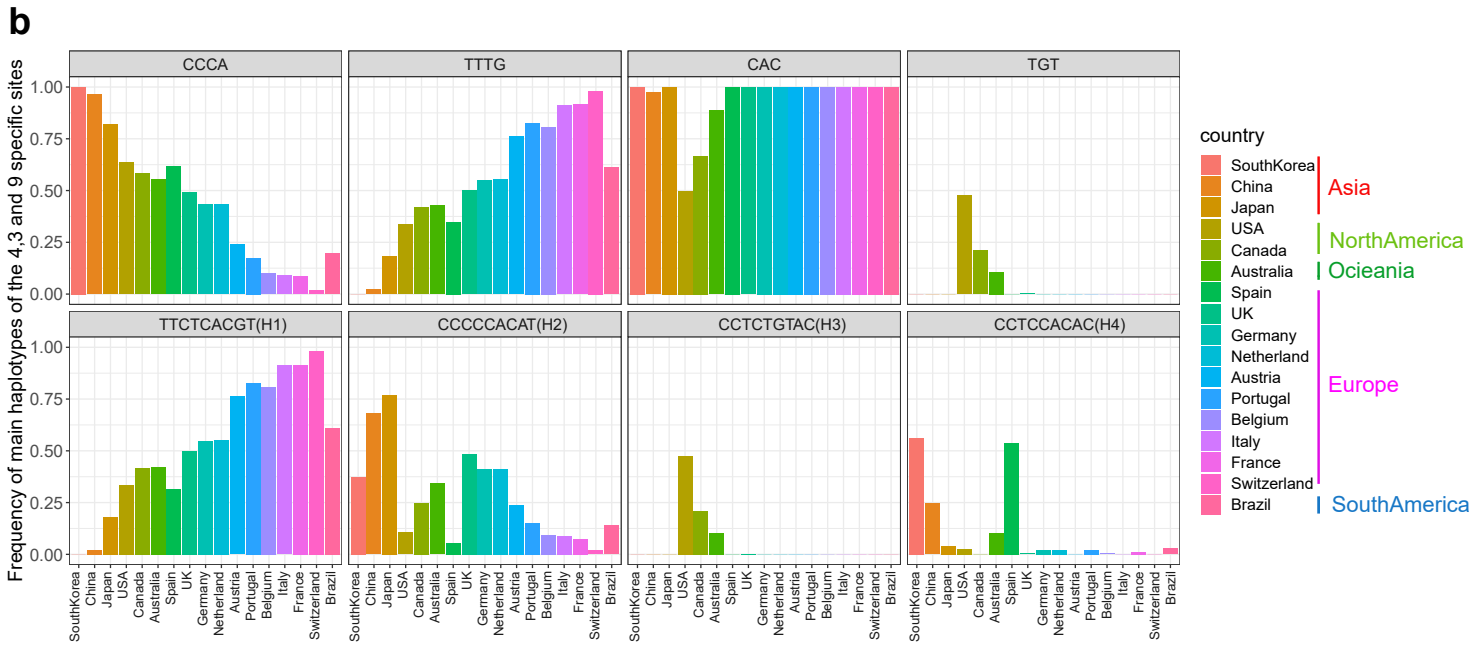
b

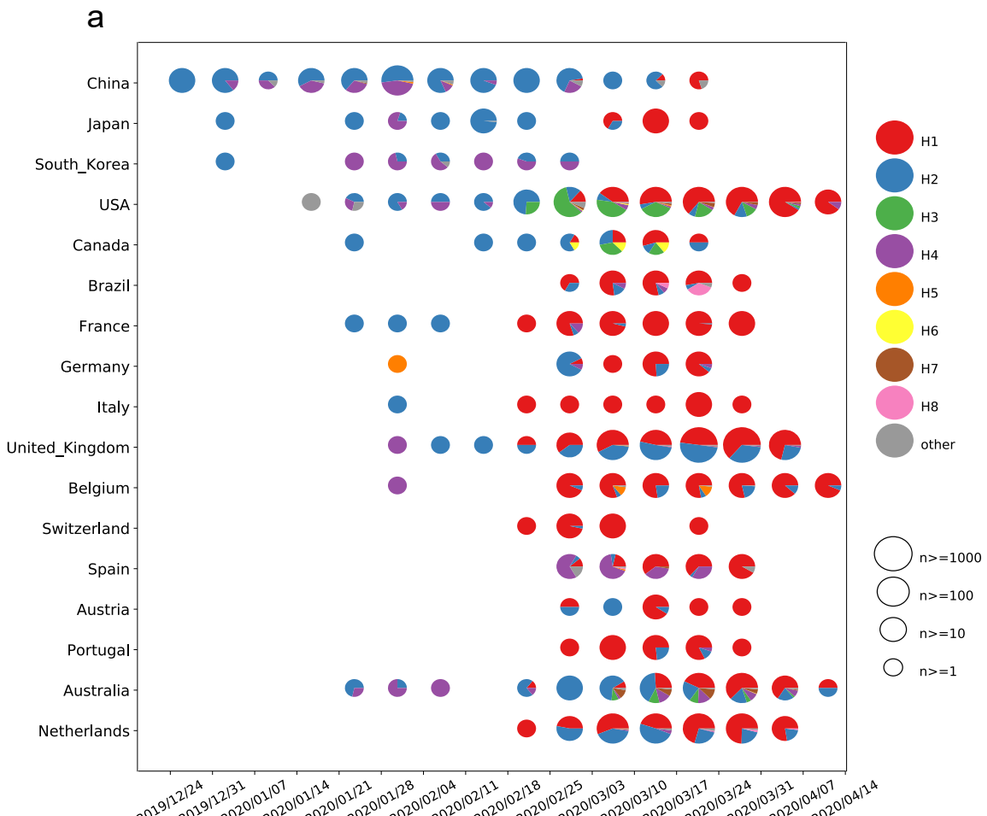

C241T
C3037T
C8782T
C14408T
C17747T
A17858G
C18060T
A23403G
T28144C
Block 1 (27 kb)
1 2 3 4 5 6 7 8 9
97 99 99 98 98 99 99
99 98 97 98 98
97 97 99 98
99 97 99 98
99 99 98
99 99 98
99
99

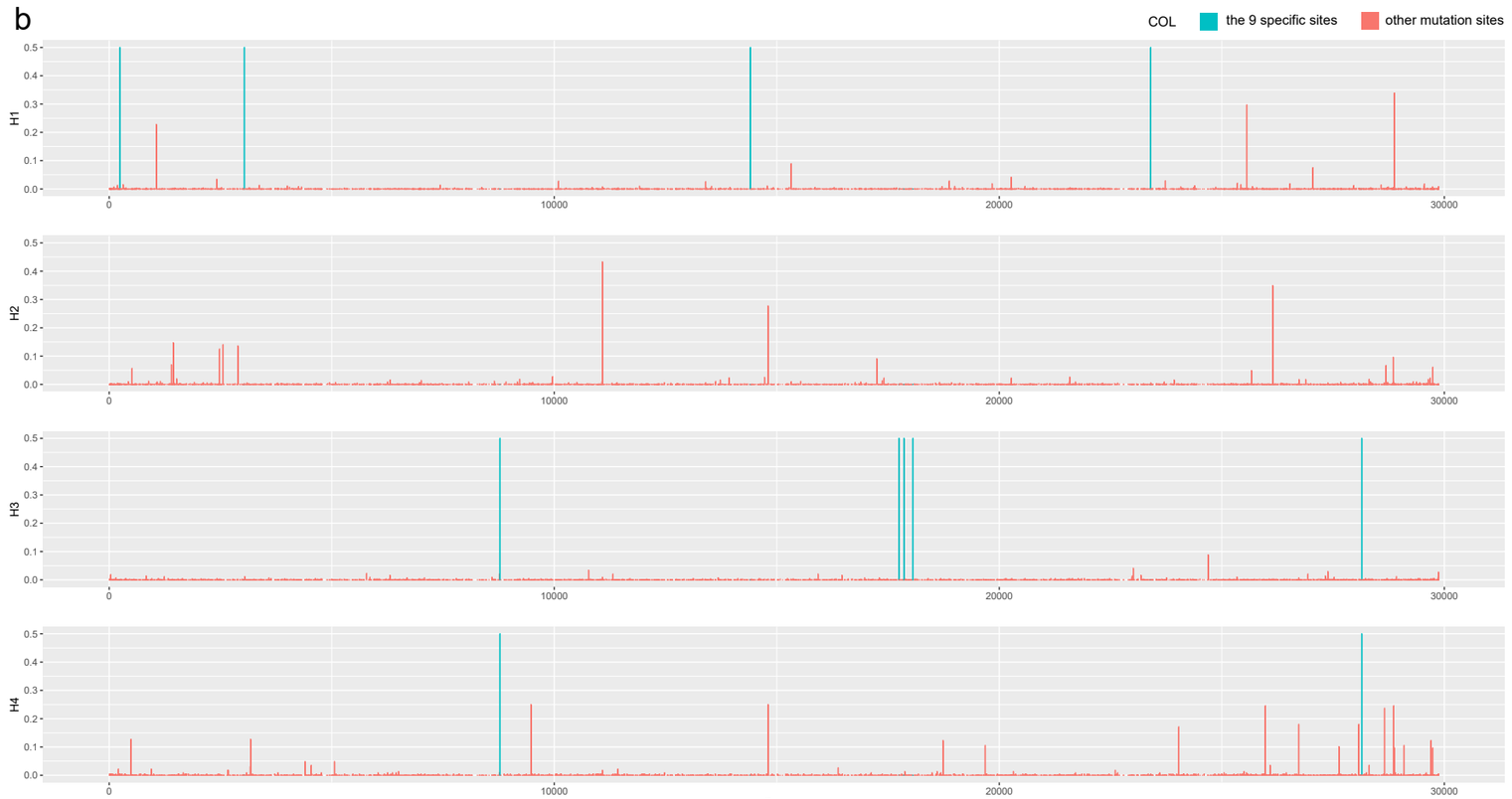

1 2 3 4 5 6 7 8 9
T T C T C A C G T .454
C C C C C A C A T .324
C C T C T G T A C .123
C C T C C A C A C .063


1 2 3 4 5 6 7 8 9
.454
.324
.123
.063


a
241T
3037T
8782T
14408T
17747T
17858G
18060T
23403G
28144C
Frequency of each specific site
1.00
0.75
0.50
0.25
0.00
SouthKorea
China
Japan
USA
Canada
Australia
Spain
UK
Germany
Netherland
Austria
Portugal
Belgium
Italy
France
Switzerland
Brazil
country
Asia
NorthAmerica
Ocieania
Europe
SouthAmerica


b
CCCA
TTTG
CAC
TGT
TTCTCACGT(H1)
CCCCCACAT(H2)
CCTCTGTAC(H3)
CCTCCACAC(H4)
Frequency of main haplotypes of the 4,3 and 9 specific sites
1.00
0.75
0.50
0.25
0.00
SouthKorea
China
Japan
USA
Canada
Australia
Spain
UK
Germany
Netherland
Austria
Portugal
Belgium
Italy
France
Switzerland
Brazil
country
Asia
NorthAmerica
Ocieania
Europe
SouthAmerica


a
China
Japan
South_Korea
USA
Canada
Brazil
France
Germany
Italy
United_Kingdom
Belgium
Switzerland
Spain
Austria
Portugal
Australia
Netherlands
2019/12/24
2019/12/31
2020/01/07
2020/01/14
2020/01/21
2020/01/28
2020/02/04
2020/02/11
2020/02/18
2020/02/25
2020/03/03
2020/03/10
2020/03/17
2020/03/24
2020/03/31
2020/04/07
2020/04/14
H1
H2
H3
H4
H5
H6
H7
H8
other
n>=1000
n>=100
n>=10
n>=1


b
COL
the 9 specific sites
other mutation sites
H1
H2
H3
H4
0.0
0.1
0.2
0.3
0.4
0.5
0
10000
20000
30000
Pos

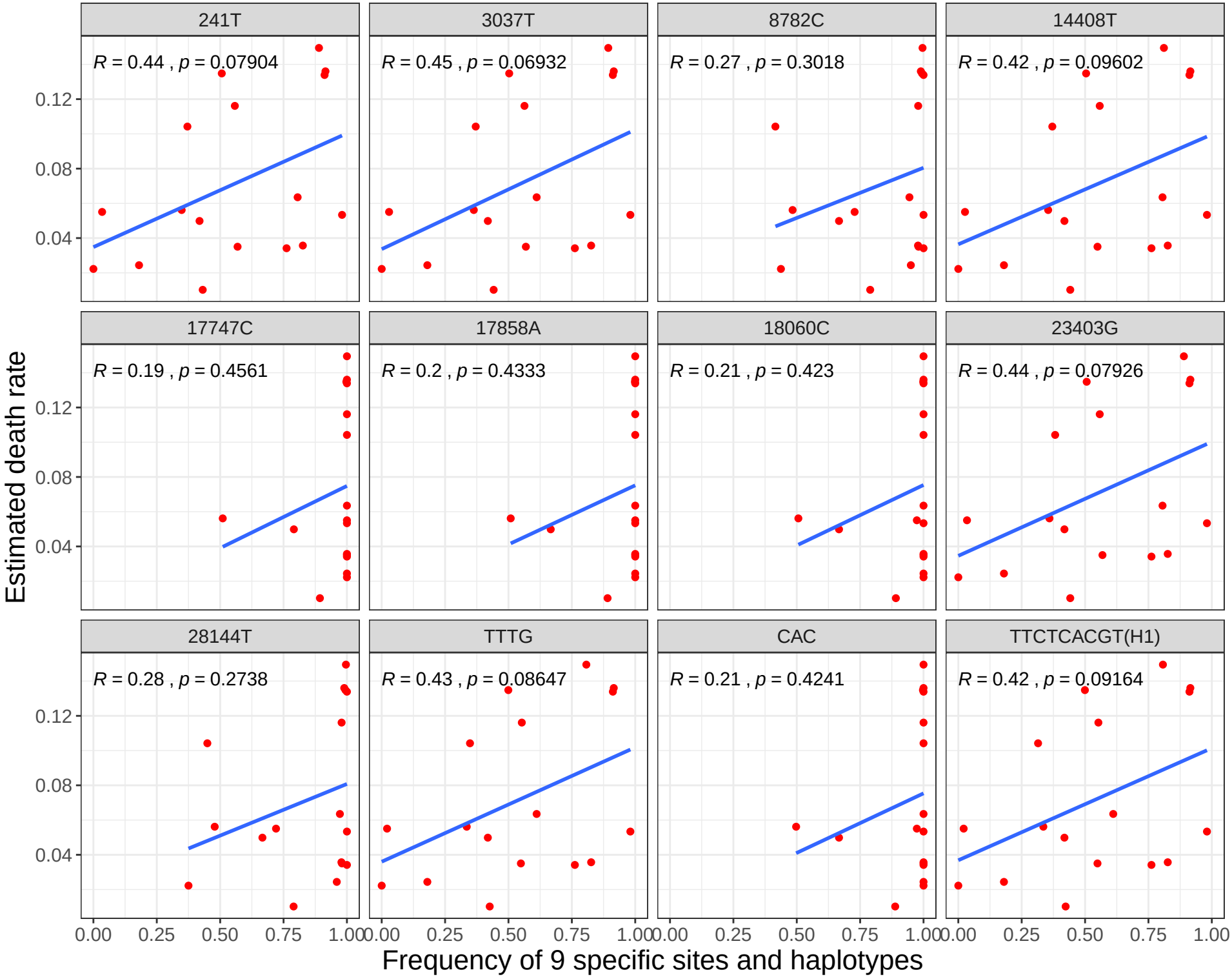

241T
R = 0.44 , p = 0.07904
3037T
R = 0.45 , p = 0.06932
8782C
R = 0.27 , p = 0.3018
14408T
R = 0.42 , p = 0.09602
17747C
R = 0.19 , p = 0.4561
17858A
R = 0.2 , p = 0.4333
18060C
R = 0.21 , p = 0.423
23403G
R = 0.44 , p = 0.07926
28144T
R = 0.28 , p = 0.2738
TTTG
R = 0.43 , p = 0.08647
CAC
R = 0.21 , p = 0.4241
TTCTCACGT(H1)
R = 0.42 , p = 0.09164
Estimated death rate
0.04
0.08
0.12
0.00
0.25
0.50
0.75
1.00
Frequency of 9 specific sites and haplotypes

Table 1. Statistics of Ka, Ks and Ka/Ks ratios for all coding regions of the SARS-CoV, MERS-CoV and SARS-CoV-2 genome sequences

	Ka (mean) 10^{-3}	***s.e. (Ka)10^{-3}***	***Ks (mean) 10^{-3}***	***s.e. (Ks)10^{-3}***	***Ka/Ks (mean)***	***s.e. (Ka/Ks)***	***Hypothesis***	***P-value****
SARS-CoV	0.985	0.018	1.310	0.049	0.760	0.038	Ka/Ks (SARS-CoV < SARS-CoV-2)	0.007
MERS-CoV	1.319	0.040	4.887	0.096	0.260	0.003	Ka/Ks (MERS-CoV < SARS-CoV-2)	0.000
SARS-CoV-2	0.231	0.004	0.265	0.005	1.008	0.020		

*One-sided T-test for the means of two independent samples.

Table 2. Substitution rate and tMRCA estimated by BEAST v2.6.2

	substitution rate (10^{-3})	***95% CI (10^{-3})***	***tMRCA***	***95% CI***	***references***
SARS-CoV	1.05	0.49, 1.65	2002-6-28	2002/1/19, 2002/11/3	this study
	-	0.80, 2.38	spring of 2002	-	[26]
MERS-CoV	1.52	1.39, 1.63	2012-6-7	2012/6/4, 2012/6/9	this study
	1.12	0.88; 1.37	2012-3	2011/12, 2012/6	[27]
SARS-CoV-2	1.6	1.42, 1.80	2019-9-27	2019/8/28, 2019/10/26	this study

26 Zhao, Z. et al. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC Evol Biol 4, 21, doi:10.1186/1471-2148-4-21 (2004).

27 Cotten, M. et al. Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. mBio 5, doi:10.1128/mBio.01062-13 (2014).

Table 3. The information of the 9 specific sites in each Cluster.

Pos	***Ref***	***Alt***	***Cluster 1***			***Cluster 2***			***Cluster 3***			***Fst***	***Gene region***	***Mutation type***	***Protein changed***	***codon changed***	***Impact***
			pi	Major allele	frequency	pi	Major allele	frequency	pi	Major allele	frequency						
241	C	T	0.0000	C	1.0000	0.0109	C	0.9945	0.0000	T	1.0000	0.991240	5'UTR	upstream	NA	NA	MODIFIER
3037	C	T	0.0000	C	1.0000	0.0109	C	0.9945	0.0000	T	1.0000	0.991240	gene-orf1ab	synonymous	924F	2772ttC>ttT	LOW
8782	C	T	0.0000	T	1.0000	0.3863	C	0.7390	0.0000	C	1.0000	0.582113	gene-orf1ab	synonymous	2839S	8517agC>agT	LOW
14408	C	T	0.0000	C	1.0000	0.0055	C	0.9973	0.0000	T	1.0000	0.995609	gene-orf1ab	missense	4715P>L	14144cCt>cTt	MODERATE
17747	C	T	0.0735	T	0.9620	0.0000	C	1.0000	0.0000	C	1.0000	0.975245	gene-orf1ab	missense	5828P>L	17483cCt>cTt	MODERATE
17858	A	G	0.0497	G	0.9747	0.0000	A	1.0000	0.0000	A	1.0000	0.983573	gene-orf1ab	missense	5865Y>C	17594tAt>tGt	MODERATE
18060	C	T	0.0000	T	1.0000	0.0164	C	0.9918	0.0000	C	1.0000	0.976100	gene-orf1ab	synonymous	5932L	17796ctC>ctT	LOW
23403	A	G	0.0000	A	1.0000	0.0164	A	0.9918	0.0000	G	1.0000	0.986895	gene-S	missense	614D>G	1841gAt>gGt	MODERATE
28144	T	C	0.0000	C	1.0000	0.3915	T	0.7335	0.0000	T	1.0000	0.578530	gene-ORF8	missense	84L>S	251tTa>tCa	MODERATE

Table 4. Statistics of Ka, Ks and Ka/Ks ratios for all coding regions of each 4 main haplotype genomes

	Ka (mean) 10^{-3}	***s.e. (Ka)10^{-3}***	***Ks (mean) 10^{-3}***	***s.e. (Ks)10^{-3}***	***Ka/Ks (mean)***	***s.e. (Ka/Ks)***	***Hypothesis***	***P-value*(Ka/Ks)***
H1	0.510	0.057	0.498	0.056	1.099	0.010	Ka/Ks (H3 < H1)	0.000
H2	0.347	0.042	0.334	0.044	1.268	0.023	Ka/Ks (H3 < H2)	0.000
H3	0.298	0.007	0.397	0.009	0.795	0.010	-	-
H4	0.334	0.023	0.414	0.019	0.796	0.019	-	-

*One-sided T-test for the means of two independent samples.