Clustering analysis of single nucleotide polymorphism data reveals population structure of SARS-CoV-2 worldwide

Yawei Li[1], Qingyun Liu[2], Yuan Luo[1*]

[1] Department of Preventive Medicine, Northwestern University, Feinberg School of Medicine, Chicago, IL 60611, USA

[2] Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA.

[*] Correspondence: yuan.luo@northwestern.edu

Abstract

The SARS-CoV-2 virus has been spreading rapidly and across the globe since first being reported in December 2019. To understand the evolutionary trajectory of the coronavirus, phylogenetic analysis is needed to study the population structure of SARS-CoV-2. As sequencing data worldwide is accruing rapidly, grouping them into clusters helps to organize the landscape of population structures. To effectively group these data, computational methodologies are needed to provide more productive and robust solutions for clustering. In this study, using the single nucleotide polymorphisms of the viral sequences as input features, we utilized three clustering algorithms, namely K-means, hierarchical clustering and balanced iterative reducing and clustering using hierarchies to partition the viral sequences into six major clusters. Comparison of the three clustering results reveals that the three methods produced highly consistent results, but K-means performed best and produced the smallest intra-cluster pairwise genetic distances among the three methods. The partition of the viral sequences revealed that the six clusters differed in their geographical distributions. Using comprehensive approaches to compare the diversity and selective pressure across the clusters, we discovered a high genetic diversity between the clusters. Based on characteristics of the mutation profiles in each cluster along with their geographical distributions and evolutionary histories, we identified the extent of molecular divergence within and between the clusters. The identification of the mutations that are strongly associated with clusters have potential implications for diagnosis and pathogenesis of COVID-19. In addition, the clustering method will enable further study of variant population structures in specific regions of these fast-growing viruses.

Introduction

Since the COVID-19 pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (1) was first reported in December 2019 (2), it is rapidly spreading throughout the world (3) with over 17 million cases and 0.65 million deaths by the end of July 2020. In the meantime, more than 70 thousand virus genomes have been published by the Global Initiative on Sharing All Influenza Data (GISAID) (4), helping researchers to better study the evolution of SARS-CoV-2.

As an emerging virus, very limited information is available regarding the genetic diversity, evolutionary trajectory and possible routes of transmission of SARS-CoV-2 from the natural reservoir to humans. Most studies have looked into the aspects of real-world SARS-CoV-2 evolution and strain diversification through phylogenetic trees (5-8). Phylogenetic tree is a graph that shows the evolutionary relationships among various biological entities based on their genetic closeness (9, 10). The distances of one sequence from the other sequences indicate the degree of relationships. As population genomic datasets grow in size, grouping the entities into clusters helps to simplify the analysis. A cluster is a subset of these entities such that the similarity among the entities in the subset is generally higher than the similarity among the objects in the full set. Traditionally, using the distance matrix and the branches of the phylogenetic tree, entities can be grouped into clusters. However, when the number of entities becomes large, it is not easy to directly and accurately partition the clades in the phylogenetic tree.

In order to identify the best way to effectively group the entities, clustering methods emerge as more productive and robust solutions. Clustering is the process of partitioning biological entities into meaningful clusters, based on distances in their phenotypic and genetic characteristics. The objective of clustering is automatically minimizing intra-cluster distances and maximizing inter-cluster distances (11). Accurate clustering helps to better understand the inner relationships between data and inform downstream analysis. Clustering methods have been widely used as a good supplementary tool in phylogenetic analysis, including phylogenetic tree construction (12-14), ancestral relationship identification (15), evolutionary rate estimation (16, 17), gene evolutionary mechanisms research (18) and population structure analysis (19).

To analyze the population structure of SARS-CoV-2, we focused on 16,873 public genomes collected from GISAID and used three clustering methods to group them into six clusters based on their mutation profiles. The partition of the viral sequences revealed that the six clusters showed specific geographic preferences. Using comprehensive approaches to analyze the diversity, selective pressure and evolutionary histories across the clusters, we identified the extent of molecular divergence within and between the clusters. This study provides a perspective of the SARS-CoV-2 population structural analysis, helping to investigate the evolutionary trajectory and transmission of the virus spreading across the human populations worldwide.

Materials and Methods

SARS-CoV-2 sample collection

A set of African, Asian, European, North American, Oceanian and South American SARS-CoV-2 genome sequences marked as "high coverage" were downloaded from GISAID. The "high coverage" was defined as entries with <1% Ns and <0.05% unique amino acid mutations (not seen in other sequences in databases) and no insertion/deletion unless verified by the submitter. In addition, all sequences with a non-human host and all assemblies of total genome length less than 29,000 bps were removed from our analysis. Ultimately, our dataset consisted of 16,873 sequences.

Mutation calls and phylogenetic reconstruction

All downloaded genomes were mapped to the complete genome of SARS-CoV-2 Wuhan-Hu-1 isolate (GenBank Accession Number: NC_045512.2). Multiple sequence alignments and pairwise alignments were constructed using CLUSTALW 2.1 (20). Considering many putatively artefactual mutations and the gaps in many sequences are located at the beginning and end of the alignment, we masked the first 130 bps and last 50 bps in mutation calling. We used substitutions as features to reconstruct the phylogenetic tree using FastTree 2 (21). The phylogeny is rooted following the Nextstrain pipeline (22) using FigTree v1.4.4 (23). The phylogenetic trees were visualized using the online tool Interactive Tree Of Life (iTOL v5) (24).

Region analysis and data visualization

For each country with SARS-CoV-2 data available, clustering proportions were calculated and plotted on the world map using the tool Tableau Desktop 2020.2. Other figures and statistical analyses were generated by the ggplot2 library in R 3.6.1, the seaborn package in Python 3.7.6 and GraphPad Prism 8.0.2.

Data clustering

In this study, we used K-means clustering, hierarchical clustering and balanced iterative reducing and clustering using hierarchies (BIRCH) clustering to partition the SARS-CoV-2 strains.

The K-means clustering (25) is one of the most common unsupervised machine learning algorithms. It aims to partition N observations into a pre-specified cluster number (K) in which each observation belongs to the cluster with the nearest center (cluster centroid, serving as a prototype of the cluster). The K-means algorithm starts with randomly selected centroids as the starting prototypes for each cluster, and then performs iterative calculations to compute the new center of each cluster (i.e. centroid

update) until the centroids become stabilized or the defined number of iterations has been achieved. The model was implemented using the Python package sklearn with the KMeans function.

Another unsupervised method, hierarchical clustering does not require the user to pre-specify the number of clusters to be generated as K-means does. Hierarchical clustering starts by treating each observation as a separate cluster. It then iteratively executes the following two steps: (a) identify the two clusters that are closest together, and (b) merge the two closest clusters. This iterative process continues until all the clusters are merged. The model was implemented using the Python package sklearn with the AgglomerativeClustering function.

BIRCH (26, 27) is used to perform hierarchical clustering over particularly large data-sets. It obtains the clustering-feature summarization of the dataset by incrementally building a height-balanced clustering feature (CF) tree. A CF is defined as a tuple (*N*, LS, SS), where *N* represents the number of sample points in the CF, LS represents the vector sum of the feature dimensions of the sample points in the CF and SS is the square sum of the feature dimensions of the sample points in the CF. As such, the clustering features are organized in a CF tree and each leaf node is a subcluster. Each new sample descends the tree by following its closest CF until a leaf node is reached and merges into the leaf node. The distance between the incoming sample and the existing subclusters is limited by the parameter threshold. If the radius of the subcluster obtained by merging the new sample and the nearest subcluster is greater than the square of the threshold, the two farthest subclusters are taken and the subclusters are separated into two clusters on the basis of the distance between these subclusters. After the CF tree is built, a selected clustering algorithm is applied to cluster these subclusters into global clusters (the global clustering step). The selected clustering algorithm is adapted to work with a set of subclusters, rather than a set of data points. The model was implemented using the Python package sklearn with the Birch function.

Simpson's diversity index

Simpson's Diversity Index (*D*) is a measure of diversity that takes into account the number of entities as well as their abundance. The index measures the probability that two randomly selected individuals are the same. The formula for calculating the value of the index is:

$$\mathrm{D} = 1 - \frac{\sum_{all\ traits} n(n-1)}{N(N-1)}$$

where n is the number of individuals displaying one trait and N is the total number of all individuals. The value of *D* ranges between 0 and 1. With this index, 1 represents infinite diversity and 0 denotes no diversity.

Pairwise dN/dS ratio

To test whether there is positive selection or negative selection in each cluster that accelerates the divergence of the virus genomes, we used the pairwise nonsynonymous to synonymous substitution ratio (pairwise dN/dS ratio) (28) to evaluate the selective pressure within clusters. The dN/dS ratio is used to compare two real sequences that exist in nature, and is calculated as the ratio of the number of nonsynonymous substitutions per non-synonymous site (dN), in a given period of time, to the number of synonymous substitutions per synonymous site (dS), in the same period. Pairwise dN/dS ratio is similar to that of standard dN/dS ratio, but this ratio is calculated using all polymorphic sites within clusters to express both within-cluster substitutions and transient polymorphisms. To avoid having undetermined pairwise dN/dS values due to dN or dS being zero, a mean dN/dS value was then calculated for each sequenced isolate by dividing its mean pairwise dN by its mean pairwise dS with respect to all other sequenced isolates within each cluster. The estimated dN and dS were estimated using R package seqinr with the kaks function and the final pairwise dN/dS ratio was calculated using Python 3.7.6.

Inferring positive/purifying selection of individual sites

To test which position was under selective pressure, we used a set of programs available in HyPhy (29) to calculate nonsynonymous (dN) and synonymous (dS) substitution rates on a per-site basis to infer pervasive selection. Fast Unconstrained Bayesian AppRoximation (FUBAR) (30) was applied to detect overall sites under positive selection. The positively selected sites were identified using a probability larger than 0.95 using the FUBAR method.

Pairwise mutation dependency score

To trace the evolutionary history within a cluster, we calculated the pairwise mutation dependency scores for polymorphic mutations. For two selected mutations X and Y, the scores S(X|Y) and S(Y|X) can be calculated using the following functions:

$$\mathrm{S(X|Y)} = \frac{\sum_{all\ samples} S_X = 1\ \&\ S_Y = 1}{\sum_{all\ samples} S_Y = 1}$$

$$\mathrm{S(Y|X)} = \frac{\sum_{all\ samples} S_X = 1\ \&\ S_Y = 1}{\sum_{all\ samples} S_X = 1}$$

where $S_X = 1$ denotes that the sequence has a mutation X. Pairwise mutation dependency score displays the correlation and the timescale relationship of the two mutations. The value of S(X|Y) and S(Y|X) ranges between 0 and 1. With this index,

S(X|Y) = 1 with S(Y|X) < 1 represents that mutation Y occurs after mutation X. In contrast, S(X|Y) = 1 with S(Y|X) = 1 represents that the two mutations occur simultaneously and are genetically linked. Statistical analyses and data presentations were generated using Python 3.7.6.

Results

General statistics for SARS-CoV-2 genomes

We obtained 16,873 (98 from Africa, 1324 from Asia, 9527 from Europe, 4765 from North America, 1040 from Oceania and 119 from South America) SARS-CoV-2 whole-genome sequencing data from GISAID, aligned the sequences, and identified the genetic variants. We used the SARS-CoV-2 isolate Wuhan-Hu-1 complete genome (GenBank Accession Number: NC_045512.2) as the reference genome. A total of 7,970 substitutions were identified, including 4,908 non-synonymous mutations, 2,748 synonymous mutations and 314 intronic mutations. The average mutation count per genome was 6.99 (Figure S1). The frequency spectrum of substitutions illustrated that more than half (54.05%) of the mutations were singletons and 15.35% were doubletons. The proportion of the mutations below 0.01 was 99.28% (Figure S2). The high percentage of these low-frequency mutations suggested that SARS-CoV-2 occurred recently and displayed rapidly proliferating pattern (31). In addition, there were 8,706 unique genomes across the 16,873 strains (Figure S3), and most unique genomes (7,078) were singletons, yielding high diversity of the virus genomes. In particular, Simpson's diversity index of the strains was 0.8222, indicating that two random strains would have a high probability of being genetically different.

Clustering of SARS-CoV-2 reveals six major clusters

Grouping the strains into clusters will help us analyze the population structure of the virus. However, the general statistics for SARS-CoV-2 genomes indicated high genetic diversity within the virus, thus it is not easy to directly and accurately partition the strains. For this reason, we applied clustering techniques to the measure similarity between these strains and effectively group them.

We first employed the K-means clustering method. The K-means algorithm requires one to pre-specify the number of clusters (*K*), but we have little prior knowledge about the number of subtypes formed by the heterogeneous SARS-CoV-2 genomes. To determine the number of clusters, we plotted the curves of the sum of squared errors (SSE) and Bayesian information criterion (BIC) (32) under different cluster numbers ranging from 2 to 12 (Figure S4). We used the elbow method and choose the elbow of the curve as the number of clusters (33). This approach resulted in K=6 for both SSE and BIC curves. We also employed two other clustering methods:

hierarchical clustering and BIRCH clustering to group the strains, and compared their results with those of K-means. The confusion matrix indicated that the similarity of K-means and BIRCH exceeded 99% (Figure 1A), whereas the similarity of K-means and hierarchical clustering was 90% (Figure 1B). The differences between K-means and hierarchical clustering results may be due to their different algorithms. The merges and splits of hierarchical clustering are determined in a greedy manner. Each step of the algorithm is based on local decisions, merging the most "compact" cluster available (34). The compactness criterion (or how close together objects are) is defined by the linkage criteria and proximity measure used. On the other hand, K-means maximizes an objective that is a combination of cluster compactness and cluster separation (35).

To determine which method is best, we aligned the partitions of the six clusters against the phylogenetic tree for the three methods (Figure 1C). The differences between the hierarchical clustering results and the two other clustering results were mainly on the boundary of the clusters. Of the three methods, strains clusters by K-means and BIRCH were more compact in the phylogenetic tree than those by hierarchical clustering. For example, the strains in both K-means cluster D and BIRCH cluster D were split into two clusters by hierarchical clustering. However, such split was not supported by the phylogenetic tree (Figure 1C). The objective of clustering is minimizing intra-cluster distances and maximizing inter-cluster distances. Therefore, we calculated the intra-cluster pairwise genetic distances. The average number of genetic distances for K-means, hierarchical clustering and BIRCH were 4.886, 5.062 and 4.900, respectively. The average number of genetic distances of K-means was significantly lower than that of hierarchical clustering (P-value < 0.001, Wilcoxon rank-sum test) and BIRCH (P-value < 0.001, Wilcoxon rank-sum test). In general, K-means exhibited the best performance compared to the other methods.

In the meantime, we used complementary approaches to validate the K-means clustering results. First, we compared the pairwise genetic distances between intra-cluster and inter-cluster. In all six clusters, the average number of intra-cluster genetic distances was significantly lower (P-value < 0.001, Wilcoxon rank-sum test, Figure 1D) than inter-cluster. Second, we applied T-distributed Stochastic Neighbor Embedding (t-SNE) to visualize the K-means clustering results. In the t-SNE plot, the strains were well isolated between clusters (Figure 1E).

The specific geographical distributions of SARS-CoV-2 across clusters

For each cluster, the proportions of strains from different continents show that the clusters differ in their geographical distributions (Figure 2, Table S1). Of the six clusters, cluster C spreads globally. By contrast, cluster A and cluster F occur at high frequencies in specific regions. 81.92% of the strains in cluster A and 85.73% of the strains in cluster F are from Europe. The geographical spread of each of the three remaining clusters is

intermediate. Cluster E occurs at higher frequencies in North America and Europe, and lower frequencies in Asia and Oceania. Cluster D occurs at higher frequencies in North America, and lower frequencies in Asia, Europe and Oceania. The strains in cluster B are mainly in Asia and Europe and partially in North America and Oceania. However, considering the sampling bias of the SARS-CoV-2 (Figure S5), we evaluated the concentration of the six continents in the clusters. Except for the strains collected from Oceania that distribute uniformly in the six clusters, the distributions of the strains collected from the other continents are concentrated in one or two clusters, including Asia (49% in cluster B), Africa (66% in cluster C), South America (78% in cluster C and F), North America (74% in cluster D and E) and Europe (64% in cluster C and F), see Table S1. It is noteworthy that the clusters in the USA also display geographic preferences (Figure S6). Strains in cluster D are mainly on the west coast and strains in cluster E are mainly on the east coast.

The diversity and the variance of selective pressures between clusters

The average mutation counts for the six clusters were 6.38, 3.49, 6.57, 7.09, 7.89 and 8.96 (Figure S7), respectively. Considering the different collection dates (Figure 3A) of the strains, mutation rates instead of mutation counts were more effective for describing the genetic variations between clusters. We defined the date 24 December 2019 when the earliest strain (EPI_ISL_402123) was collected as the index day. The average mutation rates for the six clusters were 25.55, 15.91, 25.44, 31.64, 30.99 and 34.12 substitutions per year. Specifically, the average mutation rate in cluster B was significantly lower (P-value < 0.001, Wilcoxon rank-sum test) than those in other clusters. In contrast, the average mutation rate in cluster F was significantly higher (P-value < 0.001, Wilcoxon rank-sum test) than those in other clusters. The Simpson's diversity indexes of the six clusters were 0.7616, 0.7608, 0.8398, 0.8466, 0.8082 and 0.8502, respectively. Both the average mutation rate and Simpson's index were highest in cluster F, suggesting that the diversity of cluster F was higher than the other clusters. The nucleotide diversities per site for the six clusters were 0.0196%, 0.0222%, 0.0171%, 0.0256%, 0.0131% and 0.0132%. The high mutation rates but low nucleotide diversity in cluster E and cluster F suggests that these two clusters may have more fixed mutations than the other clusters. We then calculated the nucleotide diversity of each gene across all clusters (Figure 3B-G). Except for some short genes that are unlikely to be informative, the diversity of most genes was close to the diversity of their genome-wide variants.

To test whether these clusters were subject to positive selection or purifying selection, we calculated the pairwise ratios of rates of nonsynonymous to synonymous substitutions (pairwise dN/dS, see Materials and Methods) in each cluster. Typically, purifying selection tends to remove amino acid-altering mutations to reduce dN/dS, and

303 positive selection has the opposite effect that increases dN/dS. The average pairwise
304 dN/dS in cluster B (P-value = 8.42×10^{-8}, Wilcoxon rank-sum test) was significantly
305 higher than 1, and similarly for cluster E (P-value = 1.35×10^{-4}, Wilcoxon rank-sum test),
306 suggesting positive selection in these two clusters (Figure S8). In contrast, the pairwise
307 dN/dS ratios in the other four clusters were significantly lower than 1 (P-value < 0.001,
308 Wilcoxon rank-sum test), suggesting that the nonsynonymous mutations tended to be
309 selected against. In particular, the pairwise dN/dS in cluster D was much lower than
310 those in the other clusters.
311
312 **Identify significant mutations for SARS-CoV-2 clustering**
313 To analyze the divergence and track the evolutionary history of clusters, we applied
314 the analysis of variance (ANOVA) to identify the statistically significant mutations that
315 characterize the clustering results. Across the 7,970 substitutions, 26.27% (2,094
316 substitutions) of them achieved P-values < 0.05 (Figure S9). We selected the 2%
317 (42/2094) substitutions that achieved the lowest P-values (Table 1) and analyzed their
318 distributions in the clusters. Of the 42 substitutions, there were 25 nonsynonymous
319 mutations, 15 synonymous mutations and 2 intronic mutations. However, mutation
320 G28882A was not a synonymous mutation. It was a trinucleotide mutation from
321 position 28881 to 28883 that spans two codons and results in an RG (arginine-glycine)
322 to KR (lysine-arginine) amino acid change (36, 37). The strains with this trinucleotide
323 mutation were all in cluster F in the K-means clustering results. In addition to the
324 trinucleotide mutation, many of the extracted substitutions were reported to be related
325 to the evolution of SARS-CoV-2 (38-41). Tang et al (42) defined "L" type (defined as
326 "L" type because T28,144 is in the codon of Leucine) and "S" type (defined as "S" type
327 because C28,144 is in the codon of Serine) of SARS-CoV-2 through two genetically
328 linked mutations C8782T and T28144C. In the meantime, Forster et al (5) used these
329 two mutations to define one of the three subtypes in their study. The proportion of "S"
330 type was only 11.36% and mainly located in North America in our downloaded strains,
331 indicating that "L" type was more prevalent and transmissible than "S" type. The "S"
332 type strains were almost exclusively in cluster B according to the K-means clustering
333 results. Another mutation A23403G (D614G, Asparticacid to Glycine) in the *spike*
334 protein domains was reported to show significant variation in cytopathic effects and
335 viral load, and substantially change the pathogenicity of SARS-CoV-2 (43).
336 Furthermore, this mutation was accompanied by two other mutations: a silent mutation
337 T3037C, and a mutation T14408C which results in a RNA-dependent RNA polymerase
338 (RdRp) amino acid change (44).
339 Previous studies have reported that recombination is common in coronavirus
340 evolution (45-47). Given that recombinations in SARS-CoV-2 may perturb the
341 clustering, we used Haploview (48) to analyze the linkage disequilibrium (LD) by

342 calculating standardized disequilibrium coefficients (D') and squared allele-frequency
343 correlations (r^2) of the 42 substitutions. D′ is affected solely by recombination and not
344 by differences in allele frequencies between sites, and r^2 is also affected by differences
345 in allele frequencies at the two sites (49). In the heatmap of D' (Figure S10) and r^2
346 (Figure S11), most substitution pairs showed very high D' and low r^2, indicating no
347 recombination between the SARS-CoV-2 strains.
348 To assess the selective pressure of these substitutions, we used HyPhy to infer the
349 probabilities of positive selection across the 42 substitutions. There are nine mutations
350 (asterisks in Table 1) with a positive probability >0.95. Of the inferred mutations,
351 mutations G2891A, G11083T, C14408T, C17747T and A23403G were reported as
352 recurrent mutations (39, 50). The recurrence of these mutations agrees with the
353 assumption that they may confer selective advantages in the population. In addition, the
354 inferred selective advantaged mutation G25563T may also be a recurrent mutation: the
355 earliest sequence carrying both C1059 T and G25563T mutations was discovered on 21
356 February 2020, but the earliest sequence carrying only G25563T was reported five days
357 later.
358
359 **Tracking evolutionary trajectory of SARS-CoV-2 across clusters**
360 Integrating the geographic distributions of the extracted mutations with their daily
361 running counts, we can track the evolutionary trajectory of SARS-CoV-2 and its
362 divergence between the clusters. Figure 4 displays the frequencies of the mutations in
363 each cluster and their daily counts. The clustering results show that cluster C, cluster E
364 and cluster F are closer than the other clusters as they share four common fixed
365 mutations: A23403G (D614G), C241T, C3037T and C14408T. As the dominant
366 mutations in the world, the four genetically linked mutations are high frequency in all
367 continents in our downloaded sequences, including South America (87%), Africa (86%),
368 Europe (75%), North America (65%), Oceania (55%) and Asia (32%). The earliest time
369 when sequences carrying mutation A23403G were collected was in late January 2020.
370 About a month later, these mutations were discovered throughout the world. Cluster E
371 is grouped as an independent cluster by K-means as strains in this group that accumulate
372 two mutations: C1059T and G25563T. These two mutations were first collected on 21
373 February 2020. Compared with cluster C and cluster E, cluster F has three unique fixed
374 mutations: the trinucleotide mutation from position 28881 to position 28883, which was
375 initially discovered on 24 February 2020.
376 In contrast to cluster C, cluster E and cluster F, there are no common fixed
377 mutations between the other three clusters. Cluster D has two fixed mutations (C8782T,
378 T28144C). Mutations C8782T and T28144C were two of the earliest mutations that
379 were collected in early January. Cluster A has three fixed mutations: G11083T,
380 G14805T and G26144T. Two of the three mutations (G11083T, G26144T) were first

collected in late January, and the third one was discovered about two weeks later. In addition, we found that mutation G14805T may be a recurrent mutation. The earliest sequence carrying both C2558T and G14805T mutations was discovered on 9 February 2020, but the earliest sequence carrying only G14805T was discovered about ten days later. Furthermore, the standardized disequilibrium coefficients (D') of mutation G14805T with two genetically linked mutations C8782T and T28144C are very low (Figure S10), which is consistent with our assumption that G14805T is a recurrent mutation.

Of the six clusters, cluster B is the only cluster that has no fixed mutations, suggesting that the population structure in cluster B is complex. We calculated the pairwise dependency scores (see Materials and Methods) of all the mutations with frequencies >0.05 in cluster B. Figure S12 shows that these mutations can group the cluster B into five subclusters. We mapped the geographical distributions of the five subclusters and found that these subclusters were dispersed across many countries. The highest frequency mutation G11083T was isolated into two subclusters: the subcluster B1 (G1397A, T28688C and G29742T) was mainly in Australia, Turkey and India, and the subcluster B2 (C6312A, C13730T, C23929T and C28311T) was mainly in Singapore, India and Australia. The subcluster B3 containing three mutations (G1440A, G2891A and G28851T) was concentrated in the United Kingdom. The subcluster B4 (T514C) was mostly collected from the Netherlands, and the remaining subcluster B5 (G29711T) was almost exclusively in the USA.

Discussion

Understanding of the evolutionary trajectory of SARS-CoV-2 is important in evaluating future risks of novel infections. To analyze the population structure and diversity of the pandemic viruses, we cataloged 16,873 publicly available SARS-CoV-2 genomes and applied multiple clustering methods to partition the genomes. Using the SNPs as input features and three different clustering methods: K-means clustering, hierarchical clustering and BIRCH, we consistently identified six major clusters of SARS-CoV-2 strains. By comparing the distributions of the clustering results in the phylogenetic tree and intra-cluster pairwise genetic distances across the three methods, we found that K-means exhibited the best performance.

The mutation profile characterized clusters of the viral sequences displayed specific geographic preferences. The proportion of the dominant continent exceeds 60% in all clusters except for cluster B, with the highest being the Europe proportion of 85.73% in cluster F. The strains collected from all continents are mainly distributed in one or two clusters apart from Oceania. To evaluate the heterogeneity between clusters, we identified the important substitutions that characterize the clusters, investigated the

LD structure and the selective pressure of the substitutions, and tracked their evolutionary trajectory. Our results indicated high genetic diversity in SARS-CoV-2, which was consistent with an earlier study (51). The topology and the divergence of the clusters in the phylogenetic tree illustrated a relatively recent common ancestor, similar to the fact that the emergence and the spread of the virus was highly concentrated in a short time (2, 52-54).

However, due to the sampling bias of SARS-CoV-2, more than 60% of the strains are from the United Kingdom and the USA. In contrast, the overall proportion of strains from Africa and South America is less than 2% (Figure S5). Sampling biases can lead to biased parameter estimation and affect the clustering results we observed. For example, the frequency of mutation C15324T reached 41.84% in Africa, but only 2.21% outside Africa. The frequency of mutation T29148C reached 15.13% in South America, but only 0.12% outside South America. Another mutation T27299C with frequency 10.92% in South America is only found with frequency 0.08% in other regions. More than 95% of strains accumulating mutation C15324T are in group C, and mutations T29148C and T27299C are almost exclusively in group F, indicating these mutations are highly concentrated. However, due to the small proportion of the strains from these two continents, these mutations were unable to be extracted by ANOVA. To address this issue, collecting more sequences from these continents is needed.

Despite the limited number of SARS-CoV-2 genome sequences, our work, as well as previous studies (55, 56) that used clustering techniques to study the evolution of the SARS-CoV-2 virus, has shown to be a good supplement tool in phylogenetic analyses. In addition, the clustering methods can be used for further study of variant population structures in specific regions of these fast-growing viruses.

Competing Interests

The authors declare that they have no conflict of interests.

References

1. A. E. Gorbalenya *et al.*, Severe acute respiratory syndrome-related coronavirus: The species and its viruses – a statement of the Coronavirus Study Group. *bioRxiv* 10.1101/2020.02.07.937862, 2020.2002.2007.937862 (2020).
2. N. Zhu *et al.*, A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **382**, 727-733 (2020).
3. R. Li *et al.*, Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489-493 (2020).

4. Y. L. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Eurosurveillance* **22**, 2-4 (2017).
5. P. Forster, L. Forster, C. Renfrew, M. Forster, Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A* **117**, 9241-9243 (2020).
6. T. Koyama, D. Platt, L. Parida, *Variant analysis of COVID-19 genomes* (2020), 10.2471/BLT.20.253591.
7. S. U. Rehman, L. Shafique, A. Ihsan, Q. Liu, Evolutionary Trajectory for the Emergence of Novel Coronavirus SARS-CoV-2. *Pathogens* **9** (2020).
8. T.-Y. Yeh, G. Contreras, *Faster de novo mutation of SARS-CoV-2 in shipboard quarantine* (2020), 10.2471/BLT.20.255752.
9. G. Mahapatro, D. Mishra, K. Shaw, S. Mishra, T. Jena, Phylogenetic Tree Construction for DNA Sequences using Clustering Methods. *Procedia Engineer* **38**, 1362-1366 (2012).
10. A. Sharma, S. Jaloree, R. Thakur, "Review of Clustering Methods: Toward Phylogenetic Tree Constructions". (2018), 10.1007/978-981-10-8198-9_50, pp. 475-480.
11. T. F. Gonzalez, Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science* **38**, 293-306 (1985).
12. J. Wang, J. Soininen, J. He, J. Shen, Phylogenetic clustering increases with elevation for microbes. *Environ Microbiol Rep* **4**, 217-226 (2012).
13. J. Ning, R. G. Beiko, Phylogenetic approaches to microbial community classification. *Microbiome* **3**, 47 (2015).
14. D. Fioravanti *et al.*, Phylogenetic convolutional neural networks in metagenomics. *Bmc Bioinformatics* **19**, 49 (2018).
15. L. Qin, Y. X. Chen, Y. Pan, L. Chen, A novel approach to phylogenetic tree construction using stochastic optimization and clustering. *Bmc Bioinformatics* **7** (2006).
16. J. Felsenstein, G. A. Churchill, A hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution* **13**, 93-104 (1996).
17. A. Siepel *et al.*, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050 (2005).
18. M. H. Medema, P. Cimermancic, A. Sali, E. Takano, M. A. Fischbach, A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Comput Biol* **10**, e1004016 (2014).
19. E. Han *et al.*, Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nat Commun* **8**, 14238 (2017).
20. M. A. Larkin *et al.*, Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).

21. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
22. J. Hadfield *et al.*, Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123 (2018).
23. A. Rambaut (2012) FigTree v1. 4.
24. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127-128 (2007).
25. J. MacQueen (1967) Some methods for classification and analysis of multivariate observations. in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (University of California Press, Berkeley, Calif.), pp 281-297.
26. T. Zhang, R. Ramakrishnan, M. Livny (1996) BIRCH: an efficient data clustering method for very large databases. in *Proceedings of the 1996 ACM SIGMOD international conference on Management of data* (Association for Computing Machinery, Montreal, Quebec, Canada), pp 103–114.
27. T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: A New Data Clustering Algorithm and Its Applications. *Data Mining and Knowledge Discovery* **1**, 141-182 (1997).
28. D. Stucki *et al.*, Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat Genet* **48**, 1535-1543 (2016).
29. S. L. K. Pond *et al.*, HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Molecular Biology and Evolution* **37**, 295-299 (2020).
30. B. Murrell *et al.*, FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. *Molecular Biology and Evolution* **30**, 1196-1205 (2013).
31. W. B. Yu, G. D. Tang, L. Zhang, R. T. Corlett, Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2 / HCoV-19) using whole genomic data. *Zool Res* **41**, 247-257 (2020).
32. G. Schwarz, Estimating the Dimension of a Model. *Ann. Statist.* **6**, 461-464 (1978).
33. R. L. Thorndike, Who Belongs in the Family? *Psychometrika* **18**, 267-276 (1953).
34. A. K. Jain, R. C. Dubes, *Algorithms for clustering data*, Prentice Hall advanced reference series (Prentice Hall, Englewood Cliffs, N.J., 1988), pp. xiv, 320 p.
35. M. C. P. de Souto, I. G. Costa, D. S. A. de Araujo, T. B. Ludermir, A. Schliep, Clustering cancer gene expression data: a comparative study. *Bmc Bioinformatics* **9** (2008).
36. M. Ayub, *Reporting Two SARS-CoV-2 Strains Based on A Unique Trinucleotide-*

534 *Bloc Mutation and Their Potential Pathogenic Difference* (2020),
535 10.20944/preprints202004.0337.v1.
536 37. C. B. F. Vogels *et al.*, Analytical sensitivity and efficiency comparisons of
537 SARS-COV-2 qRT-PCR primer-probe sets. *medRxiv*
538 10.1101/2020.03.30.20048108, 2020.2003.2030.20048108 (2020).
539 38. H. Cai, K. Cai, J. Li (2020) Identification of Novel Missense Mutations in a
540 Large Number of Recent SARS-CoV-2 Genome Sequences. (Preprints.org).
541 39. L. van Dorp *et al.*, Emergence of genomic diversity and recurrent mutations in
542 SARS-CoV-2. *Infect Genet Evol* **83**, 104351 (2020).
543 40. X. YANG, N. Dong, W.-C. CHAN, S. CHEN, Identification of super-
544 transmitters of SARS-CoV-2. *medRxiv* 10.1101/2020.04.19.20071399,
545 2020.2004.2019.20071399 (2020).
546 41. C. Yin, Genotyping coronavirus SARS-CoV-2: methods and implications.
547 *Genomics* 10.1016/j.ygeno.2020.04.016 (2020).
548 42. X. Tang *et al.*, On the origin and continuing evolution of SARS-CoV-2. *National*
549 *Science Review* 10.1093/nsr/nwaa036 (2020).
550 43. H. Yao *et al.*, Patient-derived mutations impact pathogenicity of SARS-CoV-2.
551 *medRxiv* 10.1101/2020.04.14.20060160, 2020.2004.2014.20060160 (2020).
552 44. B. Korber *et al.*, Spike mutation pipeline reveals the emergence of a more
553 transmissible form of SARS-CoV-2. *bioRxiv* 10.1101/2020.04.29.069054,
554 2020.2004.2029.069054 (2020).
555 45. R. L. Graham, R. S. Baric, Recombination, reservoirs, and the modular spike:
556 mechanisms of coronavirus cross-species transmission. *J Virol* **84**, 3134-3146
557 (2010).
558 46. S. U. Rehman, L. Shafique, A. Ihsan, Q. Y. Liu, Evolutionary Trajectory for the
559 Emergence of Novel Coronavirus SARS-CoV-2. *Pathogens* **9** (2020).
560 47. E. Rouchka, J. Chariker, D. Chung (2020) Phylogenetic and Variant Analysis of
561 1,040 SARS-CoV-2 Genomes. (Preprints.org).
562 48. J. C. Barrett, B. Fry, J. Maller, M. J. Daly, Haploview: analysis and visualization
563 of LD and haplotype maps. *Bioinformatics* **21**, 263-265 (2005).
564 49. D. L. Remington *et al.*, Structure of linkage disequilibrium and phenotypic
565 associations in the maize genome. *Proc Natl Acad Sci U S A* **98**, 11479-11484
566 (2001).
567 50. M. Pachetti *et al.*, Emerging SARS-CoV-2 mutation hot spots include a novel
568 RNA-dependent-RNA polymerase variant. *J Transl Med* **18**, 179 (2020).
569 51. X. Li *et al.*, Transmission dynamics and evolutionary history of 2019-nCoV. *J*
570 *Med Virol* **92**, 501-511 (2020).
571 52. J. F. W. Chan *et al.*, A familial cluster of pneumonia associated with the 2019
572 novel coronavirus indicating person-to-person transmission: a study of a family

cluster. *Lancet* **395**, 514-523 (2020).

53. J. Sun *et al.*, COVID-19: Epidemiology, Evolution, and Cross-Disciplinary Perspectives. *Trends Mol Med* **26**, 483-495 (2020).

54. P. Zhou *et al.*, A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273 (2020).

55. X. Li *et al.*, Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. *J Med Virol* **92**, 602-611 (2020).

56. A. Mishra *et al.*, Mutation landscape of SARS-CoV-2 reveals three mutually exclusive clusters of leading and trailing single nucleotide substitutions. *bioRxiv* 10.1101/2020.05.07.082768, 2020.2005.2007.082768 (2020).

Table 1. The information of the 42 mutations using ANOVA.

Mutation	Substitution	Amino Acid Substitution	Type	GENE	Frequency	Cluster					
						A	B	C	D	E	F
C241T	C > T	Intron	Intron	Intron	66.37%	10	10	4238	2	3548	3391
T490A	T > A	D > E	N	ORF1ab	1.04%	0	0	1	174	0	0
T514C	T > C	H > H	S	ORF1ab	0.97%	0	162	1	0	0	0
C1059T*	C > T	T > I	N	ORF1ab	21.69%	1	8	2	0	3645	3
G1397A	G > A	V > I	N	ORF1ab	1.12%	0	186	0	0	1	2
G1440A	G > A	G > D	N	ORF1ab	1.92%	0	324	0	0	0	0
A2480G	A > G	I > V	N	ORF1ab	3.60%	608	0	0	0	0	0
C2558T	C > T	P > S	N	ORF1ab	3.83%	646	1	0	0	0	0
G2891A*	G > A	A > T	N	ORF1ab	1.77%	0	298	0	0	0	0
C3037T	C > T	F > F	S	ORF1ab	67.26%	2	7	4277	3	3611	3448
C3177T	C > T	P > L	N	ORF1ab	1.05%	0	0	1	171	6	0
C6312A	C > A	T > K	N	ORF1ab	1.14%	0	189	1	0	0	3
C8782T	C > T	S > S	S	ORF1ab	11.42%	1	21	5	1898	1	1
T9477A	T > A	F > Y	N	ORF1ab	1.17%	0	3	0	195	0	0
G11083T*	G > T	L > F	N	ORF1ab	11.81%	1342	485	52	21	54	39
C14408T*	C > T	P > L	N	ORF1ab	67.47%	1	8	4301	2	3636	3436
C14805T	C > T	Y > Y	S	ORF1ab	9.39%	1352	8	1	195	0	28
T17247C	T > C	R > R	S	ORF1ab	3.00%	500	5	1	0	0	0
C17747T*	C > T	P > L	N	ORF1ab	6.92%	1	0	0	1165	1	0
A17858G	A > G	Y > C	N	ORF1ab	7.05%	1	1	0	1187	0	0
C18060T	C > T	L > L	S	ORF1ab	7.16%	0	3	2	1202	1	0
T18736C	T > C	F > L	N	ORF1ab	1.01%	0	0	1	169	0	0
C18877T	C > T	L > L	S	ORF1ab	2.67%	2	2	440	4	0	2
A20268G	A > G	L > L	S	ORF1ab	4.61%	0	1	773	3	0	1
A23403G*	A > G	D > G	N	S	67.65%	4	4	4316	6	3634	3451
C23731T	C > T	T > T	S	S	1.68%	0	0	0	0	1	282
C23929T	C > T	Y > Y	S	S	1.13%	0	186	1	0	1	2
C24034T	C > T	N > N	S	S	1.16%	0	2	1	187	4	1
G25563T*	G > T	Q > H	N	ORF3a	26.44%	1	3	829	2	3625	2
G25979T	G > T	G > V	N	ORF3a	1.16%	0	2	1	193	0	0
G26144T*	G > T	G > V	N	ORF3a	8.61%	1387	62	0	1	1	1
T26729C	T > C	A > A	S	M	1.07%	0	1	1	179	0	0
C27046T	C > T	T > M	N	M	2.13%	0	1	5	0	0	353
G28077C	G > C	V > L	N	ORF8	1.13%	0	1	1	188	0	0
T28144C*	T > C	L > S	N	ORF8	11.36%	0	10	1	1903	2	0
C28657T	C > T	D > D	S	N	1.21%	0	3	3	196	1	2
T28688C	T > C	L > L	S	N	1.07%	0	178	1	0	1	0
C28863T	C > T	S > L	N	N	1.19%	1	2	2	193	2	0
G28881A	G > A	R > K	N	N	20.54%	4	3	3	1	1	3453
G28882A	G > A	R > K[1]	N	N	20.49%	1	2	0	0	0	3454
G28883C	G > C	G > R	N	N	20.49%	1	2	1	0	0	3453
A29700G	A > G	Intron	Intron	Intron	1.04%	0	0	4	167	4	1

[1] G28881A and G28882A occur within the same codon. Amino acid annotation (R > K) is based on the co-occurrence of these mutations.

* Under positive selection inferred by HyPhy.

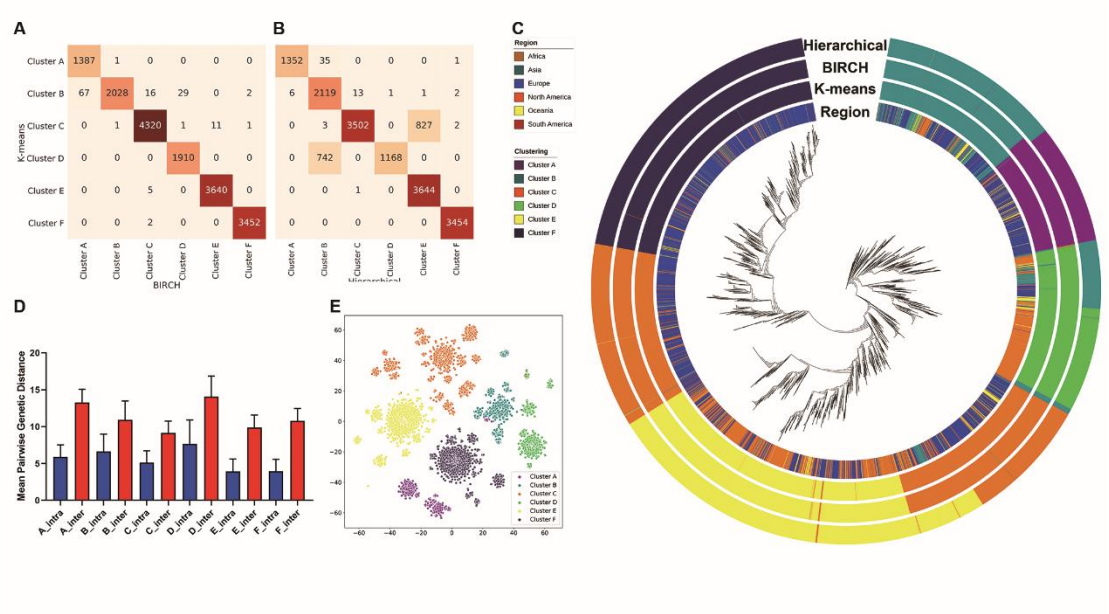

Figure 1. Clustering of SARS-CoV-2 reveals six major clusters. (A and B) Confusion matrix for the similarity of K-means clustering with BIRCH clustering (A) and Hierarchical clustering (B). Each row corresponds to the results of K-means clustering and each column corresponds to the clustering results of BIRCH (Hierarchical) clustering. (C) Phylogenetic tree of 16,873 SARS-CoV-2 strains. Four colored panels outside the phylogenetic tree are used to identify auxiliary information for each virus strain. The inner panel represent the distributions of the continents in the three. The outer three panels represent the partitions of the six clusters across the three clustering methods in the tree. (D) Mean pairwise genetic distances for intra-clustered and inter-clustered genetic distances. The blue bars represent mean pairwise genetic distances between pairs of isolates within the clusters, and the red bars represent mean pairwise genetic distances between pairs of isolates of outside the clusters. The error bar represents the standard deviation. The mean distance between pairs of strains of intra-clusters was significantly lower (P-value < 0.001, Wilcoxon rank-sum test) than that of inter-clusters. (E) The t-SNE plot of the K-means clustering results. Each dot represents one strain and each color represents to the corresponding cluster.

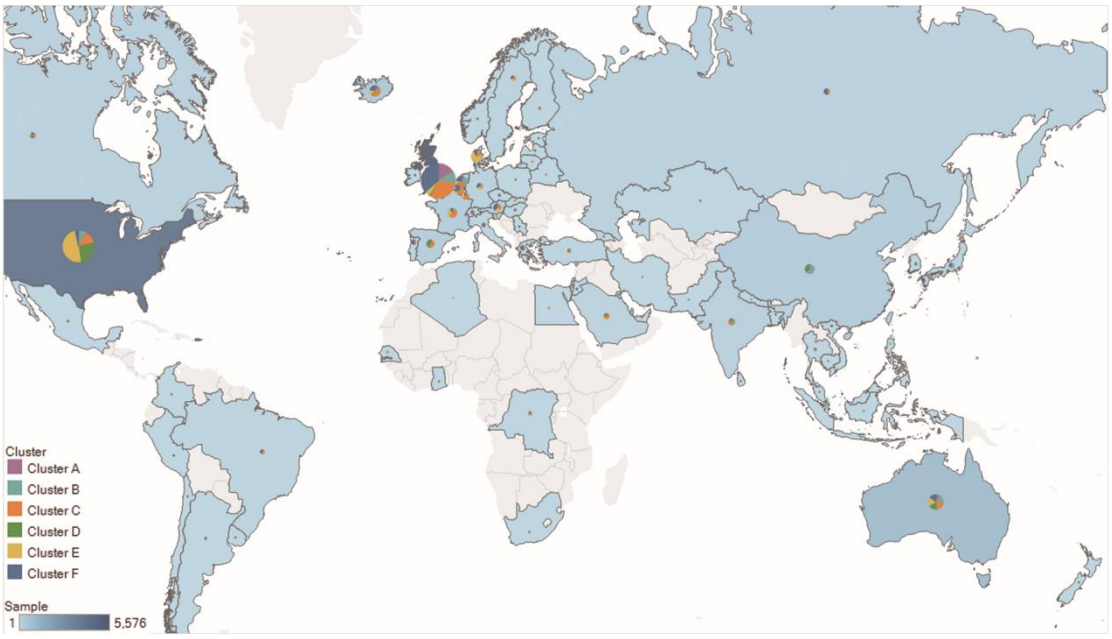

Figure 2. Geographic distributions of the six clusters. Pie charts display the proportions of six clusters among all SARS-CoV-2 strains in each country. Circle sizes and the color scales correspond to the number of strains analyzed per country.

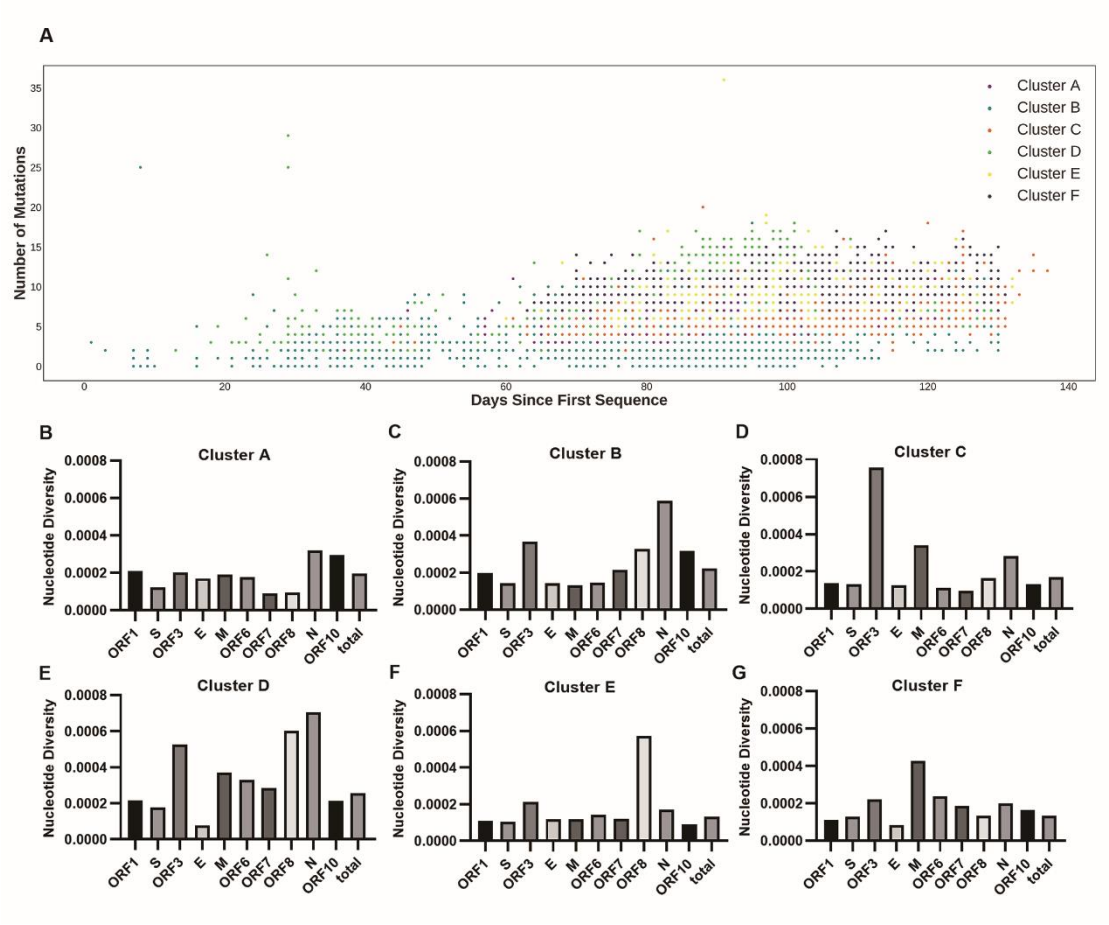

Figure 3. The genetic diversity between clusters. (A) The mutation counts over days of 16,873 SARS-CoV-2 strains. The X axis represents the days from the corresponding collection date of strains to 24 December 2019 when the earliest strain (EPI_ISL_402123) was collected. The Y axis represents the number of mutations of each collected strain. A mutation is defined by a nucleotide change from the original nucleotide in the reference genome to the alternative nucleotide in the studied viral genome. (B-G) The nucleotide diversity (π) per site for each gene and genome-wide across six clusters.

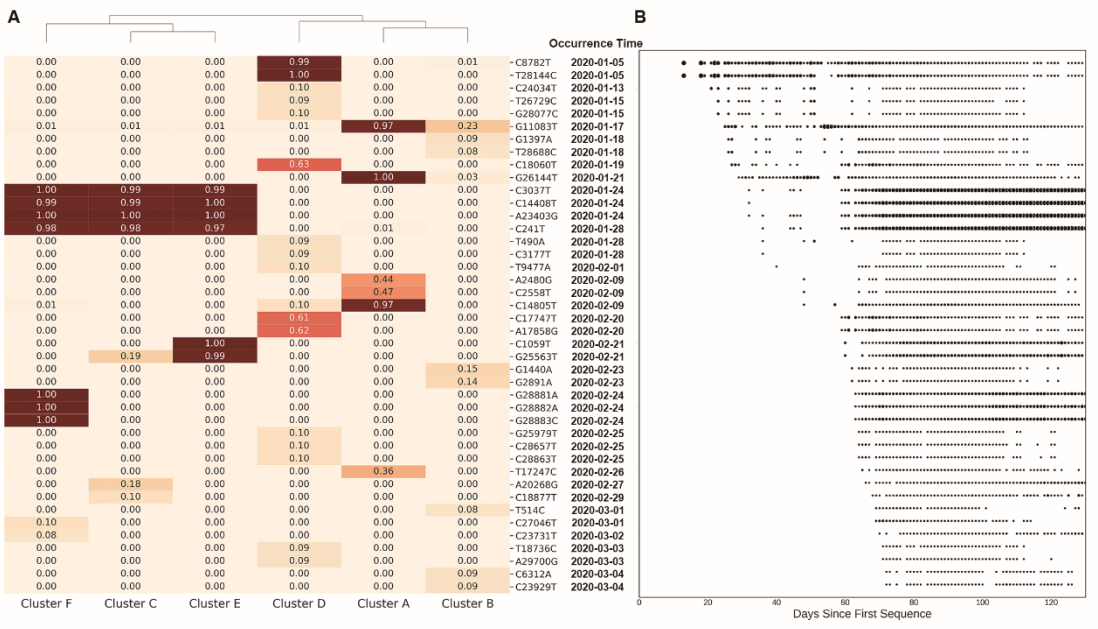

Figure 4. The clustering of the six clusters by the extracted mutations. (A) The heatmap displays mutation frequency of the 42 mutations across six clusters. The colors and values represent different frequencies of the corresponding mutations in each cluster. The collected days of the mutations are represented in (B). The X axis represents the days from the corresponding collection date of strains to 24 December 2019 when the earliest strain (EPI_ISL_402123) was collected. Circle sizes represent the frequency the of the mutations on each collection day.

Table S1. Geographic distribution of six continents for each cluster.

Cluster	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E	Cluster F	Total
Africa	3	4	65	7	10	9	98
Asia	38	648	248	217	57	116	1,324
Europe	1,137	990	3,119	212	1,108	2,961	9,527
North America	94	334	625	1,268	2,274	170	4,765
Oceania	110	161	233	196	191	149	1,040
South America	6	5	44	10	5	49	119
Total	1,388	2,142	4,334	1,910	3,645	3,454	16,873

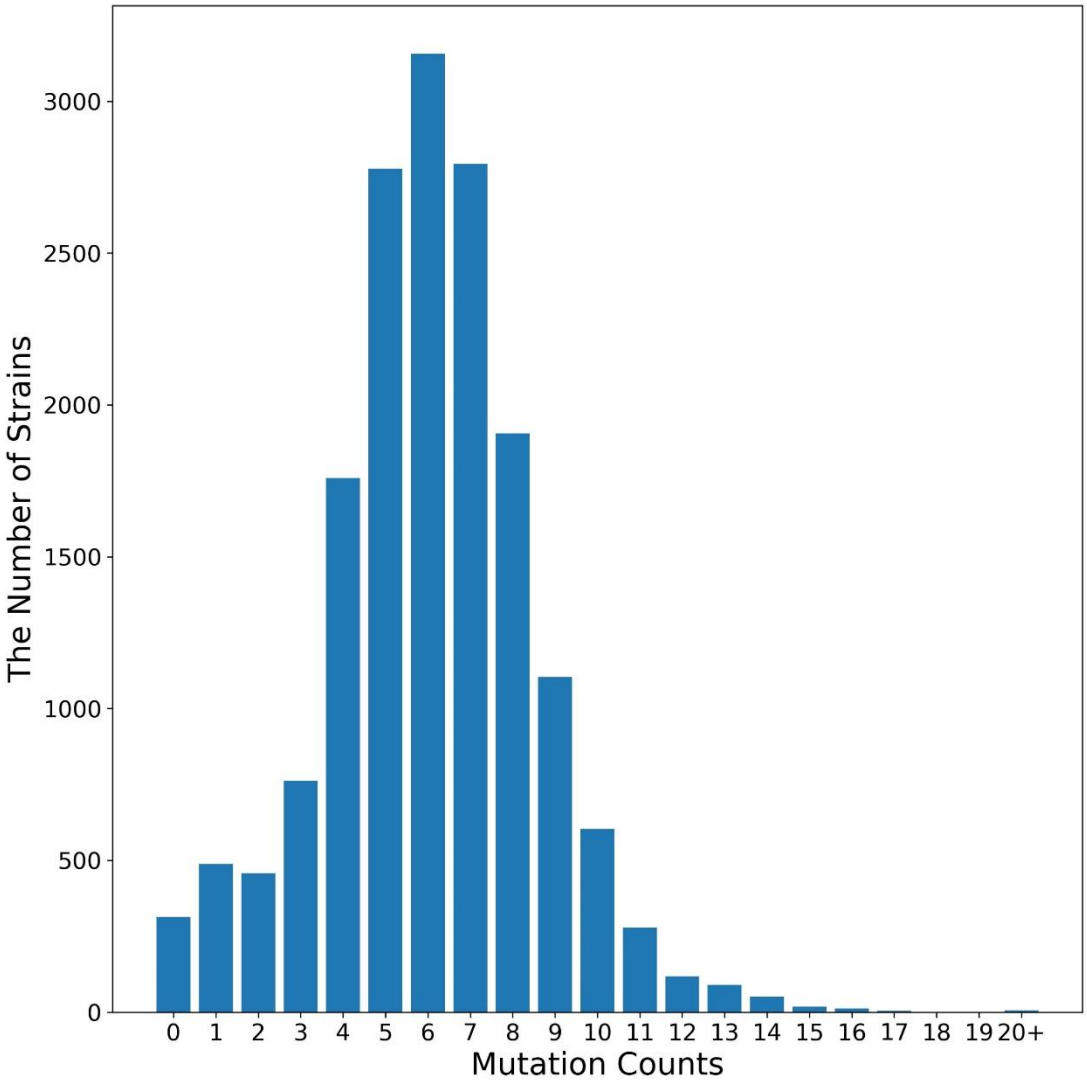

Figure S1. The distribution of the mutation counts of the 16,873 SARS-CoV-2 strains.

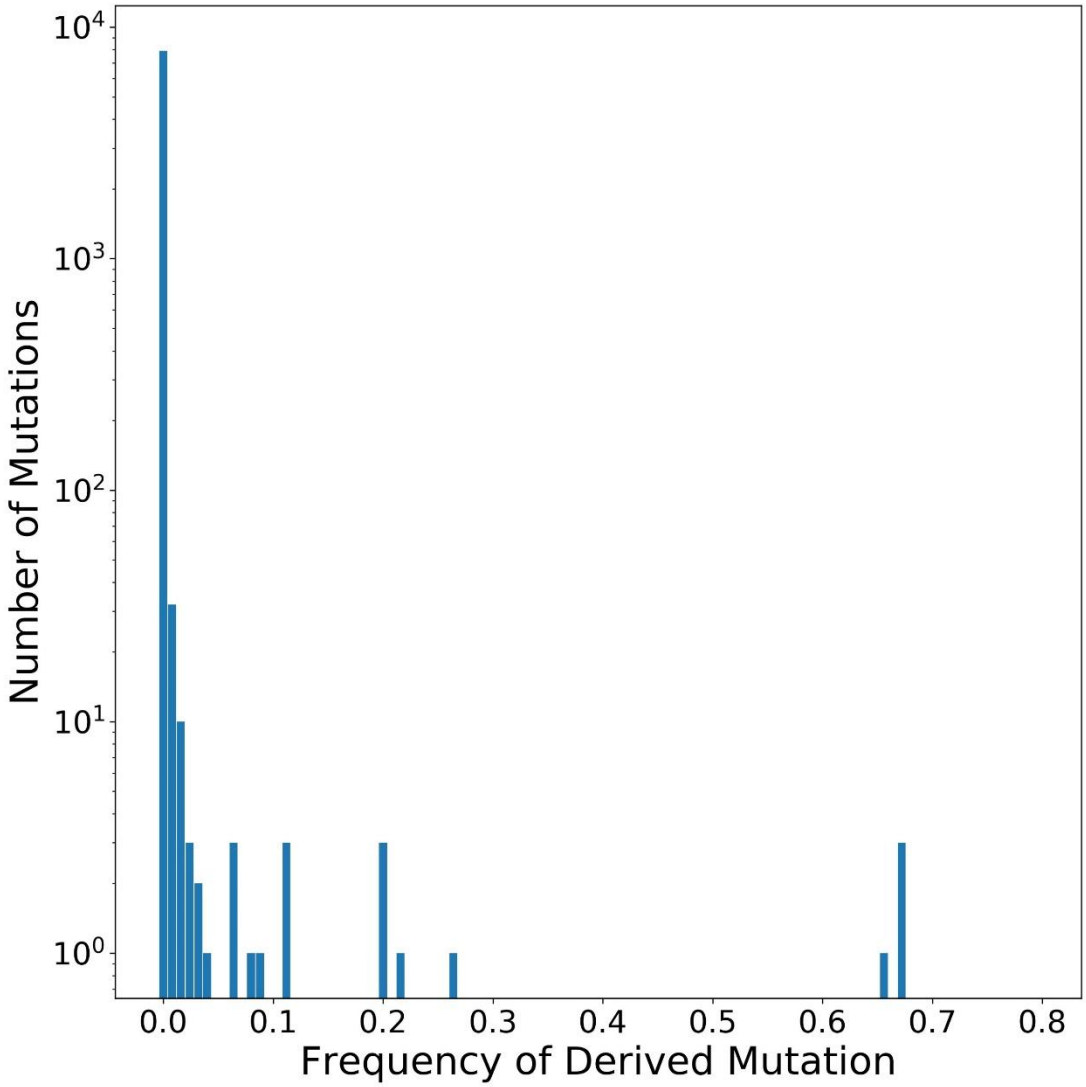

Figure S2. Frequency spectra of SARS-CoV-2. The mutation frequency of derived mutations of 16,873 SARS-CoV-2 stains is depicted on the X axis, and the number of mutations in which strains occurred is displayed on the Y axis. A log-10 scale is used for the Y axis of the graph, and the Y axis ranges from 1 to 10,000.

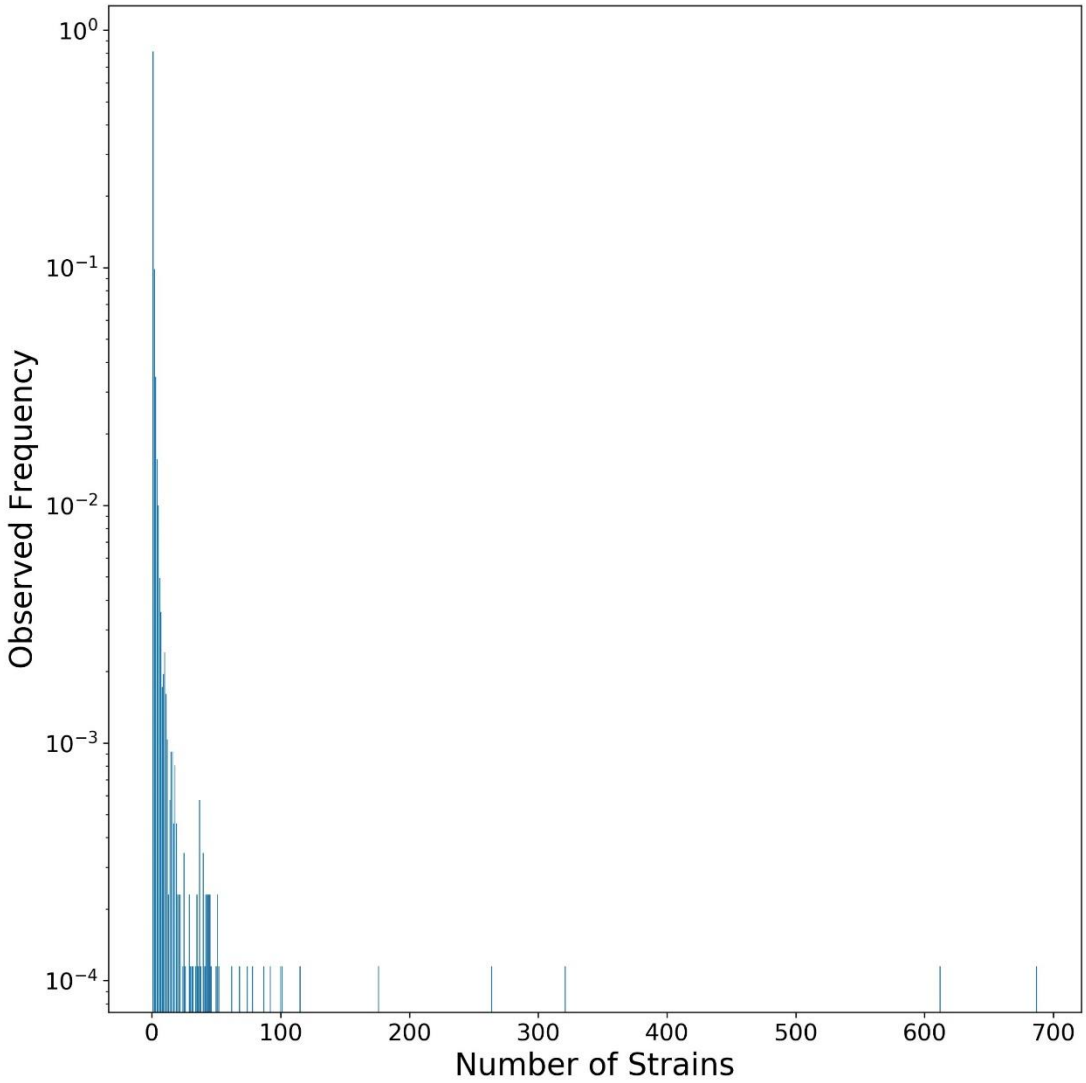

Figure S3. Normalized allele frequency of 16,873 SARS-CoV-2 strains. There are 8,706 unique genomes across the 16,873 strains. The X axis is the number of strains for each unique genome and the Y axis is the proportion of the unique genomes. A log-10 scale is used for the Y axis of the graph, and the Y axis ranges from 0.0001 to 1.

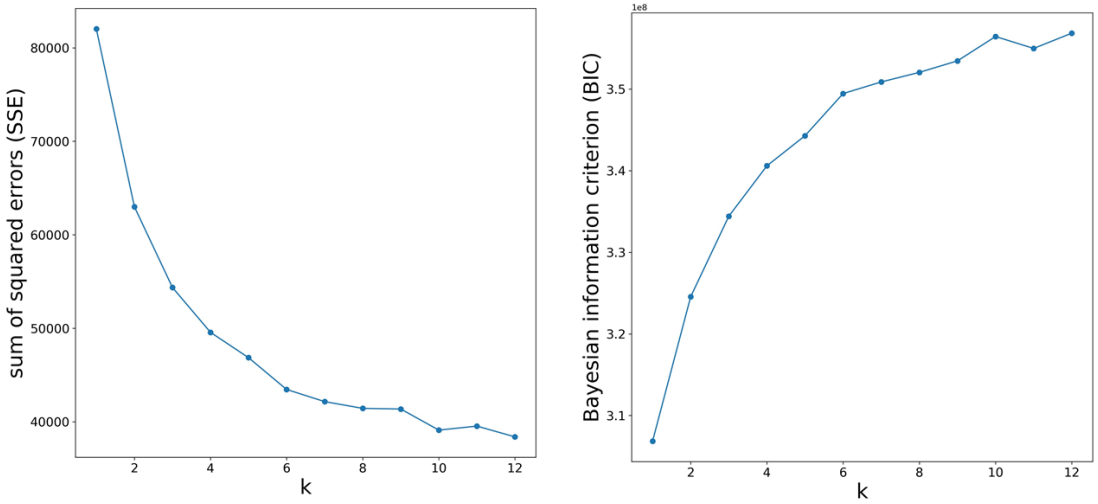

Figure S4. The evolution of the sum of squared errors (SSE; left) and Bayesian information criterion (BIC; right) for the number of clusters (K) in the K-means runs.

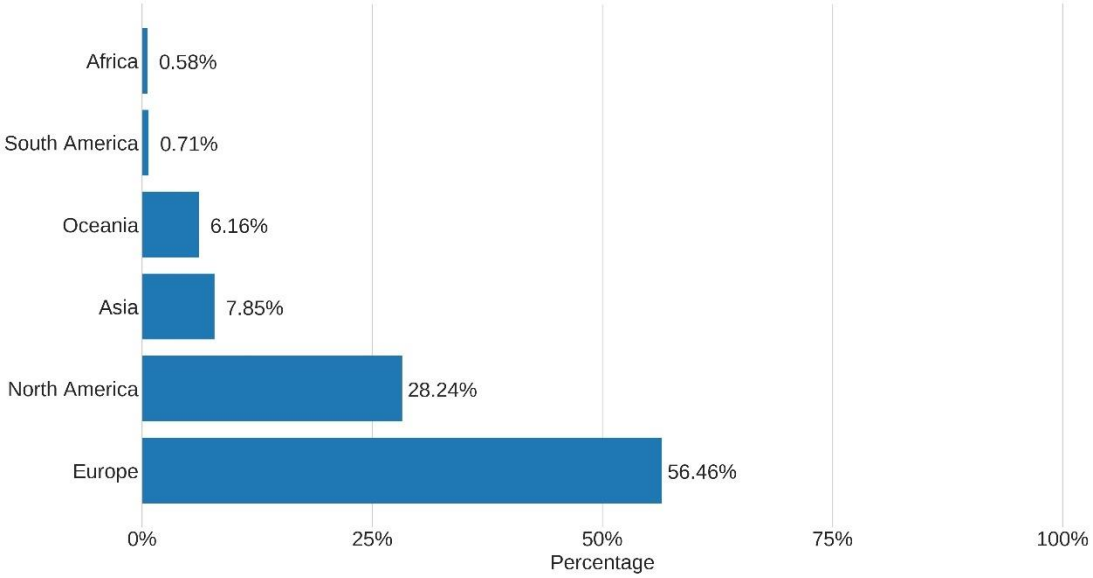

Figure S5. The bar chart displays the global proportions of the strains collected from each continent. More than half of the strains were collected from Europe.

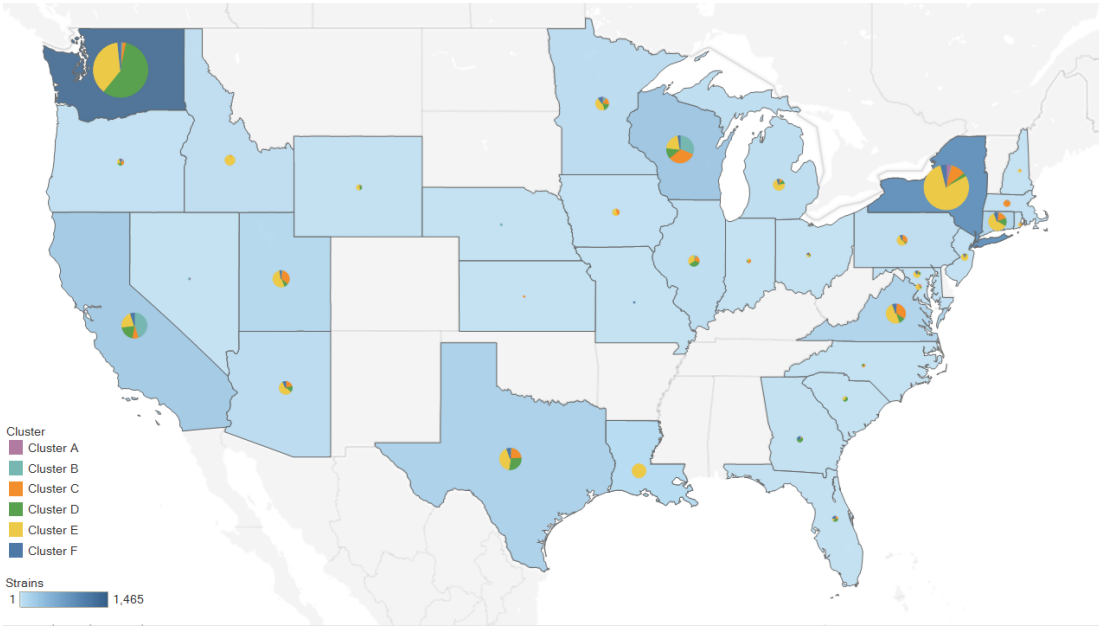

Figure S6. Geographic distribution of six clusters in the United States. Pie charts display the proportions of six clusters among all SARS-CoV-2 strains in each state. Circle sizes and the color scales correspond to the number of strains analyzed per state.

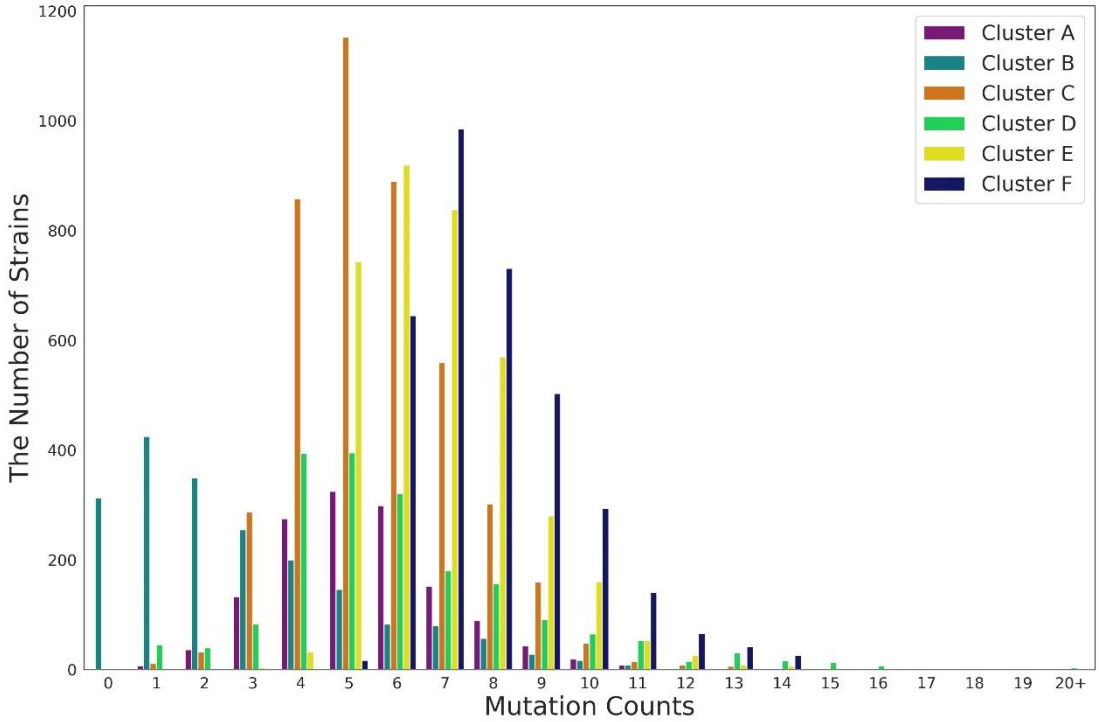

Figure S7. The distribution of the mutation counts of the strains for the six clusters.

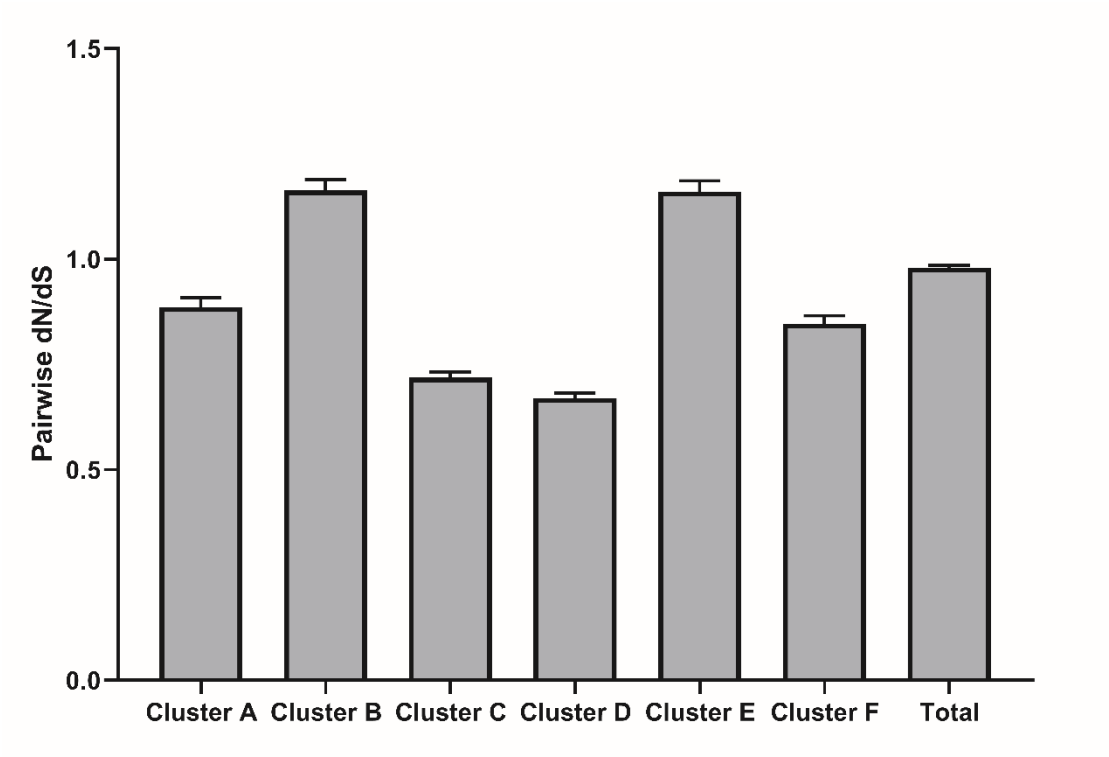

Figure S8. Pairwise ratios of rates of nonsynonymous to synonymous substitutions (dN/dS) for the six clusters and the total strains. Error bars represent the 95% confidence interval (CI) of the mean dN/dS. The average pairwise dN/dS in cluster B (P-value = 8.42×10^{-8}, Wilcoxon rank-sum test) was significantly higher than 1, and similarly for cluster E (P-value = 1.35×10^{-4}, Wilcoxon rank-sum test), suggesting positive selection in these two clusters. In contrast, the pairwise dN/dS ratios in the other four clusters were significantly lower than 1 (P-value < 0.001, Wilcoxon rank-sum test), suggesting that the nonsynonymous mutations tended to be selected against.

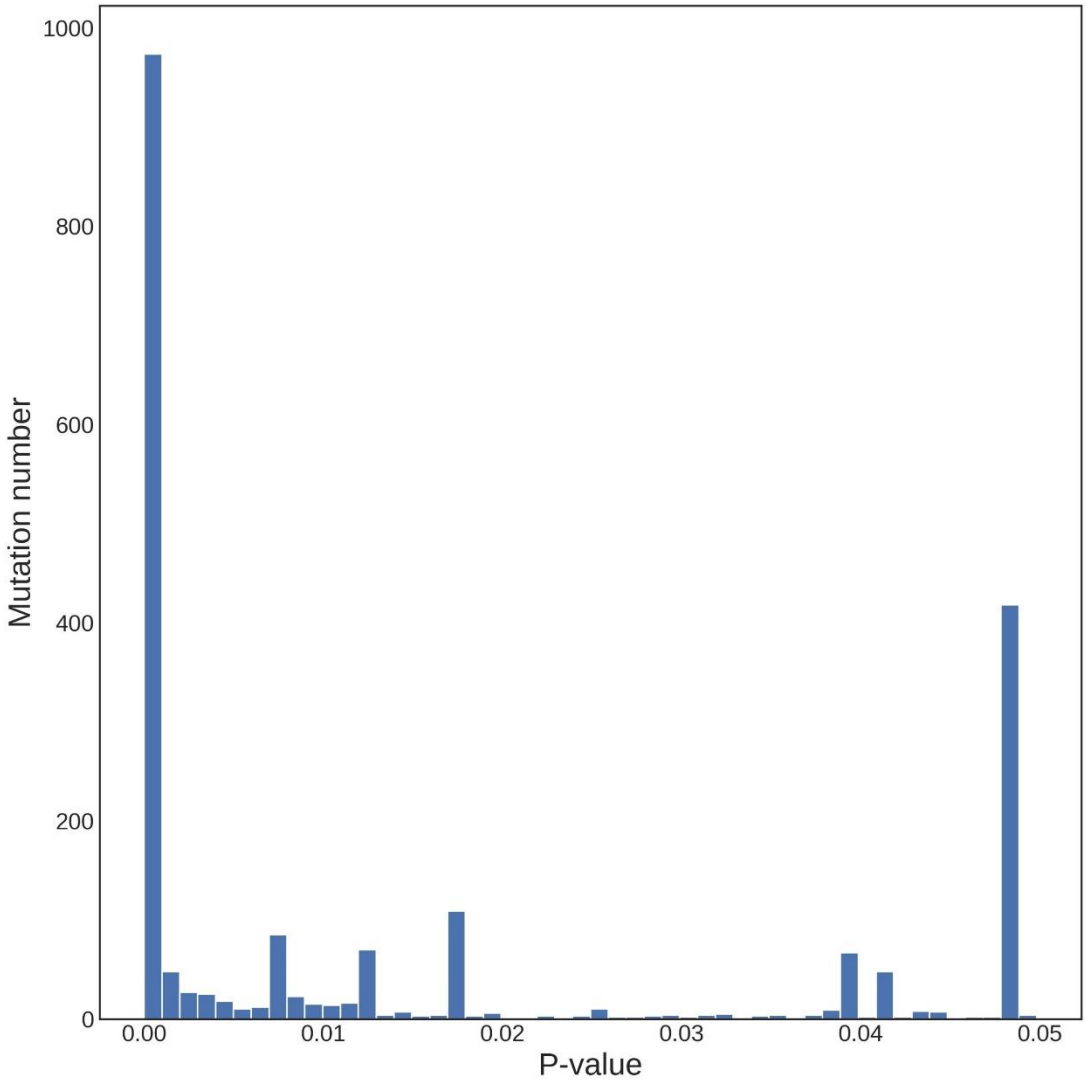

Figure S9. The distribution of P-values from the 2,094 mutations with a P-value <0.05 by ANOVA.

Figure S10. The heatmap displays the D' of the 42 mutations. D' values that correspond to substitution pairs are expressed as percentages and are shown within the respective squares. Higher D' values are indicated with a brighter red color. Higher values indicate stronger pairwise LD between the two mutations.

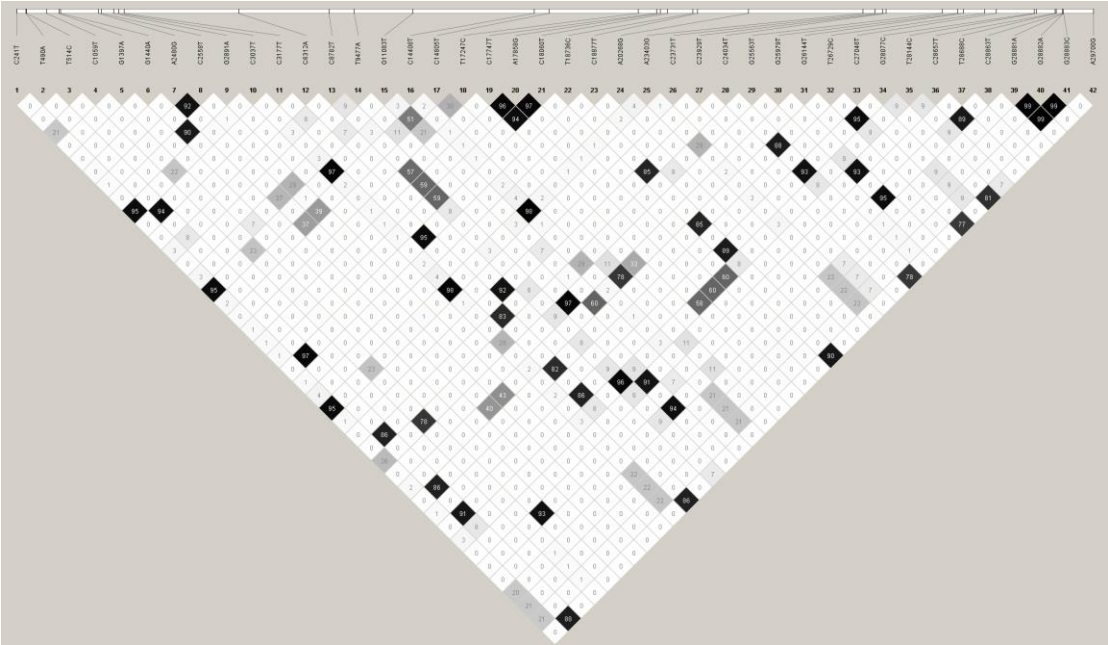

Figure S11. The heatmap displays the r^2 of the 42 mutations. The numbers within the squares represent the r^2 scores for pairwise LD. r^2 values are represented by white for $r^2 = 0$, with intermediate values for $0 < r^2 < 1$ indicated by shades of grey.

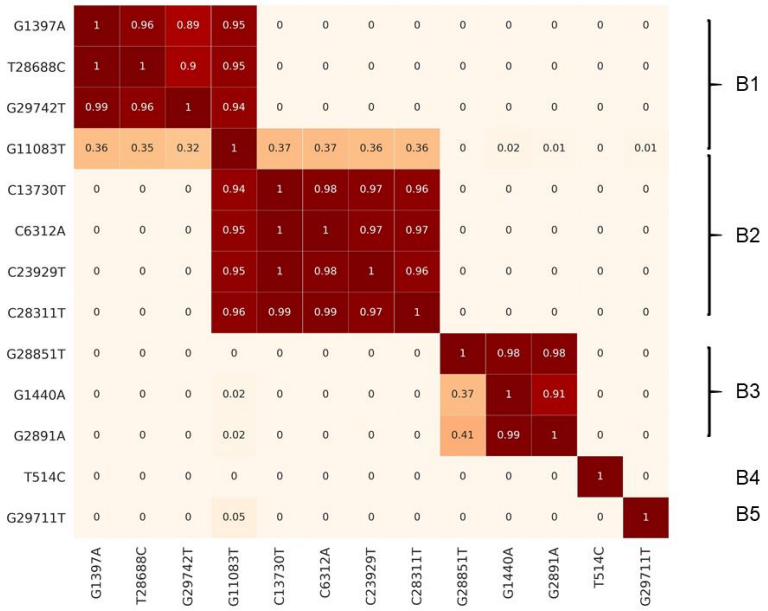

Figure S12. The pairwise dependency score (see materials and methods) of the mutations with frequency >0.05 within group B.