# Unsupervised clustering analysis reveals global population structure of SARS-CoV-2

Yawei Li[1], Qingyun Liu[2], Zexian Zeng[3], 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.

[3] Department of Data Science, Dana Farber Cancer Institute, Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA

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

## Abstract

Identifying the population structure of the newly emerged coronavirus SARS-CoV-2 has significant potential to inform public health management and diagnosis. As SARS-CoV-2 sequencing data is accruing rapidly, grouping them into clusters is important for organizing the landscape of the population structure of the virus. Since we have little prior information about the newly emerged coronavirus, we applied unsupervised clustering algorithms to group 16,873 SARS-CoV-2 strains, which automatically minimize intra-cluster distances and maximize inter-cluster distances. Six distinct genomic clusters were identified, with mutation profiles as input features. The varied proportions of the six clusters within different continents revealed specific geographical distributions. Comprehensive analysis indicated that genetic factors and human migration played an important role in shaping the specific geographical distribution of population structure. This study provides a concrete framework for the use of clustering methods in phylogenetic analysis to study the global population structure of SARS-CoV-2. In addition, clustering methods can be used for future studies of variant

population structures in specific regions of these fast-growing viruses.

## Introduction

COVID-19 pandemic was caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (1, 2). It is rapidly spreading throughout the world (3) with over 24 million cases and 0.85 million deaths by the end of August 2020. In an effort to understand the molecular characteristics of the virus, viral genomes have been abundantly sequenced and presented at the Global Initiative on Sharing All Influenza Data (GISAID) (4). To update, more than 70,000 genomes were collected and published.

As an emerging virus, it is important to understand 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 from one entity to the other entities indicate the degree of relationships. However, as population genomic datasets grow in size, simply using pairwise genetic distances cannot present an explicit structure of the total population in phylogenetic analysis. Grouping similar entities into the same cluster and identifying the number of main subtypes (clusters) makes it easier to understand the main characteristics of the population. Traditionally, using the distance matrix and the bifurcations between branches of leaves on 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 a better way to effectively group entities, clustering methods emerge as more productive and robust solutions. 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

supplemental 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).

In this study, to identify the population structure of the newly emerged coronavirus SARS-CoV-2, we employed clustering methods to group a total of 16,873 strains. The clustering results showed that there were six major clusters of SARS-CoV-2. In particular, we found that the proportions of six clusters in each continent showed a specific geographical distribution. Our analysis revealed that the unique geographical distributions across the clusters are both influenced by intrinsic genetic factors and migration of humans. This study provides a perspective of the SARS-CoV-2 population structural analysis, helping to investigate the evolution and spread of the virus 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 strains marked as “high coverage” were downloaded from GISAID. The “high coverage” was defined as strains 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 strains 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 strains.

### Mutation calls and phylogenetic reconstruction

All downloaded genomes were mapped to the reference genome of SARS-CoV-2 (GenBank Accession Number: NC_045512.2) following Nextstrain pipeline (20). Multiple sequence alignments and pairwise alignments were constructed using CLUSTALW 2.1 (21). Considering many putatively artefactual mutations and the gaps in sequences are located at the beginning and end of the alignment, we masked the first

130 bps and last 50 bps in mutation calling following Nextstrain pipeline. We used substitutions as features to reconstruct the phylogenetic tree using FastTree 2 (22). The phylogeny is rooted following Nextstrain pipeline 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 (25), hierarchical clustering and BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) (26, 27) clustering to partition the SARS-CoV-2 strains. The models were implemented using the Python package sklearn with the KMeans function, AgglomerativeClustering function and Birch function, respectively.

**Simpson's diversity index**

Simpson's Diversity Index (*D*) is a measure of diversity that considers the number of entities as well as their abundance. The index measures the probability that two randomly selected individuals are the same. The formula to calculate 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.

117 **Inferring positive/purifying selection of individual sites**

118 To test which position was under selective pressure, we used a set of programs
119 available in HyPhy (28) to calculate nonsynonymous (dN) and synonymous (dS)
120 substitution rates on a per-site basis to infer pervasive selection. Fast Unconstrained
121 Bayesian AppRoximation (FUBAR) (29) was applied to detect overall sites under
122 positive selection. The positively selected sites were identified using a probability larger
123 than 0.95 using the FUBAR method.

124

125 **Pairwise mutation dependency score**

126 Pairwise mutation dependency scores can measure the order in which genetic
127 mutations are acquired within a cluster. For two selected mutations X and Y, the score
128 S(X|Y) represents the proportion of strains that accumulated both X among the strains
129 that accumulated mutation Y. S(X|Y) and S(Y|X) can be calculated using the following
130 functions:

131 $$\mathrm{S(X|Y)} = \frac{\sum_{all\ samples} S_X = 1\ \&\ S_Y = 1}{\sum_{all\ samples} S_Y = 1}$$

132 $$\mathrm{S(Y|X)} = \frac{\sum_{all\ samples} S_X = 1\ \&\ S_Y = 1}{\sum_{all\ samples} S_X = 1}$$

133 where $S_X = 1$ denotes that the sequence has a mutation X. Pairwise mutation
134 dependency score displays the correlation and the timescale relationship of the two
135 mutations. The value of S(X|Y) and S(Y|X) ranges between 0 and 1. With this index,
136 S(X|Y) = 1 with S(Y|X) < 1 represents that mutation Y occurs after mutation X. In
137 contrast, S(X|Y) = 1 with S(Y|X) = 1 represents that the two mutations occur
138 simultaneously and are genetically linked. Statistical analyses and data presentations
139 were generated using Python 3.7.6.

140

141 ## Results

142 **Genetic analysis of SARS-CoV-2 indicates high diversity and rapidly proliferating**
143 **pattern of the virus**

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. 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 a rapidly proliferating pattern (30). In addition, there were 8,706 unique strains across the 16,873 strains (Figure S3), and most unique strains (7,078) were singletons, yielding high diversity of the virus. 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. The frequency spectrum of substitutions and high Simpson's diversity index indicated high genetic diversity of SARS-CoV-2.

**Clustering of SARS-CoV-2 reveals six major clusters**

To clarify the main population structure of the virus, grouping these strains into clusters is necessary, as these clusters displayed the major types of the virus. However, the genetic analysis of SARS-CoV-2 showed that there were 8,706 unique strains across the 16,873 strains (Figure S3), it is not easy to directly and accurately partition the strains. For this reason, we applied clustering techniques to measure similarities between these strains and effectively group them.

Because SARS-CoV-2 exhibits limited number of SNPs per virus strains and little ongoing horizontal gene exchange, making SNPs ideal clustering input features. We first employed the most common and well-known method, K-means clustering (25), to group the strains. 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 genome. To determine the number of clusters, we

174 plotted the curves of the sum of squared errors (SSE) and Bayesian information
175 criterion (BIC) (31) under different cluster numbers ranging from 2 to 12 (Figure S4).
176 We used the elbow method and chose the elbow of the curve as the number of clusters
177 (32). This approach resulted in $K$=6 for both the SSE and BIC curves. We also
178 employed two other clustering methods: hierarchical clustering and BIRCH clustering
179 (26, 27) to group the strains. The confusion matrix indicated that the similarity of K-
180 means and BIRCH exceeded 99% (Figure 1A), whereas the similarity of K-means and
181 hierarchical clustering was 90% (Figure 1B). The differences between K-means and
182 hierarchical clustering results may be due to their different algorithms. The merges
183 and splits of hierarchical clustering are determined in a greedy manner. Each step of
184 the algorithm is based on local decisions, merging the most "compact" cluster
185 available (33). The compactness criterion (or how close together objects are) is
186 defined by the linkage criteria and proximity measure used. On the other hand, K-
187 means maximizes the objective that is a combination of cluster compactness and
188 cluster separation (34).

189 The objective of clustering is minimizing intra-cluster distances and maximizing
190 inter-cluster distances. To determine which method is performed best, we calculated
191 intra-cluster pairwise genetic distances. The average intra-cluster pairwise genetic
192 distances in K-means (4.886) was significantly lower than that in hierarchical clustering
193 (5.062, P-value < 0.001, Wilcoxon rank-sum test) and BIRCH (4.900, P-value < 0.001,
194 Wilcoxon rank-sum test). In addition, we aligned the partitions of the six clusters
195 against the phylogenetic tree for the three methods (Figure 1C). The differences
196 between the hierarchical clustering results and the two other clustering results were
197 mainly at the boundary of the clusters. Of the three methods, strains grouped by K-
198 means and BIRCH were more compact in the phylogenetic tree than those by
199 hierarchical clustering. For example, the strains in both K-means cluster D and BIRCH
200 cluster D were split into two clusters by hierarchical clustering. However, such a split
201 was not supported by the phylogenetic tree (Figure 1C). In general, K-means exhibited
202 the best performance in comparison to the other methods.

203 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 varied proportions of the six clusters in different continents showed specific geographical distributions.**

Mapping the proportions of strains from each continent showed that the clusters differed in their geographical distributions (Figure 2, Table S1). Of the six clusters, cluster C spread globally. By contrast, cluster A and cluster F occurred at high frequencies in specific regions. 81.92% of the strains in cluster A and 85.73% of the strains in cluster F were from Europe. The geographical spread of each of the three remaining clusters was intermediate. Cluster E occurred at higher frequencies in North America and Europe, and lower frequencies in Asia and Oceania. Cluster D occurred at higher frequencies in North America, and lower frequencies in Asia, Europe and Oceania. The strains in cluster B were mainly in Asia and Europe and partially in North America and Oceania.

However, due to the sampling bias of the SARS-CoV-2 (Figure S5), 85% of the strains were collected from Europe and North America, making the proportion of the continents in each cluster not informative. Therefore, we evaluated the proportion of the clusters on each continent. In most continents, the distributions of the strains were 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. Strikingly, Oceania was the only continent that was uniformly separated into the six clusters, indicating strains in Oceania were more diverse than in the other continents.

The different geographical distributions for the six clusters could be due to intrinsic genetic factors, extrinsic factors such as the migration of humans, or both. Hence, we

next aimed to explore the genomic characteristics of these clusters, as well as the transmission and human migration of the virus across the globe.

**The genetic variance analyses indicated high diversity between clusters**

If the different geographical distributions for the six clusters were due to intrinsic genetic factors, there would be high genetic variance between the clusters. The average mutation counts for the six clusters were 6.38, 3.49, 6.57, 7.09, 7.89 and 8.96 (Figure S6), respectively. Considering the different collection dates (Figure 3A) of the strains, mutation rates as opposed to mutation counts were more effective for describing the genetic variations between clusters. We defined the date when the reference strain was collected as the index date. 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, respectively. 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 for 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 diversity per site for the six clusters was 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. The nucleotide diversity of each gene across all clusters is displayed in 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. Our analysis showed that intra-cluster genetic diversity differed between clusters, suggesting that selective pressures were different between clusters. These different selective pressures will affect the geographical distribution of each cluster.

**Explore mutations that shaped the geographical distribution of population**

**structure.**

The high genetic diversity between clusters indicated that the frequencies of the mutations across clusters were very different. In order to explore whether there are mutations that affect the genetic structure within the clusters, we applied ANOVA to identify the statistically significant mutations that were strongly associated with clusters. Across the 7,970 substitutions, 26.27% (2,094 substitutions) of them achieved P-values <0.05 (Figure S7). We found that some of these mutations were fixed in one or several clusters. Cluster C, cluster E and cluster F shared four common fixed substitutions: A23403G, C241T, C3037T and C14408T. Cluster E had two additional fixed substitutions: C1059T and G25563T, and cluster F had three additional substitutions from position 28,881 to position 28,883. For the remaining three clusters, there were two fixed substitutions (C8782T, T28144C) and three fixed substitutions (G11083T, G14805T and G26144T) in cluster A. It is noteworthy that the fixed mutation numbers in cluster E (six) and cluster F (seven) were higher than in any of the other clusters, which was consistent with our conclusion of the high mutation rates but low nucleotide diversity in cluster E and cluster F.

We selected the 2% (42/2094) substitutions that achieved the lowest P-values (Table 1) and analyzed their distributions in the clusters. Of the 42 substitutions, there were 26 nonsynonymous mutations (mutation G28882A was in a trinucleotide mutation from position 28881 to 28883 that spans two codons and results in an RG (arginine-glycine) to KR (lysine-arginine) amino acid change). We focused on these nonsynonymous mutations as these mutations may be under selection that affect the population structure (35). Some of these substitutions were reported to impact the evolution of SARS-CoV-2 (36-39). For example, mutation A23403G (D614G, Asparticacid to Glycine) in the *spike* protein domains was reported to show significant variation in cytopathic effects and viral load, and substantially change the pathogenicity of SARS-CoV-2 (40). This mutation was accompanied by a mutation (T14408C) that results in an RNA-dependent RNA polymerase (RdRp) amino acid change (41). In addition, Tang et al (42) used mutation T28144C to define “L” type (defined as “L” type because T28,144 is in the codon of Leucine) and “S” type (defined as “S” type

because C28,144 is in the codon of Serine) of SARS-CoV-2. They found that the “L” type was more transmissible and aggressive than the “S” type.

Previous studies have reported that recombination is common in coronavirus (43-45). Given that recombinations in SARS-CoV-2 may perturb the clustering, we used Haploview (46) to analyze the linkage disequilibrium (LD) by calculating standardized disequilibrium coefficients (D’) and squared allele-frequency correlations ($r^2$) of the 42 substitutions. D′ is affected solely by recombination and not by differences in allele frequencies between sites, and $r^2$ is also affected by differences in allele frequencies at the two sites (47). In the heatmap of D’ (Figure S8) and $r^2$ (Figure S9), no obvious LD blocks were discovered, indicating that our clustering of SARS-CoV-2 strains using substitutions was not distorted by recombination.

Selection usually affects the distribution of the mutations in the population. Purifying selection tends to remove amino acid-altering mutations, while positive selection tends to increase the frequency of the mutations. Considering the rapidly proliferating pattern of SARS-CoV-2 that strengthened the power of drift relative to the power of purifying selection (48-50), we mainly focused on the positive selective sites. We applied HyPhy to infer the probabilities of the extracted 26 nonsynonymous mutations that were under positive selection. There are nine mutations (asterisks in Table 1) with a positive probability >0.95. In particular, mutations G2891A, G11083T, C14408T, C17747T and A23403G (D614G) were reported as recurrent mutations (37, 51). The recurrence of these mutations agrees with the assumption that they may confer selective advantages in the population. These possible positively selected mutations may result in greater diversity among clusters with different population structures of SARS-CoV-2 across geographical regions.

**The global spread of SARS-CoV-2 suggests that the unique geographical distribution may also be influenced by the human migration**

Regardless of the genetic factors, the travel of humans could also lead to unique geographical distributions in today's highly globalized world. By analyzing the frequencies of the extracted 42 mutations in each cluster (Figure 4A) and their collected

daily counts (Figure 4B), we can trace the dynamics of substitutions in the SARS-CoV-2 genome. The four genetically linked mutations, A23403G (D614G), C241T, C3037T and C14408T that were fixed across three clusters (C, E and F) had become the highest frequency mutations in the world, with a high frequency on all continents in our downloaded sequences, including South America (87%), Africa (86%), Europe (75%), North America (65%), Oceania (55%) and Asia (32%). The earliest time when sequences carrying these mutations was collected was in late January 2020. About a month later, these mutations were discovered worldwide. Though the mutation A23403G (D614G) has been reported and estimated to be a positive selective mutation, it is almost impossible to spread to the world without human migration in such a short time. Besides these high frequency mutations, some lower frequency mutations also provided some evidence of human migration. We explored the geographical distributions of mutations with global frequencies <0.05 in Table 1. Though most of these low frequency mutations were mainly collected within a single continent, we still find two mutations, T28688C and G1397A, were discovered in Asia, Europe and Oceania with high proportion. In addition, the spatial geographical distributions of some substitutions also provide the evidence that human migration may have influenced the spread of the virus. For example, on the west coast of the USA, most strains accumulated the mutations C8782T and T28144C (cluster D), and these mutations were also found in high frequencies in east Asia. In contrast, on the east coast of the USA, most strains accumulated the mutations A23403G, C241T, C3037T, C14408T, C1059T and G25563T (cluster E), and the similar strains were mainly discovered in Europe (Figure S10).

## Discussion

Understanding the population structure of SARS-CoV-2 is important in evaluating future risks of novel infections. To precisely analyze its population structure, we used clustering methods in the phylogenetic analysis to group a total of 16,873 publicly available SARS-CoV-2 strains. We employed three different clustering methods: two

353 of which (K-means clustering, hierarchical clustering) were widely used, and the third
354 one (BIRCH) was a much newer one and suggested to be applied to large datasets. By
355 comparing the distributions of the clustering results in the phylogenetic tree and intra-
356 cluster pairwise genetic distances across the three methods, we found that K-means
357 exhibited the best performance.

358 Our clustering results indicated six major clusters of SARS-CoV-2. The mutation
359 profile characterizing clusters of the viral sequences displayed specific geographical
360 distributions. Most continents were mainly concentrated in one or two clusters, but we
361 also found that in Oceania, the strains were uniformly separated across the six clusters.
362 To evaluate whether the geographical distributions for the clusters were due to genetic
363 factors or travel of humans. The varied intra-cluster genetic diversity across the clusters
364 suggested different selective pressures between clusters, which would affect the
365 geographical distributions across the clusters. By analyzing the statistically significant
366 mutations that were strongly associated with the clusters we identified that some
367 mutations might be under positive selection, indicating different geographical
368 distributions between the clusters were partially affected by these mutations. In addition,
369 the dynamics and the spatial geographical distributions of some substitutions suggested
370 that human migration may also have affected the different geographical distributions.
371 In general, our findings indicate that the geographical distributions for the clusters are
372 the result of both genetic factors and migration of humans.

373 It is noteworthy that our study is limited due to the sampling bias of SARS-CoV-2,
374 with more than 60% of the strains being from the United Kingdom and the USA. In
375 contrast, the overall proportion of strains from Africa and South America is less than 2%
376 (Figure S5). Sampling biases can lead to biased parameter estimation and affect the
377 clustering results we observed. For example, the frequency of mutation C15324T
378 reached 41.84% in Africa, but was only 2.21% outside Africa. The frequency of
379 mutation T29148C reached 15.13% in South America, but was only 0.12% outside
380 South America. Another mutation T27299C with frequency 10.92% in South America
381 was only found with frequency 0.08% in other regions. In fact, all three mutations were
382 mostly grouped in single clusters, indicating these mutations were highly concentrated.

However, due to the small proportion of the strains from these two continents, these mutations were unable to affect the clustering of samples. To address this issue, more strains were needed to be collected from these continents. In addition, we found that in cluster B, there were no fixed mutations. We calculated the pairwise dependency scores (see Materials and Methods) of all the mutations with frequencies >0.05 in cluster B and discovered five main subclusters (Figure S11). Other than the mutation G11083T that was discovered in two subclusters, there were no common mutations between either of the five clusters. As shown in Figure 3A, these strains were grouped in one cluster mainly because these strains had smaller mutation counts than strains in other clusters. The genetic distance between two strains was still small, though they shared no common mutations. To address this issue, another clustering can be used for more further analyses.

Despite the limited number of SARS-CoV-2 genome sequences, our analysis of population genetics is formative. Our discovery of high genetic diversity in SARS-CoV-2 is consistent with an earlier study (52). The topology and the divergence of the clusters in the phylogenetic tree illustrate 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, 53-55). Our work, as well as previous studies (56-58) that use clustering techniques to study the population structure of the SARS-CoV-2 virus, has proved to be a valuable supplemental tool in phylogenetic analyses. In addition, clustering ideas can be used for further study of variant population structures in specific regions of these fast-growing viruses.

## Data Availability

The publicly available SARS-CoV-2 datasets in this study are available at GISAID (https://www.gisaid.org). The reference SARS-CoV-2 is available at the NCBI GenBank (GenBank Accession Number: NC_045512.2, https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2).

## Competing Interests

The authors declare that they have no conflict of interests.

## Authors' contributions

Y.Luo and Y.Li designed the research; Y.Li and Q.L. analyzed data; Y.Luo, Y.Li and Q.L. contributed to the theory; and Y.Luo and Y.Li drafted the manuscript. All authors have read, edited and approved the final manuscript.

## References


1. Coronaviridae Study Group of the International Committee on Taxonomy of V. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020;5(4):536-44.
2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020;382(8):727-33.
3. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science. 2020;368(6490):489-93.
4. Shu YL, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Eurosurveillance. 2017;22(13):2-4.
5. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci U S A. 2020;117(17):9241-3.
6. Koyama T, Platt D, Parida L. Variant analysis of COVID-19 genomes2020.
7. Rehman SU, Shafique L, Ihsan A, Liu Q. Evolutionary Trajectory for the Emergence of Novel Coronavirus SARS-CoV-2. Pathogens. 2020;9(3).
8. Yeh T-Y, Contreras G. Faster de novo mutation of SARS-CoV-2 in shipboard quarantine2020.
9. Mahapatro G, Mishra D, Shaw K, Mishra S, Jena T. Phylogenetic Tree Construction for DNA Sequences using Clustering Methods. Procedia Engineer. 2012;38:1362-6.
10. Sharma A, Jaloree S, Thakur R. Review of Clustering Methods: Toward

Phylogenetic Tree Constructions. 2018. p. 475-80.
11. Gonzalez TF. Clustering to minimize the maximum intercluster distance. Theoretical Computer Science. 1985;38:293-306.
12. Wang J, Soininen J, He J, Shen J. Phylogenetic clustering increases with elevation for microbes. Environ Microbiol Rep. 2012;4(2):217-26.
13. Ning J, Beiko RG. Phylogenetic approaches to microbial community classification. Microbiome. 2015;3:47.
14. Fioravanti D, Giarratano Y, Maggio V, Agostinelli C, Chierici M, Jurman G, et al. Phylogenetic convolutional neural networks in metagenomics. Bmc Bioinformatics. 2018;19(Suppl 2):49.
15. Qin L, Chen YX, Pan Y, Chen L. A novel approach to phylogenetic tree construction using stochastic optimization and clustering. Bmc Bioinformatics. 2006;7.
16. Felsenstein J, Churchill GA. A hidden Markov Model approach to variation among sites in rate of evolution. Molecular Biology and Evolution. 1996;13(1):93-104.
17. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034-50.
18. Medema MH, Cimermancic P, Sali A, Takano E, Fischbach MA. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. PLoS Comput Biol. 2014;10(12):e1004016.
19. Han E, Carbonetto P, Curtis RE, Wang Y, Granka JM, Byrnes J, et al. Clustering of 770,000 genomes reveals post-colonial population structure of North America. Nat Commun. 2017;8:14238.
20. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121-3.
21. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23(21):2947-8.
22. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3):e9490.
23. Rambaut A. FigTree v1. 4 2012 [Available from:

http://tree.bio.ed.ac.uk/software/figtree/.

24. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. 2007;23(1):127-8.

25. MacQueen J, editor Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics; 1967 1967; Berkeley, Calif.: University of California Press.

26. Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases. Proceedings of the 1996 ACM SIGMOD international conference on Management of data; Montreal, Quebec, Canada: Association for Computing Machinery; 1996. p. 103–14.

27. Zhang T, Ramakrishnan R, Livny M. BIRCH: A New Data Clustering Algorithm and Its Applications. Data Mining and Knowledge Discovery. 1997;1(2):141-82.

28. Pond SLK, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, et al. HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. Molecular Biology and Evolution. 2020;37(1):295-9.

29. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Pond SLK, et al. FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. Molecular Biology and Evolution. 2013;30(5):1196-205.

30. Yu WB, Tang GD, Zhang L, Corlett RT. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2 / HCoV-19) using whole genomic data. Zool Res. 2020;41(3):247-57.

31. Schwarz G. Estimating the Dimension of a Model. Ann Statist. 1978;6(2):461-4.

32. Thorndike RL. Who Belongs in the Family? Psychometrika. 1953;18(4):267-76.

33. Jain AK, Dubes RC. Algorithms for clustering data. Englewood Cliffs, N.J.: Prentice Hall; 1988. xiv, 320 p. p.

34. de Souto MCP, Costa IG, de Araujo DSA, Ludermir TB, Schliep A. Clustering cancer gene expression data: a comparative study. Bmc Bioinformatics. 2008;9.

35. Hartl DL, Clark AG. Principles of population genetics. 4th ed. Sunderland, Mass.: Sinauer Associates; 2007. xv, 652 p. p.

501 36. Cai H, Cai K, Li J. Identification of Novel Missense Mutations in a Large Number
502 of Recent SARS-CoV-2 Genome Sequences. Preprints.org; 2020.
503 37. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence
504 of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol.
505 2020;83:104351.
506 38. YANG X, Dong N, CHAN W-C, CHEN S. Identification of super-transmitters of
507 SARS-CoV-2. medRxiv. 2020:2020.04.19.20071399.
508 39. Yin C. Genotyping coronavirus SARS-CoV-2: methods and implications. Genomics.
509 2020.
510 40. Yao H, Lu X, Chen Q, Xu K, Chen Y, Cheng L, et al. Patient-derived mutations
511 impact pathogenicity of SARS-CoV-2. medRxiv. 2020:2020.04.14.20060160.
512 41. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al.
513 Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity
514 of the COVID-19 Virus. Cell. 2020;182(4):812-+.
515 42. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing
516 evolution of SARS-CoV-2. National Science Review. 2020.
517 43. Graham RL, Baric RS. Recombination, reservoirs, and the modular spike:
518 mechanisms of coronavirus cross-species transmission. J Virol. 2010;84(7):3134-46.
519 44. Rehman SU, Shafique L, Ihsan A, Liu QY. Evolutionary Trajectory for the
520 Emergence of Novel Coronavirus SARS-CoV-2. Pathogens. 2020;9(3).
521 45. Rouchka E, Chariker J, Chung D. Phylogenetic and Variant Analysis of 1,040
522 SARS-CoV-2 Genomes. Preprints.org; 2020.
523 46. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD
524 and haplotype maps. Bioinformatics. 2005;21(2):263-5.
525 47. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, et
526 al. Structure of linkage disequilibrium and phenotypic associations in the maize
527 genome. Proc Natl Acad Sci U S A. 2001;98(20):11479-84.
528 48. Nowak MA, Michor F, Iwasa Y. The linear process of somatic evolution. Proc Natl
529 Acad Sci U S A. 2003;100(25):14966-9.
530 49. Wu CI, Wang HY, Ling S, Lu X. The Ecology and Evolution of Cancer: The Ultra-

Microevolutionary Process. Annu Rev Genet. 2016;50:347-69.
50. Chen Y, Tong D, Wu CI. A New Formulation of Random Genetic Drift and Its Application to the Evolution of Cell Populations. Mol Biol Evol. 2017;34(8):2057-64.
51. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med. 2020;18(1):179.
52. Li X, Wang W, Zhao X, Zai J, Zhao Q, Li Y, et al. Transmission dynamics and evolutionary history of 2019-nCoV. J Med Virol. 2020;92(5):501-11.
53. Chan JFW, Yuan SF, Kok KH, To KKW, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet. 2020;395(10223):514-23.
54. Sun J, He WT, Wang L, Lai A, Ji X, Zhai X, et al. COVID-19: Epidemiology, Evolution, and Cross-Disciplinary Perspectives. Trends Mol Med. 2020;26(5):483-95.
55. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270-3.
56. Li X, Zai J, Zhao Q, Nie Q, Li Y, Foley BT, et al. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. J Med Virol. 2020;92(6):602-11.
57. Mishra A, Pandey AK, Gupta P, Pradhan P, Dhamija S, Gomes J, et al. Mutation landscape of SARS-CoV-2 reveals three mutually exclusive clusters of leading and trailing single nucleotide substitutions. bioRxiv. 2020:2020.05.07.082768.
58. Seemann T, Lane CR, Sherry NL, Duchene S, Gonçalves da Silva A, Caly L, et al. Tracking the COVID-19 pandemic in Australia using genomics. Nature Communications. 2020;11(1):4376.

**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.

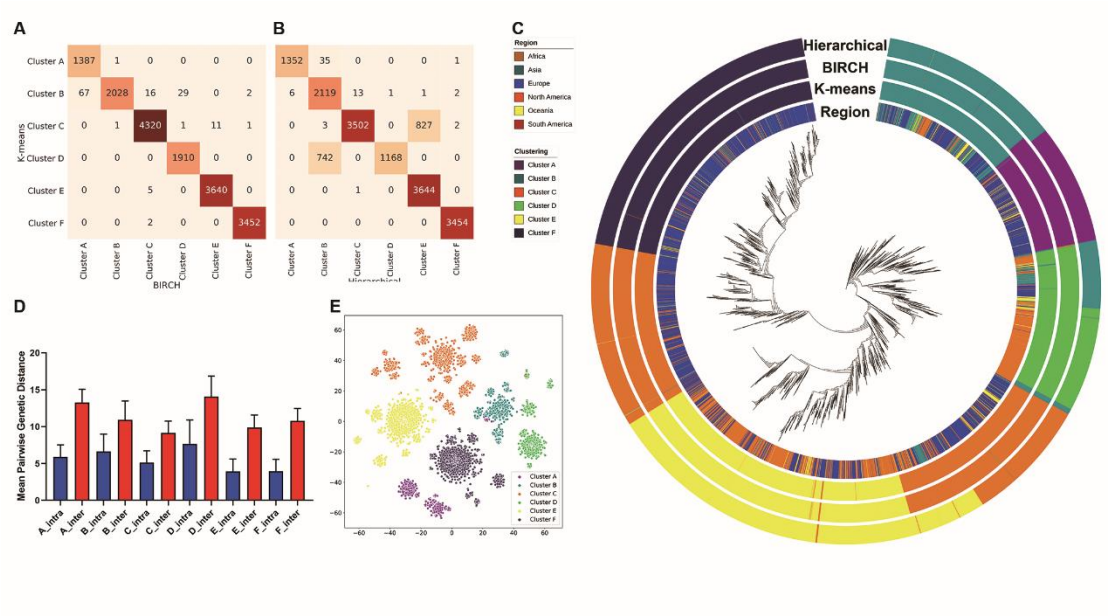


**Figure 1.** Clustering of SARS-CoV-2 reveals six major clusters. (A and B) Confusion matrix for the similarity of K-means clustering with BIRCH clustering (A) and Hierarchical clustering (B). Each row corresponds to the results of K-means clustering and each column corresponds to the clustering results 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.

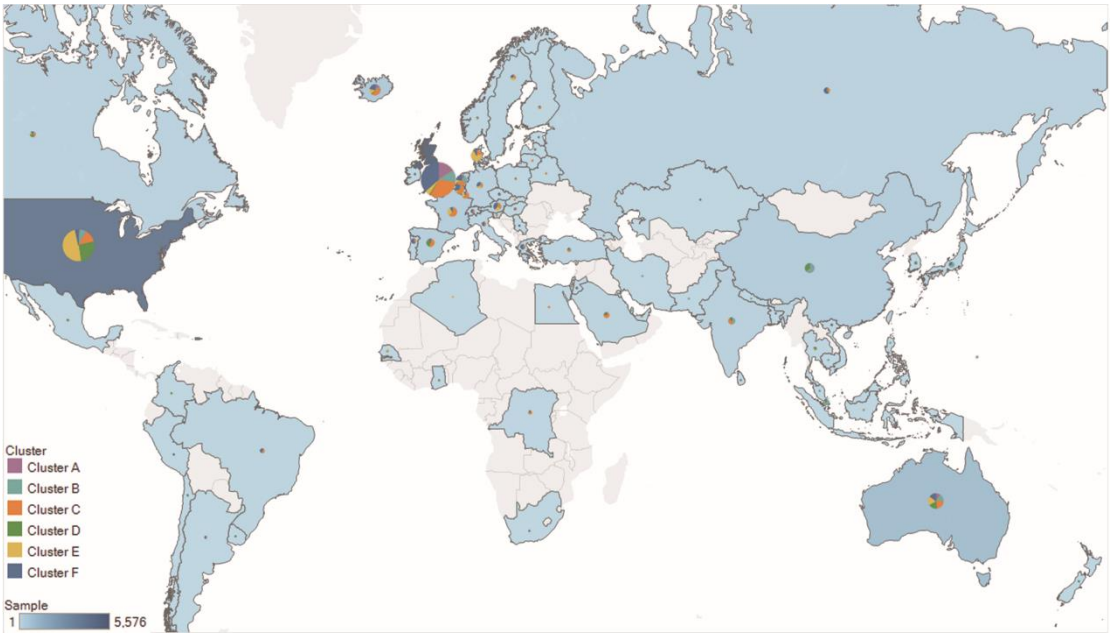


**Figure 2.** Geographic distributions of the six clusters. Pie charts display the proportions of six clusters among all SARS-CoV-2 strains in each country. Circle sizes and the color scales correspond to the number of strains analyzed per country.

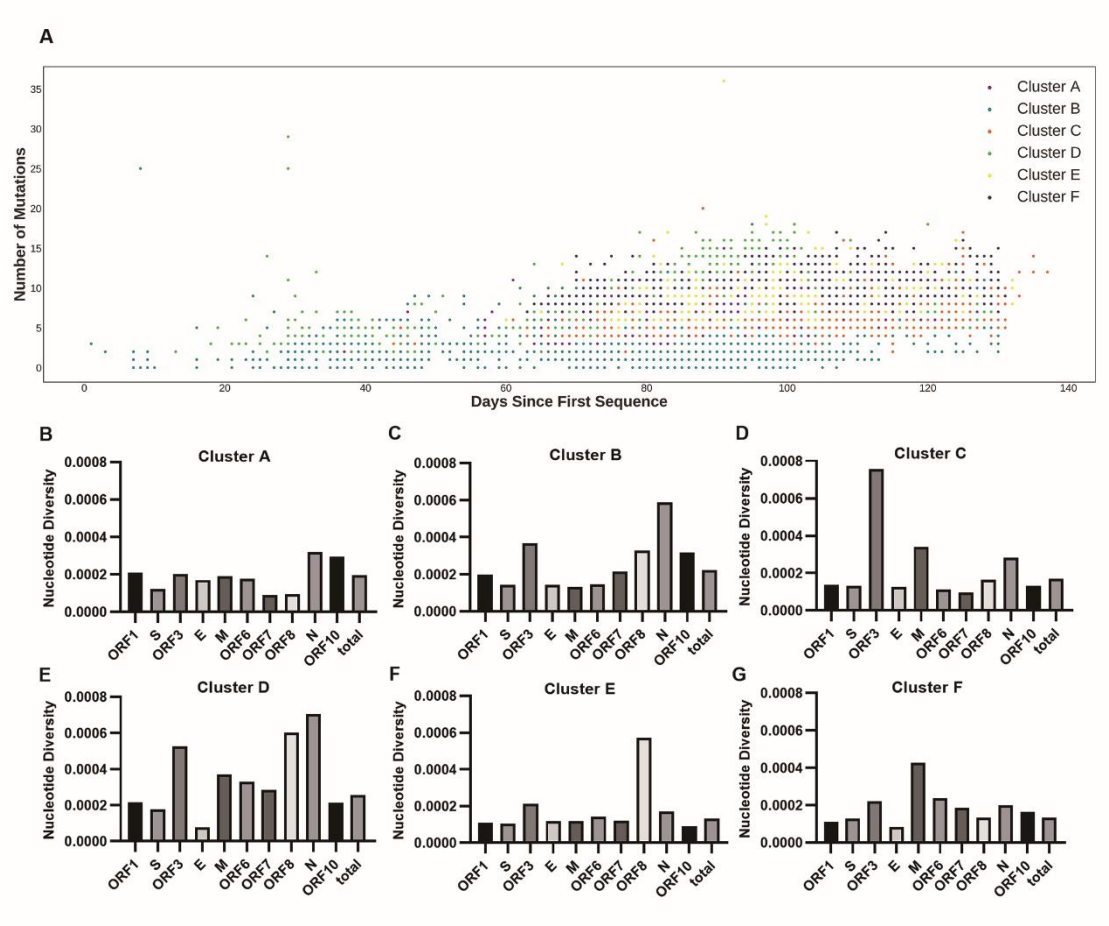


**Figure 3.** The genetic diversity between clusters. (A) The mutation counts over days of 16,873 SARS-CoV-2 strains. The X axis represents the days from the corresponding collection date of strains to 24 December 2019 when the earliest strain (EPI_ISL_402123) was collected. The Y axis represents the 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.

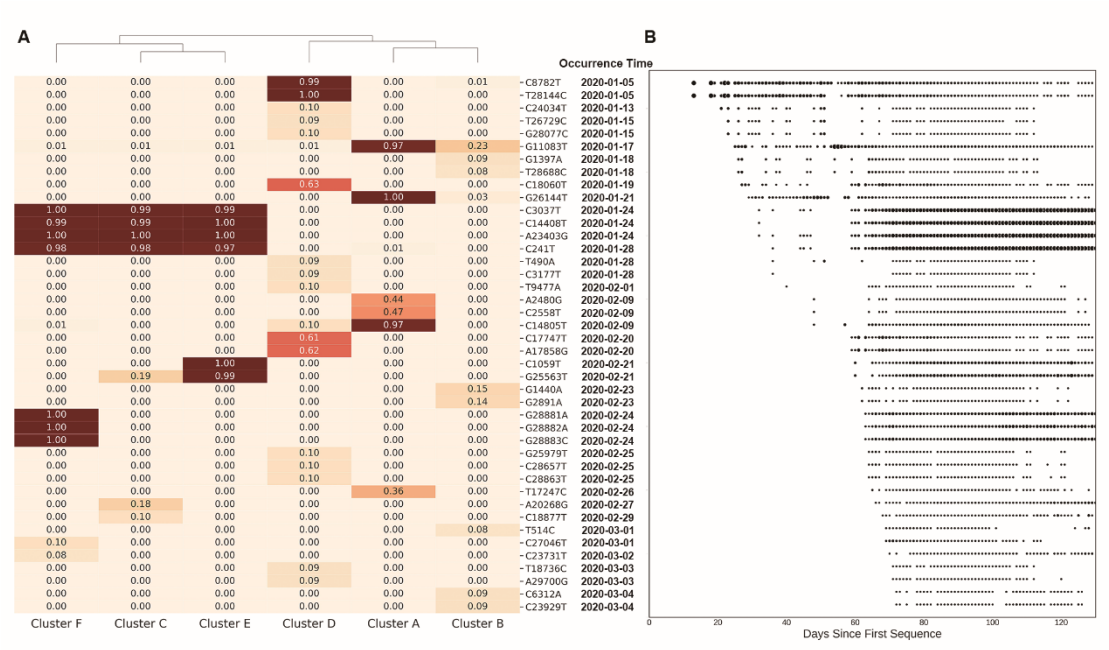


**Figure 4.** The clustering of the six clusters by the extracted mutations. (A) The heatmap displays mutation frequency of the 42 mutations across six clusters. The colors and values represent different frequencies of the corresponding mutations in each cluster. The collected days of the mutations are 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 |

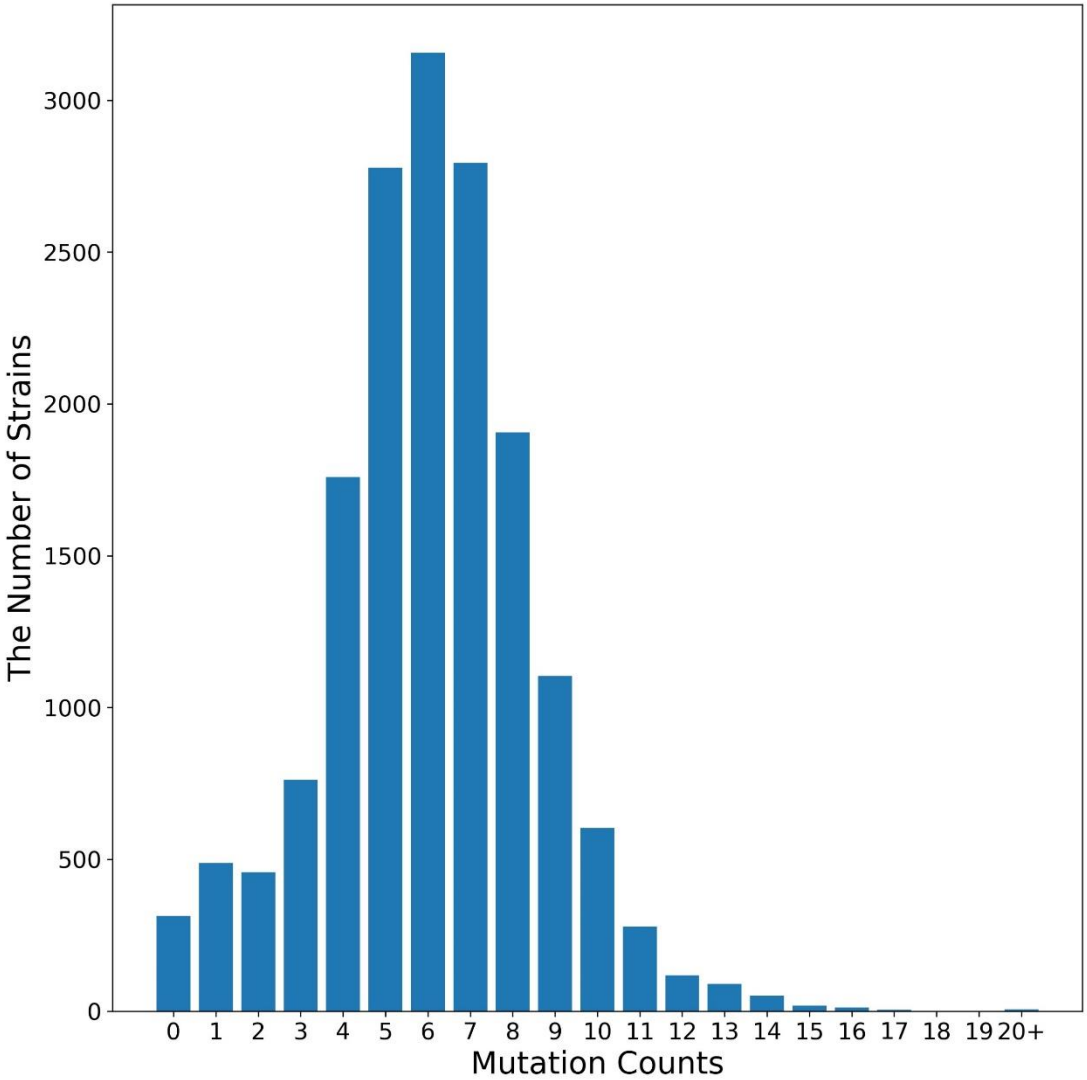


**Figure S1.** The distribution of the mutation counts of the 16,873 SARS-CoV-2 strains.

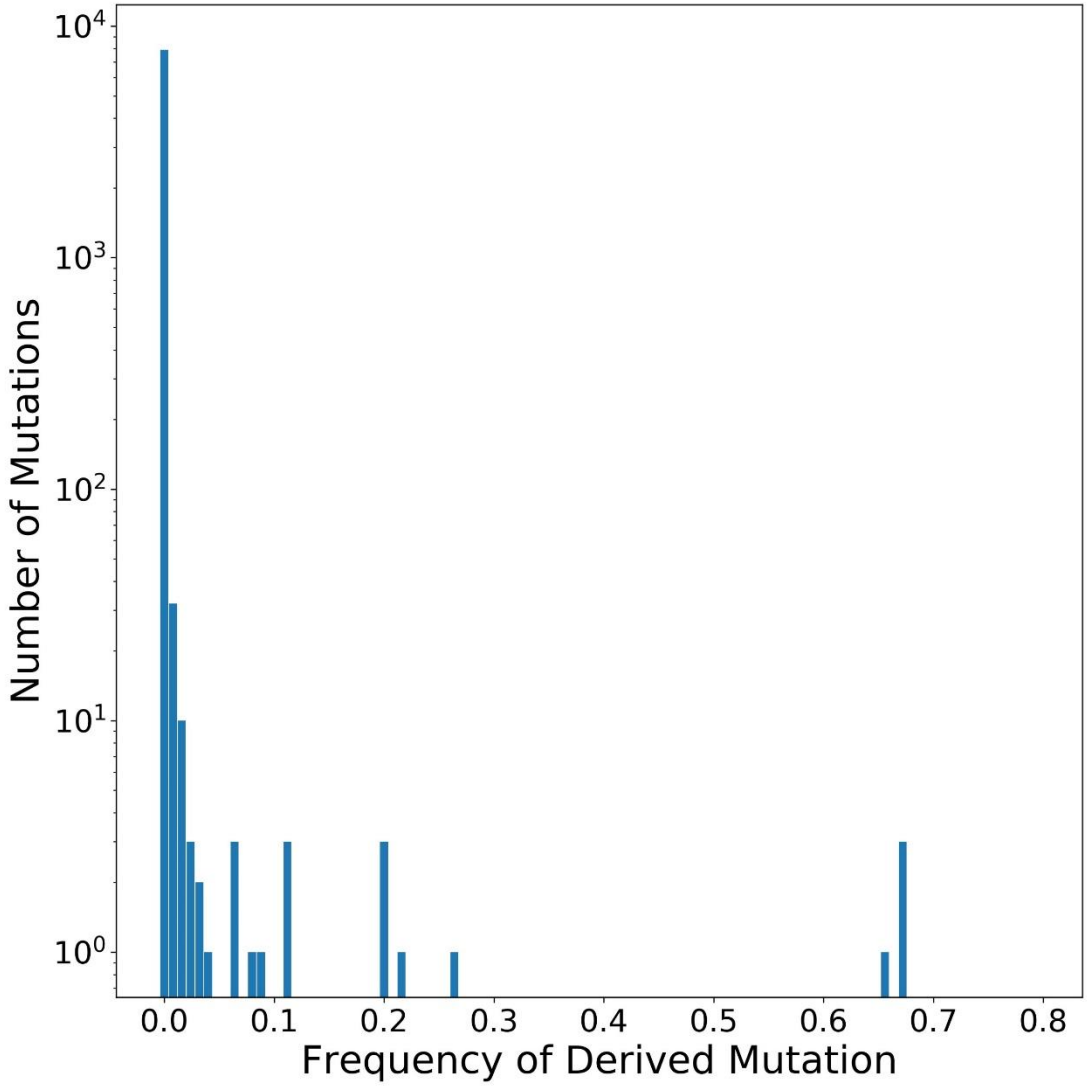


**Figure S2.** Frequency spectra of SARS-CoV-2. The mutation frequency of derived mutations of 16,873 SARS-CoV-2 stains is depicted on the X axis, and the number of mutations in which strains occurred is displayed on the Y axis. A log-10 scale is used for the Y axis of the graph, and the Y axis ranges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

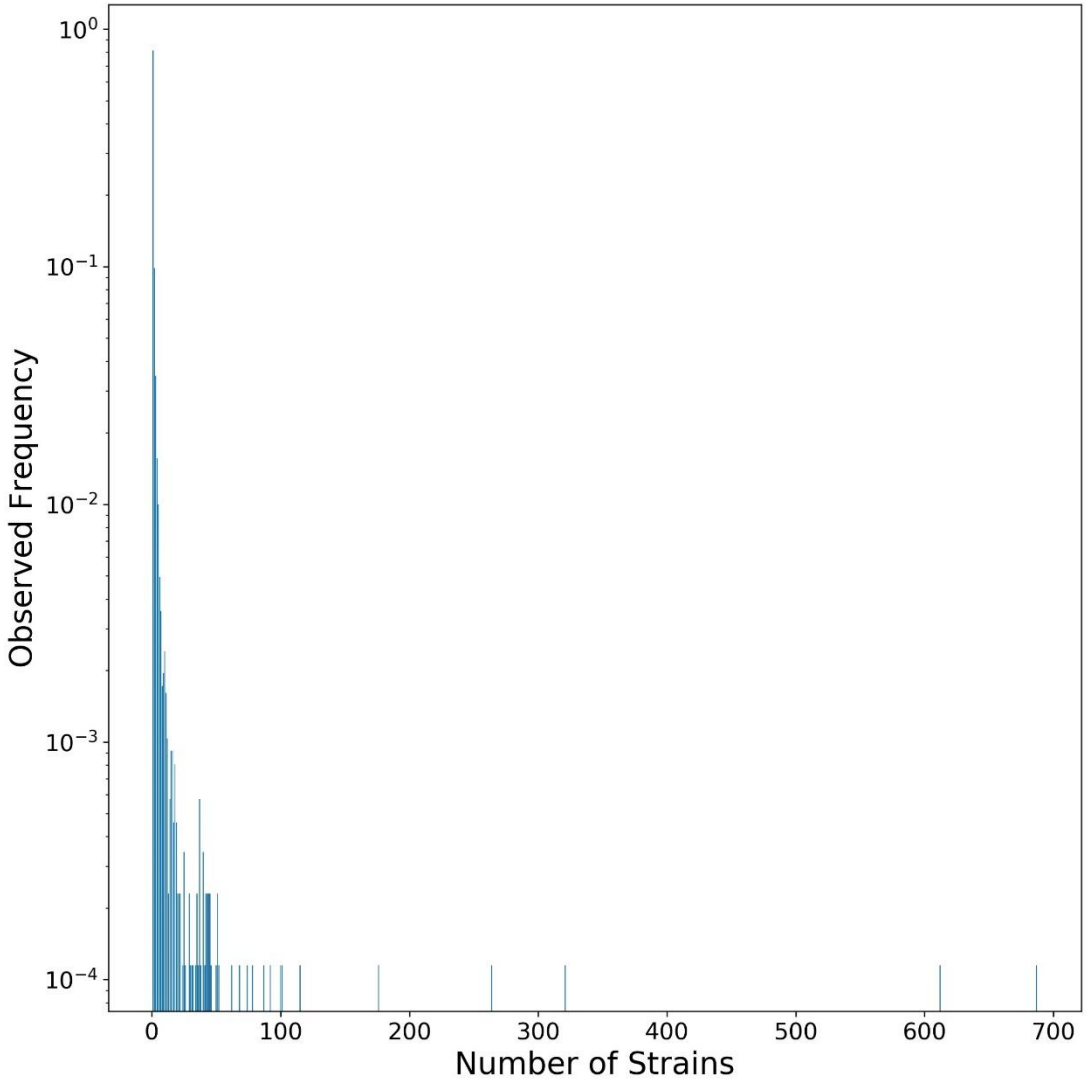


**Figure S3.** Normalized allele frequency of 16,873 SARS-CoV-2 strains. There are 8,706 unique genomes across the 16,873 strains. The X axis is the number of strains for each unique genome and the Y axis is the proportion of the unique genomes. A log-10 scale is used for the Y axis of the graph, and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

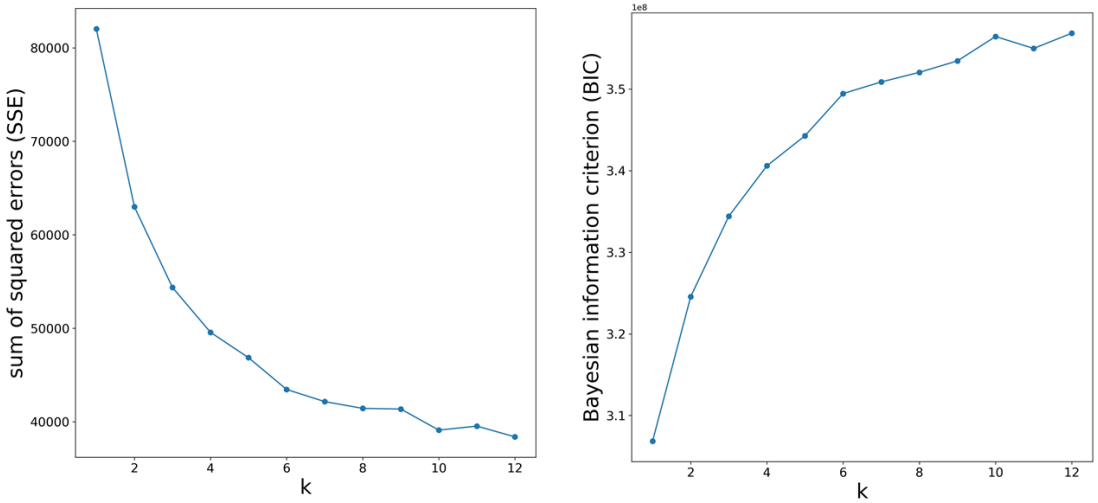


**Figure S4.** The evolution of the sum of squared errors (SSE; left) and Bayesian information criterion (BIC; right) for the number of clusters (K) in the K-means runs.

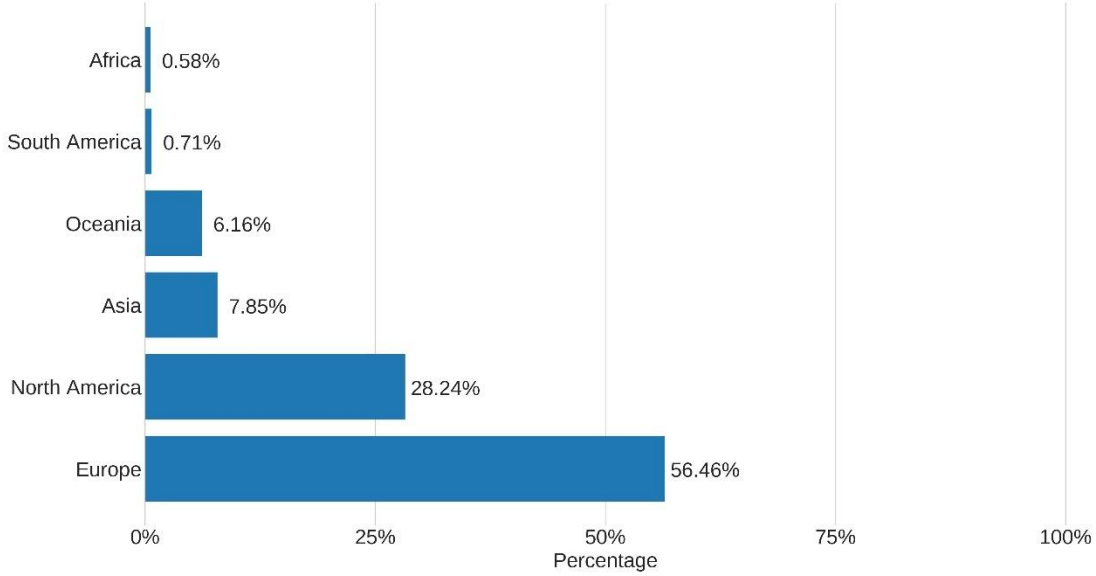


**Figure S5.** The bar chart displays the global proportions of the strains collected from each continent. More than half of the strains were collected from Europe.

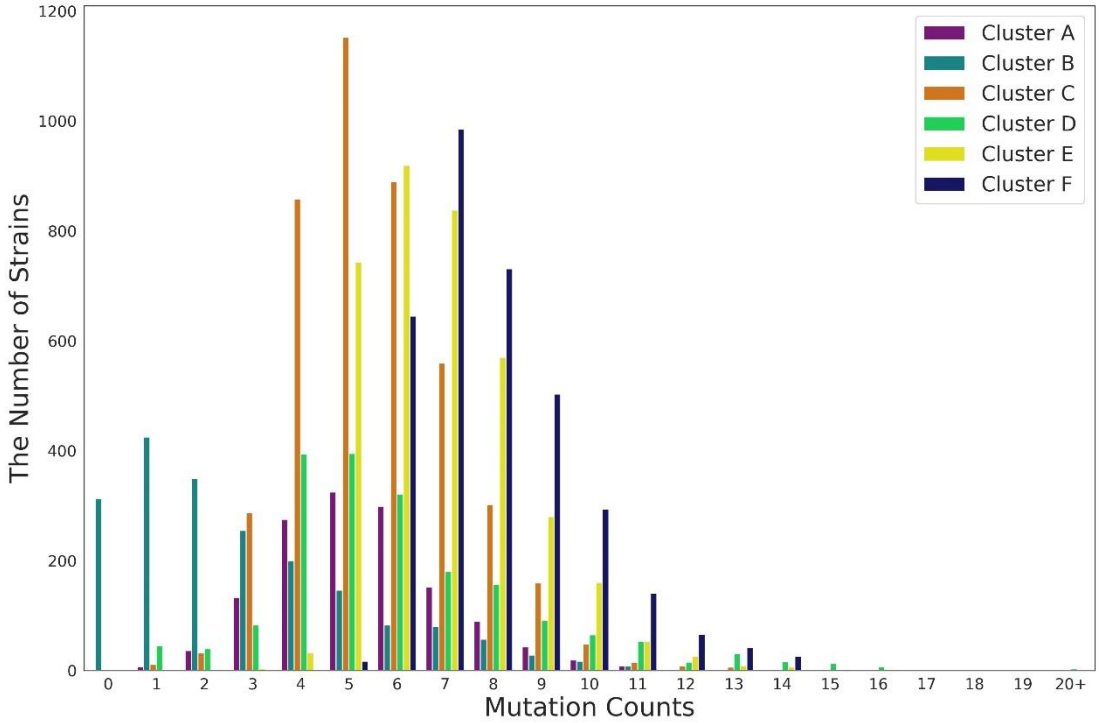


**Figure S6.** The distribution of the mutation counts of the strains for the six clusters.

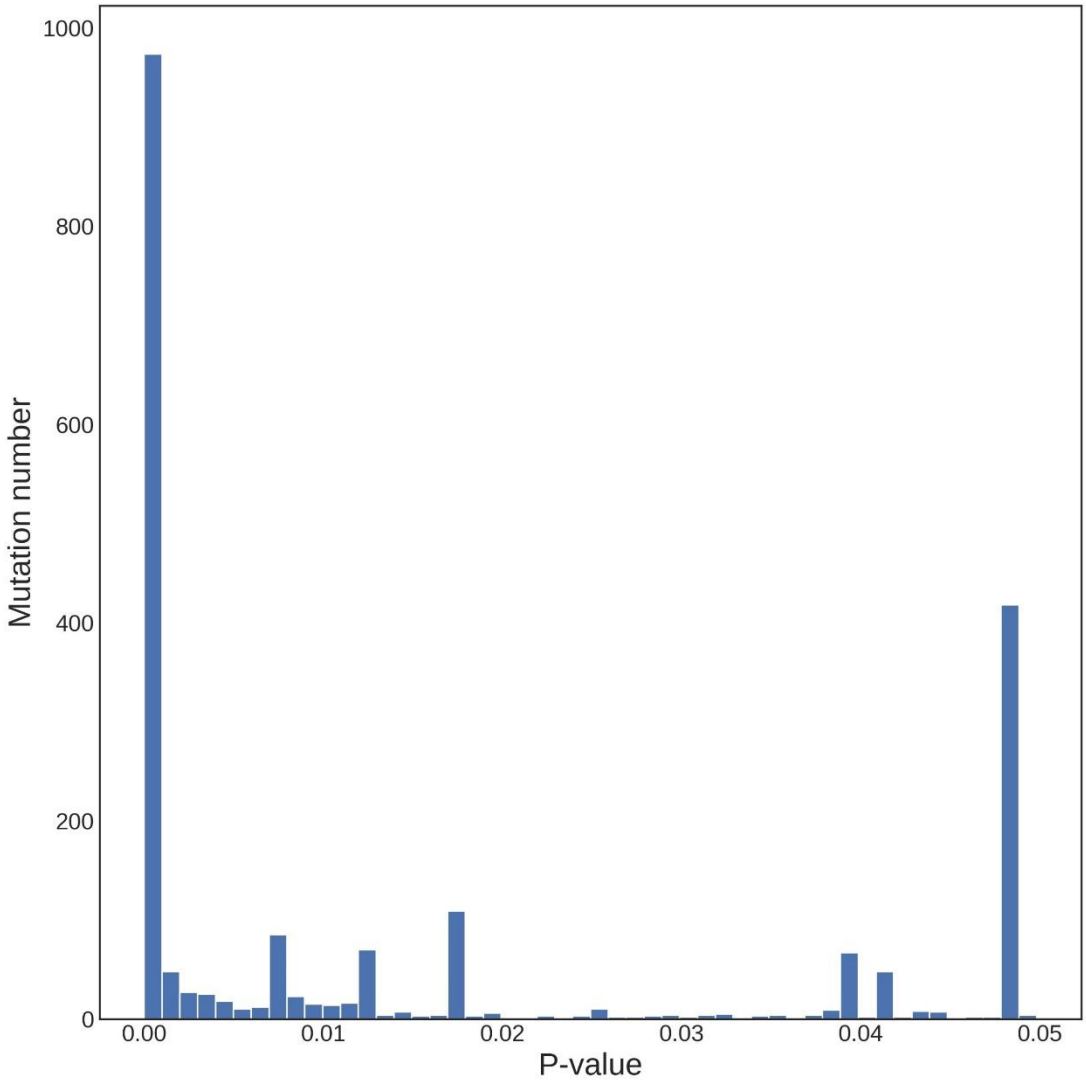


**Figure S7.** The distribution of P-values from the 2,094 mutations with a P-value <0.05 by ANOVA.

**Figure S8.** 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.

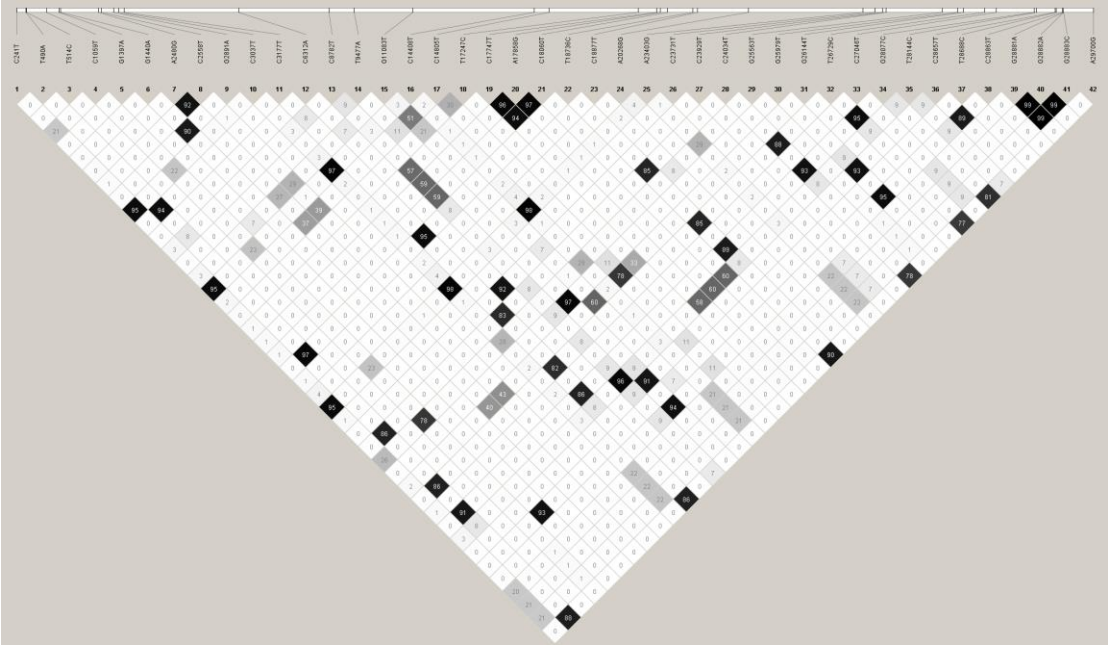

**Figure S9.** The heatmap displays the $r^2$ of the 42 mutations. The numbers within the squares represent the $r^2$ scores for pairwise LD. $r^2$ values are represented by white for $r^2 = 0$, with intermediate values for $0 < r^2 < 1$ indicated by shades of grey.

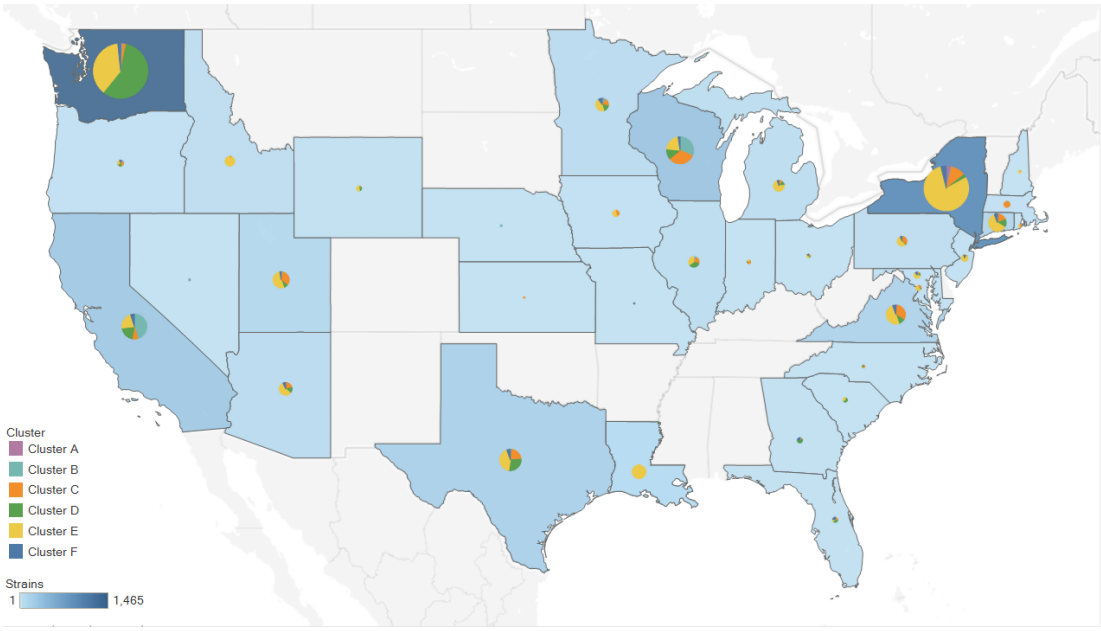


**Figure S10.** Geographic distribution of six clusters in the United States. Pie charts display the proportions of six clusters among all SARS-CoV-2 strains in each state. Circle sizes and the color scales correspond to the number of strains analyzed per state.

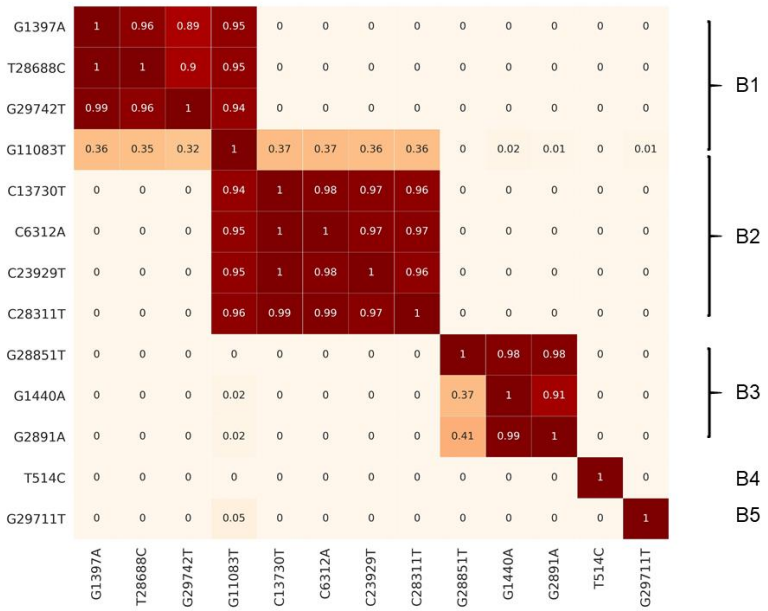


**Figure S11.** The pairwise dependency score (see materials and methods) of the mutations with frequency >0.05 within group B.