Supplementary Information: 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

[*] Corresponding author:

Email: yuan.luo@northwestern.edu

Supplementary Information

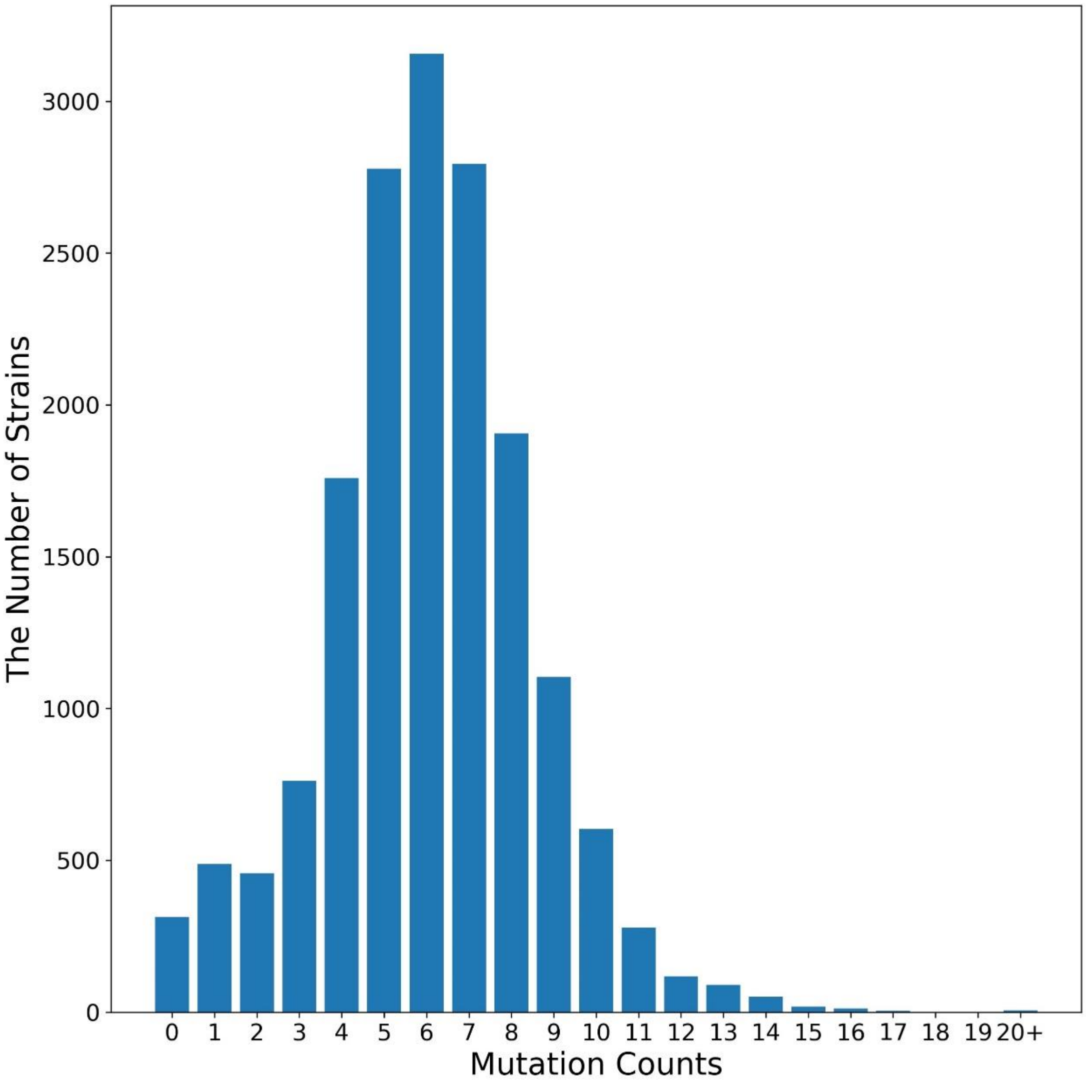

Figure S1. The distribution of the mutation counts of the 16,873 SARS-CoV-2 strains.

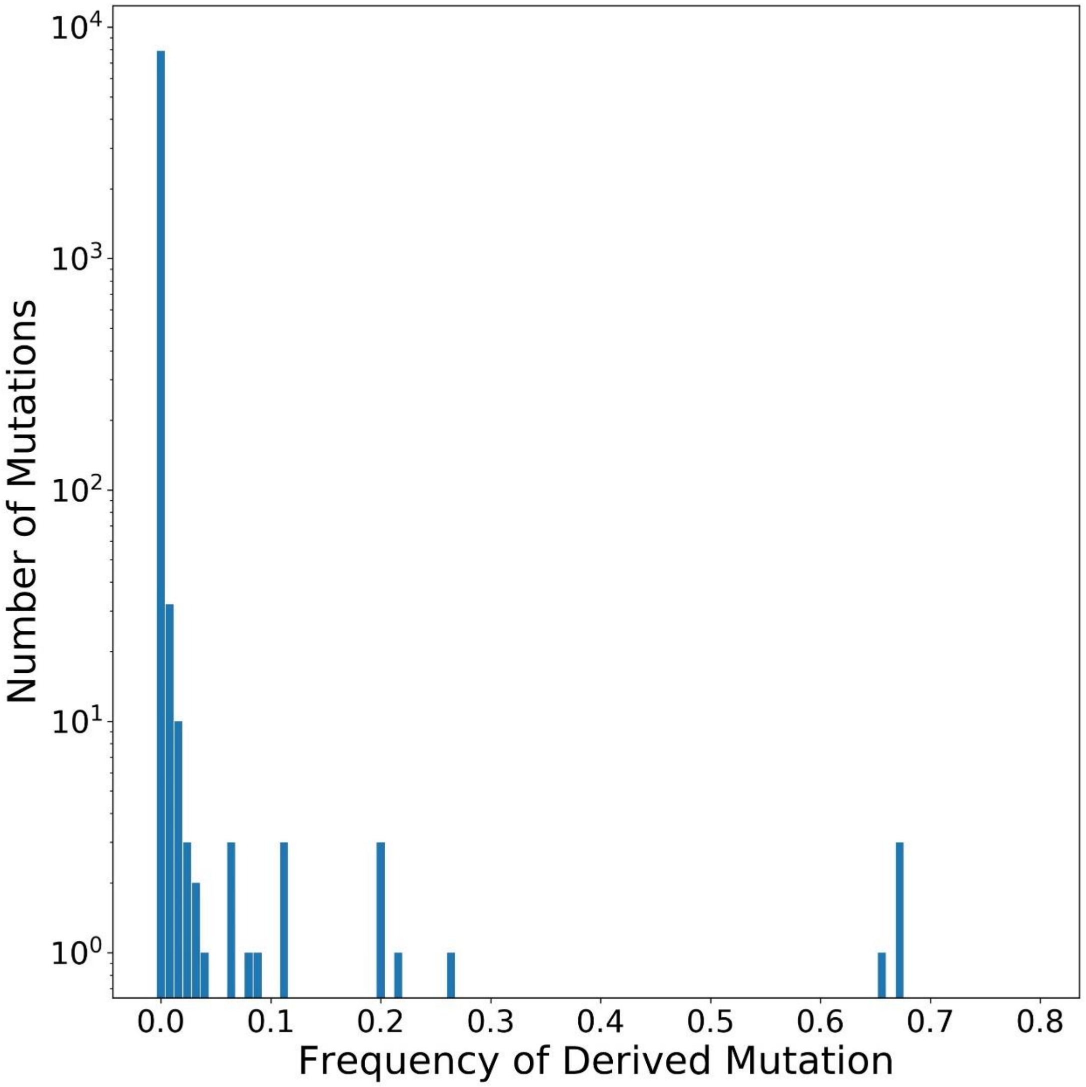

Figure S2. Frequency spectra of SARS-CoV-2. The mutation frequency of derived mutations of 16,873 SARS-CoV-2 stains is depicted on the X axis, and the number of mutations in which strains occurred is displayed on the Y axis. A log-10 scale is used for the Y axis of the graph, and the Y axis ranges from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

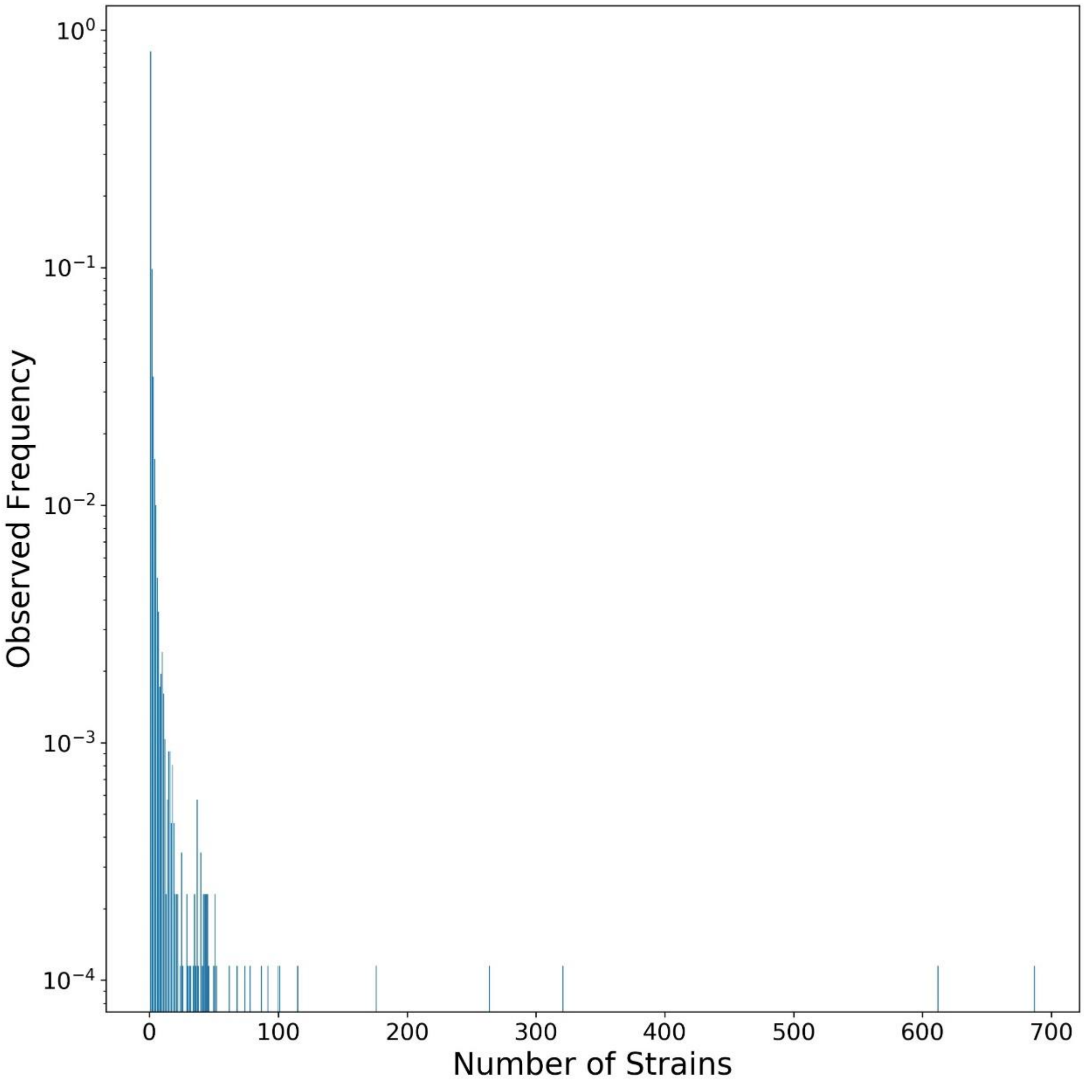

Figure S3. Normalized allele frequency of 16,873 SARS-CoV-2 strains. There are 8,706 unique genomes across the 16,873 strains. The X axis is the number of strains for each unique genome and the Y axis is the proportion of the unique genomes. A log-10 scale is used for the Y axis of the graph, and the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

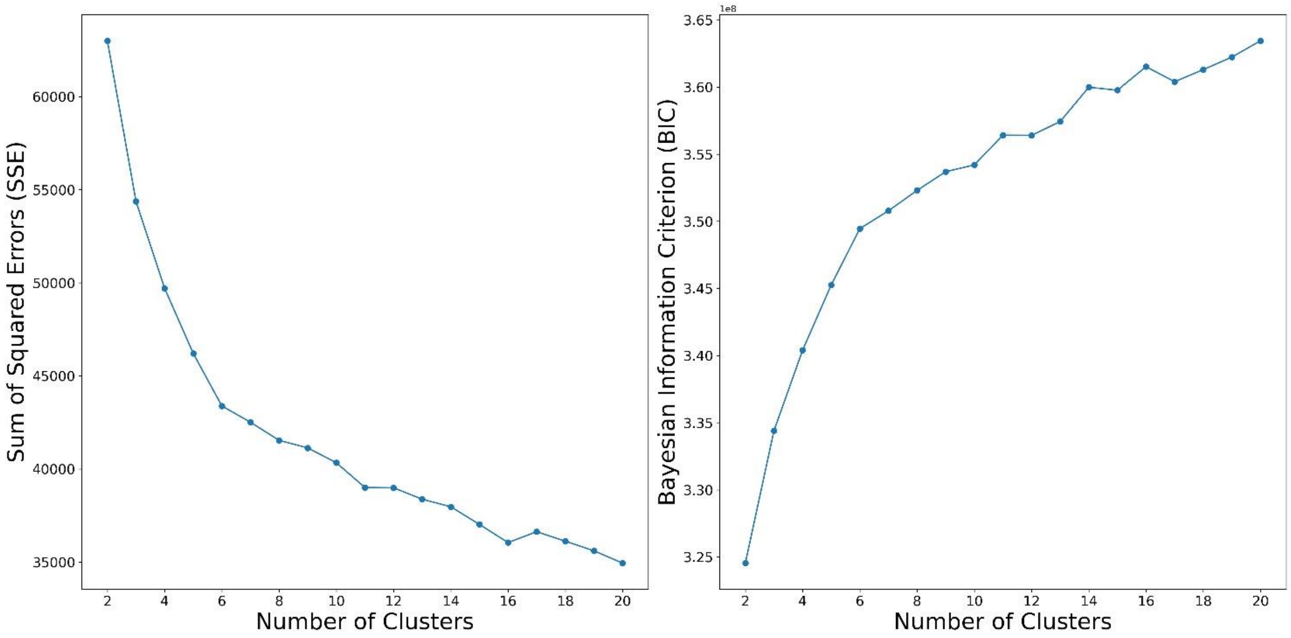

Figure S4. Evaluation of the number of clusters. The evolution of the sum of squared errors (SSE; left) and Bayesian information criterion (BIC; right) for the number of clusters in the deep learning clustering runs. We used the elbow method and chose the elbow of the curve as the number of clusters. The elbow method indicated that the number of clusters is six.

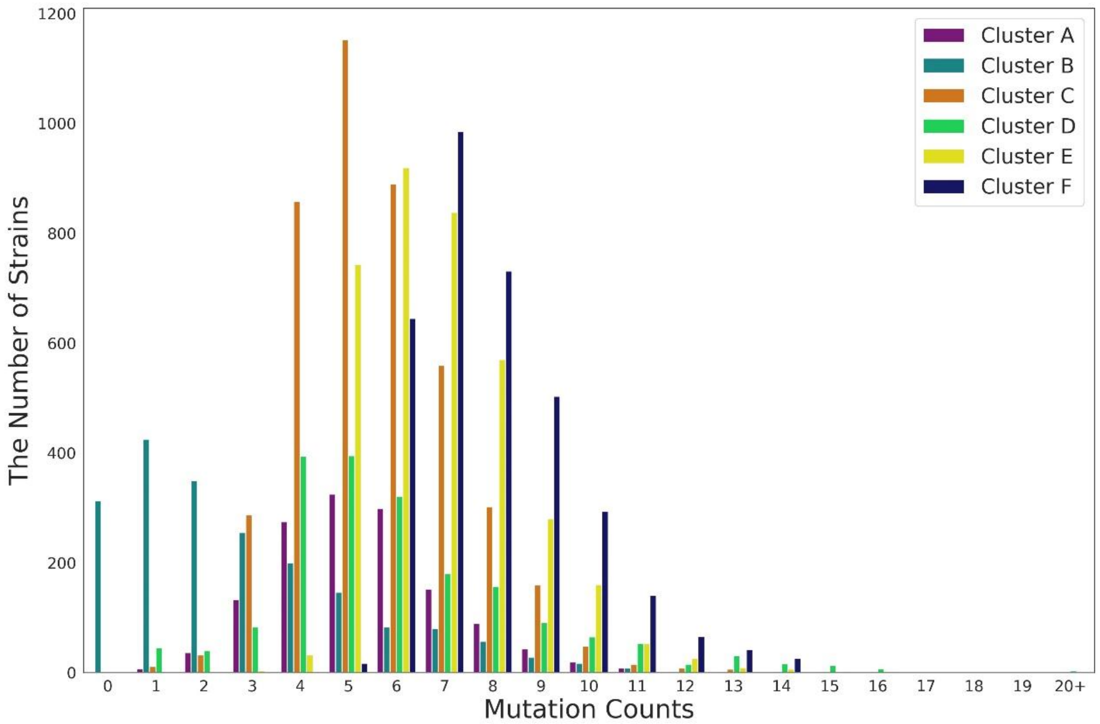

Figure S5. The distribution of the mutation counts of the strains for the six clusters.

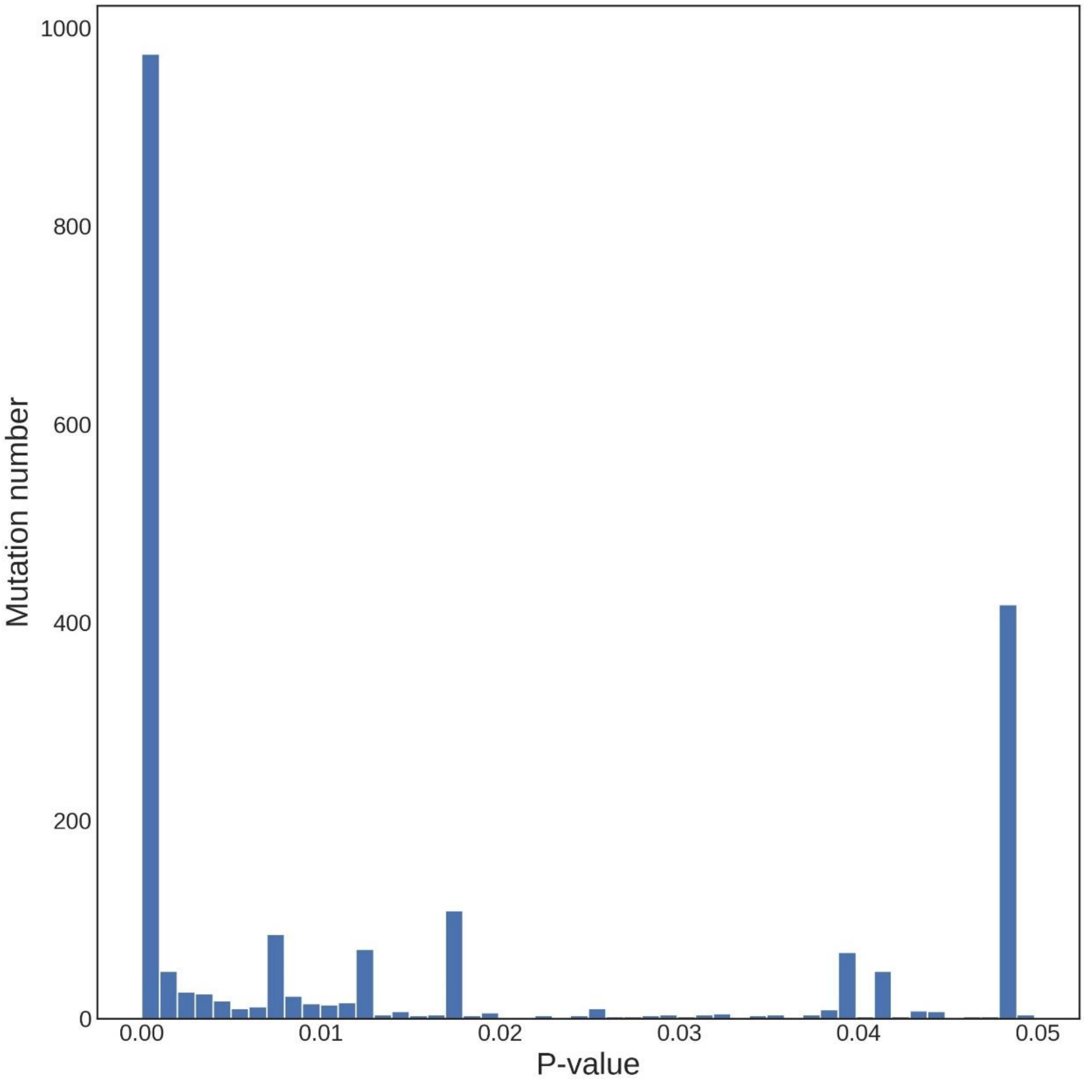

Figure S6. The distribution of P-values from the 2,094 mutations with P-values <0.05 by ANOVA.

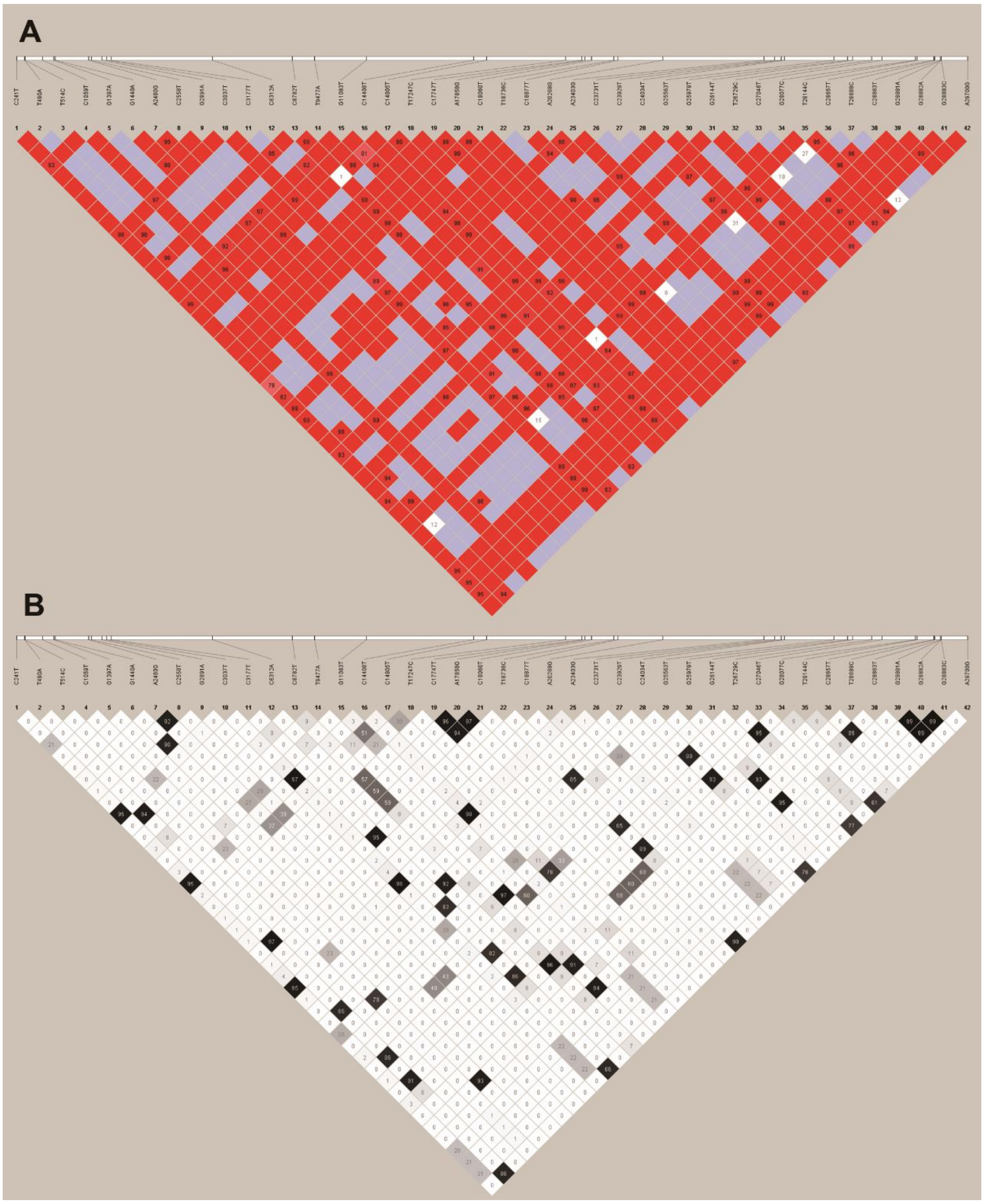

Figure S7. The D' and r^2 of the 42 mutations. (**A**) D' values that correspond to substitution pairs are expressed as percentages and are shown within the respective squares. Higher D' values are indicated with a brighter red color. (**B**) The numbers within the squares represent the r^2 scores for 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.

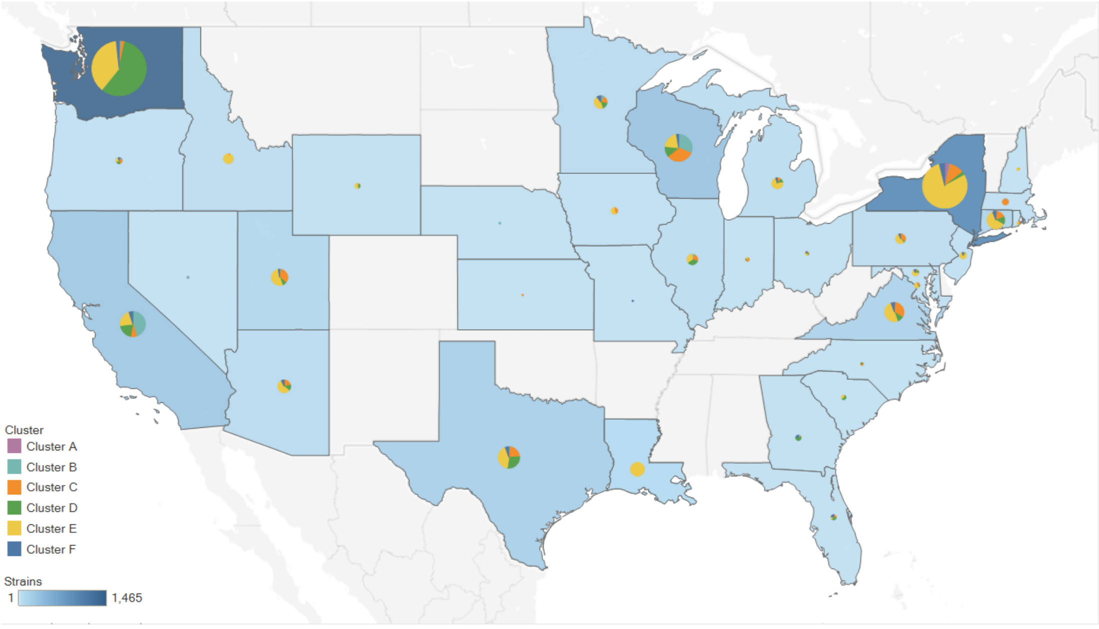

Figure S8. Geographic distribution of six clusters in the United States. Pie charts display the proportions of six clusters among all SARS-CoV-2 strains in each state. Circle sizes and the color scales correspond to the number of strains analyzed per state.

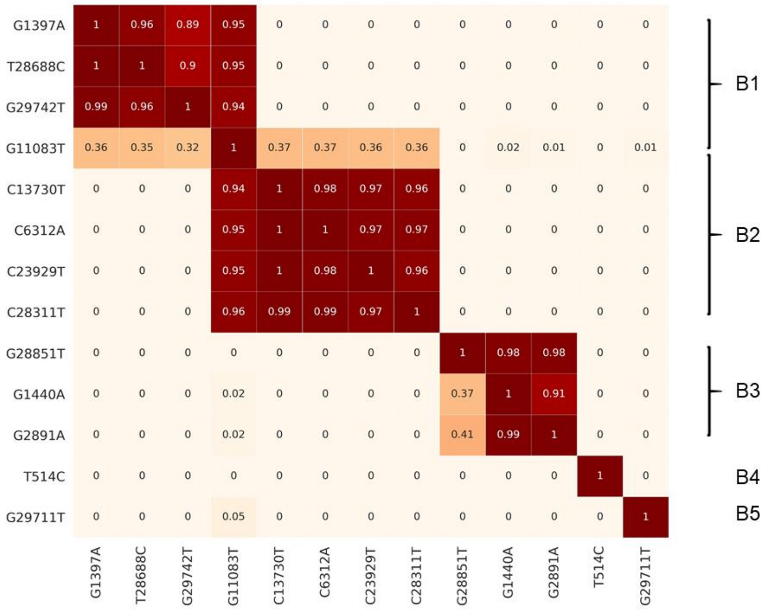

Figure S9. The pairwise dependency score (see Materials and Methods) of the mutations with frequency >0.05 within cluster B. The heatmap shows that there are five major subclusters within cluster B.

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