Supporting Information
Decoding SARS-CoV-2 Transmission, Evolution, and Ramifications for COVID-19 Diagnosis, Vaccine, and Medicine

Rui Wang[1], Yuta Hozumi[1], Changchuan Yin[2] *, and Guo-Wei Wei[1,3,4] †

[1] Department of Mathematics, Michigan State University, MI 48824, USA

[2] Department of Mathematics, Statistics, and Computer Science,
University of Illinois at Chicago, Chicago, IL 60607, USA

[3] Department of Biochemistry and Molecular Biology
Michigan State University, MI 48824, USA

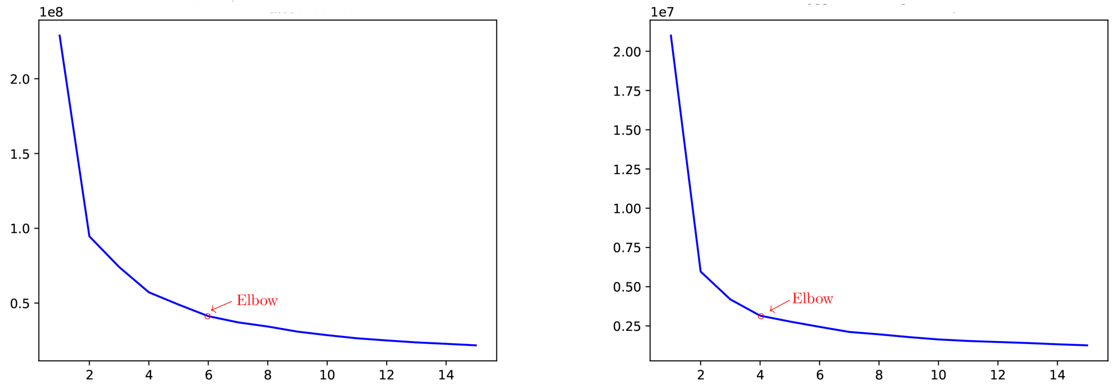
[4] Department of Electrical and Computer Engineering
Michigan State University, MI 48824, USA

* Address correspondences to Changchuan Yin. E-mail:cyin1@uic.edu

† Address correspondences to Guo-Wei Wei. E-mail:wei@math.msu.edu

S1 K-mean clustering for optimal groups

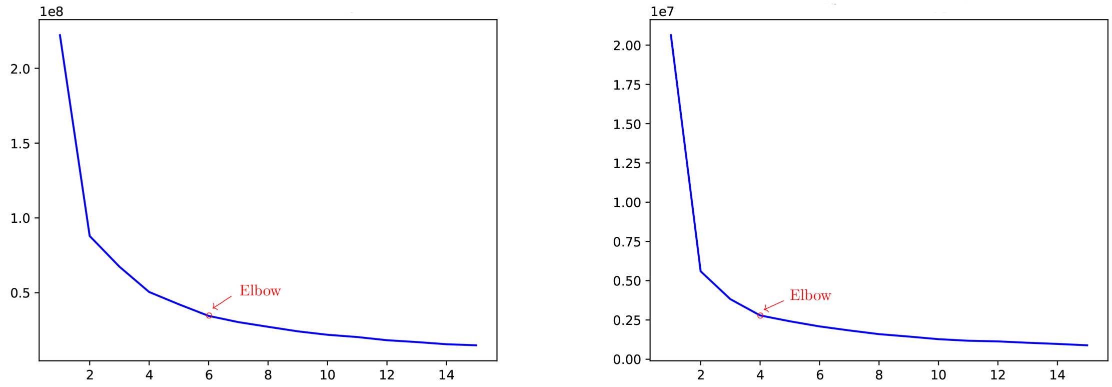
The K-means clustering is used to classify the SARS-CoV-2 genetypes. The Elbow and results demonstrate five main clusters of SARS-CoV-2 SNP variants in the world, and three main clusters in the U.S. Figure S1 is the plot of the within-cluster sum of squares according to the number of clusters k for the SNP variants in the world (left chart) and in the US (right chart) based on Jaccard distance metric. The optimal values of k-mean clusters are shown as the turning point in the in the elbow plots. Figure S2 shows the plot of WCSS according to the number of clusters based on PCA method.



(a) World

(b) US

Figure S1: The plot of WCSS according to the number of clusters based on Jaccard distance metric. Here, Jaccard distance-based representation is taken as the input feature. The arrows point out the optimal number of clusters. (a)The within-cluster sum of squares against the number of clusters for the SNP variants in the world. In the plot above the Elbow is at $k = 6$, which indicates that the optimal number of clusters worldwide is six. (b) The within-cluster sum of squares against the number of clusters for the SNP variants in the US. The optimal number of clusters in the US is four.



(a) World

(b) US

Figure S2: The plot of WCSS according to the number of clusters based on PCA method (Principle components = 30). Here, Jaccard distance-based representation is taken as the input feature. The arrows point out the optimal number of clusters. (a)The within-cluster sum of squares against the number of clusters for the SNP variants in the world. In the plot above the Elbow is at $k = 6$, which indicates that the optimal number of clusters worldwide is six. (b) The within-cluster sum of squares against the number of clusters for the SNP variants in the US. The optimal number of clusters in the US is four.

S2 Supplementary Tables

Total 26 spreadsheets are merged in the SupportingInformation.xlsx.

Table S1: S1_snpRecords_06012020_World: The SNP profiles in the world (Up to June 1, 2020).

Table S2: S2_snpRecords_06012020_US: The SNP profiles in the U.S. (Up to June 1, 2020).

Table S3: S3_World_Cluster_I: The Cluster I in the world.

Table S4: S4_World_Cluster_II: The Cluster II in the world.

Table S5: S5_World_Cluster_III: The Cluster III in the world.

Table S6: S6_World_Cluster_IV: The Cluster IV in the world.

Table S7: S7_World_Cluster_V: The Cluster V in the world.

Table S8: S8_World_Cluster_VI: The Cluster VI in the world.

Table S9: S9_US_Cluster_A: The Cluster A in the U.S.

Table S10: S10_US_Cluster_B: The Cluster B in the U.S.

Table S11: S11_US_Cluster_C: The Cluster C in the U.S.

Table S12: S12_US_Cluster_D: The Cluster D in the U.S.

Table S13: S13_3CL protease: The records of single mutation site, mutation site on protein and frequencies in different clusters on 3CL protease.

Table S14: S14_Spike protein: The records of single mutation site, mutation site on protein and frequencies in different clusters on spike protein.

Table S15: S15_RNA polymerase: The records of single mutation site, mutation site on protein and frequencies in different clusters on RNA polymerase.

Table S16: S16_Papain-like protease: The records of single mutation site, mutation site on protein and frequencies in different clusters on Papain-like protease.

Table S17: S17_Endoribo-nuclease: The records of single mutation site, mutation site on protein and frequencies in different clusters on Endoribo-nuclease.

Table S18: S18_Envelope protein: The records of single mutation site, mutation site on protein and frequencies in different clusters on Envelope protein.

Table S19: S19_Membrane protein: The records of single mutation site, mutation site on protein and frequencies in different clusters on Membrane protein.

Table S20: S20_Nucleocaspid protein: The records of single mutation site, mutation site on protein and frequencies in different clusters on Nucleocaspid protein.

Table S21: Acknowledgement table provided by GISAID in Jan 2020.

Table S22: Acknowledgement table provided by GISAID in Feb 2020.

Table S23: Acknowledgement table provided by GISAID in March 2020.

Table S24: Acknowledgement table provided by GISAID in April 2020.

Table S25: Acknowledgement table provided by GISAID in May 2020.

Table S26: Acknowledgement table provided by GISAID in June 2020.