Modeling and visualizing the SARS-CoV-2 mutation based on geographical regions and time

Bomin Wei

Abstract

The Coronavirus Disease 2019 (COVID-19) epidemic was first detected in late-December 2019 in Wuhan, China. So far, it has caused more than one hundred million confirmed cases and over two million deaths in the world. The original genome of SARS-CoV-2 contains several open reading frames (ORFs) that encode the spike (S) glycoprotein, the replicase polyprotein, membrane (M), envelope (E), nucleocapsid (N) proteins, and accessory protein. 150,659 COVID-19 sequences were collected from China National Center for Bioinformation 2019 Novel Coronavirus Resource. Based on the previous phylogenomic analysis, I found three major branches of the virus RNA genomic mutation located in Asia, America, and Europe which is consistent with other studies. I selected sites with high mutation frequencies from Asia, America, and Europe. There are only 13 sites have a high mutation rate in all of these three regions. It infers that the viral mutations are highly dependent on their location and different locations have specific mutations. Most mutations can lead to amino acid substitutions. These substitutions occurred in 3/5'UTR, S/N/M protein, and ORF1ab/3a/8/10. Thus, the mutations may affect the pathogenesis of the virus. Additionally, I developed a prediction model for the frequency change of these top mutation sites during the spread of the disease. The MSE values show that the prediction is reliable. It provides a reference for researchers on different continents. I plan to develop a website to visualize the prediction model of SARS-CoV-2 mutation in future work. Vaccines developed in different countries may have diverse efficiencies in other countries. The similarity of genomes in different countries can indirectly reflect the efficiency of vaccines while there is no direct information. Our system can provide local residents with an alternative way to assess the availability of vaccines.

Keywords: COVID-19, SARS-CoV-2, genomic mutation, prediction model, website

INTRODUCTION

The outbreak of the Coronavirus Disease 2019 (COVID-19) has become a severe epidemic, claiming more than 100,000,000 cases and 2,000,000 death worldwide until now[1]. The COVID-19 is caused by a novel evolutionary divergent RNA virus, called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which triggers a respiratory tract infection and spreads mainly through person-to-person contact [2]. The genetic information of SARS-CoV-2 mutates much more dramatically than DNA due to RNA viruses' mechanisms[3]. The worldwide outbreak happens to provide good environments for SARS-CoV-2 mutations. The accumulation of these mutations may

cause the COVID-19 to develop in an uncertain direction, which will have a huge impact on society and personal life [4].

According to other epidemiological studies, mutations in the genome of an epidemic will be inherited from the spreader to the next generation of patients. The spread of the disease generally has regional characteristics, which leads to the diversity of the genome with regional traits. Thus, the SARS-CoV-2 genome mutation should have divergent mutation patterns in different geographic locations. Our purpose is to study the mutation patterns of SARS-CoV-2 in the world and try to predict the trend of the mutation so that to provide a reference for other researchers and may be helpful to the choice of the vaccine. Finally, I shall develop a visualizing system to model the SARS-CoV-2 mutation trend based on geographical regions and time.

METHODS

Sample data filtering

As of January 31, 2021, the China National Center for Bioinformation (CNCB) database hosted 528,611 SARS-COV-2 sequences. I filtered out low-quality data (as assessed by the database) and excluded the data that required authorization by the submitting agencies, and I was able to download 184,475 entries of raw sequence data[5]. An entry of raw sequence data directly downloaded from the CNCB database does not contain all the necessary metadata information such as host, sampling location, or date. The information is documented in a set of sequence metadata information, which needs to be downloaded separately. I performed a complete paring search to match the raw sequence data to the sequence metadata information and obtained 183,850 data entries with metadata information.

I discarded irrelevant metadata information and kept the following information for analysis and modeling: sequence name, detection time, detection region, base name before mutation, location of mutation site, base name after mutation. The base names and site also guarantee an easy retrieval of information needed for amino acid substitution analysis.

The data was divided based on the continent of the detection region and then sorted by time. I performed a further data cleaning and dropped those data with the non-standard format or incomplete metadata information. For example, I dropped a group of data reported from Japan which labeled detection time with only the month and the date but not the year. After the cleaning, 150,659 data entries remained.

I focused on data from Europe, Asia, and North America due to sufficient data quantity and high data quality. For analysis and modeling, I selected those sites which have more than 0.1% mutation rate on the last day and larger than 10% mutation rate on average.

Development of prediction model

For the prediction model, the ARIMA (i.e, Autoregressive Integrated Moving Average) model is used because of its success in time series forecasting. An ARIMA model contains three parameters – p, d, and q, or written as ARIMA(p,d,q).

The p represents the number of lag observations in the auto-regression (AR) part of the model, indicating the relation between an observation (or data) to the past observations.

The q represents the size of the moving average window in the moving average (MA) part of the model, indicating the relation between an observation to the past error. The d represents the integration order of the I part of the model, indicating the number of times that the raw observations are differenced. If we let $\hat{y}$ be the d^{th} difference of Y (the observation or data), then the model can be expressed as $\widehat{y_t} = \mu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}$.

Each site should have its model parameters fitted independently, and an automatic parameter scan was performed for about 90 groups of data using a computer program developed by the author. For the mutation rate data of each site, the program tests 5 different parameters for p and q, respectively, and number 1 or 2 for parameter d. In total, there were 50 parameter combinations tested for every site of the 30 high rate sites. Values of p are first determined from ACF (autocorrelation function) of the mutation rate. The best parameter set (p, q, d) which has the lowest MSE value will be used as parameters for the prediction.

RESULTS

Data source and distribution

Since COVID-19 has been circulating for more than a year, virus samples from many countries have been sequenced. The samples range dozens of countries and the sampling time covers several months. This gives us great convenience to study gene mutations in different regions and their changing trends over time. I downloaded mutation data of disease sequences from a public database: China National Center for Bioinformation, 2019nCoVR (https://bigd.big.ac.cn/ncov/?lang=en) [5]. In total, SARS-CoV-2 sequence mutation information from 150,659 patients was downloaded from the database. The reference sequence used in this study is NC_045512 with a length of 29,903 bp ss-RNA in National Center for Biotechnology Information (NCBI) [6]. The corresponding sample location (country) and sampling time were also obtained. According to the source of the sample, 74% of the samples are from Europe, 15% of the samples are from North America, and 7% of the samples are from Asia (Figure 1A). In terms of time, the sample starts from January 2020 to January 2021 (Figure 1B). The samples in Europe increase dramatically after July 2020, while samples in Asia and North America slightly increase from April 2020. The number of cases is very inconsistent with the fact that the United States in North America and India in Asia has the largest number of COVID-19 patients in their continents. Although the database seems to lack sufficient samples to reflect the real situation in America and Asia, there are still 22,599 sequences in America and 10,546 sequences in Asia, and the mutation frequency calculation should be precise.

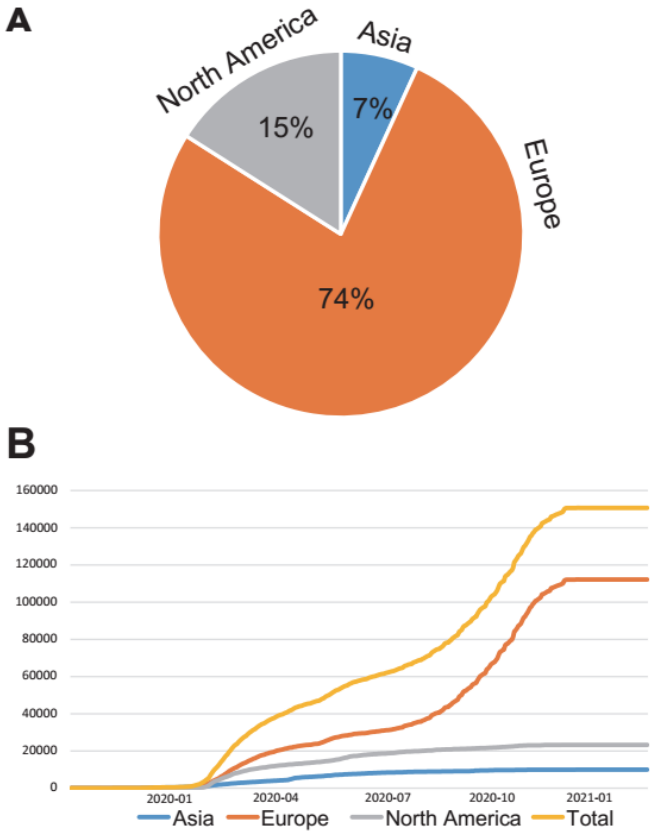

Figure 1. Distribution of downloaded SARA-Cov-2 sequence mutation data.

Single nucleotide polymorphisms (SNPs) in the SARS-CoV-2 genome

I focused on the single-site mutation (SNP) and calculated the frequencies of all mutation sites (Figure 2). The mutations are ubiquitous in most regions of the genome and the sites with the highest mutation frequencies are in Polyprotein (ORF1ab), S protein, ORF3a, M protein, ORF8, N protein, ORF10 (Figure 2).

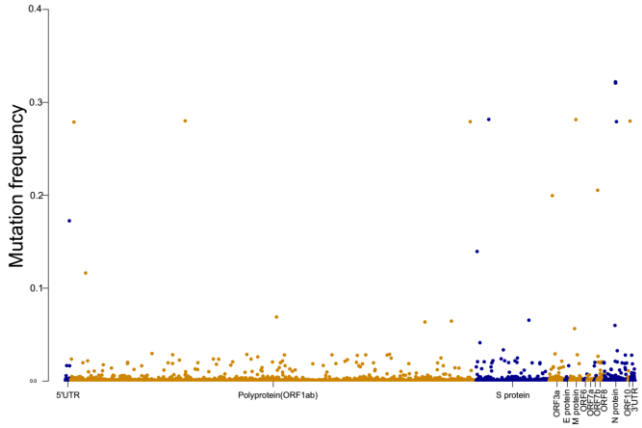

Figure 2. Distribution of SNPs on SARS-CoV-2 genome.

Different patterns in Asia, Europe, and North America

I selected the sites with top mutation frequencies in Asia, Europe, and North America. There are 27 sites in Asia, 30 sites in Europe, and 30 sites in North America (Figure 3). Only 13 sites are shared by all three regions suggesting that the mutation patterns may be different in three regions. I also calculated the mutation frequency changes with time (Figure S1). Three main clusters can be observed: 1) The mutation frequency gradually increases with time, and finally reaches a plateau. 2) The mutation frequency has a high peak at the beginning of the outbreak, and then the frequency decreases to a low level with time. 3) The mutation frequency has a high peak at the beginning of the outbreak, and then the frequency decreases with time but maintains a high level. Cluster 1 takes for a little more than 30% of each group. Cluster 2 takes about 30% of Asia and Europe, while more than two thirds of North America. Cluster 3 takes around 30% in Asia and 20% in Europe. It confirms the different patterns in Asia, Europe, and North America.

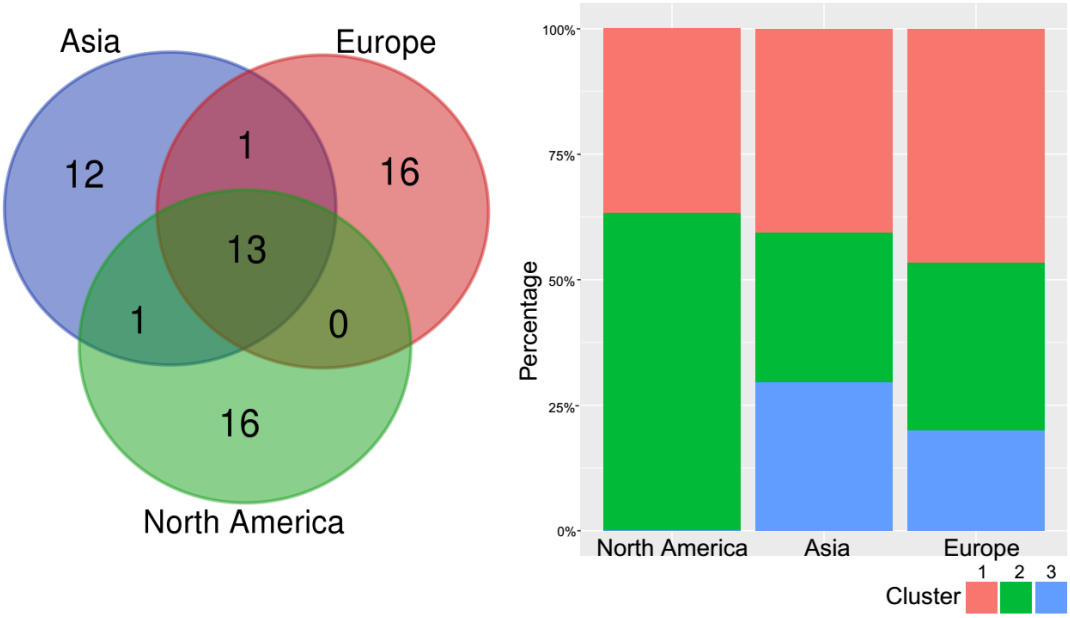

Figure 3. Sites with top mutation frequencies in Asia, Europe, and North America

Most of the mutation can lead to amino acid substitution

Non-synonymous and synonymous mutation can lead to different results. I can use the change of codon to see whether SNP can cause synonymous or non-synonymous mutations. Most of the SNPs in the SARS-CoV-2 genome can contribute to amino acid substitution (Table 1). Three genes contain the most non-synonymous mutations, including ORF1ab, Spike protein, and N protein (Table 2). Spike protein mediates host cell receptor recognition and binding and is key for vaccine design and development against SARS-CoV-2 infection [7]. The enrichment of SNPs in the Spike protein gene could lead to the subsequent evolution of the virus. Three regions display different mutation directions which make the prediction of the mutation valuable (Table 2).

Table 1. Most mutated sites in the SARS-CoV-2 genome.

Position	Gene	SNP	Codon change	Amino acid position	Amino acid change	Non-synonymous
313	ORF1ab	C -> T	CTC->CTT	16	L->L	no
445	ORF1ab	T -> C	GTT->GTC	60	V->V	no
490	ORF1ab	T -> A	GAT->GAA	75	D->E	yes
1059	ORF1ab	C -> T	ACC->ATC	265	T->I	yes
1397	ORF1ab	G -> A	GTA->ATA	378	V->I	yes
2416	ORF1ab	C -> T	TAC->TAT	717	Y->Y	no
2480	ORF1ab	A -> G	ATT->GTT	739	I->V	yes
2558	ORF1ab	C -> T	CCA->TCA	765	P->S	yes
3037	ORF1ab	C -> T	TTC->TTT	924	F->F	no
3177	ORF1ab	C -> T	CCT->CTT	971	P->L	yes
5572	ORF1ab	G -> T	ATG->ATT	1769	M->I	yes
6286	ORF1ab	C -> T	ACC->ACT	2007	T->T	no
6310	ORF1ab	C -> A	AGC->AGA	2015	S->R	yes
6312	ORF1ab	C -> A	ACA->AAA	2016	T->K	yes
6446	ORF1ab	G -> T	GTT->TTT	2061	V->F	yes
8782	ORF1ab	C -> T	AGC->AGT	2839	S->S	no
9891	ORF1ab	C -> T	GCT->GTT	3209	A->V	yes
11083	ORF1ab	G -> T	TTG->TTT	3606	L->F	yes
13730	ORF1ab	C -> T	CTA->TTA	4489	A->L	yes
14408	ORF1ab	C -> T	CTA->TTA	4715	P->L	yes
14805	ORF1ab	C -> T	ACT->ATT	4847	Y->I	yes
17747	ORF1ab	C -> T	CTG->TTG	5828	P->L	yes
17858	ORF1ab	A -> G	ATG->GTG	5865	Y->V	yes
18060	ORF1ab	C -> T	TCT->TTT	5932	L->F	yes
18877	ORF1ab	C -> T	GTC->GTT	6204	C->V	yes
19524	ORF1ab	C -> T	TCG->TTG	6420	L->L	no
20268	ORF1ab	A -> G	TAG->TGG	6668	L->W	yes
21255	ORF1ab	G -> C	CGT->CCT	6997	A->P	yes
21614	Spike protein	C -> T	CTT->TTT	18	L->F	yes
21707	Spike protein	C -> T	CAT->TAT	49	H->Y	yes
22227	Spike protein	C -> T	GCT->GTT	222	A->V	yes
23403	Spike protein	A -> G	GAT->GGT	614	D->G	yes
23929	Spike protein	C -> T	TAC->TAT	789	Y->Y	no
24034	Spike protein	C -> T	AAC->AAT	824	N->N	no
25563	ORF3a	G -> T	CAG->CAT	57	Q->H	yes

26144	ORF3a	G -> T	GGT->GTT	251	G->V	yes
26729	M protein	T -> C	GCT->GCC	69	A->A	no
26735	M protein	C -> T	TAC->TAT	71	Y->Y	no
26801	M protein	C -> G	CTC->CTG	93	L->L	no
27046	M protein	C -> T	ACG->ATG	175	T->M	yes
27944	ORF8	C -> T	CAC->CAT	17	H->H	no
27964	ORF8	C -> T	TCA->TTA	24	S->L	yes
28077	ORF8	G -> C	GTG->CTG	62	V->L	yes
28144	ORF8	T -> C	TTA->TCA	84	L->S	yes
28253	ORF8	C -> T	TTC->TTT	120	F->F	no
28311	N protein	C -> T	CCC->CTC	13	P->L	yes
28688	N protein	T -> C	TTG->CTG	139	L->L	no
28854	N protein	C -> T	TCA->TTA	194	S->L	yes
28881	N protein	G -> A	AGG->AAA	203	R->K	yes
28882	N protein	G -> A	AGG->AAA	203	R->K	yes
28883	N protein	G -> C	GGA->CGA	204	G->R	yes
28932	N protein	C -> T	GCT->GTT	220	A->V	yes
29095	N protein	C -> T	TTC->TTT	274	F->F	no
29645	ORF10	G -> T	GTA->TTA	30	V->L	yes

Table 2. Non-synonymous mutations enriched in three gene loci.

Gene	Asia	Europe	America
ORF1ab	9	9	11
Spike protein	1	3	2
N protein	4	4	4
Cluster (1/2/3)	7/2/5	6/5/5	7/10/0

Model of predicting the mutation frequency

Based on a regression method, I developed a model to predict the frequency change with time in a short period (Figure S1). The prediction period is from November 1, 2020, to January 11, 2021. The MSE values represent the accuracy and stability of the model (Table 3). The model proposes the mutation trends of each site in three regions. The method considers multiple parameters, such as disease outbreak region, and predicted the mutation frequency for specific sites. It could be a reference for researchers on different continents. The mutations in Spike protein may affect the off-target effect of the vaccine. Based on the similarity of genomes and the prediction of mutation trends, I hope that our work can provide an alternative reference for residents to choose vaccines produced in different brands and different countries. Overall, I expect our approach to be a fundamental solution in the literature and contribute as reliable quantitative benchmarks.

Table 3. MSE of the prediction model.

MSE	Average ($-\log_{10}$)	SD
Asia	6.97	0.95
Europe	4.62	1.50
America	6.93	0.86

Future Work

In the future, I would like to improve our prediction model so that the model can automatically make predictions of short-term frequency changes based on database updates. I shall focus on the Spike protein which is the most important region in the SARS-CoV-2 genome related to human immune response (Figure 4) [8]. Additionally, I will develop a website to visualize our results. The website will contain a variety of functions, including the number and distribution of samples from the database, the mutation status of important sites in each continent, and the prediction of mutations in important sites, and the impact of these mutations on the function of S protein.

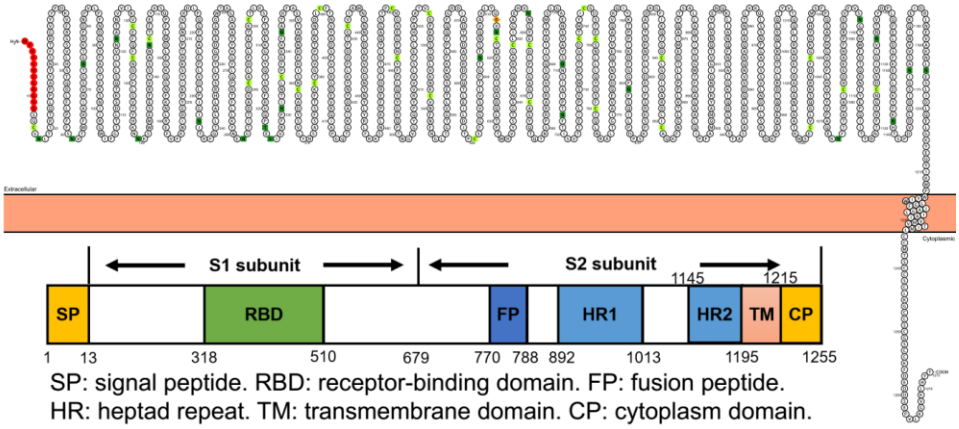

Figure 4. Structure of Spike protein in SARS-CoV-2 genome.

Reference

1 Giorgi, G. *et al.* COVID-19-Related Mental Health Effects in the Workplace: A Narrative Review. *Int J Environ Res Public Health* **17**, doi:10.3390/ijerph17217857 (2020).

2 Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292 e286, doi:10.1016/j.cell.2020.02.058 (2020).

3 Abdullahi, I. N. *et al.* Implications of SARS-CoV-2 genetic diversity and mutations on pathogenicity of the COVID-19 and biomedical interventions. *J Taibah Univ Med Sci* **15**, 258-264, doi:10.1016/j.jtumed.2020.06.005 (2020).

4 Makizako, H. *et al.* Exercise and Horticultural Programs for Older Adults with Depressive Symptoms and Memory Problems: A Randomized Controlled Trial. *J Clin Med* **9**, doi:10.3390/jcm9010099 (2019).

5 Gong, Z. *et al.* An online coronavirus analysis platform from the National Genomics Data

Center. *Zool Res* **41**, 705-708, doi:10.24272/j.issn.2095-8137.2020.065 (2020).

6 Wu, F. *et al.* Author Correction: A new coronavirus associated with human respiratory disease in China. *Nature* **580**, E7, doi:10.1038/s41586-020-2202-3 (2020).

7 Shang, J. *et al.* Cell entry mechanisms of SARS-CoV-2. *Proc Natl Acad Sci U S A* **117**, 11727-11734, doi:10.1073/pnas.2003138117 (2020).

8 Sternberg, A. & Naujokat, C. Structural features of coronavirus SARS-CoV-2 spike protein: Targets for vaccination. *Life Sci* **257**, 118056, doi:10.1016/j.lfs.2020.118056 (2020).

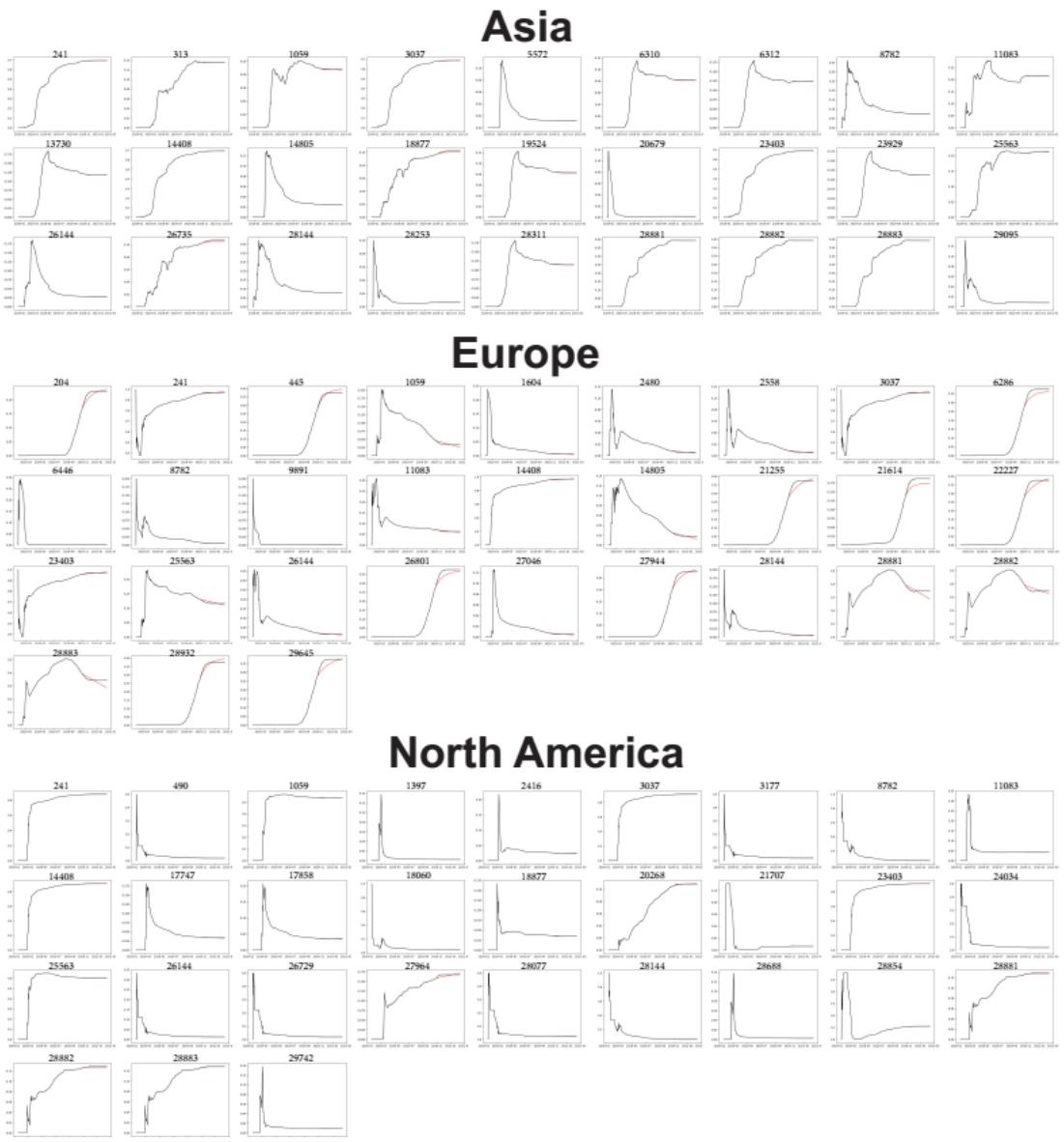

Figure S1. Frequencies and prediction of top mutated sites in Asia, Europe, and North America.