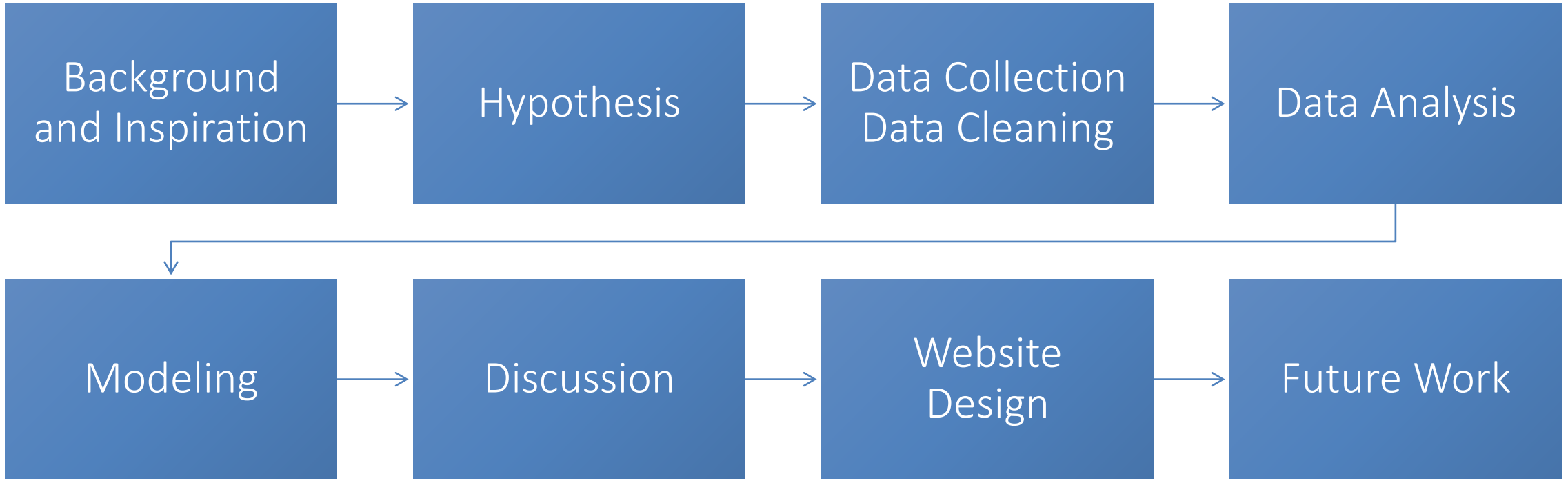
Modeling and visualizing the SARS-CoV-2 mutation based on geographical regions and time

Bomin (David) Wei

Princeton International School of Math and Science

March 2021

Outline


Background and Inspiration
Hypothesis
Data Collection Data Cleaning
Data Analysis
Modeling
Discussion
Website Design
Future Work

Background and Inspiration

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.

The COVID-19 is caused by a novel evolutionary divergent RNA virus, called severe acute respiratory syndrome coronavirus 2 **(SARS-CoV-2)**

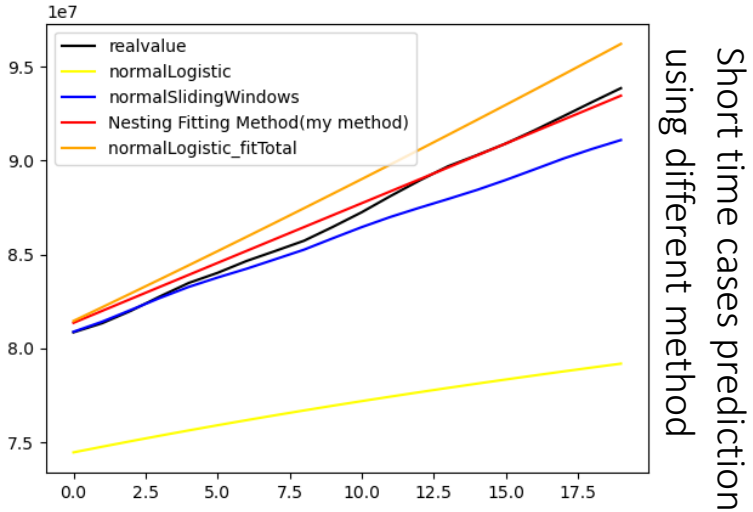
Schools had to close throughout the US.

I had to stay away from campus and friends, taking online classes.

Figure captured from https://coronavirus.jhu.edu/map.html

Interests and Inspiration

I'm a computer enthusiast, last summer, I compared many models to do the COVID-19 case prediction. I felt it was very interesting and effective.



Short time cases prediction using different method

Later, from the news, I learned that the virus had continued mutating. And this gave me an idea to use my experience in predicting the cases to model and predict the mutation trend of the COVID-19 virus, which could help scientists to estimate the effectiveness of the vaccine in the long term and the possibility of using one vaccine which was developed in one region in another region.

"Can I use my skills in CS and math modelling to predict the end of the pandemic?"

Virus mutations are important

Why does it mutate so quickly?

- The genetic information of SARS-CoV-2 mutates much more dramatically than DNA due to RNA viruses' mechanisms.
- The worldwide outbreak happens to provide good environments for SARS-CoV-2 mutations.

Why is it important?

- 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 (Makizako et al., 2019)
- Such as development of vaccine

Can I predict *mutations*?

Do all the genes have equal chances to mutate?

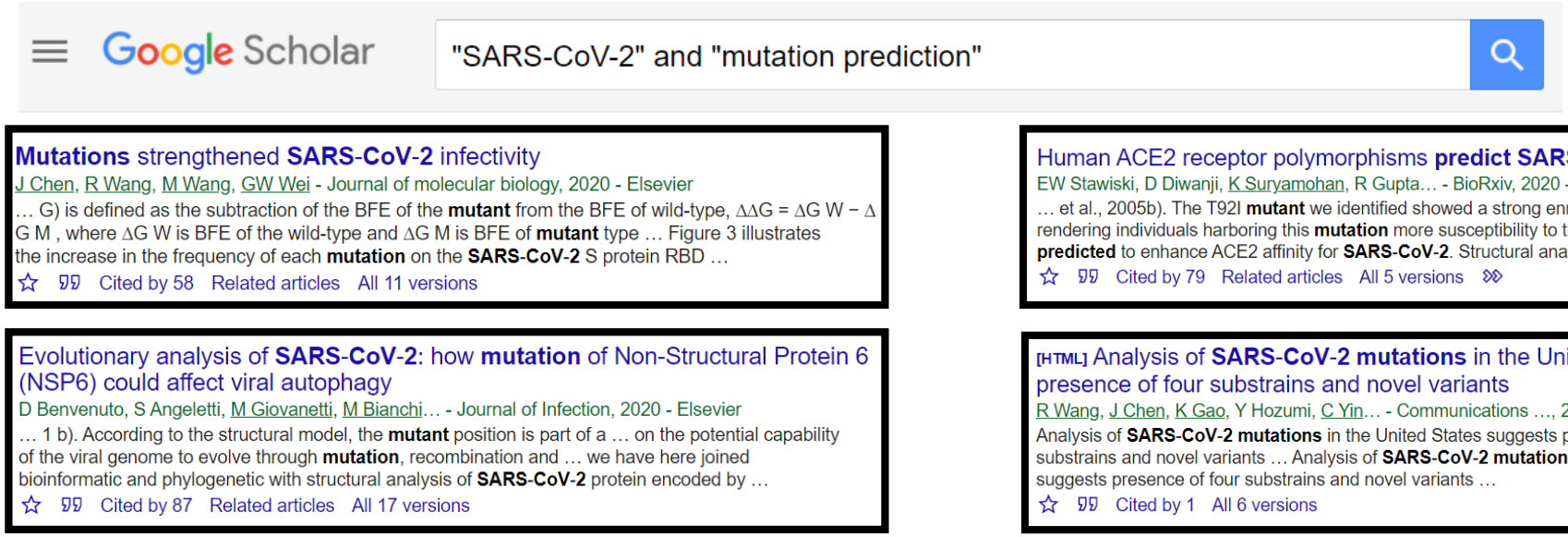
Do they mutate in a similar trend?

Do mutations vary from one continent to another?

Would any of the mutations really affect the function of the virus?

- and therefore, affect the vaccine development?
- or affect which vaccine we should choose when we travel back to school?

Literature search on mutation prediction of the virus



The majority of related research that combines with computer sciences focus on the **prediction of patient number**.

The majority of research on "SARS-CoV-2" and genome mutation are about
1. Distribution of mutations: what regions most mutations are found and
2. Influence on the functions of the virus.

Only a few research work done for "SARS-CoV-2" and "mutation".
A google scholar search gave us 47 results; and we didn't see a work on time series prediction on mutation chance.

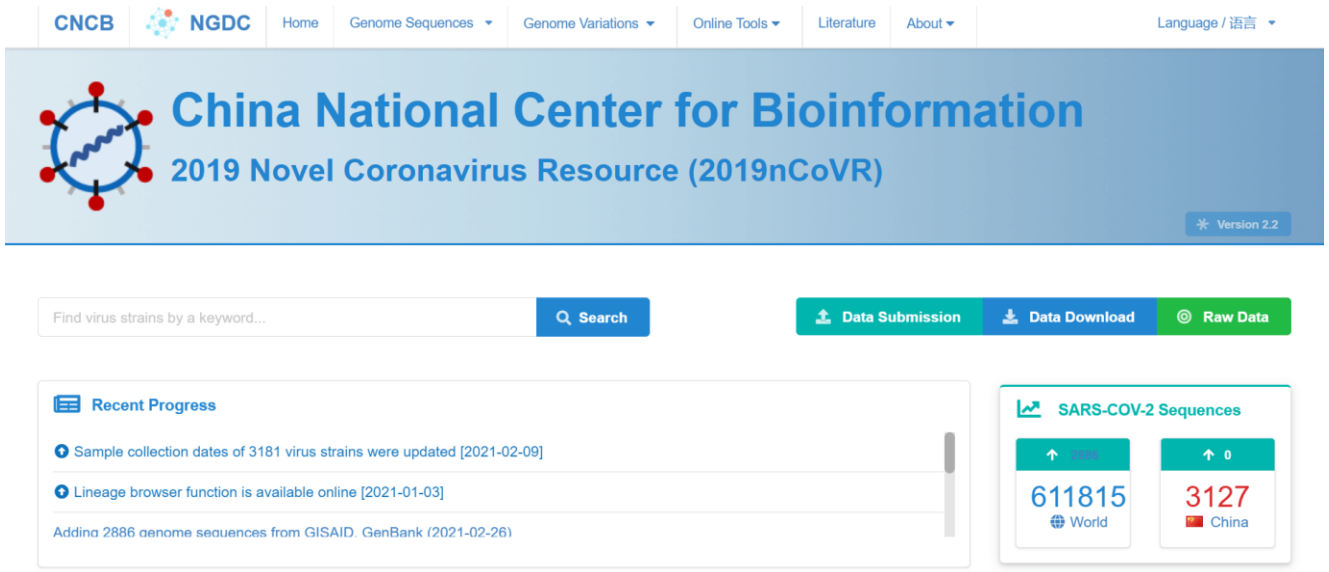
Hypothesis and Proposal

The mutation of the virus is highly related to the geographical location.

We plan to use ARIMA model to predict the trend of the mutation. The model parameters should also be different for different regions.

Through checking the synonymous and nonsynonymous mutation, we also should identify important mutations that affect the protein function of the SARS-CoV-2 virus.

We plan to build a website to show prediction model and the mutation trend.



Data Collection

Downloaded mutation data of sequences from a public database

China National Center for Bioinformation, 2019nCoVR

https://bigd.big.ac.cn/ncov/?lang=en

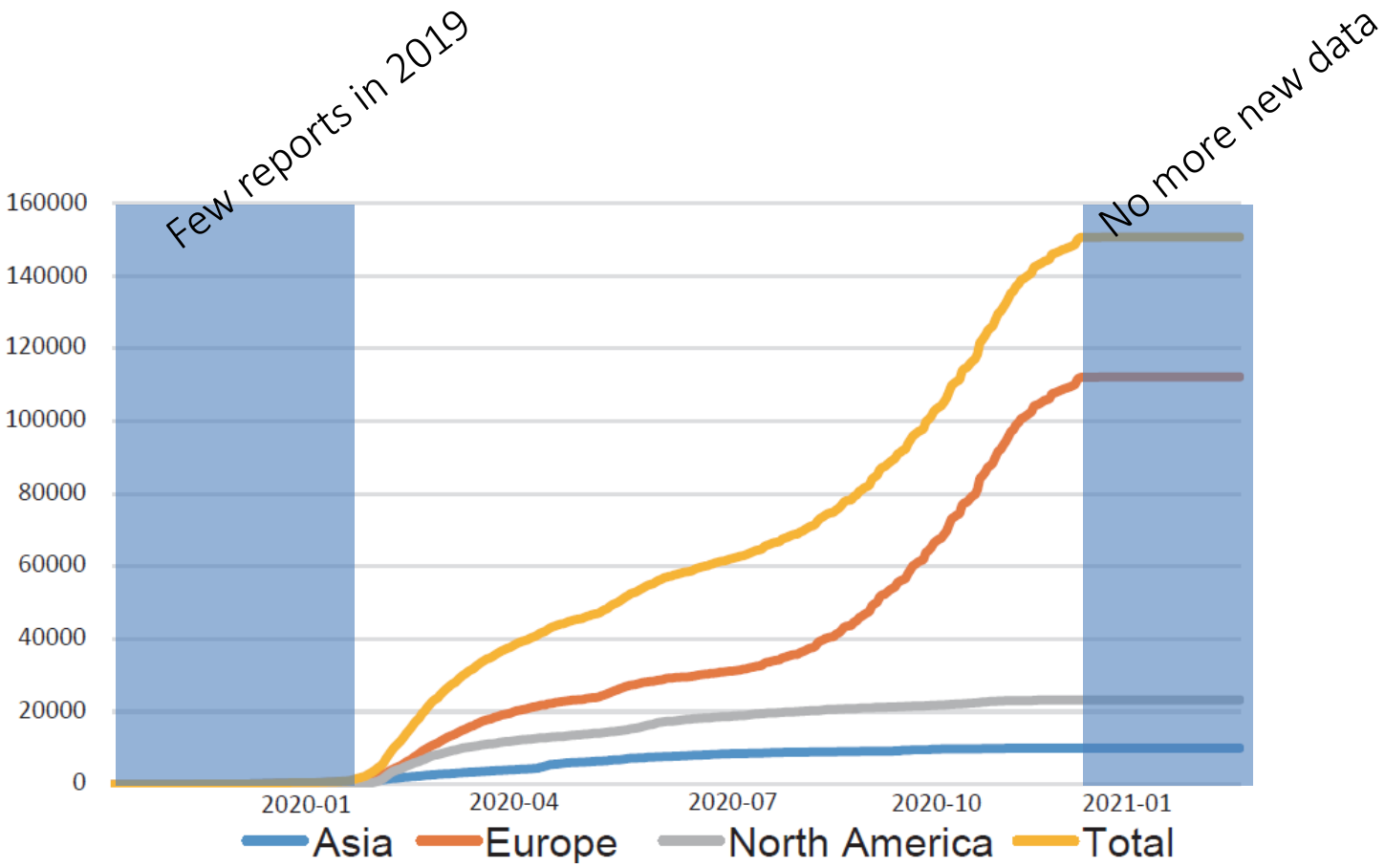
→ Already asked permission from Dr. Zhao, the database director.

US database: restricted access

528,611	• All sequence metadata (sampling location, time, host, age, and etc.) hosted on the CNBC database as of 2021/2
184,475	• **High quality data; Rule out back mutations; full sequence**
183,850	• Pair the raw sequence (i.e., atcg letter code) with corresponding metadata;
150,659	• Filter out incomplete metadata (e.g. only month w/o date in sampling time)

Automatic Data Cleaning

Data Analysis



With 22,000+ in North America and 10,000+ in Asia, the calculations and predictions would still be precise.

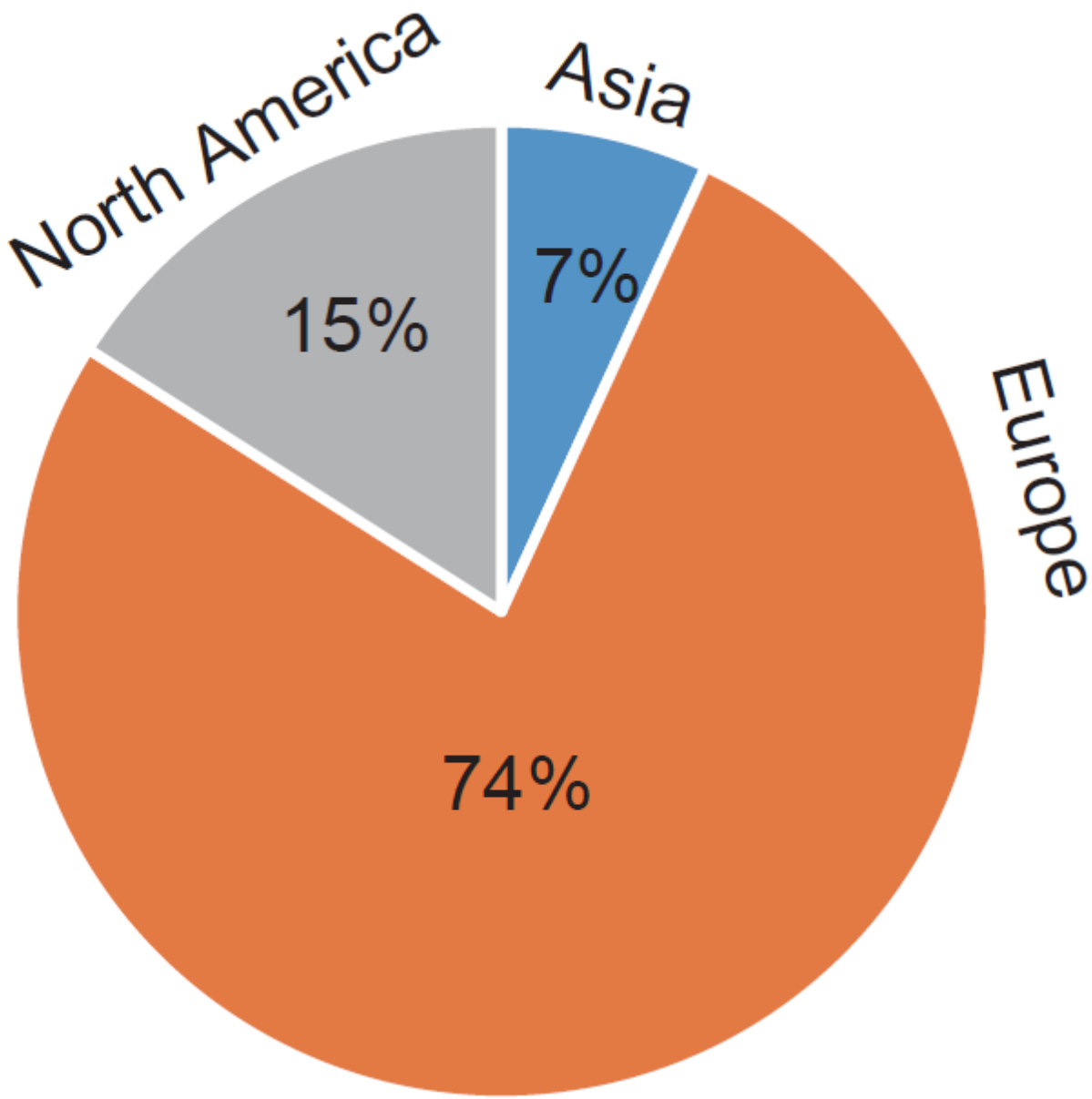
The statistic of mutation data

A total number of **150,659** data entries in the study from Jan. 2020 to Jan. 2021.

Europe sample number is consistent with its reported cases.

The number of cases is very **inconsistent** with the fact that the **United States** in North America and **India** in Asia have the largest number of COVID-19 patients in their continents.

Less reported cases
Many US reported cases requires extra authorization to be downloaded.



Overview of data distribution

Data distribution is mainly from Europe.

The sample sizes of Asia and North America are relatively small

- 22599 in North America
- 10546 in Asia
- The calculation of rate would be precise.

Modeling will be dependent on location, so there is no regional bias in our prediction.

A reference for mutation comparison

The reference sequence used in this study is NC_045512 with the length of 29,903 bp ss-RNA in National Center for Biotechnology Information (NCBI) (Wu et al., 2020).

Because the reference is always the same, sometimes the mutation rate at the beginning of a site in another region could be very high.

mutation frequency: converting letter code of gene info into time series

Mutation frequency at t (date) on a specific site is defined as:

- Total count of **mutation events to date** divided by **total count of sample**
- **Only focus on the single site mutation (SNP)**

The mutation rate at any time is ***cumulative*** in time

Counting of mutation events for a specific site includes all possible directions

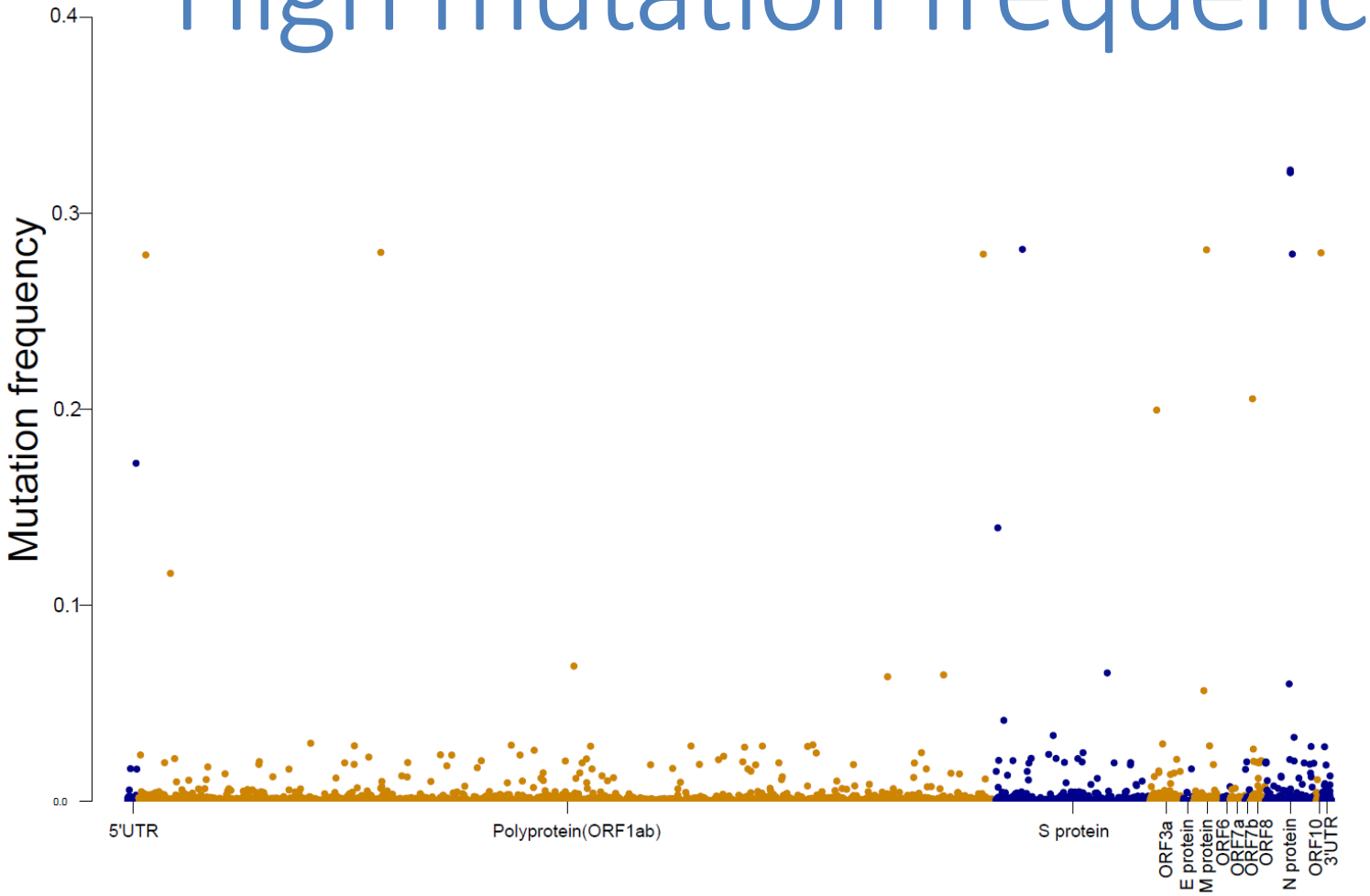
- It **doesn't affect** the modeling significantly
- Most mutations are in one direction.
- e.g., a few multi-directional mutations is Asia

241C	T	15264
241C	A	3
241C	G	4

6310C	T	11
6310C	A	764
6310C	G	1

28882G	A	11920
28882G	T	1
28882G	C	1

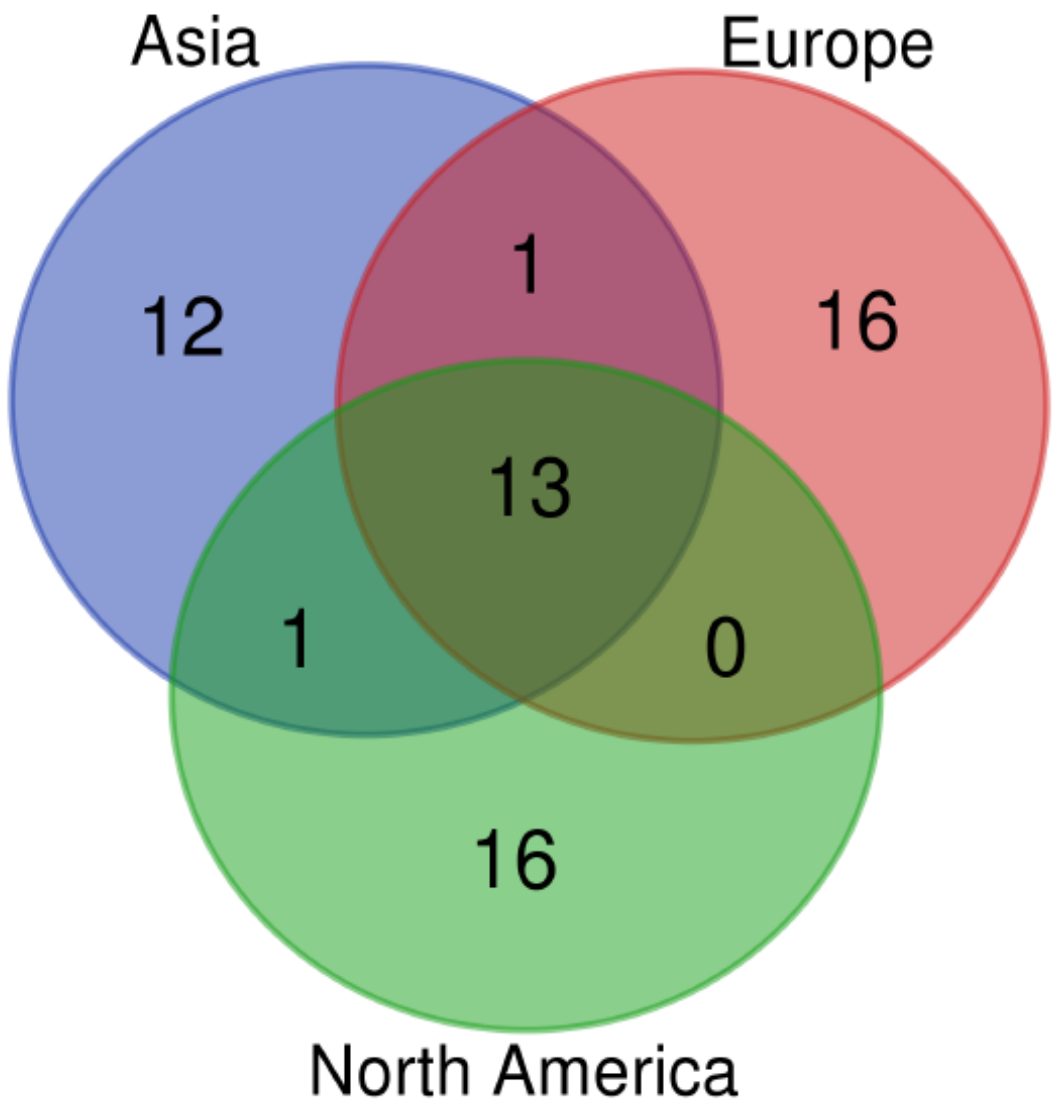
High mutation frequency sites



We calculated the frequencies of all mutation sites (the mutation of site which mutation is larger than 0.5 will be considered as ‘mutation = 1 – mutation’) .

The sites with highest mutation frequencies are in

Polyprotein (ORF1ab), S protein, ORF3a, M protein, ORF8, N protein, ORF10



High frequency mutations are different in three major continents

Among top mutation sites (minimum > 0.1% && average > 10%)

27 sites in Asia

30 sites in Europe

30 sites in North

Only 13 sites are shared by all three regions

This suggests that the mutation patterns may be different in three regions.

Data Modeling

What is ARIMA model?

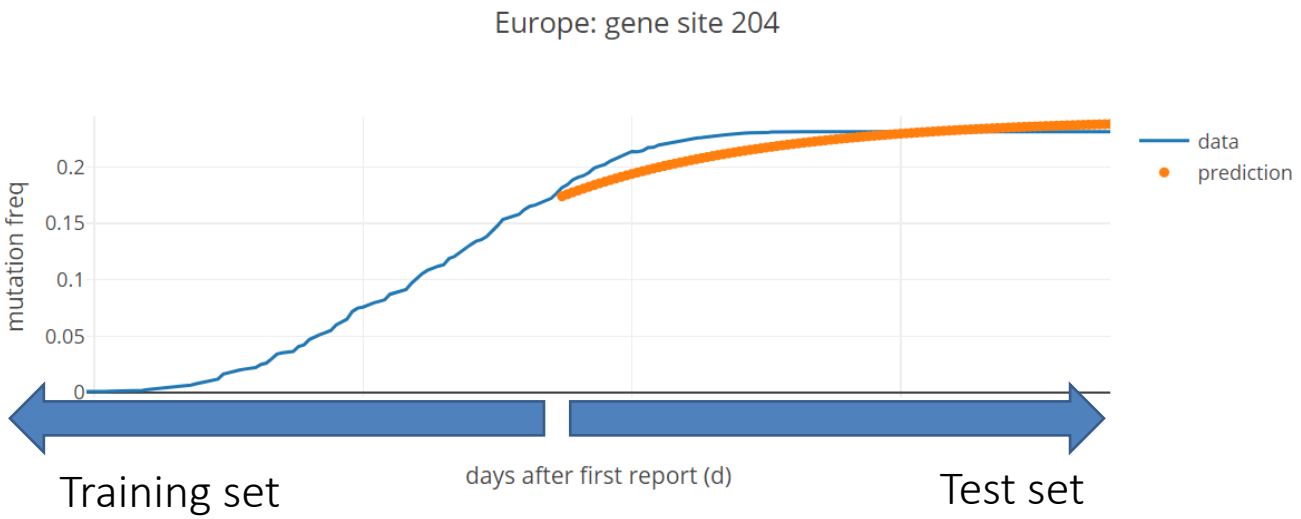
Auto**r**egressive **I**ntegration **M**oving **A**verage

Three parameters – p, d, and q

- p: the number of lag observations in the auto-regression (AR) part of the model, indicating relation between an observation (or data) to the past observations.
- q: 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.
- d: the integration order of the I part of the model, indicating the number of times that the raw observations are differenced.

If we let y be the d^{th} difference of Y (the observation or data), then the model can be expressed

$$\widehat{y_t} = \mu + \phi_1 y_{t-1} + \cdots + \phi_p\, y_{t-p} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}$$



Educated initial guess of p is first determined from ACF (auto-correlation function) of a mutation frequency time series.

Parameter fitting and data prediction

Training set data:
from 2019/12 to 2020/11

Test set data:
from 2020/12 to 2021/2

Automatic parameter search ensures good prediction quality

Each site should have its model parameters fitted independently.

I developed a program for automatic parameter scan and performed the search for about 90 groups of data.

For each site, the program tests 5 different parameters for p and q, respectively, and 1 or 2 for d,

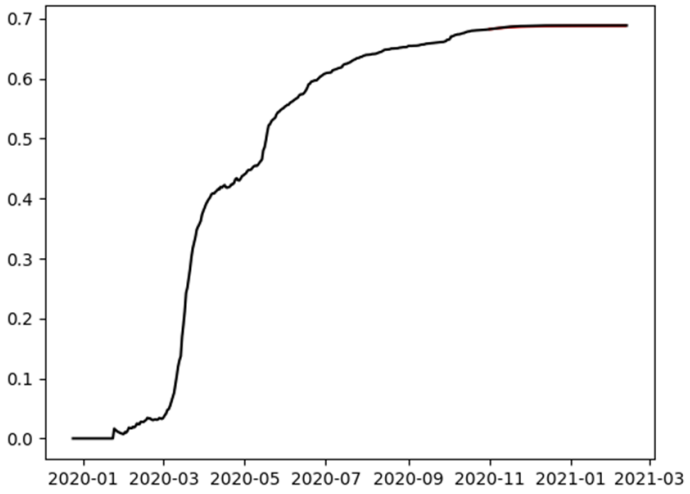
i.e, 50 parameter combinations are tested for each site.

In total, about 1500 automatic model fitting jobs.

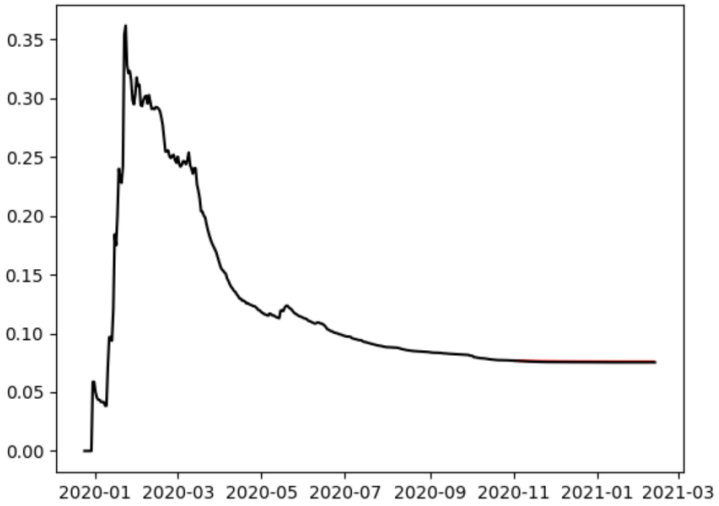
The best set of parameters (p, d, q) which has the lowest MSE will be automatically picked.

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

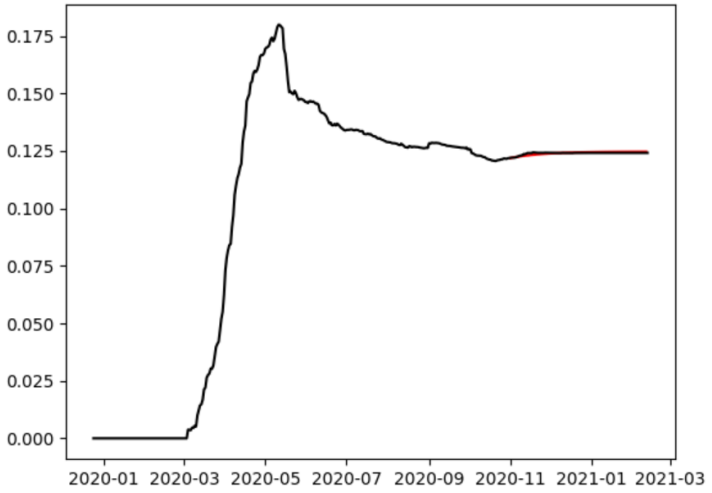
Asia; Genome site 23403



Asia; Genome site 8782

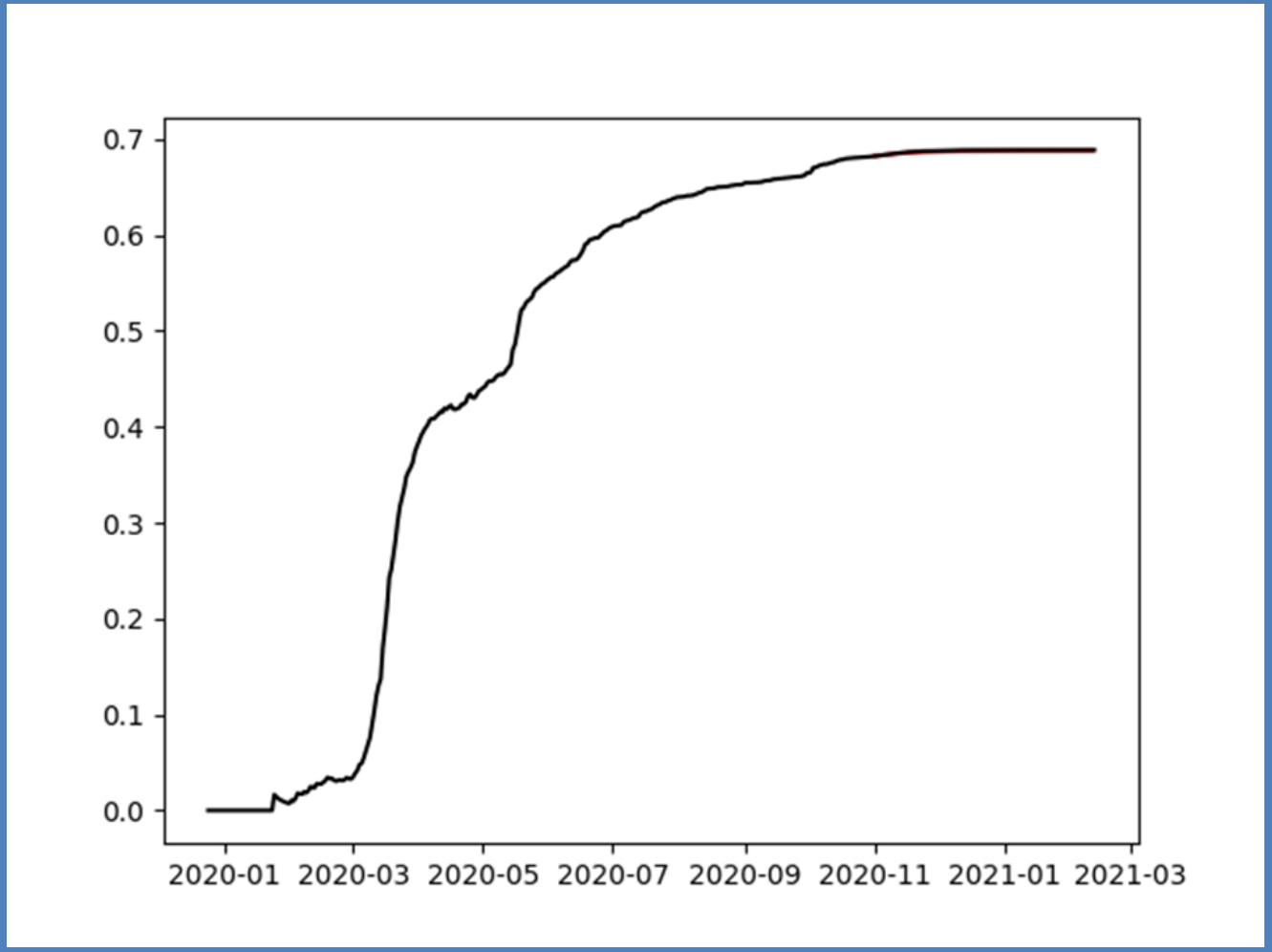


Asia; Genome site 6312



Mutation frequency time series show different features at different sites.

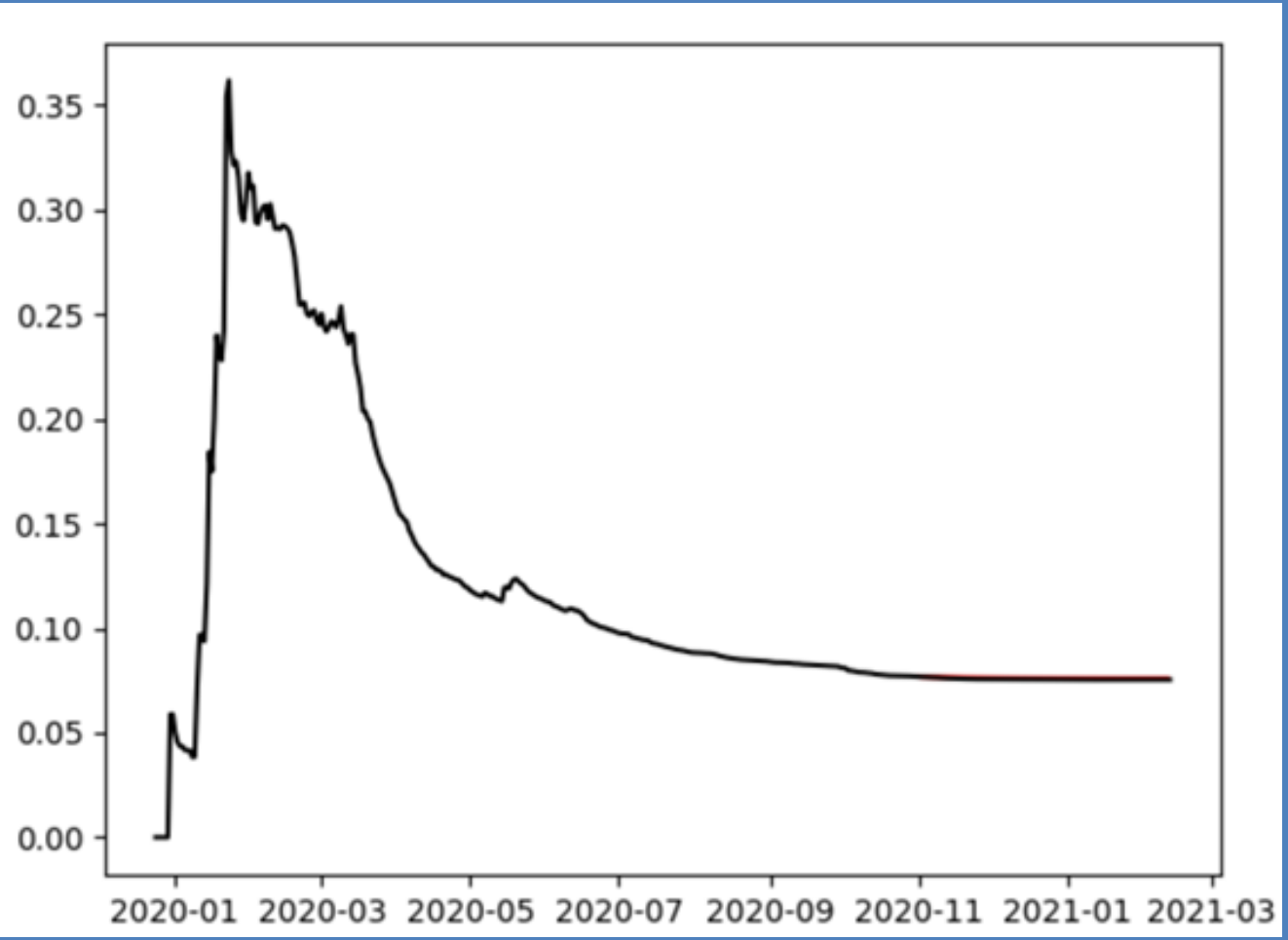
Example: Asia; Genome site 23403



Cluster 1:

The mutation frequency gradually increases with time, and finally reaches a plateau.

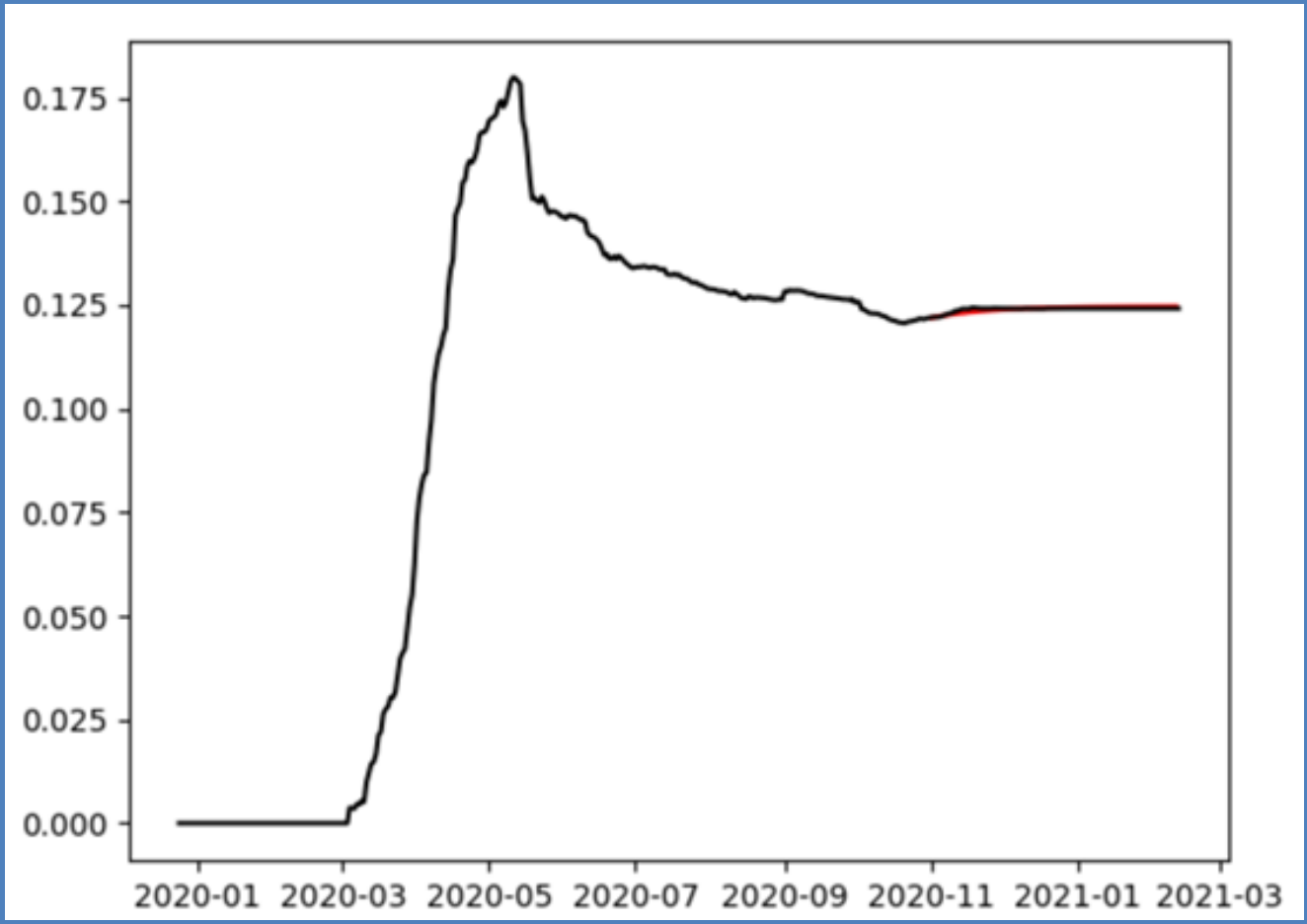
Asia; Genome site 8782



Cluster 2:

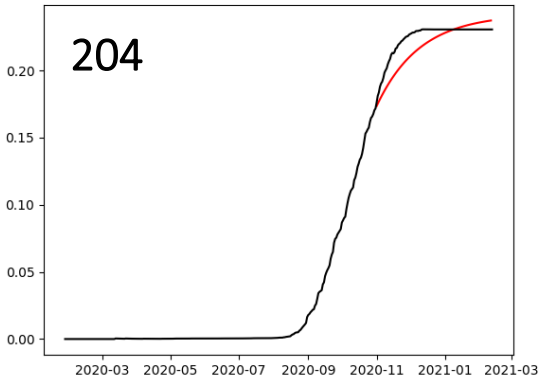
The mutation frequency has a high peak at the beginning of the outbreak, and then the frequency decreases to a low level with time.

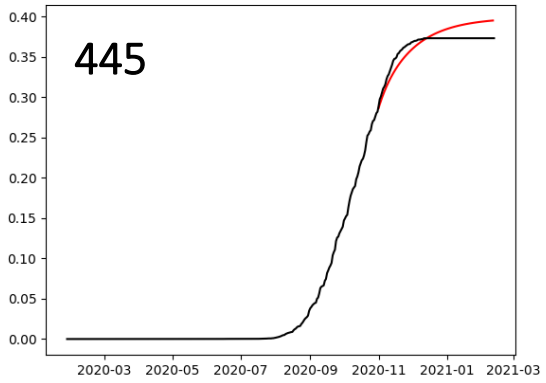
Example: Asia; Genome site 6312

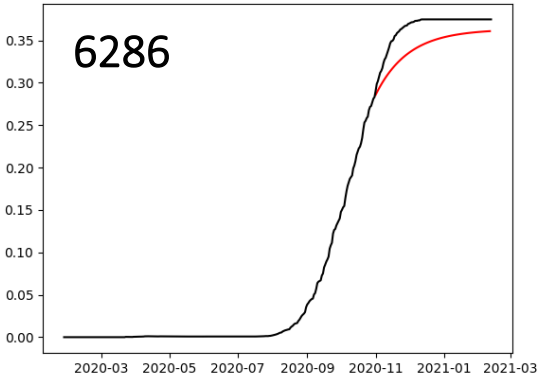


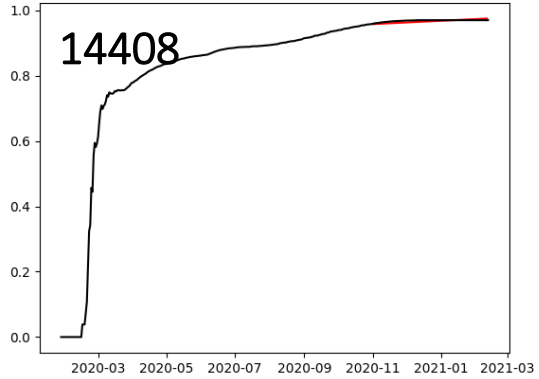
Cluster 3:

The mutation frequency has a high peak at the beginning of the outbreak, and then the decreases with time but maintains a high level.

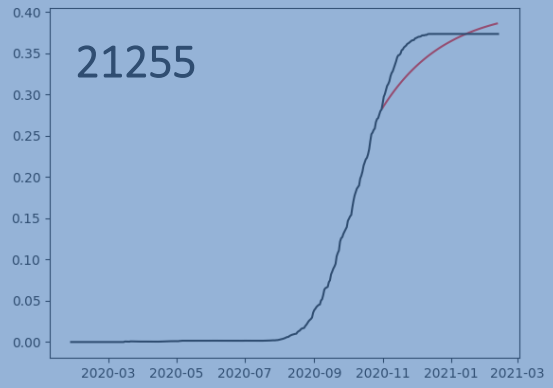


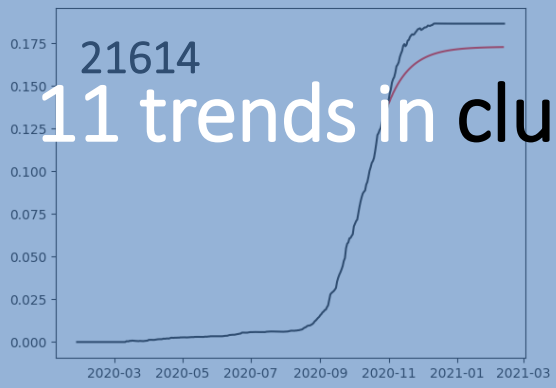


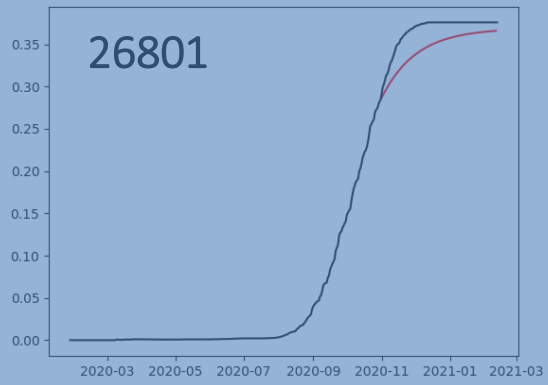


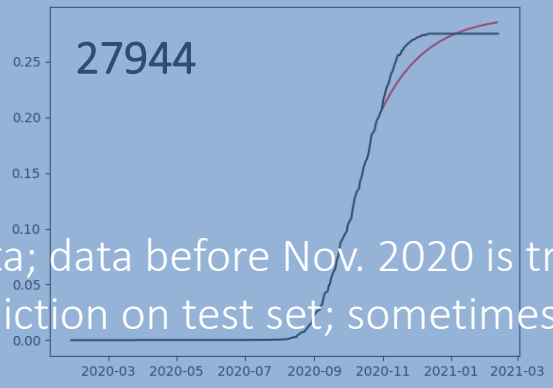


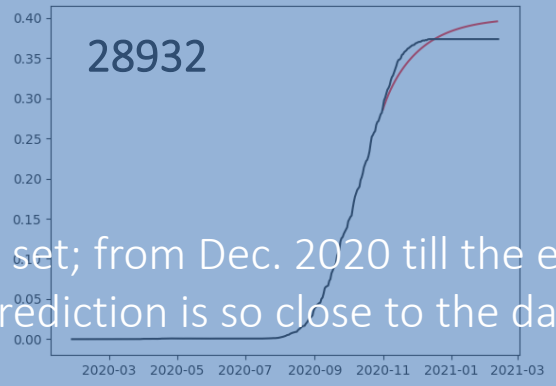
11 trends in cluster 1 in Europe





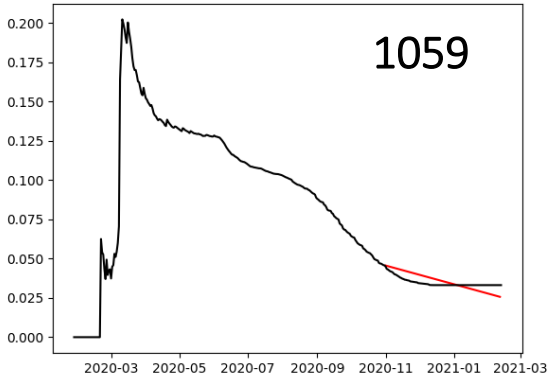


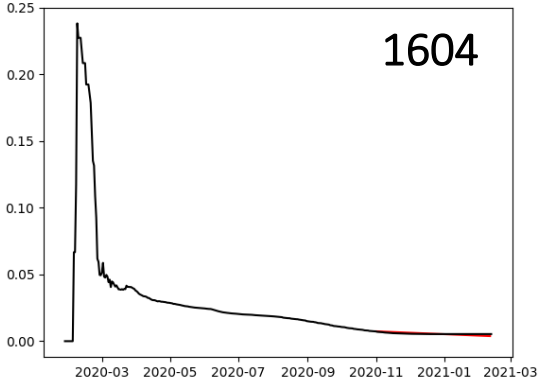


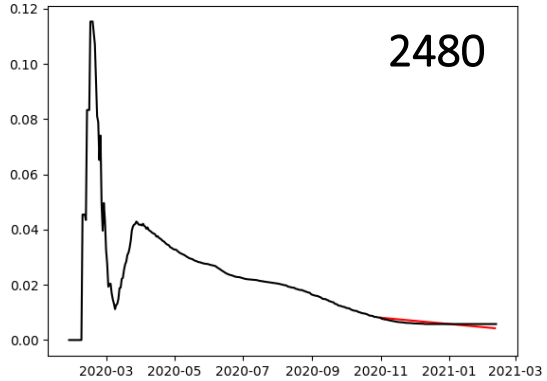


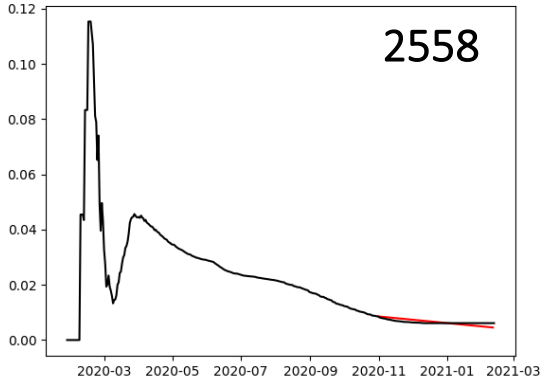
Black curve: data; data before Nov. 2020 is training set; from Dec. 2020 till the end is test set

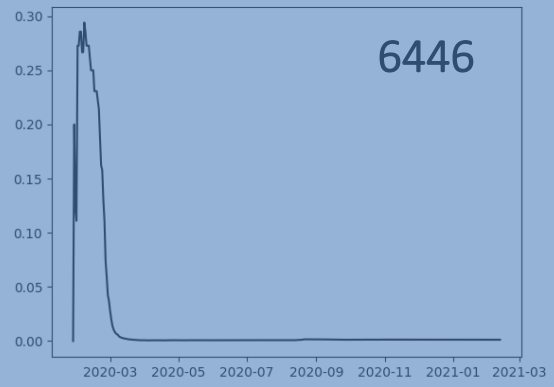
Red curve: prediction on test set; sometimes the prediction is so close to the data that you can not see very well.

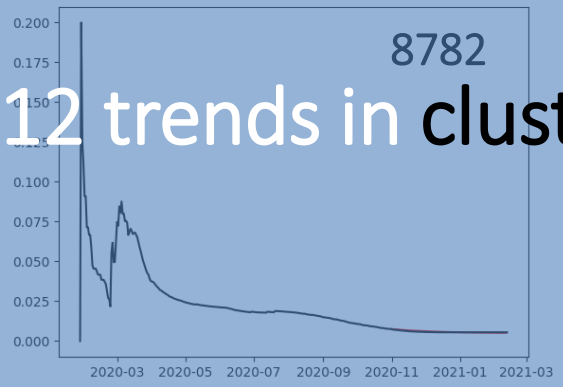


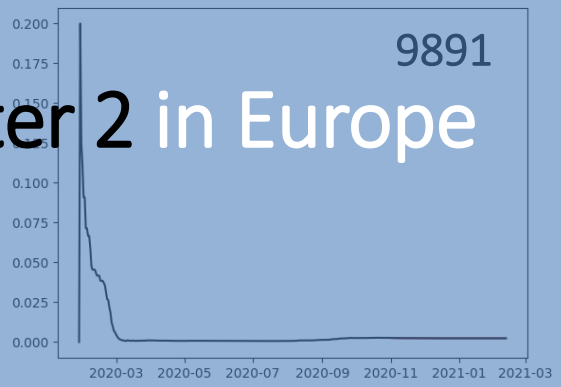


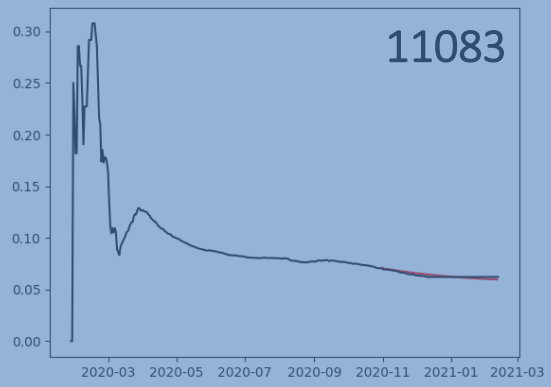




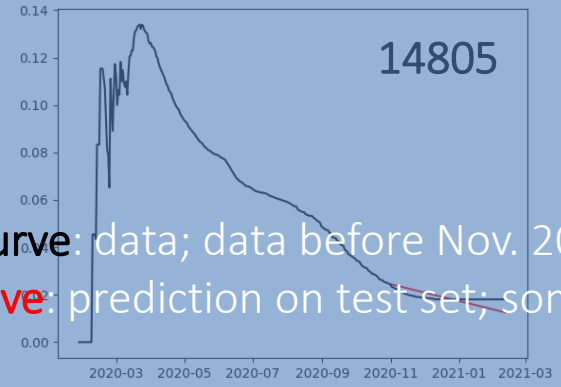


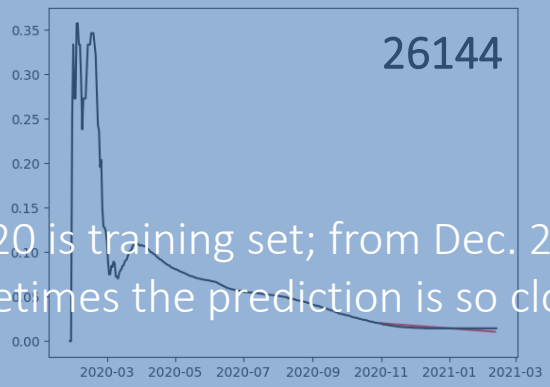


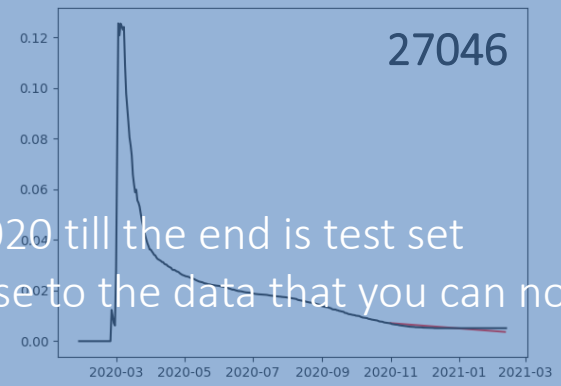


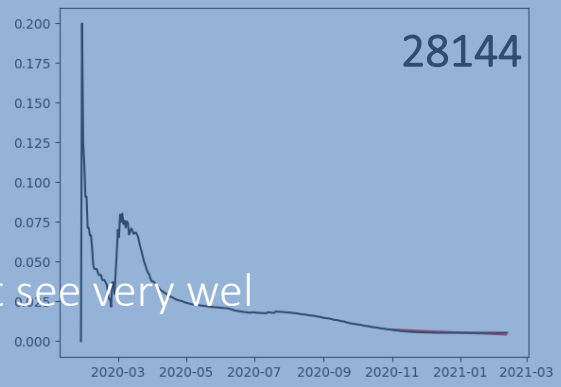


12 trends in cluster 2 in Europe





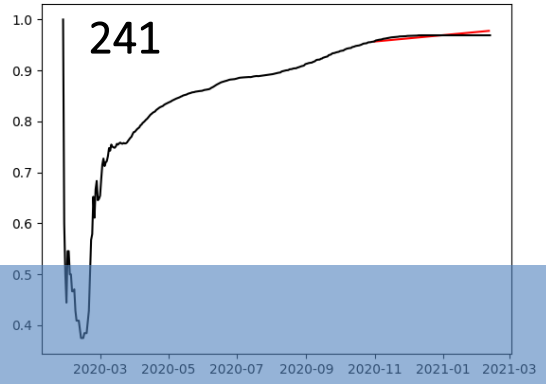


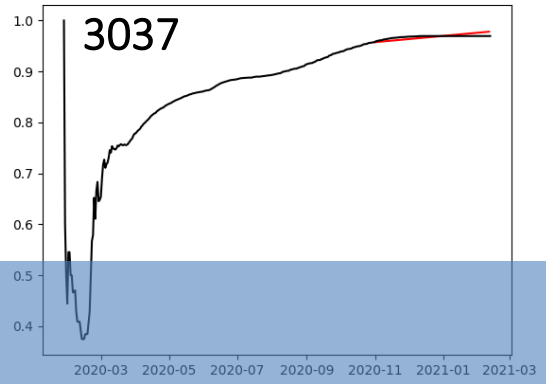


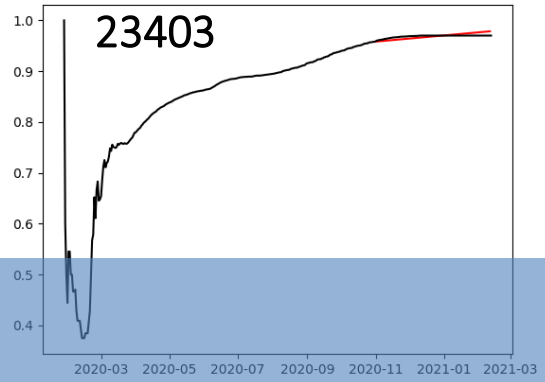
Black curve: data; data before Nov. 2020 is training set; from Dec. 2020 till the end is test set

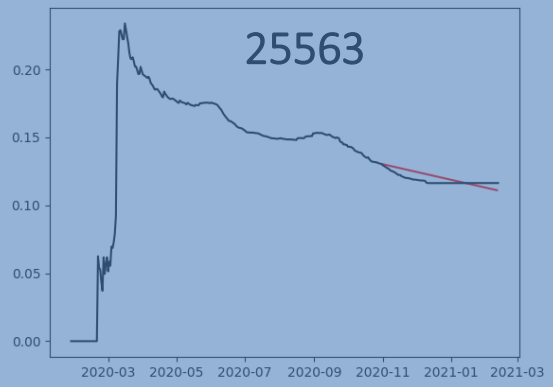
Red curve: prediction on test set; sometimes the prediction is so close to the data that you can not see very wel

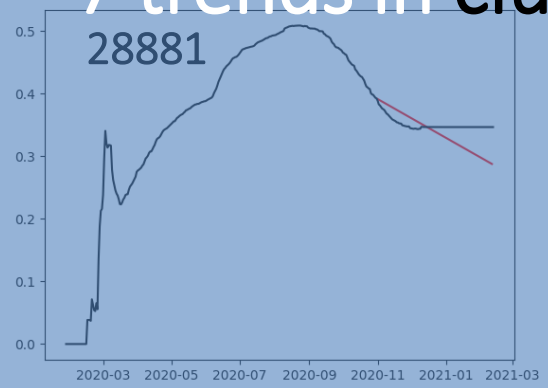
7 trends in cluster 3 in Europe





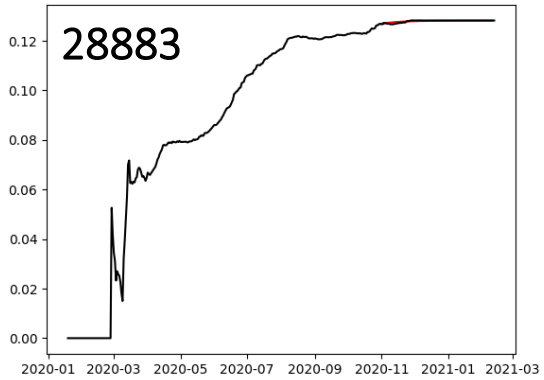


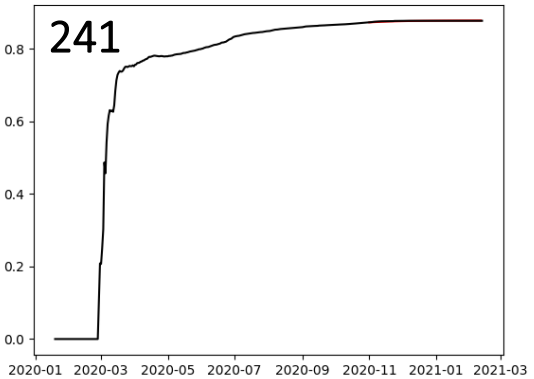


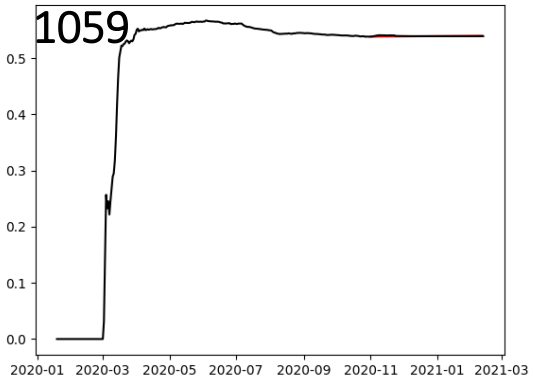


Black curve: data; data before Nov. 2020 is training set; from Dec. 2020 till the end is test set

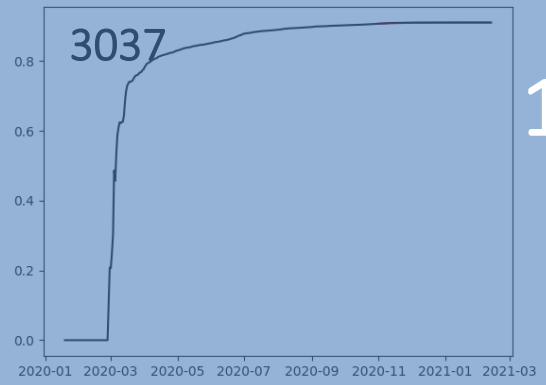
Red curve: prediction on test set; sometimes the prediction is so close to the data that you can not see very wel

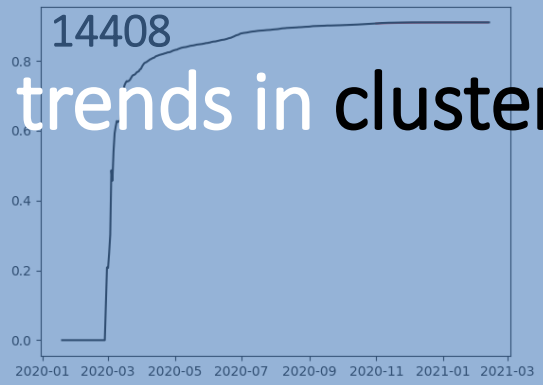


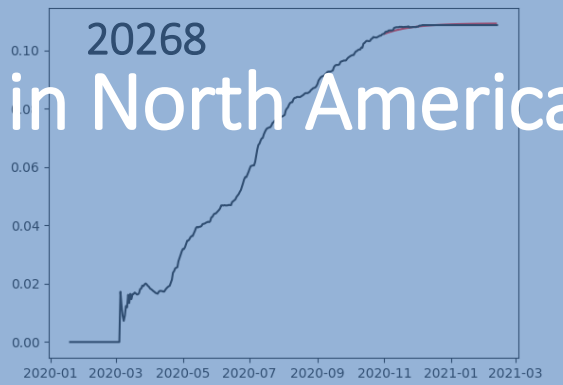


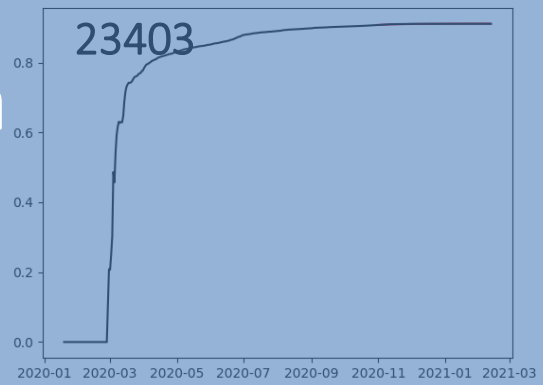


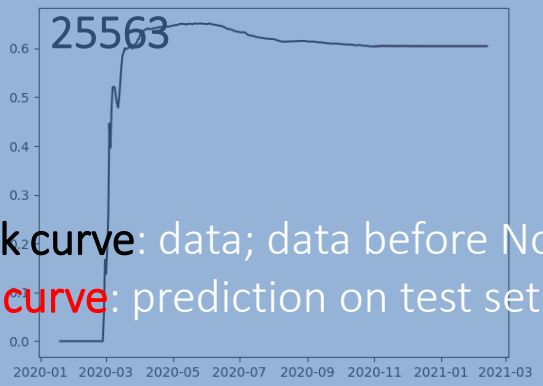
11 trends in cluster 1 in North America

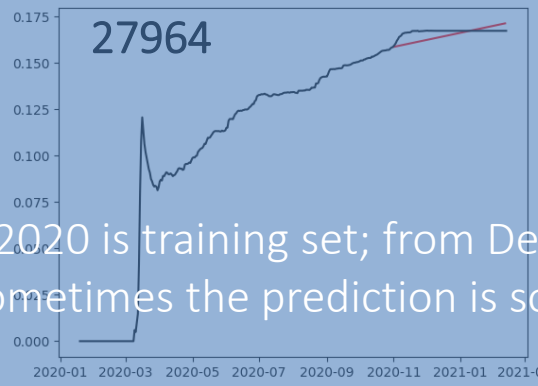


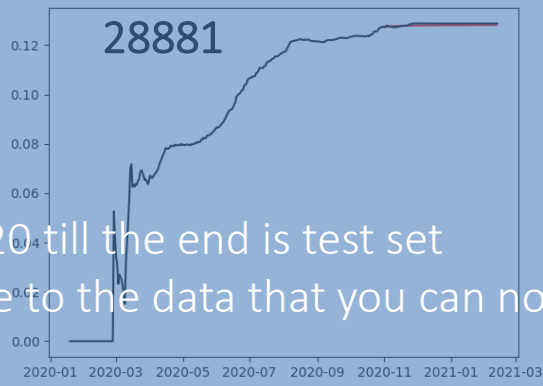






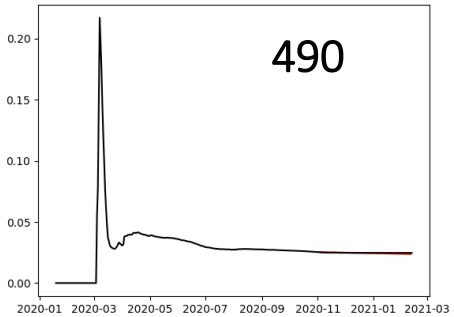


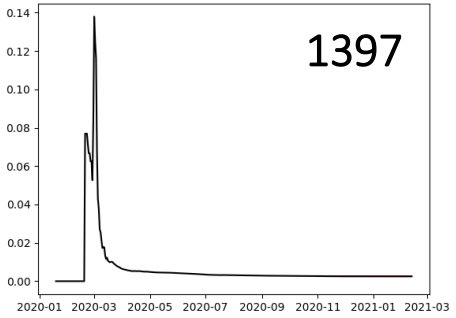


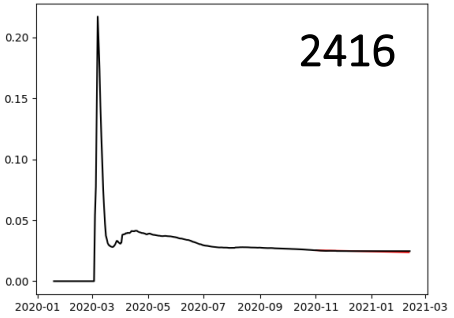


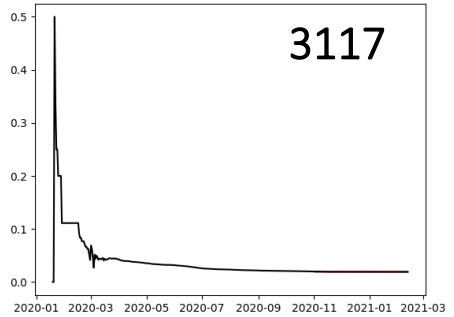
Black curve: data; data before Nov. 2020 is training set; from Dec. 2020 till the end is test set

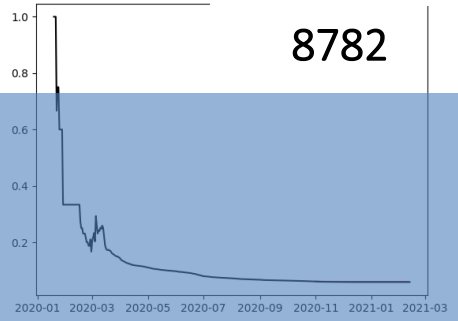
Red curve: prediction on test set; sometimes the prediction is so close to the data that you can not see very wel

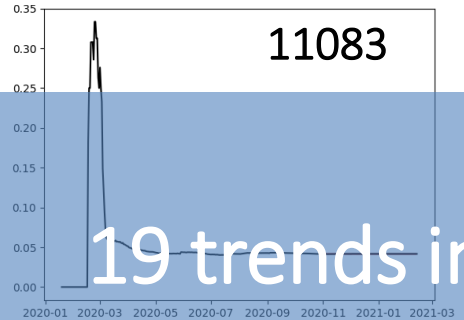


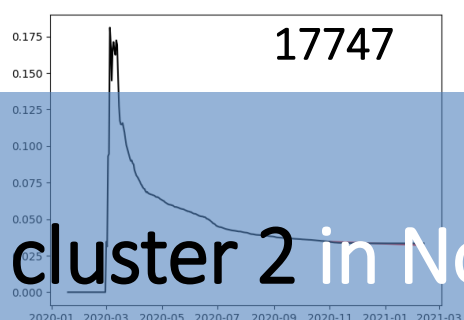


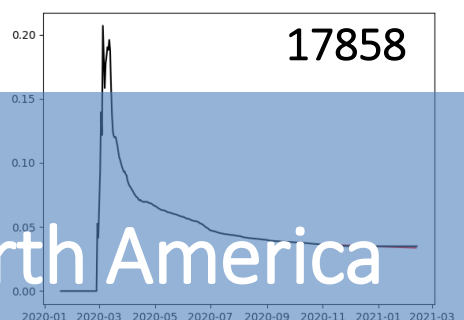


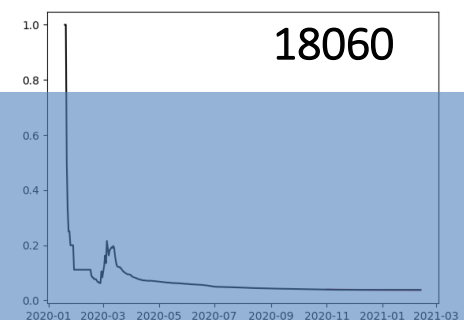




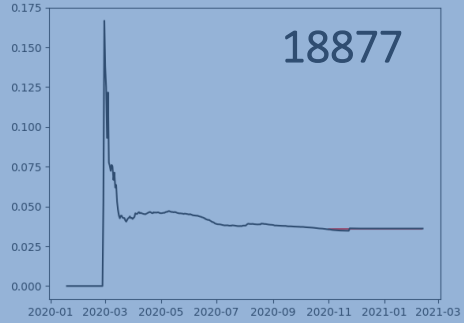


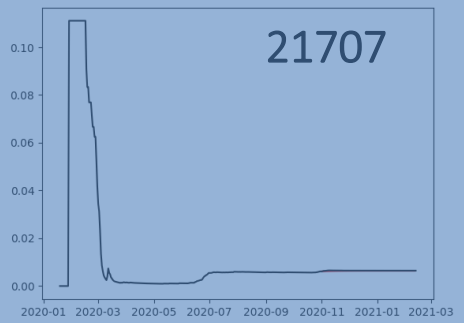


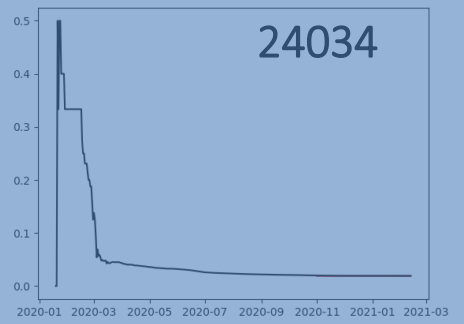


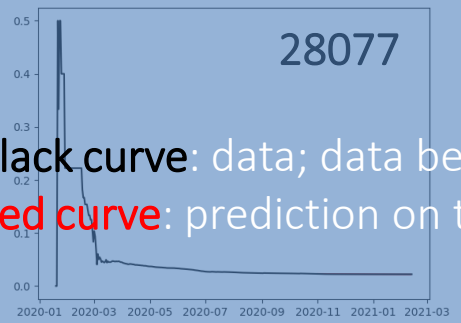


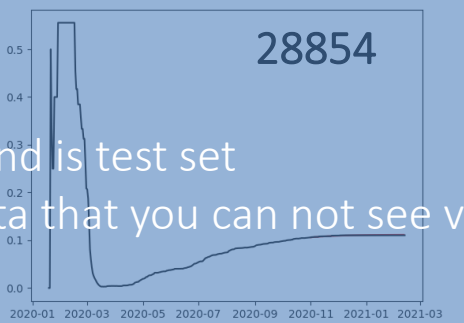
19 trends in cluster 2 in North America





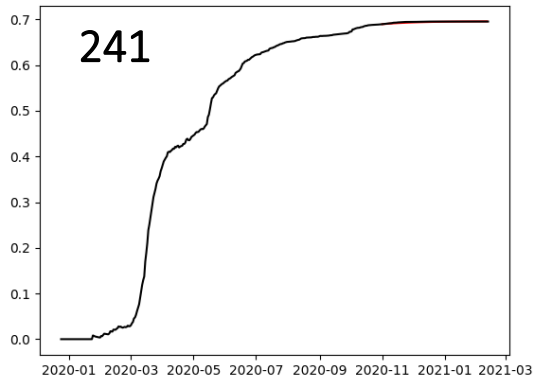


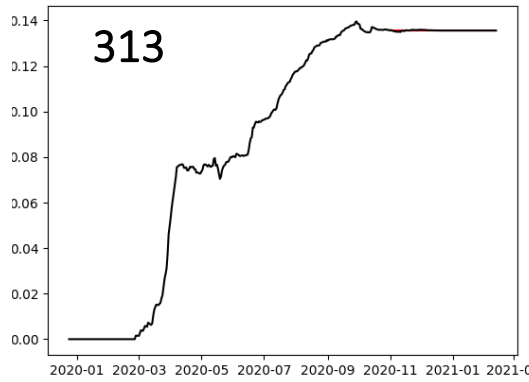


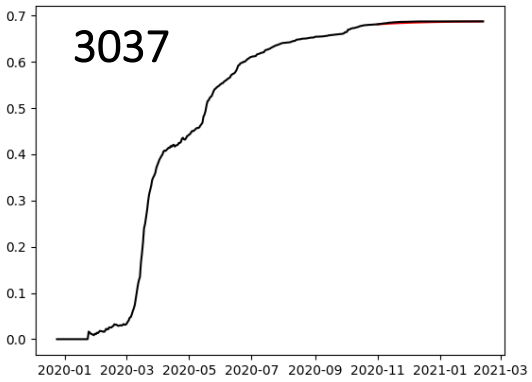


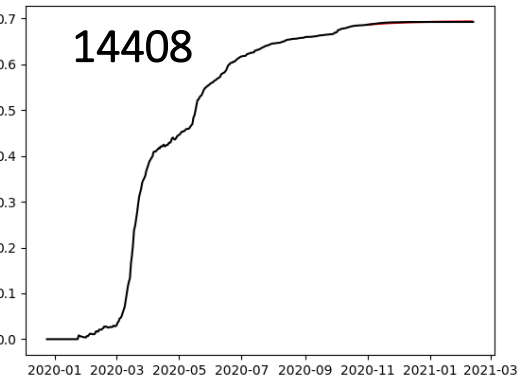
Black curve: data; data before Nov. 2020 is training set; from Dec. 2020 till the end is test set

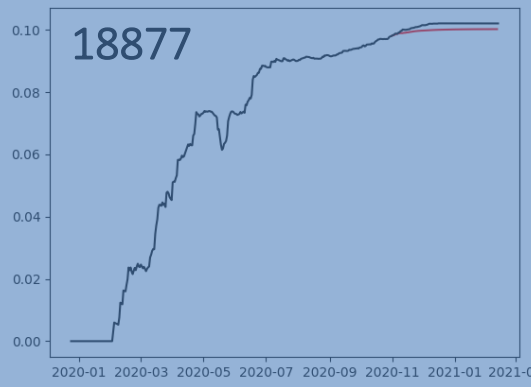
Red curve: prediction on test set; sometimes the prediction is so close to the data that you can not see very wel

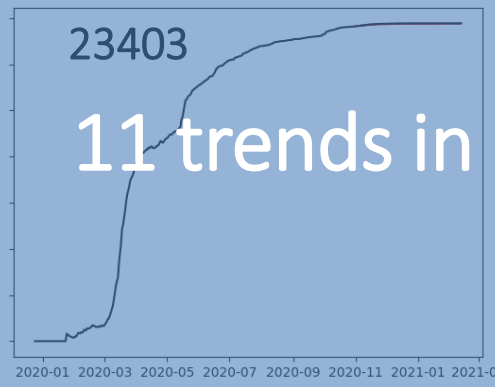


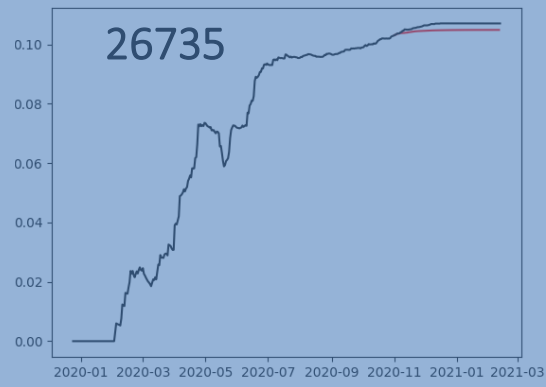








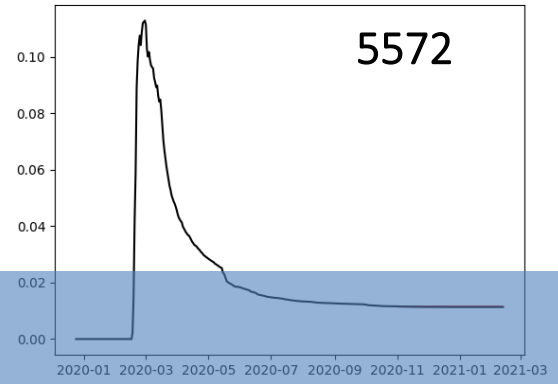


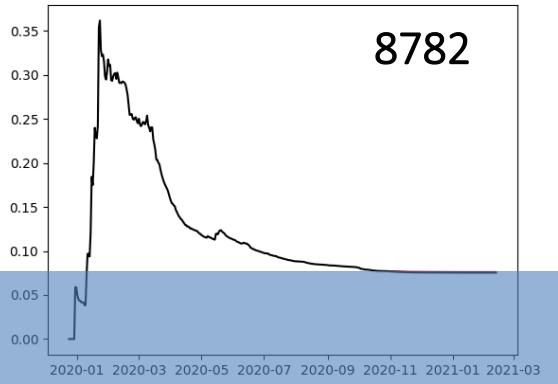


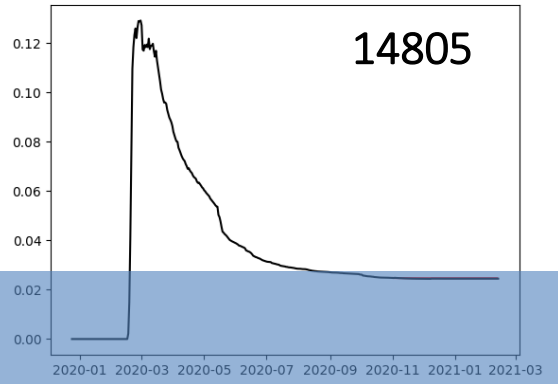
11 trends in cluster 1 in Asia

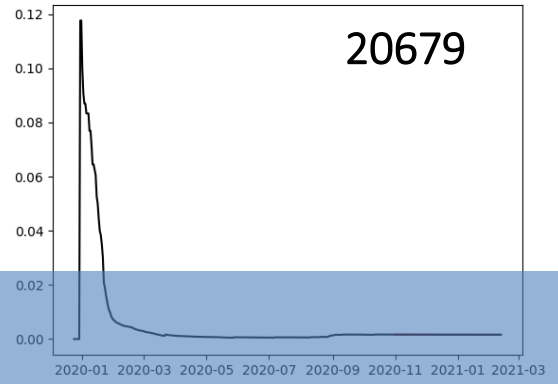
Black curve: data; data before Nov. 2020 is training set; from Dec. 2020 till the end is test set

Red curve: prediction on test set; sometimes the prediction is so close to the data that you can not see very wel

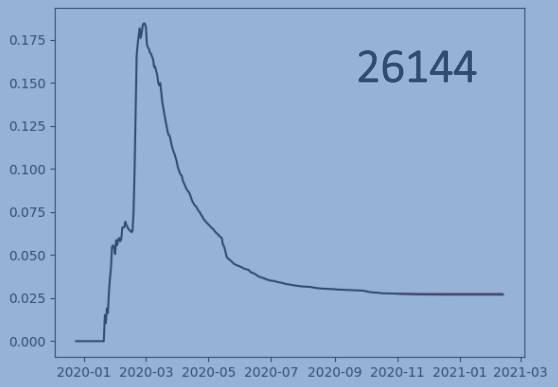


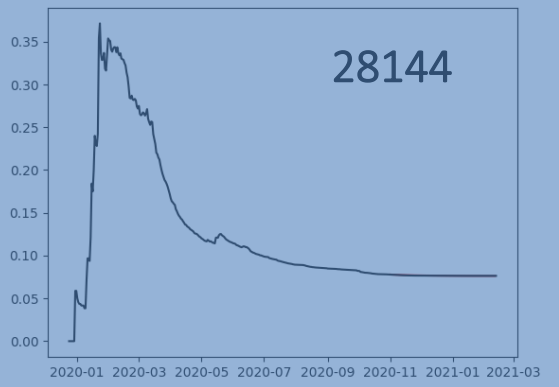


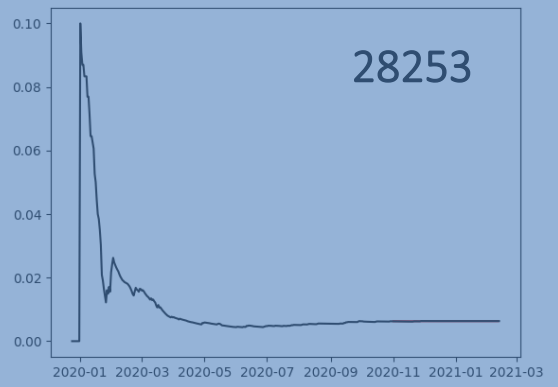


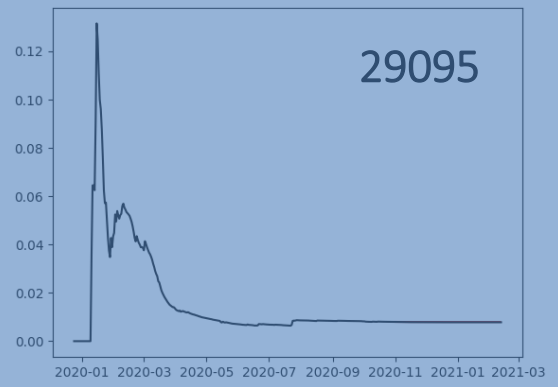


8 trends in cluster 2 in Asia





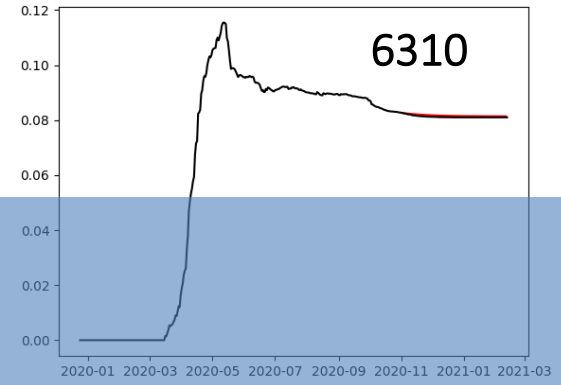


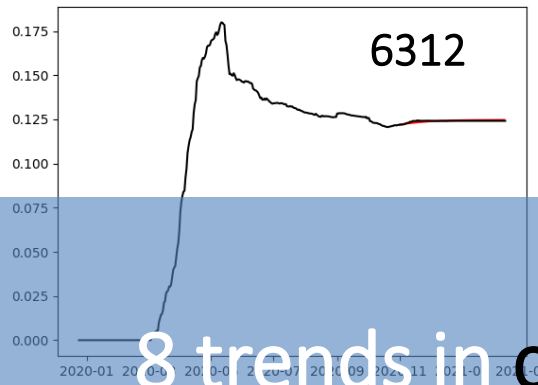


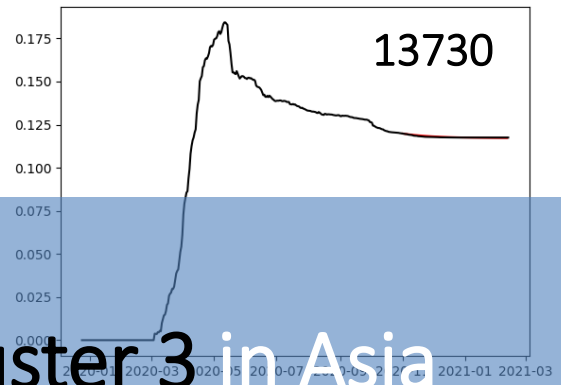
Black curve: data; data before Nov. 2020 is training set; from Dec. 2020 till the end is test set

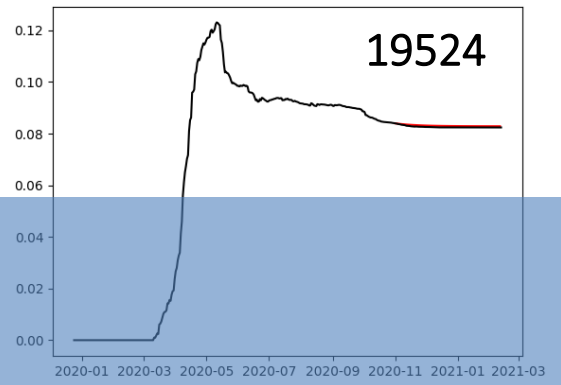
Red curve: prediction on test set; sometimes the prediction is so close to the data that you can not see very wel

8 trends in cluster 3 in Asia





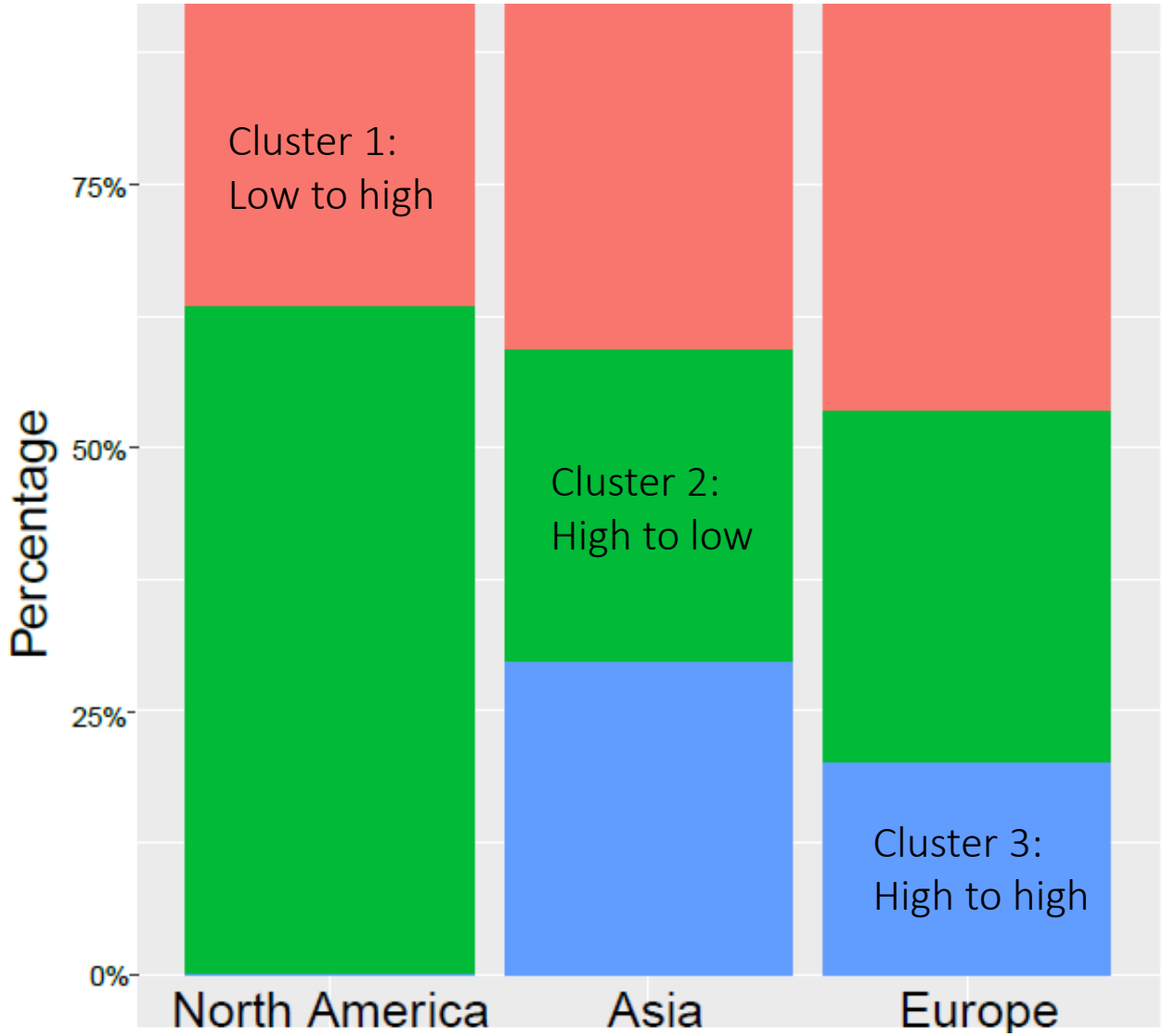




Black curve: data; data before Nov. 2020 is training set; from Dec. 2020 till the end is test set

Red curve: prediction on test set; sometimes the prediction is so close to the data that you can not see very wel

Result & Novel Contribution



Cluster 1 takes for a little more than 30% of each group, but the site in the cluster is different.

Cluster 2 takes about 30% of Asia and Europe, more than two thirds of North America.

Cluster 3 takes around 30% in Asia and 20% in Europe.

Result

Distinct distributions of three clusters occur in Europe, Asia and North America

With the statistic graph, It confirms that the patterns in Asia, Europe, and North America are different.

First Part of the Table for 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

** Full table can be found at the end of all slides

Result

A lot of mutations lead to amino acid change and most important non-synonymous mutations enriched in three gene loci

Non-synonymous and synonymous mutation can lead to different results. We use the change of codon to see whether SNP can cause synonymous or non-synonymous mutations. Most of the SNPs in SARS-CoV-2 genome can contribute to amino acid substitution

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

A lot of mutations lead to amino acid change and most important non-synonymous mutations enriched in three gene loci

Novel Contribution

→ We proposed a series of genome RNA sites that mutated differently in Asia, Europe, and North America.

→ We developed a model to predict the genome RNA mutation in three major continents.

→ We find that the S protein contains multiple high mutated sites which could affect the efficiency of current vaccines.

So, by using this prediction, scientists can have a alternative reference for assessing the COVID-19 and existing vaccines.

Summary

Data:

- We collected gene mutation data and cleaned the data for modeling as time series mutation frequency.

Analysis

- We observed three clusters of mutation trends
- We found different mutation patterns on different continents
- We found 30 high frequency mutations and 13 shared by all continents
- Most of them lead to amino acid changes

Modeling and Prediction

- ARIMA model can forecast the mutation frequency with good quality.

What I have learned in doing the project outside of scientific knowledge.

Rigorous scientific attitude	Be scientifically rigorous and always validate your data and results
Details are very important	A simple mistake in detail can slow you down for weeks.
Learn from the professionals	Professional advice can be hard to fully understood and requires efforts.
Always eager for new knowledge	It was the desire for knowledge that drives a person through difficulties.

Future Work

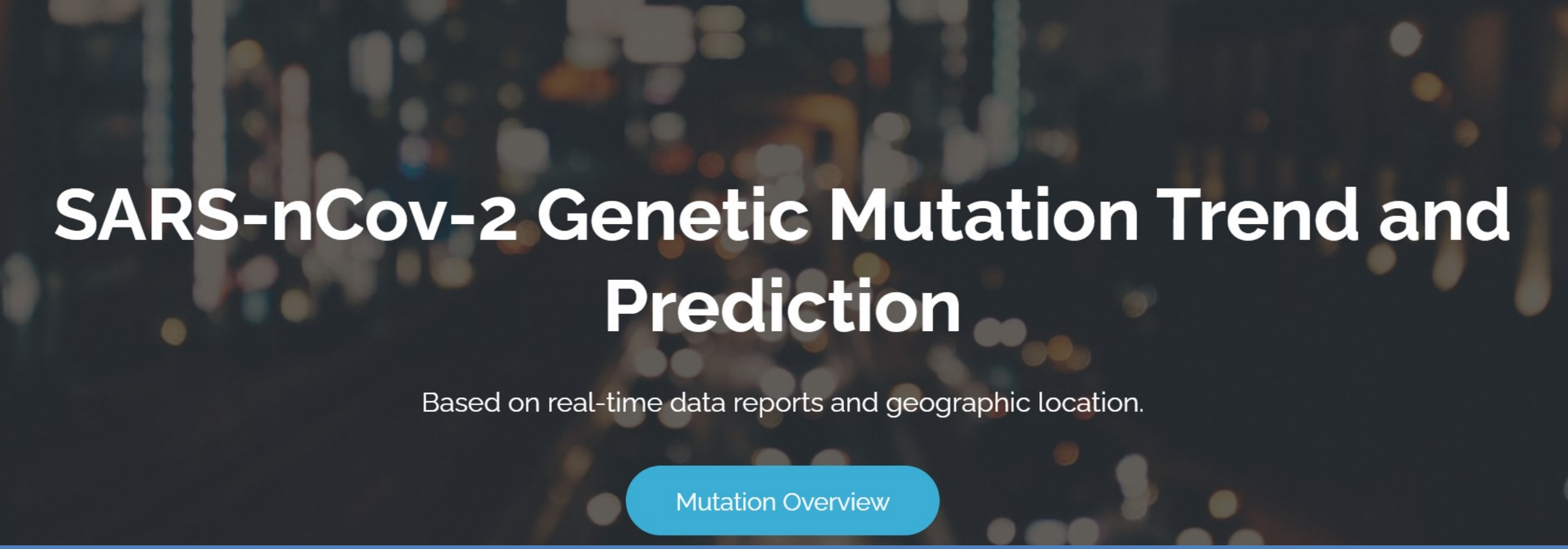
Improve prediction modeling

- Try different criteria for clustering, e.g., using tail/peak ratio
- Try dividing the data into countries and look for new patterns
- Try different cutoff value for picking high frequency mutation sites

Focus the prediction on the mutations that would change S protein

Construction of a website that updates data and visualization in real time, whose functions includes

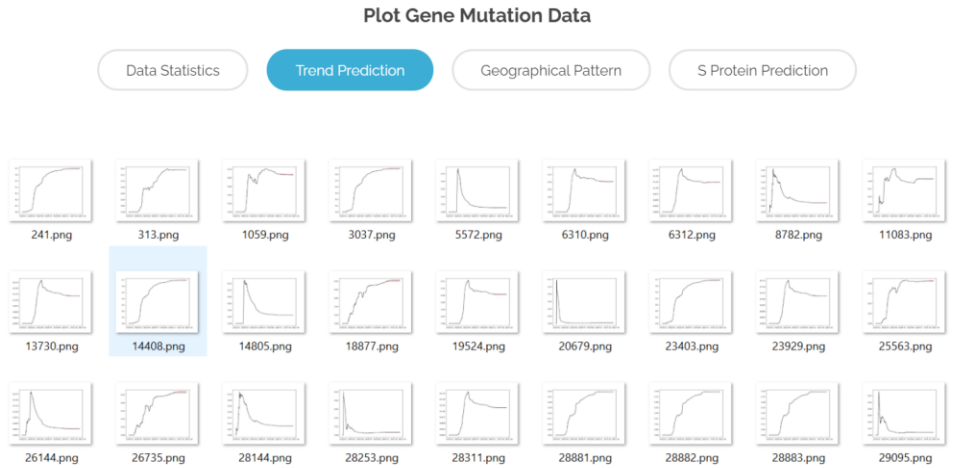
- The number and distribution of samples from the database
- The mutation status and prediction of mutation trends of important sites in each continent
- The impact of these mutations on the function of S protein.
- The demonstration of website is shown on the “extra files” video

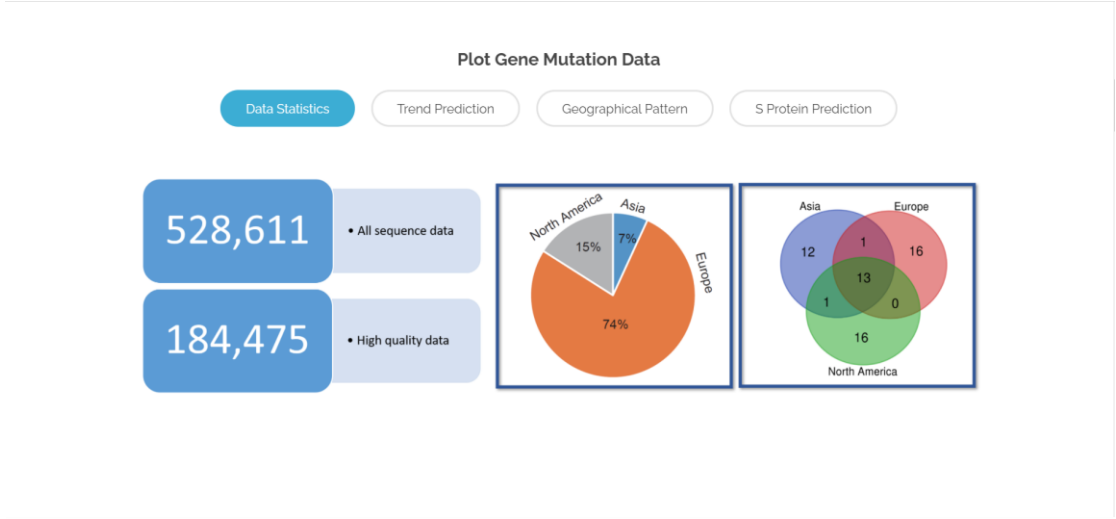


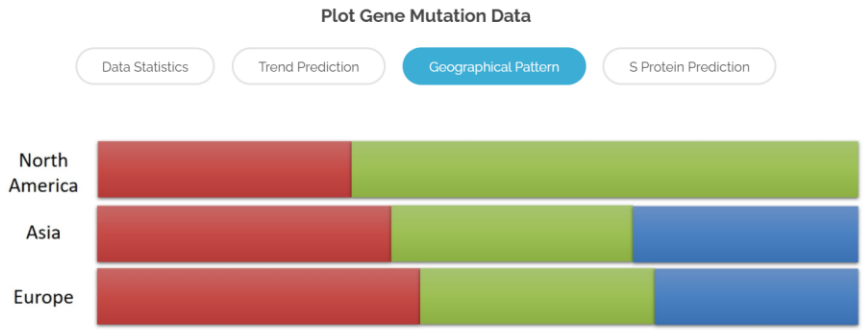
Website Design

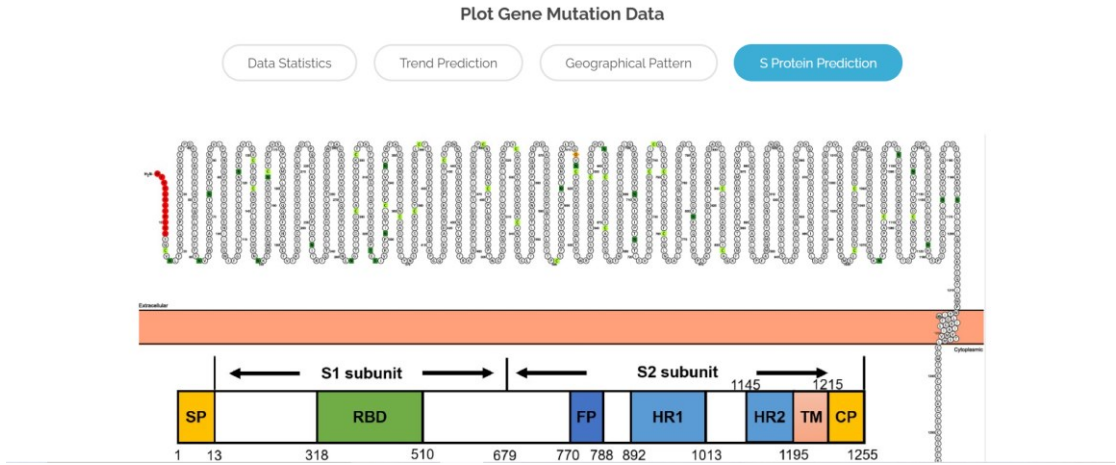
The demonstration of website is shown on the “extra files” video

The webserver can fetch real time data, and updates important mutation predictions.









Acknowledgement

Thanks to Dr. Xiang Gong from Princeton International School of Math and Science for his help on ARIMA modeling and on general scientific advice.

Thanks to Dr. Yuzhe SUN from the University of Hong Kong, for his help on bioinformatics knowledge and for educating me on how to understand the genetic mutations.

Thanks to Dr. Wenming Zhao from China National Center for Bioinformation for his advice on data collection.

References

1. Gong, Z., Zhu, J.W., Li, C.P., Jiang, S., Ma, L.N., Tang, B.X., Zou, D., Chen, M.L., Sun, Y.B., Song, S.H., *et al.* (2020). An online coronavirus analysis platform from the National Genomics Data Center. Zool Res *41*, 705-708.

2. Makizako, H., Tsutsumimoto, K., Doi, T., Makino, K., Nakakubo, S., Liu-Ambrose, T., and Shimada, H. (2019). Exercise and Horticultural Programs for Older Adults with Depressive Symptoms and Memory Problems: A Randomized Controlled Trial. J Clin Med *9*.

3. Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., and Li, F. (2020). Cell entry mechanisms of SARS-CoV-2. Proc Natl Acad Sci U S A *117*, 11727-11734.

4. Sternberg, A., and Naujokat, C. (2020). Structural features of coronavirus SARS-CoV-2 spike protein: Targets for vaccination. Life Sci *257*, 118056.

5. Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T., and Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell *181*, 281-292 e286.

6. Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., *et al.* (2020). Author Correction: A new coronavirus associated with human respiratory disease in China. Nature *580*, E7.

Special 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

Attachment

Special mutated sites in the SARS-CoV-2 genome						
Position	Gene	SNP	Codon change	Amino acid position	Amino acid change	Non-synonymous
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

Special mutated sites in the SARS-CoV-2 genome						
Position	Gene	SNP	Codon change	Amino acid position	Amino acid change	Non-synonymous
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

Special mutated sites in the SARS-CoV-2 genome						
Position	Gene	SNP	Codon change	Amino acid position	Amino acid change	Non-synonymous
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

Special mutated sites in the SARS-CoV-2 genome						
Position	Gene	SNP	Codon change	Amino acid position	Amino acid change	Non-synonymous
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

Special mutated sites in the SARS-CoV-2 genome						
Position	Gene	SNP	Codon change	Amino acid position	Amino acid change	Non-synonymous
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

Special mutated sites in the SARS-CoV-2 genome						
Position	Gene	SNP	Codon change	Amino acid position	Amino acid change	Non-synonymous
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
End of Table						