Mathematical Principles of COVID-19 Infections

Zhaobin Xu[1]*, Qiangcheng Zeng[1]，Zuyi Huang[2]

[1] Department of Life Science, Dezhou University, ShanDong, P.R. China, 253023
[2] Department of Chemical and Biological Engineering, Villanova University, PA, USA, 19085
* Corresponding author: zhaobin23@126.com;

Abstract:

Through statistical analysis of COVID-19 infection cases in many countries, it is noticed the virulence of COVID-19 is indeed decreasing over time. A virulence attenuation theory is proposed due to the corrosion of virus UTR (Untranslated Region) region. Though the statistic analysis of COVID-19 meta-genomic data, the COVID-19 UTR region was confirmed to experience a significant truncation through time but could be stabilized around the length of 29782bp. It was also discovered the virus UTR corrosion probability was dependent on its own length. This might due to the secondary structure of its UTR region. Therefore, it is inferred COVID-19 virulence would not be able to disappear naturally without strong human intervention although its virulence engaged a significant declination compared to the first wave. Instead of using the traditional SIR model, a new mathematical model that could take the virulence decay of COVID-19 into consideration is proposed. The microscopic virus population proliferation model displayed good fitting results in almost all infection cases globally. By constructing a microscope virus replication model, we propose that the major factor driving COVID-19 infection is the very sensitive relationship between the maintenance of virus genome active length and the host population contact density, rather than the traditional virus gene mutation. Due to the existence of this sensitive relationship, strict epidemic prevention measures can reduce the number of viruses in a short time, and finally realize the complete extinction of COVID-19 in human society. This mathematical model could also help explain lots of mystical phenomenon during this epidemic such as how asymptomatic infections happened and why young people drove a second surge of COVID-19 cases.

Keywords:

Covid-19, mathematical modeling, virulence attenuation, genome corrosion

1. Introduction

Since the outbreak in December 2019 and up to July 31, 2020, COVID-19 has caused more than 16 million cases of infection worldwide, making it the largest global public health threat after World War II [1-3]. COVID-19 poses a great threat to human life and health, causing great economic losses and social panic. Currently, the research on COVID-19 mainly focuses on two aspects: the laboratory research on molecular mechanisms of the virus and the research on epidemiological models. Molecular mechanism research includes structural analysis of important protein complexes of virus [4-7], pathogenic mechanism study [4-12], classification of virus gene subtypes, source tracing, evolution diversity [13-17], and potential drugs development [18-24]. Epidemiological modeling mainly refers to the establishment and prediction of the change trend of

COVID-19 cases with time through data fitting [25-31]. However, current epidemiological models cannot accurately predict the relationship between the number of infected people in different regions and time. Most of the existing mathematical models are built on the level of simple regression fitting [43-45]. These types of mathematical models generally do not consider the physical mechanism of virus transmission and attenuation. They thus lack the broad-spectrum prediction capability. Developing epidemiological models of novel coronavirus is helpful for the prevention and control of epidemic situation. It is also of value for exploring and discovering the mechanism of its transmission and pathogenesis.

At present, the most widely used epidemiological model is the SIR model [27-31], which is mainly based on the relationships between the total number of people N, the number of infected people I, the number of susceptible people S, the number of cured people R and the time T in the epidemic area.

$$\frac{dS}{dt} = -\beta \frac{IS}{N} \quad (1)$$

$$\frac{dI}{dt} = \beta \frac{IS}{N} - \gamma I \quad (2)$$

$$\frac{dR}{dt} = \gamma I \quad (3)$$

The SIR model has achieved good results in some epidemiological studies [40-42], but the prediction results for the COVID-19 infections are not ideal [30, 31]. The main problem in the SIR model is the fixed infection coefficient β cannot describe the change in the toxicity of COVID-19 over time. In order to fit the SIR model to the data for the epidemic surging period, a very small initial susceptible number S (i.e., less 100,000) has to be set. This is obviously inconsistent with the actual situation. If the data for the early infection stage was used to fit the SIR model, the obtained γ (γ stands for the recovery percentage of infected population) value is often too small, causing a significant long tail after the peak point of COVID-19 cases. Specifically, the predicted time for the epidemic to end is more than 6 months after the peak point, which is not consistent with actual case either [31]. A model that fits the change of the total number of infected persons over time can make the total number of infected persons decrease rapidly after reaching the peak value. However, that model cannot effectively predict the arrival time of the inflection point from the rising data in the early stage. It cannot predict the peak value of the total number of infected persons either [30, 31].

The trend that describes the overall infection cases with time, especially the rapid decrease of dI/dt after the turning point, fully illustrates the fact that the overall infection coefficient β decreases with time. However, the decrease in the infection coefficient cannot be attributed to the decrease in the size of susceptible population. For the epidemic in any country, the final infection number occupies a very small proportion in the whole population, and the overall susceptible population S will not change significantly with time [27-31]. The decrease of the overall infection coefficient β*S can only be attributed to the decrease of β itself, which may be caused by two reasons: the enhancement of quarantine and the attenuation of virus toxicity. All kinds of data are more inclined to support the attenuation of virus toxicity, although quarantine will also affect β. There

are at least four reasons: different countries and regions have adopted different intensity of prevention, but there is no strict correlation between infections and intervention rigidity. Secondly, β shows a continuous downward trend. The decrease of β caused by government intervention will stabilize β at a lower level, but it will not cause a continuous downward trend. Thirdly, there is relatively very high percentage of severe patients in the initial stage of the virus outbreak. The proportion of severe patients in the newly infected population continues to decrease, and most patients show mild symptoms or even no symptoms in the late stage of epidemic. Finally, the infection rate of susceptible people in close contact, such as medical personnel, has shown an obvious downward trend over time. As far as the situation in China is concerned, the infection rate of medical personnel mostly occurs in the early stage, and none of the 42,600 medical personnel assisting Hubei in the later stage was infected. This astonishing statistical number cannot be simply explained by the enhancement of protective measures. They may imply a decreasing virus toxicity (i.e., decreasing β).

Although statistical curve fitting supports the hypothesis that β decays with time, there is no direct evidence on whether COVID-19 decays with time at the molecular level. If there was a COVID-19decay, the specific mechanism would remain to be studied. At present, the mutation of novel coronavirus over time is widely studied. Many RNA viruses in nature have a high error rate since their RNA polymerase use single strand RNA as template, which is easy to cause the occurrence of mutation catastrophe [32-35]. However, according to the results of gene sequencing, the fidelity of RNA polymerase of coronavirus is relatively high. The mutation frequency is only 1.65 * 10-3/BP/year [14-16]. Although this mutation frequency is higher than that of DNA, it is still unable to cause mutation disaster in a short time (< 3 months). Although some experiments have pointed out that individual point mutations may cause obvious changes in virus activity [39], the experimental results lack repeatability. Moreover, the viral load of the virus strain in cultivated cells also fluctuates randomly, which cannot be explained by the mutation theory. The mutation theory holds that if certain mutations can change the toxicity of the virus, the change in toxicity will not change over time, and the detrimental virulent strain will show higher viral load than the mild virulent strain at any different time points. In addition, the highly toxic strains isolated in this experiment [39] do not come from severe patients, and there is no obvious positive correlation between the toxicity of each strain and the symptoms of the collected sample. Therefore, we speculate that the fluctuation of viral load of different strains in different time periods is caused by other reasons, rather than site mutations. The evidences tend to support the conclusion that in a short period, coronavirus does not engage a decrease or increase in toxicity due to mutation [33]. Mutation-based research can study the differentiation process and evolutionary relationship among different virus strains [13-17]. Nevertheless, based on current sequencing results, the frequency of mutation does not significantly affect any process of virus replication, assembly and transfection.

Since low mutation rate cannot indicate the assumption that virus toxicity will not decay with time, there may be other molecular mechanisms for virus toxicity decline. Based on the analysis of the sequencing results of the whole gene of the new coronavirus in GISAID and Genebank databases, that genome corrosion of the new coronavirus may be the cause of the attenuation of virus toxicity. In the central area of the outbreak, COVID-19 display strong toxicity and the mortality rate is relatively high. The corresponding UTR length is generally longer at that time. With the increase of time, it is noticed that the virus genome engaged significant corrosion at both ends. For

example, the gene sequencing results of Washington state patients show that many virus genome sizes are significantly shorter than that of the reference genome. Some reductions are more than 300 nucleotides, and those already affect the coding sequence of the virus functional gene Orf1ab. Many RNA virus studies have shown that with the progress of replication, UTR region of the virus will have the problem of length attenuation [36-38]. By analyzing the sequencing results of the COVID-19 in different month, gene truncation was noticed to be commonly existed among the collected samples. For this coronavirus, currently there is no experiment research on whether the attenuation of genome length will affect the replication and transcription of virus, but data based on statistics and gene sequencing supports the virus attenuation hypothesis. Since virus gene length is the only variable that changes rapidly with time at molecular level, it is necessary to further study the influence of virus length attenuation on virus replication and translation efficiency at a molecular level.

A hypothesis that the toxicity of coronavirus decays with time is studied in this work. Through mathematical modeling, we established a microscopic virus population proliferation model which has a good prediction in many cases. In order to further study its decay mechanism, we applied the microscopic virus population proliferation model to simulate the virus population variation with time and generations. The model was able to well explain the mortality polymorphism among regions and time. It could also be used to explain the polymorphism of epidemic trends in different countries and the causes of epidemic resurgence. This study is beneficial to enhance the cognition of COVID-19 for the research community. Moreover, it is helpful to accurately predict the epidemic growth trend and provide scientific suggestions to policy makers. Specifically speaking, it is helpful to reveal a possible mechanism on the outbreak and transmission of COVID-19.

2. Methods and Results

2.1 Investigate the Genome Deletion of COVID-19 virus

By sequencing and analyzing the SARS-CoV-2 genome in different months, its genome length was noticed to display an obvious downward trend with time. The sequence data was extracted as of September 16th, 2020 from the SARS-CoV-2 database of Chinese Academy of Sciences, and analyzed the distribution of its genome length with time. For clarity, 20bp was used as the group distance, and its genome size distribution is shown in the Figure 1:

Figure 1 displayed virus genomes of different sequencing dates have been deleted in different degrees. This deletion is also shown a good correlation with time. The genome length of the early virus is normally long, mainly distributed around 29,900bp, and the genome of the later virus is mainly distributed below 29,800bp. If smaller group distance such as 1bp was selected for mapping, it will show that the distribution frequency of viruses are mainly concentrated at 29903bp in the early stage and 29782bp in the later stage. There are two possible origins of these short genome strains such as 29782bp. The first is that they have sibling relationship with long genome strains, and they evolved in parallel. The second is that they are the descendants of these long genome strains, which are derived from the deletion of the long genome virus. If it is the second relationship, can they transform each other, or can they only change from a long genome to

a short genome? Taking 29903bp length strains and 29782bp length strains as examples, we randomly selected 25 29903bp strain samples and 170 29782bp strain samples. By calculating the similarity of genome strains with different lengths, we got the similarity matrix between them, and the specific values are shown in Table 1. Statistical analysis showed that the pair-wise similarity between 29903bp was significantly higher than that between 29782bp and 29903bp ($P = 2.85 * 10^{(-21)}$), and the similarity between 29782bp and 29903bp was significantly higher than that between 29782bp ($P = 5.83 * 10^{(-92)}$) Therefore, statistical analysis proves that short genome length strains are derived from long genome length strains, and short UTR length viruses are not siblings with long UTR length viruses in evolution, but are caused by the deletion of both ends of long fragment strains. This deletion is irreversible, that is to say, short genome strains cannot be transformed into long genome strains, although some evolutionary tree mapping may put long genome length viruses as descendants of short genome length viruses. The figure of genome length changes with time also shows that although the length of virus genome reduced as time, its genome truncation trend is not permanent with time, and it can form a steady state at about 29782bp.

2.2 Develop Virus Micro-Amplification Model

In order to better study the process of virus increment and diffusion, we further study the change process of virus population from the microscopic level. The establishment of this model is based on the following two assumptions:

First, according to the sequencing results, UTR regions at both ends of the virus genome will be deleted in different degrees with time. The reasons for deletion may be corrosion by the host nucleic acid degradation system or poor conservation of its own RNA replicase, etc. Whatever the reasons, the deletion is irreversible, and short-length viruses cannot replicate to produce long-length viruses. Therefore, we need to determine the parameter of virus genome deletion rate. We assume that the length of each deletion is the absolute value of random numbers following the normal distribution with mean value of 0 and standard deviation of $\boldsymbol{\theta}$, which is described in equation (4). The probability of small number base deletions is obviously higher than that of a large number of base deletions. In fact, the deletion probabilities of UTR regions of viruses with different genome lengths are different, which may be related to their secondary structure.

Second, because of the loss of URT at both ends, the replication efficiency of the deleted short-length virus will be affected. The shorter the genome, the lower the replication efficiency would be and the longer the replication cycle would be. There is a certain mathematical relationship between the replication period and the deletion length, so we establish a mathematical function between the deletion length and the variation of replication period. This relationship is described in equation (5).

$$\boldsymbol{Genome_{length} = Genome_{length} - \mathbf{fix}\left(\mathbf{abs}\left(\boldsymbol{normrnd}(\mathbf{0}, \boldsymbol{\theta})\right)\right)} \qquad (4)$$

$$\boldsymbol{Replication_{cycle}} = \frac{\boldsymbol{Replication_{cycle_{initial}}}}{\mathbf{e}^{(\boldsymbol{Genome_{length_{initial}}} - \boldsymbol{Genome_{length}}) * \mathbf{para1}}} \quad (\mathbf{5})$$

Establishment of model

First, assuming that the initial number of viruses is ***n***, the genome length of primary virus is 29903bp. Consider the simulation time, we use a small value of ***n***.

Second, the virus would die in the host cell, that is, it has a half-life. Due to the solvent environment and immune environment of the host cell, its half-life will not be too long. Nevertheless it should be significantly larger than the initial replication cycle of the virus. If the change of the total amount of viruses in a large population is considered, its half-life is also affected by the transmission coefficient $\boldsymbol{R_0}$. If the transmission is blocked, although the replication cycle of viruses does not change, the total amount of viruses will decrease significantly after several generations. In terms of mechanism, the transmission coefficient does not directly affect the half-life of a single virus, but in terms of simulation effect, the reduction of the total number of viruses caused by the smaller transmission coefficient can be explained by the relatively shortened half-life of the virus population. We established a virus half-life parameter ***d*** under different environmental conditions. ***d*** represents the surviving possibility of individual virus within certain time interval. We assumed the $\mathrm{Replication}_{\mathrm{cycl\,e_{initial}}}$ to be 10 min, and time interval was also selected as 10min. So ***d*** represents the surviving possibility of individual virus within 10 min. strict prevention and control measures can lead to a smaller $\boldsymbol{R_0}$ value, thus shortening the half-life of the virus in the whole population which will lead to smaller value of ***d*** in a group population modeling. Meanwhile, individual with strong immunity will display a relatively smaller ***d*** value compared to weak immunity person in single body infection stimulation.

Third, the UTR region of the virus may be deleted in every replication process. We construct a deletion model which is independent of its parent UTR length. Equation (4) is used to calculate the new genome length after one round of replication. Actually, the deletion probability is UTR length dependent, so at different UTR length, different equations will be used to describe its corrosion rate.

Fourth, the deletion of UTR region will affect the replication efficiency of virus, which shows that it increases the replication cycle of virus. There is a certain mathematical relationship between the deletion length and the replication period. Equation (5) is adopted to calculate the new replication cycle with different length. The exponential relationship is preferred compared to the linear relationship. There are two reasons. Firstly, the recognition and binding force of virus replicase and UTR region is often nonlinear correlated with its UTR length. Secondly, the exponential model shows that the virus replication cycle is increasing faster and faster with the increase of deleted fragments, and the deletion of front-end UTR fragments will not significantly affect the virus replication cycle. For the model of population infection, it often shows exponential growth in the early stage of the initial epidemic area, which is consistent with the actual situation.

Fifth, for the population infection model, the number of newly infected people every day is directly proportional to the total number of viruses in the population. For the individual infection model, the severity of symptoms is directly proportional to the virus load in the host. For example, if the virus load in the host passes the nucleic acid test valve, it will be tested positive. If the virus load is humongous, it might turn out to be severe case and fatal. It is difficult for us to estimate the values of parameters ***θ, d,*** and ***para1*** without experiments. However, given the proper relation among those parameters, we can qualitatively predict the changing trend of epidemic situation and explain various phenomena in the development of epidemic situation through simulation.

In Figure2, the virus has ***d*** chance to pass the surviving valve. If (***virus.waiting_time*** + ***Time_interval***) > = ***virus.replication_cycle***, the virus will pass the replication valve to replicate itself, or else, ***New_virus.waiting_time = virus.waiting_time + Time_interval***. If the virus passed the replication valve, it will generate two offspring, with new genome length following the equation (4)
Waiting time of the new generated offspring is 0. Those new generated generations will go to the next round cycle.

As can be seen from Figure 3A, if the model assumes that the virus genome will undergo undifferentiated attenuation with time, that is, the rate and probability of attenuation are irrelevant to the current UTR length, the epidemic situation cannot reach a plateau period, and will drop rapidly after reaching a peak value, which is obviously inconsistent with the actual epidemic situation in many countries. As can be seen from Figure 3B, the viral genomes in different periods of the epidemic will be linearly deleted with time, which conform to the normal distribution on the whole, and will not be concentrated in certain length to form peaks. However, the sequencing results imply that although the genome of the virus as a whole displayed deletion trend in the first few months of epidemic, the meta-genomes data was concentrated in some specific lengths, the most obvious of which are 29903bp and 29782bp. The ratio of virus length distribution at 29782bp gradually increased, but did not form an obvious peak after 29782, indicating that the virus length can maintain a steady state at 29782bp, which also indicates that the deletion probability of virus is related to the current UTR length. In some peak regions such as 29903bp and 29782bp, the deletion possibility in UTR region is significantly lower than that in other regions

Considering the relationship between UTR deletion and its length, we construct a simple model. We only consider the probability difference between 29903bp and 29782bp deletion, that is, the frequency of UTR deletion is divided into three parts, namely, the frequency of deletion at 29903 bp, the frequency of deletion at 29782bp and the frequency of deletion in other regions. The real base decay probability will be more complicated, because apart from 29903bp and 29782bp, the genome also has small peaks at other lengths. For the sake of simplicity, our model does not consider these.

It can be seen from figure 4A that when the UTR deletion probability is affected by the UTR length, the overall epidemic situation can reach a steady state at a high level, and there will be no

downward trend after reaching the first peak, which is consistent with the actual situation of the epidemic situation. At the same time, it can be seen from Figure 4B that when the UTR deletion probability is affected by its UTR length, the length of virus genome will decline with time in the early stage, but it will not decline all the time. After the epidemic develops to a certain stage, the length distribution of its genome reaches a steady state, which is consistent with the current sequencing results. At the same time, its genome will show the characteristics of high frequency distribution in a specific length. Since the length of UTR region is closely related to virus activity, the UTR region is relatively long in the early stage of epidemic, the virus replication cycle is short which is corresponded with a high mortality rate. With the development of time, most of SARS-CoV-2's genome decayed to 29782bp, and the lost UTR will affect the virus proliferation efficiency, so the mortality rate will drop obviously. However, because the attenuation of virus UTR region will be affected by the current UTR length, the potential complementary strand might form a stable secondary structure at certain node point, such as 29782bp length. This may effectively prevent the further corrosion from host nucleic acid degradation system. The distribution of virus genome length will not be attenuated all the time, so the mortality rate in the later period of epidemic will not decrease all the time, but present a relatively stable situation. It is not plausible to explain the changing relationship of mortality in SARS-CoV-2 by mutation theory. If the direction of mutation continues to develop in the direction of weak toxicity, then the mortality rate in the middle and late period of the epidemic should show a continuous downward trend, but the actual phenomenon is not the case. Therefore, we believe that in general, the dominant factor for the change of virus toxicity is the attenuation of UTR region length, not base mutation. Moreover, the length of UTR region will not show infinite attenuation all the time, so SARS-CoV-2's virulence will maintain at certain range. SARS-CoV-2 cannot die out naturally in a short time before reaching group immunity threshold.

Under the condition of the same population contact density, the initial growth rate of the number of infected people in the original epidemic area was significantly higher than that in the post-epidemic area. The epidemic trend mainly depends on the actual population contact density.

It can be seen from Figure 5A that the epidemic situation caused by different initial virus lengths will have different development trends. The initial virus length is normally long. Long UTR region will make the virus have higher proliferation efficiency, so the epidemic situation will have a rapid growth trend in the early stage, as shown by the blue curve in Figure 5A. This situation can be clearly reflected in the epidemic curves of Europe and the United States. The epidemic situation caused by late imported cases often shows a slow growth in the early stage, and the virus UTR region of late imported cases is often attenuated to varying degrees, so the virus replication efficiency is low, resulting in a slow growth in the early stage of the epidemic situation, as shown by the red curve in Figure 5A. This is well reflected in some countries where the epidemic started late. It can be seen from Fig. 5B that the epidemic trend mainly depends on the actual population contact density, which affects the transmission coefficient $\boldsymbol{R_0}$ of the virus, and then affects the population half-life decay rate of the virus. Excellent prevention and control measures can effectively reduce the d value, and the smaller d value can lead to the gradual growth of epidemic situation, as shown by the red curve in the figure. For individuals, $\boldsymbol{d}$ value mainly depends on the strength of immunity, and stronger immunity corresponds to smaller $\boldsymbol{d}$ value. It can be seen from Figure 5C that premature unsealing may lead to the second wave of epidemic, but the rising trend

of the second wave of epidemic is obviously slower than that of the first wave of epidemic, and the mortality rate will also be significantly lower than that of the first wave of epidemic.

As far as individuals are concerned, people with strong immunity are less likely to be infected with viruses than people with weak immunity, and the mortality rate of virus infection is significantly lower than that of people with weak immunity. However, when the virus concentration in the environment rises to a certain degree, the number of inhaled viruses will increase greatly, all people will lose their barrier function against viruses, and infections will be positive detected by nucleic acid detection. However, people with strong immunity are not easy to develop into symptomatic and severe patients. For different periods, the virus in the early stage is more toxic, and for people with the same immunity, the fatality rate of high-length virus is obviously higher than that of late-length virus.

It can be seen from the figure6 A, for people with strong immunity, in the random simulation of 5000 population samples, the maximum viral load in the host body is obviously smaller than that of people with weak immunity. The proportion of individuals with strong immunity who eventually develop into individuals with high viral load is very small, while the proportion of individuals with low immunity who eventually develop into individuals with high viral load is significantly higher than those with strong immunity. This can also explain the significant age difference in mortality rate of SARS-CoV-2 infection.

It can be seen from Fig. 6B that when people with the same immunity intensity are invaded by a virus with a long UTR length of 29900bp with a small dose (***n***=1), the random simulation of 5,000 individual samples shows that the majority of individuals have not significantly proliferated their viruses in their bodies. However, a small proportion of samples still show the characteristics of high virus load, as shown in the blue part of the figure. When invaded by a virus with a short UTR length of 29850bp (***n***=20), the random simulation of 5,000 individual samples showed that the virus of most individuals proliferated significantly in their bodies, but almost no samples showed the characteristics of high viral load, as shown in the brown part of the figure. This can better explain the trend of younger infected people in the later period of the epidemic. At the beginning of the epidemic, the UTR region of virus is long and has strong toxicity, but the concentration in the environment is low, and the number of invading human body is relatively small. For people with strong immunity, the maximum reproductive load of most viruses cannot reach the concentration of positive nucleic acid detection. Although infection is not easy to occur, once infection occurs, compared with those infected in the later period of the epidemic, there is a greater risk of turning into severe condition or even death. The UTR region of virus is generally short and its toxicity is weak in the late epidemic period. However, a large amount of concentration has accumulated in the environment, and the number of invading viruses has increased relatively. The increase of the number of invading viruses will make the maximal viral load in host exceed the threshold of positive nucleic acid detection. Therefore, young people have lost their shielding effect on SARS-CoV-2, and the epidemic situation is characterized by youthfulness. However, due to the reduced toxicity of the virus, the viral load in infected people often does not reach the scale of severe illness or death, so the mortality rate and severe illness rate of infected people in the later period are obviously reduced.

The datasets and codes for generating these figures are uploaded as supplementary materials.

3. Discussion

We put forward a mathematical model of virus toxicity attenuation for the first time. Although the attenuation of virus toxicity with time has been confirmed by many researchers, we put forward the molecular mechanism of coronavirus toxicity attenuation for the first time. In addition to generating immune response to eliminate the virus, the replication system of the host will irreversibly cause corrosion and deletion of the virus genome. The UTR sequence plays an extremely important role in virus activity, and this deletion will affect the occurrence of virus transcription or replication, thus affecting the replication efficiency of the virus.

This hypothesis can well explain the following phenomena:

First, masks and social distance play a decisive role in epidemic prevention and control. Under the condition of high population density, the virus cannot disappear naturally without group immunization. There may be great differences in the percentage requirements of group immunity under different population density conditions, and the requirements of group immunity in areas with high population density may be significantly higher than those in areas with low population density. Given the fact that COVID-19 antibody in human body may not exist for a long time, in some areas with high population density and frequent individual contact, if the population cannot be intensively vaccinated, the group immunization criterionmay never be reached. Masks and social distance actually play a role in reducing the contact density of population. A magical feature of novel coronavirus is that it is very sensitive to the contact density of host population. This can also explain why bats are the source of many coronaviruses. The bat genome is highly similar to the human genome, and there is a similar RNA corrosion mechanism in its body. However, due to its high population density, super-strong mobility and super-high contact frequency, the population contact density is second to none in mammals, so the coronavirus genome can achieve a steady state and a long-term coexistence in bat population. The population of human beings is increasing significantly in recent years, and the progress of science and technology makes the population mobility and contact frequency become higher and higher, which greatly increases the population contact density of human beings. If COVID-19 is not well controlled, we may coexist with COVID-19 for a long time like bats.

Second, this could explain the randomness of individual infection. For example, in group infection and family infection events, not all close contact people will be infected. Our hypothesis is that the virus particles released into the environment by patients have great toxicity differences due to various core RNA genome size. If a susceptible individual comes into contact with virus particles with strong toxicity, it will cause infection with critical symptoms. If one contacts with virus particles with long-deleted fragments, mild symptom or asymptomatic infection may occur because the low toxicity of the virus is not able to cause the immune system to overreact. For the situation that people inhale severely deleted virus particles, that is, virus particles with extremely low toxicity, antibodies may be generated. However, since the virus concentration in the body is always at an extremely low level, nucleic acid detection may always be negative. Meanwhile, we can explain the experimental results of Li Lanjuan’s research group [39] in a better way. Because

the deletion length of the initial transfected virus is different, viral load at different times will display significant differences. Since the deletion in the host cell is a random process, the virus loading originated from different virus strains might fluctuate at different time points, that is, the toxicity of the virus will change randomly with time. The viruses isolated from patients are random, and viruses isolated from severe patients may also be low-toxicity viruses with large fragment deletions. This will lead to no obvious positive correlation between virus toxicity and symptoms of isolated patients in the experiment results.

Third, why does the epidemic show various trends in different countries? As for the epidemic curve, we can roughly classify them into these situations: ① It rose rapidly, and then declined continuously until the epidemic basically disappears. ② It rose rapidly, and then reached a plateau period and existed stably for a long time. ③ It rose slowly in the early stage but kept rising for long time before the inflection point finally reached. ④The second wave or the third wave of epidemic occurred. Our microscopic virus population proliferation model shows that in the original place of the epidemic, under strict epidemic prevention conditions, due to the long length of the initial virus genome, the growth in the early stage showed a horrible exponential growth. However, with the further corrosion of the virus genome, the UTR sequence will be gradually deleted. At this time, it will gradually affect the toxicity of the virus, resulting in a more obvious decline relationship. Combined with the reproductive ability of the virus itself, its growth rate will slow down and finally reach the peak or inflection point. This will eventually experience a decline after the inflection point. Due to the decrease of virus toxicity, not only the number of infected people decreased, but also the proportion of severe cases became lower and lower. This situation can be applied to China, South Korea, and some European countries. Moderate epidemic prevention measures can reduce the population contact density to a certain range, which can maintain the number of viruses at a certain level for a long time. When the control is relaxed, there will be a second high platform of epidemic situation, which is reflected as the cases in the United States. Medium-level prevention and control measures with super high population density will lead to large population contact density, but due to the short initial virus length and poor initial toxicity, it will not experience exponential growth in the early epidemic period. However, this kind of growth is often slow and long-lasting. Although the mortality rate is low, it is the most terrible growth mode. This situation applies to India. If it is not strictly controlled, the new infection cases per day may reach an extremely horrible level. Our model reveals that premature prevention relaxation can lead to a second epidemic, which has already happened in many countries, such as the United States, Israel, Iran and so on.

Fourth, as far as individuals are concerned, there are subtle differences in immunity among individuals of different races and ages, which can cause weak changes in $\boldsymbol{d}$ value, so there may be significant differences in mortality among people of different races and ages. However, for the same region, especially the original place of the epidemic and the epidemic area caused by early imported cases, the toxicity of the virus will decrease with time. Therefore, the mortality rate at the initial stage of the epidemic is extremely high, and with the extension of time, the mortality rate will show different degrees of decline. For example, in terms of mortality rate, the initial mortality rate was extremely high in Italy, Spain, Wuhan, Northeast America and other regions, and then there was a significant downward trend. However, in the absence of strict prevention and

control measures, the toxicity of the virus will not always show a rapid decline trend, but remain in a stable range. Therefore, it is impossible for COVID-19 to die out naturally over time.

Fifth, why are there a large number of asymptomatic infected people in the later period, and are asymptomatic infected people contagious? ***d*** value is personal dependent, strong immune response will lead to smaller ***d*** value, therefore it might lead to the quick elimination of infected virus. In this case, the viruses carried by those asymptomatic people are infectious. However, when the virus genome is deleted to a certain extent, it will not cause obvious harm to infected people. Because the partial deletion of its genome affects its normal biological activity, its reproduction ability will be greatly weakened. The immune system will not overreact, so it is mild or asymptomatic. Many asymptomatic infected people are not really asymptomatic infected people, but symptomatic infected people who are still in the latent period. Because the viral load is still in a low range, the immune system in the body has not been activated, so there is no obvious symptom. For the real asymptomatic infected case, the average infectivity should be lower than that of ordinary infected case. Apart from the influence of self-immunity, the genome deletion degree of virus carried by asymptomatic infected people is more serious than that of ordinary infected people, so the virus carried by asymptomatic infected people is less toxic, and it may not even activate the autoimmune system to remove a small amount of virus for a long time, so some of these asymptomatic infected people will have long-term positive nucleic acid detection. In any country or region, the real infected population is much higher than the statistical number of infected people. If antibodies are used for detection, the statistical number of infected people is much higher than the number of people using nucleic acid detection. The reason for this is that the sensitivity of the immune system is much higher than that of nucleic acid detection, and nucleic acid detection can only detect whether there is a virus in the body at present, while antibody detection can detect whether it has been infected with a virus, which also shows that asymptomatic infected people actually occupy a very large proportion.

4. Conclusions

Through the simulation of system biology, we put forward a mathematical model of virus toxicity attenuation for the first time. Although gene mutation occurs widely, we propose that the major factor driving COVID-19 infection is the very sensitive relationship between the maintenance of virus genome active length and the host population contact density, rather than the traditional virus gene mutation. Due to the existence of this sensitive relationship, COVID-19 will attenuate firstly but would be able to maintain its genome size at about 29782bp. Strict epidemic prevention measures can reduce the number of viruses in a short time, and finally realize the complete extinction of COVID-19 in human society. Unfavorable prevention and control measures can not reduce the population contact density, which would cause a stable or even worse epidemic. Before reaching the threshold of group immunity, COVID-19 will not die out naturally with the extension of time. Asymptomatic infected people will occupy a large and stable proportion in the later stage of the epidemic, but there are still a certain proportion of symptomatic patients. Young people will lose their shielding effect on SARS-CoV-2, and the late epidemic situation is characterized by youthfulness when the virus concentration increases. Without manual intervention, this large-scale infection will last for a long time before reaching the group immunization threshold. Early

relaxation of epidemic prevention measures will probably lead to a resurgence of epidemic situation.

In addition, by comparing the simulation results with the actual sequencing, we found that the non-coding UTR region of COVID-19 played a very important role in virus replication. The extremely high population contact density (such as bat population density) may maintain its original sequence length (29903bp). The current population contact density of human beings cannot maintain its original sequence length, but its genome can be maintained around 29782bp. If strict prevention and control measures are given to effectively reduce the population contact density, the epidemic situation can be effectively controlled and eventually the virus can be completely eliminated. On the contrary, if it is not strictly controlled, it is difficult for COVID-19 to die out naturally, and it is very likely that mankind will coexist with novel coronavirus for a long time.

References:

[1] Novel C P E R E. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China[J]. Zhonghualiuxingbingxue za zhi= Zhonghualiuxingbingxuezazhi, 2020, 41(2): 145.

[2] Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019[J]. New England Journal of Medicine, 2020.

[3] Wang C, Horby P W, Hayden F G, et al. A novel coronavirus outbreak of global health concern[J]. The Lancet, 2020, 395(10223): 470-473.

[4] Shang J, Wan Y, Liu C, et al. Structure of mouse coronavirus spike protein complexed with receptor reveals mechanism for viral entry[J]. PLoS pathogens, 2020, 16(3): e1008392.

[5] Gao Y, Yan L, Huang Y, et al. Structure of RNA-dependent RNA polymerase from 2019-nCoV, a major antiviral drug target[J]. bioRxiv, 2020.

[6] Yang T J, Chang Y, Ko T, et al. Cryo-EM analysis of a feline coronavirus spike protein reveals a unique structure and camouflaging glycans[J]. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117(3): 1438-1446.

[7] Wrapp D, Wang N, Corbett K S, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation[J]. Science, 2020, 367(6483): 1260-1263.

[8] Zhang L, Lin D, Sun X, et al. X-ray Structure of Main Protease of the Novel Coronavirus SARS-CoV-2 Enables Design of α-Ketoamide Inhibitors[J]. bioRxiv, 2020.

[9] Chen Y, Liu Q, Guo D. Emerging coronaviruses: genome structure, replication, and pathogenesis[J]. Journal of medical virology, 2020.

[10] Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China[J]. The Lancet, 2020, 395(10223): 497-506.

[11] Zhou P, Yang X L, Wang X G, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin[J]. Nature, 2020: 1-4.

[12] Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding[J]. The Lancet, 2020, 395(10224): 565-574.

[13] Xiong C, Jiang L, Chen Y, et al. Evolution and variation of 2019-novel coronavirus[J]. bioRxiv, 2020.

[14] Yu W B, Tang G D, Zhang L, et al. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2) using whole genomic data[J]. ChinaXiv, 2020, 202002: v2.

[15] Fang B, Liu L, Yu X, et al. Genome-wide data inferring the evolution and population demography of the novel pneumonia coronavirus (SARS-CoV-2)[J]. bioRxiv, 2020.

[16] Zehender G, Lai A, Bergna A, et al. GENOMIC CHARACTERISATION AND PHYLOGENETIC ANALYSIS OF SARS-COV-2 IN ITALY[J]. medRxiv, 2020.

[17] He J, Tao H, Yan Y, et al. Molecular mechanism of evolution and human infection with the novel coronavirus (2019-nCoV)[J]. bioRxiv, 2020.

[18] Li G, De Clercq E. Therapeutic options for the 2019 novel coronavirus (2019-nCoV)[J]. 2020.

[19] Khan R J, Jha R K, Amera G M, et al. Targeting Novel Coronavirus 2019: A Systematic Drug Repurposing Approach to Identify Promising Inhibitors Against 3C-like Proteinase and 2'-O-Ribose Methyltransferase[J]. 2020.

[20] Prajapat M, Sarma P, Shekhar N, et al. Drug targets for corona virus: A systematic review[J]. Indian Journal of Pharmacology, 2020, 52(1): 56.

[21] Li G, De Clercq E. Therapeutic options for the 2019 novel coronavirus (2019-nCoV)[J]. 2020.

[22] Pillaiyar T, Meenakshisundaram S, Manickam M. Recent discovery and development of

inhibitors targeting coronaviruses[J]. Drug Discovery Today, 2020.

[23] Wang M, Cao R, Zhang L, et al. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro[J]. Cell research, 2020, 30(3): 269-271.

[24] Matsuyama S, Kawase M, Nao N, et al. The inhaled corticosteroid ciclesonide blocks coronavirus RNA replication by targeting viral NSP15[J]. bioRxiv, 2020.

[25] Wang H, Wang Z, Dong Y, et al. Phase-adjusted estimation of the number of Coronavirus Disease 2019 cases in Wuhan, China[J]. Cell Discovery, 2020, 6(1): 1-8.

[26] He S, Tang S, Rong L. A discrete stochastic model of the COVID-19 outbreak: Forecast and control[J]. Mathematical Biosciences and Engineering, 2020, 17(4): 2792.

[27] Nesteruk I. Estimations of the coronavirus epidemic dynamics in South Korea with the use of SIR model[J]. Preprint.] ResearchGate, 2020.

[28] Nesteruk I. Statistics based predictions of coronavirus 2019-nCoV spreading in mainland China[J]. MedRxiv, 2020.

[29] Götz T. First attempts to model the dynamics of the Coronavirus outbreak 2020[J]. arXiv preprint arXiv:2002.03821, 2020.

[30] Zhang J, Wang L, Wang J. SIR Model-based Prediction of Infected Population of Coronavirus in Hubei Province[J]. arXiv preprint arXiv:2003.06419, 2020.

[31] Fanelli D, Piazza F. Analysis and forecast of COVID-19 spreading in China, Italy and France[J]. Chaos, Solitons & Fractals, 2020, 134: 109761.

[32] Eigen M. Error catastrophe and antiviral strategy[J]. Proceedings of the National Academy of Sciences, 2002, 99(21): 13374-13376.

[33] Denison M R, Graham R L, Donaldson E F, et al. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity[J]. RNA biology, 2011, 8(2): 270-279.

[34] Imbert I, Guillemot J C, Bourhis J M, et al. A second, non-canonical RNA-dependent RNA polymerase in SARS Coronavirus[J]. The EMBO journal, 2006, 25(20): 4933-4942.

[35] Grande-Pérez A, Sierra S, Castro M G, et al. Molecular indetermination in the transition to error catastrophe: systematic elimination of lymphocytic choriomeningitis virus through mutagenesis does not correlate linearly with large increases in mutant spectrum complexity[J]. Proceedings of the National Academy of Sciences, 2002, 99(20): 12938-12943.

[36] Kao C C, Singh P, Ecker D J. De novo initiation of viral RNA-dependent RNA

synthesis[J]. Virology, 2001, 287(2): 251-260.

[37] Agrawal S, Gupta D, Panda S K. The 3′ end of hepatitis E virus (HEV) genome binds specifically to the viral RNA-dependent RNA polymerase (RdRp)[J]. Virology, 2001, 282(1): 87-101.

[38] Mahilkar S, Paingankar M S, Lole K S. Hepatitis E virus RNA-dependent RNA polymerase: RNA template specificities, recruitment and synthesis[J]. Journal of General Virology, 2016, 97(9): 2231-2242.

[39] Yao H P, Lu X, Chen Q, et al. Patient-derived mutations impact pathogenicity of SARS-CoV-2[J]. CELL-D-20-01124, 2020.

[40] Bjørnstad O N, Finkenstädt B F, Grenfell B T. Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model[J]. Ecological monographs, 2002, 72(2): 169-184.

[41] Side S, Noorani M S M. A SIR model for spread of dengue fever disease (simulation for South Sulawesi, Indonesia and Selangor, Malaysia)[J]. World Journal of Modelling and Simulation, 2013, 9(2): 96-105.

[42] Osthus D, Hickmann K S, Caragea P C, et al. Forecasting seasonal influenza with a state-space SIR model[J]. The annals of applied statistics, 2017, 11(1): 202.

[43] Mohamadou Y, Halidou A, Kapen P T. A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19[J]. Applied Intelligence, 2020, 50(11): 3913-3925.

[44] Sameni R. Mathematical modeling of epidemic diseases; a case study of the COVID-19 coronavirus[J]. arXiv preprint arXiv:2003.11371, 2020.

[45] Kucharski A J, Russell T W, Diamond C, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study[J]. The lancet infectious diseases, 2020.

Pair-wise sequence similarity score	Among 29903bp	Between 29903bp and 29782bp	Among 29782bp
Mean Value	69303	68371	67646
Standard Deviation	14.65	1695.9	2119.6
Max Value	69336	69329	69337
Min Value	69265	59460	56348
Sample Size	300	4250	14365

Table 1

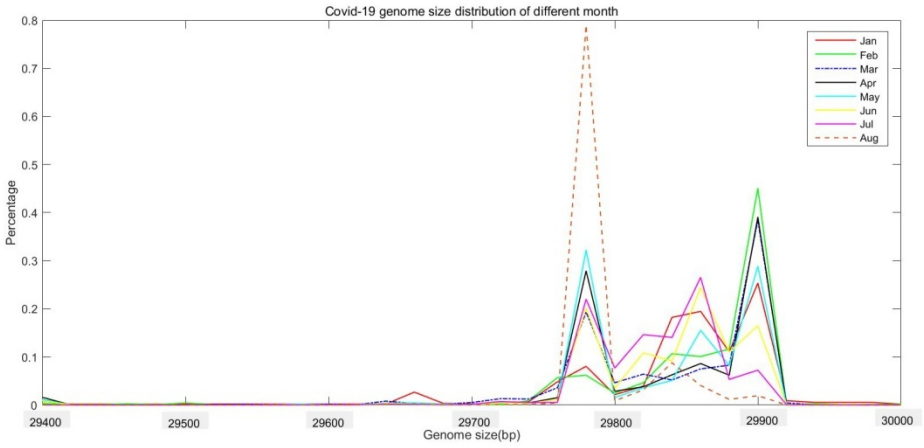

Figure 1

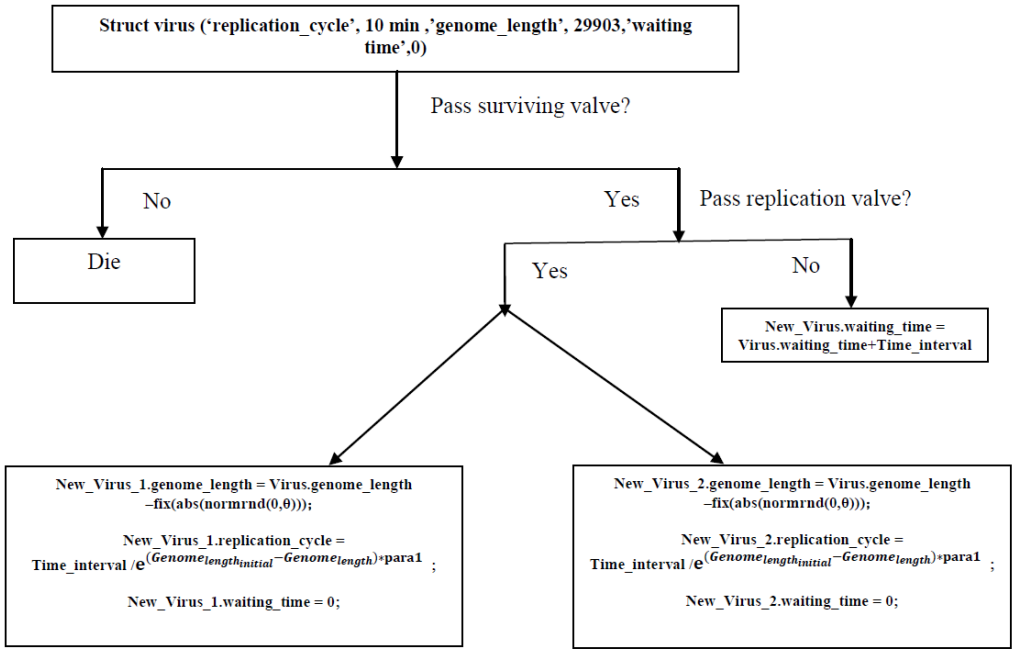

Figure 2

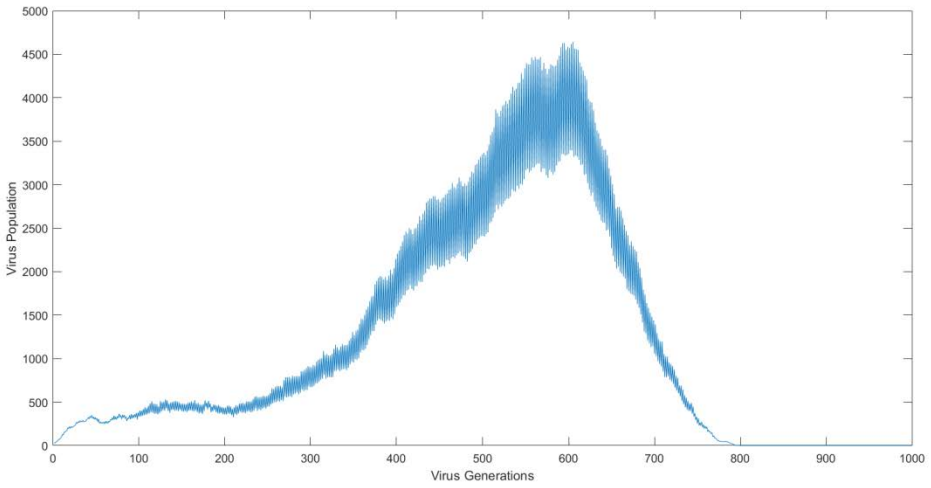

Figure 3 A)

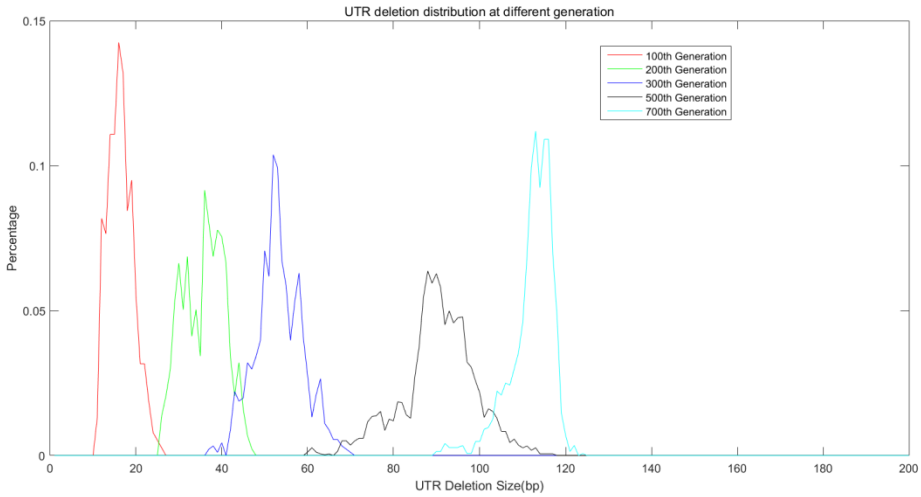

Figure 3 B)

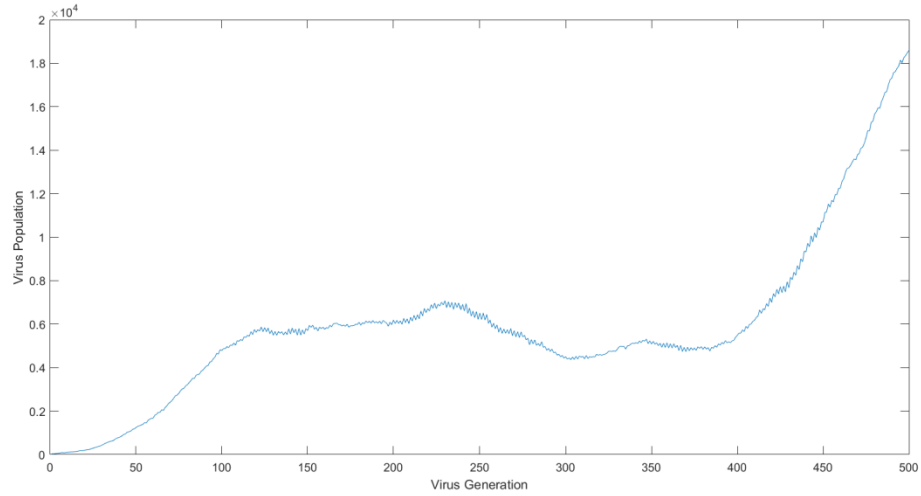

Figure 4 A)

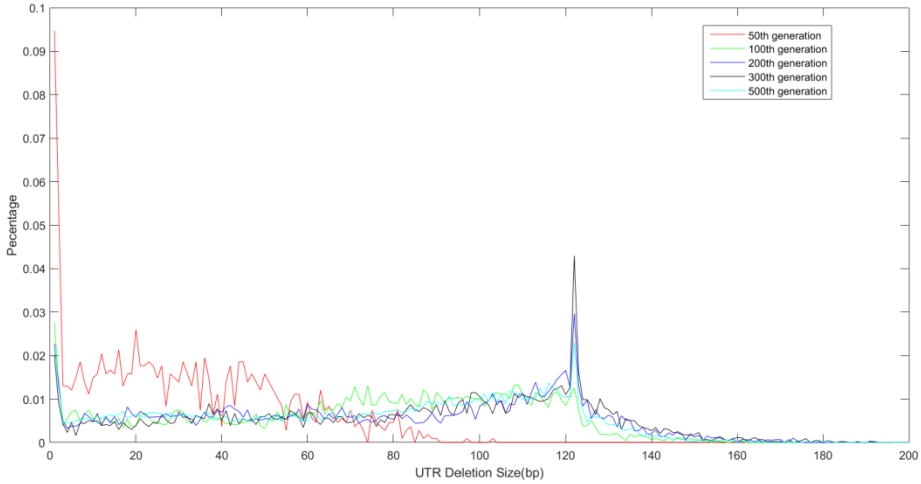

Figure 4 B)

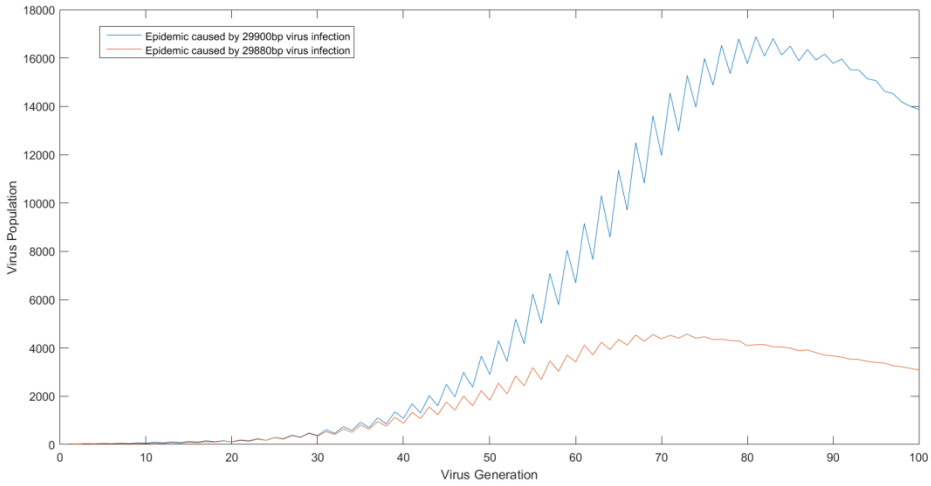

Figure 5 A)

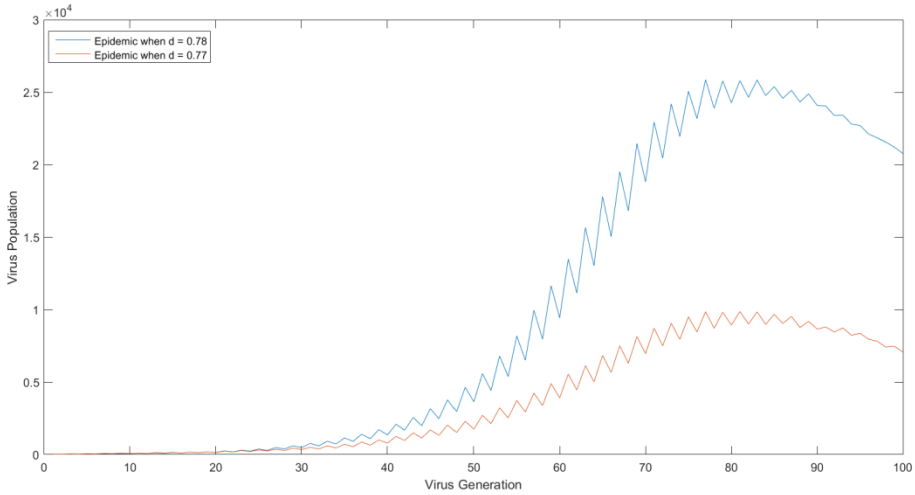

Figure 5 B)

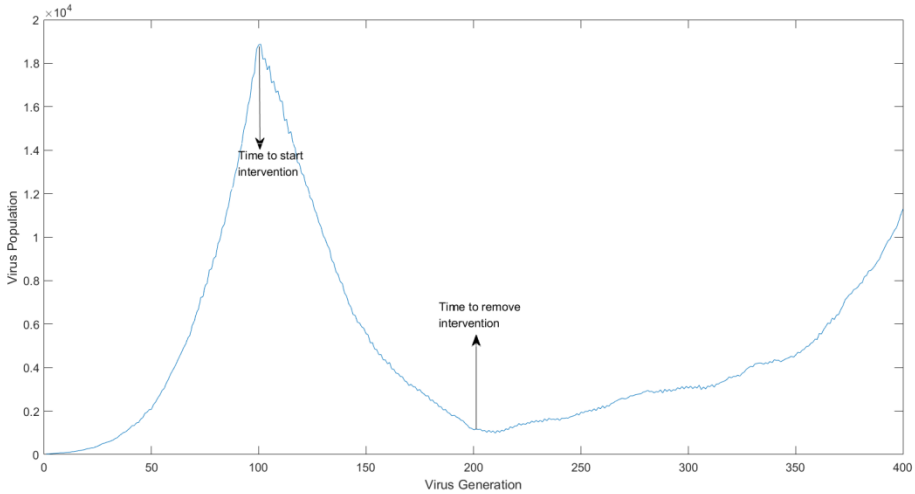

Figure 5 C)

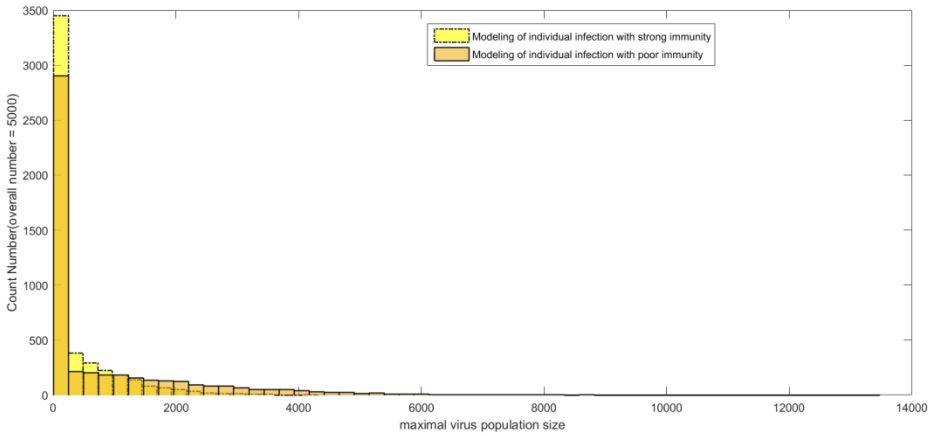

Figure 6 A)

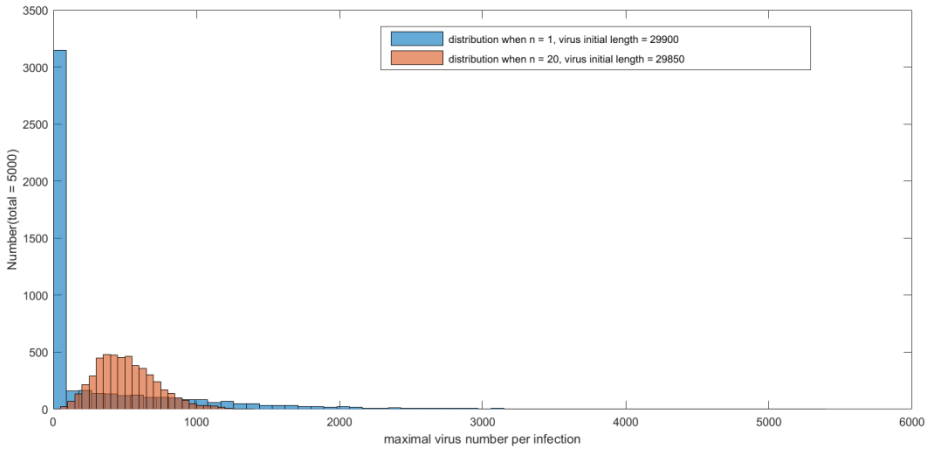

Figure 6 B)

Legend：

Table 1: Pair-wise sequence similarity score between 29903bp group and 29780bp group.

Fig. 1: COVID-19 genome size distribution of different month. 20bp was selected as group interval to plot the figure. The distribution of January, February, March, April, May, June, July and August is each represented in red line, green line, dashed blue line, black line, cyan line, yellow line, purple line and dashed brown line.

Fig. 2: Algorithm flowchart.

Fig.3 A): Modeling of COVID-19 proliferation through time. X axis is the generations which from 1 to 1000 in this stimulation. Y axis represents the virus population size which is described as the overall number of virus. The ancestor virus are 20 29903bp viruses with $\boldsymbol{d} = 0.72$.

Fig. 3 B): UTR deletion size distribution at different generation. 100^{th}, 200^{th}, 300^{th}, $500^{th}$and 700^{th} generation from Fig.3 A was selected to further analyze their UTR region deletion situation. 100^{th}, 200^{th}, 300^{th}, $500^{th}$and 700^{th} generation was marked in red line, green line, blue line, black line and cyan line respectively.

Fig.4 A): Modeling of COVID-19 proliferation through time with the consideration that UTR deletion probability is its own length dependent. X axis is the generations which from 1 to 500 in

this stimulation. Y axis represents the virus population size which is described as the overall number of virus. The ancestor virus is 20 29903bp viruses with ***d*** = 0.72.

Fig. 4 B): UTR deletion size distribution at different generation. 50^{th}, 100^{th}, 200^{th}, $300^{th}$and 500^{th} generation from Fig.4 A was selected to further analyze their UTR region deletion situation. 50^{th}, 100^{th}, 200^{th}, $300^{th}$and 500^{th} generation was marked in red line, green line, blue line, black line and cyan line respectively.

Fig.5 A): Modeling of COVID-19 proliferation through time with different ancestor UTR length. X axis is the generations which from 1 to 100 in this stimulation. Y axis represents the virus population size which is described as the overall number of virus. Virus population alternation caused by single 29900bp virus proliferation is represented in blue line and Virus population alternation caused by single 29880bp virus proliferation is represented in red line.

Fig. 5 B): Modeling of COVID-19 proliferation through time with different ***d*** value. X axis is the generations which from 1 to 100 in this stimulation. Y axis represents the virus population size which is described as the overall number of virus. Virus population alternation caused by single 29900bp virus proliferation with a bigger d value (***d*** = 0.78) is represented in blue line and Virus population alternation caused by single 29900bp virus proliferation with a smaller d value (***d*** = 0.77) is represented in red line.

Fig. 5 C): Modeling of COVID-19 second wave after removing intervention policies. X axis is the generations which from 1 to 400 in this stimulation. Y axis represents the virus population size which is described as the overall number of virus. No intervention is applied during the early stage (***d*** = 0.74). Intervention is started at 100^{th} generation (***d*** =0.70) and is relaxed at 200^{th} generation (***d*** =0.73).

Fig.6 A): Modeling of individual infection with different immunity. X axis is maximal viral load per infection. Y axis represents count number with overall number equal to 5000. The light yellow color in fig. 6A represents the virus proliferation of people with strong immunity after being attacked by Covid-19, and the brown color represents the virus proliferation of people with weak immunity after being attacked by the virus.

Fig. 6 B): Modeling of individual infection with different invading virus amount and length. X axis is maximal viral load per infection. Y axis represents count number with overall number equal to 5000. The blue color in fig. 6B represents the virus proliferation in host with a single 29900bp virus invasion, and the brown color represents the virus proliferation in host with 20 29850bp viruses invasion.