

Statistical Analysis Supports UTR Deletion Theory in SARS-CoV-2

Zhaobin Xu (✉ zhaobin23@126.com)

Dezhou University

Yinghui Sun

Dezhou University

Dongying Yang

Dezhou University

Liyan Wang

Dezhou University

Research Article

Keywords: SARS-COV-2, deletion of the untranslated region, nucleic acid degradation system, viral toxicity

Posted Date: October 15th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-967894/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Statistical Analysis Supports UTR (untranslated region) Deletion Theory in SARS-CoV-2

Zhaobin Xu^{1*}, Yinghui Sun¹, Dongying Yang², Liyan Wang¹

1 Department of Life Science, Dezhou University, Shandong, P.R. China, 253023

2 Department of Medicine, Dezhou University, Shandong, P.R. China, 253023

* Corresponding address: zhaobin23@126.com;

Abstract

The statistical analysis found that the mortality rate of COVID-19 infection experienced a significant decline in the early stage of the epidemic. We suspect that the sharp deterioration of virus toxicity is related to point mutation and the deletion of the untranslated region of the virus genome. Through sequence analysis of mega-genome data, we found that the genome length of COVID-19 was deleted, which mainly occurred in the untranslated regions at both ends. Sequence similarity analysis further indicated that short UTR length strain emerged by deleting strain with long sequence length. This process is irreversible; the genome with a short sequence length could not restore to the long sequence length. By studying the relationship between genome length and mortality, we found a good correlation between them statistically, which demonstrated that the deletion of the untranslated region of the virus significantly affected the toxicity of the virus. We extracted the viral genome length of patients with different symptoms from the GISAID database for analysis to confirm this relationship. It discovered that the viral genome length of hospitalized patients was significantly more extensive than that of asymptomatic patients. In contrast, the viral genome length of asymptomatic patients was considerably longer than that of ordinary patients with symptoms. To further prove this idea, we performed a genome-level mutation scanning to systematically evaluate the influence of mutations at each position on virulence. After Pearson correlation analysis and chi-square investigation, UTR deletion was the primary driving force in alternating virus virulence. All those statistical evidence support the UTR deletion theory in SARS-COV-2. In the end, we proposed a mathematical model to explain why its UTR deletion was not continuous, which indicated humans could not eliminate SARS-COV-2 in a short time without robust intervention procedures.

Keywords: SARS-COV-2, deletion of the untranslated region, nucleic acid degradation system, viral toxicity

1 Introduction

Since the outbreak of COVID-19, it has caused a very high number of infections and deaths worldwide and has become the most prominent public

security crisis in the world since World War II [1-3]. Concerning COVID-19, we urgently need to answer a question, that is, why did it not disappear naturally like SARS and MERS but became a global pandemic disease? Is this merely due to its high reproduction capacity R_0 ? Answering this question helps to solve the problem itself and helps to guide our future countermeasures.

Before studying this problem, we first noticed that coronaviruses such as SARS and COVID-19 experienced a sharp decline in toxicity after the epidemic outbreak. To confirm this phenomenon, we chose the global mortality rate as the basis for analysis. This data has a sufficient sample size and high reliability in statistics. We can't explain this phenomenon with site point mutation theory based on the following four reasons. First, sars-cov-2 replication is very conservative, and the mutation rate is meager, so it is difficult to evolve a mutant with sharp virulence declination in a short time [4-6]. Even if this mutant evolved out after a short-time environment selection, this mutant could not evolve into a mainstream strain in a short time. Secondly, the statistical and epidemiological analysis did not establish a significant correlation between different variants and the degree of illness of patients. Although mutations can enhance their infectivity, there was no experimental and statistical evidence to show that the lethality of any strain has significantly increased or decreased [7-11]. Third, if the first and second reasons are wrong, that is to say, point mutation will cause a sharp decline in toxicity in a short time, then why is this downward trend not sustainable in COVID-19? After May 2020, the global mortality rate has remained relatively stable, and there has not been an apparent continuous decline. Fourthly, by comparing mortality in different countries, we can see that the influence of medical intervention on the mortality of virus-infected diseases is not apparent. We could notice that the mortality in many developed countries is significantly higher than that in developing countries [12-14]. For COVID-19, there is no specific drug so far, so the improvement of medical treatment can't play a leading role in the sharp reduction of mortality in early COVID-19. Therefore, considering all the above factors, we think that the significant decline in the mortality rate in the early stage of the COVID-19 epidemic is due to the decrease in virus toxicity.

We cannot simply interpret this declination by point mutation. There must be other reasons for the decline in virus toxicity. We believe that the toxicity of RNA viruses is closely related to the length of UTR at both ends. The untranslated regions of viruses generally exist as internal ribosome entry sites(IRES) during translation. Although the deletion of UTR regions of viruses will not affect the properties of encoded proteins, it will significantly affect their translation efficiency. This deletion would reduce virus replication efficiency and toxicity, which is elucidated in many viruses such as coxsackievirus[15-17] and HCV [18-20]. Although the function of the UTR region of coronavirus has been studied [21-22], and some scholars

have pointed out its essential role in virus replication [23], no one has found that UTR region deletion will naturally occur during SARS-CoV-2 infection. This deletion may correspond to lower virus activity.

Therefore, we analyzed and demonstrated bioinformatics and statistics from the following directions. First of all, we confirmed that the genome length of COVID-19 managed a significant decline in the early stage of the epidemic. Then, employing sequence alignment, we prove that there is an evolutionary relationship between virus genomes. Viruses with long UTR fragments can evolve into viruses with short UTR fragments. There is an irreversible parent-child relationship between them, rather than a sibling relationship with parallel evolution. In the third part of the results, through studying the relationship between global mortality and viral genome length, we found a strong positive correlation between them. In the fourth part of the results, we selected the virus sequence length in patients with different symptoms for analysis to further confirm this viewpoint. We eventually confirmed that there was a significant statistical correlation between the sequence length and symptoms. In the fifth part of the results, we performed a genome-level mutation scanning to investigate each point mutation's influence on virulence systematically. The fifth part of the results further demonstrated the importance of UTR on COVID-19 virulence. In the sixth part of the results, utilizing mathematical simulation, we further studied why this deletion is unsustainable at the population level. It could well explain why COVID-19's genome did not undergo continuous deletion to a complete extinction like SARS and MERS but could exist in a relatively stable length plateau for quite a long time.

2 Results

2.1 Investigate the Genome Deletion of COVID-19

The genome data of COVID-19 has been increasing with time. As of July 6, 2021, more than 2.23 million COVID-19 genomes information have been uploaded to the GISAID database. We extracted the basic information from the GISAID database. After filtered out the samples with uncomplete genomes, its genome length distribution with time was shown from Figure 1A to Figure 1U

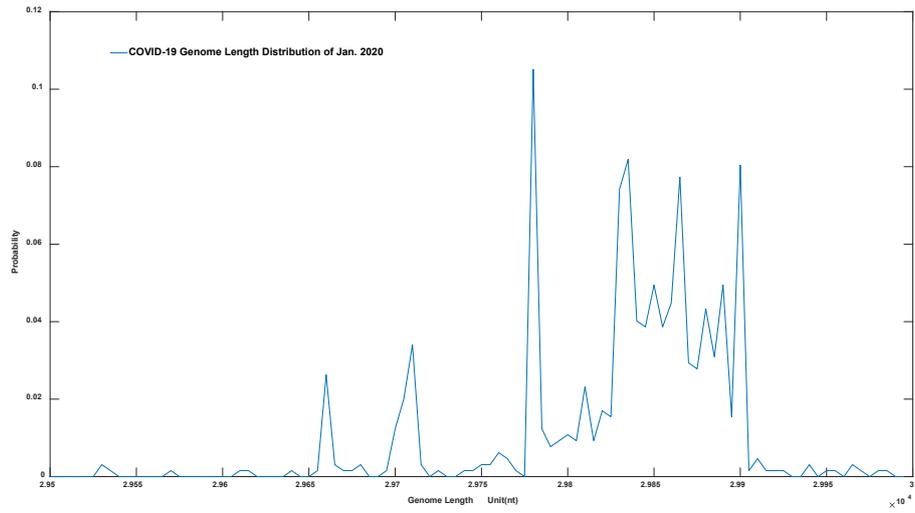


Figure1.A: COVID-19 Genome Length Distribution of Jan. 2020

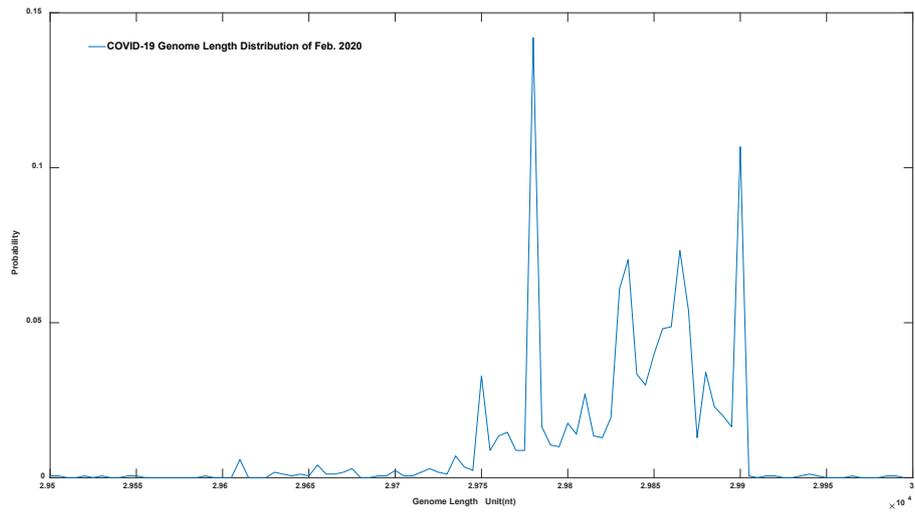


Figure1.B: COVID-19 Genome Length Distribution of Feb. 2020

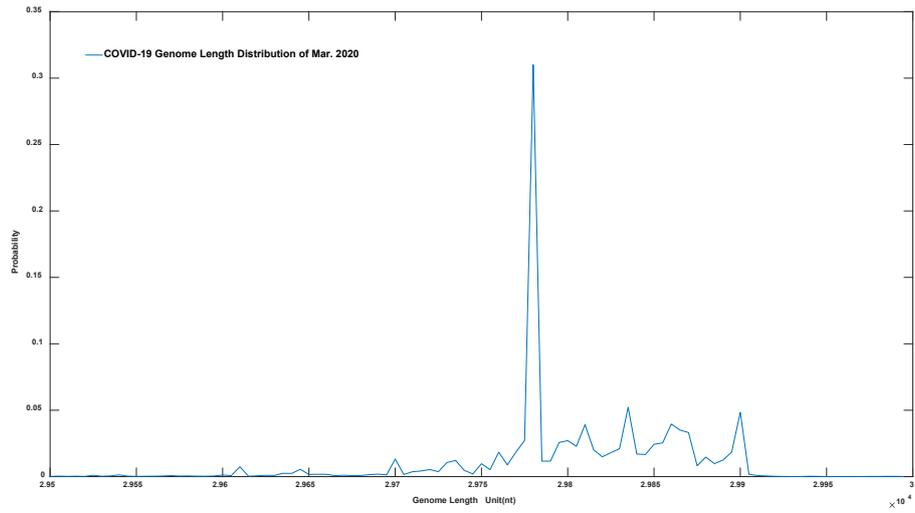


Figure1.C: COVID-19 Genome Length Distribution of Mar. 2020

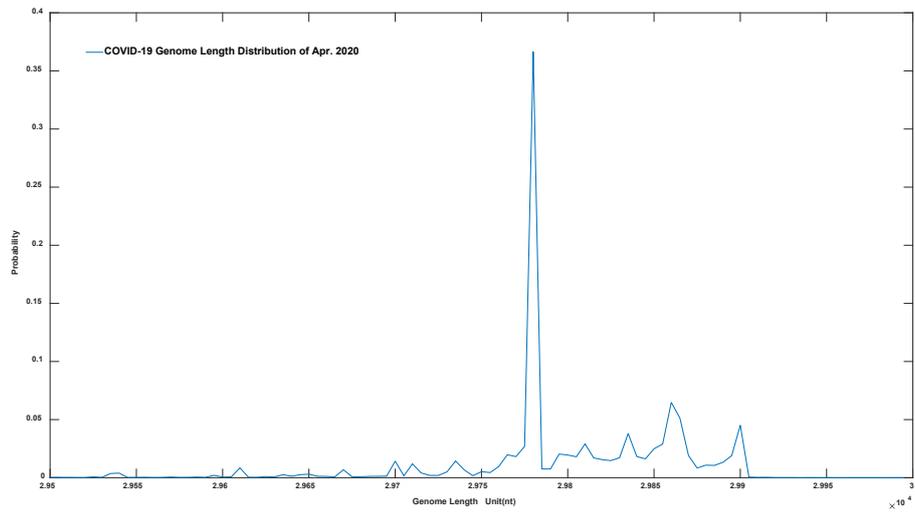


Figure1.D: COVID-19 Genome Length Distribution of Apr. 2020

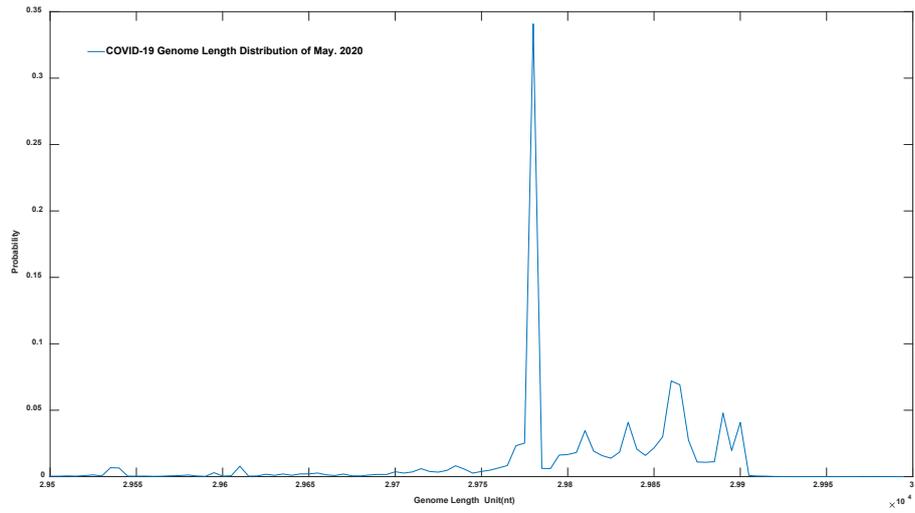


Figure 1.E: COVID-19 Genome Length Distribution of May. 2020

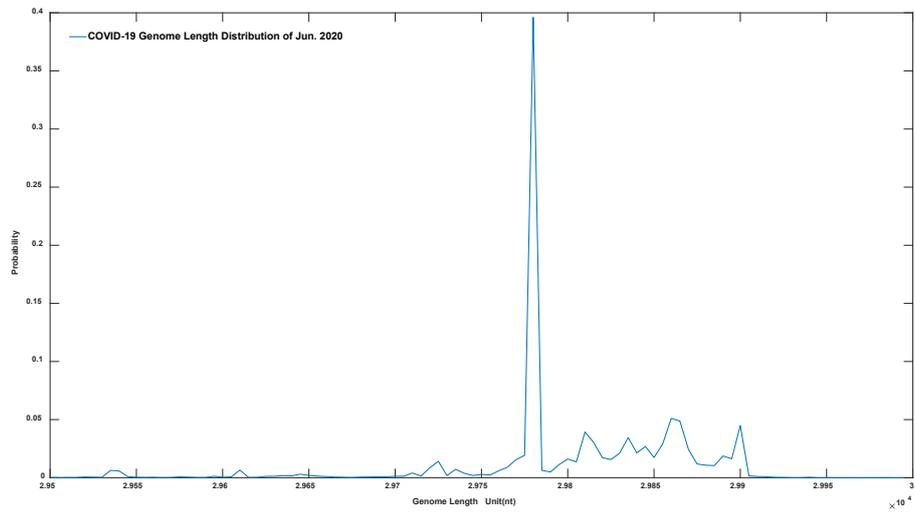


Figure 1.F: COVID-19 Genome Length Distribution of Jun. 2020

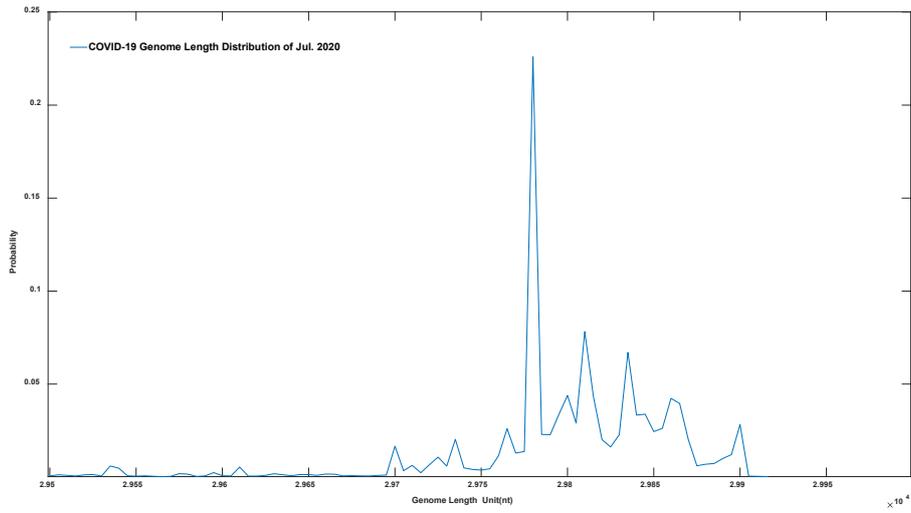


Figure 1.G: COVID-19 Genome Length Distribution of Jul. 2020

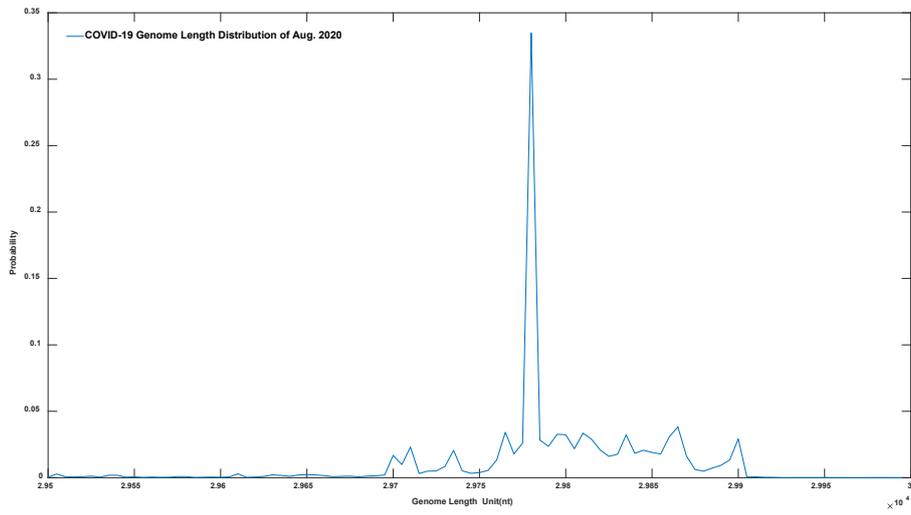


Figure 1.H: COVID-19 Genome Length Distribution of Aug. 2020

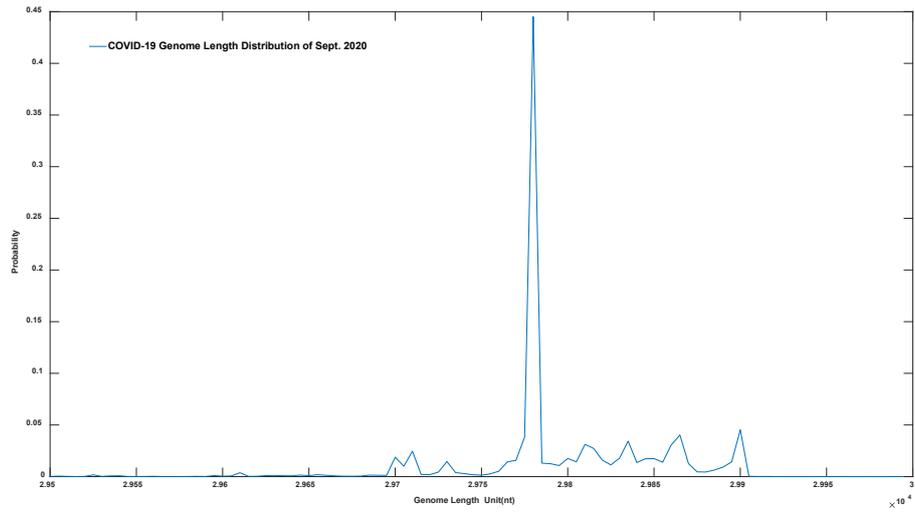


Figure 1.I: COVID-19 Genome Length Distribution of Sept. 2020

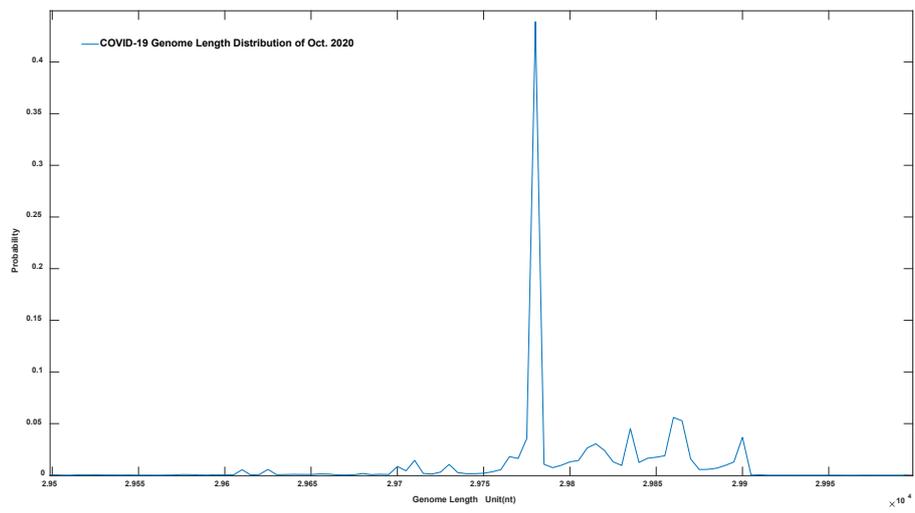


Figure 1.J: COVID-19 Genome Length Distribution of Oct. 2020

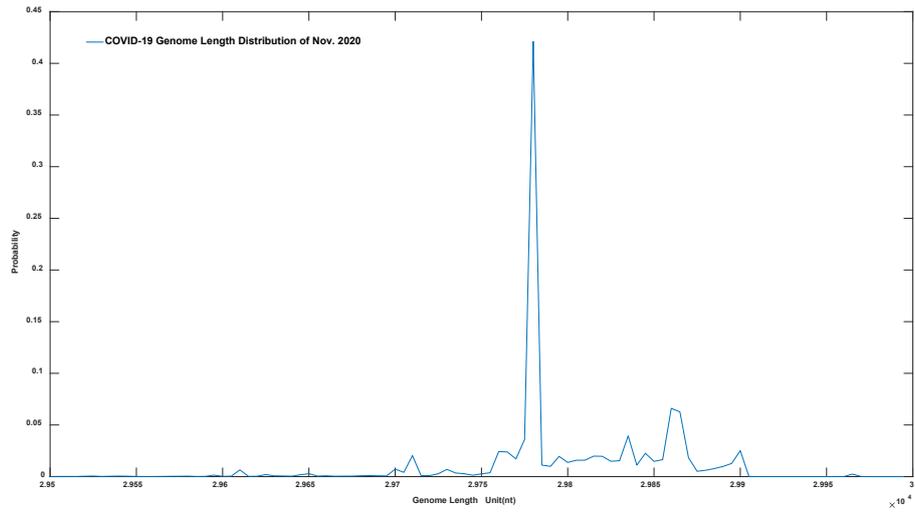


Figure 1.K: COVID-19 Genome Length Distribution of Nov. 2020

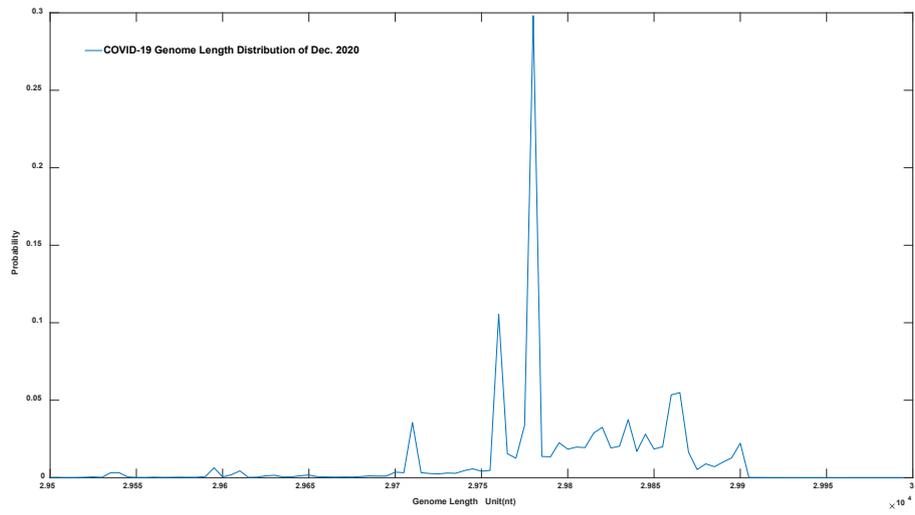


Figure 1.L: COVID-19 Genome Length Distribution of Dec. 2020

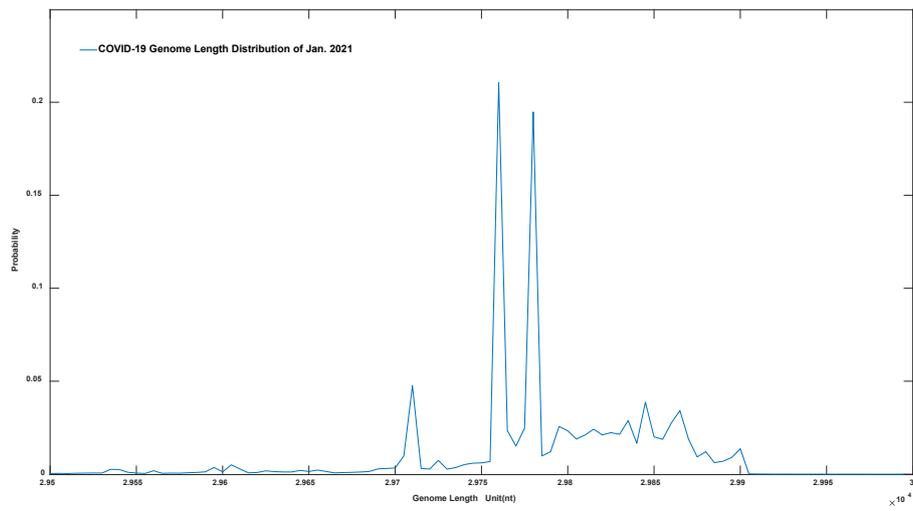


Figure 1.M: COVID-19 Genome Length Distribution of Jan. 2021

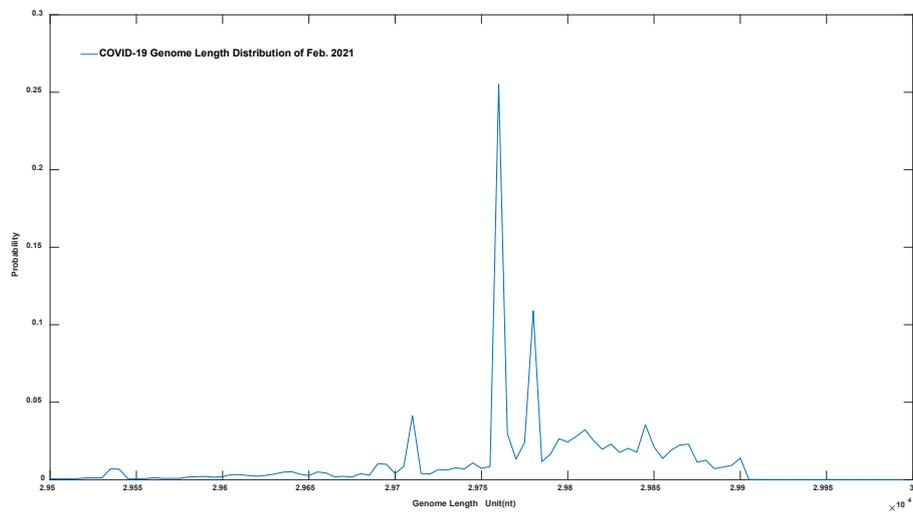


Figure 1.N: COVID-19 Genome Length Distribution of Feb. 2021

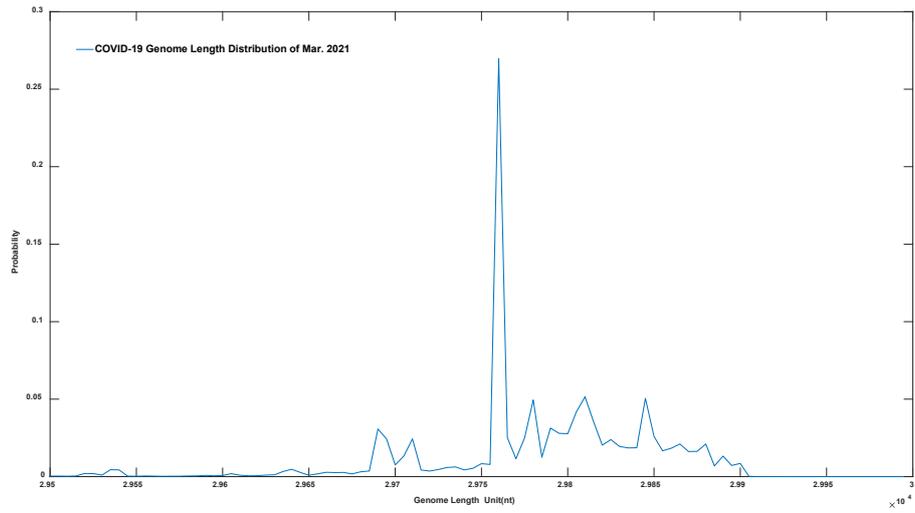


Figure 1.O: COVID-19 Genome Length Distribution of Mar. 2021

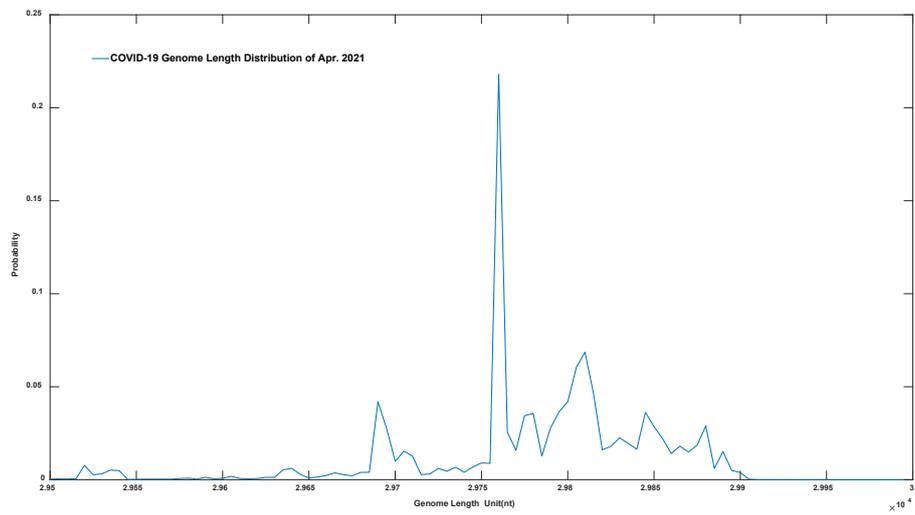
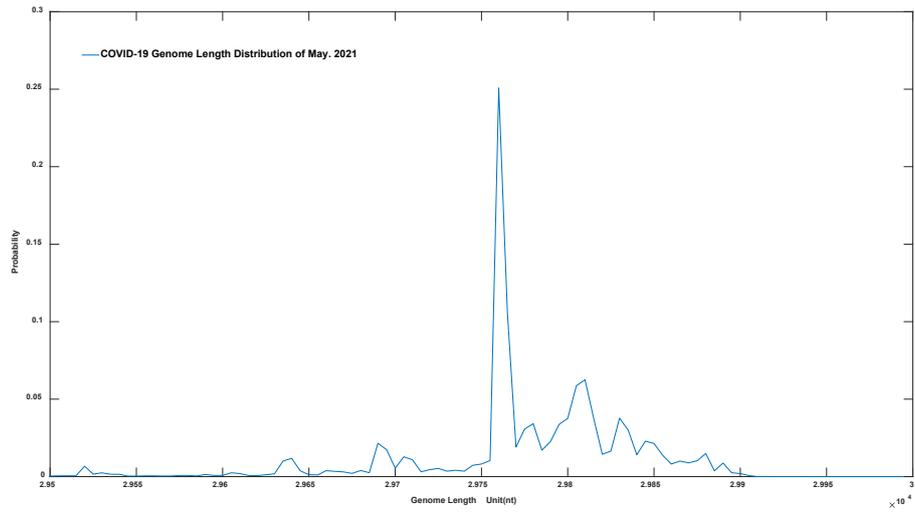
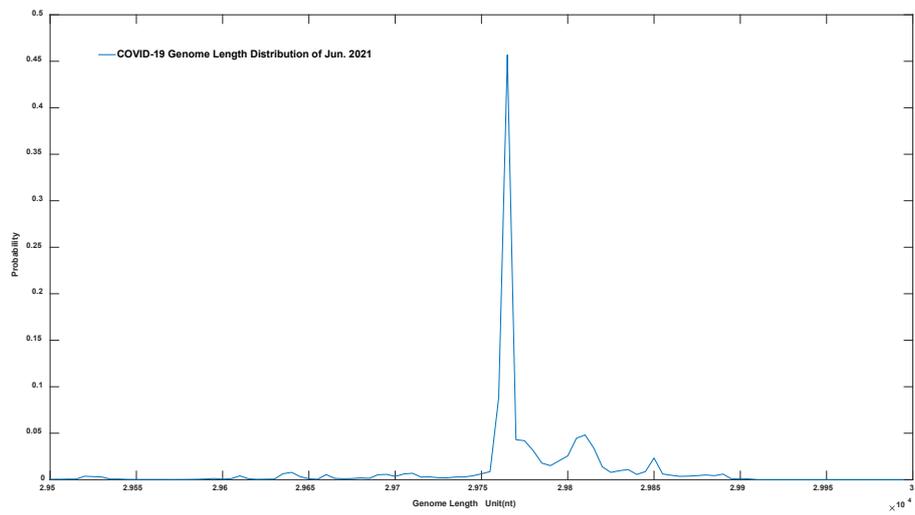


Figure 1.P: COVID-19 Genome Length Distribution of Apr. 2021



: COVID-19 Genome Length Distribution of May, 2021



: COVID-19 Genome Length Distribution of Jun, 2021

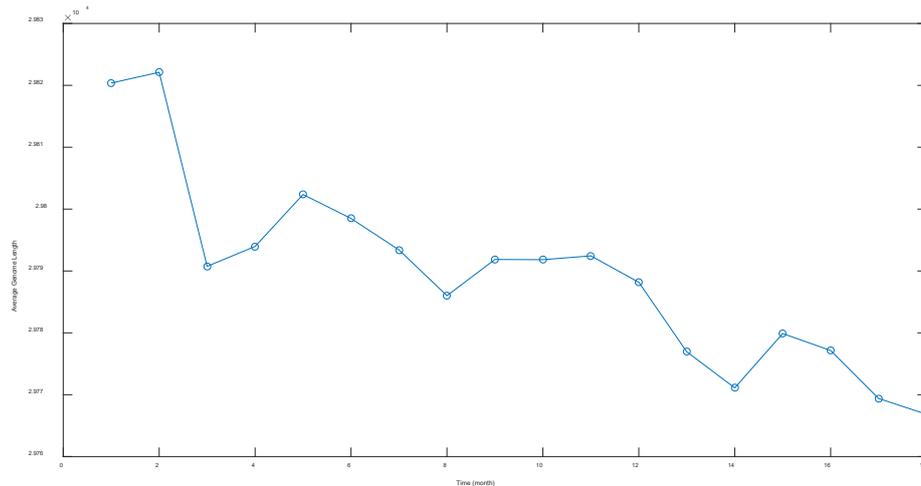


Figure 1.S: COVID-19 average genome length at different time points. 1st month starts from Jan.2020.

For the sake of clarity, we select 5nt as the group distance interval, and its genome size distribution is shown from Fig. 1A to Fig. 1R. It can be seen that virus genomes at different sample collection months have been deleted to varying degrees. Interestingly, this kind of deletion is undeniable in the initial stage of the epidemic. With the extension of time, the deletion rate of the virus genome drops significantly, and sometimes there is a weak rebound, making it reach a relatively equilibrium state around 29800nt. This declination can be more intuitively reflected in Figure 1.S. It is worth noting that the distribution of viral genome length is neither a normal distribution nor an average distribution. There are differences in the probability at different length points, and the likelihood of some length loci is significantly higher than that of other length loci. Generally speaking, the length of the virus genome will form two prominent high-frequency distributions in the vicinity of 29900nt and 29780nt. It should be pointed out that Figure 1.S can't fully reflect the change of virus genome length but can only reflect the changing trend of an average value. We roughly use this data to calculate the correlation coefficient with mortality in Section 2.3. Fig. 1A-R shows the distribution and change process of genome length of virus population more accurately. After July 2020, the proportion of long genome viruses near 29900nt no longer shows an apparent downward trend, although the average genome length still shows a slow gradient downward trend.

2.2 Bioinformatics Analysis Indicate that a UTR Deletion in COVID-19

In part 2.1, we show that the genome length of the COVID-19 population is changing with time. However, we can't directly prove the deletion of its untranslated region because there may be a possibility that viruses of different lengths may evolve independently. The proportion of viruses of varying genome lengths in the whole population might change with time. Viruses with short genome length might not be offspring of long UTR region length strains

but gradually occupy a dominant position in evolution, increasing their proportion. To verify the deletion of the UTR region in COVID-19, we randomly selected the sequencing results before 15th-March, 2020 in the NCBI VIRUS database. The reason to choose the early sequencing results is that the earlier data contains a higher proportion of viruses with long UTR lengths. We finally selected 178 29903nt sequences and 108 29782nt short sequences to study their evolutionary relationship. All those sequences have high sequence quality. We evaluated the evolutionary relationship between them through sequence similarity analysis. If there is a parallel evolutionary relationship between them, the sequence homology will be significantly higher than that between them. Suppose the long UTR virus is deleted to produce the short UTR virus. In that case, the sequence similarity between 29903nt is the highest. The sequence similarity between them is the second, while the sequence similarity of the short genome virus with the length of 29782nt is the lowest. If the virus with a short UTR length produces the virus with a long UTR length, the sequence homology of 29780nt length virus is the highest, and the sequence similarity between them is the middle, while the sequence similarity within the long genome virus with 29903nt is the lowest. The details are described in section 3.2. The results are shown in Table 1.

Pair-wise sequence mutation score	Among 29903nt	Between 29903nt and 29782nt	Among 29782nt
Mean Value	8.4651	10.0457	10.9576
Standard Deviation	4.8509	5.5880	6.3711
Max Value	40	54	52
Min Value	0	0	0
Sample Size	15753	19224	5778

Table 1: Statistical characteristics of mutation score of two different length groups.

It can be seen from Table 1 that the sequence similarity within 29903nt is significantly higher than that between them (Kolmogorov-Smirnov test $p = 1.07 \times 10^{-135}$), while the sequence similarity between them is considerably higher than that within short sequences (Kolmogorov-Smirnov test $p = 8.6 \times 10^{-22}$). Therefore, we can prove that the virus with a short sequence length originates from the deletion of the long sequence virus, and this process is irreversible. This phenomenon is difficult to find and confirm through wet experiments.

2.3 Statistical Analysis of COVID-19 Mortality and UTR Length Suggest a

Strong Correlation Exists between UTR Length and COVID-19 Toxicity

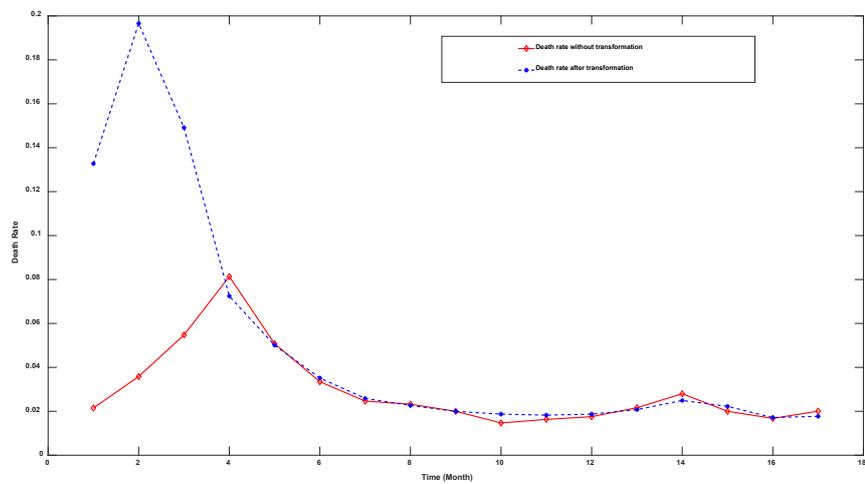


Figure2: COVID-19 mortality calculated using two different approaches. The red line stands for the mortality using death data explicitly. The blue

stripe stands for the mortality calculated after transformation.

The variation trend of genome length in COVID-19 is shown in Fig. 1S. Pearson correlation coefficients between the two mortality curves were calculated to be 0.2501 and 0.7293, respectively, indicating that the correlation between COVID-19 genome length and its actual mortality reached a significant level. Through this study, we can judge a significant positive correlation between the genome length of COVID-19 and its toxicity statistically. The correlation between the transformed accurate mortality data and the ratio of high-length virus strains is further studied, and the results are shown in Table 2 below:

Threshold Set	Pearson Correlation factor
>=29850nt	0.8191
>=29855nt	0.8082
>=29860nt	0.7980
>=29865nt	0.8371
>=29870nt	0.8508
>=29875nt	0.8217
>=29880nt	0.8351

Table 2: Pearson correlation between genome length and death rate at different threshold sets.

It can be seen from Table 2 that there is a better positive correlation between the ratio of high-length strains and the mortality rate of COVID-19. When the threshold value is taken as 29870nt, the Pearson correlation coefficient can exceed 0.85. From a statistical point of view, it can be further explained that there is indeed a significant positive correlation between its genome length and its toxicity.

2.4 Statistical Analysis of The Correlation Between Symptoms of Patients and UTR Length

In the above three parts, we proved that the UTR region in COVID-19 is indeed deleted, and there is an apparent statistical correlation between its UTR length and its virus toxicity. To further establish the relationship between UTR length and virus toxicity, we selected patients with different clinical

symptoms for research. The specific method is described in detail in 3.3. In brief, we extracted the data of strains containing patient information from the GISAID database, and we divided the patients into three categories: asymptomatic infected persons, symptomatic infected persons, and hospitalized patients. The genome length distribution of COVID-19 in these three types of patients is shown in Figure 3:

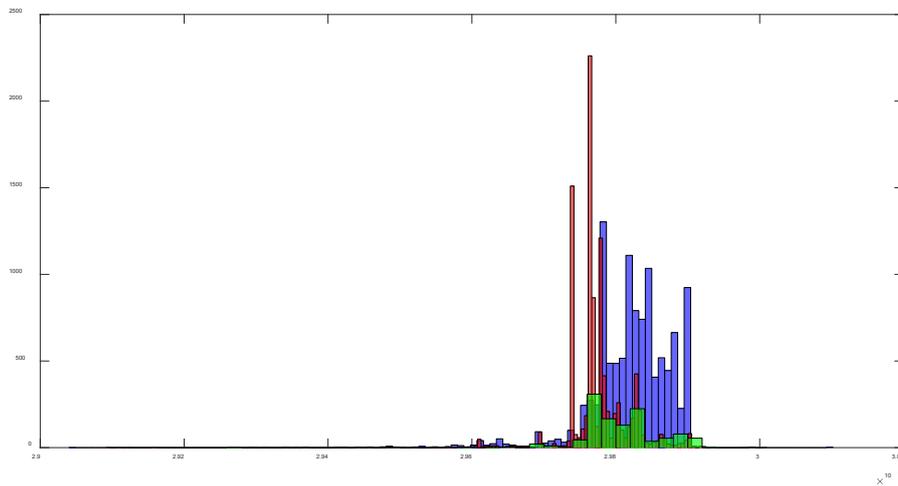


Figure3: The genome length distribution of COVID-19 in three different types of patients. The red box stands for symptomatic patients; The blue one stands for hospitalized patients; the green one stands for asymptomatic patients.

The statistical Kolmogorov-Smirnov test P-value within and between their populations is shown in Table 3.

<i>P</i>	Virus genome length in hospitalized patient	Virus genome length in asymptomatic patient	Virus genome length in symptomatic patient
Virus genome length in hospitalized patient	0.5311	6.2923e-41	0
Virus genome length in asymptomatic patient	6.2923e-41	0.5395	6.3969e-117
Virus genome length in symptomatic patient	0	6.3969e-117	0.8739

Table 3: Heterogeneity test of COVID-19 genome length among different symptom patients.

As shown from Figure 3, the genome length of COVID-19 in hospitalized patients is generally longer, while the genome length of COVID-19 in

commonly infected patients with symptoms is the shortest, and that in asymptomatic infected patients is in the middle. Table 3 can better reflect the statistically significant differences among the three distributions. To prove that the randomness of sampling does not cause this difference, we also carried out the Kolmogorov-Smirnov test after random sample within the population [27], which shows that the genome length within the population is not heterogeneous. However, the size of the virus in patients with different symptoms is significantly heterogeneous. We put forward the following theory to explain this statistically significant difference. First, hospitalized patients tend to be infected with highly toxic COVID-19, which corresponds to a more extended UTR region, a longer genome length. Asymptomatic infected people should be inclined to be infected with the weakest virus, that is, the virus with the shortest genome length, but the actual situation is not the case. Lots of asymptomatic infected people in the database are the people who display symptoms in the later infection phase. Therefore, in most cases, samples from asymptomatic infected people were taken at the initial phase of infection. At that time, the host nucleic acid degradation system had not seriously eroded the virus genome. Therefore, its genome length is longer than that of viruses sampled in the middle and late stages of infection. This interference could explain why the average size of viruses in asymptomatic infected people is significantly longer than that in symptomatic infected people.

2.5 Mutation scanning analysis at genome level indicates UTR deletion is the major driving force in COVID-19 virulence evolution

Although many studies have previously investigated the effect of genetic mutations on viral virulence by statistical means, most of these studies are one-sided and unsystematic [38-43]. For example, Nagy et al. [38] identified 15 mutations that enhance viral virulence, including D614G, P323L, etc... Through a chi-square analysis of the genomes of patients with different symptom statuses, many in-depth experimental studies were subsequently conducted on the D614G mutation. However, although experimental studies [44-46] demonstrated that the 614G mutation could significantly enhance upper respiratory virus load, it was shown that this mutation does not cause significant changes in virulence. There are three reasons for the inaccuracy of those statistical prediction methods. First is the limitation of sampling size. The second reason is the biased sampling procedure, and the third reason is that the correlation of a single indicator cannot establish a causal relationship favorably. The first point is well understood, and with the current expansion of the sample size, we can overcome the first defect. The second point refers to the unevenness of patient information collected at different periods. For

example, the over-recruitment of genomes in deceased patients in a given period and the proportion of death rate in genome databank significantly exceeds the overall mortality rate in the general population. This unevenness would lead to incorrect conclusions using statistical analysis. For example, in the research done by Nagy et al. [38] mentioned above, their sampling deadline was before September 2020. If the time point were shifted back and analyzed using the current extended patient database, then most of their toxicity enhancement predictions would be wrong. The third reason is also an essential factor in generating erroneous judgments. So in our study below, we use the correlation of multiple indicators to evaluate the association between mutation and toxicity. We only consider a strong association between the two if all these indicators are more significant than a threshold value.

It was necessary to scan all loci at the meta-genome level to systematically evaluate the effect of mutations at different loci on virulence to verify that deletions in the UTR region are essential in influencing changes in virulence of SARS-CoV-2. We used wuhan03 as the original strains. Virus sequences with patient information in the GISAID database before September 1, 2021 (90,000 sequences) were extracted to study their mutation situation at each locus. We applied three indicators in our study.

First: the mutation frequency of individual bases, the vast majority of these 29903 bases did not undergo severe mutations and were found to be highly conserved after sequence alignment. However, low mutation frequency also causes a higher Pearson correlation on mortality. Therefore, we need to integrate the first indicator, which means we only need to study the effect of loci with higher mutation frequencies on virulence.

Second: the temporal Pearson correlation between the frequency of mutations in genetic variation and mortality. We can first obtain the mutation frequency of all bases at all loci in different months by sequence comparison. After that, we performed a correlation analysis to obtain the Pearson correlation coefficient between single mutation and mortality rate. The Pearson correlation factor is positive for mutations that attenuate virulence and negative for loci that enhance virulence. The larger the absolute value, the stronger the correlation.

Third: Significance of the ratio between deceased patients and asymptomatic patients. We extracted information from 90,000 patients to identify 2468 dead cases and 1386 asymptomatic infected patients. We used chi-square analysis to determine whether a single base mutation could cause a significant change in the proportion between those two. For mutations that attenuate virulence, it will cause a substantial decrease in the ratio between deceased and asymptomatic patients, the significance of which can be found by chi-square analysis. For mutations that enhance viral virulence, mutating a single base causes a significant increase in that proportion.

The whole-genome scanning results are shown in supplementary materials.

Combining the three factors, we set the threshold for the mutation frequency

to greater than 20% (equal to 0.8 conversation score), the Pearson correlation threshold to greater than 0.2, the chi-square significance to positive, and the chi-square significance threshold to 0.01. The base loci screened out were positively correlated with viral virulence significance.

Positions that are filtered out using Pearson correlation threshold and mutation threshold (Pearson > 0.2, mutation rate > 0.2) are shown in Fig 4A.

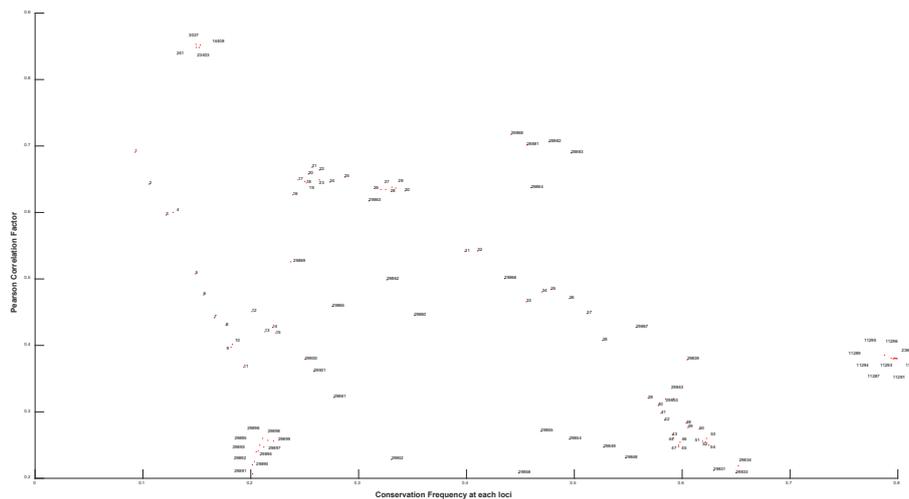


Figure 4.A: Positions that are filtered out using Pearson correlation threshold and mutation threshold (Pearson > 0.2, mutation rate > 0.2)

Positions that are filtered out using chi-square test and mutation threshold (P < 0.01, mutation rate > 0.2) are shown in Fig 4B

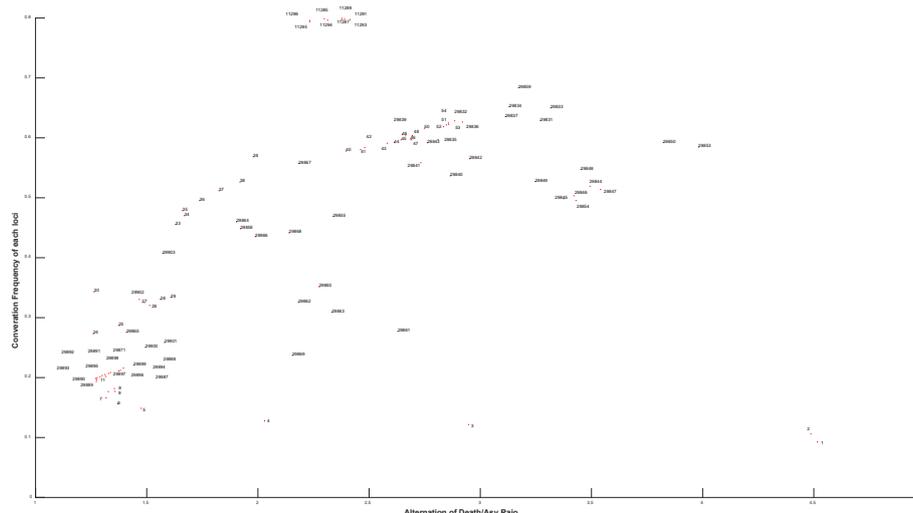


Figure 4.B: Positions that are filtered out using chi-square test and mutation threshold (P < 0.01, mutation rate > 0.2)

Table 4: Locus that meet all of those three thresholds

position	Conservation score	Pearson correlation	P-value using the chi-square test with two extreme symptom groups	Is UTR or not
1	0.0927	0.6907	0	Y
2	0.1059	0.6425	0	Y
3	0.1216	0.596	0	Y
4	0.1278	0.6001	0	Y
5	0.1488	0.5076	0.0004	Y
6	0.1563	0.476	0.0029	Y
7	0.1662	0.4416	0.0065	Y
8	0.177	0.4307	0.0018	Y
9	0.1817	0.3974	0.0016	Y
11	0.1935	0.3672	0.0095	Y
24	0.2732	0.6458	0.0044	Y
25	0.2871	0.653	0.0001	Y
26	0.3205	0.6346	0	Y
27	0.3251	0.6346	0	Y
28	0.3308	0.6381	0	Y
29	0.3345	0.6367	0	Y
30	0.343	0.6325	0.0008	Y
33	0.4557	0.4655	0	Y
34	0.4702	0.4803	0	Y
35	0.4784	0.4845	0	Y
36	0.4955	0.4704	0	Y
37	0.5118	0.4479	0	Y
38	0.5261	0.4075	0	Y
39	0.5685	0.3208	0	Y
40	0.5778	0.3094	0	Y
41	0.5804	0.2977	0	Y
42	0.5838	0.2875	0	Y
43	0.5909	0.2652	0	Y
44	0.5922	0.2603	0	Y
45	0.5968	0.2476	0	Y
46	0.5981	0.2543	0	Y
47	0.5968	0.2499	0	Y
48	0.6038	0.2831	0	Y
49	0.6055	0.2763	0	Y
50	0.616	0.2743	0	Y

51	0.6188	0.2557	0	Y
52	0.6215	0.2546	0	Y
53	0.6226	0.26	0	Y
54	0.6248	0.2498	0	Y
11285	0.7988	0.3803	0	N
11287	0.7994	0.3802	0	N
11289	0.7974	0.3805	0	N
11291	0.7967	0.3811	0	N
11293	0.7982	0.3802	0	N
11294	0.7964	0.3802	0	N
11295	0.7957	0.3798	0	N
11296	0.7938	0.3806	0	N
29831	0.6291	0.2119	0	Y
29833	0.6499	0.2084	0	Y
29834	0.6521	0.2188	0	Y
29839	0.6047	0.3783	0	Y
29843	0.592	0.3198	0	Y
29848	0.5475	0.2309	0	Y
29849	0.5271	0.2467	0	Y
29853	0.5846	0.3191	0	Y
29854	0.4954	0.2589	0	Y
29855	0.4692	0.2707	0	Y
29858	0.4484	0.2083	0	Y
29860	0.3515	0.4457	0	Y
29861	0.2771	0.3218	0	Y
29862	0.3262	0.4985	0	Y
29863	0.3095	0.6174	0	Y
29864	0.46	0.6373	0	Y
29865	0.2757	0.459	0	Y
29866	0.4353	0.5005	0	Y
29867	0.5573	0.4274	0	Y
29868	0.4415	0.7173	0	Y
29869	0.2372	0.5257	0	Y
29890	0.2015	0.2203	0.0044	Y
29891	0.2017	0.2067	0.002	Y
29892	0.2034	0.2253	0.0032	Y
29893	0.205	0.2399	0.0022	Y
29894	0.2069	0.2409	0.0013	Y
29895	0.2084	0.25	0.001	Y
29896	0.211	0.2599	0.0003	Y
29897	0.2121	0.2469	0.0002	Y
29898	0.2159	0.2571	0.0001	Y

29899	0.2211	0.256	0	Y
29900	0.2504	0.3796	0	Y
29901	0.2585	0.3614	0	Y
29902	0.3304	0.2282	0	Y

Combining the three factors, we set the threshold for the mutation frequency to greater than 20% (equal to 0.8 conversation score), the Pearson correlation threshold to greater than 0.2, the chi-square significance to positive, and the chi-square significance threshold to 0.01. we collected those loci by filtering those positively correlated with viral virulence significance. The results are shown in Table 4.

These bases were all in the UTR region except for the 11285 11287 11289 11291 11293 11294 11295 11296 loci. The dominant mutation types of those UTR loci were deletions. However, after carefully checking those non-UTR positions, we further filter those eight positions out since those mutations start to emerge after Nov 2020. Those mutations contribute to the alpha variant firstly identified in England. Therefore, those mutations are not correlated with fatality in the first 10-month epidemic.

Similarly, we set the threshold for the mutation frequency of bases to greater than 20%, the Pearson correlation threshold to less than -0.2, the chi-square significant to negative, and the correlation threshold to 0.05. We did not find any loci that were significantly negatively correlated with viral virulence. That is, no mutations at any sites were found to increase viral virulence substantially.

Therefore, we can boldly conclude that the deletion of the UTR region dominates the change in virulence of COVID-19 and is the main driving force of viral virulence.

2.6 A Mathematical Model to Explain the Unsustainability of UTR Deletion in COVID-19 Population

From the first part, we can see that COVID-19's UTR region experienced a significant reduction in the early stage. Still, this reduction was unsustainable and entered a relatively long platform period in the following period. The speed of deletion is greatly truncated, and sometimes there is a weak rebound phenomenon. This UTR deletion theory can also explain why COVID-19 did not die naturally like other coronaviruses such as SARS and MERS but developed into a global epidemic of respiratory infectious disease. In this part, we try to explain the various characteristics of UTR region length in COVID-19 using the mathematical model. We believe that the degradation probability of viral RNA at both ends caused by the host nucleic acid degradation system is related to its RNA secondary structure [28,29,47]. This hypothesis can also explain why the gene length distribution of the COVID-19 population does not conform to the characteristics of orthographic distribution or random

distribution but has significant high-frequency features at some specific length loci. We think that when the node forming a stable secondary structure in the UTR region is the bottleneck site of degradation, the degradation rate will be significantly lower than that of straight-chain RNA near this site. Therefore, these secondary structures help enhance the binding ability with ribosomes and play the ribosome entry site, which is more conducive to preventing viruses from eroding by the host nucleic acid degradation system. Because the host's nucleic acid degradation system is a part of its immunity and an essential part of population protection, the average UTR length of the virus population will be continuously deleted. Finally, the toxicity is fundamentally eliminated. However, this situation cannot be applied to all scenarios because the replication ability of the virus will decline after deletion, so it is not an inevitable event whether the gene length of the virus population will continuously reduce due to the existence of a degradation mechanism. Unfortunately, the UTR region in COVID-19 has a very stable secondary structure at some sites, which significantly reduces the deletion probability at these sites. In addition, the reduced replication ability of the deleted virus has led to a continuous decline in the overall genome length of COVID-19 and even a slight increase in specific periods.

To better study the virus increment and diffusion process, we further model virus population dynamics at 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 have been deleted with different degrees. The reasons for deletion may be the host nucleic acid degradation system's corrosion, poor conservation of its 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 a mean value of 0 and a standard deviation of θ , which is described in equation (1) in the materials and methods section. The probability of a small number of base deletions is higher than that of a large number of base deletions. 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 UTR at both ends, the replication efficiency of the deleted short-length virus will be affected. The shorter the genome is, the lower the replication efficiency and the longer the replication cycle would be. There is a mathematical relationship between the replication cycle and the deletion length, so we establish a mathematical function between the deletion length and the variation of the replication period. This relationship is described in equation (2) in the materials and methods section. The details are described in the materials and methods section.

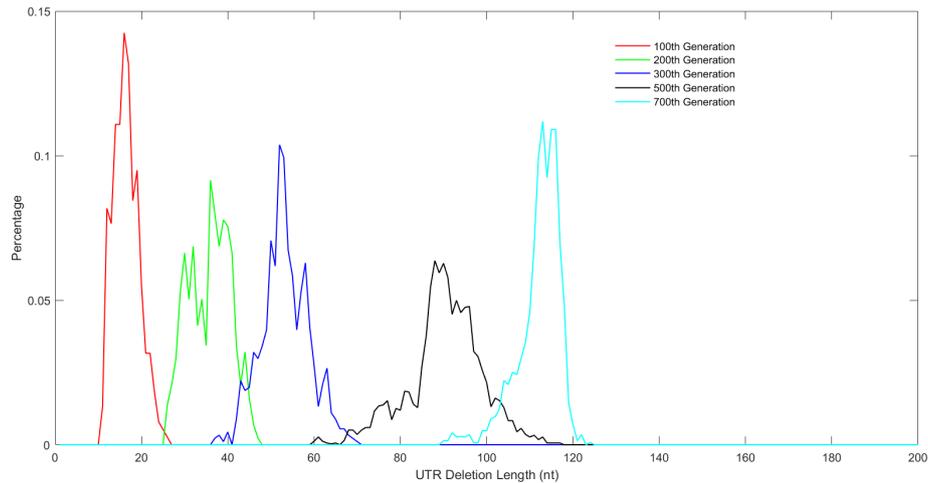


Figure 5.A: UTR deletion size distribution at different generations based on undifferentiated attenuation model. 100th, 200th, 300th, 500th and 700th generation was selected to analyze their UTR region deletion degree further. 100th, 200th, 300th, 500th and 700th generation was marked in the red line, green line, blue line, black line, and cyan line, respectively.

As shown from Fig.5A, if the model assumes that the virus genome will undergo undifferentiated attenuation with time, the viral genomes in different epidemic periods will be linearly deleted with time. However, the sequencing results imply that although the genome of the virus displayed a deletion trend in the first few months of the epidemic, it reached homeostasis in the last few months of 2020. The ratio of virus length distribution at 29782nt gradually increased but did not form a prominent peak after 29782nt, indicating that the virus length can maintain a steady state around 29782nt. This also suggests that the deletion probability of the virus is related to the current UTR length. In some peak regions such as 29903nt and 29782nt, the deletion possibility in the UTR region is significantly lower than that in other sizes.

Considering the relationship between UTR deletion and its length, we construct a simplified model. We only consider the deletion probability difference at 29903nt and 29782nt. The frequency of UTR deletion is divided into three parts:

The possibility of deletion at 29903nt

The possibility of deletion at 29782nt

The possibility of deletion in other regions

UTR base deletion probability will be more complicated because apart from 29903nt and 29782nt, the genome maps also have small peaks at other lengths. For the sake of simplicity, our model does not consider this.

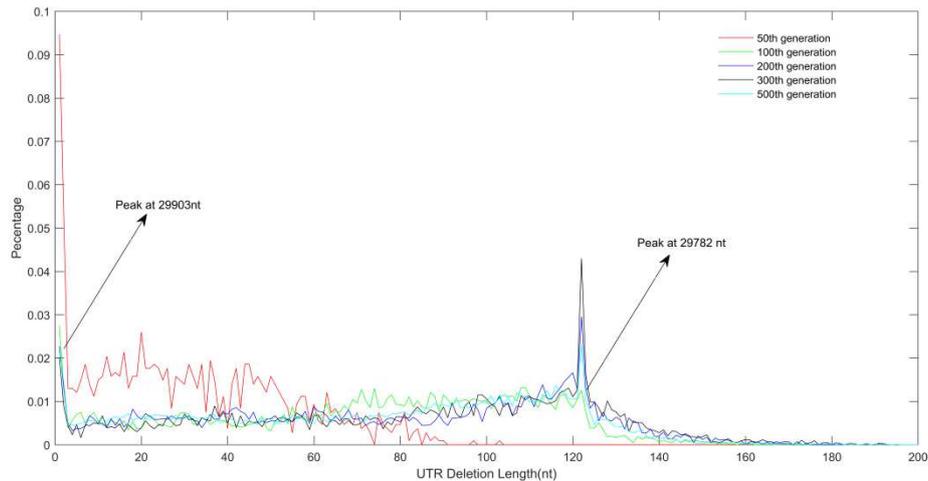


Figure 5.B: UTR deletion size distribution at different generations considering reduced deletion probability at certain bottleneck points. 50th, 100th, 200th, 300th and 500th generation was selected to further analyze their UTR region deletion situation. 50th, 100th, 200th, 300th and 500th generation was marked in the red line, green line, blue line, black line, and cyan line, respectively.

It can be seen from Fig.5B that when the UTR deletion probability is affected by the UTR length, the length of the virus genome will decline with time in the early stage, but it will not decline all the time. After the epidemic develops into a particular scene, the length distribution of its genome reaches an equilibrium state, which is consistent with the current sequencing results. At the same time, its genome will show the characteristics of high-frequency distribution at a specific length, as shown in Fig.1A-R. The size of the UTR region is closely related to virus activity, and the UTR region is relatively long in the early stage of the epidemic; the virus replication cycle is short, which corresponds with a high mortality rate. With the development of time, most of SARS-CoV-2's genomes are deleted to 29782nt, and the lost UTR will affect the virus proliferation efficiency so that the mortality rate will drop obviously. However, because the attenuation of the virus UTR region will be affected by the current UTR length, UTR might form a stable secondary structure at a specific node point, such as 29782nt length. The existence of a loop structure may effectively prevent further corrosion from the 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 the epidemic will not decrease all the time. Therefore, we believe that in general, the dominant factor for the change of virus toxicity is the deletion of UTR, not site mutation. Moreover, the length of the UTR region will not show infinite attenuation all the time, so SARS-CoV-2's virulence will maintain at a particular range. SARS-CoV-2 cannot die out naturally in a short time before reaching the community immunity threshold.

3 Materials and methods

3.1 Database Selection

We mainly selected the NCBI VIRUS database [30] and the GISAID database [31] for data analysis. GISAID database contains more information about strains, but it is not easy to extract them, so we directly selected fixed-length sequences from the NCBI database for extraction during sequence alignment. We selected the GISAID database for the change of virus genome length and extraction of patient information. As of July 6, 2021, the database we used contained 2.23 million pieces of COVID-19 data, including virus identity, sample collection location, sample collection date, genome length, genome integrity, and so on. First, we filter out the virus information with an incomplete genome and then draw Figure 1a and Figure 1b according to the sampling date. The specific data information and MATLAB codes are provided in the supplementary materials.

3.2 Sequence Alignment Approach

In the sequence similarity analysis, we selected two virus genomes with different lengths from the NCBI virus database before March 15, 2020, for sequence alignment analysis, which were 178 of 29903nt and 108 of 29782nt, respectively. To prevent alignment errors caused by UTR length differences, we uniformly cut the sequences into sequences of the same length. That is to say, the 5'UTR and 3' UTR of two virus strains with different lengths were cut to same length simultaneously. A sequence alignment algorithm *nwalign* in MATLAB could be used to score the sequence similarity. However, the operation speed is about 5s per alignment @ Intel(R) Core(TM) i9-10900KF CPU 3.70 GHz. The total running time would be up to 80 hours. Since all those sequences have the same length without deletion or insertion, we set a new simple algorithm to compare the difference between two strings. This approach greatly accelerates the computational speed. Specific strain information and MATLAB codes are provided in supplementary materials.

3.3 Calculation of Actual Death Rate

According to the gamma distribution relationship of hospitalization time of death cases.

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} * e^{-x} dx;$$
$$f(x) = \frac{\lambda^{\alpha} * x^{\alpha-1} * e^{-\lambda x}}{\Gamma(\alpha)}; \quad (1)$$

α was set to be 2.2, λ was set to be 0.14 [37].

The overall death number of the first N day is represented as `death_original`,

which is a vector that contains N elements. $\text{death_original}(i)$ is the death number of *the i -th* day.

We are going to transform this data into the death number caused by *the j -th* day infection. For example, if 500 death was reported for a specific day, and 10000 infections was reported for the same day. The actual death rate is not 500/10000 for that day. What we need to know is how many deaths would be among those 10000 infections. Since we cannot obtain these data, we have to estimate the death number caused by the *j -th* day infection.

$$\begin{aligned} \text{new}_{\text{death_number}(i,j)} &= \frac{f(j-i) * \text{death_original}(i)}{\sum_1^j f(k)} ; \\ \text{final}_{\text{death_number}(j)} &= \sum_{i=1}^N \text{new}_{\text{death_number}(i,j)} ; \end{aligned} \quad (2)$$

$\text{new}_{\text{death_number}(i,j)}$ is the contribution of the death number $\text{death_original}(i)$ of the *i -th* day on the *j -th* day. $\text{final}_{\text{death_number}(j)}$ is the final transformed death number caused by *j -th* day infection.

The detailed MATLAB codes are provided in supplementary materials.

3.4 Extraction of Virus Information Based on Infection Symptom

We used the GISAID database and extracted and downloaded the patient information data by using the restriction conditions of sequence integrity, high quality, patient information, and specific sampling date. We extracted the virus length information of 1149 asymptomatic infected persons, 11117 hospitalized patients, and 9626 symptomatic infected persons. Detailed patient information and strain information are provided in supplementary materials.

3.5 Develop Virus Micro-Amplification Model Considering UTR Deletion Effects

The following processes are used to construct the model:

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

Second, the virus would die randomly in the host cell since it has a half-life. Due to the host cell's solvent and immune environments, its half-life will not be too long. Nevertheless, the half-life should be significantly longer than the initial replication cycle of the virus. If we model the population behavior

at a significant host society level, its half-life is also affected by the transmission coefficient 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. Still, in terms of the 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 $t_{1/2}$ under different environmental conditions. d represents the surviving possibility of an individual virus within a specific time interval, as depicted in equation (5). We set the replication cycle to be 10 min, and the time interval is also selected as 10min. So d represents the surviving possibility of an individual virus within 10 min. Strict prevention and epidemic control can lead to a smaller R_0 value, thus shortening the half-life of the virus in the whole population, which will lead to a smaller d value in the dynamic population model. Meanwhile, individuals with solid immunity will display a relatively more minor d value than weak immunity in single body infection simulation.

$$Genome_{length} = Genome_{length} - \text{fix}(\text{abs}(\text{normrnd}(0, \theta))) \quad (3)$$

$$Replication_{cycle} = \frac{Replication_{cycle_{initial}}}{e^{(Genome_{length_{initial}} - Genome_{length}) * para1}} \quad (4)$$

$$d = 0.5^{Time_{interval}/t_{1/2}} \quad (5)$$

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 (3) is used to calculate the new genome length after one round of replication. The deletion probability is UTR length-dependent, so at different UTR lengths, different equations will be used to describe its corrosion rate. Fourth, the deletion of the UTR region will affect the replication efficiency of the virus, which shows that it increases the replication cycle of the virus. There is a certain mathematical relationship between the deletion length and the replication period. Equation (4) is adopted to calculate the new replication cycle with different lengths. The exponential relationship is preferred compared to the linear relationship. There are two reasons. Firstly, the recognition and binding affinity between virus RNA replicase and the UTR are often nonlinear correlated with its UTR length. Secondly, the exponential model shows that the virus replication cycle is increasing 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, epidemic data 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 a severe case and fatal. It is difficult for us to estimate parameters θ , d , and $para1$ without experiments. However, given the proper relation among those parameters, we can qualitatively predict the changing trend of the epidemic situations and explain various phenomena in the development of epidemic situations through simulation. $para1$ is set to be -0.006 in all simulations since this value will reflect a good match with the actual epidemic growth situation.

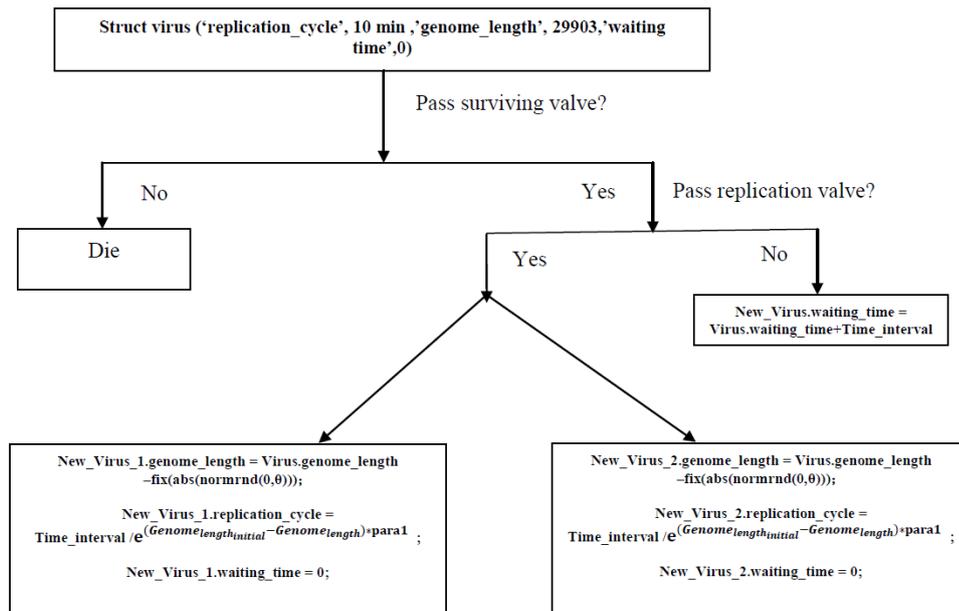


Figure 6: Algorithm workflow diagram

In Fig 6, the virus has d chance to pass the surviving valve. If $(virus.waiting_time + Time_interval) \geq virus.replication_cycle$, 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 a new genome length following the equation (1). Waiting time of the newly generated offspring is set to be 0. Those newly generated generations will go

to the next round cycle.

4 Discussion

Although the deletion of untranslated regions at both ends of the virus genome has been verified in many viruses in the past research, we put forward the hypothesis of UTR deletion theory in SARS-COV-2 employing biostatistics analysis for the first time. Moreover, we found that the length of the UTR region corresponds to the significant correlation of its biological activity through systematic statistical research. We infer that viruses with high UTR deletion degrees have weak activity, while viruses with low UTR deletion degrees have substantial toxicity. The UTR deletion theory is derived based on the correlation between its length and the change of global mortality together with patients' symptoms. Therefore, we can better explain this phenomenon that the mortality rate in COVID-19 dropped significantly in the early stage of the epidemic, and this downward trend is not sustainable. For the first time, we propose that the degradation of the UTR region of RNA virus by the host nucleic acid degradation system can lead to a decline in the overall toxicity of the virus population. At the same time, we believe that the trend of UTR length alternation is determined by the comprehensive effects of host degradation and virus replication. The large-scale deletion of UTR will significantly affect the replication activity of the virus, which is not conducive to obtaining an evolutionary advantage. At the same time, high degradation efficiency will accelerate the attenuation of the virus population genome. Unfortunately, on the COVID-19, we didn't see the continuous attenuation of the genome. Our mathematical model demonstrated how those two factors could reach a balance and maintain a plateau of its genome length. Therefore, it is difficult for COVID-19 to die naturally like other coronaviruses.

We have to admit that there is currently no experiment report explicitly describing the UTR deletion effect in COVID-19. However, our UTR deletion theory is strongly supported by statistical analysis. In today's complicated and confusing epidemic situation, COVID-19's UTR deletion theory has significant academic research value and practical significance. The idea of herd immunity has been challenged repeatedly by opposite reports [32-34]. On the theoretical level, our research group has also reiterated the limitations of the theory of group immunity in COVID-19 [35-36]. Because of the mutation effect of viruses and the time attenuation effect of antibodies, the idea of exterminating viruses by mass vaccination is not feasible. So, whether should we be inspired by this study and eliminate COVID-19 from another perspective? Many scholars have put forward this idea in previous studies [23]. That is, to accelerate the deletion of virus UTR. Suppose RNA interference, enhancement of nucleic acid exonuclease activity, or other drug treatments are used to expedite the deletion of the virus genome. In that case, it can directly help patients rescue and reduce the toxicity and risk of its

transmission. Because this deletion is irreversible, accelerating the deletion of virus UTR can successfully break through the degradation bottleneck of secondary structure, which will help foster a weak replication activity of COVID-19 in the population level, thus eventually realizing the eradication of the virus. If the natural host degradation system cannot efficiently degrade the secondary structure's bottleneck sites, can we accelerate this process through manual intervention?

Supplementary materials

All supplementary materials and Matlab codes can be accessed at:

https://github.com/zhaobinxu23/UTR_Deletion_theory

Declaration of Interest

We declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere.

We know of no conflicts of interest associated with this publication, and there has been no significant financial support for this work that could have influenced its outcome. As Corresponding Author, I confirm that the manuscript has been read and approved for submission by all the named authors.

Acknowledgment

We thank Professor Wei Dongqing from Shanghai Jiaotong University in China for the discussion. This work was supported by Dezhou university.

References

- [1] Xiao Y, Torok M E. Taking the right measures to control COVID-19[J]. *The Lancet Infectious Diseases*, 2020, 20(5): 523-524.
- [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] 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.
- [5] 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.
- [6] Zehender G, Lai A, Bergna A, et al. GENOMIC CHARACTERISATION AND PHYLOGENETIC ANALYSIS OF SARS-COV-2 IN ITALY[J]. *medRxiv*, 2020.
- [7] 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.
- [8] Yao H P, Lu X, Chen Q, et al. Patient-derived mutations impact pathogenicity of SARS-CoV-2[J]. *CELL-D-20-01124*, 2020.

- [9] Grubaugh, Nathan D., William P. Hanage, and Angela L. Rasmussen. "Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear." *Cell* 182.4 (2020): 794-795.
- [10] Plante, Jessica A., et al. "Spike mutation D614G alters SARS-CoV-2 fitness." *Nature* (2020): 1-6.
- [11] Davies, Nicholas G., et al. "Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England." *medRxiv* (2020).
- [12] Khafaie M A, Rahim F. Cross-country comparison of case fatality rates of COVID-19/SARS-CoV-2[J]. *Osong Public Health and Research Perspectives*, 2020, 11(2): 74.
- [13] Jain V K, Iyengar K, Vaish A, et al. Differential mortality in COVID-19 patients from India and western countries[J]. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 2020, 14(5): 1037-1041.
- [14] Hashim M J, Alsuwaidi A R, Khan G. Population risk factors for COVID-19 mortality in 93 countries[J]. *Journal of epidemiology and global health*, 2020, 10(3): 204.
- [15] Chapman, Nora M., et al. "5' terminal deletions in the genome of a coxsackievirus B2 strain occurred naturally in human heart." *Virology* 375.2 (2008): 480-491.
- [16] Lévêque, Nicolas, et al. "Functional consequences of RNA 5'-terminal deletions on coxsackievirus B3 RNA replication and ribonucleoprotein complex formation." *Journal of virology* 91.16 (2017).
- [17] Hunziker, Isabelle P., Christopher T. Cornell, and J. Lindsay Whitton. "Deletions within the 5' UTR of coxsackievirus B3: consequences for virus translation and replication." *Virology* 360.1 (2007): 120-128.
- [18] AlMalki W H, Shahid I, Abdalla A N, et al. Consensus small interfering RNA targeted to stem-loops II and III of IRES structure of 5' UTR effectively inhibits virus replication and translation of HCV sub-genotype 4a isolates from Saudi Arabia[J]. *Saudi Journal of Biological Sciences*, 2021, 28(1): 1109-1122.
- [19] Collier A J, Tang S, Elliott R M. Translation efficiencies of the 5'untranslated region from representatives of the six major genotypes of hepatitis C virus using a novel bicistronic reporter assay system[J]. *Journal of General Virology*, 1998, 79(10): 2359-2366
- [20] Tang S, Collier A J, Elliott R M. Alterations to both the primary and predicted secondary structure of stem-loop IIIc of the hepatitis C virus 1b 5' untranslated region (5' UTR) lead to mutants severely defective in translation which cannot be complemented in trans by the wild-type 5' UTR sequence[J]. *Journal of virology*, 1999, 73(3): 2359-2364.
- [21] Lin Y J, Zhang X, Wu R C, et al. The 3'untranslated region of coronavirus RNA is required for subgenomic mRNA transcription from a defective interfering RNA[J]. *Journal of Virology*, 1996, 70(10): 7236-7240..
- [22] Zhang, Jian Jun, et al. "Promoter activity of SARS coronavirus 5'UTR sequence in eukaryotic cells." *Sichuan da xue xue bao. Yi xue ban= Journal of Sichuan University. Medical science edition* 37.1 (2006): 5-9.
- [23] Baldassarre, Antonella, et al. "Potential use of noncoding RNAs and innovative therapeutic strategies to target the 5'UTR of SARS-CoV-2." *Epigenomics* 12.15 (2020): 1349-1361.
- [24] Yang T, Shen K, He S, et al. CovidNet: To bring data transparency in the era of COVID-19[J]. *arXiv preprint arXiv:2005.10948*, 2020.
- [25] Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study[J]. *The lancet*, 2020, 395(10229): 1054-

1062.

- [26] Faes C, Abrams S, Van Beckhoven D, et al. Time between symptom onset, hospitalisation and recovery or death: statistical analysis of Belgian COVID-19 patients[J]. *International journal of environmental research and public health*, 2020, 17(20): 7560.
- [27] Chakravart L, Laha R G, Roy J W. Handbook of methods of applied statistics. Vol. I: techniques of computation, descriptive methods and statistical inference[J]. 1967.
- [28] Narayanan K, Makino S. Interplay between viruses and host mRNA degradation[J]. *Biochimica Et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 2013, 1829(6-7): 732-741.
- [29] Miao Z, Tidu A, Eriani G, et al. Secondary structure of the SARS-CoV-2 5'-UTR[J]. *RNA biology*, 2021, 18(4): 447-456.
- [30] Hatcher E L, Zhdanov S A, Bao Y, et al. Virus Variation Resource—improved response to emergent viral outbreaks[J]. *Nucleic acids research*, 2017, 45(D1): D482-D490.
- [31] Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data—from vision to reality[J]. *Eurosurveillance*, 2017, 22(13): 30494.
- [32] Murray C J L, Piot P. The potential future of the COVID-19 pandemic: will SARS-CoV-2 become a recurrent seasonal infection?[J]. *Jama*, 2021, 325(13): 1249-1250.
- [33] Sabino E C, Buss L F, Carvalho M P S, et al. Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence[J]. *The Lancet*, 2021, 397(10273): 452-455.
- [34] Poustchi H, Darvishian M, Mohammadi Z, et al. SARS-CoV-2 antibody seroprevalence in the general population and high-risk occupational groups across 18 cities in Iran: a population-based cross-sectional study[J]. *The Lancet Infectious Diseases*, 2021, 21(4): 473-481.
- [35] Zhaobin Xu, Hongmei Zhang, Yonggang Niu. A Continuous Bayesian Model for the Stimulation COVID-19 Epidemic Dynamics[J]. *medRxiv*, 2021.
- [36] Zhaobin Xu, Hongmei Zhang. If we cannot eliminate them, should we tame them? Mathematics underpinning the dose effect of virus infection and its application on COVID-19 virulence evolution[J]. *medRxiv*, 2021.
- [37] Hawryluk I, Mellan T A, Hoeltgebaum H, et al. Inference of COVID-19 epidemiological distributions from Brazilian hospital data[J]. *Journal of the Royal Society Interface*, 2020, 17(172): 20200596.
- [38] Nagy Á, Pongor S, Györfy B. Different mutations in SARS-CoV-2 associate with severe and mild outcome[J]. *International journal of antimicrobial agents*, 2021, 57(2): 106272.
- [39] Biswas S K, Mudi S R. Spike protein D614G and RdRp P323L: the SARS-CoV-2 mutations associated with severity of COVID-19[J]. *Genomics & informatics*, 2020, 18(4).
- [40] Oulas A, Zanti M, Tomazou M, et al. Generalized linear models provide a measure of virulence for specific mutations in SARS-CoV-2 strains[J]. *PloS one*, 2021, 16(1): e0238665.
- [41] Majumdar P, Niyogi S. ORF3a mutation associated with higher mortality rate in SARS-CoV-2 infection[J]. *Epidemiology & Infection*, 2020, 148.
- [42] Voss J D, Skarzynski M, McAuley E M, et al. Variants in SARS-CoV-2 associated with mild or severe outcome[J]. *Evolution, medicine, and public health*, 2021, 9(1): 267-275.
- [43] Toyoshima Y, Nemoto K, Matsumoto S, et al. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19[J]. *Journal of human genetics*, 2020, 65(12): 1075-1082.
- [44] Volz E, Hill V, McCrone J T, et al. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity[J]. *Cell*, 2021, 184(1): 64-75. e11.
- [45] Zhang L, Jackson C B, Mou H, et al. SARS-CoV-2 spike-protein D614G mutation increases

virion spike density and infectivity[J]. *Nature communications*, 2020, 11(1): 1-9.

[46] Hou Y J, Chiba S, Halfmann P, et al. SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo[J]. *Science*, 2020, 370(6523): 1464-1468.

[47] Wacker A, Weigand J E, Akabayov S R, et al. Secondary structure determination of conserved SARS-CoV-2 RNA elements by NMR spectroscopy[J]. *Nucleic acids research*, 2020, 48(22): 12415-12435.