Variants and mutation rate in SARS-CoV-2

Zhang Leyi[1], Jinqi Duan[1]

[1]MSc in BiRC, Aarhus University, Aarhus, Denmark

*Corresponding author: Palle Villesen, Associate Professor; palle@birc.au.dk

2023-01

Abstract

Severe Acute Respiratory Syndrome Coronavirus 2(SARS-CoV-2) is the virus that causes COVID-19, a respiratory illness that outbreaks worldwide. It is critical to interpret the mutation of SARS-CoV-2 and the selection by the environment in understanding and predicting the prevalence of past and further outbreaks and developing informed strategies to prevent more severe global spreading.

Our research focuses on the mutation of SARS-CoV-2 and the natural selection by the environment(human immune system). SARS-CoV-2 viruses are highly spread and show a high rate of transmissibility and neutralizing antibody escape. We quantified the strength of natural selection by using an alternative method of dNdS. And we observed that, in Omicron lineages, 47.2% of lineages are positively selected by the environment, including BA.2(dNdS = 1.13), BA.4(dNdS = 1.09), BA.5(dNdS = 1.99), which leads to the strong differentiation and generation of new lineages. In different countries, the strengths of natural selection are still divergent, though there is high globalization worldwide, while the selection pressure is proportional to the number of lineages. In different Variants of Concern, we observed that the older variants were positively selected at a stronger rate than newer variants(Omicron is the least).

Genetic mutation drives the variation in amino acids, leading to the variation in phenotypes. The differences in the frequency of amino acid change can be used for detecting the level of differentiation of different lineages. Adaptive evolution can be kept, thus we see a higher frequency. The frequencies of amino acid changes are dynamic through time. Thus, we designed a method to detect and compare the frequencies among some lineages and Delta variants. This can explain the prevalence of some variants. In comparing the frequencies of L452R change in BA.2 and BA.5, we also found proof of the existence of convergence evolution.

Our research highlights and quantifies the strength of natural selection(especially positive selection), and explains the reasons for different transmissibility and prevalence of different variants. The findings in this research can help us gain a better understanding of SARS-CoV-2 mutations and the selection by the environment, and can be used to develop more informed strategies to prevent further global spreading. We also provide a new method for detecting the level of differentiation of different lineages and variants and a more accurate method for detecting the prevalence of variants.

In conclusion, SARS-CoV-2 is highly mutated and driven by the environment. The strength of natural selection is divergent in different countries and Variants of Concern, which leads to the generation of new lineages. The frequencies of amino acid changes can be used to explain the prevalence of some variants and detect the level of differentiation of different lineages. Our research provides a few insights for understanding and predicting the prevalence of past and further outbreaks and developing informed strategies to prevent more severe global spreading.

Keywords: SARS-CoV-2; Mutation; Natural Selection Strength, Amino Acid Change Frequency

1 Introduction

Severe Acute Respiratory Syndrome Coronavirus 2(SARS-CoV-2) is the virus that causes COVID-19, a respiratory illness which was first identified in December 2019. The symptoms can range from mild to severe, including fever, cough, shortness of breath, body aches, fatigue, and loss of taste and smell. The mutation rate of SARS-CoV-2 is not high, but SARS-CoV-2 can widely spread worldwide. The nucleotide mutation rate of the whole genome was $1.3\times10^{-6}\pm0.2\times10^{-6}$ per base per infection cycle, Amicone et al. (2022). Measurements show that mutation rates for RNA virus can range of 10^{-3} to 10^{-6}, Domingo, García-Crespo, Lobo-Vega, and Perales (2021). It means that SARS-CoV-2 can have a lot of advantageous mutations and be positively selected by the environment. Therefore, interpreting the evolution of SARS-CoV-2 and the selection by the environment is critical to understanding the dynamics of future outbreaks and developing informed strategies to prevent more severe global spread. In this analysis, we focused on the strength of natural selection of different lineages in the genetic diversity of SARS-CoV-2 and explains the reasons for divergence and convergence through the frequency of amino acid change.

For the evolution of SARS-CoV-2, our research shows that the mutation is under both purifying selection and strong positive selection. Some analysis shows that the evolution of some genes(e.g., S, ORF3a, and N) is driven by positive selection actively, leading to the molecular processes involving pathogen–host interactions, including viral invasion into and egress from host cells, viral inhibition and evasion of host immune response, Y. Hou et al. (2022). We quantified the strength of natural selection by using the ratio of nonsynonymous substitution accumulation to synonymous substitution accumulation, also known as "dNdS". As dNdS analysis requires both the number of nonsynonymous sites and synonymous sites when comparing two sequences, which is lacking in our dataset and cannot be calculated directly. Therefore, based on the dNdS analysis, we propose a new method to quantify evolutionary strength, **see Section 3.4**.

We observed that, in different variants and lineages, the strength of natural selection were distinct, which proved the different circumstances of various transmissibility rate and replacement rate of newer variants to older variants. In different countries and different Variants of Concern, due to differentiated measurements of the treatment, the selection pressures were also different.

Mutation of genes drives the change in amino acids, which leads to the different appearance of phenotypes. Mutations are random, and the environment will keep those adaptive evolutions and eliminate those deleterious evolutions. Therefore, by calculating the frequency of amino acid change of different variants, it can explain why some variants are adaptively selected and some are eliminated by the environment. We believe that the amino acid with increased frequency is generally affected by positive selection, and the decreased frequency of an amino acid is also considered to be changed due to the environment change(vaccination, medication, etc.) or new variant outbreaking that replaced the older gene.

2 Basic Terminologies

2.1 Amino acid change

In the case of SARS-CoV-2, the mutation of genes lead to changes in amino acid sequences. Amino acids are the building blocks of proteins, and they are coded by specific sequences of nucleotides in a gene. Amino acid monomers are the basis for protein synthesis. Thus changes in the expression of viral traits can be achieved by changing amino acids. Multiple amino acid sequences can be translated into the same protein, and some of these mutations may have no effect on the trait expression, which are called synonymous mutations, Another type of mutation, nonsynonymous mutation, can cause a range of effects on the protein's function, including affecting the virus' ability to infect, replicate, and evade from the immune system.

In the case of this project dataset, D614G means that Aspartic Acid in a specific strain was replaced by Glycine at the 614th amino-acid position of the spike protein.

2.2 Variants of Concern(VOCs)

A variant for which there is evidence of an increase in transmissibility, more severe disease (e.g. the increment of hospitalization and death), strong neutralizing antibody escape, reduced effectiveness of treatment or vaccination, and diagnostic detection failures.

The current list of the Variants of Concern that are currently being closely monitored and characterized by WHO are as follows:

Pangolin Lineage: B.1.1.529, BA.1, BA.1.1, BA.2, BA.3, BA.4 and BA.5 lineages.

The following are the main variants of concern quantified for this project. Delta, Omicron(BA.1, BA.2, BA.4,

BA.5, BF.7, XBB)

2.3 Adaptive evolution

Adaptive evolution is the process by which populations of organisms change over time in response to selective pressures in their environment.

Mutations that could alter the traits of organisms. As mutations occur randomly, natural selection will preserve traits compatible with the organism's habitat, and the individual has an increased chance of survival and reproduction. Alleles in a population become more common through natural selection, but the frequency of harmful alleles becomes less common through selection against them. This is adaptive evolution at work.

3 Methods and Tools

3.1 Data

The findings of this study are based on metadata associated with individual sequences and mutations associated with sequences available on GISAID up to November 07, 2022.

The metadata includes the collection date of the case, synonymous mutation accumulation, non-synonymous mutation accumulation, pangolin lineage, etc. The mutation dataset includes information such as detected base change, corresponding amino acid change, etc. We merge the two datasets according to the corresponding patient id information, using (id, position) as the primary key.

We collected three sub-datasets: 299,941 sequences of individual cases and 14,758,805 sequences of mutation sequences, dated from 2020-01-01 to 2022-10-31; 999,826 sequences of individual cases and 49,213,687 sequences of mutation sequences from 2020-01-01 to 2022-11-08; 956,311 sequences of individual cases and 50,849,777 sequences of mutation sequences, dated from 2020-04-23 to 2022-11-07. In the third dataset, we balanced the number of collected Omicron variants between older and closer lineages, so that some new lineages, such as BA.5 and BF.7, can have enough acquisition, which helps us in our analysis.

Phylogenetic methods impose associations on the data, so the results are very sensitive to the presence of problematic sequences or data. Some data in Metadata have problems such as wrong collection date, data pollution, and label confusion. For example, in reality, the earliest BA.5 was detected on February 25, 2022, while the data contained BA.5 marked in November 2021. We have carried out a series of Data Cleaning to delete obviously wrong data. We selected the 95% data in between of each lineage for analysis to avoid labelling irregularities of early data. After screening, the average difference between synonymous and non-synonymous mutations increased linearly over time, so that the estimate of the mutation rate would be robust. The non-synonymous mutation accumulation and synonymous mutation accumulation of different lineages after QC are shown in Fig.2.

3.2 Linear Regression

When estimating the evolution rate of various VOCs or Pangolin Lineages, we mainly used linear regression to quantify this trend. We focused on the trend of Non-synonymous or synonymous substitutions change over time, grouping them by Pangolin Lineages, and then using linear regression to fit them to obtain trend quantification.

Since multiple lineages share ancestors, the differences in sequences are not independent; while there may be multiple lineages from different ancestors at the same time, mutation accumulation does not grow linear over time. Therefore, it is difficult to obtain confidence intervals using normal linear regression. We observed that, within each lineage, non-synonymous and synonymous mutation accumulations were independent and positively correlated with time. For some lineages, our sample sizes were insufficient for analysis. But the effective sample size was sufficient for most lineages to analyze and perform linear regression analysis. Therefore we perform linear regression on different lineages or VOCs respectively.

$$Y = W \times X + Z \tag{1}$$

In a given dataset where X $= [x1, x2, ..., xn]^T \in R^N$, where X denotes the set of individual samples, N is the number of samples in a variant or lineage, Y is the number of synonymous or non-synonymous mutations in total, W is the trend we wish to get, Z is the number of synonymous or non-synonymous mutations in its parent lineage.

Another reason to use linear regression is that W in this equation would be used in calculating *new dNdS* in the method below, see Equation 5.

3.3 Frequency by time

Spike mutation has received widespread attention, mainly due to the reactive investigation of the frequency of amino acid changes, that is, the increase or decrease in frequency often has certain epidemiological characteristics. Genome analysis shows that the host environment changes and selection pressure changes will cause changes in amino acid change frequency, and can cause convergence mutation, Harvey et al. (2021a). In addition, the study of spike mutation also includes exploring how it affects neutralization. For example, in an immune escape study, E484K was identified as a mutation that can reduce the neutralization ability of mAb combinations to an undetectable level, Baum et al. (2020).

When the nucleotide is mutated in the corresponding DNA sequence, the amino acid in the protein will be replaced. In the study of SARS-CoV-2, continuous long-term tracking of viral genomics is generally required. Changes in a single amino acid will lead to virus mutation, and those that adapt to the environment will dominate in the world. For example, D614G change has the advantage to be more adaptable, Korber et al. (2020). Here we designed a set of methods to track amino acids change: By analyzing the relationship between the frequency of amino acid each day and the date, it is judged whether the frequency of the amino acid change is rising or dropping in different time periods.

$$Freq = \frac{SAA}{ALL}, \tag{2}$$

where SAA denotes the number of individuals catching the variant with the Specific Amino Acid change in a day, and ALL denotes the number of individuals getting positive on the same day.

3.4 *new dNdS*

SARS-CoV-2 variants show punctuated evolution, but not gradualism. After different pangolin lineages are produced, they will be subject to positive or negative selection from the environment, as well as the competition between lineages, thus the spread and daily detection ratio are different. dNdS is a method used to estimate selection. Since synonymous mutations are largely unaffected by natural selection, non-synonymous mutations may be under strong selective pressure. Compare the rates of these two types of mutations Can be used to quantify the strength of evolution by natural selection, Yang and Nielsen (2000).

The traditional dNdS method is to use the maximum likelihood extension of the McDonald-Kreitman test. By comparing non-synonymous and synonymous divergence in different lineages, we expect to get the values of dN and dS, Obbard, Welch, Kim, and Jiggins (2009).

In this experiment, our data is the accumulation of non-synonymous and synonymous substitutions. Different lineages may infect different people at the same date, and different lineages do not necessarily have a linear relationship with the phylogeny tree. Therefore, here we cannot simply use the formula to calculate dNdS. We propose an alternative scheme that uses linear regression to find the accumulation rate of non-synonymous and synonymous substitutions over time. It is not possible to directly compare the accumulation of non-synonymous and synonymous substitutions of each person to get CN/CS, or directly calculate dNdS. The reasons are as follows:

$$\frac{\mathrm{d}\frac{CountN}{CountS}}{\mathrm{d}t} = \frac{CountN' \times CountS - CountN \times CountS'}{CountS^2} \tag{3}$$

We set the left derivative $\frac{\mathrm{d}\frac{CountN}{CountS}}{\mathrm{d}t}$ to be 0, as if it is the case that dNdS = 0, so we'll get:

$$\frac{CountN^{'}}{CountN} = \frac{CountS^{'}}{CountS} \tag{4}$$

It means that the slope of $\frac{CountN}{CountS}$ is controlled by:

1. Increment: $CountN^{'}$ and $CountS^{'}$
2. Aggregate: CountN and CountS

Alternatively, we use the accumulation data of non-synonymous and synonymous substitutions to get the following values:

$$\frac{\Delta CountN}{\Delta CountS} = \frac{(\Delta CountN + \delta) - CountN}{(\Delta CountS + \delta) - CountS} \tag{5}$$

Where δ is an infinitesimal.

$$\frac{\Delta CountN}{\Delta CountS} = \frac{\frac{(\Delta CountN+\delta)-CountN}{\Delta t}}{\frac{(\Delta CountS+\delta)-CountS}{\Delta t}} \tag{6}$$

Thus we have our *new dNdS* as:

$$\frac{\Delta CountN}{\Delta CountS} = \frac{CountN^{'}}{CountS^{'}} \tag{7}$$

So when $\frac{CountN^{'}}{CountS^{'}} > 1$, it is positive selection; $\frac{CountN^{'}}{CountS^{'}} = 1$, it is neutral selection; $\frac{CountN^{'}}{CountS^{'}} < 1$, it is negative selection.

Where CN and CS are the accumulation of non-synonymous and synonymous substitutions, and they are respectively partial derivatives to obtain dN and dS. We call the ratio of the two values *new dNdS*.

If the *new dNdS* ratio is greater than 1, it suggests that there has been an excess of non-synonymous mutations and that positive selection has been acting on the gene. Positive selection refers to the process by which certain genetic variants become more common in a population over time because they provide a reproductive advantage. In this case, the excess of non-synonymous mutations may be due to the fact that they have provided some advantages to the organisms carrying them, such as increased survival or reproductive success.

On the other hand, if the *new dNdS* ratio is less than 1, it suggests that there has been an excess of synonymous mutations and that the gene may be subject to purifying selection. Purifying selection refers to the process by which harmful or deleterious mutations are removed from a population through natural selection. In this case, the excess of synonymous mutations may be due to the fact that they have not had a significant impact on the fitness of the organisms carrying them.

Therefore, the steps are as follows:

1) Perform linear regression on non-synonymous and synonymous mutation accumulations of different lineages.
2) Record the slopes separately, as k1 and k2.
3) *New dNdS* $= \frac{k1}{k2}$

3.5 The estimate of *new dNdS* by Bootstrapping

The Bootstrapping algorithm is a method of re-sampling based on limited sample data that is sufficient to represent the distribution of the parent sample.

The basic idea of Bootstrap is to continuously take small samples randomly for the existing data, process the data for each small sample, and obtain the estimator to obtain the distribution of the estimator. It is a method of re-sampling that takes independent samples from existing sample data and replaces the same number of samples to make inferences on these re-sampled data. It has steps as follows:

1) Draw a sample from the population, with the sample size n.
2) Draw a replacement sample of size n from the sample data and replicate it B times. Each re-drawn sample is called a Bootstrap sample, and there will be a total of B Bootstrap samples.
3) Evaluate the statistic of θ for each Bootstrap sample, there will be a total of B estimates of θ. Use these B Bootstrap statistics to construct a sampling distribution, and use it to make further statistical inferences, for example: estimating the statistical standard error of θ, and obtaining the confidence interval of θ.

In this paper, due to the differences in the detection level of the new crown in different countries and the time difference of lineages from emergence to the present, the number of samples we collected for different lineages or variants varies greatly. The effective sample size of some lineages is not enough, thus the calculated *new dNdS* would be biased. Therefore, we wish to use Bootstrapping to calculate the accuracy of the *new dNdS* estimate.

We repeat the sampling of different lineages 1000 times, and the sample size of each sampling is 10 times the length of the specific lineage sample, to ensure that the deviation calculated each time is negligible. Then we use the above method to calculate *new dNdS* for 1000 samples and use a histogram to present the frequency and calculate the mean, median, and confidence interval to judge whether the *new dNdS* we calculated is acceptable.

4 Results

In the three years since the outbreak of SARS-CoV-2 till today, there have been 14 different Variants of Concern. In addition, in the Omicron variant, 451 species have been defined based on the mainstream strains using the pangolin lineage Fig.1.

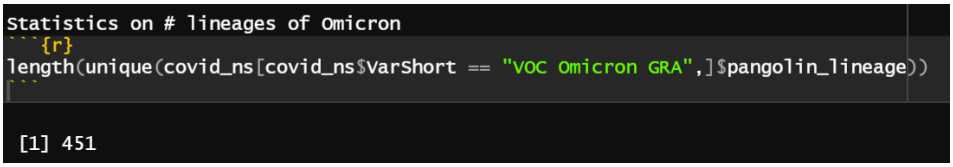
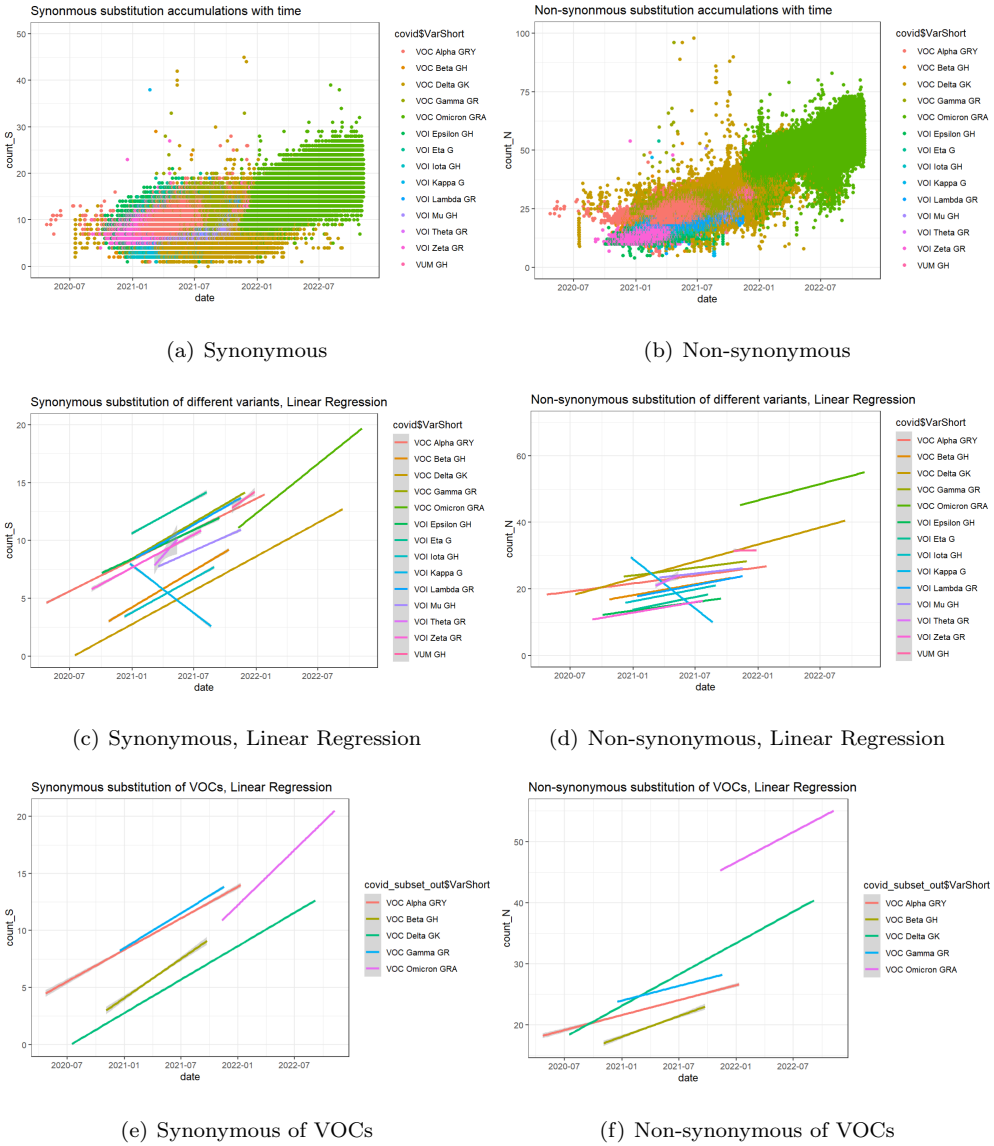


Figure 1: **Number of different lineages in Omicron.**

We tracked the mutation and natural selection of SARS-CoV-2 in the COVID-19 pandemic, focusing on analyzing some Variants of Concern (VOCs) by WHO reports, Table 1, and making a horizontal comparison of different lineages. We analyzed patient and virus data from January 2020 to November 2022. When different lineages or VOCs change with date, the growth rates of non-synonymous substitution accumulation and synonymous substitution accumulation are different. Then we made linear regression on the substitution accumulation of these variants and drew a more intuitive trend judgment Fig.2.

Alpha(Previous VOC)	Beta(Previous VOC)
Gamma(Previous VOC)	Delta(Previous VOC)
Omicron BA.1(Current VOC)	Omicron BA.2(Current VOC)
Omicron BA.4(Current VOC)	Omicron BA.5(Current VOC)
Omicron BF.7(Current VOC)	Omicron XBB(Current VOC)

Table 1: The list of VOCs



(a) Synonymous

(b) Non-synonymous

(c) Synonymous, Linear Regression

(d) Non-synonymous, Linear Regression

(e) Synonymous of VOCs

(f) Non-synonymous of VOCs

Figure 2: **Substitution accumulations with time.**

Each panel shows the trend of mutations in different variants over time (Synonymous(A), Non-synonymous(B), Synonymous with linear regression(C), Non-synonymous with linear regression(D), Synonymous of VOCs(e), Non-synonymous of VOCs(f).

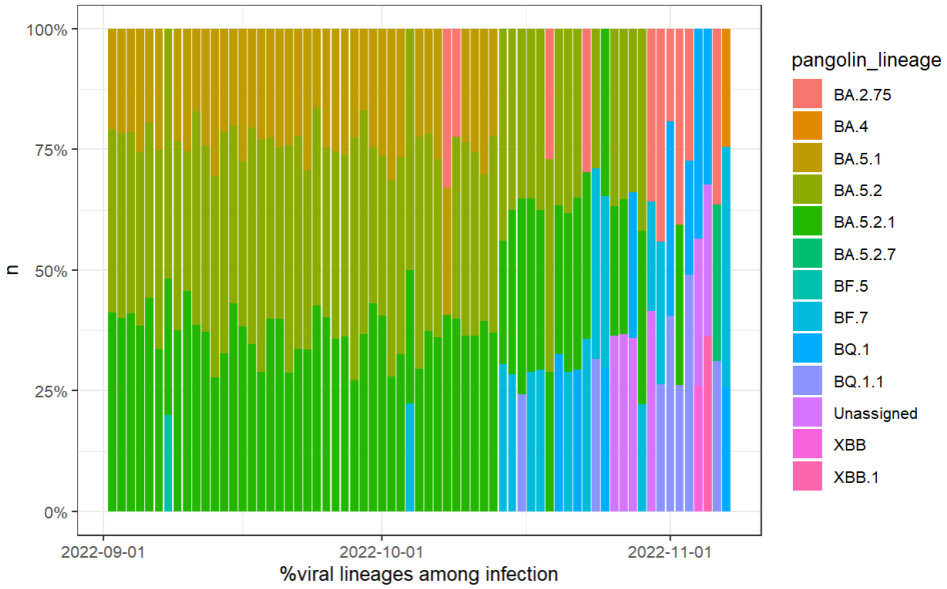
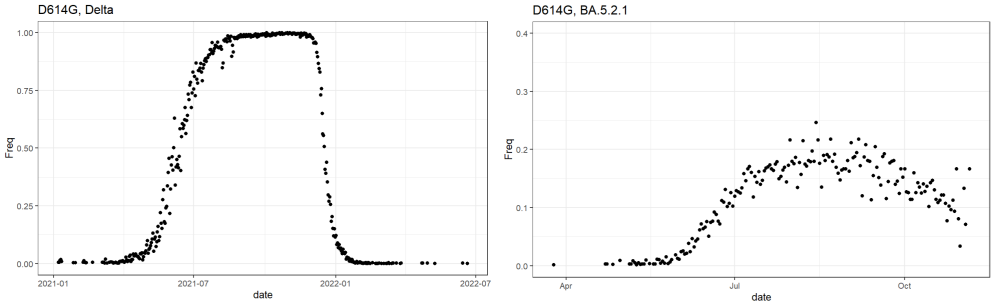


Figure 3: **Time of mainstream lineages replacement.**

Where the x-axis represents dates from 2022-09-01 to 2022-11-01, "n" on the y-axis represents the proportion of different types of strains detected on the same day. It can be seen that it took 32 days for BF.7 to become mainstream from its appearance.

The differences in the mutation accumulation rate of different lineages are determined by both the mutation itself and the environment. If beneficial genes are mutated to produce beneficial amino acids and synthesize proteins, the mutations can be accumulated. For example, some amino acid changes may lead to spike protein mutations, so some spike proteins are cleaved more or less frequently, Callaway (2021). Therefore, different amino acid changes can be used to explain whether the variant adapts to the environment. We calculate the frequency of a specific amino acid change in daily patient data to determine whether the amino acid change has an impact on the variant or lineage or is "hitchhiking", Fig.4.

Using the *frequency by time* method, we judge whether the amino acid change brought by a certain mutation is an advantageous mutation according to the joint influence of environmental selection and intraspecific competition. The amino acid with increased frequency is positively selected by the environment and presents an advantageous level in intraspecific competition, while the decreased frequency of an amino acid is also considered to be a change in the environment (vaccination, medication, etc.) or a new variant that replaces the mutation.

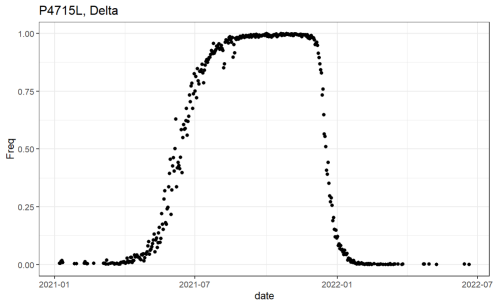


(a) D614G, Delta

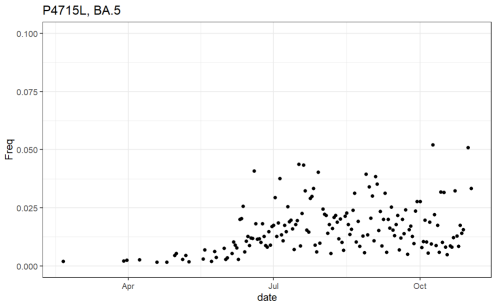
(b) D614G, BA.5.2.1

Figure 4: **Frequency of amino acid change(D614G) with the date by lineages.**
Where Freq on the y-axis represents the frequency of this specific amino acid change detected on this day compared to all amino acid changes detected(**D614G of Delta(a), D614G of BA.5.2.1(b), L452R of BA.4(c), L452R of BA.5(d)**). In plot a, D614G showed a rising frequency at first, meaning that Delta with D614G mutation had some advantageous effect. It dropped between 2021-11 to 2022-01, which means that Delta with D614G got out-competed by other variants. In reality, that variant out-competed Delta was Omicron. In BA.5.2.1, however, since April, D614G mutation frequency increased again.

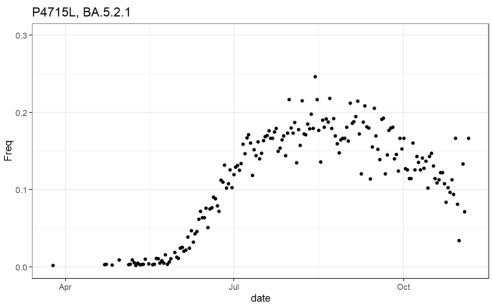
We also analyzed that the frequency of P4715L increased during the Delta period and decreased as the Delta variant was replaced by Omicron variants; while in BA.5 and BA.5.2.1, the frequency of P4715L increased, Fig.5. Based on this, we predicted that the new lineages with P4715L change would become a more successful variant than the virus of the same lineage. The amino acid changes, as mentioned above, D614G, T478K, L452R, S704L, R346T, are the same.



(a) The frequency of P4715L in Delta



(b) The frequency of P4715L in BA.5



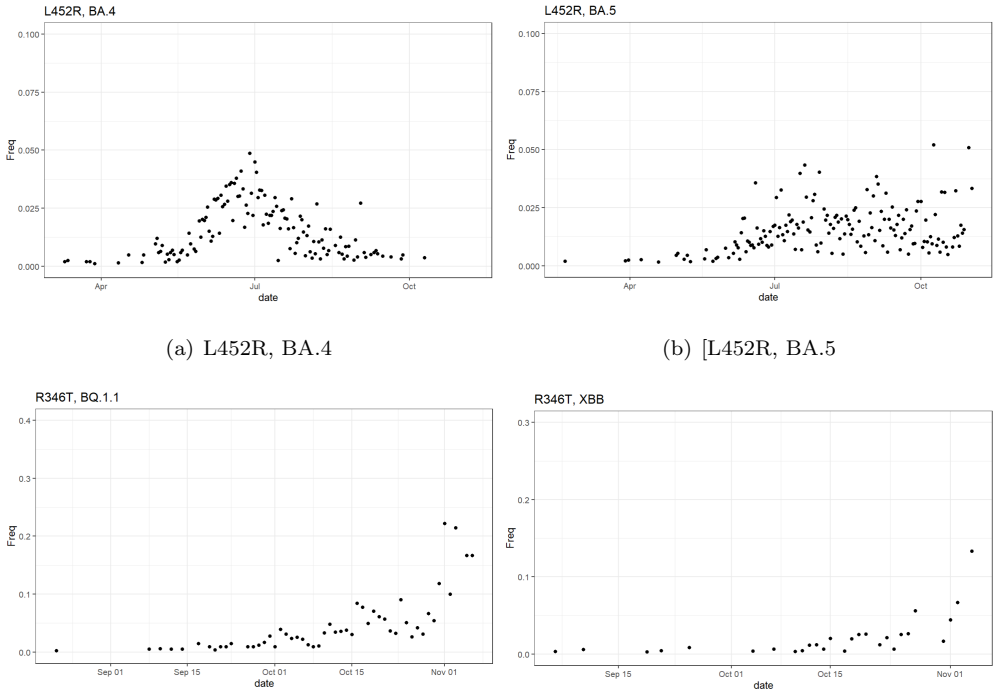
(c) The frequency of P4715L in BA.5.2.1

Figure 5: **The frequency of P4715L in different variants**

Through some other analysis, in the spike protein S of the new coronavirus in the early stage of the outbreak, amino acid 614 was aspartic acid (D), but starting from March 2020, this amino acid was mutated to glycine (G). By June, the frequency of D614G had increased to 75%, and now, the D614G mutant has replaced the earlier virus as the dominant strain worldwide. Clinically, nasopharyngeal swabs from G614-infected individuals detected higher levels of viral RNA than D614-infected individuals but did not alter the pathogenicity of the virus, Shi and Xie (2021).

In Fig. 4, we detected the frequency of D614G based on all Delta variant data. D614G in Delta showed an advantageous level before July 2021, and almost fixed since then. When Omicron rose in November 2021, the reason for the frequency of D614G dropped is that the Delta variant was also replaced by Omicron. The frequency of D614G in the Omicron variant rose, taking BA.5.2.1 as an example, which means that D614G would be a reason that new variants can get better neutralizing antibody escape, and be in an advantageous level of selection.

We also detected frequencies of some other amino acid changes, Figure 7, Figure 8. Spike mutation L452R is from antigenic drift, which is hard to accumulate. But in only a month(Feb 2022), it was accumulated. it also shows convergence evolution, that BA.2, and BA.5 both have L452R change.



(a) L452R, BA.4

(b) [L452R, BA.5

(c) R346T, BQ.1.1

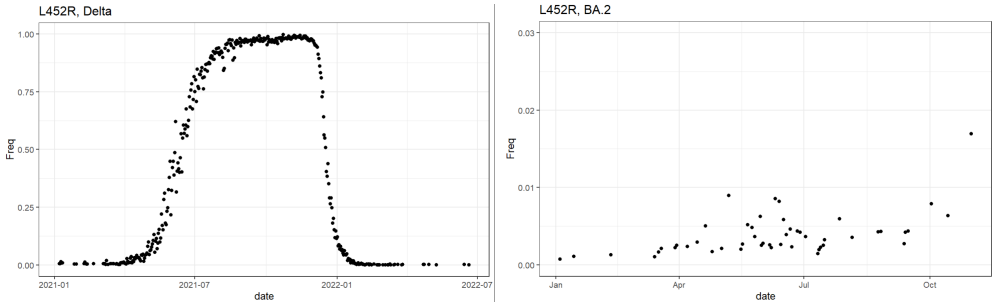
(d) R346T, XBB

Figure 6: **The existence of Convergence evolution in Omicron**

Plot a and b show that L452R in BA.4 and BA.5 rose at almost the same time. **Plot c and d** show that R346T in BQ.1.1 and XBB rose together.

Both BA.4 and BA.5 were first discovered in South Africa. BA.4 was discovered on January 10, 2022, and BA.5 was discovered on February 25, 2022. Both are the sub-generation of BA.2 on the phylogeny tree. Plots c and d of Fig. 4 show that the frequency of L452R of BA.4 and BA.5 started to increase almost simultaneously in April 2022. The mutation was also previously detected in the Delta variant and is thought to make the virus more infectious by enhancing its ability to attach to human cells, Y. J. Hou et al. (2020). Breakthrough infection of BA.2 and BA.5 due to humoral immune imprinting reduced the diversity of Neutralizing antibody binding(NAb) sites and increased the proportion of non-neutralizing antibody clones, which in turn concentrates humoral immune pressure and promotes RBD Convergent evolution, Cao et al. (n.d.). We also analyzed the frequency changes of R346T in XBB and BQ.1.1 for this purpose, and also demonstrated the existence of convergent evolution, Fig. 6, these observations raise the possibility that the new variants could be more fusion and pathogenic than the old variants, Ito et al. (n.d.).

We observe the amino acids change, and count the number and frequency of some specific synonymous and non-synonymous amino acid changes in the daily new patients so that we can judge whether these amino acid changes are positively or negatively selected for the lineages adaptability effect. We focused on D614G, T478K, L452R, S704L, R346T, Fig.6(a)(b), Fig.7, Fig.8. The Freq of some amino acid changes also increased first, then decreased, and then rose again to a certain extent in the Omicron stage. We believe that the emergence of this form of amino acid change needs to be paid attention to, it will cause the newly emerged Omicron subtypes to produce some traits in the Delta stage, and it can explain the reason for its positive selection by the environment.

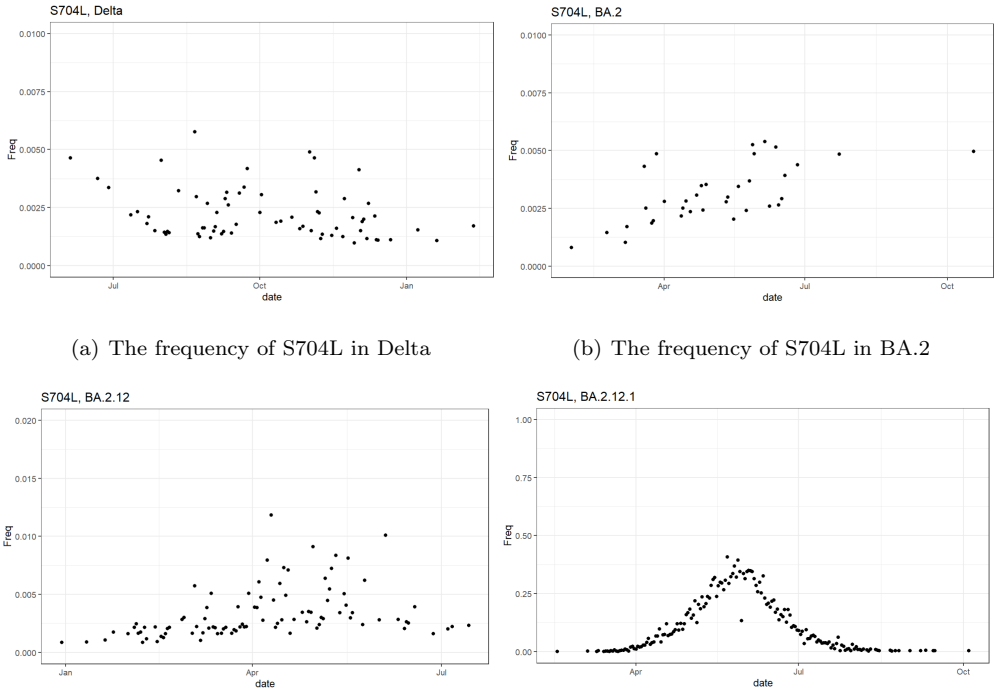


(a) The frequency of L452R in Delta

(b) The frequency of L452R in BA.2

Figure 7: **The frequency of L452R in different variants**

In different lineages, the frequency of L452R increases. It shows the convergence evolution.



(a) The frequency of S704L in Delta

(b) The frequency of S704L in BA.2

(c) The frequency of S704L in BA.2.12

(d) The frequency of S704L in BA.2.12.1

Figure 8: **The frequency of S704L in different variants**

We, by improving the *new dNdS* method, quantified the selection pressure on different lineages. We used the *new dNdS* method to calculate the *new dNdS* of all pangolin lineages of Omicron and Delta, and removed the outliers, leaving the *new dNdS* of 355 lineages (Supp Table 1). Since the amount of data in the dataset is too small for some lineages, there are obvious outliers, e.g. *new dNdS* is negative or *new dNdS* is greater than 10, we will remove them. Among the remaining dNdS values, the proportion of lineages being positively selected by the environment, with *new dNdS* > 1, is 0.538. We screened some pangolin lineages that have been mainstream or are becoming mainstream in the world, calculated their *new dNdS* and drew a bar plot Fig.9.

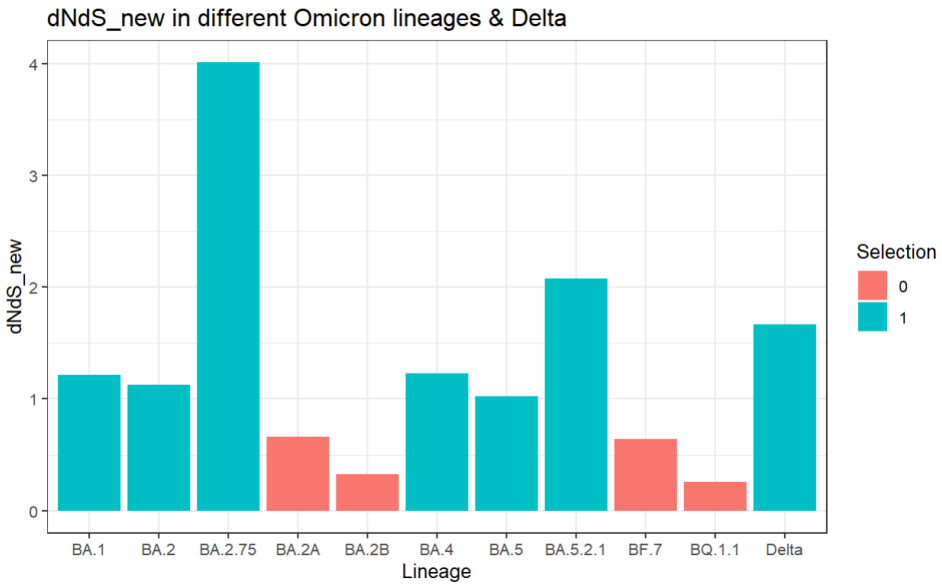


Figure 9: **The new dNdS in different Omicron lineages and Delta.**
Where we selected some mainstream Omicron subtypes and whole Delta data for calculation, quantifying the strength of their natural selection pressure. *new dNdS* > 1, which is positive selection, is marked in blue; *new dNdS* < 1, that is, a negative selection, and is marked in red.

The 7/11 lineages in Fig.9 are all subject to positive selection. Since the non-synonymous substitution increment of the BA.2 virus showed obvious differences around February 15, Fig.10, we divided BA.2 into two parts, BA.2A and BA.2B, and calculated *new dNdS*.

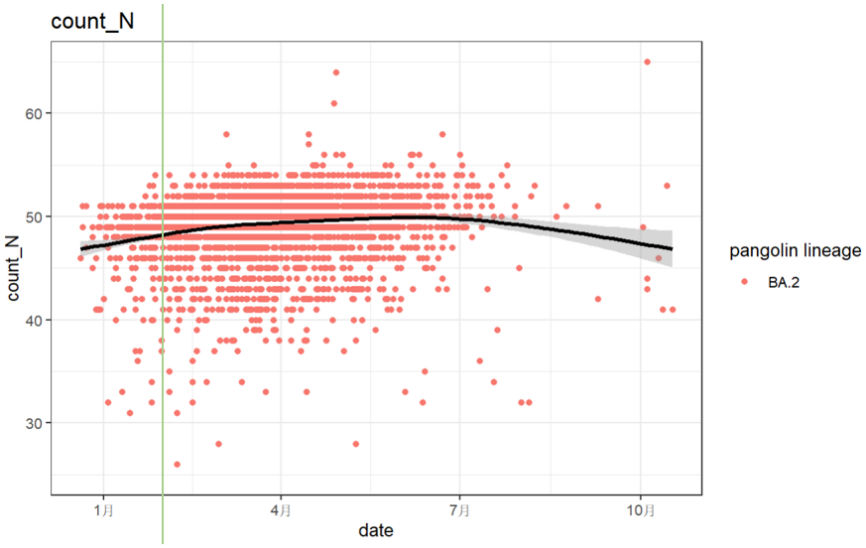
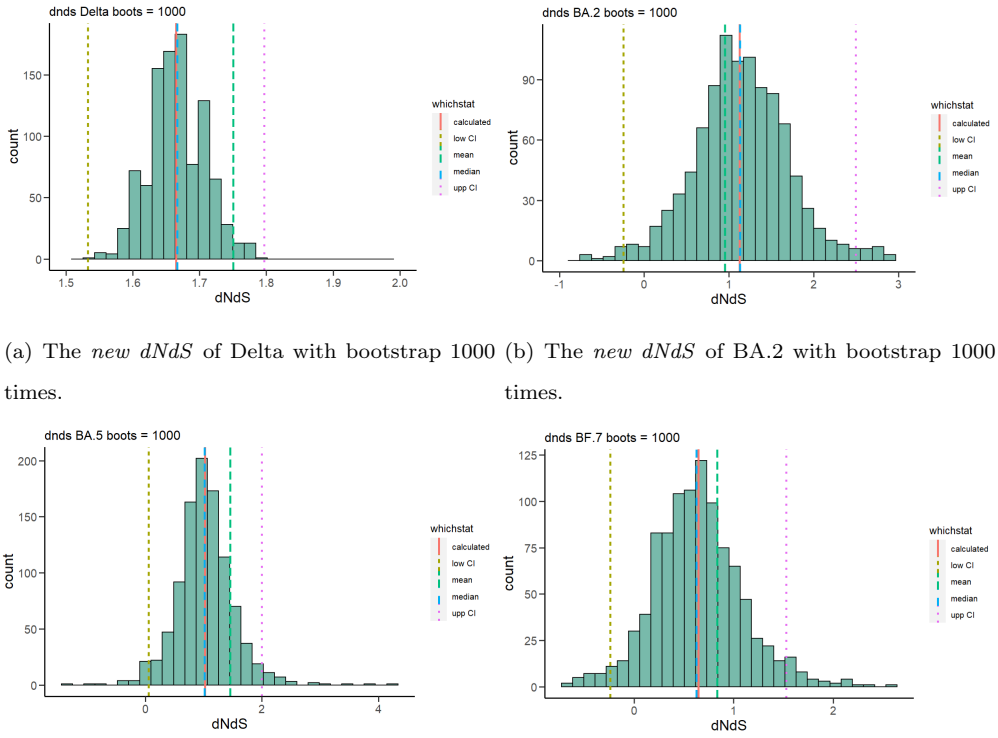


Figure 10: **Non-synonymous accumulation rate in BA.2 before Feb. 15 and after Feb. 15.**
Where the black line is a smooth spline to the non-synonymous accumulation rate of BA.2, and the green vertical line is the dividing line on February 15th. BA.2 on and before the 15th of February, is marked as BA.2A, and BA.2 after the 15th, of February, is marked as BA.2B.

Due to the limited number of samples of each lineage, there will be a large error when calculating dN and dS. We hope to see the error within an acceptable range, so we did Bootstrap sample processing on the data. We performed random sampling 1000 times on the Omicron lineages and Delta samples appearing in Fig.5 respectively, each time extracting the same amount of data of the sample size to obtain the *new dNdS* Fig.11.



(a) The *new dNdS* of Delta with bootstrap 1000 times.

(b) The *new dNdS* of BA.2 with bootstrap 1000 times.

(c) The *new dNdS* of BA.5 with bootstrap 1000 times.

(d) The *new dNdS* of BF.7 with bootstrap 1000 times.

Figure 11: **The dNdS of different lineages with t=1000 bootstrap.**

Where the x-axis represents the *new dNdS* value calculated by bootstrap, and the y-axis represents the number of occurrences of the dNdS value, which generally forms similar to normal distributions. The five lines are **red**: calculated by the original data, **green**: mean of *new dNdS* by Bootstrap, **blue**: median *new dNdS* by Bootstrap, **yellow**: Lower Bound of Confidence Interval, **purple**: Upper Bound of Confidence Interval. The four graphs respectively calculate the *new dNdS* by Bootstrap of Delta, BA.2, BA.5, and BF.7.

We then compared the *new dNdS* calculated for the original sample with the median of *new dNdS* calculated by Bootstrap Fig.12. Their values in detail are shown in Table.2. In the table, the median by bootstrapping and the *new dNdS* value obtained by direct calculation are relatively close, indicating that the *new dNdS* method estimates acceptable results. However, it should be noted that the accuracy of the Bootstrapping estimation statistics depends on the distribution characteristics of the sample data. Since the number of some variants in the dataset is small, the confidence interval of the Bootstrapping estimation result is relatively large.

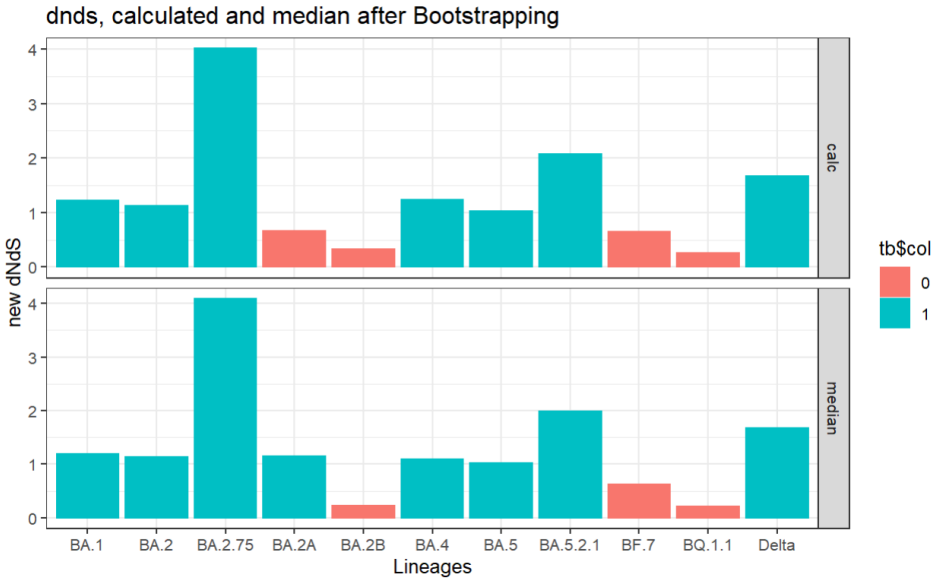


Figure 12: **The *new dNdS* of Calculated and Median by Bootstrap.**

Where the upper figure represents the *new dNdS* of each lineage calculated from the original sample, and the figure below represents the median of the *new dNdS* of each lineage calculated by Bootstrap. $dNdS > 1$, which is a positive selection, is marked in blue; $dNdS < 1$, which is a negative selection, and is marked in red.

	BA.2	BA.2A	BA.2B	BA.2.75	BA.4	BA.5	BA.5.2.1	BF.7	BQ.1.1
calculated	1.124543	0.661823	0.32689	4.011184	1.228655	1.024231	2.076075	0.64414	0.256505
mean	0.950299	1.014871	0.083325	2.680127	0.756001	1.450628	1.37681	0.835844	1.184935
median	1.131758	1.149846	0.225703	4.083851	1.091861	1.014175	1.98824	0.625818	0.209091
low CI	-0.24481	-2.4792	-1.39471	2.672016	0.094159	0.051279	1.140487	-0.23938	-2.51883
upp CI	2.493896	4.508945	1.561364	5.350353	2.363151	1.997182	3.011663	1.527662	3.031839

Table 2: **The dnds values of different lineages by Bootstrapping.**

We calculated the *new dNdS* of each lineage of Delta and Omicron in each country and counted the proportion of lineages with *new dNdS* > 1 in all lineages, 13. In the Delta group, the numbers of lineages of each country are: the United States (164), the United Kingdom (144), Japan (62), France (138), the Netherlands (97), South Korea (29); in the Omicron group, the numbers of lineages of each country are: United States (346), United Kingdom (306), Japan (223), France (230), Netherlands (191), South Korea (118).

In our selected countries, the number of lineages is not proportional to the number of cases, Table.3, Table.4. The proportion of positive selection is roughly proportional to the number of lineages. Japan in the Omicron group is an exception, but compared with Japan in the Delta group, the proportion of position selection is still proportional to the number of lineages.

	United States of America	United Kingdom	Japan	France	Netherlands	South Korea
dnds>1	0.5574713	0.5590062	0.3382353	0.472973	0.4074074	0.2941176
Kinds of Lineages	164	144	62	138	97	29
Number of Data	132065	73949	10695	14393	8535	8906

Table 3: **The proportions of positive selection in Delta lineages by countries.**

	United States of America	United Kingdom	Japan	France	Netherlands	South Korea
dnds>1	0.4754098	0.4006211	0.3803419	0.4031621	0.4433962	0.4328358
Kinds of Lineages	346	306	223	230	191	118
Number of Data	137226	63924	55725	17725	6190	4459

Table 4: **The proportions of positive selection in Omicron lineages by countries.**

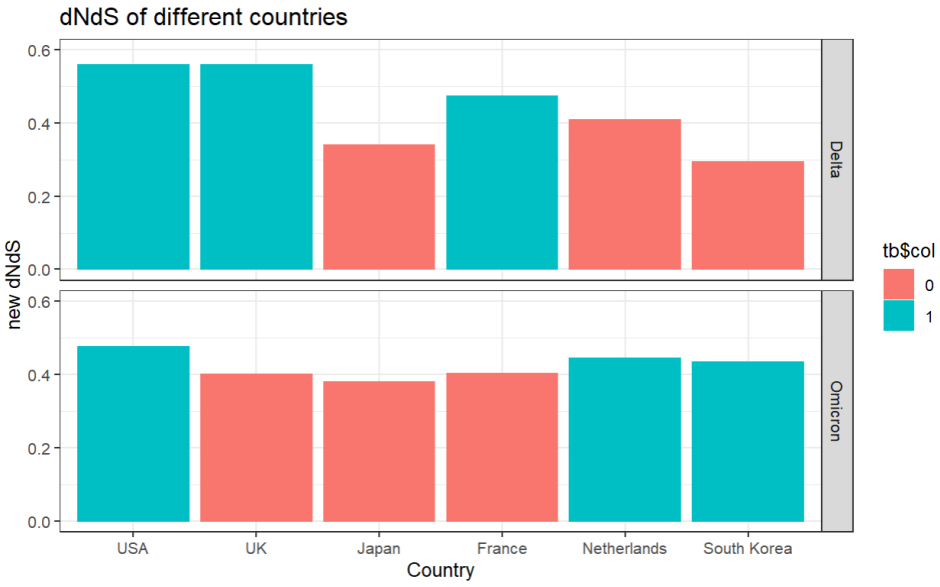


Figure 13: **The proportion of dNdS > 1 by country in Delta and Omicron groups.** Where the upper figure represents the proportions of positive selection on lineages in Delta group, and the lower figure represents the proportions of positive selection on lineages in Omicron group; **blue**: three countries with highest three proportion values; **red**: three countries with lowest three proportion values.

We selected several Variants of Concern(Alpha, Beta, Delta, Gamma, Omicron) and calculated the proportion of positive selection in lineages by VOCs. We observed that the proportion of early variants subject to positive selection was significantly higher than that of Omicron, Fig.14. The detailed values are shown in the Table.5.

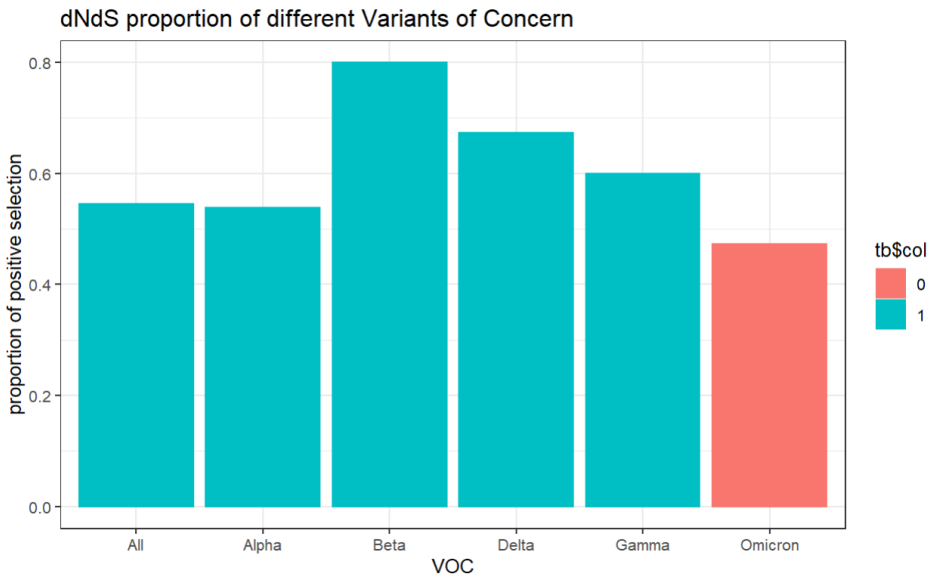


Figure 14: **The proportion of dNdS > 1 by VOCs.**

Where **blue**: with proportion > 0.5; **red**: with proportion < 0.5.

	All	Alpha	Beta	Delta	Gamma	Omicron
dnds>1	0.544837	0.538462	0.8	0.6733871	0.6	0.4721604
Kinds of Lineages	719	13	5	247	23	436
Number of Data	955905	16608	5771	453538	25591	421361

Table 5: **The proportions of positive selection on lineages by VOCs.**

5 Discussion

There are many lineages and variants worldwide. The adaptability of these different variants in the environment is different. Basepair substitutions that do not cause amino acid changes are called synonymous mutations and basepair substitutions that cause changes in encoded amino acids are also called non-synonymous substitutions. Synonymous substitution does not change the protein sequence, and it is generally believed that it does not change the fitness of the organism. Synonymous substitution is widely used to study mutation rate, effective population size, etc., and compared with non-synonymous substitution to study DNA natural selection Yang and Nielsen (2000).

We focused on the cumulative number of synonymous and non-synonymous substitutions and analyzed the relationship between their rate change and time in different lineages. In different lineages, it shows changes in the mutation growth rate. The time of mainstream replacement of newer strains is gradually shortened. For example, BF.7 shown in Fig.11 took 32 days to complete the mainstream replacement. The new subtype can replace BA.5, mainly because it is better at evading the immune protection established by humans through vaccination or previ-

ous infection, and thus has stronger infectivity, that is, it has a stronger ability to escape neutralizing antibodies. It coincides with other researches, that in two cohorts of individuals who received monovalent or bivalent mRNA vaccine boosters, BQ.1.1 NAb titers were 1/7 of BA.5 NAb titers, Miller et al. (n.d.).

As for the differences between different lineages, we mainly cared about the amino acid change caused by the accumulation of gene mutations, and the different adaptive performances of different variants. Gene mutations lead to changes in the expression of some amino acids, such as synonymous and non-synonymous, which in turn lead to changes in viral phenotypes. Although most mutations are deleterious and rapidly purged or relatively neutral, a small number of mutations can still seriously affect functional properties, Harvey et al. (2021b) and lead to changes such as infectivity, disease severity, escape antibodies elicited, Cao et al. (2022), etc. The current latest lineages have accumulated about 75 non-synonymous amino acid changes compared to the original strain. Different phenotypes will be affected by the environment, such as infected population, vaccination, medical condition level, etc., resulting in positive or negative selection, López-Cortés et al. (2022). In our analysis, the frequency of D614G, T478K, and S704L showed a significant trend, that these emerged twice as time passed. Based on our findings, we predicted the new lineages with P4715L change would become a more successful variant than the virus of the same lineage. The amino acid changes, as mentioned above, D614G, T478K, L452R, S704L, R346T, are the same.

Importantly, the *frequency by time* approach offers the possibility to quantify and predict the level of mutation. We observed that when a certain amino acid change frequency was high, it was not necessarily an advantageous mutation, but may also be mutated and selected along with other phenotypes, like "hitchhiking". However, by combining the data of Delta and Omicron, we can find out those traits whose frequency rises and fall during the Delta period. If they start to appear obviously in a certain lineage of Omicron, it is robust to say that such amino acid change is an advantageous phenotype. This type of amino acid change is exactly what we need to focus on.

We then quantified the strength of natural selection on different pangolin lineages, and thus judged whether the lineages were adapted to the environment or inhibited by the environment. Adaptation and Purifying selection have an impact on SARS-CoV-2 evolution at the same time, and those VOCs that we can pay attention to are because of their substitution rate in a short term increasing so that we can pay attention to some specific mutations, Tay, Porter, Wirth, and Duchene (2022). While slow within-variant evolution and rapid adaptive evolution produce new variants, this pattern only appears in non-synonymous changes, Neher (n.d.).

In our analysis, the proportion of *new dNdS* > 1 of Omicron lineages is 0.472, and the proportion of *new dNdS* 1 of All Variants is 0.545. Since the effective number of data collected by some lineages is too small, the bias is too large when doing linear regression. We extracted lineages (Omicron only) with a valid data volume, greater than 100, and recalculated *new dNdS*. The proportion of *new dNdS* 1 is 0.672. In Delta variants with an efficient number ¿ 100, this value is 0.698. One explanation is that the dataset we used collected 950,000 patient cases. Among these lineages, most of the lineages that can have more than 100 data are also subject to at least a certain degree of positive selection. We think that the proportion of dNdS 1 in all Omicron lineages should be between 0.538 and 0.672. During the three years of global SARS-CoV-2 governance, the measures taken against SARS-CoV-2, such as vaccination and quarantine, have increased the selection pressure. Changes in the environment have increased the pressure of natural selection and accelerated the rate of evolution. This is consistent with the

current phenomenon that Omicron variants have strong Neutralizing Antibody Escape.

The measurement of genetic diversity is highly sensitive to sample size, and populations with larger sample sizes tend to have higher genetic diversity and stronger adaptive evolution, Bashalkhanov, Pandey, and Rajora (2009). Positive selection will lead to the expansion of genetic differentiation. When the natural environment changes, we tend to see pervasive adaptive molecular evolution in multiple species, Booker, Jackson, and Keightley (2017). We selected three countries with the highest number of cases in the data set: the United States (297068 samples), the United Kingdom (141897 samples), Japan (67894 samples); and the three countries with relatively small numbers: France (34067 samples), the Netherlands (15874 samples), South Korea (13435 samples), and calculate their *new dNdS*, Fig.13. Considering that the absolutely small sample size will lead to a large error in the calculation of *new dNdS*, we chose the three countries with a relatively small number of countries that still rank high in this data set, but they are still 1/10 of the countries with the highest numbers. Due to factors such as globalization and high population mobility, some lineages do not appear to be significantly differentiated between countries. However, due to the differences in culture, policies, and medical capabilities of various countries, *new dNdS* still have differences.

In the pandemics of SARS-CoV-2, the order of the original strain, Alpha, Beta, Delta, Gamma, and Omicron replaced each other. In the second half year of 2020 of the spread of the VOCs of SARS-CoV-2 with Alpha, Beta and Gamma variants, the most response was Quarantine. In the Delta period, from the second half of 2020 to December 2021, the major responses were vaccination + isolation; and in the Omicron period, from November 2021 till the present, people have been experiencing "coexistence" + vaccination countermeasures. Under different circumstances, the degrees of adaptive evolution of SARS-CoV-2 variants are different. The main result is that the positive selection of Beta and Delta is the highest, and the positive selection of Omicron is the lowest, Table.5. We think that the positive selection effect of quarantine + vaccination on SARS-CoV-2 was statistically significant, followed by the positive selection effect of quarantine, and the positive selection effect of vaccination is the lowest.

In summary, our study focuses on the strength of natural selection of different lineages in the genetic diversity of SARS-CoV-2 and explains the reasons for divergence and convergence through the frequency of amino acid change. We provide a method to quantify the selection pressure of different lineages and provide a method to detect whether the amino acid change frequency changes significantly to become a new variant. We calculated the strength of selection pressure on the lineages of Omicron, and all the most concerned lineages of Omicron showed positive selection by the environment, which is consistent with the high rate of differentiation and outbreaks of new variants of these lineages. By calculating the different Variants of Concern and variants in various countries, we found that the positive selection presented different selection pressures in various countries and VOCs. The characteristic result is that the selection pressures received by the early strains were higher than those of the closer strains; the selection pressures are proportional to the number of lineages. For the difference of mutations of diverse lineages, we analyzed the changing trend of the frequency of the corresponding amino acid changes. Frequencies have behaved differently in different lineages over time, and the rise and fall of some frequencies of amino acid change can obviously cause them to adapt to the environment or be eliminated by the environment. The consistency of some amino acid changes through some lineages can also prove the existence of convergence evolution.

Our results provide a useful tool for understanding the dynamic driving force of the SARS-CoV-2 virus and will help to further understand the evolution of SARS-CoV-2. The method proposed in this paper can be used to monitor the mutation of SARS-CoV-2 and can be used to detect whether the mutation frequency of a certain amino acid changes significantly become a new Variant of Concern.

However, the sequences included in this analysis were relatively limited in both sample and loci size. In particular, time scales may be relatively short for the analysis of population statistics events. Further studies with larger, multi-site and longer timescale samples may be required to confirm the results and generalize the findings.

6 Acknowledgement

We highly appreciate all staff and classmates at Bioinformatics Research Center of Aarhus University for the technical support and discussion.

We especially acknowledge Palle Villesen for his collaboration and instruction.

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

References

Amicone, M., Borges, V., Alves, M. J., Isidro, J., Zé-Zé, L., Duarte, S., ... Gordo, I. (2022, 1). Mutation rate of sars-cov-2 and emergence of mutators during experimental evolution. *Evolution, Medicine, and Public Health*, *10*, 142-155. DOI: 10.1093/emph/eoac010

Bashalkhanov, S., Pandey, M., & Rajora, O. P. (2009, 12). A simple method for estimating genetic diversity in large populations from finite sample sizes. *BMC Genetics*, *10*, 84. DOI: 10.1186/1471-2156-10-84

Baum, A., Fulton, B. O., Wloga, E., Copin, R., Pascal, K. E., Russo, V., ... Kyratsous, C. A. (2020, 8). Antibody cocktail to sars-cov-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science*, *369*, 1014-1018. DOI: 10.1126/science.abd0831

Booker, T. R., Jackson, B. C., & Keightley, P. D. (2017, 12). Detecting positive selection in the genome. *BMC Biology*, *15*, 98. DOI: 10.1186/s12915-017-0434-y

Callaway, E. (2021, 8). The mutation that helps delta spread like wildfire. *Nature*, *596*, 472-473. DOI: 10.1038/d41586-021-02275-2

Cao, Y., Jian, F., Wang, J., Yu, Y., Song, W., Yisimayi, A., ... wangyc, Y. W. (n.d.). Imprinted sars-cov-2 humoral immunity induces convergent omicron rbd evolution. Retrieved from https://doi.org/10.1101/2022.09.15.507787 DOI: 10.1101/2022.09.15.507787

Cao, Y., Yisimayi, A., Jian, F., Song, W., Xiao, T., Wang, L., ... Xie, X. S. (2022, 8). Ba.2.12.1, ba.4 and ba.5 escape antibodies elicited by omicron infection. *Nature*, *608*, 593-602. DOI: 10.1038/s41586-022-04980-y

Domingo, E., García-Crespo, C., Lobo-Vega, R., & Perales, C. (2021, 9). Mutation rates, mutation frequencies, and proofreading-repair activities in rna virus genetics. *Viruses*, *13*, 1882. DOI: 10.3390/v13091882

Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., ... Robertson, D. L. (2021a, 7). Sars-cov-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, *19*, 409-424. DOI: 10.1038/s41579-021-00573-0

Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., ... Robertson, D. L. (2021b, 7). Sars-cov-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, *19*, 409-424. DOI: 10.1038/s41579-021-00573-0

Hou, Y., Zhao, S., Liu, Q., Zhang, X., Sha, T., Su, Y., ... Chen, H. (2022, 6). Ongoing positive selection drives the evolution of sars-cov-2 genomes. *Genomics, Proteomics Bioinformatics*. DOI: 10.1016/j.gpb.2022.05.009

Hou, Y. J., Chiba, S., Halfmann, P., Ehre, C., Kuroda, M., Dinnon, K. H., ... Baric, R. S. (2020, 12). Sars-cov-2 d614g variant exhibits efficient replication ex vivo and transmission in vivo. *Science*, *370*, 1464-1468. DOI: 10.1126/science.abe8499

Ito, J., Suzuki, R., Uriu, K., Itakura, Y., Zahradnik, J., Deguchi, S., ... Ikeda, T. (n.d.). Conver-

gent evolution of the sars-cov-2 omicron subvariants leading to the emergence of bq.1.1 variant. *Gideon Akatsuki Saito*, *11*, 23. Retrieved from https://doi.org/10.1101/2022.12.05.519085 DOI: 10.1101/2022.12.05.519085

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., ... Wyles, M. D. (2020, 8). Tracking changes in sars-cov-2 spike: Evidence that d614g increases infectivity of the covid-19 virus. *Cell*, *182*, 812-827.e19. DOI: 10.1016/j.cell.2020.06.043

López-Cortés, G. I., Palacios-Pérez, M., Veledíaz, H. F., Hernández-Aguilar, M., López-Hernández, G. R., Zamudio, G. S., & José, M. V. (2022, 5). The spike protein of sars-cov-2 is adapting because of selective pressures. *Vaccines*, *10*, 864. DOI: 10.3390/vaccines10060864

Miller, J., Hachmann, N. P., Collier, A.-R. Y., Lasrado, N., Mazurek, C. R., Patio, R. C., ... Barouch, D. H. (n.d.). Substantial neutralization escape by the sars-cov-2 omicron variant bq.1.1 authors for print edition. Retrieved from https://doi.org/10.1101/2022.11.01.514722 DOI: 10.1101/2022.11.01.514722

Neher, R. A. (n.d.). Contributions of adaptation and purifying selection to sars-cov-2 evolution. Retrieved from https://doi.org/10.1101/2022.08.22.504731 DOI: 10.1101/2022.08.22.504731

Obbard, D. J., Welch, J. J., Kim, K.-W., & Jiggins, F. M. (2009, 10). Quantifying adaptive evolution in the drosophila immune system. *PLoS Genetics*, *5*, e1000698. DOI: 10.1371/journal.pgen.1000698

Shi, A. C., & Xie, X. (2021, 7). Making sense of spike d614g in sars-cov-2 transmission. *Science China Life Sciences*, *64*, 1062-1067. DOI: 10.1007/s11427-020-1893-9

Tay, J. H., Porter, A. F., Wirth, W., & Duchene, S. (2022, 2). The emergence of sars-cov-2 variants of concern is driven by acceleration of the substitution rate. *Molecular Biology and Evolution*, *39*. DOI: 10.1093/molbev/msac013

Yang, Z., & Nielsen, R. (2000, 1). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution*, *17*, 32-43. DOI: 10.1093/oxfordjournals.molbev.a026236