Evolving Fitness and Immune Escape: A Retrospective Analysis of SARS-CoV-2 Spike Protein (2020-2024) Using Protein Language Model

Sihua Peng1,3, Ludy Registre Carmola1,3, Sachin Subedi1,2,3, M H M Mubassir1,2,3, Mohamed Bakheet1,3, Leke Lyu1,2, and Justin Bahl1,2,3,4\*†

1 Center for Ecology of Infectious Diseases, University of Georgia, Athens, GA 30602.

2Institute of Bioinformatics, University of Georgia, Athens, GA 30602.

3Department of Infectious Diseases, University of Georgia, Athens, GA 30602.

4Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA 30602.

†Corresponding author

\*Email: justin.bahl@uga.edu

Abstract

Background: The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has led to unprecedented global health challenges. Understanding the virus's evolutionary dynamics, particularly its fitness and ability to escape immune responses, is crucial for developing effective public health strategies. Utilizing new innovations in advanced protein language models, this study examines the influence of genetic variations on the virus's adaptability and its resistance to immunological defenses.

Methods: We employed the CoVFit model to predict the Fitness and Immune Escape Index (IEI) values. The dataset comprises over 2,504,278 cases /160,892 variant sequences, providing a broad base for observing temporal and geographical variations in viral behavior.

Results: The findings indicate significant increases in both fitness and immune escape indices over the study period, suggesting continuous viral adaptation despite extensive public health interventions and vaccination efforts. Notably, North America, with the largest volume of sequence data, exhibited the most significant evolutionary changes, which may be indicative of the virus transitioning towards a more stable phase of evolution.

Conclusions: The use of protein language models like CoVFit has proven instrumental in offering deeper insights into the molecular dynamics of SARS-CoV-2 evolution. By accurately predicting the impacts of specific mutations, these models serve as crucial tools for forecasting future viral behavior and aiding in vaccine development. The results underscore the importance of continued surveillance and the adaptation of vaccine strategies to address the evolving challenges posed by the virus, highlighting the critical role of advanced computational tools in pandemic preparedness and response.

Keywords: SARS-CoV-2; spike protein; protein language models; protein fitness; immune escape; retrospective analysis.

1. Introduction

The COVID-19 pandemic, caused by SARS-CoV-2, has presented an unprecedented global challenge1. A key element in the virus's ability to infect humans is the SARS-CoV-2 spike (S) protein, which facilitates viral entry by binding to the ACE2 receptor in human cells. This protein is also the primary target for most vaccines, making it the focus of intense study. Given its pivotal function, it is crucial to comprehend how the S protein evolved since it has an immediate effect on the virus's ability to infect hosts and escape immune responses.

In this perspective, one of the most important concerns is: What drives the evolution of SARS-CoV-2? The virus's evolutionary trajectory is significantly shaped by a number of elements, such as immune system evasion, environmental stresses, and genetic alterations. In addition to being a scholarly endeavor, tracking these evolutionary shifts is crucial for creating adaptive tactics, such as novel vaccinations and medical treatments.

A comprehensive exploration of the fitness and immune escape landscape is crucial for several reasons 2,3. First, it aids in forecasting how the virus may change in reaction to broad immunity, whether from vaccination or natural infection. Second, it informs us of the virus’s adaptability—how efficiently it can spread despite these immune pressures. Finally, understanding virus's evolutionary dynamics is critical for public health preparedness 4, as it enables proactive measures against potential future variants.

This study seeks to retrospectively analyze the fitness and immune escape trends of the SARS-CoV-2 S protein using advanced protein language models 5. By examining the evolutionary history of this critical viral component, we aim to shed light on how mutations have shaped the virus’s fitness and immune evasion capabilities from 2020 to 2024.

The advent of protein language models (PLMs) has revolutionized computational biology, offering unprecedented insights into protein structure, function, and interactions 6. Drawing inspiration from natural language processing (NLP) 7,8, these models treat amino acid sequences analogously to sentences, leveraging advanceed machine learning techniques to predict various protein attributes. The core idea is to harness large-scale sequence data to learn representations that capture the biochemical and biophysical properties of proteins, much like how language models capture semantic and syntactic structures of text 9,10.

The concept of using machine learning to analyze protein sequences dates back to the early applications of neural networks in bioinformatics 11. However, it was the introduction of transformer-based architectures that truly marked a turning point 12. Rao et al. presented the first transformer-based model specifically tailored for protein sequences. Their model, known as TAPE (Tasks Assessing Protein Embeddings), adapted BERT (Bidirectional Encoder Representations from Transformers) for the biological domain, achieving remarkable performance across various protein prediction tasks 13. Elnaggar et al. further extended the capabilities of transformer models with ProtTrans, a suite of transformer models trained on the largest collection of protein sequences available at the time. This work highlighted the scalability of transformers and their ability to learn meaningful representations across diverse protein families 14. Rives et al. introduced ESM-1b, a transformer model trained on the entire UniProt database. This model showcased the effectiveness of large-scale training and transfer learning, significantly improving performance on downstream tasks such as secondary structure prediction and remote homology detection 15.

Some researchers have developed methods for predicting the protein fitness based on variant patterns, using statistical modeling approaches 16-18. However, these statistical models represent protein fitness merely as a linear combination of individual mutation effects and do not account for interactions between mutations. Recently, Ito J. and others addressed this issue using protein language models 19.

Recently, Lin et al. utilized large language model technology to directly infer complete atomic-level protein structures from amino acid sequences. They developed the ESM-2 protein language model, which boasts up to 15 billion parameters 20. Ito et al. fine-tuned the ESM-2 model using genotype-fitness data and mutation effect information on evasion ability from humoral immunity, determined by high-throughput deep mutational scanning (DMS) experiments 21,22. They developed the CoVFit model. Using the CoVFit model, they mapped the fitness landscape of SARS-CoV-2.

DMS experiments are a biological technique used to systematically study the functional effects of all possible mutations in protein or nucleic acid sequences 23. This method combines high-throughput screening techniques with deep sequencing technologies, enabling the rapid and accurate analysis of the phenotypes and functional characteristics of a large number of mutants, so we can use DMS experiments to study and measure the ability of mutations in protein sequences to evade humoral immunity (such as antibody responses) 24.

It is worth noting that after the launch of ESM-2, ESM-3 has now been released. The largest pre-trained model of ESM-3 has 98 billion parameters, however currently, only a smaller pre-trained model with 1.4 billion parameters has been made open source 25.

In this study, we employed the most recently trained CoVFit model to investigate the fitness and immune evasion levels of the SARS-CoV-2 S protein sequences collected from April 1, 2024, to May 15, 2024. Additionally, as a focal point of our research, we conducted a retrospective study on the entire evolutionary history of SARS-CoV-2 from early 2020 to the present. We reviewed the evolutionary dynamics of SARS-CoV-2 fitness and immune evasion capabilities and identified unique patterns and key features of the fitness and immune evasion abilities of the S protein.

2. Materials and methods

2.1 Data Collection

SARS-CoV-2 S protein sequences were downloaded from NCBI SARS-CoV-2 Data Hub (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType\_s=Nucleotide&VirusLineage\_ss=taxid:2697049). Data containing more than five unidentified characters were excluded. To analyze the evolutionary trajectory of SARS-CoV-2 fitness and immune evasion, we segmented the period from January 1, 2020 to May 15, 2024 into three-month intervals and downloaded the S protein amino acid sequences for each interval.

A total of 2,504,278 S protein sequences were downloaded from NCBI, and 160,892 variants were obtained from the above downloaded sequences (Supplementary Table 1).

2.2 Variant sequence identification and mutation count

In the epidemiology of SARS-CoV-2, the terms "mutation," "variant," and "strain" are often used interchangeably, but they have distinct scientific meanings. A mutation refers to an actual change in the sequence, such as D614G, which is the substitution of aspartic acid for glycine at position 614 of the spike protein 26. Genomes differing in sequence are typically referred to as variants. This term can be vague since two variants might differ by one or many mutations. A variant is considered a strain when it exhibits a significantly different phenotype, such as differences in antigenicity, transmissibility, or virulence 27,28. This paper discusses variants as sequences of the spike protein that differ by one or more mutations. Some of these variants may not show differences in antigenicity, transmissibility, or virulence, but any two variants referred to in this text do differ in their amino acid sequences, commonly termed variant sequences or simply as variants. In this study, we define “variants” as spike protein S amino acid sequences that differ by one or more mutations.

The downloaded S protein data over various time periods contained many duplicate sequences. Duplicate sequences within each of the 18 time periods were removed. As a result, only one sample of each unique amino acid sequence was retained within each time period, meaning that the non-duplicated S protein amino acid sequences are the variants within a given time period. However, some samples might be duplicated across different time periods.

During each time period, we extracted variants from all cases. For each variant, the positions and frequencies of mutations are different. To study the changes in mutation frequency over time. Therefore, we count the occurrence frequency of mutations in each variant sequence. A cubic spline interpolation was used to connect the mean values of mutation counts for each time segment. The code developed to remove duplications, and count occurrence and frequency of mutations is available at https://github.com/pengsihua2023/VFIEI-SARS-cov-2/tree/main/code.

2.3 CoVFit Protein Language Model

Ito et al. first established the ESM-2 Coronaviridae model 19 by pre-training (i.e., domain adaptation) on S protein sequences obtained from 1,506 types of coronaviruses. They then fine-tuned the model on genotype–fitness (Re) data and DMS data to evaluate antibody neutralization escape capabilities. Consequently, for a given S protein sequence, CoVFit can predict the fitness value in a specific country and its ability to escape from each monoclonal antibody (mAb). Their dataset included data from 17 countries: Australia, Belgium, Brazil, Canada, Denmark, France, Germany, India, Italy, Japan, the Netherlands, South Korea, Spain, Sweden, Switzerland, the United Kingdom, and the United States.

We adopted the CoVFit protein model (<https://github.com/TheSatoLab/CoVFit>). The model files were downloaded from Zenodo (<https://zenodo.org/records/10911205>). CoVFit model was finetuned from ESM-2 protein model using two data sets: 1) genotype–fitness data, obtained from virus genome surveillance and ii) mutation effect data on evasion ability from humoral immunity, determined by high-throughput deep mutational scanning (DMS) experiments. Then, using a multinomial logistic model fitted to genome surveillance data from GISAID (<https://gisaid.org/>) up to November 2, 2023, the Re of each genotype in each country was estimated. As a result, 21,751 genotype-fitness data points were obtained, covering 12,914 genotypes across the 17 countries, indicating that the fitness of S protein originally derived from publicly available data set of Effective Reproduction Number (Re), which is an epidemiological parameter used to measure the transmission potential of an infectious disease within a specific population 29,30. Finally, the derived fitness data were scaled to 0-1 for the convenience of model training.

We used Clustal Omega-1.2.3 31 to perform multiple sequence alignment of amino acid sequences to input the aligned sequences into the CoVFit model.

2.4 Fitness values analysis

When using the CoVFit model, aligned S protein sequences are inputted into the model, which then enables the prediction of fitness values for the S protein. The average fitness value for North America was derived as the average of the United States and Canada. For Europe, the average was derived from Germany, the United Kingdom, Switzerland, Sweden, Spain, the Netherlands, Italy, France, Denmark, and Belgium. The average of South Korea, Japan, and India represent was used for Asia. The Fitness value of Australia was regarded as the average for Oceania, and the fitness value of Brazil was regarded as the average for South America.

Since the number of genotypes from Africa did not exceed 300, the CoVFit model did not include data from African countries during training. Consequently, there are no fitness prediction values for Africa. For the sake of completeness, the average fitness prediction values of the 17 countries were used as the fitness average for Africa. Therefore, when using CoVFit for model predictions, the S protein sequence data from Africa was input, and the average fitness values of the 17 countries were used as the fitness average for Africa.

2.5 Immune Escape Index (IEI)

The CoVFit model can predict the relative binding affinities of different monoclonal antibodies (mAbs). By inputting a SARS-CoV-2 S protein sequence, the model can predict the relative binding affinities for 1,548 mAbs. We averaged these predicted affinity values and denoted this average as the Immune Escape Index (IEI), which describes the immune escape capability of a variant. The higher the IEI, the greater the immune escape capability of the variant.

2.6 Prediction study on the fitness and immune escape of the S protein in recent prevalent variants in North America

A total of 2,170 S protein amino acid sequences, with 543 variants, were downloaded from the United States between April 1, 2024, and May 15, 2024. We then conducted a prediction study on the fitness and immune escape capabilities of all the variants' S proteins. It should be noted that the downloaded U.S. case samples actually represent the sample size for all of North America.

2.7 Global retrospective analysis of S protein fitness and immune escape

We divided the time periods into three-month intervals and downloaded the S protein data for six continents from January 1, 2020, to May 15, 2024, removing duplicate sequences from the same time-period. We then conducted retrospective studies on the fitness and immune escape capabilities of the S proteins for each period and continent.

The temporal trends in fitness and immune escape indices were analyzed, and the geographical distribution of SARS-CoV-2 variants across different continents were examined.

3. Results

3.1 Prediction study in America: January 1, 2024, to May 15, 2024

The CoVFit model was used to predict the Fitness and IEI values for the 543 variants in the United States. Table 1 displays the top 20 highest fitness values. The case with highest fitness value was from a sample in Iowa. The difference in fitness values among these 20 variants was small, ranging from 0.921 to 0.925, and the differences in IEI were also minor, ranging from 0.581 to 0.589. Among these 20 variants, 11 were identified as SARS-CoV-2 lineages. Some lineages contained multiple samples with different mutations, for example, three samples were included in lineages JN.1.16 l, JN.1.11.1, and JN.1.7, respectively (Table 1), indicating that different numbers of mutations can lead to different fitness values. This may be because different amino acid mutations may significantly alter the protein conformation, affecting the binding affinity of the S protein to the human ACE2 receptor.

Compared to the original samples discovered in Wuhan, the top 20 SARS-CoV-2 variants in North America have shown higher adaptability and immune evasion capabilities, suggesting that the virus has reached a higher level of adaptation to human hosts. However, the pace of the virus's evolution has slowed, likely because it has found a relatively stable state of adaptation within the current biological and social environments. Nevertheless, whether it will gradually evolve into a seasonal flu virus requires ongoing observation of the virus's long-term behavioral patterns and its impact on public health. Thus, claiming that the virus will evolve into a seasonal flu remains premature and requires further scientific evidence to support such a prediction.

Table 1 Top 20 predicted Fitness values and IEI for SARS-CoV-2 variants in the United States

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Accession | Lineage | Collection date | Fitness | IEI |
| XAN64366.1 | JN.1.16 | 2024-04-29 | 0.925 | 0.585 |
| XBA97060.1 | JN.1.11.1 | 2024-05-09 | 0.924 | 0.584 |
| XAW33708.1 | JN.1.16 | 2024-05-06 | 0.924 | 0.584 |
| XAU78949.1 | JN.1.11.1 | 2024-04-20 | 0.924 | 0.584 |
| XAU78842.1 | JN.1.7 | 2024-04-17 | 0.923 | 0.584 |
| XAJ36120.1 | JN.1.11.1 | 2024-04-21 | 0.923 | 0.584 |
| XAW33862.1 | JN.1.4.2 | 2024-04-29 | 0.923 | 0.586 |
| XAJ04662.1 | JN.1.9 | 2024-04-10 | 0.923 | 0.585 |
| XAN64414.1 | JN.1 | 2024-04-25 | 0.922 | 0.589 |
| XAW33674.1 | BA.2.86.1 | 2024-05-04 | 0.922 | 0.589 |
| XAO61989.1 | JN.1 | 2024-04-02 | 0.922 | 0.581 |
| XAJ04710.1 | XDD | 2024-04-14 | 0.922 | 0.588 |
| XAU78878.1 | JN.1.9 | 2024-04-24 | 0.921 | 0.588 |
| WZH70794.1 | JN.1.16 | 2024-04-08 | 0.921 | 0.581 |
| XBA97012.1 | JN.1.7.2 | 2024-04-30 | 0.921 | 0.588 |
| XAW19132.1 | JN.1.8.1 | 2024-04-17 | 0.921 | 0.588 |
| XAJ29041.1 | JN.1.7 | 2024-04-16 | 0.921 | 0.588 |
| XAU78770.1 | JN.1.7 | 2024-04-05 | 0.921 | 0.588 |
| XAU78782.1 | JN.1.4 | 2024-04-03 | 0.921 | 0.588 |
| XAU78794.1 | JN.1.4 | 2024-04-03 | 0.921 | 0.588 |

We then conducted a historical examination of the 11 lineages, identifying all instances of these lineages among global S protein variants from 2020 to 2024 and tallied the occurrences of different mutations within the same lineage (Table 2). We found that the JN.1 lineage had the highest number of mutations, reaching 1,082, while JN.1.7.2 had the fewest, with only 10 mutations. In some lineages, there is an obvious difference between the fitness and IEI of different mutations. For example, the highest fitness for JN.1.16 is 0.914, and the lowest is 0.863 (Table 2).

Table 2 Historical examination of fitness and IEI across 11 lineages of SARS-CoV-2 variants

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Variants | Max Fitness/Date/Country Accession | Min Fitness/Date/Country  Accession | Max IEI/Date/Country  Accession | Min IEI/Date/Country  Accession |
| JN.1.16 | 88 | 0.914|2024-03-19|USA  WZC46762.1 | 0.863|2024-05-02|USA  XAX97799.1 | 0.603|2024-03-19|USA  WZC46762.1 | 0.555|2024-05-02|USA  XAX97799.1 |
| JN.1.11.1 | 33 | 0.908|2024-05-01|USA XBA92584.1 | 0.885|2024-05-11|USA  XBG56702.1 | 0.655|2024-04-22|USA  XAN95867.1 | 0.630|2024-05-11|USA  XBG56702.1 |
| JN.1.7 | 143 | 0.913|2024-04-15|USA  XAN71841.1 | 0.827|2024-04-01|USA  WZH69489.1 | 0.616|2024-05-15|USA  XBL92471.1 | 0.553|2024-04-01|USA  WZH69489.1 |
| JN.1.4.2 | 99 | 0.915|2024-05-15|USA  XBL92342.1 | 0.913|2023-12-07|USA  WQQ58013.1 | 0.615|2024-02-15|USA  WWQ17088.1 | 0.563|2024-01-16|USA  WWZ24569.1 |
| JN.1.9 | 63 | 0.907|2024-04-10|USA  XAJ04662.1 | 0.871|2024-02-14|USA  WWQ16291.1 | 0.599|2024-01-17|USA  WVH24328.1 | | 0.569|2024-02-14|USA  WWQ16291.1 |
| JN.1 | 1082 | 0.890|2023-12-09|USA  WVH05687.1 | 0.811|2024-01-28|USA  WVW94984.1 | 0.623|2023-12-23|USA  WWQ99983.1 | 0.506|2024-03-29|USA  WYX99223.1 |
| BA.2.86.1 | 69 | 0.903|2024-01-02| USA  WWZ24049.1 | 0.860|2024-02-03| USA  WWB25296.1 | 0.624/2024-01-22|Japan  BFH89706.1 | 0.582|2024-02-03|USA  WWB25296.1 |
| XDD | 16 | 0.903|2024-04-14|USA  XAJ04710.1 | 0.899|2024-01-19|USA  WZC33827.1 | 0.601|2024-02-15|USA  WXH69581.1 | 0.588|2023-12-22|USA  WVQ25143.1 |
| JN.1.7.2 | 10 | 0.904|2024-04-30|USA  XBA04631.1 | 0.899|2024-03-19|USA  WYX95543.1 | 0.594|2024-04-04|USA  XAO62227.1 | 0.580|2024-03-19|USA  WYX95543.1 |
| JN.1.8.1 | 121 | 0.902|2024-04-22|USA  XAN95821.1 | 0.800|2024-03|USA  WYK03097.1 | 0.680|2024-02-12|USA  WWZ27898.1 | 0.585|2023-12-26|USA  WRO52942.1 |
| JN.1.4 | 562 | 0.914|2023-12-14|USA  WRI41262.1 | 0.748|2024-01-03|USA  WXB52480.1 | 0.701|2024-01-05|USA  WWA06107.1 | 0.598|2024-01-23|USA  WVO05100.1 |

In summary, we concluded that the above 20 variants are the more concerning variants in North America, and the variant with the accession number XAN64366.1 is the most concerning variant belonging to the JN.1.16 lineage.

3.2 Retrospective study from January 1, 2020, to May 15, 2024

We analyzed the distribution of lineages and mutations within the variants, and identified the dominant lineage, mean mutations per variant, maximum and minimum of fitness, and IEI for the variants, as well as its percentage among all lineages during each specific time period. Among the six continents, North America accounted for 95.46% of the cases and 89.66% of the variants. Therefore, we primarily present the analysis results for North America here (Table 4). The results of the other five continents are provided in the supplementary material (Table S2-S6).

Table 3 Profiles of SARS-cov-2 variants in North America: lineage, mutation details, Fitness, and IEIs

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Case /  Variant sequence | Dominant Lineage | Dominant Percentage | Unique  Lineages | MMut | MaxFit | MFit | MaxIEI | MIEI |
| Jan-Mar, 2020 | 10,055/400 | B.1 | 37.5% | 73 | 2 | 0.963 | 0.227 | 0.591 | 0.171 |
| Apr-Jun, 2020 | 19,662/968 | B.1 | 35.85% | 154 | 2 | 0.325 | 0.211 | 0.373 | 0.164 |
| Jul-Sep, 2020 | 18,434/1,159 | B.1 | 16.65% | 169 | 2 | 0.329 | 0.212 | 0.513 | 0.164 |
| Oct-Dec, 2020 | 45,478/3,264 | B.1.2 | 30.91% | 213 | 3 | 0.531 | 0.229 | 0.384 | 0.196 |
| Jan-Mar, 2021 | 140,838/10,916 | B.1.2 | 25.53% | 283 | 6 | 0.534 | 0.286 | 0.466 | 0.258 |
| Apr-Jun, 2021 | 177,356/11,233 | B.1.1.7 | 39.6% | 220 | 14 | 0.534 | 0.286 | 0.442 | 0.297 |
| Jul-Sep, 2021 | 382,456/21,392 | AY.44 | 11.53% | 203 | 20 | 0.645 | 0.363 | 0.372 | 0.254 |
| Oct-Dec, 2021 | 461,734/25,974 | AY.103 | 17.9% | 210 | 28 | 0.694 | 0.421 | 0.387 | 0.275 |
| Jan-Mar, 2022 | 247,674/9,632 | BA.1.1 | 39.11% | 185 | 49 | 0.773 | 0.527 | 0.436 | 0.313 |
| Apr-Jun, 2022 | 261,252/8,397 | BA.2.12.1 | 31.19% | 240 | 39 | 0.795 | 0.642 | 0.451 | 0.367 |
| Jul-Sep, 2022 | 253,298/11,225 | BA.5.2.1 | 12.71% | 390 | 40 | 0.806 | 0.715 | 0.459 | 0.398 |
| Oct-Dec, 2022 | 147,966/11,044 | BQ.1.1 | 8.82% | 626 | 42 | 0.837 | 0.755 | 0.476 | 0.424 |
| Jan-Mar, 2023 | 90,796/8,895 | XBB.1.5 | 23.66% | 706 | 53 | 0.891 | 0.780 | 0.518 | 0.441 |
| Apr-Jun, 2023 | 22,424/3,707 | XBB.1.5 | 27.65% | 562 | 50 | 0.901 | 0.808 | 0.532 | 0.459 |
| Jul-Sep, 2023 | 38,369/5,797 | FL.1.5.1 | 5.49% | 629 | 64 | 0.911 | 0.857 | 0.532 | 0.489 |
| Oct-Dec, 2023 | 44,399/6,486 | HV.1 | 14.75% | 534 | 63 | 0.924 | 0.895 | 0.551 | 0.522 |
| Jan-Mar, 2024 | 26,275/3,216 | JN.1 | 27.92% | 247 | 72 | 0.913 | 0.901 | 0.545 | 0.536 |
| Apr-May, 2024 | 2,170/543 | JN.1 | 15.65 | 46 | 68 | 0.940 | 0.930 | 0.563 | 0.555 |

Mean Mutation: MMut; Maximum Fitness: MaxFit; Mean Fitness: MFit;

Maximum Immune Escape Index: MaxIEI; Mean Immune Escape Index: MIEI

3.2.1 Evolution of Mutations in S Protein Variant Sequences

We found that the distribution of the number of mutations per variant during the entire SARS-CoV-2 pandemic could be divided into three distinct stages (Figure 1, Figure S1-S5). In North America, the first stage, spanning approximately from January 2020 to December 2021, is characterized by a large number of mutations in many samples, resulting in a significant number of outliers. This period marks a phase of rapid viral mutation, with these outliers representing samples with an extremely high number of mutations. During the middle stage of the pandemic, from January 2022 to March 2023, the total number of mutations continued to increase, yet the number of outliers above the median decreased, indicating a slowdown in the rate of rapid mutations. However, this stage also recorded a substantial number of mutations below the median, suggesting that cases with fewer mutations from the earlier stage persisted. In the third stage, from April 2023 onward, the number of outliers, both those with fewer and those with a large number of mutations, significantly decreased, signaling a potential winding down of the pandemic (Figure 1). Similar patterns are observed for the other five continents (Figures S1-S5).

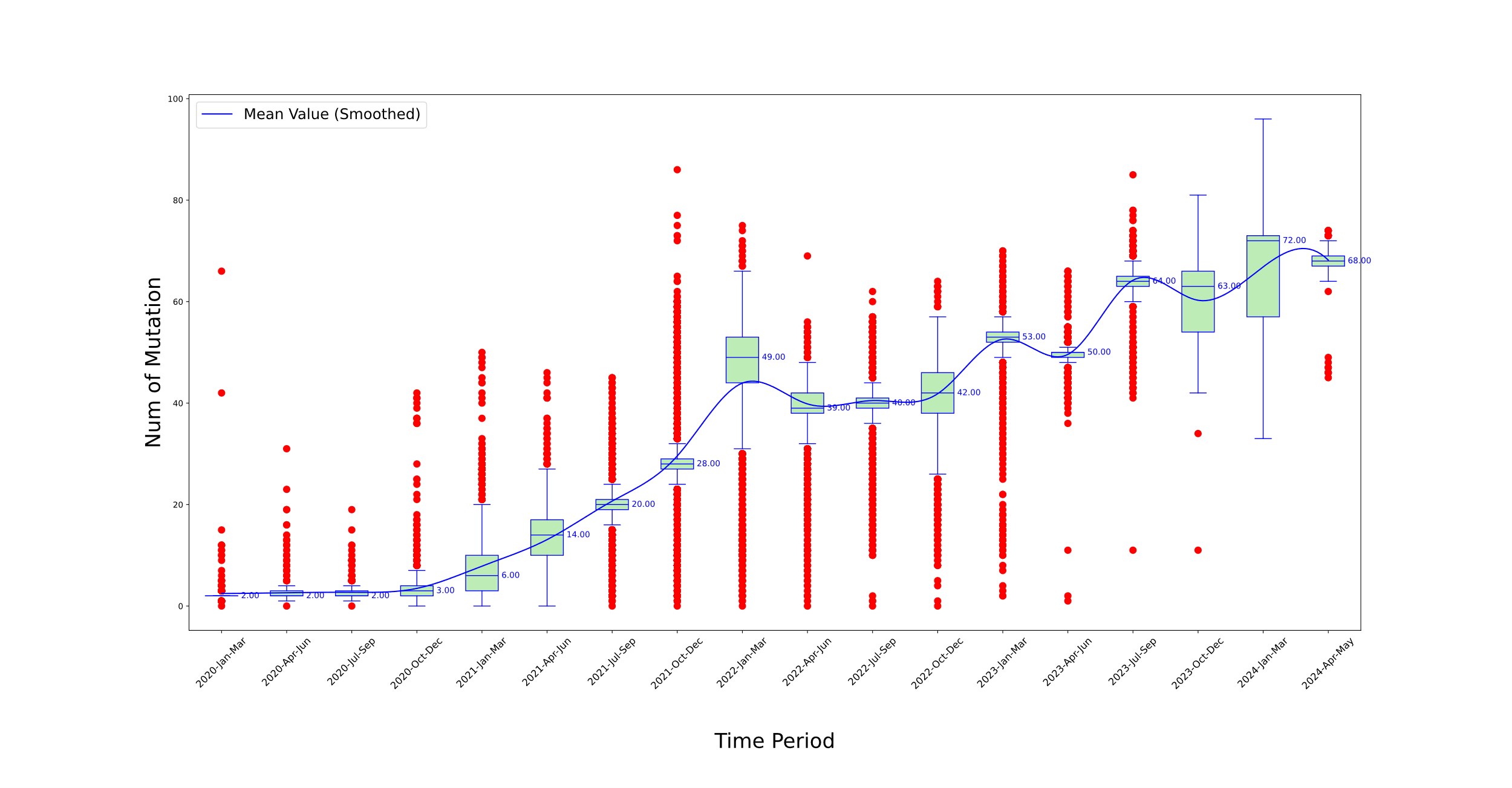


Figure 1 Temporal analysis of mutational frequency per variant sequence in North America from 2020 to 2024

3.2.2 Evolution of global SARS-CoV-2 lineage numbers over various time periods

Among the global dataset, 2,442 lineages were represented. The lineage counts in North America were consistently higher than those in other continents throughout the observed period (Figure 2). This pronounced difference is likely attributable to the larger number of samples sequenced in North America, leading to the detection of more lineages.

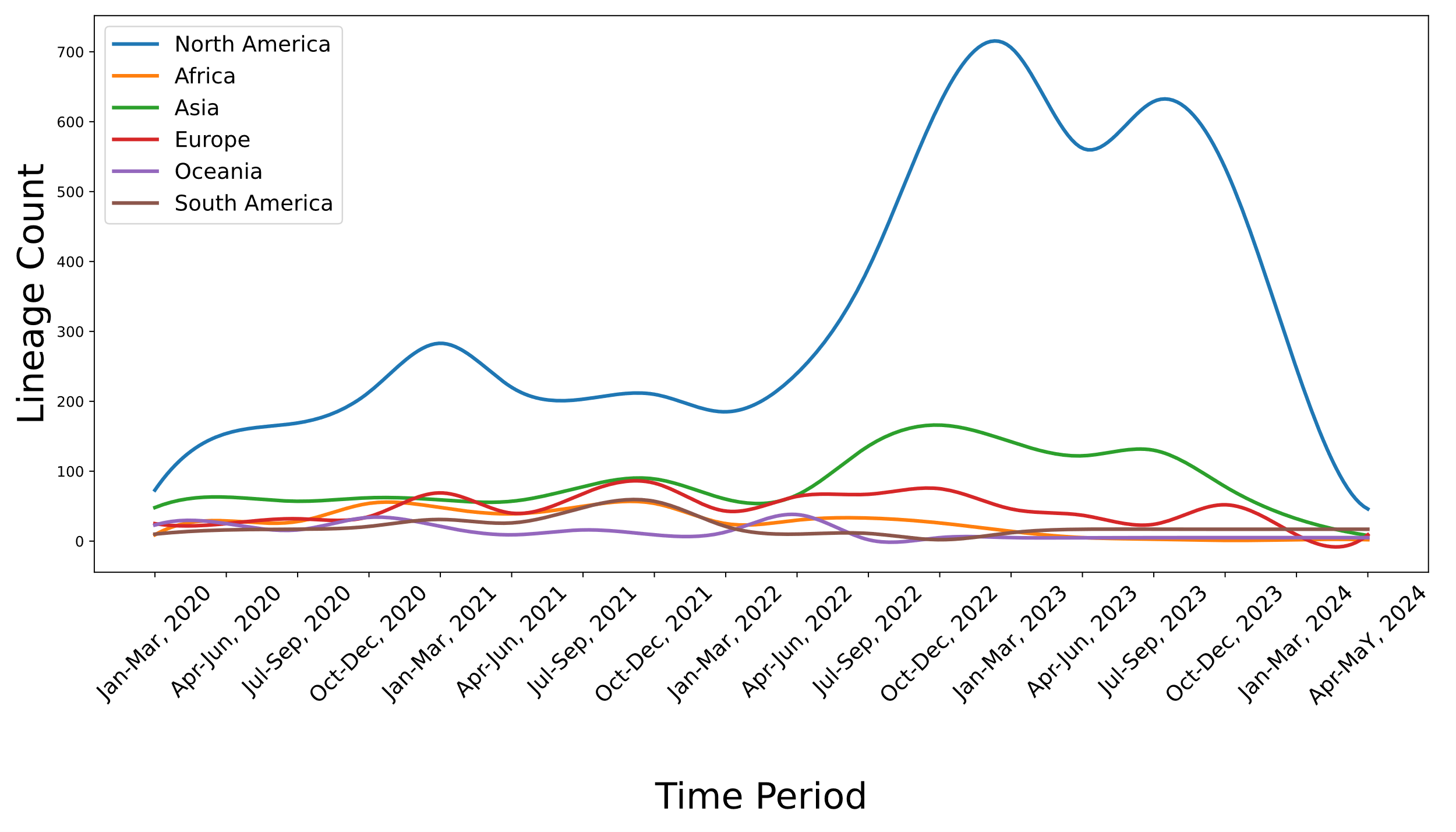


Figure 2 Temporal trends in virus lineage counts across six continents from 2020 to 2024. This chart depicts the trends in the number of virus lineages observed on each of six continents over a five-year period, highlighting considerable fluctuations and a pronounced peak observed in the North America region. The variations and peak may suggest differences in viral evolution or the effectiveness of regional response strategies.

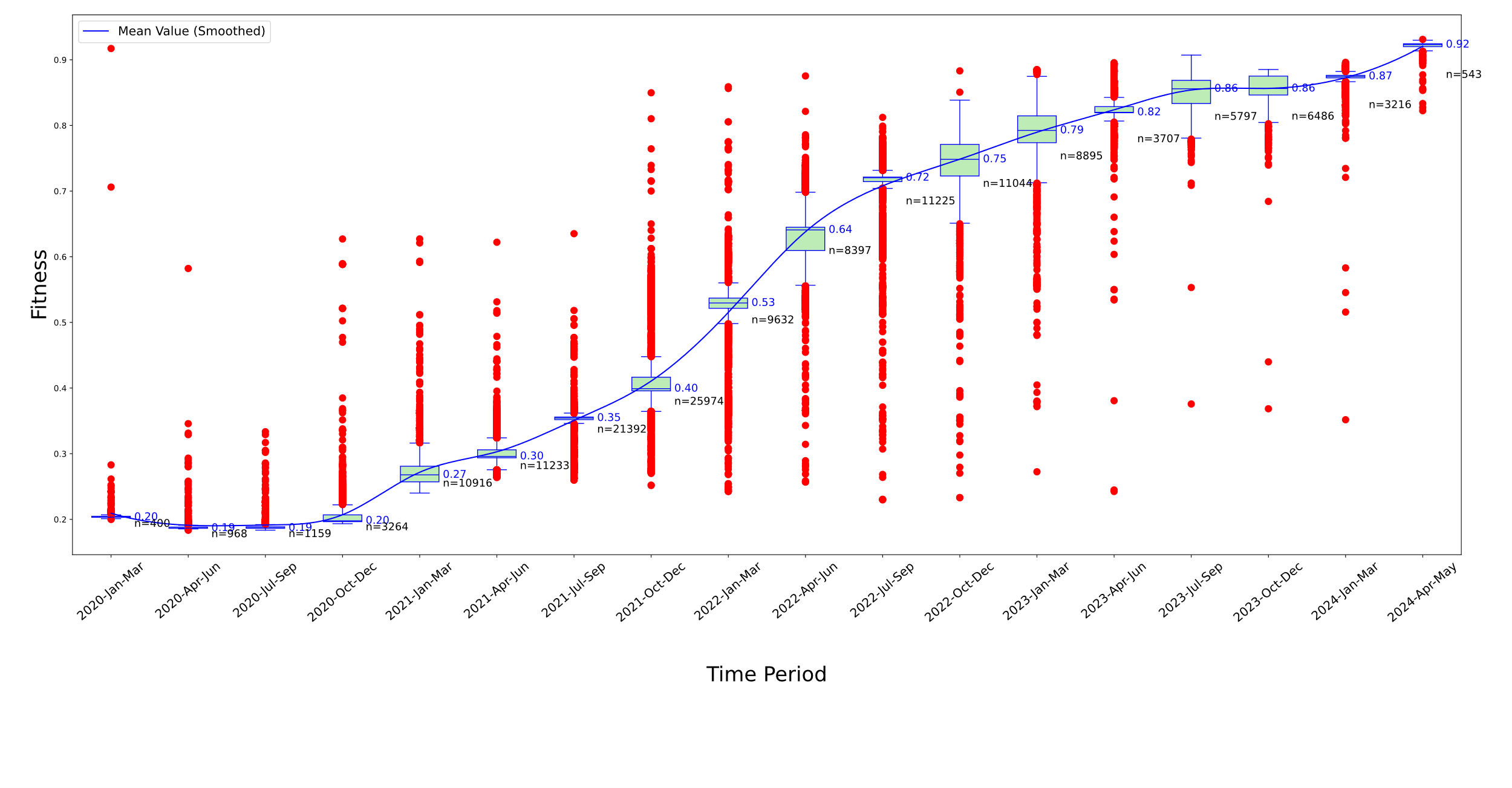
Specifically, the lineage counts in North America exhibit three distinct peaks: The first peak occurs at the end of 2021, reaching a notable high; the second peak is observed at the end of 2022, again showing significant growth; and the third peak emerges around September 2023, indicating a third substantial increase.

In contrast, the lineage counts in other continents remain relatively stable over time. For instance: Europe and Asia showed relatively stable lineage counts throughout the period, with only minor fluctuations and no significant peaks comparable to those in North America; and Africa, Oceania, and South America exhibit relatively low lineage counts with minimal fluctuations, which might be due to fewer sequencing samples from these regions (Figure 2).

Overall, the lineage counts in North America are significantly higher than those in other continents, with three notable peaks, reflecting differences in viral surveillance and sequencing sample sizes across regions.

3.2.3 Temporal fitness evolution of SARS-CoV-2 S protein variants

The four-and-a-half-year period from January 2020 to May 2024 can be roughly divided into three stages (Figure 3): The First stage (Jan 2020 - Mar 2022), the second stage (Mar 2022 - Mar 2023), and the third stage (Mar 2023 - May 2024).

****

**Figure 3** Temporal analysis of Fitness levels for S protein variants in North American over various time periods.

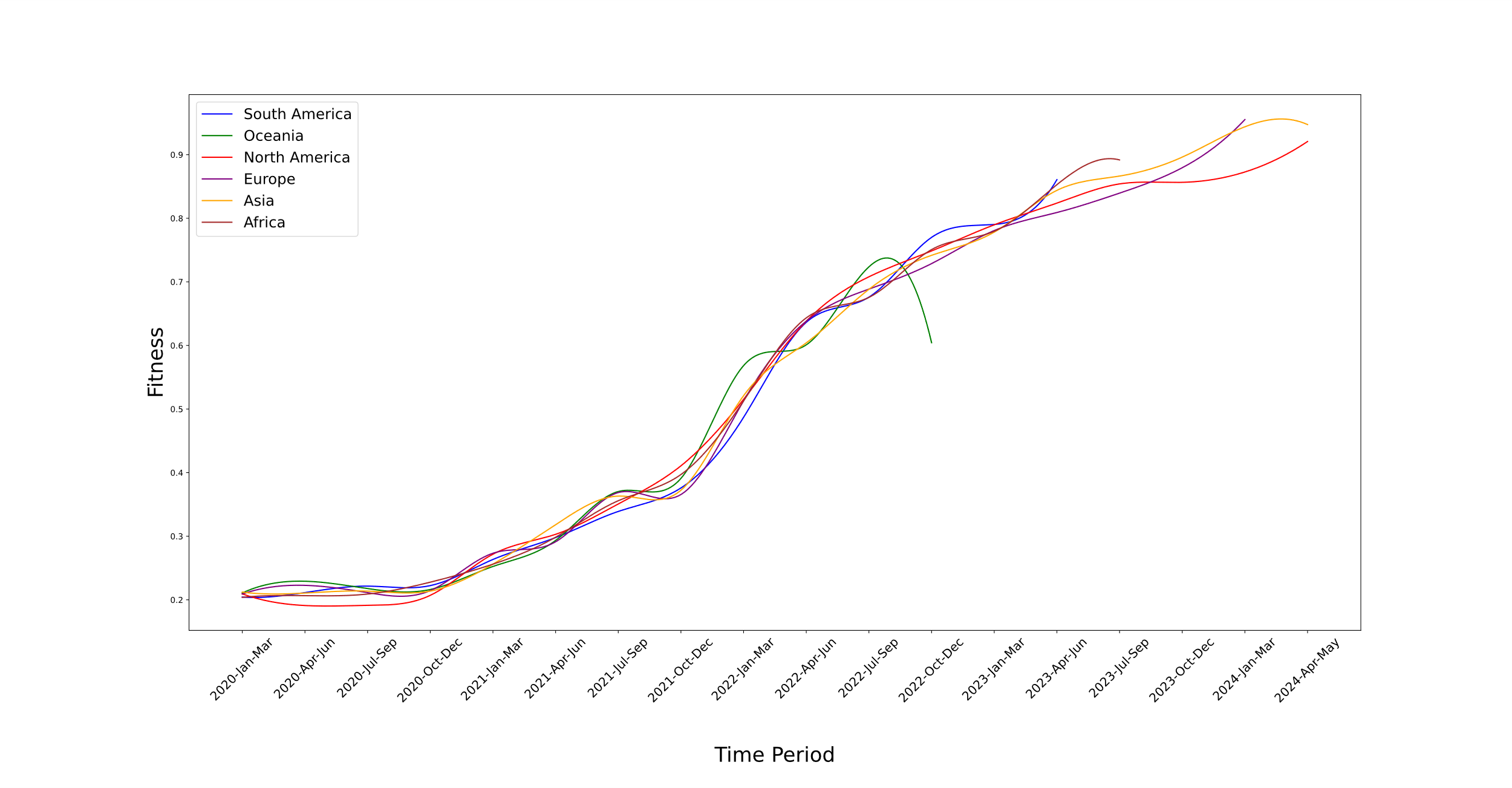
The figure illustrates an increasing trend in the fitness values of S protein variants over time. The lowest fitness of the virus is from China (ancestral type, QZA85478.1, collect date 2020-02-23), with a fitness value of 0.234 (the lowest globally). Therefore, the fitness values of all samples are compared to the wild type from China. In the early stage of the outbreak, the rate change in fitness (the slope of the curve) was not high, and the fitness values were not high. In the middle stage, the rate of change in fitness sharply increased, enhancing the virus's immune escape capability. In the late stage, the growth rates of fitness values and immune escape levels both slowed down, but their values had reached very high levels.

In the first stage, there were many outliers with fitness values higher than the maximum value of the boxplot, which is a significant characteristic of the early outbreak of SARS-CoV-2. The numerous outliers indicated that the virus evolved rapidly during the early stage of the outbreak, with a large accumulation of mutations. The more mutations there were, the higher the fitness of the S protein, resulting in many variants with fitness values significantly exceeding the mean fitness during this phase (Figure 3).

In the second stage, there was a notable decrease in the number of outliers exceeding the boxplot's maximum value, while a substantial proportion of outliers fell below the minimum value. Throughout this stage, not only did these low-fitness variants continue to circulate, but high-fitness variants also emerged more frequently. The fitness levels increased at the fastest rate during this phase, as evidenced by the steep slope of the smoothed mean fitness curve (Figure 3).

In the third stage, the prevalence of low-fitness variants was very rare, and there were also fewer variants with fitness values above the average, suggesting that this is a major characteristic of the end of the pandemic. During this stage, although the fitness values were large, the slope of the smoothed fitness curve became very low. At the same time, there were very few variants with fitness values above the mean, indicating that the rate of virus evolution had slowed down.

Similar characteristics were also observed in the results of the other five continents (Figure 4, Figures S6-S10).



**Figure 4** Longitudinal comparison of Fitness trends for S protein variants across six continents from 2020 to 2024. This graph displays the smoothed mean fitness values over time, highlighting significant regional variations and trends in the evolutionary adaptation of S protein variants. Each line represents a different continent, illustrating comparative rises in fitness levels, which may suggest differences in variant adaptability and potential immune escape efficiency.

Overall, the above results highlight the evolutionary trends in the fitness of S protein variants in all the six continents, demonstrating an increase in fitness over time with reduced variability among newer variants.

3.2.4 Temporal evolution of immune escape capacity in SARS-CoV-2 S protein variants

The smoothed mean immune escape index showed a general increasing trend over time, indicating that the immune escape capability of S protein variants has progressively improved (Figure 5). Early time periods exhibited lower immune escape index values with significant variability and numerous outliers, suggesting diverse immune escape capabilities among early variants. As time progressed, the immune escape index values increased, with the variability and number of outliers decreasing, reflecting more consistent and higher immune escape capabilities among newer variants.

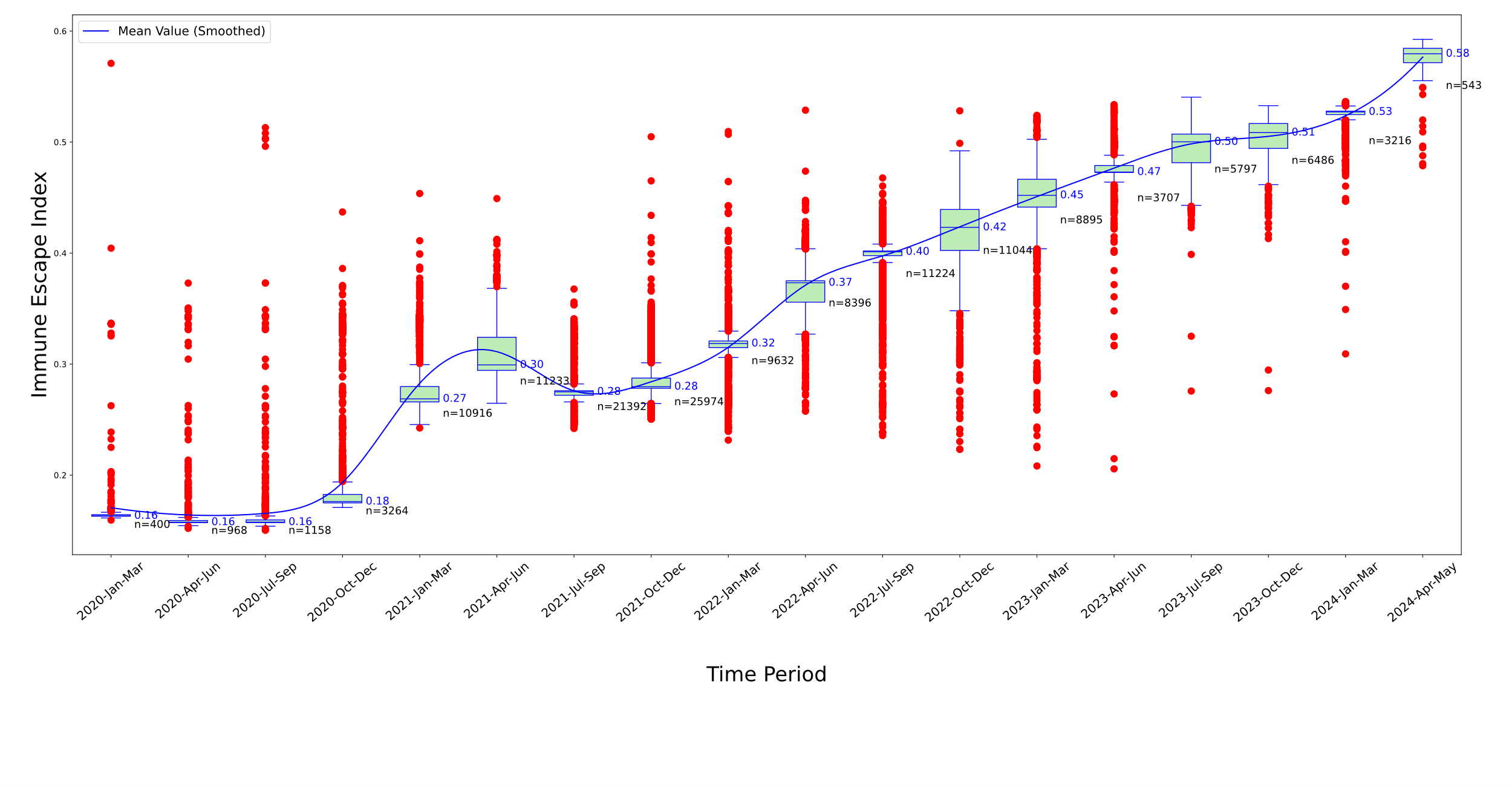


Figure 5 Temporal analysis of Immune Escape Index for S protein variants in North America from 2020 to 2024. This boxplot graphically represents the distribution and temporal progression of the Immune Escape Index for S protein variants in North America over a five-year period. The blue line traces the smoothed mean value of the Immune Escape Index, illustrating an overall trend of increasing immune escape capabilities over time. The graph highlights the variability of the index, with some periods showing a high density of outlier values, suggesting episodes of significant evolutionary changes in the virus's immune escape mechanisms. Each plotted point represents an individual variant's measured Immune Escape Index, providing a comprehensive overview of the changing landscape of viral resistance against immune responses over the specified timeframe.

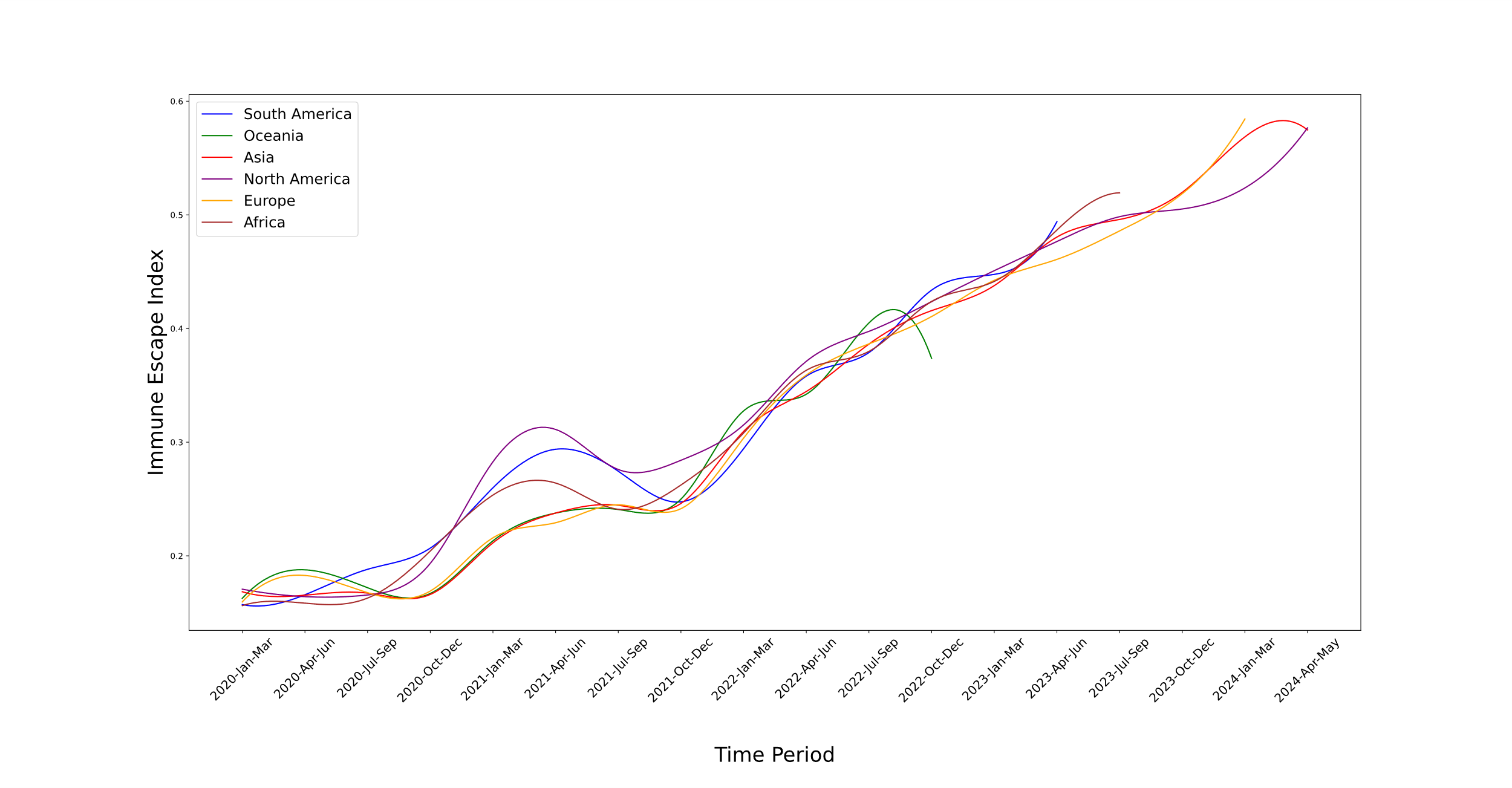
In the immune escape index curves, the trajectories for all six continents generally exhibited an upward trend, yet each showed one or more periods of decline or stagnation over shorter intervals. For instance, the immune escape index for North America reached a local trough in 2021 (Figure 5). This could be attributed to several factors:

Firstly, starting in 2021, various public health interventions were implemented by countries and health organizations, such as lockdowns and travel restrictions. These measures likely curtailed the spread of the virus, particularly variants with high immune escape capabilities.

Secondly, with the widespread rollout of vaccines, variants with higher immune escape capabilities may have been effectively suppressed for a period, resulting in a temporary decrease in the overall immune escape index.

However, despite the significant increase in global vaccination rates over time, the immune escape index of the virus has not shown a sustained downward or stagnant trend. This indicates that the continuous accumulation of mutations in the S protein has enhanced the virus's adaptability, leading to a persistent rise in immune escape capabilities. The rapid evolution of the virus means that newly developed vaccines quickly become less effective, thereby contributing to the ongoing increase in the immune escape index.

Overall, this boxplot highlights the evolutionary trends in the immune escape capabilities of S protein variants across the six continents, showing an increase in immune escape index over time with reduced variability among more recent variants. However, in 2021, the immune escape index for the other five continents also exhibited short-term declines, similar to the results observed in North America (Figure 6, Supplementary Figures 11-15).



**Figure 6** Comparative analysis of the Immune Escape Index for S protein variants across six continents over time. This graph displays the smoothed mean values of the Immune Escape Index for S protein variants across five continents, charting the trends from 2020 to 2024. Each colored line represents the trajectory of immune escape capabilities in South America, Oceania, North America, Europe, Asia, and Africa, indicating how these capabilities have evolved over the years. The chart highlights regional differences in the evolution of the virus's immune escape mechanisms, with some continents showing more pronounced rises in immune escape indices than others. This visualization aids in understanding the geographical variation in viral adaptation and the potential implications for global public health strategies.

4. Discussion

This retrospective study elucidates the dynamic evolution of the SARS-CoV-2 spike protein from January 2020 to May 2024, revealing significant shifts in viral adaptability and immune escape capabilities. Utilizing advanced protein language models, our research emphasizes the role of genetic mutations in shaping the trajectory of the COVID-19 pandemic, revealing intricate mechanisms by which the virus adapts to the human immune system.

Geographic and temporal variations indicate that the virus's adaptability and immune escape indices vary with environmental conditions and the genetic diversity of host populations. These variations might reflect different evolutionary pressures, such as those exerted by high population densities accelerating virus mutation and transmission, while extensive public health interventions could limit the spread of these variants.

Additionally, our findings during the study period (from Apr 1, 2024 to May 15, 2024) indicated a higher adaptability of spike protein variants in North America as of early 2024, suggesting that the virus in this region may be evolving towards a more stable phase of adaptability. This stabilization might signal the virus transitioning towards an endemic phase, potentially manifesting a periodic outbreak pattern similar to seasonal influenza.

Notably, despite increasing global vaccination coverage, the virus's immune escape index continues to rise. This highlights the high adaptability of SARS-CoV-2 under immune pressure and its capacity to accumulate new mutations to evade immune responses. Therefore, ongoing genetic surveillance and timely adjustments in vaccine strategies are crucial to manage potential outbreaks.

This study employed a protein language model to conduct a retrospective analysis of the spike protein of the SARS-CoV-2 virus. Methodologically, CoVFit utilized historical data to develop a deep learning model, on which basis our study predicted the protein fitness and immune evasion capabilities of historical spike protein (S protein) sequences. To address potential inquiries, we clarify that the training data for the CoVFit model comprised 21,751 genotype-fitness data points, covering 12,914 genotypes across 17 countries . Due to the presence of different mutations or variant combinations that can constitute distinct genotypes, many of the genotypes contain repeated mutations. Thus, the number of variants used in training the CoVFit model is estimated to be in the hundreds to thousands, which can be precisely quantified using CoVFit’s original dataset. Additionally, in our study, the total number of global variants analyzed was 160,892, and the variant amino acid sequences used did not include any uncertain 'X' entries. Consequently, in this study's retrospective analysis, only about 2% of the data overlaps with the model training data. To ensure the integrity of the sample, we did not exclude this very small proportion of overlapping data. Therefore, although the retrospective study may include a minimal portion of the data used during model training, this does not affect the primary conclusions drawn from our research.

There are two data points from North America that are exceptionally high in terms of the number of mutations, fitness values, and IEI values. The Fitness/IEI/Mutation values are 0.944/0.571/66 (WZD59850.1, JN.1) and 0.712/0.404/42 (WIJ15993.1, BA.4.6), with collection dates of 2020-01-20 and 2020-02-02 respectively. Firstly, we speculate that the collection dates of these two samples may have been recorded incorrectly. This is because other samples with a fitness value greater than 0.9 occurred after April 2023, and the remaining samples with fitness values above 0.7 appeared after October 2021.

In the case of WZD59850.1, if the recorded collection date is correct, this would imply that the sample underwent an astounding number of 66 mutations in an incredibly short period during the early stages of the virus outbreak. Could it be that the source of this sample was not Wuhan, China? Could they be the result of a longer period of mutations (2-3 years?) accumulated locally in the USA? Is the ancestor of this strain not the same as that of the Wuhan strain? These are puzzling questions that call for more precise phylogenetic studies.

5. Limitations of this study

Despite providing comprehensive analysis, this study has several limitations that need to be considered. First, the accuracy of our predictions largely depends on the quality of the data used. Although our dataset includes a substantial number of variant sequences, there might still be biases due to overrepresentation from certain regions, particularly North America. This could potentially skew the understanding of global viral evolution patterns.

Secondly, although the CoVFit protein language model represents significant progress in predicting adaptability and immune escape capabilities, it is fundamentally limited by its training data. Changes in viral properties not adequately captured in the training sequences might not be accurately reflected in the model's predictions.

Moreover, the retrospective nature of this study means that conclusions are drawn based on past data, which may not necessarily predict future viral evolution trends well, especially as the virus continues to evolve and new variants emerge. This might lead to outdated predictions or reduced accuracy over time.

Lastly, while we strive for precision in our analyses, computational predictions of immune escape capabilities cannot fully substitute for empirical validation in a laboratory setting. Continuous verification of computational results with experimental data is necessary to ensure the accuracy and relevance of the predictions made.

These limitations underline the importance of ongoing research, continuous data collection, and model updates to enhance the predictive accuracy and utility of computational tools in virology.

6. Conclusions

Our thorough investigation into the development of the SARS-CoV-2 spike protein between January 2020 and May 2024 has provided important new information on the virus's adaptability and its consequences for international public health efforts.

Our study highlights the remarkable adaptability of the virus, demonstrated by the continuous increase in fitness and immune escape capabilities. This adaptability underscores the critical need for ongoing genetic surveillance and the development of responsive public health measures to manage emerging and existing variants effectively.

The geographic variability in viral evolution observed in our data necessitates region-specific public health strategies. The distinct evolutionary trajectories identified across different regions emphasize the importance of localized responses that consider regional viral behavior to optimize outbreak control and prevention efforts.

The data suggest a potential stabilization in the adaptability of the virus, particularly in regions like North America, hinting at a possible shift towards endemicity. This transition suggests that SARS-CoV-2 may follow a seasonal outbreak pattern similar to influenza, necessitating adjustments in public health responses, such as periodic vaccination booster schedules and ongoing surveillance.

The application of advanced protein language models, such as CoVFit, has proven invaluable in our understanding of viral evolution. These models offer a powerful tool for early prediction of viral evolutionary trends and potential immune escape mechanisms, enhancing our preparedness and response strategies against current and future viral threats.

The insights provided by this study are instrumental for guiding ongoing research and refining public health strategies and vaccine development. We can improve our ability to forecast viral variations and successfully lessen their influence on global health by utilizing state-of-the-art computational tools and incorporating them into public health infrastructure.

AUTHOR CONTRIBUTIONS

Sihua Peng performed this study and wrote the manuscript. Justin Bahl supervised the study and secured the funding. All authors contributed to the manuscript review and approved the final version of the manuscript.

ACKNOWLEDGMENTS

The study was funded by a grant from the CDC (75D301-21-R-71738) to Justin Bahl.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interests.

DATA AND CODE AVAILABILITY STATEMENT

The data and code are publicly available at GitHub: https://github.com/pengsihua2023/VFIEI-SARS-cov-2.

ORCID

Sihua Peng: https://orcid.org/0009-0007-3189-5408

Justin Bahl: https://orcid.org/0000-0001-7572-4300

REFERENCES

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

2. Yu X, Juraszek J, Rutten L, et al. Convergence of immune escape strategies highlights plasticity of SARS-CoV-2 spike. *PLoS Pathog.* 2023;19(5):e1011308.

3. Willett BJ, Grove J, MacLean OA, et al. SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway. *Nat Microbiol.* 2022;7(8):1161-1179.

4. Lamb KD, Luka MM, Saathoff M, et al. Mutational signature dynamics indicate SARS-CoV-2's evolutionary capacity is driven by host antiviral molecules. *PLoS Comput Biol.* 2024;20(1):e1011795.

5. Nijkamp E, Ruffolo J, Weinstein E, Naik N, Madani A. ProGen2: Exploring the Boundaries of Protein Language Models. *arXiv.* 2022;arXiv:2206.13517

6. Bepler T, Berger B. Learning the protein language: Evolution, structure, and function. *Cell Syst.* 2021;12(6):654-669 e653.

7. Raffel C, Shazeer N, Roberts A, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research.* 2020;21.

8. Devlin J, Chang MW, Lee K, Toutanova K, Assoc Computat L. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Paper presented at: Conference of the North-American-Chapter of the Association-for-Computational-Linguistics - Human Language Technologies (NAACL-HLT); Jun 02-07, 2019; Minneapolis, MN.

9. Brandes N, Goldman G, Wang CH, Ye CJ, Ntranos V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics.* 2023;55(9):1512-+.

10. Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. *Nature Methods.* 2019;16(8):687-694.

11. Lin K, Simossis VA, Taylor WR, Heringa J. A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics.* 2005;21(2):152-159.

12. Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017; Long Beach, CA, USA.

13. Rao R, Bhattacharya N, Thomas N, et al. Evaluating Protein Transfer Learning with TAPE. *Adv Neural Inf Process Syst.* 2019;32:9689-9701.

14. Elnaggar A. ProtTrans: Towards Cracking the Language of Life’s Code Through Self-Supervised Deep Learning and High Performance Computing. *bioRxiv.* 2020.

15. Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A.* 2021;118(15).

16. Obermeyer F, Jankowiak M, Barkas N, et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science.* 2022;376(6599):1327-1332.

17. Ito J, Suzuki R, Uriu K, et al. Convergent evolution of SARS-CoV-2 Omicron subvariants leading to the emergence of BQ.1.1 variant. *Nat Commun.* 2023;14(1):2671.

18. Masuda Y, Nasser H, Zahradnik J, et al. Characterization of the evolutionary and virological aspects of mutations in the receptor binding motif of the SARS-CoV-2 spike protein. *Frontiers in Virology.* 2023;3.

19. Ito J, Strange A, Liu W, et al. A Protein Language Model for Exploring Viral 1 Fitness Landscapes. *bioRxiv.*<https://doi.org/10.1101/2024.03.15.584819>.

20. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science.* 2023;379(6637):1123-1130.

21. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. *Nature Methods.* 2014;11(8):801-807.

22. Wei HJ, Li XH. Deep mutational scanning: A versatile tool in systematically mapping genotypes to phenotypes. *Frontiers in Genetics.* 2023;14.

23. Starr TN, Greaney AJ, Hilton SK, et al. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell.* 2020;182(5):1295-1310 e1220.

24. Cao Y, Jian F, Wang J, et al. Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. *Nature.* 2023;614(7948):521-529.

25. Hayes T, Rao R, Akin H, et al. Simulating 500 million years of evolution with a language model. *bioRxiv.* 2024.

26. Korber B, Fischer WM, Gnanakaran S, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell.* 2020;182(4):812-+.

27. Young M, Crook H, Scott J, Edison P. Covid-19: virology, variants, and vaccines. *Bmj Medicine.* 2022;1(1).

28. Lauring AS, Hodcroft EB. Genetic Variants of SARS-CoV-2-What Do They Mean? *Jama-Journal of the American Medical Association.* 2021;325(6):529-531.

29. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B-Biological Sciences.* 2007;274(1609):599-604.

30. Cori A, Ferguson NM, Fraser C, Cauchemez S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *American Journal of Epidemiology.* 2013;178(9):1505-1512.

31. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol.* 2011;7:539.

SUPPORTING INFORMATION

Supplementary Table 1 Variant, Lineage, Mutation, Fitness, and Immune Escape Profiles of COVID-19 in Africa

Supplementary Table 2 Variant, Lineage, Mutation, Fitness, and Immune Escape Profiles of COVID-19 in Asia

Supplementary Table 3 Variant, Lineage, Mutation, Fitness, and Immune Escape Profiles of COVID-19 in Europe

Supplementary Table 4 Variant, Lineage, Mutation, Fitness, and Immune Escape Profiles of COVID-19 in Oceania

Supplementary Table 5 Variant, Lineage, Mutation, Fitness, and Immune Escape Profiles of COVID-19 in South America

Supplementary Figure 1 Temporal analysis of mutational frequency per variant sequence in Africa from 2020 to 2023

Supplementary Figure 2 Temporal analysis of mutational frequency per variant sequence in Asia from 2020 to 2024

Supplementary Figure 3 Temporal analysis of mutational frequency per variant sequence in Europe from 2020 to 2024

Supplementary Figure 4 Temporal analysis of mutational frequency per variant sequence in Oceania from 2020 to 2022

Supplementary Figure 5 Temporal analysis of mutational frequency per variant sequence in South America from 2020 to 2022

Supplementary Figure 6 Temporal analysis of Fitness levels for S protein variants in Africa over various time periods

Supplementary Figure 7 Temporal analysis of Fitness levels for S protein variants in Asia over various time periods

Supplementary Figure 8 Temporal analysis of Fitness levels for S protein variants in Europe over various time periods

Supplementary Figure 9 Temporal analysis of Fitness levels for S protein variants in Oceania over various time periods

Supplementary Figure 10 Temporal analysis of Fitness levels for S protein variants in South America over various time periods

Supplementary Figure 11 Temporal analysis of Immune Escape Index for S protein variants in Africa

Supplementary Figure 12 Temporal analysis of Immune Escape Index for S protein variants in Asia

Supplementary Figure 13 Temporal analysis of Immune Escape Index for S protein variants in Europe

Supplementary Figure 14 Temporal analysis of Immune Escape Index for S protein variants in Oceania

Supplementary Figure 15 Temporal analysis of Immune Escape Index for S protein variants in South America