Longitudinal Analysis of SARS-CoV-2 Genome Reveals Changing Mutation Patterns and T-cell epitopes

Dongsun Kim

The Graduate School
Yonsei University
Department of Medical Science

Longitudinal Analysis of SARS-CoV-2 Genome Reveals Changing Mutation Patterns and T-cell epitopes

A Master's Thesis Submitted
to the Department of Medical Science
and the Graduate School of Yonsei University
in partial fulfillment of the
requirements for the degree of
Master of Medical Science

Dongsun Kim

June 2024

This certifies that the Master's Thesis
of Dongsun Kim is approved.

Thesis Supervisor Sangwoo Kim

Thesis Committee Member Jun-Young Seo

Thesis Committee Member Eui-Cheol Shin

The Graduate School
Yonsei University
June 2024

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to Professor Sangwoo Kim for guiding and helping me to successfully complete my master's degree. I am truly thankful for the opportunity to participate in research since my undergraduate days, for the encouragement when I transitioned to the master's program from the integrated course, and for the trust and valuable advice you provided during my time in the lab. I deeply regret not being able to fully repay the trust you placed in me. It was a happy time to receive your guidance during my graduate studies.

I am also grateful to Professor Euicheol Shin for being a chair of my thesis committee, helping me significantly with my COVID research, and enabling me to gain more insights into the study. Additionally, I would like to express my gratitude to Professor Junyoung Seo for serving as the chair of my thesis defense and providing me with excellent advice.

I am thankful to my parents, who have silently supported me throughout my graduate studies. Because you always believed in and supported whatever decisions I made, I was able to pursue what I wanted freely. I am always grateful and love you very much!

My four years of laboratory periods have been a precious experience in my life. I was happy to make good connections at TGIL and to gain a lot of experience in research, which made my time there very enjoyable. Watching the TGIL researchers' enthusiasm for research was truly inspiring. Although my journey ends here, I will always be cheering for TGIL wherever I am. I also want to thank the immunology team, who endured with me for a long time, and I hope only good things come your way. I will strive forward with the memories and experiences from TGIL as my foundation. Once again, thank you to everyone at TGIL!

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Longitudinal Analysis of SARS-CoV-2 Genome Reveals Changing Mutation Patterns and T-cell epitopes

During the pandemic, people worldwide suffered greatly from SARS-CoV-2, and even after vaccination, infections continued. This virus is known to accumulate mutations, enabling it to evade immune responses and alter interactions with human cells. To enhance understanding of the mutations and evolution of the coronavirus, we analyzed 6,392,101 sequences collected from 2019 to 2022, identifying 4,590,994 mutations. By examining the occurrence and accumulation of these mutations, we identified an increasing number of mutations that rose in proportion over time, predominantly clustered in the RBD and NTD regions of the spike protein. These mutations were mostly located at antibody-binding sites, suggesting that they are an attempt to evade B-cell mediated immune pressure through mutations. Consequently, these mutations likely enhance viral fitness. T-cell-mediated immune pressure was found to have a minimal impact on mutations, as shown through affinity and strong binder analyses across different variants and HLA types. This study confirms that antibody-mediated immune pressure significantly influences SARS-CoV-2 mutation formation, providing insights valuable for addressing future viruses and variants.

Key words : SARS-CoV-2, Mutation, B-cell, T-cell, Accumulation of mutations, Immune pressure

1. Introduction

The COVID-19 pandemic, caused by the SARS-CoV-2, persisted for over three years from December 2019 to March 2023, profoundly impacting global health, the economy, and society. Initially discovered in 2019, SARS-CoV-2 continuously infected humans, accumulating mutations in its genome over these years. The nucleotide mutation rate of SARS-CoV-2 was estimated at 1.7926×10^{-3} to 1.8266×10^{-3} substitutions per site per year in the first month of the pandemic and 6.677×10^{-4} substitutions per site per year from February 2020 to April 2021(1,2). Although the mutation rate of SARS-CoV-2 is lower than that of most other RNA viruses (3,4), likely due to the exonuclease proofreading function encoded by nsp14 (5), several variants have emerged due to the extensive spread of the pandemic. Following the global replacement with SARS-CoV-2 harboring the D614G substitution in the spike protein, the Delta (B.1.617.2) variant became the overwhelmingly predominant strain (6). Subsequently, the Omicron variant emerged and became widespread(7).

Most random mutations of SARS-CoV-2 are expected to be either deleterious and quickly eliminated or relatively neutral, with only a minor fraction impacting the functional properties of the virus (8). Over three years, numerous mutations accumulated in the SARS-CoV-2 spike protein, resulting in phenotypic changes. These phenotypic changes led to the emergence of various variants, each utilizing its mutations to enhance human infection (9). Mutations that enhance viral fitness and those that enable immune escape, particularly from neutralizing antibodies, act as selection pressures driving viral evolution. Andreano et al. showed that escape mutations were induced in the N-terminal domain and the receptor-binding domain (RBD) of SARS-CoV-2 under immune selective pressure by co-incubating of the wild-type virus with neutralizing plasma from a convalescent COVID-19 patient for more than 90 days (10). They observed the E484K substitution in the RBD of SARS-CoV-2 under strong immune pressure, which is found in current variants such as Beta (B.1.351), Gamma (P.1), and Mu (B.1.621).

However, most studies have been limited to specific proteins or particular variants. In contrast, it is much harder for SARS-CoV-2 to escape T-cell responses because multiple T-cell epitopes are scattered across viral proteins, and human leukocyte antigen (HLA) alleles are diverse across the population. Several studies demonstrated that SARS-CoV-2 variants of concern (VOCs), such as Alpha (B.1.1.7), Delta, and Omicron, rarely escape memory T cell responses in COVID-19

convalescent patients or vaccine recipients (11–15).

The relative preservation of T-cell epitopes across major SARS-CoV-2 variants is partly explained by the highly polymorphic nature of HLA allotypes among human populations. Even if mutations in T-cell epitopes abrogate their binding to HLA molecules in an infected individual, those mutations would not confer an advantage in other infected individuals carrying different HLA allotypes. Thus, nullifying the HLA-binding capacity of T-cell epitopes cannot be a unidirectional pressure for the emergence of variants. However, SARS-CoV-2 may evolve to have mutations that abrogate HLA-binding of T-cell epitopes in an infected individual, although these mutations may revert or further mutate when the virus is transmitted to other individuals. Under this speculation, we hypothesize that T-cell immune pressure results in non-synonymous mutations in HLA-binding epitopes in each infected individual. Accordingly, we also hypothesize that non-synonymous mutations in epitope peptides binding to diverse HLA allotypes can be found across reported SARS-CoV-2 genomes, even if these mutations do not contribute to the emergence of variant strains.

In the present study, we aimed to achieve a comprehensive understanding of viral evolution by identifying new mutations affecting viral fitness, elucidating the mechanisms by which these mutations enhance viral fitness, and exploring the interactions between these mutations and T-cell immune pressure. As the COVID-19 pandemic subsides, this study offers a valuable opportunity to observe the evolutionary dynamics of the entire virus population, contributing to a broader understanding of viral adaptation and immune evasion strategies.

2. Materials and Methods

2.1. Sample collection, sequence alignment and mutation detection

Full genomes of SARS-CoV-2 were obtained from GISAID(16) and NCBI(17). From GISAID, we downloaded samples collected between December 2019 and December 2022 that had complete sequences, high coverage, and complete collection dates. From NCBI, we downloaded samples with complete nucleotide sequences of SARS-CoV-2. These downloaded sequences were aligned to the reference genome (NC_045512) using the Nexclade (18), which also identified mutations based on the alignment results. Samples that did not have a "good" overall QC status as provided by NextClade were excluded from the analysis. Out of a total of 7,380,361 samples downloaded, 6,392,010 samples were ultimately used for the analysis.

2.2. Determination of Mutations order by phylogenetic tree and ancestral sequence prediction

To determine the occurrence of mutations rather than their accumulation, we constructed a phylogenetic tree to understand how sequences evolved. Samples aligned to the reference genome (NC_045512) using NextClade were used to generate a phylogenetic tree with FastTree (19). The phylogenetic tree was then analyzed with Augur Ancestral to predict the expected sequences at each node (20). By comparing the differences between successive sequences, we were able to identify the specific mutations that occurred at each node by custom scripts.

2.3. Selecting increasing mutations for viral fitness And IM(impact of mutation) score

After classifying the virus into strains, we determined the proportion of samples with specific mutations within each group on a monthly basis. Using these proportions and the corresponding months, we performed linear regression. Mutations present in more than 50% of the samples were defined as preserved mutations. Mutations with a positive slope and a p-value less than 0.05 were classified as increasing mutations, as their proportion increased over time, suggesting a positive impact on viral fitness.

To normalize the impact of mutations emerging simultaneously, we considered the time-dependent portion of each mutation as the variant allele frequency (VAF) in multi-sample analyses and used PyClone-vi for multi-sample analysis(21). Clonal analysis allowed us to identify the co-

movement of mutations. To quantify the positive impact of mutations, we developed the IM score, which considers the maximum value of portion change, the frequency of the same domain mutations in different variants, and the number of mutations within a clone.

$$\text{IM score} = \frac{Max\ portion\ change\ x\ detection\ frequency}{clone\ size}$$

2.4. Identify the positions of mutations within the structure

To identify the positions of mutations in the RBD and NTD regions, we used PDB IDs 7T9K and 7W94 obtained from RCSB PDB. The protein structures were analyzed and visualized using PyMOL.

2.5. Prediction of immunogenicity

To calculate the immunogenicity of SARS-CoV-2, we acquired HLA alleles with global frequencies from the CIWD database version 3.0.0(22). From a total of 691 HLA-A alleles and 1,027 HLA-B alleles, we selected those present in netMHCpan version 4.1, excluding non-supertype alleles, and only included HLA alleles with a global frequency of 0.01% or more (23). This resulted in a final selection of 101 MHC-I alleles, consisting of 34 HLA-A alleles and 67 HLA-B alleles, for our study.

The prediction of CD8+ T cell epitopes was carried out using netMHCpan. For each unique mutation, we generated short sequence windows consisting of a 9-mer epitope on either side of the mutation site, containing either the reference or mutated amino acid. Epitope immunogenicity was defined as strong binders with a % rank below 0.5. Agretopicity is defined as the ratio of mutant binding affinity to wild-type binding affinity.

The weighted immunogenicity (wIG) or weighted epitope gain (wEG) is calculated as the sum of global HLA frequencies for immunogenic mutations. Conversely, the weighted immune evasiveness (wIE) or weighted epitope loss (wEL) is calculated as the HLA coverage for immune-evasive mutations. The value of wIE - wIG (wEL - wEG) represents the difference between the sum of the global frequencies of HLA alleles that lose immunogenicity due to the mutation and the sum of global frequencies of HLA alleles that gain immunogenicity.

$$wEG = \sum_{Mi}\sum_{Hj} A(Mi,Hj) \quad , \quad A(Mi,Hj) = \begin{cases} f(Hj), & if\ immunogenic\ mutation \\ 0, & if\ not\ immunogenic\ mutation \end{cases}$$

$$wEL = \sum_{Mi}\sum_{Hj} E(Mi,Hj) \, , \quad E(Mi,Hj) = \begin{cases} f(Hj), & if\ immune\ evasion\ mutation \\ 0, & if\ not\ immune\ evasion\ mutation \end{cases}$$

2.6. Statistics

Statistical analyses of all data were performed with R. Statistical significance between groups was determined using the Mann-Whitney test to compare ranks. Differences in the number of mutations between groups were assessed using the Fisher exact test. Linear regression analysis between quantitative variables was conducted using the 'lm' module. P values < 0.05 were considered significant.

3. Results

3.1. The mutation landscape of SARS-CoV-2 genome

We obtained 7,380,361 sequences collected between December 2019 and December 2022 from NCBI (17) and GISAID (16). These samples were aligned to the reference genome (hCoV-19/Wuhan/WIV04/2019), and mutations were identified using the aligned sequences. After filtering out low-quality samples, we used 6,392,010 sequences for subsequent analyses (Fig. 1A). The distribution of samples highlighted the sequential dominance of the Alpha, Delta, and Omicron variants, with a substantial number of sequences: 1,036,010 for Alpha, 2,468,495 for Delta, and 1,165,450 for Omicron. (Fig. 1B).

The number of mutations per sample consistently increased over time, with an average accumulation of approximately 2.37 mutations per month (Fig. 1C). Notably, the rate of mutation accumulation was steeper for the Omicron variant compared to the Alpha and Delta variants (Fig. 1C). To understand the characteristics of mutation accumulation, we calculated the dN/dS ratio for each sample (15). The Alpha variant had a dN/dS ratio of 0.48, Delta 1.28, Omicron 0.87, and the rest of the SARS-CoV-2 genome sequences exhibited the lowest dN/dS ratio of 0.47 (Fig. 1C). Additionally, the dN/dS ratio decreased over time across all strains, excluding non-variant strains, suggesting that negative pressure is reducing the ratio of nonsynonymous to synonymous mutations (all p-values < 0.05 in linear regression). Despite differences in starting points, the consistent decline in dN/dS ratios indicates that negative pressure is acting on the SARS-CoV-2 genome.

We conducted phylogenetic tree and ancestral sequence prediction analyses by country and strain to identify the occurrence of mutations. This allowed us to observe the emergence of multiple mutations at specific locations (Methods). When normalized by protein length, ORF7a had the highest number of mutations, while ORF1b had the lowest. This was positively correlated with the expression levels of the proteins ($p < 0.04$ in regression). This result is consistent with previous studies showing that mutations are influenced by expression levels, and it was confirmed that, similar to humans, coronaviruses also have more mutations in highly expressed genes (24). The ratio of nonsynonymous to synonymous mutations varied among proteins, with ORF8 showing the highest ratio at approximately 3.32, and the M protein showing the lowest ratio at 0.69 (Fig. 1F). The five-fold difference in these ratios suggests that mutation generation is influenced differently across proteins. We also examined whether mutation generation varied by strain by analyzing the

ratios of G>T/C>A (ROS related), C>T/G>A (APOBEC related), and A>G/T>C (ADAR related) over time. As ADAR is known to act on double-stranded RNA, its ratio remained consistently around 1. In contrast, ROS and APOBEC-related ratios showed fluctuations and a general decline over time (Fig. 1G) (16). During the periods dominated by the Alpha, Delta, and Omicron strains, we observed that the APOBEC-related ratio remained constant across variants, while the ROS-related ratio increased during Alpha, decreased during Delta, and remained low during Omicron (Fig. 1G). The different mutation rates observed in proteins and variants suggest that factors beyond ROS, APOBEC, and ADAR contribute to mutation generation. Therefore, by examining the accumulated mutations, we believed that it would be possible to identify additional factors influencing viral fitness beyond the known mechanisms.

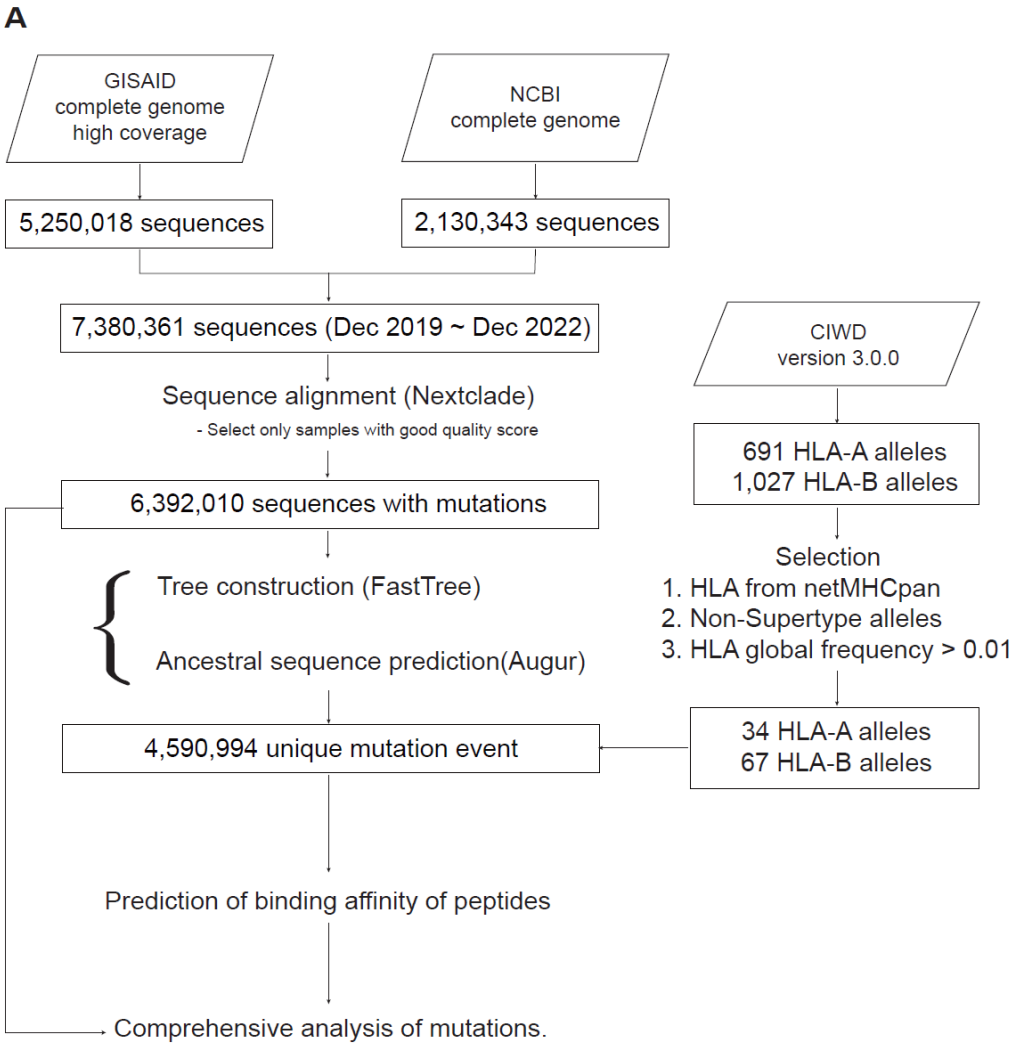

Figure 1. workflow of analysis.

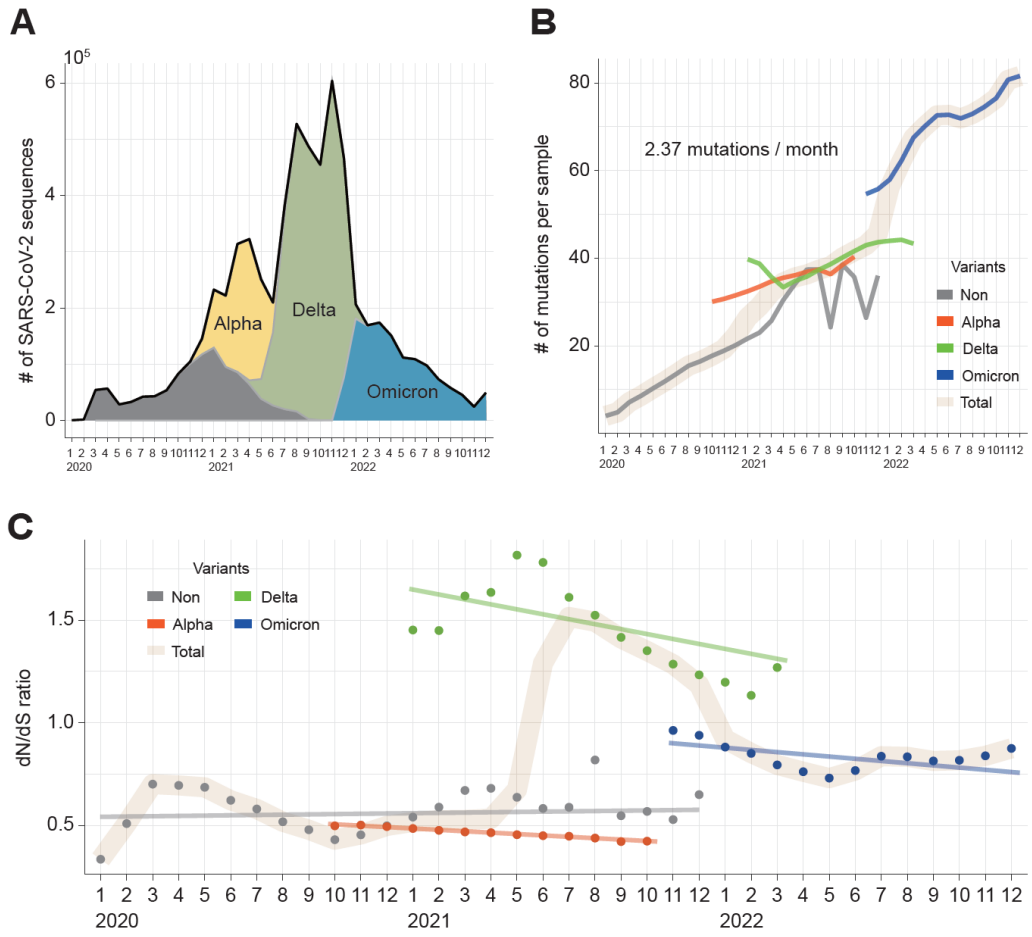

Figure 2. Mutational landscape of SARS-CoV-2. (A) Sample distribution over time. (B) Mutations accumulating in the SARS-CoV-2 genome over time by variant. (C) Changes in dN/dS ratio over time by variant.

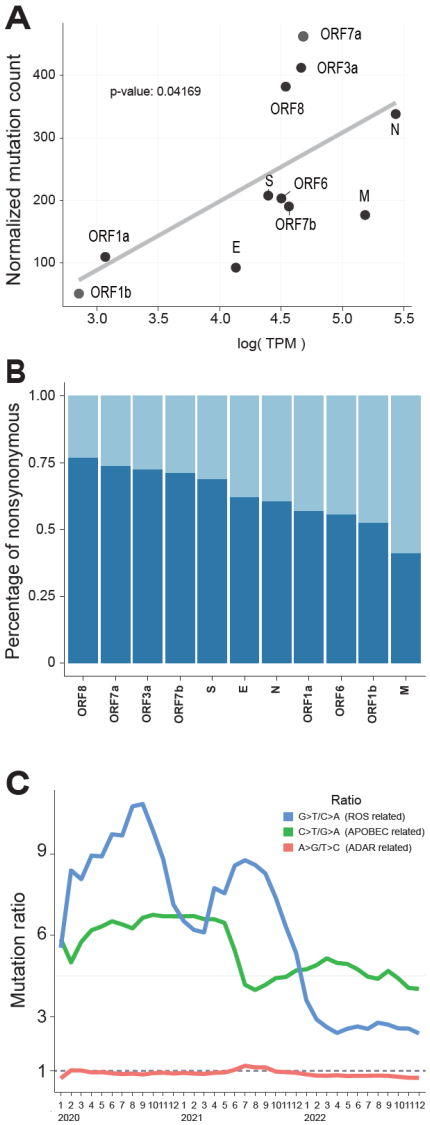

Figure 3. Characteristics of SARS-CoV-2 mutations (A) The correlation between expression levels (TPM) and mutation occurrence rates. (B) The ratio of nonsynonymous mutations varies by protein. (C) Mutation ratios related to etiology over time. Mutation ratios differ over time and by variant.

3.2. Mutation accumulation affects virus fitness

We observed that the occurrence of mutations varies by strain and protein, and we aimed to understand the factors influencing how these mutations accumulate. We believed that understanding the characteristics of accumulated mutations would reveal the factors influencing their development. Therefore, we quantified the impact of accumulated mutations on viral fitness by developing the IM score, which considers the rate of increase, frequency of occurrence in variants, and the size of the mutation cluster (Fig. 2A, Method). We hypothesized that mutations with higher rates of increase, higher frequencies, and smaller clusters would have a greater impact. Additionally, we investigated the potential structural effects of mutations with high IM scores on protein structures.

By examining the mutation rates for each strain, we found that certain mutations consistently increased over time (Fig. 2B, 2C). Using linear regression, we identified mutations that showed significant increases and defined these as "increasing mutations". Based on the movement patterns of mutations similar to increasing mutations, we categorized mutations into four types: increasing, preserved, decreasing, and non-type (Method). Previous studies indicated that preserved mutations enhance viral fitness (17). We considered increasing mutations to be an intermediate stage leading to preserved mutations, and we expected that, like preserved mutations, they would ultimately boost viral fitness.

When analyzing the IM scores of increasing mutations, we found that mutations in the spike protein had an average IM score of 3.22, significantly higher than the average score of 0.52 for mutations in other proteins (Fig. 2D, p-value=0.0002 in Mann-Whitney test). The distribution of increasing mutations revealed a significant concentration in the spike protein compared to non-type or decreasing mutations (p-values of 3.05×10^{-5} and 0.043 in Fisher exact test). There was no significant difference between increasing and preserved mutations (p-value=0.57 in Fisher exact test). Thus, we infer that increasing mutations positively affect viral fitness similarly to preserved mutations, and that the spike protein plays a critical role in enhancing viral fitness.

We examined the distribution of mutations within the spike protein and found that 92% of increasing mutations were located in the S1 region, particularly in the receptor-binding domain (RBD) and N-terminal domain (NTD), which are targets for vaccines or externally protruding regions. In contrast, preserved and decreasing mutations had concentrations of 75.6% and 75%, respectively, while non-type mutations had a concentration of 44.4%. The non-type mutations

showed a significant difference compared to the increasing mutations (Fisher exact test, p-value=0.0005). Notably, 11 of the 13 increasing mutations in the spike protein were found in Omicron sub-strains, with 11 mutations post-vaccination, suggesting they arose due to vaccine-induced immune pressure. To further investigate this hypothesis, we conducted structural analyses.

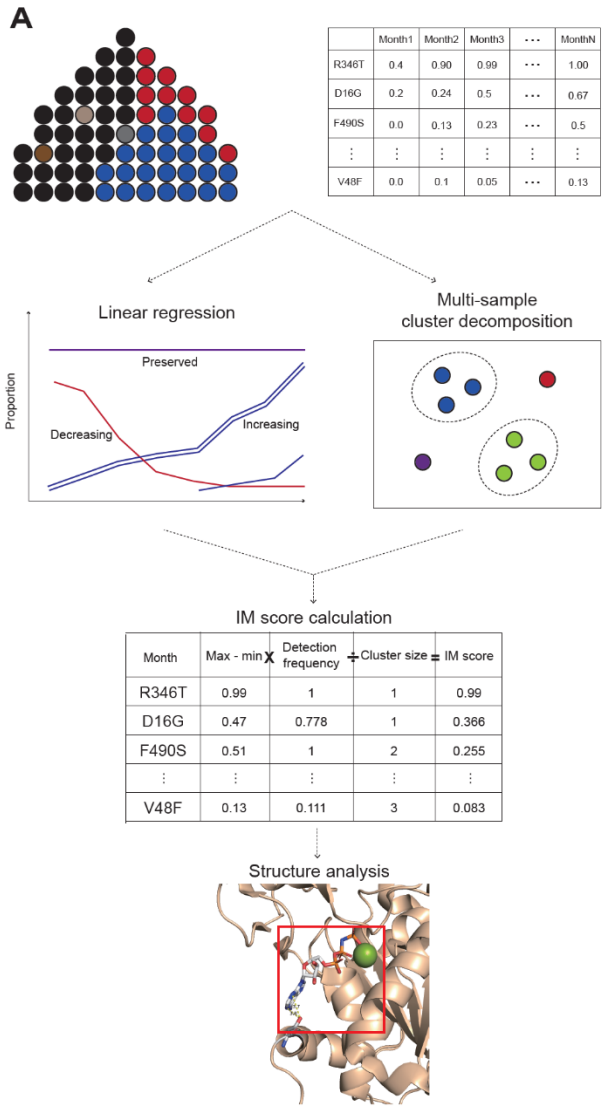

Figure 4. IM score calculation and structural analysis.

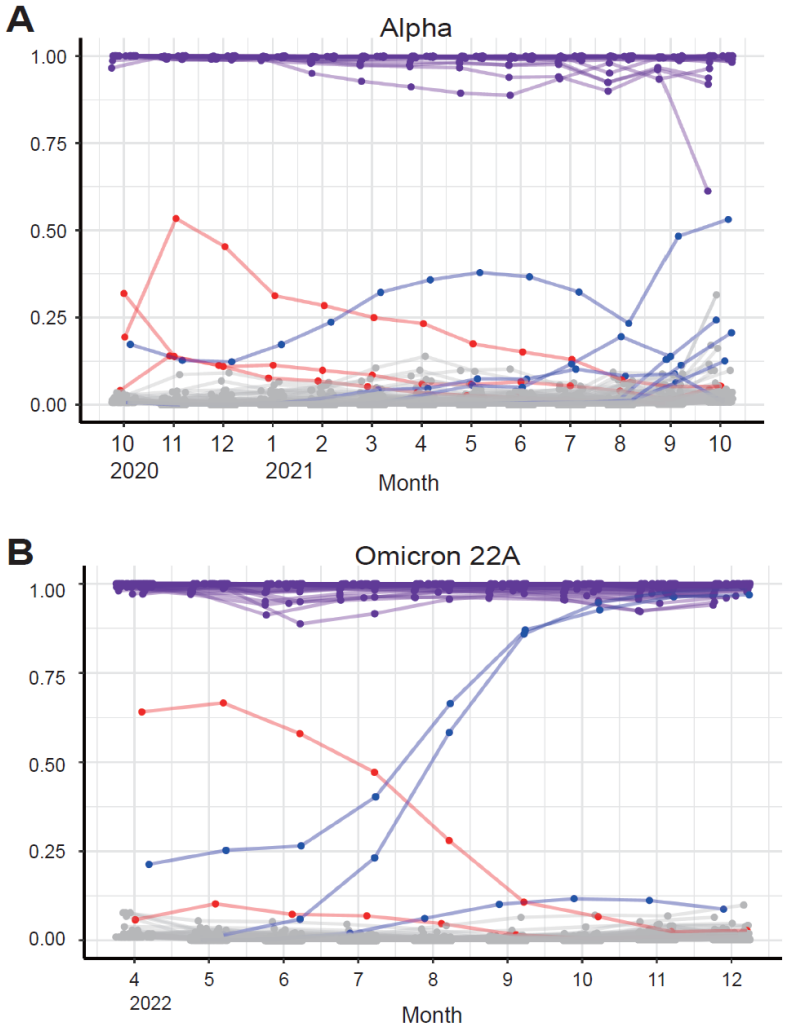

Figure 5. Mutation proportion change over time (A) Changes in mutation ratios in the Alpha variant. There are 4 increasing mutations and 3 decreasing mutations. (B) Changes in mutation ratios in the Omicron 22A variant. There are 3 increasing mutations and 2 decreasing mutations.

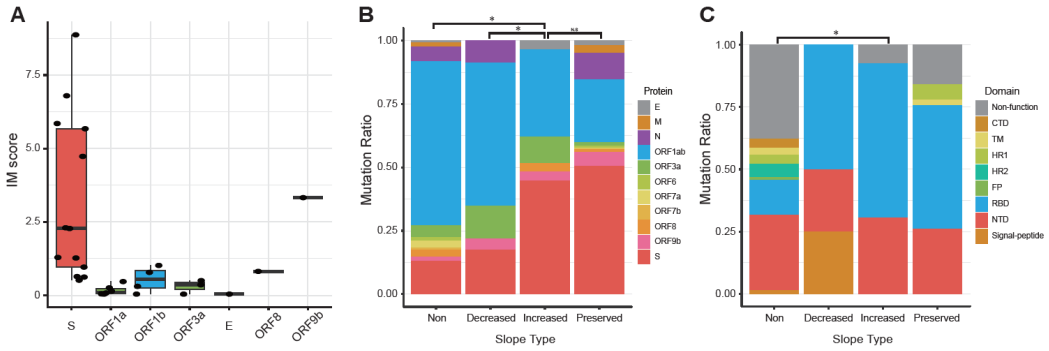

Figure 6. IM scores and distribution of increasing mutations (A) The IM scores of increasing mutations by protein. (B) The distribution of mutations for each mutation type. (C) The distribution of mutations in the spike protein for each mutation type.

3.3. Protein structure of increasing mutations

To investigate how increasing mutations might enhance viral fitness, we conducted a protein structure analysis. Given the significant presence of increasing mutations in the spike protein, we hypothesized that these mutations would exhibit distinct structural patterns. Among the 13 increasing mutations identified in the spike protein, 8 were located in the receptor-binding domain (RBD), and 4 were located in the N-terminal domain (NTD). Statistical tests confirmed the significance of this clustering, prompting us to map these mutations onto the known spike protein structure (Fig.3A). We observed that, with the exception of N460K, the mutations were concentrated on one face of the RBD. According to Kathryn et al., this face corresponds to the outer face of the RBD(25). The six mutations—R346K, R346T, K356T, G446S, E484K, and F490S (with R346K appearing in both Omicron 22A and Omicron 21K)—were identified on this outer face. Except for R356T, these mutations are known to interact with antibodies (Fig. 3B, yellow region) (25). N460K, although not clustered, was located in a region known to interact with antibodies (Fig. 3C, green region) (25). These increasing mutations are likely to interfere with antibody binding, potentially preventing spike protein neutralization or disrupting the spike-ACE2 interaction.

In addition to the RBD, we found four mutations (G142D, N164K, and del144 (with del144 appearing in both Omicron 22E and Omicron 22B)) clustered at the top of the N-terminal domain (NTD) (Fig.3D). Kathryn et al. also identified this region as an antibody-binding site (Fig. 3D, blue region) (25). Notably, except for G142D and E484K, all 11 increasing mutations in the spike protein were found in the Omicron variant. This suggests that the late-emerging Omicron variant, which appeared after widespread vaccine deployment, developed these increasing mutations to evade the immune response. As vaccines enabled many individuals to rapidly produce antibodies, the virus adapted through these mutations to escape immune detection and maintain its fitness.

Beyond the spike protein, the ORF1ab mutation also appears to impact viral fitness. While other mutations did not show a clear functional relevance, E5585D is located in the Rec1A domain of the NSP13 helicase, directly interacting with ATP. This mutation likely alters ATP binding, enhancing viral fitness by increasing the portion of the virus with this mutation(Fig.4E,F). In summary, increasing mutations in SARS-CoV-2 likely enhance viral fitness by either evading antibodies or altering protein functions. These findings underscore the adaptive potential of SARS-CoV-2 in response to immune pressures.

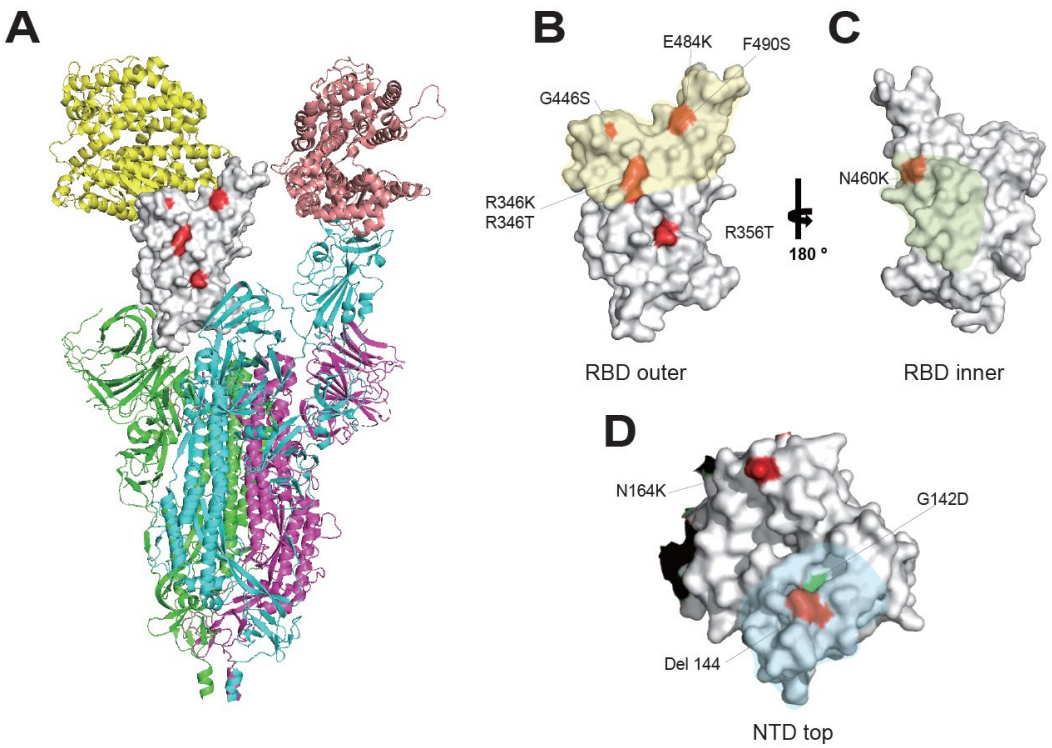

Figure 7. Increasing mutations in Spike protein (A) The positions of increasing mutations in the RBD while binding with the ACE receptor (the protein above). (B) Increasing mutations occurring in the outward-facing surface of the RBD region. (C) Increasing mutations occurring in the inward-facing surface of the RBD region. (D) The top view of the NTD region and the positions of the increasing mutations located in the region.

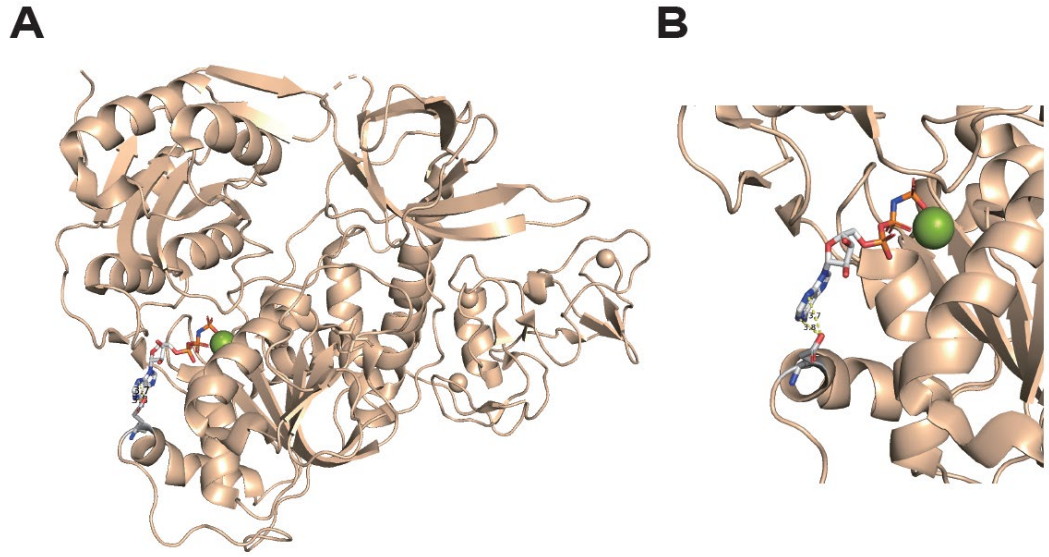

Figure 8. Increasing mutation E5585D in NSP13(helicase) (A) The structure of NSP13 (helicase) within ORF1ab and the position of the E5585D mutation. (B) The E5585D mutation is located at a position that directly interacts with the ATP binding domain.

3.4. T-cell immune pressures on SARS-CoV-2 genome

To confirm the presence of immune pressure from T-cells, known to play a critical role in viral clearance, we investigated whether T-cell-mediated immune pressure, similar to B-cell immune pressure, affects the SARS-CoV-2 genome (18). After determining the sequence of mutations using a phylogenetic tree, we identified how these mutations alter amino acids and calculated strong binder and agretopicity values, weighted by global HLA frequencies (Fig.4A, B, Methods). We also examined changes in strong binders using weighted Epitope Gain (wEG) and weighted Epitope Loss (wEL) to track the impact on strong binders (Method).

Our analysis of all mutations revealed that both HLA-A and HLA-B showed significantly higher wEG values compared to wEL, suggesting that mutations often increase the likelihood of peptide fragments binding to HLA (Fig.4C). Similarly, agretopicity values for HLA-A and HLA-B were less than zero, indicating an overall increase in binding affinity between HLA and peptides due to mutations (Fig.4D). When comparing different strains, we found that all strains exhibited higher wEG than wEL for both HLA-A and HLA-B, indicating that mutations are enhancing immunogenicity across all strains. (Fig. 4E, G). Agretopicity values also remained below zero across all strains, confirming increased HLA binding affinity in every strain examined (Fig. 4F, H). This means that mutations increase the binding between HLA and the peptide, thereby increasing the likelihood that the virus's peptide will be recognized by T-cells.

To determine if the immune pressure was more pronounced in frequent HLA types, we compared wEG and wEL between frequent and infrequent HLA alleles. In HLA-A, frequent types had lower wEG compared to infrequent types, while in HLA-B, there was no significant difference. For wEL, both HLA-A and HLA-B showed higher values in infrequent types. Despite significant differences in wEL and wEG individually, net Gain was greater than 0 and agretopicity was less than 0, and neither differed significantly between frequent and infrequent HLA types, indicating that mutations generally promote epitope gain regardless of strain or HLA frequency.

For the Alpha, Delta, and Omicron strains, we observed an increase in the average number of strong binders due to mutations relative to their references in both HLA-A and HLA-B. This finding aligns with previous studies suggesting that T-cell epitopes are maintained, indicating that T-cell immune pressure is not the dominant force influencing epitope gain or loss (11). Additionally, the cumulative number of strong binders, compared to the reference for each strain, consistently remained above zero, further supporting the notion that SARS-CoV-2 mutations

enhance HLA binding over time.

Overall, these results demonstrate that as mutations accumulate in the SARS-CoV-2 genome, the binding affinity with HLA increases, suggesting a trend towards enhanced immunogenicity through epitope gain.

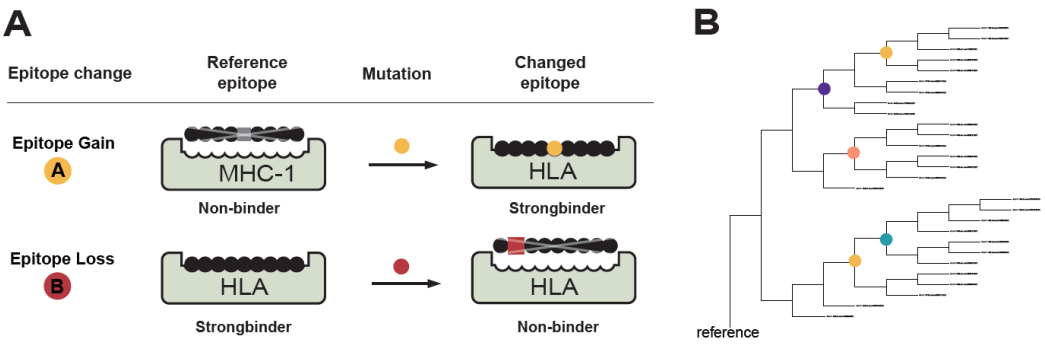

Figure 9.wEG,wIL and phylogenetic tree of SARS-CoV-2 (A) An epitope gain occurs when a mutation enhances the binding affinity of a protein fragment to HLA, turning it into a strong binder. Conversely, an epitope loss occurs when a mutation causes a strong binder to lose its binding affinity. (B) To accurately identify the occurrence of mutations, we divided the samples by country and variant and constructed a phylogeny.

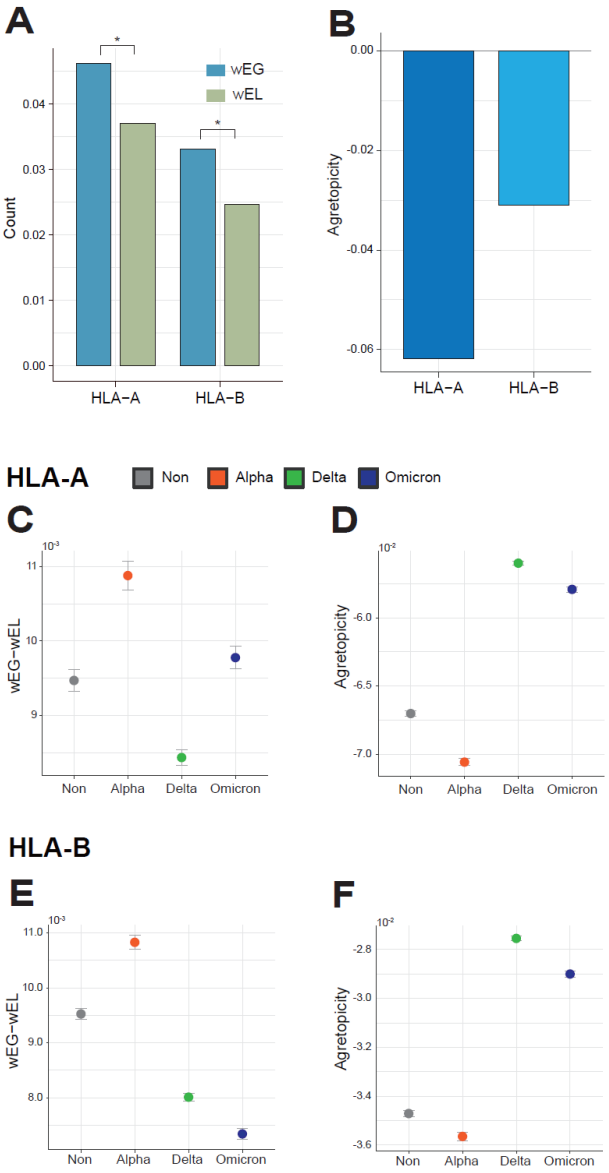

Figure 10. Epitope changes of SARS-CoV-2 genome in each variant. (A) The wEG and wEL by HLA type for all mutations. (B) The changes in binding affinity (agretopicity) due to mutations for all mutations. (C) The net increase in strong binders (wEG - wEL) for HLA-A type by variant. (D) Changes in binding affinity (agretopicity) of peptides due to mutations by variant for HLA-A type. (E) The net increase in strong binders (wEG - wEL) for HLA-B type by variant. (F) Changes in binding affinity (agretopicity) of peptides due to mutations by variant for HLA-B type.

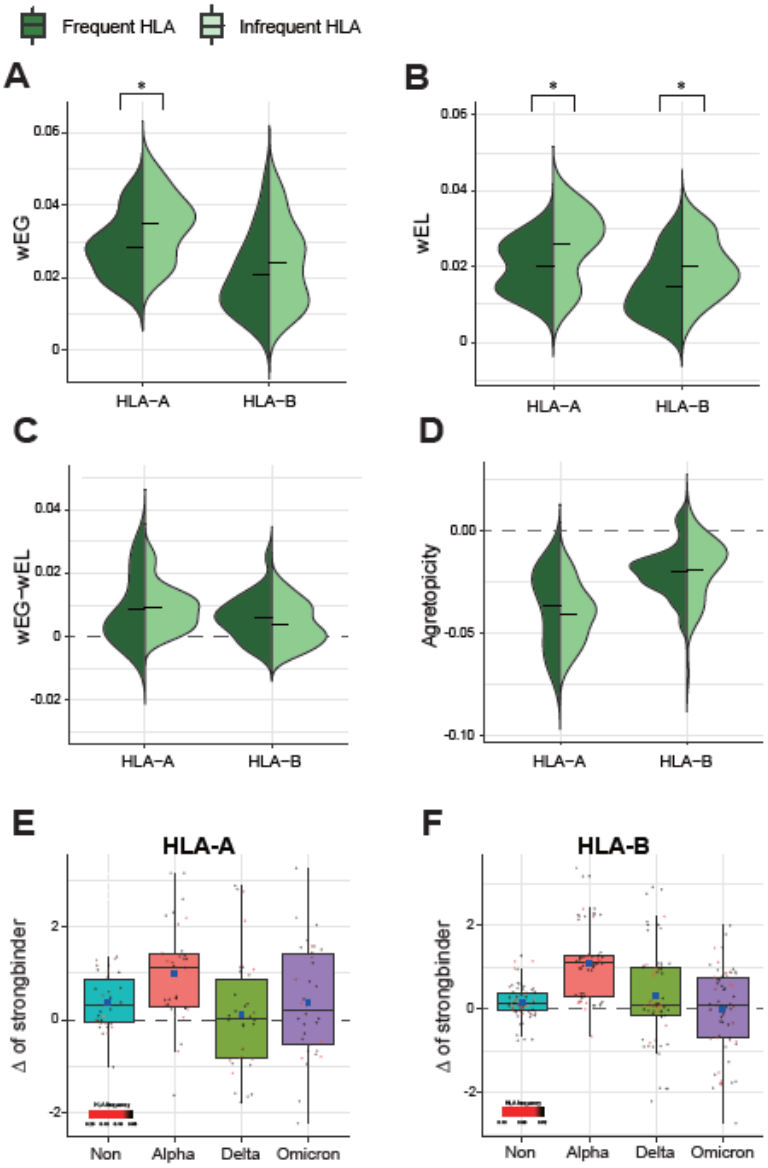

Figure 11. Epitope change in SARS-CoV-2 genome (A) Difference in wEG by HLA frequency type. (B) Difference in wEL by HLA frequency type. (C) Comparison of net strong binder gain (wEG wEL) by HLA frequency type. (D) Difference in binding affinity (agretopicity) of peptides due to mutations by HLA frequency type. (E) Changes in strong binders due to mutations relative to each variant reference for HLA-A. (F) Changes in strong binders due to mutations relative to each variant reference for HLA-B.

4. Discussion

From the initial discovery of SARS-CoV-2 in 2019 to December 2022, we observed the occurrence and accumulation of mutations in the coronavirus genome. This study allowed us to identify factors influencing mutation accumulation, with a particular focus on the impact of immune pressure. By analyzing the dN/dS ratio and mutation ratios, we determined known mutation etiologies are not the sole factors at play; additional elements influence mutation accumulation differently across proteins and strains.

To identify factors affecting mutation accumulation, we examined increasing mutations—mutations that consistently rose in frequency across all strains. We found that these increasing mutations significantly accumulated in the spike protein, especially in the receptor-binding domain (RBD) and N-terminal domain (NTD). Structural analysis of these regions revealed that 11 out of 13 increasing mutations were located at antibody binding sites, with most mutations appearing in the Omicron variant. This suggests that increasing mutations have accumulated to evade immune pressure from antibodies, particularly following widespread vaccination efforts.

After confirming the impact of B-cell-mediated immune pressure, we investigated the influence of T-cell-mediated immune pressure by examining the affinity between HLA molecules and peptides resulting from mutations. Both emergent and accumulated mutations showed an increase in the number of strong binders and enhanced binding affinity between peptides and HLA molecules. Our findings align with Shin et al.'s study, which reported the preservation of T-cell epitopes in the Omicron variant (11). The current findings suggest that B-cell-mediated antibody pressure plays a more significant role in driving mutation accumulation compared to T-cell immune pressure. While the impact of T-cell pressure on mutation formation was not as evident, this may be attributed to the substantial variability in HLA types among individuals. The diversity of HLA molecules, which present viral targets to T cells, complicates the detection of consistent mutation patterns driven by T-cell pressure across different populations. Consequently, the observed increase in mutations in the spike protein, particularly in regions targeted by antibodies, underscores the adaptive evolution of SARS-CoV-2 in response to antibody-mediated immune pressure.

According to studies, T-cell immune pressure can indeed drive the selection of viral mutations. For instance, research on HIV has shown that T-cell responses can lead to adaptations within T-cell epitopes presented by HLA molecules, resulting in immune escape (26). However, due to the

broad heterogeneity in HLA genotypes, these adaptations vary significantly between individuals, making it challenging to observe consistent patterns at the population level (27).

In conclusion, while B-cell-mediated immune pressure appears to play a more prominent role in driving mutation accumulation, the role of T-cell pressure remains significant yet obscured by the variability in HLA types among individuals. This underscores the need for more targeted studies to better understand the nuanced contributions of T-cell immune responses to mutation evolution

5.Conclusion

This study comprehensively examined the occurrence and accumulation of mutations in the SARS-CoV-2 genome from its initial discovery in 2019 through December 2022. Our findings highlight the significant role of immune pressure, particularly from B-cell-mediated antibodies, in shaping the viral mutation landscape. In conclusion, the adaptive evolution of SARS-CoV-2 is primarily driven by antibody-mediated immune pressure, leading to the accumulation of mutations that enhance viral fitness by evading immune detection. Understanding these dynamics is crucial for developing effective strategies to combat the virus and mitigate the impact of emerging variants.

References

1. Li, X., Wang, W., Zhao, X., Zai, J., Zhao, Q., Li, Y. (2020). Transmission dynamics and evolutionary history of 2019-nCoV. Journal of Medical Virology, 92(5), 501-511.

2. Wang, S., Xu, X., Wei, C., Li, S., Zhao, J., Zheng, Y. (2022). Molecular evolutionary characteristics of SARS-CoV-2 emerging in the United States. Journal of Medical Virology, 94(1), 310-317.

3. Worobey, M., Pekar, J., Larsen, B. B., Nelson, M. I., Hill, V., Joy, J. B. (2020). The emergence of SARS-CoV-2 in Europe and North America. Science, 370(6516), 564-570.

4. Peck, K. M., & Lauring, A. S. (2018). Complexities of viral mutation rates. Journal of Virology, 92(14), e01031-17.

5. Robson, F., Khan, K. S., Le, T. K., Paris, C., Demirbag, S., Barfuss, P. (2020). Coronavirus RNA proofreading: Molecular basis and therapeutic targeting. Molecular Cell, 79(5), 710-727.

6. Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W. (2020). Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus. Cell, 182(4), 812-827.e19.

7. Anand, U., Pal, T., Zanoletti, A., Sundaramurthy, S., Varjani, S., Rajapaksha, A. U. (2023). The spread of the omicron variant: Identification of knowledge gaps, virus diffusion modelling, and future research needs. Environmental Research, 225, 115612.

8. Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M. (2021). SARS-CoV-2 variants, spike mutations and immune escape. Nature Reviews Microbiology, 19(7), 409-424.

9. Carabelli, A. M., Peacock, T. P., Thorne, L. G., Harvey, W. T., Hughes, J., de Silva, T. I. (2023). SARS-CoV-2 variant biology: immune escape, transmission and fitness. Nature Reviews Microbiology, 21(3), 162-177.

10. Andreano, E., Piccini, G., Licastro, D., Casalino, L., Johnson, N. V., Paciello, I. (2021). SARS-CoV-2 escape from a highly neutralizing COVID-19 convalescent plasma. Proceedings of the National Academy of Sciences of the United States of America, 118(36), e2103154118.

11. Choi, S. J., Kim, D. U., Noh, J. Y., Kim, S., Park, S. H., Jeong, H. W. (2022). T cell epitopes in SARS-CoV-2 proteins are substantially conserved in the Omicron variant. Cellular & Molecular Immunology, 19(3), 447-448.

12. Woldemeskel, B. A., Garliss, C. C., & Blankson, J. N. (2021). SARS-CoV-2 mRNA vaccines induce broad CD4+ T cell responses that recognize SARS-CoV-2 variants and HCoV-NL63. Journal of Clinical Investigation, 131(10), e149335.

13. Geers, D., Shamier, M. C., Bogers, S., den Hartog, G., Gommers, L., Nieuwkoop, N. N.

(2021). SARS-CoV-2 variants of concern partially escape humoral but not T cell responses in COVID-19 convalescent donors and vaccine recipients. Science Immunology, 6(59), eabj1750.

14. Tarke, A., Sidney, J., Methot, N., Yu, E. D., Zhang, Y., Dan, J. M. (2021). Impact of SARS-CoV-2 variants on the total CD4+ and CD8+ T cell reactivity in infected or vaccinated individuals. Cell Reports Medicine, 2(7), 100355.

15. Jordan, S. C., Shin, B. H., Gadsden, T. A. M., Chu, M., Petrosyan, A., Le, C. N. (2021). T cell immune responses to SARS-CoV-2 and variants of concern (Alpha and Delta) in infected and vaccinated individuals. Cellular & Molecular Immunology, 18(11), 2554-2556.

16. Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. EuroSurveillance, 22(13), 30494.

17. Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. Nucleic Acids Research, 41(D1), D36-D42.

18. Aksamentov, I., Roemer, C., Hodcroft, E. B., & Neher, R. A. (2021). Nextclade: clade assignment, mutation calling and quality control for viral genomes. Journal of Open Source Software, 6(67), 3773.

19. Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately maximum-likelihood trees for large alignments. PLOS ONE, 5(3), e9490.

20. Huddleston, J., Hadfield, J., Sibley, T. R., Lee, J., Fay, K., Ilcisin, M. (2021). Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. Journal of Open Source Software, 6(57), 2906.

21. Gillis, S., & Roth, A. (2020). PyClone-VI: scalable inference of clonal population structures using whole genome data. BMC Bioinformatics, 21(1), 571.

22. Hurley, C. K., Kempenich, J., Wadsworth, K., Sauter, J., Hofmann, J. A., Schefzyk, D., Schmidt, A. H., Galarza, P., Cardozo, M. B. R., Dudkiewicz, M., Houdova, L., Jindra, P., Sorensen, B. S., Jagannathan, L., Mathur, A., Linjama, T., Torosian, T., Freudenberger, R., Manolis, A., Mavrommatis, J., Cereb, N., Manor, S., Shriki, N., Sacchi, N., Ameen, R., Fisher, R., Dunckley, H., Andersen, I., Alaskar, A., Alzahrani, M., Hajeer, A., Jawdat, D., Nicoloso, G., Kupatawintu, P., Cho, L., Kaur, A., Bengtsson, M., & Dehn, J. (2020). Common, intermediate and well-documented HLA alleles in world populations: CIWD version 3.0.0. HLA, 95(6), 516-531.

23. Reynisson, B., Alvarez, B., Paul, S., Peters, B., & Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Research, 48(W1), W449-W454.

24. Martincorena, I., Seshasayee, A. S. N., & Luscombe, N. M. (2012). Evidence of non-

random mutation rates suggests an evolutionary risk management strategy. Nature Reviews Genetics, 13(2), 123-130.

25. Hastie, K. M., Li, H., Bedinger, D., Schendel, S. L., Dennison, S. M., Li, K. (2021). Defining variant-resistant epitopes targeted by SARS-CoV-2 antibodies: A global consortium study. Science, 374(6569), 472-478.

26. Currenti, J., Chopra, A., John, M., Leary, S., McKinnon, E., Alves, E., et al. (2019). Deep sequence analysis of HIV adaptation following vertical transmission reveals the impact of immune pressure on the evolution of HIV. PLOS Pathogens, 15(12), e1008177.

27. Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., et al. (2020). Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. Immunity, 53(5), 882-897.e7

Abstract in Korean

시간에 따른 SARS-CoV-2의 돌연변이 분포 변화 및 T-cell epitope 변화

코로나바이러스의 유행 기간 전 세계 사람들은 SARS-CoV-2에 의해 많은 고통을 받고 있었고 백신을 맞은 이후에도 계속해서 걸리는 현상이 있었습니다. 이 바이러스는 돌연변이를 쌓아가면서 사람의 면역을 회피하기도 하고 사람의 세포와 상호작용을 바꾸기도 했다고 알려져 있습니다. 코로나바이러스의 돌연변이와 진화에 대한 이해를 높이고자 2019년부터 2022년까지 6,392,101개의 서열을 이용해 4,590,994개의 돌연변이를 찾아 분석을 진행했습니다. 돌연변이들의 발생과 누적을 확인해 시간에 따라 비율이 증가하는 "증가" 돌연변이를 찾을 수 있었고, 이 돌연변이들은 Spike 단백질의 수용체 결합 부위와 N 말단 영역에 몰려있음을 확인했습니다. 특히 대부분의 돌연변이가 항체가 붙는 위치였음을 확인해 B세포에 의한 면역 압력이 작용했음을 알 수 있었습니다. 이를 통해 유의미하게 증가하는 돌연변이들은 바이러스의 적합도를 높이리라는 것을 알 수 있었습니다. T세포에 의한 면역 압력은 돌연변이에 크게 영향을 주지 못하는 것을 변종과 HLA의 입장에서 친화도와 strong binder를 통해 확인할 수 있었습니다. 이번 연구를 통해 SARS-CoV-2의 돌연변이 형성에 크게 영향을 주는 것은 항체에 의한 면역 압력이라는 것을 확인할 수 있었고, 이후에 새로운 바이러스 또는 변이에 대한 대응에 도움이 될 것으로 생각합니다.

핵심되는 말: SARS-CoV-2, 돌연변이, B세포, T세포, 돌연변이 누적, 면역 압력