




Conserved Pattern and Potential Role of Recurrent Deletions in SARS-CoV-2 Evolution

Shenghui Weng,^{a,b} Hangyu Zhou,^{a,b} Chengyang Ji,^{a,b} Liang Li,^c Na Han,^{a,b} Rong Yang,^{a,b} Jingzhe Shang,^{a,b}  Aiping Wu^{a,b}

^aInstitute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

^bSuzhou Institute of Systems Medicine, Suzhou, China

^cLinyi People's Hospital, Shandong, China

ABSTRACT SARS-CoV-2 continues adapting to human hosts during the current worldwide pandemic since 2019. This virus evolves through multiple means, such as single nucleotide mutations and structural variations, which has brought great difficulty to disease prevention and control of COVID-19. Structural variation, including multiple nucleotide changes like insertions and deletions, has a greater impact relative to single nucleotide mutation on both genome structures and protein functions. In this study, we found that deletion occurred frequently in not only SARS-CoV-2 but also in other SARS-related coronaviruses. These deletions showed obvious location bias and formed 45 recurrent deletion regions in the viral genome. Some of these deletions showed proliferation advantages, including four high-frequency deletions (nsp6 Δ 106-109, S Δ 69-70, S Δ 144, and Δ 28271) that were detected in around 50% of SARS-CoV-2 genomes and other 19 median-frequency deletions. In addition, the association between deletions and the WHO reported variants of concern (VOC) and variants of interest (VOI) of SARS-CoV-2 indicated that these variants had a unique combination of deletion patterns. In the spike (S) protein, the deletions in SARS-CoV-2 were mainly in the N-terminal domain. Some deletions, such as S Δ 144/145 and S Δ 243-244, have been confirmed to block the binding sites of neutralizing antibodies. Overall, this study revealed a conservative regional pattern and the potential effect of some deletions in SARS-CoV-2 over the whole genome, providing important evidence for potential epidemic control and vaccine development.

IMPORTANCE Mutations in SARS-CoV-2 were studied extensively, while only the structure variations on the spike protein were discussed well in previous studies. To study the role of structural variations in virus evolution, we described the distribution of structure variations on the whole genome. Conserved patterns were found of deletions among SARS-CoV-2, SARS-CoV-2-like, and SARS-CoV-like viruses. There were 45 recurrent deletion regions (RDRs) in SARS-CoV-2 generated through the integration of deleted positions. In these regions, four high-frequency deletions parallelly appeared in multiple strains. Furthermore, in the spike protein, the deletions in SARS-CoV-2 were mainly in the N-terminal domain, blocking the binding sites of some neutralizing antibodies, while the structural variations in SARS-related coronavirus were mainly in the N-terminal domain and receptor binding domain. The receptor binding domain is highly related to hosting recognition. The deletions in the receptor binding domain may play a role in host adaptation.

KEYWORDS SARS-CoV-2, recurrent deletion, mutation, structural variation, adaptive evolution

Since the outbreak of COVID-19 caused by SARS-CoV-2 in late 2019, this virus has spread globally for nearly two years. In some regions, although the vaccination rate or infection rate has reached a relatively high level, great concerns have been raised for the continuous variation in SARS-CoV-2 on their ability to escape immune neutralization (1, 2). Recently, a

Editor Mathilde Richard, Erasmus MC

Copyright © 2022 Weng et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Jingzhe Shang, sjz@ism.cams.cn, or Aiping Wu, wap@ism.cams.cn.

The authors declare no conflict of interest.

Received 30 November 2021

Accepted 6 February 2022

Published 7 March 2022

SARS-CoV-2 variant, B.1.617.2, also named the Delta strain by World Health Organization (WHO), has shown increased transmission and immune escape capabilities (3, 4). People with previously induced antibodies still have the risk of infection by this variant (5). Therefore, there is an urgent need to understand the molecular mechanism underlying the adaptive evolution of SARS-CoV-2.

SARS-CoV-2 can take advantage of genome variation to evolve rapidly, including single nucleotide polymorphisms (SNPs) and structural variations (SVs). SVs consist of short fragment insertions, deletions, sequence reversals, and recombination, etc. Current research mainly focused on SNPs (6), but SV changes can include more nucleotides, which may have a greater impact on genomic structure or protein function. Many SVs arise during a viral passage, while only a small part can be retained and spread. These preserved deletions may have played a potential role during the evolution of SARS-CoV-2 (7).

Previous studies have shown that fragment deletions have the possibility to affect the proliferation and transmission of SARS-CoV-2 (8, 9). For instance, a 382-nucleotide deletion in the ORF8 protein weakening the virulence of SARS-CoV-2 was reported in the early stages of the SARS-CoV-2 epidemic in Singapore (8). A Δ 500-532 deletion event was shown to reduce the host INF- β response, a mutation that seemed to occur early in this epidemic and can be found on the nonstructural protein 1 (nsp1) (10). Another 34-nucleotide deletion was found in France on the ORF6 protein. This variant was shown to induce the overexpression of several specific cytokines, including CCL2/MCP1, PTX3, and TNF α , etc., which are involved in the regulation and transduction of NF-kb signaling (11). Recently, in the B.1.1.7 lineage of SARS-CoV-2, Δ 69-70 and Δ 144 were found in the S protein. S Δ 69-70 was shown to increase the viruses' ability to release the S2 structure, which can augment viral infectivity and improve viral syncytium production (12). Based on a bioinformatic analysis, Reham et al. found S Δ 144 can alter the pocket structure on the N-terminal (NTD) of the S protein and reduce the affinity between the NTD and endogenous host antibodies (13).

With the accumulation of site information and structural variations, more SARS-CoV-2 variants with divergent mutations continue to appear in the literature. For instance, the B.1.1.7 variant (the Alpha strain) outbreak occurred in the United Kingdom first, then the B.1.617.2 variant (the Delta strain) outbreak happened in India. These variants have been observed to evade vaccine immunity (3, 14). Four recurrent deletion regions (RDRs), including S Δ 69-70 and S Δ 144, in the S protein, prevent the virus from being bound by neutralizing antibodies (15). The above observations suggest that deletion is one of the ways for SARS-CoV-2 to escape from adaptive immunity and to adapt to their host. Therefore, a systematic analysis of the pattern of deletions in SARS-CoV-2 and their potential effect on immune escape is urgently needed.

In this study, we comprehensively analyzed deletions and insertions in SARS-CoV-2, together with those in SARS-CoV-2-like and SARS-CoV-like viruses. We found that there were conserved patterns of deletions not only in SARS-CoV-2 but also in SARS-CoV-2-like and SARS-CoV-like viruses. Among all recurrent deletion regions (RDRs), SARS-CoV-2 evolved four high-frequency deletions that were found in over 48% of sequenced strains, which were mostly the dominant Alpha strain (lineage B.1.1.7). It is worth noting that the deletions from RDRs were detected in all six variants of concern as defined by the WHO (16) with different combinations. Furthermore, the NTD and RBD regions of the S protein possess multiple RDR regions, which may promote rapid viral adaptation to the host.

RESULTS

Common RDRs in SARS-CoV-2, SARS-CoV-2-like, and SARS-CoV-like viruses. Previous studies have shown a regional preference of deletions in the NTD domain of the S protein in SARS-CoV-2, forming four recurrent deletion regions (RDR1-4) (15). Here, we systematically analyzed these deletion and insertion events in 1,289,583 high-quality SARS-CoV-2 genomes downloaded on July 05, 2021, from GISAID (see in Method). In total, 1007 unique deletions and 387 unique insertions (Table S1 and S2) were detected. The maximum number

of occurrences of these unique insertions was only 397 times in the dataset, while that of these unique deletions was 685,744 (53.18%) times, which indicates that deletions were the major structural variations in SARS-CoV-2. Across the entire genome, these deletions showed a clear regional preference (Fig. 1A), which mainly occurred in the *nsp1*, *nsp2*, *nsp3*, *nsp4*, *nsp6*, *S*, *N* proteins, and accessory proteins (Fig. S1A). After the integration of these deleted positions, 45 RDRs were generated (Table S3). Furthermore, we found that the diversity of deletions increased with time significantly (Fig. S1B).

To investigate these biased RDRs in other coronaviruses, we collected genome sequences for SARS-CoV-2-like and SARS-CoV-like viruses from different hosts and generated a set of deletions referring to SARS-CoV-2 or SARS-CoV (Table S4). In these sequences, we found that most deletions were in three regions, forming three high deletion/insertion areas (HDA 1-3) in the front part of the *nsp3*, *S*, and ORF8, respectively (Fig. 1B and C). The pattern indicated that these three HDAs were conserved among SARS-related coronaviruses. In these three proteins in the SARS-CoV-2 genome, there were twelve RDRs, seven RDRs, and one 436-nt large RDR, respectively. These facts led us to speculate the roles of these RDRs and HDAs in the evolution of coronaviruses. In addition, when we pulled out the aligned sequences of *S* Δ 69-70 in SARS-CoV-2-like viruses, we found that the high-frequency *S* Δ 69-70 deletion in SARS-CoV-2 also existed in these SARS-related sequences (Fig. 1D). This finding further indicated that deletions in HDAs were not randomly distributed. These deletions with a location preference could be the result of adaptive selection.

Diversity of deletion types in RDRs. Among the 45 RDRs in the SARS-CoV-2 genome, the distribution of the deletions in RDRs showed location-dependent characteristics. In the *S* protein, RDRs were located in its NTD domain. In the first fifth of the ORF3a sequence, there was one long 122-nt RDR. RDRs were identified out in a cluster between the *M* protein and *N* protein, also covering four accessory proteins (ORF6, ORF7Aa, ORF7b, and ORF8). Two longest RDRs were involved in this cluster, including one 402-nt RDR and one 436-nt RDR (Fig. 2A). The discontinuous transcription mechanism of SARS-CoV-2 may be the reason underlying the location preference of these RDRs (17). When we studied the association between the length and the deletion type of these RDRs, we found that more deletion types were identified in longer RDRs. In the five longest RDRs, there were more than 30 deletion types detected. Especially, among the 240-nt RDR in the *nsp3*, 5 deletion types have been identified. While in the *S* protein, among the 49-nt RDR22 and 37-nt RDR23, their deletion types were as high as 32 and 42, respectively. The relatively high diversity of these short RDRs of *S* protein could be due to the important role of the *S* protein in the adaptive evolution of the SARS-CoV-2 virus. These results showed that deletions are prone to occur in some specific RDRs.

Though there were 45 RDRs, including 842 deletion types that have been identified (Table S3) in the SARS-CoV-2 genome, only 4 high-frequency deletions were observed in over 600,000 strains (48.58%) among all genomes (Fig. 2B). They were *nsp6* Δ 106-108, Δ 28271 (a single-nucleotide intergenic deletion before the protein *N*), and two widely studied deletions, *S* Δ 69-70 and *S* Δ 144. It is worth noting that two of these deletions, *S* Δ 69-70 and *S* Δ 144, have been reported to occur spontaneously in immunodeficient patients (12, 18). In addition, 19 median-frequency deletions were identified in the *nsp1*, *nsp6*, *S*, and *N*, as well as four accessory proteins (ORF3a, ORF6, ORF7a, and ORF8) (Fig. 2C). By analyzing the temporal and spatial distribution of these four high-frequency deletions, we found that these deletions shared a similar growth pattern on six continents. They all emerged first in Europe at the end of 2020 and then spread to other regions (Fig. S2). Phylogenetic trees were used to analyze whether these mutations were distributed repeatedly and widely. The results showed that these four high-frequency deletions were distributed in parallel branches (Fig. 2D). Their repeated and independent occurrence pattern indicated their potential advantages for viral adaptation.

Relationships between deletions and SARS-CoV-2 variants. Four high-frequency deletions were found in a similar spatiotemporal pattern, which indicated there were

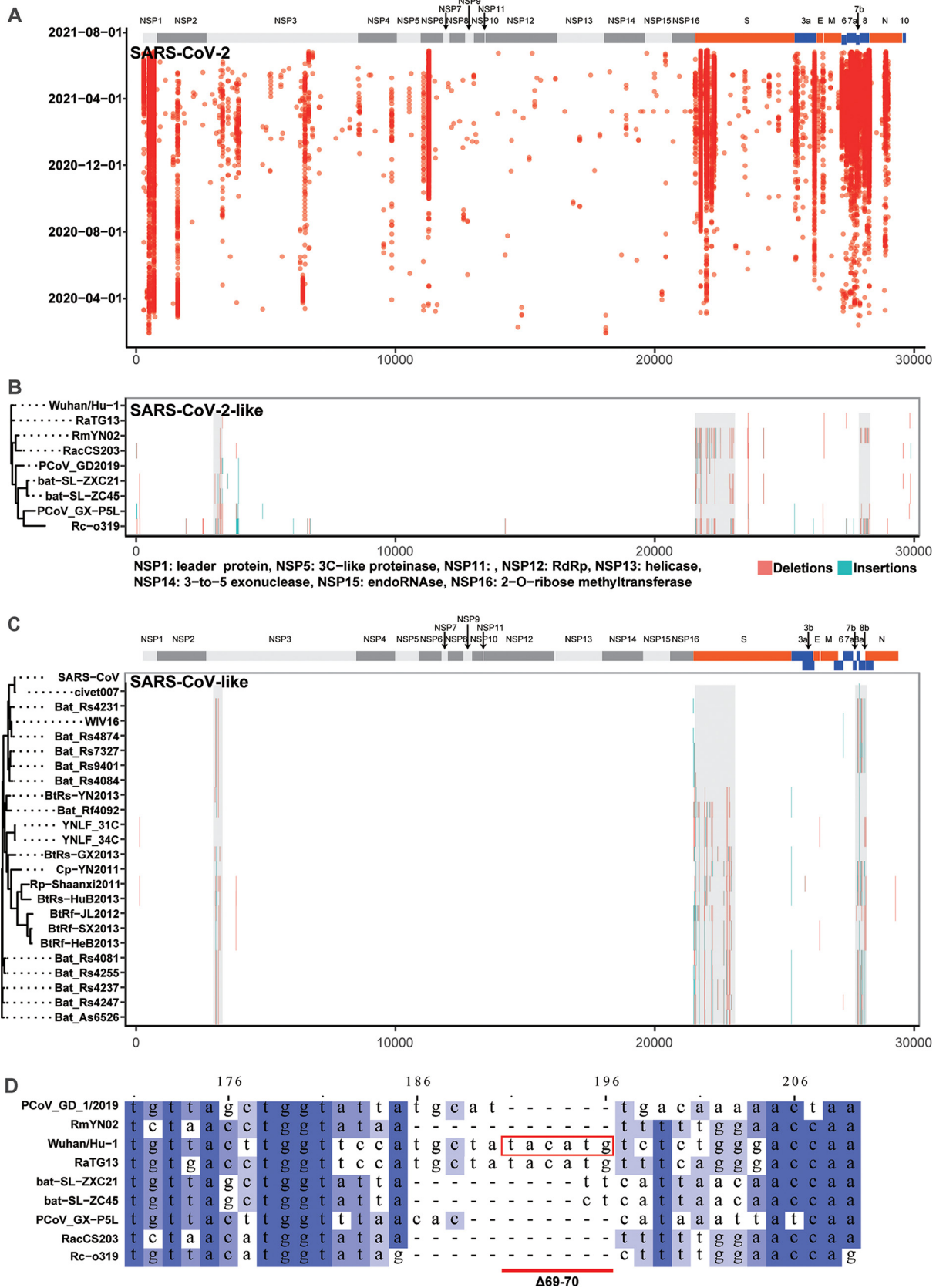


FIG 1 The genomic distribution of insertions and deletions in SARS-CoV-2 and SARS-CoV-2-like viruses. (A) Deletions are dotted in the SARS-CoV-2 genome (X-axis) with time (Y-axis). The distribution of insertions (blue) and deletions (red) in SARS-CoV-2-like (B) and SARS-CoV-like viruses (C) (Continued on next page)

internal associations between them. Venn diagram analysis showed that most of these four deletions appeared together in about 580,000 (45%) sequenced SARS-CoV-2 genomes (Fig. 3A). Except for the cooccurrence of four high-frequency deletions, other combinations among these four deletions also occurred repeatedly. We named each combination of deletions from group a to group o (Fig. 3B). Five of these combinations were observed more than 10,000 times, which were relatively frequent since this was higher than that of the middle-frequency deletions shown in Fig. 2C. Recently, the WHO reported variants of concern (VOC) and variants of interest (VOI) of SARS-CoV-2 based on their potential transmission risk and immune escape abilities. When we calculated the ratios of each combination in different variants, we found that except for the combination of four high-frequency deletions, the combination with three of four high-frequency deletions were also mainly belonged to the Alpha strain (lineage B.1.1.7). Group n with only nsp6 deletion belonged to the Beta (lineage B.1.351) and the Gamma strains (lineage P.1). Group o with the intergenic deletion mainly belonged to the Delta strain (lineage B.1.617.2). Other groups (h, i, k, and l) widely distributed in different SARS-CoV-2 lineages (Fig. 3B).

Apart from four high-frequency deletions, we observed that each SARS-CoV-2 variant had its unique combination of deletions and mutations (Fig. 3C). The Alpha variant (lineage B.1.1.7) contained all four high-frequency deletions, while other variants contained only a part of these deletions. In addition, we found that the deletion $\Delta 241-243$ in the S protein appeared in the Beta variant (lineage B.1.351), while the deletions ORF6 $\Delta 2$ and N $\Delta 2$ appeared in the Eta variant (lineage B.1.525). Notably, the recent SARS-CoV-2 variant Delta (lineage B.1.617.2), contained two intermediate-frequency deletions, namely, S $\Delta 156-157$ and ORF8 $\Delta 119-120$ (Fig. 3D). These results showed that variants formed at different times and in different environments carried their unique deletions and mutations.

Association of deletions and mutations in the S protein. To explore whether these deletions could influence viral antigenicity or change the binding regions for neutralizing antibodies, we first analyzed the relationship between deletions, mutations, antigenic sites, and the binding regions for neutralizing antibodies in the S protein (Fig. 4 and Fig. S3). We found that the high-prone mutation regions staggered to the RDRs. The NTD of the S protein was a high-risk region for deletions, while multiple high-frequency mutations were found in the S2 part of the S protein (Fig. 4A and B). In SARS-CoV-2, IgG and IgA epitopes mainly located in the S2 domain (Fig. 4C) (19). Previous studies had proved that most of the mutations cannot change the antigenic site of the virus (20). We further collected the currently reported neutralizing antibodies for the SARS-CoV-2 S protein. These antibodies could be divided into three types according to their binding sites (NTD, RBM, and HR2) (Table S5). Deletions in the NTD partially overlapped with the binding sites of these neutralizing antibodies (Fig. 4D). When we mapped RDRs and HDAs to the 3-D structure of the S protein, we found that, in SARS-CoV-2-like viruses, HDAs mainly aggregated in the RBD. In SARS-CoV-2, however, RDRs were mainly in the NTD. RDRs and HDAs partially overlapped in the NTD covering S $\Delta 69-70$. (Fig. 4E). The observed deletion site distribution was different in SARS-CoV-2 and SARS-CoV-2-like viruses. However, since all these strains belong to the same species, the observed variation may be due to the limited evolutionary time. Given a long period of evolution, deletions in SARS-CoV-2 and SARS-CoV-2-like viruses may tend to a similar pattern. All the results above indicated that some deletions in the S protein may contribute to the viral adaption, including the viral transmissibility and immune escape (20).

DISCUSSION

Deletions were frequently and widely occurring in SARS-CoV-2, yet recent studies mainly focused on the S protein. Our analyses showed an overall distribution profile of deletions

FIG 1 Legend (Continued)

are shown in the same way. Three shared high deletion/insertion areas 1-3 (HDAs 1-3) are highlighted with grey boxes. The reference genomes are EPI_ISL_402124 and NC_004718.3 for SARS-CoV-2 and SARS-CoV, respectively. (D) Context of S $\Delta 69-70$ in aligned SARS-CoV-2-like virus genome sequences. S $\Delta 69-70$ of SARS-CoV-2-like viruses is framed out by a red rectangle.

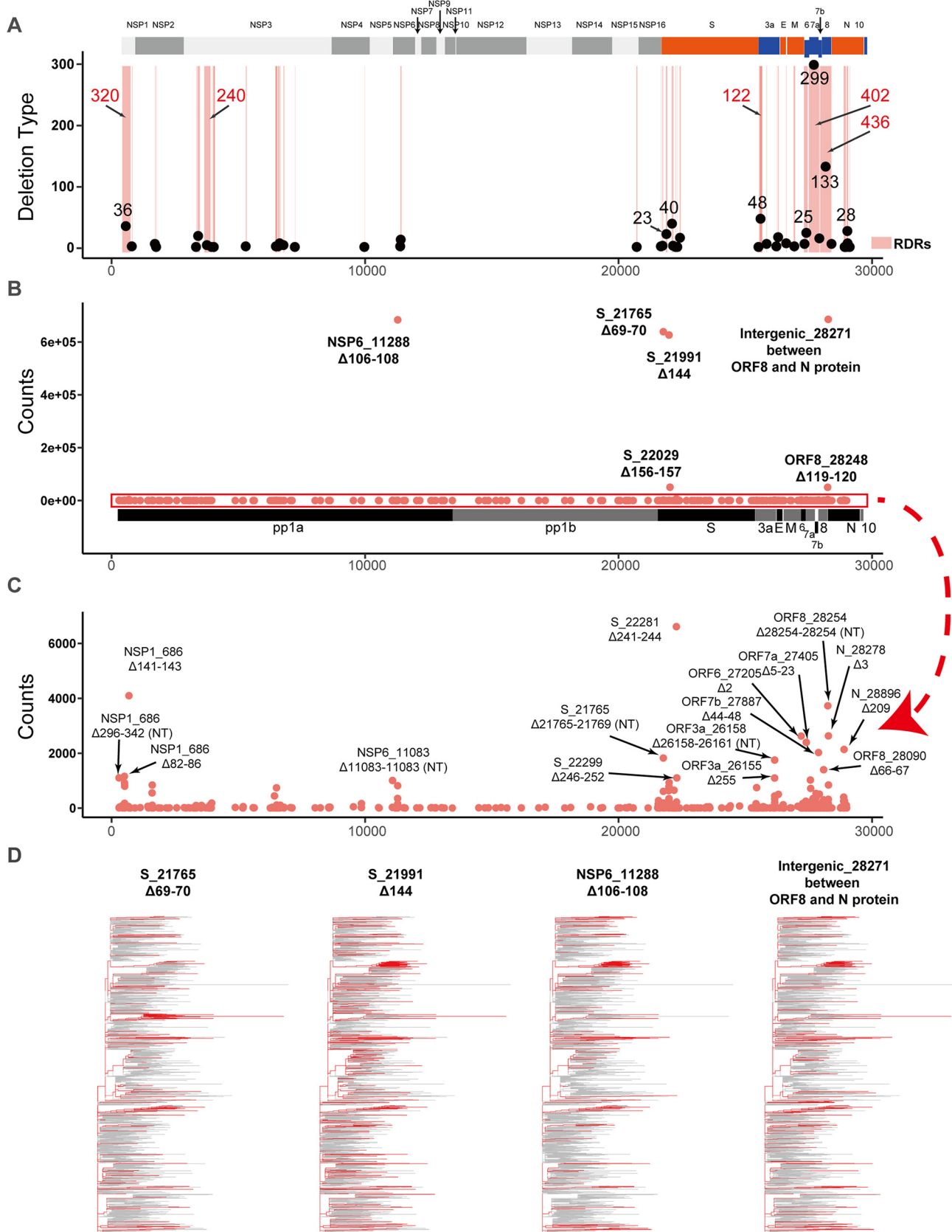


FIG 2 Deletion types in the RDR regions of the SARS-CoV-2 genome. (A) The distribution of RDR regions on the SARS-CoV-2 genome together with the number of deletion types in each RDR. The RDR regions are in red boxes. The length of the regions that are longer than 100 nt is labeled under (Continued on next page)

over the entire SARS-CoV-2 genome. We found that these deletions had a significant regional preference. We further extended this study to SARS-CoV-2-like and SARS-CoV-like viruses, finding that in these sequences deletion and insertion events also had a regional preference. These results implied that RDRs may have played a role in the evolution of SARS-CoV-2 and SARS-related coronaviruses.

Within all RDRs, four high-frequency deletions were detected majorly in the Alpha variant, which indicated the rapid increase of these deletions was because of the widespread outbreak of Alpha strain. Among these four deletions, S Δ 144 was already proved to be involved in the viral escape from neutralizing antibodies (15). S Δ 69-70 was involved in the increasement of cell entry efficiency (12). However, the function of middle-frequency deletions was still unclear. Furthermore, the cooperation of these deletions with some SNPs may play a certain role in the SARS-CoV-2 adaption and evolution. Therefore, further studies were urgently needed to understand their role in viral evolution and transmission.

The RNA-RNA interaction may trigger these deletions during viral replication. Lei et al identified the SARS-CoV-2 RNA genome structure by icSHAPE and found a large number of RNA-RNA interaction regions. Omer et al. found that SARS-CoV-2 had many short- or long-distance RNA-RNA interactions within cells (21). A study revealed the structural variants were enriched in the transcription regulatory site (TRS) of the SARS-CoV-2 genome (17). Here, we also found that a portion of identified RDRs was also located in front of the ORFs. More studies are required to reveal the reasons for the occurrence of deletions and their location preference.

The occurrence of deletions has been shown to lead to the immune escape of SARS-CoV-2 strains from neutralizing antibodies such as 4A8 (15). In this study, we found that the neutralizing antibody binding sites, which were mainly located at the NTD and RBD of SARS-CoV-2, overlapped with RDRs in the NTD domain. Furthermore, these RDRs and mutations in the S protein were present in a staggering arrangement. RDRs appeared mostly in the NTD of the S protein, while most of the high-frequency mutations presented in the S2. The complementary relationship between deletions and mutations indicated that SARS-CoV-2 evolved through using deletions to partly escape host immunity. The deletions with regional preference may work synergistically with other mutations to yield more comprehensive and rapid adaptability.

Since current SARS-CoV-2 vaccines were mainly developed against the S protein, these insertion and deletion speculations raise many questions. For instance, whether the development of vaccines can tolerate these SVs? Are the vaccines already on the market significantly weakened or partly weakened due to these deletions? Some studies have proved the role of the deletions in the S protein in viral adaption, especially in the changing NTD antigenicity from potently neutralizing convalescent plasma or specific neutralizing antibodies (20). The deletions on S Δ 144/145 and S Δ 243-244 were confirmed at the binding sites of a neutralizing antibody 4A8. These two deletions were proved to have the ability to abolish 4A8 binding (15). Therefore, these SVs (deletions and insertions) require careful monitoring and tracking in the future.

MATERIALS AND METHODS

Sequence source. The aligned SARS-CoV-2 sequences were acquired on July 8, 2021, from the GISAID database (22). All sequences were collected before July 5, 2021. The sequences longer than 29000 nt have already been aligned using MAFFT in the GISAID database. After downloading these aligned sequences, quality control was operated according to the sequence quality standard following National Information Center, together with a host screening. Sequences owing more than 15 Ns or 50 merged bases were discarded and only those isolated from human samples were kept. There were finally 1,289,583 sequences used in deletion and insertion identification and further analysis. At the same time, we collected from GISAID the pedigree information, sampling time, and location information of these SARS-CoV-2 sequences.

Insertion, deletion, and mutation identification. Sequences related to SARS-CoV and SARS-CoV-2 were collected and treated by the tool 'Genome-to-Variants' from the website of the China National Center for Bioinformatics (<https://ngdc.cncb.ac.cn/ncov/online/tool/variation>) to obtain mutation information (23). The

FIG 2 Legend (Continued)

their boxes. The black points on the red boxes represent the count of their pattern type. The counts larger than 20 times are labeled out above the points. The high-frequency (B) and medium-frequency (C) deletions were observed in the SARS-CoV-2 genome. High-frequency deletions occurred more than 600, 000 times, and middle-frequency deletions showed more than 1, 000 times. (D) The variants containing four high-frequency deletions (red) are highlighted in the phylogenetic tree of SARS-CoV-2. The other variants are colored in grey as background.

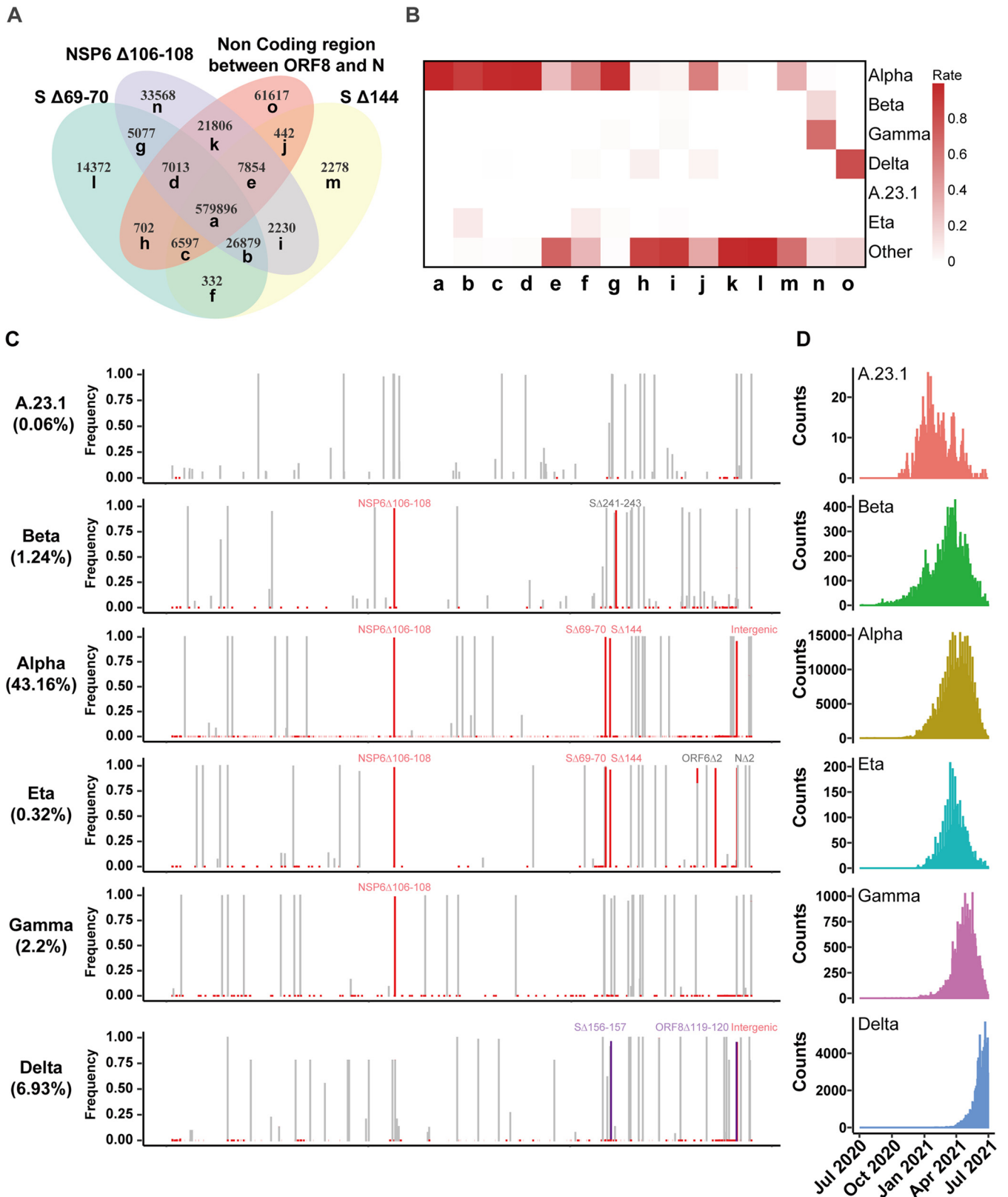


FIG 3 Deletions in different SARS-CoV-2 variants. (A) The number of overlapping sequences of high-frequency deletions in SARS-CoV-2 is shown in a Venn diagram. Group a-o represents each part in the Venn diagram. (B) Group a-o consists of various forms of deletion combination. The gradation of color represents the rate of each combination in each strain. The sum of each column is 1. (C) Deletions (red) and mutations (grey) are mapped in six VOC and VOI SARS-CoV-2 strains. The high-frequency and median-frequency deletions are colored in red and purple, respectively. (D) The number of daily samples of VOC and VOI strains is displayed in bar plots.

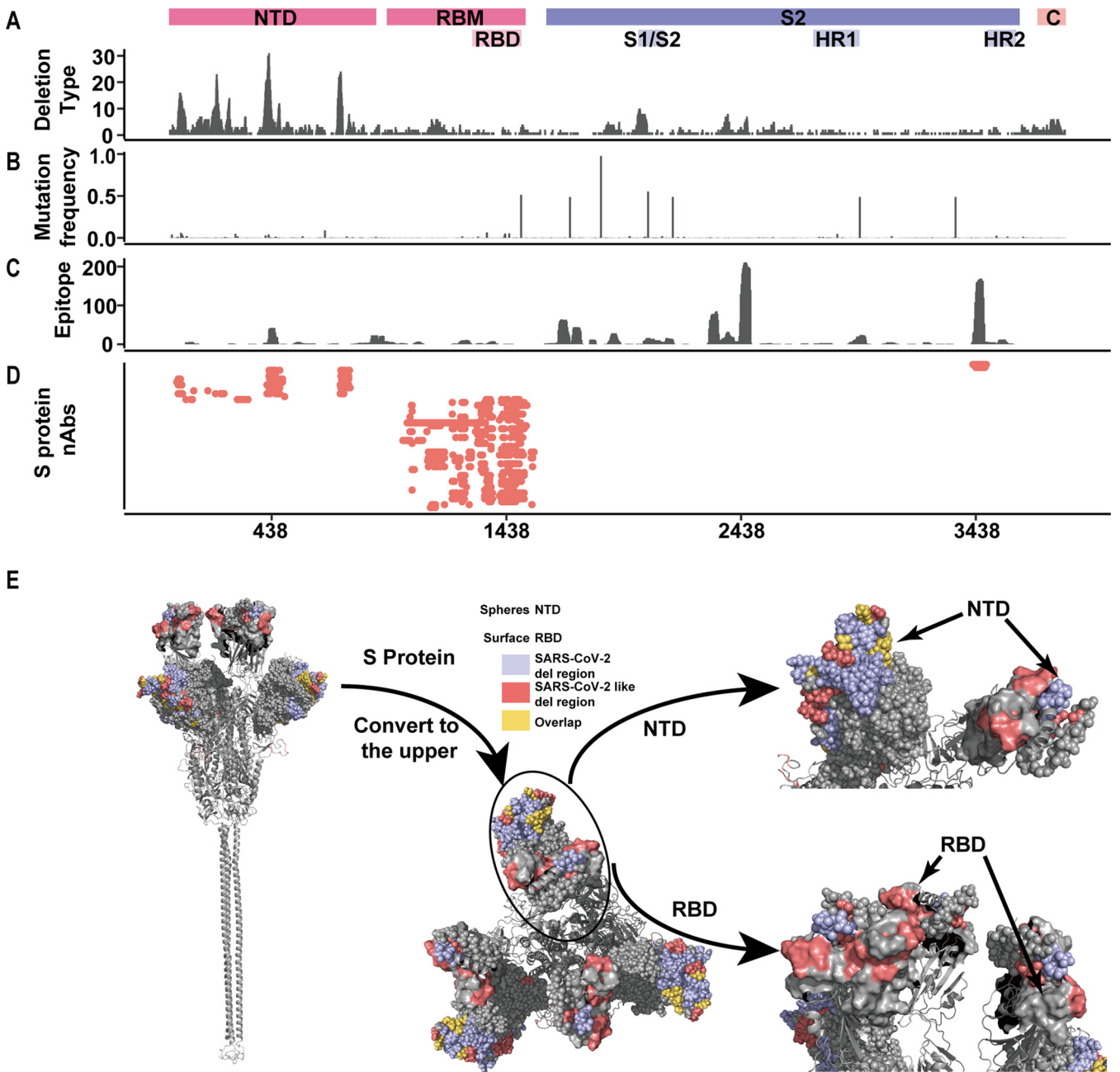


FIG 4 Relationships between deletions, mutations, and antibody binding regions in the S protein of SARS-CoV-2. (A–D) The distribution of deletions (A), mutations (B), IgA and IgG epitopes (C), and known neutralizing antibodies (D) on the SARS-CoV-2 S protein. (E) The RDRs are displayed on the 3-D model of the SARS-CoV-2 S protein. NTD and RBD are labeled as spheres and surfaces. The RDRs of SARS-CoV-2 (blue) and SARS-CoV-2-like viruses (red) are colored, and their overlapping area is in yellow.

tool aligns each sequence to its reference sequences and then lists variations in VCF format. The reference sequences for SARS-CoV-like and SARS-CoV-2-like sequences were [NC_004718.3](https://www.ncbi.nlm.nih.gov/nuccore/NC_004718.3) (NCBI Reference Sequence) and [EPI_ISL_402124](https://www.gisaid.org/sequence/EPI_ISL_402124) (GISAI Reference Sequence), respectively. The insertion and deletion locations were extracted from the VCF file. Since this online tool is hard to treat a big dataset of SARS-CoV-2, we used an R script to treat the mega sequence data which was described in Fig. S4 in the supplemental material Fig. S4. The sequence named [EPI_ISL_402124](https://www.gisaid.org/sequence/EPI_ISL_402124) was used as a reference sequence to extract the mutations and SVs information. We pulled out the reference sequence and compared it to the other sequences one by one. To avoid interference with sequencing quality, only the sites with a gap against normal bases (A, T, C, and G) will be treated as insertions or deletions. The R script is available at https://github.com/wuaipinglab/genome_treatment.

Phylogenetic tree analysis. The phylogenetic trees of SARS-CoV-like and SARS-CoV-2-like viruses were built by the ORF1b and constructed with the software 'FastTree' with version 2.1.9 (24) with the parameters "Fasttree-gtr-nt." The SARS-CoV-2 phylogenetic tree was also constructed using FastTree software using their full-length sequences with the same parameters. The phylogenetic tree for each high-frequency deletion

was shown on a background. The background sequence for viral evolution was composed of the latest sequence in each PANGO lineage. The PANGO lineage with deletions was highlighted in red.

Recurrent deletion region identification. To assemble a reasonable recurrent region, all the deletions that happened less than five times were removed. The remained deletions were joined by their location on the SARS-CoV-2 reference genome. The assembled area was defined as recurrent deletion regions (RDRs). For further analysis of the characteristic of these RDRs, the counts of the deletion types in each RDR were recorded. In SARS-CoV-like viruses, the insertion and deletion positions were uniformly corrected, based on the starting position of each protein against SARS-CoV-2.

RDR visualization in protein structures. The simulated S protein structure, which belongs to lineage A, was created by Zhang Yang lab's website (<http://zhanglab.ccmb.med.umich.edu/COVID-19/>). The 3D structure visualization was done in PyMOL (25).

Neutralizing antibody. The SARS-CoV-2 antibody information was collected from the coronavirus antibody database CoV-AbDab (26). The antibodies owing the ability to neutralize the SARS-CoV-2 virus were selected. Their target sites on the virus were collected from their original research articles which were listed in Table S3 in the supplemental material. We used an R script to display these neutralizing antibodies with detailed binding positions on the S protein.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, XLS file, 0.1 MB.

SUPPLEMENTAL FILE 2, XLS file, 0.1 MB.

SUPPLEMENTAL FILE 3, XLS file, 0.02 MB.

SUPPLEMENTAL FILE 4, XLS file, 0.03 MB.

SUPPLEMENTAL FILE 5, XLS file, 0.04 MB.

SUPPLEMENTAL FILE 6, PDF file, 4.2 MB.

ACKNOWLEDGMENTS

We acknowledge the members of the Wu laboratory for insightful discussions regarding this study. We gratefully acknowledge the laboratories who shared the sequence data via the GISAID.

This work was supported by the National key research and development program (2021YFC2301300); the CAMS Innovation Fund for Medical Sciences (2021-I2M-1-061); the National Natural Science Foundation of China (92169106, 31900472); the special research fund for central universities, Peking Union Medical College (2021-PT180-001); Suzhou science and technology development plan (szs2020311).

Conceived and designed the experiments: J.Z.S, A.P.W. Contributed to the data collection: J.Z.S, S.H.W. Data interpretation and discussion: H.Y.Z, C.Y.J, N.H, L.L, R.Y. Wrote the paper: A.P.W, J.Z.S, S.H.W. Reviewed the paper: J.Z.S, A.P.W.

We have no conflicts of interest to declare.

REFERENCES

1. Tang JW, Tambyah PA, Hui DS. 2020. Emergence of a new SARS-CoV-2 variant in the UK. *Journal of Infect* 82:E27–E28. <https://doi.org/10.1016/j.jinf.2020.12.024>.
2. Rice BL, Annapragada A, Baker RE, Bruijning M, Dotse-Gborgborsi W, Mensah K, Miller IF, Motaze NV, Raheerinandrasana A, Rajeev M. 2021. Variation in SARS-CoV-2 outbreaks across sub-Saharan Africa. *Nature Medicine* 27:447–453. <https://doi.org/10.1038/s41591-021-01234-8>.
3. Davis C, Logan N, Tyson G, Orton R, Harvey W, Haughney J, Perkins J, Peacock T, Barclay WS, Cherepanov P. 2021. Reduced neutralisation of the Delta (B. 1.617. 2) SARS-CoV-2 variant of concern following vaccination. *medRxiv*. <https://doi.org/10.1101/2021.06.23.21259327>.
4. Liu C, Ginn HM, Dejnirattisai W, Supasa P, Wang B, Tuekprakhon A, Nutalai R, Zhou D, Mentzer AJ, Zhao Y. 2021. Reduced neutralization of SARS-CoV-2 B. 1.617 by vaccine and convalescent serum. *Cell* 184:4220–4236. <https://doi.org/10.1016/j.cell.2021.06.020>.
5. Campbell F, Archer B, Laurenson-Schafer H, Jinnai Y, Konings F, Batra N, Pavlin B, Vandemaale K, Van Kerkhove MD, Jombart T. 2021. Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Eurosurveillance* 26:2100509. <https://doi.org/10.2807/1560-7917.ES.2021.26.24.2100509>.
6. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, Doolabh D, Pillay S, San EJ, Msomi N. 2021. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* 592:438–443. <https://doi.org/10.1038/s41586-021-03402-9>.
7. Wellenreuther M, Mérot C, Berdan E, Bernatchez L. 2019. Going beyond SNPs: the role of structural genomic variants in adaptive evolution and species diversification. *Molecular ecology* 28:1203–1209. <https://doi.org/10.1111/mec.15066>.
8. Young BE, Fong S-W, Chan Y-H, Mak T-M, Ang LW, Anderson DE, Lee CY-P, Amrun SN, Lee B, Goh YS. 2020. Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. *The Lancet* 396:603–611. [https://doi.org/10.1016/S0140-6736\(20\)31757-8](https://doi.org/10.1016/S0140-6736(20)31757-8).
9. Benedetti F, Snyder GA, Giovanetti M, Angeletti S, Gallo RC, Ciccozzi M, Zella D. 2020. Emerging of a SARS-CoV-2 viral strain with a deletion in nsp1. *Journal of Translational Medicine* 18:1–6. <https://doi.org/10.1186/s12967-020-02507-5>.
10. Lin J-w, Tang C, Wei H-c, Du B, Chen C, Wang M, Zhou Y, Yu M-x, Cheng L, Kuivanen S. 2021. Genomic monitoring of SARS-CoV-2 uncovers an Nsp1 deletion variant that modulates type I interferon response. *Cell host & microbe* 29:489–502. <https://doi.org/10.1016/j.chom.2021.01.015>.
11. Quéromès G, Destras G, Bal A, Regue H, Burfin G, Brun S, Fanget R, Morfin F, Valette M, Trouillet-Assant S. 2021. Characterization of SARS-CoV-2 ORF6 deletion variants detected in a nosocomial cluster during routine genomic

- surveillance, Lyon, France. *Emerging microbes & infections* 10:167–177. <https://doi.org/10.1080/22221751.2021.1872351>.
12. Meng B, Kemp SA, Papa G, Datir R, Ferreira IA, Marelli S, Harvey WT, Lytras S, Mohamed A, Gallo G. 2021. Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the variant of concern lineage B.1.1.7. *Cell Reports* 35:109292. <https://doi.org/10.1016/j.celrep.2021.109292>.
 13. Dawood RM, El-Meguid MA, Salum GM, El-Wakeel K, Shemis M, El Awady MK. 2021. Bioinformatics prediction of B and T cell epitopes within the spike and nucleocapsid proteins of SARS-CoV2. *Journal of Infection and Public Health* 14:169–178. <https://doi.org/10.1016/j.jiph.2020.12.006>.
 14. Kuzmina A, Khalaila Y, Voloshin O, Keren-Naus A, Boehm L, Raviv Y, Shemer Avni Y, Rosenberg E, Taube R. 2021. SARS CoV-2 escape variants exhibit differential infectivity and neutralization sensitivity to convalescent or post-vaccination sera. *Cell Host Microbe* 29:522–528.e2. <https://doi.org/10.1016/j.chom.2021.03.008>.
 15. McCarthy KR, Rennick LJ, Nambulli S, Robinson-McCarthy LR, Bain WG, Haidar G, Duprex WP. 2021. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* 371:1139–1142. <https://doi.org/10.1126/science.abf6950>.
 16. Rambaut A, Holmes EC, O’Toole Á, Hill V, McCrone JT, Ruis C, du Plessis L, Pybus OG. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature microbiology* 5:1403–1407. <https://doi.org/10.1038/s41564-020-0770-5>.
 17. Chrisman B, Paskov K, Stockham N, Tabatabaei K, Jung J-Y, Washington P, Varma M, Sun MW, Maleki S, Wall DP. 2020. Structural variants in SARS-CoV-2 occur at template-switching hotspots. *bioRxiv*. <https://doi.org/10.1101/2020.09.01.278952>.
 18. Kemp SA, Collier DA, Datir R, Ferreira IA, Gayed S, Jahun A, Hosmillo M, Rees-Spear C, Mlcochova P, Lumb IU. 2020. Neutralising antibodies in spike mediated SARS-CoV-2 adaptation. *medRxiv*. <https://doi.org/10.1101/2020.12.05.20241927>.
 19. Shrock E, Fujimura E, Kula T, Timms RT, Lee I-H, Leng Y, Robinson ML, Sie BM, Li MZ, Chen Y. 2020. Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity. *Science* 370. <https://doi.org/10.1126/science.abd4250>.
 20. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, Ludden C, Reeve R, Rambaut A, Peacock SJ. 2021. SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology* 19: 409–424. <https://doi.org/10.1038/s41579-021-00573-0>.
 21. Ziv O, Price J, Shalamova L, Kamenova T, Goodfellow I, Weber F, Miska EA. 2020. The short-and long-range RNA-RNA Interactome of SARS-CoV-2. *Molecular cell* 80:1067–1077. <https://doi.org/10.1016/j.molcel.2020.11.004>.
 22. Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID’s innovative contribution to global health. *Global challenges* 1:33–46. <https://doi.org/10.1002/gch2.1018>.
 23. Gong Z, Zhu J-W, Li C-P, Jiang S, Ma L-N, Tang B-X, Zou D, Chen M-L, Sun Y-B, Song S-H. 2020. An online coronavirus analysis platform from the National Genomics Data Center. *Zoological research* 41:705. <https://doi.org/10.24272/j.issn.2095-8137.2020.065>.
 24. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PloS One* 5:e9490. <https://doi.org/10.1371/journal.pone.0009490>.
 25. DeLano WL. 2002. Pymol: an open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography* 40:82–92.
 26. Raybould MI, Kovaltsuk A, Marks C, Deane CM. 2021. CoV-AbDab: the coronavirus antibody database. *Bioinformatics* 37:734–735. <https://doi.org/10.1093/bioinformatics/btaa739>.