ORIGINAL RESEARCH
published: 02 June 2022
doi: 10.3389/fgene.2022.875406

Increased Frequency of Indels in Hypervariable Regions of SARS-CoV-2 Proteins—A Possible Signature of Adaptive Selection

*Arghavan Alisoltani[1], Lukasz Jaroszewski[1], Mallika Iyer[2], Arash Iranzadeh[3] and Adam Godzik[1]**

[1]Biosciences Division, School of Medicine, University of California, Riverside, Riverside, CA, United States, [2]Graduate School of Biomedical Sciences, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, United States, [3]Computational Biology Division, Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa

Most attention in the surveillance of evolving SARS-CoV-2 genome has been centered on nucleotide substitutions in the spike glycoprotein. We show that, as the pandemic extends into its second year, the numbers and ratio of genomes with in-frame insertions and deletions (indels) increases significantly, especially among the variants of concern (VOCs). Monitoring of the SARS-CoV-2 genome evolution shows that co-occurrence (i.e., highly correlated presence) of indels, especially deletions on spike N-terminal domain and non-structural protein 6 (NSP6) is a shared feature in several VOCs such as Alpha, Beta, Delta, and Omicron. Indels distribution is correlated with spike mutations associated with immune escape and growth in the number of genomes with indels coincides with the increasing population resistance due to vaccination and previous infections. Indels occur most frequently in the spike, but also in other proteins, especially those involved in interactions with the host immune system. We also showed that indels concentrate in regions of individual SARS-CoV-2 proteins known as hypervariable regions (HVRs) that are mostly located in specific loop regions. Structural analysis suggests that indels remodel viral proteins' surfaces at common epitopes and interaction interfaces, affecting the virus' interactions with host proteins. We hypothesize that the increased frequency of indels, the non-random distribution of them and their independent co-occurrence in several VOCs is another mechanism of response to elevated global population immunity.

Keywords: indels, SARS-CoV-2, protein loop, hypervariable regions (HVR), variants of concern (VOCs)

Edited by:
*Dariusz Plewczynski,
Warsaw University of Technology,
Poland*

Reviewed by:
*Haizhou Liu,
Wuhan Institute of Virology (CAS),
China
Sayaka Miura,
Temple University, United States*

Correspondence:
*Adam Godzik
adam.godzik@medsch.ucr.edu*

Specialty section:
*This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics*

Received: *14 February 2022*
Accepted: *08 April 2022*
Published: *02 June 2022*

Citation:
*Alisoltani A, Jaroszewski L, Iyer M, Iranzadeh A and Godzik A (2022) Increased Frequency of Indels in Hypervariable Regions of SARS-CoV-2 Proteins—A Possible Signature of Adaptive Selection.
Front. Genet. 13:875406.
doi: 10.3389/fgene.2022.875406*

INTRODUCTION

Insertions/deletions (indels), are the second most common modifications in the evolution of viral genomes after single nucleotide polymorphisms (SNPs), yet receive relatively little attention in genome analyses (Palmer and Poon, 2019). One of the reasons for that is that their consequences on protein structure and function are more challenging to determine than SNPs. Examples of long, loss-of-function deletions removing entire proteins or functional domains were shown to be deleterious (Zwart et al., 2014) or attenuating (Oostra et al., 2007); however, the effects of shorter, function-refining indels are mostly unknown. Such indels tend to happen in the loops between secondary structure elements, but interestingly not in all the loops, so their distribution cannot be explained by the plasticity of protein structure alone. Such indels rarely affect the overall structure of proteins, but may alter the binding specificity or protein-protein interaction surfaces (Studer et al., 2013), in few studied examples leading to increased drug resistance and immune escape in viruses (Wood et al., 2009; Palmer and Poon, 2019). Their prevalence, evolutionary dynamics, and overall consequences for fitness of most viruses, including SARS-CoV-2, largely remain unacknowledged and unaddressed.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) first emerged in Wuhan, China and subsequently spread worldwide and infected millions of people in several waves of evolving variants. Its high mutability (van Dorp et al., 2020), typical for RNA viruses (Duffy, 2018) but exacerbated by the scale of the COVID-19 pandemic, has resulted in the emergence of multiple lineages. Higher infectivity, transmissibility and/or lower efficacy of the current vaccines have been reported for Beta (B.1.351*) (Tegally et al., 2021), Gamma (*p.**) (Jewell, 2021; Madhi et al., 2021), Delta (B.1.617.2, AY. *) (Planas et al., 2021) (Cherian et al., 2021), Lambda (C.37) (Kimura et al., 2022) and Omicron variants (B.1.1.529 and BA.*) (Karim and Karim, 2021; Viana et al., 2021). Tracking and analyzing new emerging lineages with modified disease phenotypes, dubbed variants of concern (VOCs) (Plante et al., 2021), is crucial for determining the strategies of fighting the COVID-19 pandemic. Massive sequencing of SARS-CoV-2, with over 10M genomes available today gives the United States a unique opportunity to study its evolution on the timescale of weeks or even days, as compared to much longer timescales available by comparing species. Much attention has been focused on specific mutations, such as E484K in the spike protein and their effects on host immune response (Starr et al., 2020; Jangra et al., 2021). At the same time, deletions and insertions received less attention, being less frequent, especially in the first phase of the pandemic and more challenging to interpret.

Still, several specific indels in SARS-CoV-2 in the envelope protein (Kumar et al., 2021), non-structural protein 1 (NSP1) (Lin et al., 2021), spike glycoprotein (spike or S) (McCarthy et al., 2021) and accessory ORFs (Lam et al., 2020), have been studied in detail. The NSP1 Δ79-89 was shown to be associated with lower IFN-β levels and non-severe phenotypes (Lin et al., 2021). Our analysis presented here expands on these examples and provides an overview of the dynamics of in-frame indels in the evolution of the SARS-CoV-2 genome. Regions with recurrent indels called recurrent deletion regions (RDRs) and recurrent insertion regions (RIRs) in the N-terminal domain (NTD) of the spike were shown to play a role in immune escape (McCarthy et al., 2021). Here we use the term hypervariable regions (HVRs) to refer to indel-prone regions. These concentrations of indels provide an example of a new paradigm of the effects of indels on viral genomes and proteins—instead of loss-of-function they modify it by remodeling protein surfaces, affecting major antibody epitopes (Cai et al., 2021) and, possibly, protein-protein interaction networks.

METHODS

SARS-CoV-2 Sequencing Data Collection

We retrieved multiple sequence alignment (MSA) and metadata of complete SARS-CoV-2 genomes (6,143,793) from GISAID (https://www.gisaid.org/) as of January 7th, 2022. Briefly, full alignment (msa_0106.fasta) provided by GISAID was based on 6,716,124 submissions to GISAID EpiCoV. GISAID pipeline excludes duplicate, low-quality sequences (>5% N content) and incomplete sequences (length <29,000 bp). Then, the GISAID pipeline used this cleaned data to create the MSA file of 6,143,793 sequences using MAFFT (Katoh and Standley, 2013) with hCoV-19/Wuhan/WIV04/2019 (EPI_ISL_402,124; GenBank: MN996527) used as reference (Zhou et al., 2020).

Identification of Indels

We used an in-house Perl script to identify variations in each genome based on the GISAID MSA file as of January 7th, 2022. Additionally, on top of GISAID's cutoffs for excluding low-quality genomes with high N content (0.05), we applied additional filtering to avoid spurious indels and indels with shifted positions arising from high N content. Moreover, genomes with more than 200 mutations were excluded, resulting in 4,976,200 SARS-CoV-2 genomes used in the downstream analysis in this study. Additionally, to avoid reporting spurious indels arising from sequencing errors or errors in MSA, we generated another MSA file with no gaps in reference (obtained with *keep reference length* option) (Katoh and Standley, 2013) to confirm the exact positions of all the deletions discussed in this study. Then, for visualizing and confirming the position of the indels we used the MSA file based on a representative genome for each of the indels with 0 N content.

Assessing Differences in the Rate of Indels Between SARS-CoV-2 Proteins

We adopted the method we recently used to identify significantly under-mutated and over-mutated proteins during SARS-CoV-2 evolution (Jaroszewski et al., 2021) to identify proteins with a high rate of indels. Briefly, we counted the total number of indels (except single residue deletions which are usually regarded as unreliable) for each protein (except NSP11, ORF3b, ORF9b and ORF14 as these are too short for the significance analysis). We then used a two-sided binomial test to compare the rate of indels in each protein to the rate of indels in the background (all proteins) to identify proteins with high rates of indels. Our previous study (Jaroszewski et al., 2021) showed that ORF1ab is less frequently mutated and is likely under more stringent purifying selection than the genes coding for structural and accessory proteins (ORFs2-10). Therefore, we applied an additional statistical comparison of indel rates to non-structural proteins to identify NSPs (NSP1- NSP16) with a higher rate of indels than others. We performed a separate two-sided binomial test using only ORF1ab (corresponding to proteins NSP1-NSP16) for this specific comparison as background. Adjusted *p*-values (q-values) were calculated using the false discovery rate (FDR) method. Proteins with odds ratio above one and q-values less than 0.01 were considered as having significantly increased rates of indels.

Visualization of Indels on Proteins' 3-Dimensional (3D) Structures

We used PyMol (PyMOL, 2021) and Coronavirus3D (Sedova et al., 2020) for studying and visualization of indels in the context of protein 3-dimensional (3D) structures. The 3D coordinates were downloaded from the Protein Data Bank (PDB) (Berman et al., 2000). For proteins with no available 3D structures we used, if available, models predicted by Alphafold (https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19), or homology modeling (https://zhanglab.dcmb.med.umich.edu/COVID-19/), noting in the discussion their hypothetical status. It should be noted that even for some proteins with available 3D structures we used models predicted with homology modeling when the indels were located in the regions of the protein with unresolved structures (unmodeled residues). Information on protein domain boundaries was based on 3D coordinates when available or on UniProt and the literature (**Supplementary Table S4**).

The positions of transmembrane helices for proteins with no available 3D structures were identified with the TMHMM 2.0 algorithm (Krogh et al., 2001). IEDB server (Bepipred Linear Epitope Prediction 2.0 at http://www.iedb.org/) (Jespersen et al., 2017) was used to predict B-cell epitopes for NSP1, NSP3, NSP6, spike, nucleocapsid, ORF3a, ORF7a, and ORF8 (i.e. proteins with significantly increased rates of indels).

Visualization of Indels on the Phylogenetic Tree

We mapped the number of indels for each genome (between one and six indels) on the Nextstrain time-resolved tree (Hadfield et al., 2018), which includes 3475 genomes sampled between December 2019 and Dec 27th, 2021. We used the ggtree R package (Yu, 2020) to visualize the tree.

Visualization of Indels on the Alignment File

We extracted one representative genome for each of the indels discussed in this study (i.e., the indels most frequently observed in SARS-CoV-2 genomes). These genomes were then used to visualize the indels using R packages ggmsa and Biostrings.

Analysis of Independent Occurrence of Indels in SARS-CoV-2

The independent acquisition of indels was determined using HomoplasyFinder (Crispell et al., 2019) with the same filtering criteria as used in the previous studies (van Dorp et al., 2020). To identify potential recurrent indels (independently acquired in different branches of phylogenetic tree) in SARS-CoV-2 genomes, we used the GISAID global tree that includes 4,701,022 SARS-CoV-2 genomes (GISAID as of January 7th, 2022) (Shu and McCauley, 2017) together with the input variant calling file (VCF). Briefly, HomoplasyFinder calculates the consistency index for each indel by dividing the minimum number of changes on the GISAID tree (MNCT) by the number of different indels observed at that site minus one. The most frequent indels (observed in at least 0.01% of all studied genomes) with a consistency index of <1 and MNCT >30 were reported as potentially recurrent indels if they were also independently acquired in more than two independent GISAID clades and in at least two PANGO lineages when their immediate ancestor didn't carry this indel, two-time points and two different continents (Originating lab). These filtering and stringent cutoffs were applied to address issues arising from mixed quality of assembled genomes, which in some cases are not detectable (e.g., assembly pipelines replace missing nucleotides with data from the reference genome) from the genome analysis alone. The quality issues introduce uncertainty in phylogenies, lineage assignments and underestimation of indels frequencies all lead to overestimation of independent occurrence of indels (De Maio et al., 2020; Turakhia et al., 2020; Tang et al., 2021), which we countered by increasing the cutoff thresholds. Regions with different recurrent indels which occurred in adjacent residues (up to five residues apart) were called hypervariable regions (HVRs). The HVRs observed in this study contain between 2 and 30 residues.

To calculate the recurrence of each indel as the function of time of sample collection, geographical location (originating lab), PANGO lineages, and GISAID clades, we grouped genomes into 25-time bins based on the month and year of the data collection, into six geographical locations (continents), 12 clade-based groups (G, GH, GK, GR, GRA, GRY, GV, L, O, S, V, and a non-assigned group), and 1544 different PANGO lineages. We used such relatively large groups to reduce noise arising from the difference between individual labs and from low-quality genomes.

Statistical Analysis of Co-Occurred Indels in SARS-CoV-2 Genomes

We ran cooccur R package to analyze the co-occurrence of indels in each lineage and all genomes and used ggplot2 R package (Wickham, 2011) to draw heatmap of correlation matrix. We also calculated Spearman's correlation coefficient and *p*-value of the correlation test for every two indels using hmisc (Harrell and Harrell, 2019) R package. We further checked the independent acquisition of top correlated/co-occurred indels using HomoplasyFinder (Crispell et al., 2019) based on the method explained earlier. The input VCF file includes information on the presence/absence of two co-occurred indels. We used ComplexHeatmap R package (Gu et al., 2016) to draw the heatmap of percentage of top indels in SARS-CoV-2 VOCs.

Comparing SARS-CoV-2 and SARS-CoV Genomes in Terms of Indels

Spike, NSP1, NSP3, NSP6, N, ORFs 3a, 7a, and eight protein sequences of SARS coronavirus Tor2 (NC_004,718.3) and SARS-CoV-2 (MN996527) were aligned using MAFFT (Katoh and Standley, 2013) (default parameters). We used Jalview (Waterhouse et al., 2009) to visualize alignment files and obtain the count and positions of indels.

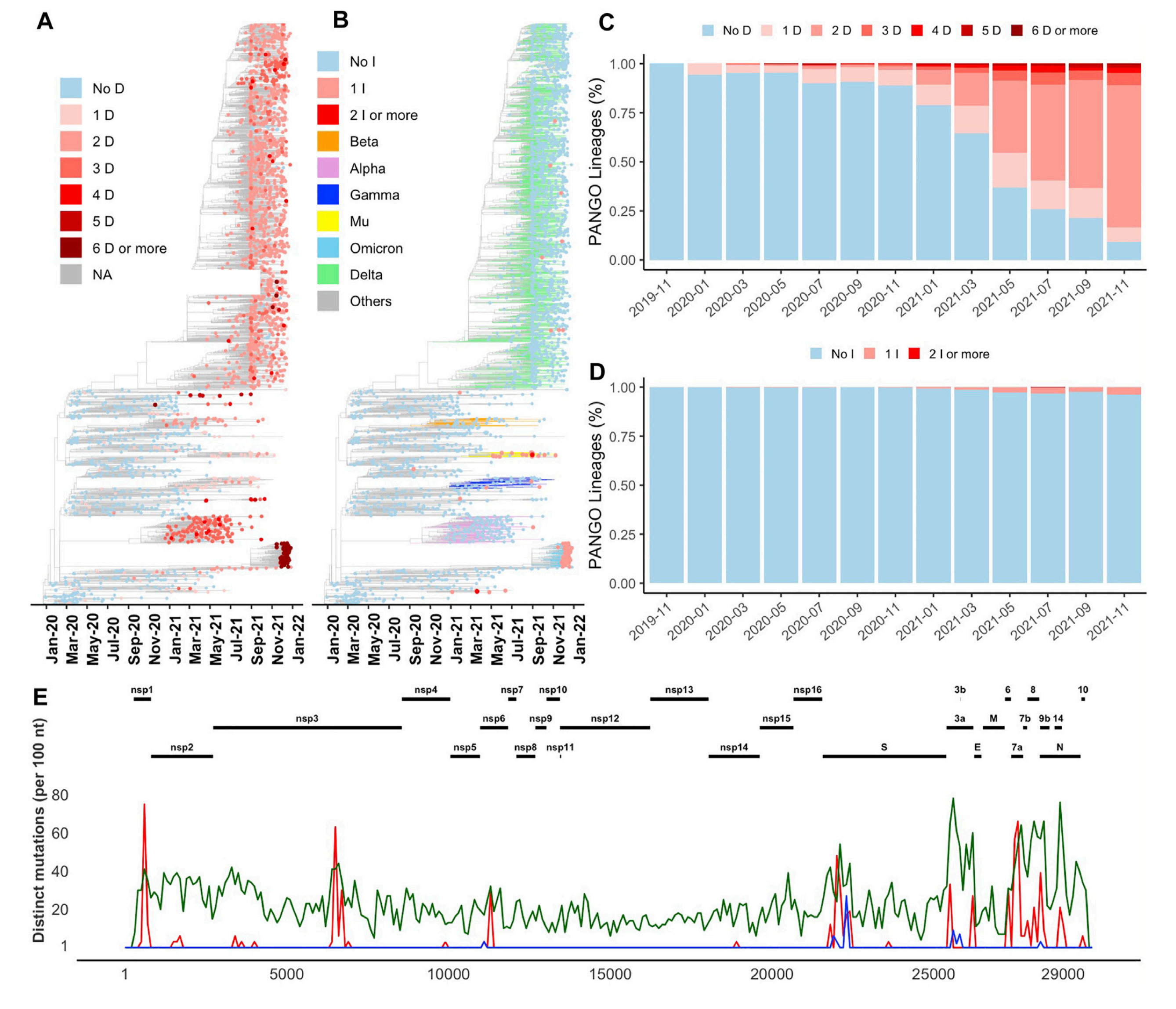


FIGURE 1 | Distribution of indels in SARS-CoV-2 genomes **(A)** and **(B)** Increase in the number of deletion **(D)** and insertion **(I)** events in newly emerged lineages illustrated on Nextstrain's time-resolved phylogenetic tree, respectively **(C)** and **(D)** Percentage of PANGO lineages with and without deletion and insertion events over time, respectively **(E)** Distribution of the most common deletions along the SARS-CoV-2 genome (red) compared to insertions (blue) and missense substitutions (green).

RESULTS

Increased Frequency of In-Frame Indels in Emerging SARS-CoV-2 Lineages

The recent increase in the number of indels (both insertions and deletions) was observed in all branches of the phylogenetic tree (**Figures 1A,B**). This increase can be seen in the percentage of both SARS-CoV-2 lineages (**Figures 1C,D**) and genomes (**Supplementary Figures S1A,B**) with at least one deletion or one insertion event (one or more than one amino acid change) growing in time. Indels were acquired in several VOCs such as Alpha (B.1.1.7 and Q.*), Beta (B.1.351*) and Omicron (B.1.1.529 and BA.*), Gamma (*P.**), and Delta (B.1.617.2, AY.*). As an example, Alpha variant (B.1.1.7) is defined by 17 signature genome modifications, including three deletions events (NSP6 Δ106–108, S Δ69-70, and S Δ144), while Omicron variant includes seven indels as shown in **Supplementary Figure S1C** (NSP3 1265:SL>I, NSP6 Δ105-107, nucleocapsid Δ31–33, S Δ69–70, S 142:GVYY>D, S 211:NL>I, and S 214: R>REPE). Additional indels and their combinations are found in other variants (**Supplementary Figure S1C**). In this study, for simplicity, genome modifications that include both indels and substitutions such as S 142:GVYY>D are only referred to as indels.

TABLE 1 | Comparison of frequencies of in-frame indels (indels) in SARS-CoV-2 proteins using the two-sided binomial test (only indels observed in at least two genomes were included to eliminate spurious mutations). Bold font indicates proteins with a significantly increased rate of indels (q-value<0.01 and Odds ratio>1).

Protein	Protein length	Number of indels	All Proteins as Background		ORF1ab as Background	
			Odds Ratio	q-value (FDR adjusted *p*-value)	Odds Ratio	q-value (FDR adjusted *p*-value)
NSP1	540	109	2.14	1.85E-12	4.40	1.44E-36
NSP2	1914	81	0.45	5.89E-17	0.92	5.02E-01
NSP3	**5835**	**442**	**0.80**	**1.78E-07**	**1.65**	**3.96E-32**
NSP4	1500	40	0.28	5.58E-24	0.58	2.93E-04
NSP5	918	6	0.07	3.74E-29	0.14	9.58E-12
NSP6	**870**	**58**	**0.71**	**6.47E-03**	**1.45**	**8.99E-03**
NSP7	249	5	0.21	1.18E-05	0.44	6.23E-02
NSP8	594	9	0.16	1.86E-14	0.33	1.65E-04
NSP9	339	7	0.22	2.31E-07	0.45	3.54E-02
NSP10	417	8	0.20	4.37E-09	0.42	1.08E-02
NSP12	2795	46	0.17	2.91E-64	0.36	5.63E-18
NSP13	1803	15	0.09	3.32E-54	0.18	2.65E-20
NSP14	1581	91	0.61	2.58E-07	1.26	3.54E-02
NSP15	1038	36	0.37	1.05E-12	0.76	9.92E-02
NSP16	894	23	0.27	6.51E-15	0.56	5.01E-03
Spike	**3822**	**459**	**1.27**	**1.22E-07**	**-**	**-**
E	228	18	0.84	5.16E-01	-	-
M	669	26	0.41	2.06E-07	-	-
N	**1260**	**159**	**1.34**	**3.43E-04**	-	-
ORF10	**117**	**7**	**0.63**	**3.01E-01**	-	-
ORF3a	**828**	**254**	**3.25**	**1.91E-57**	-	-
ORF6	**186**	**61**	**3.47**	**9.43E-16**	-	-
ORF7a	**366**	**595**	**17.22**	**0.00E+00**	-	-
ORF7b	**132**	**58**	**4.65**	**1.56E-20**	-	-

Indels are concentrated on protein surfaces near epitope regions

Most indels are significantly (q-value < 0.01 and odds ratio >1) concentrated in NSP1, NSP3, NSP6, ORF3a, ORF6, ORF7a, ORF7b, ORF8, nucleocapsid, and spike glycoprotein (**Figure 1E** and **Table 1**), all of which are involved in interactions with the host immune system (Lei et al., 2020; Liang et al., 2021; Smith et al., 2021). At the same time, proteins involved in the replication–transcription complex show very few or no indels (**Figure 1E** and **Table 1**). It is in agreement with our earlier report showing the segment of the genome coding for the non-structural proteins (Orf1ab, corresponding to proteins nsp1-nsp16) is significantly under-mutated for both missense and synonymous mutations (Jaroszewski et al., 2021). It should be noted that terms recurrent deletion regions (RDRs) and recurrent insertion regions (RIRs) are used in recent literature, indicating regions of SARS-CoV-2 proteins with frequent recurrent deletions and insertions, respectively. In this paper, we use the term "hypervariable regions (HVR)" referring to regions of proteins with frequent recurrent indels.

Aggregation and recurrence of indels in hypervariable regions of SARS-CoV-2 proteins are determined by an interplay of the protein structural constraints and functional role of specific regions. Most of the HVRs of SARS-CoV-2 proteins (except ORF7a-HVR) are found on or adjacent to loops forming either experimentally identified (Liang et al., 2021; Smith et al., 2021) or predicted antibody epitopes (**Figures 2**,**3**), suggesting SARS-CoV-2 is optimizing its interactions with the host immune system, possibly in response to the increased immunity of the population. For instance, NSP6-HVR falls on a predicted T-cells (Smith et al., 2021) and B-cells epitope (per IEDB server), forming a short loop between two transmembrane helices (**Figure 2C**). Similarly, NSP1-HVR1 and spike-HVRs (**Figure 2**), as well as HVRs in other proteins are in or near the loop forming epitope regions (**Figure 3**).

In the most studied SARS-CoV-2 protein, surface glycoprotein S (spike), NTD is one of the most genetically modified regions of spike protein and of the entire SARS-CoV-2 proteome (see **Figure 1**). Deletions in the NTD could classified as belonging to recurrent deletion regions: RDR1 (residues 60–75), RDR2 (residues 139–146), RDR3 (residues 210–213), and RDR4 (residues 242–248) (McCarthy et al., 2021). Recurrent insertions were also reported in the same regions (Gerdol, 2021). We observed that indels in NTD-HVR1 and HVR2 are more frequent as compared to HVR3 and HVR4 (**Supplementary Figure S2A,B**). Several lineages with new spike indels (expanding spike-HVR2 and HVR4) are now emerging (**Supplementary Figure S2A,B**). Comparison of spike proteins from the SARS-CoV (Tor2) and SARS-CoV-2 (one of the early Wuhan reference) viruses indicates 22 amino acid (AA) insertions and four AA deletions in SARS-CoV-2 spike protein compared to SARS-CoV that mainly occurred in NTD (**Supplementary Figure S2C**), confirming that NTD is

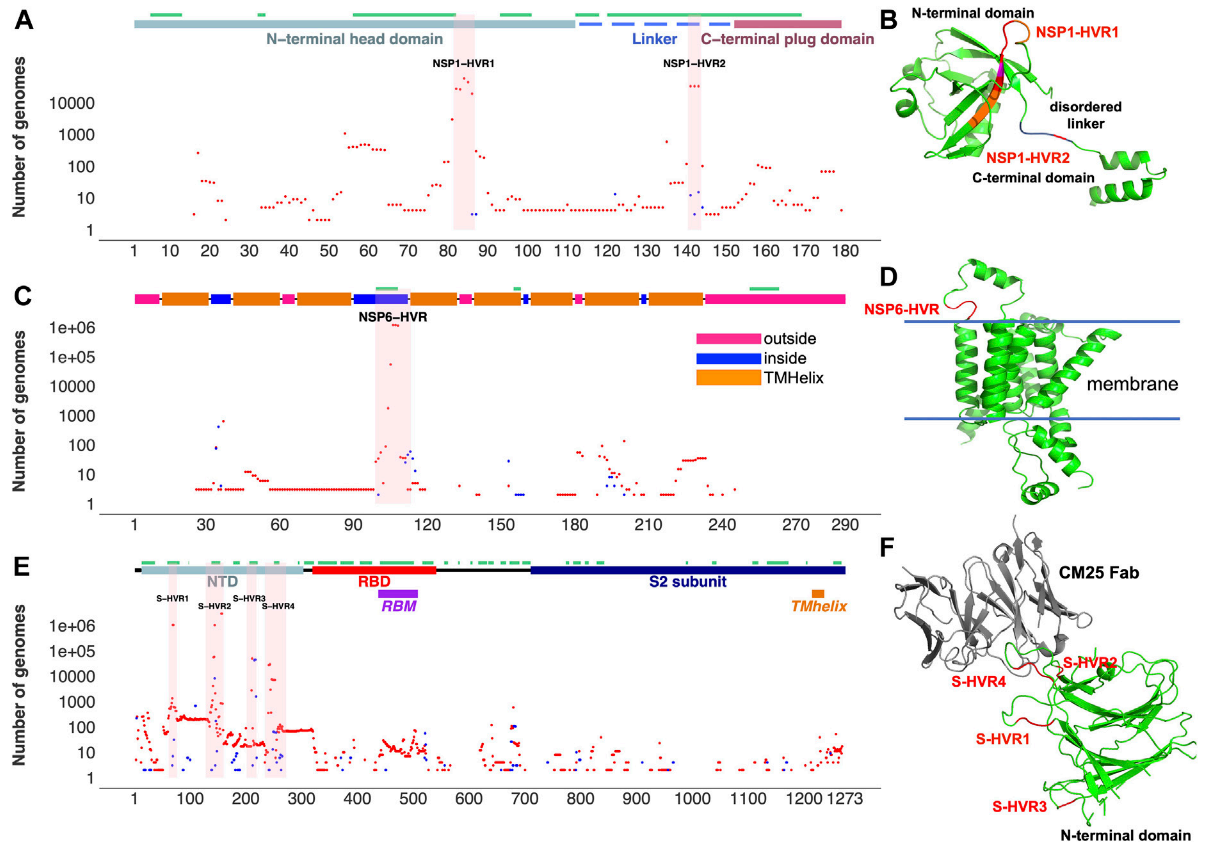


FIGURE 2 | Top SARS-CoV-2 HVRs in the context of protein 3D structures. **(A)** Distribution of indels in NSP1 **(B)** NSP1-HVRs on protein 3D structure **(C)** Distribution of indels in NSP6 **(D)** NSP6-HVR on protein 3D structure **(E)** Distribution of indels in spike glycoprotein **(F)** HVRs on the protein 3D structure of the spike glycoprotein N-terminal domain bound to human Fab CM25. Insertions, deletions, and predicted B-cell epitopes (result from the IEDB server at www.iedb.org) are represented as blue dots, red dots, and green lines, respectively. **Supplementary Table S3** provides details of structures/models used in the Figure.

generally the most indel-prone region of spike in SARS coronaviruses.

NSP3 HVR corresponds to group 2 specific marker domain (G2M), a structurally uncharacterized region of the protein (**Figures 3A,B**). Based on the NSP3 predicted model (built using D-I-TASSER/ C-I-TASSER pipeline from the Zhang lab, https://zhanggroup.org/), NSP3-HVR is in the loop and indels in this region occur near B-cell epitopes predicted using IEDB server (**Figure 3B**). Similar observations were also made for nucleocapsid protein (**Figure 3 C,D**), ORF3a (**Figure 3 E,F**), and ORF8's HVR (**Figures 3I,J**). The indels in different protein HVRs occurred independently in several lineages (**Figure 4** and **Supplementary Table S1**) as seen on the SARS-CoV-2 phylogenetic tree (Elbe and Buckland-Merrett, 2017). In the following, we will discuss in detail the independent acquisition of indels in NSP1, NSP6 and NTD of spike protein HVRs. Independently acquired indels in NSP3, ORF3a, ORF7a, and ORF8 as well as in nucleocapsid protein HVRs will be discussed in separate sections.

The independent acquisition of indels was determined using HomoplasyFinder (Crispell et al., 2019) with filtering criteria as applied in the previous study (van Dorp et al., 2020). Indels with minimum number of changes on tree (MNCT) above 30 were considered as potential recurrent deletions. We then applied additional filters (see above) and only included those that fulfilled all the criteria (**Supplementary Table S1**). These stringent cutoffs were applied to avoid overestimation of homoplasies due to sequencing errors (De Maio et al., 2020).

Two mutually exclusive NSP1 HVRs (e.g., NSP1 Δ84 and NSP1 Δ85 in NSP1-HVR1 and Δ141-143 in NSP1-HVR2) emerged independently in several lineages such as Alpha, Beta, Delta, Gamma and Omicron (**Figures 4 A, B**). A long version of the indel in NSP1-HVR1 (Δ79-89) was studied before (Lin et al., 2021), but our analysis indicates that shorter indels in this region are recurring more frequently (**Figure 5A**). The results from HomoplasyFinder (consistency index or CI) indicate that NSP1 deletions are among the potential recurrent events in SARS-CoV-2 evolution (**Figure 4B** and **Supplementary Table S1**). NSP1 (Δ79-89) was reported to induce lower IFN-I response in the infected Calu-3 cells (Lin et al., 2021), highlighting the biological importance of indels in NSP1 and other non-spike proteins. It should be noted that NSP1 deletions are not among signature genomic modifications of any SARS-CoV-2 lineage and no indel event differences were identified between NSP1 proteins of SARS-

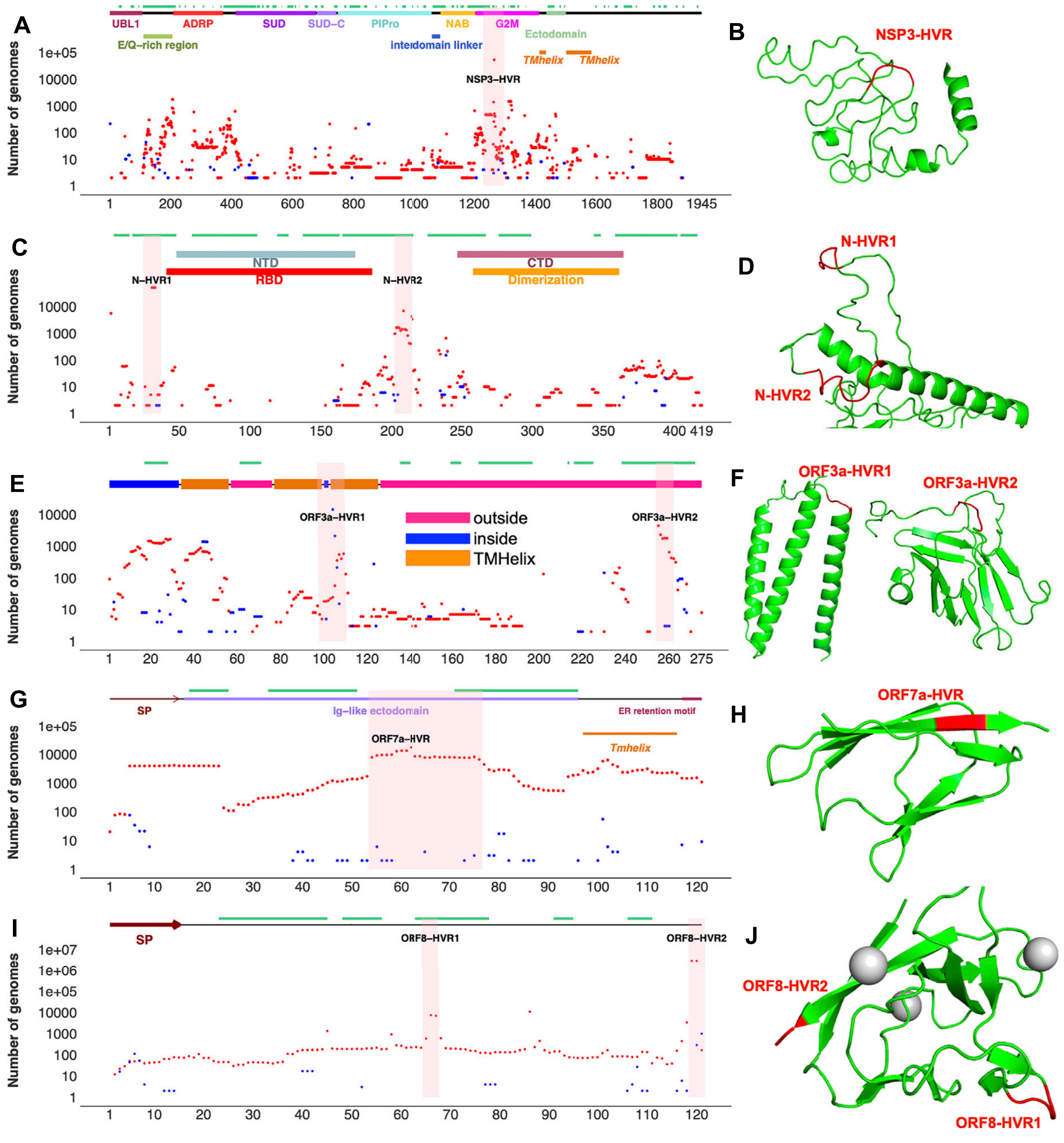
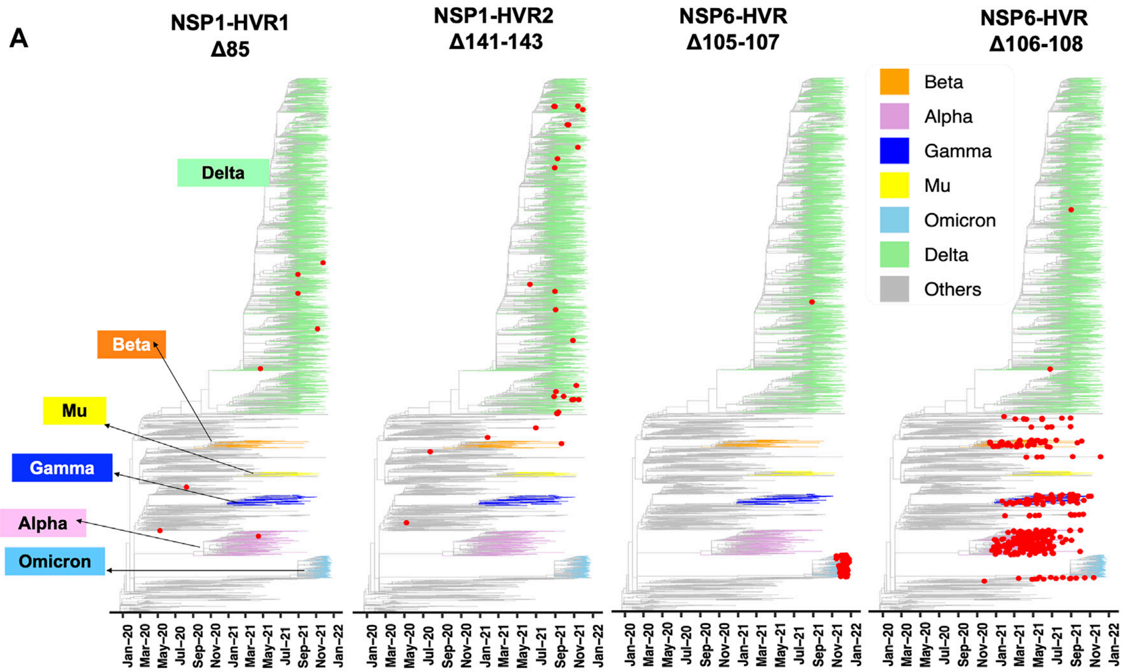


FIGURE 3 | Top SARS-CoV-2 HVRs in the context of protein 3D structures **(A)** Distribution of indels in SARS-CoV-2 non-structural protein 3 (NSP3) **(B)** NSP3 recurrent deletion region (HVR) on protein 3D structure **(C)** Distribution of indels in SARS-CoV-2 nucleocapsid (N) protein **(D)** N-HVRs on protein 3D structure **(E)** Distribution of indels in SARS-CoV-2 ORF3a **(F)** ORF3a-HVRs on protein 3D structure **(G)** Distribution of indels in SARS-CoV-2 ORF7a **(H)** ORF7-HVR on protein 3D structure **(I)** Distribution of indels in SARS-CoV-2 ORF8 **(J)** ORF8-HVRs on protein 3D structure. Deletions, insertions, and epitopes are represented as red dots, blue dots, and green lines, respectively. Pink highlighted regions represent HVRs or potential hotspots for recurrent indels in each protein. The regions of 3D structure corresponding to HVRs are colored in red. The coordinates of proteins were obtained from different sources (see **Supplementary Table S3**). Predicted 3D structural models https://zhanglab.ccmb.med.umich.edu/COVID-19/ were used for visualization of recurrent deletion regions in NSP3, ORF3a, and nucleocapsid protein. SP: signal peptide. Indels independently occur in several SARS-CoV-2 lineages in hypervariable regions.

CoV (Tor2) and SARS-CoV-2 (**Supplementary Figure S3A**). This might imply that intact NSP1 is key for the full functionality of the virus and its pathogenicity but at the same time recurrent indels could suggest the presence of intra-host variations and quasispecies (Santacroce et al., 2021).

After the spike-HVRs, the NSP6-HVR (residues 99–108) is the second most frequently modified HVR in SARS-CoV-2, with the Δ106-108 observed in more than 1M genomes as of January 2022 (**Figure 5A**). NSP6 deletions independently occurred as a signature modification for several



B

indels	Top variants (in at least 10 genomes)	CI	MNCT
NSP1: Δ82-86	Alpha, Beta, Delta, Gamma, Lambda, Omicron, Others	1.74E-04	13112
NSP1: Δ84	Alpha, Beta, Delta, Gamma, Lambda, Omicron, Others	5.78E-05	17288
NSP1: Δ85	Alpha, Beta, Delta, Gamma, Omicron, Others	2.21E-04	4526
NSP1: Δ141-143	Alpha, Beta, Delta, Gamma, Lambda, Omicron, Others	9.79E-05	10210
NSP6: Δ106-108	Alpha, Beta, Delta, Gamma, Lambda, Omicron Others	3.80E-05	26304
NSP6: Δ105-107	Alpha, Delta, Gamma, Lambda, Omicron Others	2.72E-03	368

FIGURE 4 | Recurrent Indels in NSP1 and NSP6. **(A)** Nextstrain time-resolved tree, which includes 3475 genomes sampled between December 2019 and Dec 27th, 2021) displays the presence and distribution of the most frequent deletions positioned on NSP1-HVRs and NSP6-HVR as red dots **(B)** Top SARS-CoV-2 variants harbor the most frequent and potentially recurrent deletions of NSP1 and NSP6. Minimum Number of Changes on Tree (MNCT) and Consistency Index (CI) calculated using HomoplasyFinder based on GISAID global tree (4,701,022 SARS-CoV-2 genomes as of January 7th, 2022).

VOCs—Alpha, Beta, Gamma, and Omicron but also some other lineages such as B.1.525 in Nigeria and Europe and B.1.526 in New York and Europe (**Figure 4** and **Supplementary Table S2**). Signatures of positive selection for NSP6 Δ106-108 were recently reported (Martin et al., 2021) in line with our results showing high recurrence of NSP6 deletions (**Figure 4B**). In addition to recurrent indels, overlapping indel events identified in NSP1 (**Figure 5A**), NSP6 (**Figure 5B**), and Spike NTD (**Supplementary Figure S2**) could provide additional evidence of convergent and/or parallel adaptive evolution in SARS-CoV-2 genomes. This may also offer more potential genetic routes for the rapid adaptation, immune escape and drug resistance of SARS-CoV-2. Similar evolutionary routes in HIV-1 and other RNA viruses were found to play pivotal role in drug and neutralizing antibody resistance (Menéndez-Arias et al., 2006; Gutierrez et al., 2019).

We observe an increasing number of genomes with two or more different indels in spike or other proteins. We use the term co-occurred indels for indels that appear simultaneously in at least one SARS-CoV-2 genome, and independent acquisition of top co-occurred indels was determined using HomoplasyFinder (see method section for details). Multiple spike-indels independently co-occurred with each other and with indels in other proteins, especially NSP6-indels (**Figure 6** and **Supplementary Table S2**). NSP6-indels independently co-occurred with spike indels located in HVR1 and HVR2 in Alpha (B.1.1.7, Q.*) and B.1.525, with indels located in HVR2 in B.1.526.1 and B.1.1.318, with indels in HVR4 in Beta (B.1.351*) and with several indels in HVR2 and HVR3 in Omicron (B.1.1.529 and BA.*) as shown in **Figure 5** and **Supplementary Table S2**. Based on HomoplasyFinder results, indels in the spike NTD and ORF8 are also among the top co-occurred indels. Spike Δ157-158 and ORF8 Δ119-

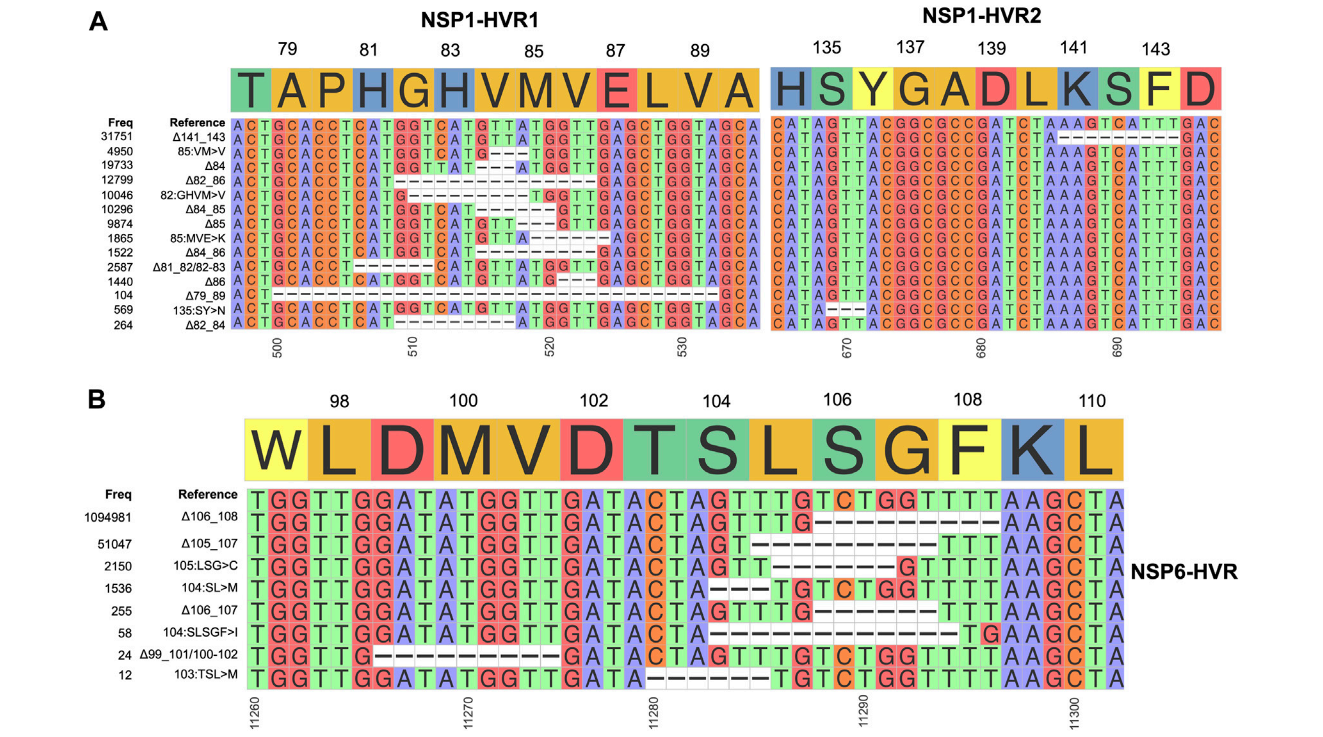


FIGURE 5 | Hypervariable regions (HVRs) of NSP1 and NSP6 **(A)** and **(B)** represent coordinates of HVRs of NSP1 and NSP6, respectively. The number of genomes containing a specific indel is provided on the left side of each plot. Indels independently co-occur in several SARS-CoV-2 lineages.

120 were found in more than 90% of the genomes assigned to Delta variant and their co-occurrences were also recorded in genomes assigned to other lineages such as Omicron and B.1.485 (**Figure 6B** and **Supplementary Table S2**).

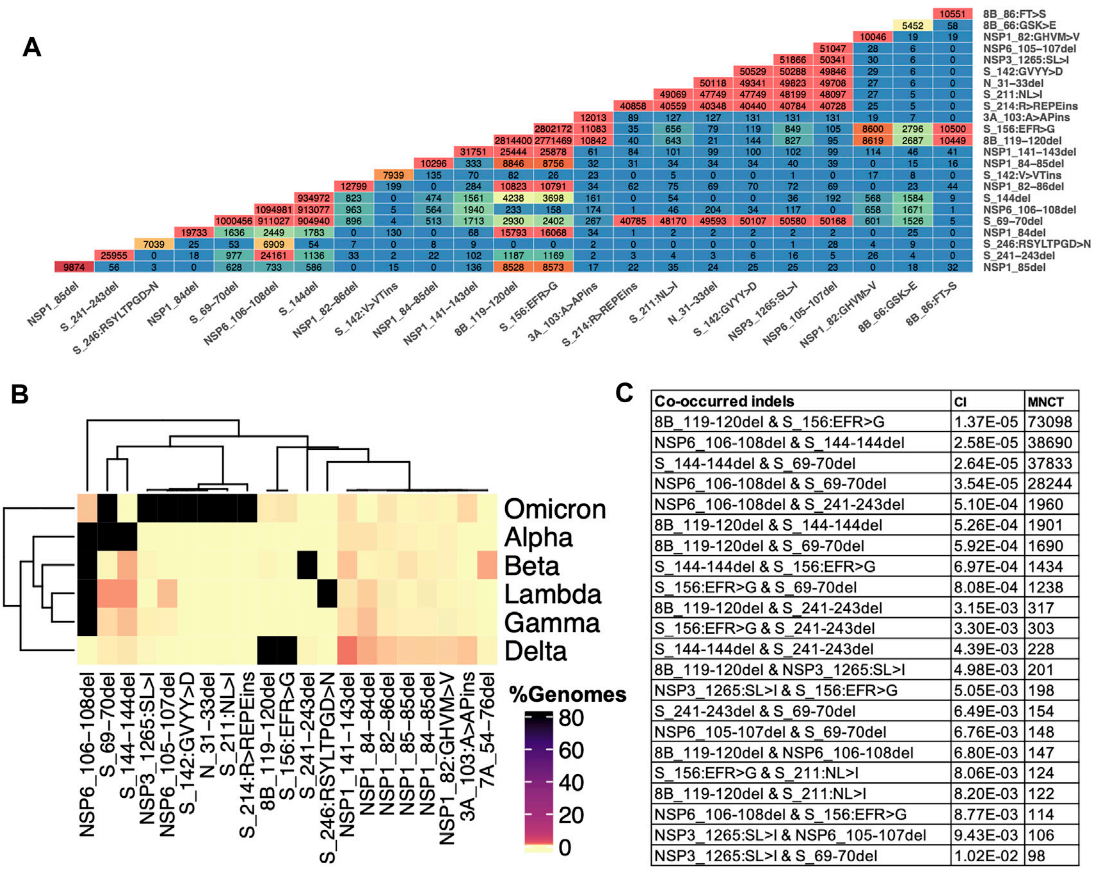
Hypervariable Region in SARS-CoV-2 Non-Structural Protein Three NSP3

NSP3 along with NSP1 and NSP6 has significantly higher number of indels when compared to the rest of NSPs (**Table 1**). As shown in **Figure 3**, indels in NSP3 are largely occurring in the loop region (1235–1270) and near epitopes (Smith et al., 2021). NSP3 deletion 1265:SL>I is a signature mutation of Omicron variant, NSP3 Δ1237-1251 was observed in L.1 PANGO lineage in Canada and NSP3 Δ1263 in B.1.1.298 variant from Denmark where the latter co-occur with NSP1 85:VM>V and spike Δ69-70. NSP3-indels are often mutually exclusive with indels in other proteins - they only co-occurred with spike and NSP6-indels in Omicron and very few genomes assigned to B.1.1.7 lineage (**Supplementary Table S3**). When compared to NSP3 of SARS-CoV, SARS-CoV-2 NSP3 had a total of 30 AA insertions and seven AA deletions which occurred mostly between residues 100–400 (**Supplementary Figure S4**) correspond to predicted epitopes (**Supplementary Figure S2).**

Although NSP2 was not identified as a significantly indel-prone protein, some indels in the NSP2 appeared independently in several lineages (**Supplementary Tables S1, S2**). The NSP2 Δ265-266 is the signature modification of the B.1.573, B.1.1.191, and AN.1 PANGO lineages (**Supplementary Table S2**), primarily seen in Canada and Denmark samples. The NSP2 Δ268 is mainly occurring in viral genomes collected from England, Scotland, Northern Ireland, and the Netherlands, and it is also the signature mutation of several lineages (**Supplementary Table S2**). The NSP2 Δ267-268 frequently appeared during the early phase of the pandemic and only a small portion of the recently collected genomes harbored other NSP2-indels positioned on NSP2-HVR (residues 260–270). NSP2 was shown to disrupt host signaling, and it might play a role in SARS-CoV-2 pathogenicity. However, more investigation is required to elucidate the role of NSP2 protein and the impact of its indels on immune evasion.

Recurrent Deletion Regions in SARS-CoV-2 Nucleocapsid and Accessory Proteins ORF3a, ORF7a, and ORF8

Indels of the nucleocapsid protein occur in two potential HVRs (HVR1: clusters around residues 28–35 and HVR2: clusters around residues 202–214) as shown in **Figures 3C,D**. Both nucleocapsid HVRs specially HVR-2 are close to experimentally identified epitopes such as 36-RSKQR-40 and 206-SPARM-210 (Liang et al., 2021; Smith et al., 2021). After Omicron signature deletion (Δ31-33 at HVR1) the second most frequent deletion in nucleocapsid protein, 208AR>G (HVR2), is a signature of B.1.1.318 and is found in some B.1.1.7 genomes (**Supplementary Table S2**). It co-occurred with three other indels



Co-occurred indels	CI	MNCT
8B_119-120del & S_156:EFR>G	1.37E-05	73098
NSP6_106-108del & S_144-144del	2.58E-05	38690
S_144-144del & S_69-70del	2.64E-05	37833
NSP6_106-108del & S_69-70del	3.54E-05	28244
NSP6_106-108del & S_241-243del	5.10E-04	1960
8B_119-120del & S_144-144del	5.26E-04	1901
8B_119-120del & S_69-70del	5.92E-04	1690
S_144-144del & S_156:EFR>G	6.97E-04	1434
S_156:EFR>G & S_69-70del	8.08E-04	1238
8B_119-120del & S_241-243del	3.15E-03	317
S_156:EFR>G & S_241-243del	3.30E-03	303
S_144-144del & S_241-243del	4.39E-03	228
8B_119-120del & NSP3_1265:SL>I	4.98E-03	201
NSP3_1265:SL>I & S_156:EFR>G	5.05E-03	198
S_241-243del & S_69-70del	6.49E-03	154
NSP6_105-107del & S_69-70del	6.76E-03	148
8B_119-120del & NSP6_106-108del	6.80E-03	147
S_156:EFR>G & S_211:NL>I	8.06E-03	124
8B_119-120del & S_211:NL>I	8.20E-03	122
NSP6_106-108del & S_156:EFR>G	8.77E-03	114
NSP3_1265:SL>I & NSP6_105-107del	9.43E-03	106
NSP3_1265:SL>I & S_69-70del	1.02E-02	98

FIGURE 6 | Indels and their co-occurrence in SARS-CoV-2. **(A)** Co-occurrence of top frequent indels **(B)** Co-occurrence of top indels in VOCs. Data for these heatmaps is provided in **Supplementary Table 2** which includes additional combinations of indels in lineages harboring them **(C)** Independent co-occurrence of indels determined based on minimum number of changes on tree (MNCT) and consistency index (CI) calculated using HomoplasyFinder based on GISAID global tree (4,701,022 SARS-CoV-2 genomes as of January 7th, 2022).

in B.1.1.318, including NSP6 Δ106-108, spike Δ144, and ORF7b 44:TNMKF>Y. According to the Coronavirus3D (Sedova et al., 2020) variant tracker, this lineage was among the top growing lineages in several countries such as the United States, United Kingdom, and France in June 2021.

The most recurrent indels of ORF3a cluster around amino acid positions 103 (ORF3a-HVR1) and 255 (ORF3a-HVR2) as shown in **Figures 3E,F**. ORF3a-HVRs are located in the structurally unresolved region of the protein. Based on the predicted structures, they correspond to loops which also contain predicted B-cell epitopes. Interestingly ORF3a-HVRs identified in our study are also near experimentally identified epitopes of ORF3a antibodies such as 100-GLEAPFLYLYALVYF-114 (Smith et al., 2021), 266-EPTTTTSVPL-275, 246-IHTID-250, and 266-EPTTTTSVPL-275 (Liang et al., 2021).

The only insertion (240P>PE) in ORF3a SARS-CoV-2, when compared to SARS-CoV is located near ORF3a-HVR2 (**Supplementary Figure 3D** and **Figures 3E,F**). Despite recurring in several lineages, ORF3a indels are not signature mutations for any lineages or sub-lineages. ORF3a Δ255 co-occurred with NSP6 and spike indels in Alpha variant (**Supplementary Table S2**).

Unlike the rest of SARS-CoV-2 proteins, accessory proteins (ORF7a and ORF8) have longer indels. The indels of ORF7a often happen in ORF7a-HVR encompassing residues 60–100 (**Figure 3 G,H**), near previously identified ORF7a epitopes such as 86-LFIRQEEVQELYSPI-100 (Liang et al., 2021). The most frequent indel in this region is 7A_62:QF>H co-occurred with NSP6 and spike indels in the Delta variant (**Supplementary Table S2**). ORF7a indels are not signature mutations of any SARS-CoV-2 lineage and protein is mostly conserved between SARS-CoV and SARS-CoV-2 when compared to ORF8 (8b) as shown in **Supplementary Figures S3E,F**.

The most recurrent and frequent indels of ORF8 is encompassing residues 63–66 (ORF8-HVR1) and 118–120 (ORF8-HVR2) as illustrated in **Figure 3 I,J**. and the latter is the signature mutation for the Delta variant and co-occurred with spike S_156:EFR>G (**Figure 6**). Interestingly, both ORF8 HVRs are near experimentally identified epitopes, including 66-GSKSP-70 and 106-EDFLE-110. The highest number of changes in terms of indels between SARS-CoV and SARS-CoV-2 proteins was recorded for ORF8 (8b) and spike proteins (**Supplementary Figure S3**), indicating they are rapidly evolving among SARS coronaviruses. Deletions of an entire ORF8 were identified during both early and late phases of SARS-CoV pandemic (2003) in China (Consortium, 2004).

Interestingly, most SARS-CoV-2 proteins have a high tendency for recurrent deletions (**Supplementary Table S1**), likely facilitating the virus adaptation to the human host. The increasing number of deletions also results in SARS-CoV-2 genome shrinkage over time, especially in the recent VOCs like Omicron (**Supplementary Figure S4**). Although the direct association of genome size with viral fitness is difficult to prove, there is evidence of replicative advantage associated with smaller genome size in RNA viruses (Tromas et al., 2014; Zwart et al., 2014; Walker et al., 2015). The results of this study should be interpreted within the context of limitations in the quality of SARS-CoV-2 genomes. Mixed quality of genomes and high numbers of Ns increases instability in lineage assignments and might underestimate indels and overestimate homoplasies. We accounted for this problem by using very stringent criteria and we hypothesize that the real extent of homoplasy in the SARS-CoV-2 evolution is likely to be even higher.

DISCUSSION

Viruses, and in particular RNA viruses, are known to undergo rapid genome modifications, but are rarely studied with frequency that would allow us to monitor their detailed dynamics. Comparison of genomes of separate species gives us only a summary of modifications that occurred over significant periods of time. The COVID-19 pandemic led to an unprecedented mobilization of the research community, which in turn provided a unique opportunity for real-time monitoring of a pathogenic virus during a pandemic. In this study, we used sequencing data provided by thousands of research groups and available in a GISAID database (Shu and McCauley, 2017) to study the dynamics of protein indels during the course of pandemic. This analysis revealed the increase in the rate of indels that started in late 2020, driven by the emergence of lineages containing deletions as signature genome modifications, such as Alpha and Beta variants which replaced most of the previous lineages without indels. These were in turn replaced by the Delta variant with even more deletions in its genome. The Omicron variant that appeared in November 2021 is the first VOC containing both insertions and deletions and it has currently replaced almost all previous variants. Some of the indels in these variants were already shown to increase immune invasion, lead to higher transmissibility and higher viral binding affinity (Karim and Karim, 2021; McCarthy et al., 2021; Viana et al., 2021), functions of others are still unknown, but we can speculate about them based on the co-occurrence and overlap with mutations at the same sites.

Different processes may contribute to the emergence of indels in viral genomes, such as replication slippage, recombination, and retrotransposition. Compared to recombination and retrotransposition, replication slippage generates short indels (Viguera et al., 2001; Domingo, 2020). Since our analysis revealed mainly short indels, we believe these indels are primarily the result of replication slippage. Another possible explanation for this hypothesis is that insertions emerged later in the pandemic consistent with a higher evolutionary cost for insertions than deletions due to higher probability of incidence of the slippage-induced deletions.

Regardless of the cause of their emergence, SARS-CoV-2 indels that were selected by evolution and contributed to the emerging lineages are predominantly found in specific regions of proteins known as hypervariable regions that typically correspond to loops in protein structures. Interestingly, not all loops in SARS-CoV-2 proteins were found to contain indels, those that do were close to either experimentally identified or predicted epitopes (Zhang et al., 2008; Liang et al., 2021; Smith et al., 2021) or were involved in protein-protein interactions, and in the case of the specific SARS-CoV-2 proteins with overabundant indels, in interactions with the host's immune system. Modeling and emerging experimental evidence (Cai et al., 2021) shows that deletions in such regions can remodel epitope surfaces, leading to immune escape. This parallels findings in HIV-1 where deletions in the spike glycoprotein regions encoding surface-exposed disordered loops were found to mediate escape from the neutralizing antibodies elicited by earlier variants of the virus (Wood et al., 2009; Palmer and Poon, 2019).

Many indel-prone regions such as the loops in the spike NTD overlap with mutation hotspots that are thought to be driven by host immune system pressure (Gerdol, 2021; McCallum et al., 2021; McCarthy et al., 2021). Therefore, we hypothesize that the emergence of indels in the same hotspots is a response to the same adaptive pressure. This is supported by the recent studies where both spike-NTD substitutions and indels were demonstrated to accelerate virus adaptation to the host and immune escape (Gerdol, 2021; McCallum et al., 2021; McCarthy et al., 2021).

Independent co-occurrence of indels in several VOCs might reflect signatures of adaptive evolution by recurrence or recombination. Several VOCs such as Alpha, Beta and Omicron which have simultaneous spike and NSP6-indels were found to have higher transmissibility, infectivity, or immune escape properties than the previously dominant lineages such as B.1.177 (Davies et al., 2021) with no indels. Such independent expansion of indels in multiple lineages and geographic locations suggests a common adaptation mechanism of SARS-CoV-2 genomes, probably to overcome host immune response, as also suggested in the recent literature (McCarthy et al., 2021; Ribes et al., 2021).

In conclusion, we conducted an in-depth analysis of indels in 4,976,200 SARS-CoV-2 genomes. We show that genomic modifications happen in a specific order, with deletions following point mutations, but growing quickly during the progress of the pandemic. In recent months we started seeing the emergence of insertions, including founder genomic modifications of the Omicron variant. Like mutations, indels are largely found in SARS-CoV-2 proteins involved in interactions with the host immune system but are preferentially located in specific regions of proteins "hypervariable regions" which overlap with structural features such as loops located close to epitopes. Indels in such regions might facilitate immune escape by remodeling the epitope surfaces and may prolong infection by these lineages. Such HVRs should be the subject of surveillance as much as common escape mutations. The increase in the number of indels and HVRs in recent lineages is likely a sign of the virus adapting to the increasing pool of resistant hosts, but other explanations, such as their role in regulating host antiviral response are also possible.

DATA AVAILABILITY STATEMENT

All sequences used in this study are accessible via the GISAID database (www.gsaid.org). All protein structures are accessible via Protein Data Bank (https://www.rcsb.org/) and models from the Zhang lab (https://zhanglab.dcmb.med.umich.edu/COVID-19/) and AlphaFold database of COVID-19 structures (https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19). All scripts are publicly available on GitHub repository (https://github.com/ArghavanAlisoltani/SARS-CoV-2-Indels.git).

AUTHOR CONTRIBUTIONS

AA, LJ, MI, AI, and AG designed the calculations; AA, LJ, and MI. performed them; AA, LJ, MI, AI, and AG analyzed data and AA, LJ, MI, and AG wrote the manuscript.

FUNDING

National Institute of Allergy and Infectious Diseases contract HHSN272201700060C (CSGID) and National Institute of General Medical Sciences Award GM118187 (to AG).

ACKNOWLEDGMENTS

We gratefully acknowledge the authors from the originating laboratories and the submitting laboratories, who generated and shared via GISAID genetic sequence data on which this research is based, as well as structural biology groups contributing their structures to the PDB.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.875406/full#supplementary-material

REFERENCES

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28 (1), 235–242. doi:10.1093/nar/28.1.235

Cai, Y., Zhang, J., Xiao, T., Lavine, C. L., Rawson, S., Peng, H., et al. (2021). Structural Basis for Enhanced Infectivity and Immune Evasion of SARS-CoV-2 Variants. *Science* 373 (6555), 642–648. doi:10.1126/science.abi9745

Cherian, S., Potdar, V., Jadhav, S., Yadav, P., Gupta, N., Das, M., et al. (2021). "Convergent Evolution of SARS-CoV-2 Spike Mutations, L452R, E484Q and P681R," in *The Second Wave of COVID-19 in Maharashtra* (India: bioRxiv).

Consortium, C. S. M. E. (2004). Molecular Evolution of the SARS Coronavirus during the Course of the SARS Epidemic in China. *Science* 303 (5664), 1666–1669. doi:10.1126/science.1092002

Crispell, J., Balaz, D., and Gordon, S. V. (2019). HomoplasyFinder: a Simple Tool to Identify Homoplasies on a Phylogeny. *Microb. Genom* 5 (1). doi:10.1099/mgen.0.000245

Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., et al. (2021). Estimated Transmissibility and Impact of SARS-CoV-2 Lineage B.1.1.7 in England. *Science* 372 (6538), eabg3055. doi:10.1126/science.abg3055

De Maio, N., Walker, C., Borges, R., Weilguny, L., Slodkowicz, G., and Goldman, N. (2020). Issues with SARS-CoV-2 Sequencing Data. Available at: https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.

Domingo, E. (2020). Molecular Basis of Genetic Variation of Viruses: Error-Prone Replication. *Virus as Populations*, 35–71. doi:10.1016/B978-0-12-816331-3.00002-7

Duffy, S. (2018). Why Are RNA Virus Mutation Rates So Damn High? *Plos Biol.* 16 (8), e3000003. doi:10.1371/journal.pbio.3000003

Gerdol, M. (2021). Emergence of a Recurrent Insertion in the N-Terminal Domain of the SARS-CoV-2 Spike Glycoprotein. *bioRxiv*.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data. *Bioinformatics* 32 (18), 2847–2849. doi:10.1093/bioinformatics/btw313

Gutierrez, B., Escalera-Zamudio, M., and Pybus, O. G. (2019). Parallel Molecular Evolution and Adaptation in Viruses. *Curr. Opin. Virol.* 34, 90–96. doi:10.1016/j.coviro.2018.12.006

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., et al. (2018). Nextstrain: Real-Time Tracking of Pathogen Evolution. *Bioinformatics* 34 (23), 4121–4123. doi:10.1093/bioinformatics/bty407

Harrell, F. E., Jr, and Harrell, M. F. E., Jr (2019). *Hmisc is R Package CRAN*, 235

Jangra, S., Ye, C., Rathnasinghe, R., Stadlbauer, D., Personalized Virology Initiative study, g., Krammer, F., et al. (2021). SARS-CoV-2 Spike E484K Mutation Reduces Antibody Neutralisation. *Lancet Microbe* 2 (7), e283–e284. doi:10.1016/S2666-5247(21)00068-9

Jaroszewski, L., Iyer, M., Alisoltani, A., Sedova, M., and Godzik, A. (2021). The Interplay of SARS-CoV-2 Evolution and Constraints Imposed by the Structure and Functionality of its Proteins. *Plos Comput. Biol.* 17 (7), e1009147. doi:10.1371/journal.pcbi.1009147

Jespersen, M. C., Peters, B., Nielsen, M., and Marcatili, P. (2017). BepiPred-2.0: Improving Sequence-Based B-Cell Epitope Prediction Using Conformational Epitopes. *Nucleic Acids Res.* 45 (W1), W24–W29. doi:10.1093/nar/gkx346

Jewell, B. L. (2021). Monitoring Differences between the SARS-CoV-2 B.1.1.7 Variant and Other Lineages. *The Lancet Public Health* 6 (5), e267–e268. doi:10.1016/S2468-2667(21)00073-6

Karim, S. S. A., and Karim, Q. A. (2021). Omicron SARS-CoV-2 Variant: a New Chapter in the COVID-19 Pandemic. *The Lancet* 398 (10317), 2126–2128. doi:10.1016/s0140-6736(21)02758-6

Katoh, K., and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* 30 (4), 772–780. doi:10.1093/molbev/mst010

Kimura, I., Kosugi, Y., Wu, J., Yamasoba, D., Butlertanaka, E. P., Tanaka, Y. L., et al. (2022). SARS-CoV-2 Lambda Variant Exhibits Higher Infectivity and Immune Resistance. *Cell Rep* 38 (2), 110218. doi:10.1016/j.celrep.2021.110218

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. L. (2001). Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete genomes11Edited by F. Cohen. *J. Mol. Biol.* 305 (3), 567–580. doi:10.1006/jmbi.2000.4315

Kumar, B. K., Rohit, A., Prithvisagar, K. S., Rai, P., Karunasagar, I., and Karunasagar, I. (2021). Deletion in the C-Terminal Region of the Envelope Glycoprotein in Some of the Indian SARS-CoV-2 Genome. *Virus. Res.* 291, 198222. doi:10.1016/j.virusres.2020.198222

Lam, J.-Y., Yuen, C.-K., Ip, J. D., Wong, W.-M., To, K. K.-W., Yuen, K.-Y., et al. (2020). Loss of Orf3b in the Circulating SARS-CoV-2 Strains. *Emerging Microbes & Infections* 9 (1), 2685–2696. doi:10.1080/22221751.2020.1852892

Elbe, S., and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* 1, 33–46. doi:10.1002/gch2.1018

Lei, X., Dong, X., Ma, R., Wang, W., Xiao, X., Tian, Z., et al. (2020). Activation and Evasion of Type I Interferon Responses by SARS-CoV-2. *Nat. Commun.* 11 (1), 3810. doi:10.1038/s41467-020-17665-9

Liang, T., Cheng, M., Teng, F., Wang, H., Deng, Y., Zhang, J., et al. (2021). Proteome-wide Epitope Mapping Identifies a Resource of Antibodies for SARS-CoV-2 Detection and Neutralization. *Signal. Transduct. Target. Ther.* 6 (1), 1–3. doi:10.1038/s41392-021-00573-9

Lin, J.-w., Tang, C., Wei, H.-c., Du, B., Chen, C., Wang, M., et al. (2021). Genomic Monitoring of SARS-CoV-2 Uncovers an Nsp1 Deletion Variant that Modulates Type I Interferon Response. *Cell. Host. icrobe.* 29 (3), 489–502. doi:10.1016/j.chom.2021.01.015

Madhi, S. A., Baillie, V., Cutland, C. L., Voysey, M., Koen, A. L., Fairlie, L., et al. (2021). Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant. *N. Engl. J. Med.* 384, 1885–1898. doi:10.1056/NEJMoa2102214

Martin, D. P., Weaver, S., Tegally, H., San, E. J., Shank, S. D., Wilkinson, E., et al. (2021). The Emergence and Ongoing Convergent Evolution of the N501Y Lineages Coincides with a Major Global Shift in the SARS-CoV-2 Selective Landscape. *Cell* 184 (20), 5189–5200.e7. doi:10.1016/j.cell.2021.09.003

McCallum, M., Marco, A. D., Lempp, F., Tortorici, M. A., Pinto, D., Walls, A. C., et al. (2021). N-terminal Domain Antigenic Mapping Reveals a Site of Vulnerability for SARS-CoV-2. *bioRxiv*. doi:10.1101/2021.01.14.426475

McCarthy, K. R., Rennick, L. J., Nambulli, S., Robinson-McCarthy, L. R., Bain, W. G., Haidar, G., et al. (2021). Recurrent Deletions in the SARS-CoV-2 Spike Glycoprotein Drive Antibody Escape. *Science* 371 (6534), 1139–1142. doi:10.1126/science.abf6950

Menéndez-Arias, L., Matamoros, T., and Cases-González, C. E. (2006). Insertions and Deletions in HIV-1 Reverse Transcriptase: Consequences for Drug Resistance and Viral Fitness. *Curr. Pharm. Des.* 12 (15), 1811. doi:10.2174/138161206776873608

Oostra, M., de Haan, C. A. M., and Rottier, P. J. M. (2007). The 29-nucleotide Deletion Present in Human but Not in Animal Severe Acute Respiratory Syndrome Coronaviruses Disrupts the Functional Expression of Open reading Frame 8. *J. Virol.* 81 (24), 13876–13888. doi:10.1128/JVI.01631-07

Palmer, J., and Poon, A. F. Y. (2019). Phylogenetic Measures of Indel Rate Variation Among the HIV-1 Group M Subtypes. *Virus. Evol.* 5 (2), vez022. doi:10.1093/ve/vez022

Planas, D., Veyer, D., Baidaliuk, A., Staropoli, I., Guivel-Benhassine, F., Rajah, M. M., et al. (2021). Reduced Sensitivity of SARS-CoV-2 Variant Delta to Antibody Neutralization. *Nature* 596, 276–280. doi:10.1038/s41586-021-03777-9

Plante, J. A., Mitchell, B. M., Plante, K. S., Debbink, K., Weaver, S. C., and Menachery, V. D. (2021). The Variant Gambit: COVID-19's Next Move. *Cell. Host. Microbe* 29 (4), 508–515. doi:10.1016/j.chom.2021.02.020

PyMOL (2021). *The PyMOL Molecular Graphics System* (Version 2.0 Schrödinger, LLC).

Ribes, M., Chaccour, C., and Moncunill, G. (2021). Adapt or Perish: SARS-CoV-2 Antibody Escape Variants Defined by Deletions in the Spike N-Terminal Domain. *Signal. Transduct. Target. Ther.* 6 (1), 164. doi:10.1038/s41392-021-00601-8

Santacroce, L., Charitos, I. A., Carretta, D. M., De Nitto, E., and Lovero, R. (2021). The Human Coronaviruses (HCoVs) and the Molecular Mechanisms of SARS-CoV-2 Infection. *J. Mol. Med.* 99 (1), 93–106. doi:10.1007/s00109-020-02012-8

Sedova, M., Jaroszewski, L., Alisoltani, A., and Godzik, A. (2020). Coronavirus3D: 3D Structural Visualization of COVID-19 Genomic Divergence. *Bioinformatics* 36 (15), 4360–4362. doi:10.1093/bioinformatics/btaa550

Shu, Y., and McCauley, J. (2017). GISAID: Global Initiative on Sharing All Influenza Data - from Vision to Reality. *Euro Surveill.* 22 (13), 30494. doi:10.2807/1560-7917.ES.2017.22.13.30494

Smith, C. C., Olsen, K. S., Gentry, K. M., Sambade, M., Beck, W., Garness, J., et al. (2021). Landscape and Selection of Vaccine Epitopes in SARS-CoV-2. *Genome Med.* 13 (1), 1–23. doi:10.1186/s13073-021-00910-1

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., et al. (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* 182 (5), 1295–1310. doi:10.1016/j.cell.2020.08.012

Studer, R. A., Dessailly, B. H., and Orengo, C. A. (2013). Residue Mutations and Their Impact on Protein Structure and Function: Detecting Beneficial and Pathogenic Changes. *Biochem. J.* 449 (3), 581–594. doi:10.1042/BJ20121221

Tang, X., Ying, R., Yao, X., Li, G., Wu, C., Tang, Y., et al. (2021). Evolutionary Analysis and Lineage Designation of SARS-CoV-2 Genomes. *Sci. Bull.* 66 (22), 2297–2311. doi:10.1016/j.scib.2021.02.012

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., et al. (2021). Detection of a SARS-CoV-2 Variant of Concern in South Africa. *Nature* 592 (7854), 438–443. doi:10.1038/s41586-021-03402-9

Tromas, N., Zwart, M. P., Forment, J., and Elena, S. F. (2014). Shrinkage of Genome Size in a Plant RNA Virus upon Transfer of an Essential Viral Gene into the Host Genome. *Genome Biol. Evol.* 6 (3), 538–550. doi:10.1093/gbe/evu036

Turakhia, Y., De Maio, N., Thornlow, B., Gozashti, L., Lanfear, R., Walker, C. R., et al. (2020). Stability of SARS-CoV-2 Phylogenies. *Plos Genet.* 16 (11), e1009175. doi:10.1371/journal.pgen.1009175

van Dorp, L., Acman, M., Richard, D., Shaw, L. P., Ford, C. E., Ormond, L., et al. (2020). Emergence of Genomic Diversity and Recurrent Mutations in SARS-CoV-2. *Infect. Genet. Evol.* 83, 104351. doi:10.1016/j.meegid.2020.104351

Viana, R., Moyo, S., Amoako, D. G., Tegally, H., Scheepers, C., Lessells, R. J., et al. (2021). Rapid Epidemic Expansion of the SARS-CoV-2 Omicron Variant in Southern Africa. *Nature* 603 (7902), 679–686. doi:10.1038/s41586-022-04411-y

Viguera, E., Canceill, D., and Ehrlich, S. D. (2001). Replication Slippage Involves DNA Polymerase Pausing and Dissociation. *EMBO J.* 20 (10), 2587–2595. doi:10.1093/emboj/20.10.2587

Walker, P. J., Firth, C., Widen, S. G., Blasdell, K. R., Guzman, H., Wood, T. G., et al. (2015). Evolution of Genome Size and Complexity in the Rhabdoviridae. *Plos Pathog.* 11 (2), e1004664. doi:10.1371/journal.ppat.1004664

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2--a Multiple Sequence Alignment Editor and Analysis Workbench. *Bioinformatics* 25 (9), 1189–1191. doi:10.1093/bioinformatics/btp033

Wickham, H. (2011). ggplot2. *Wires Comp. Stat.* 3 (2), 180–185. doi:10.1002/wics.147

Wood, N., Bhattacharya, T., Keele, B. F., Giorgi, E., Liu, M., Gaschen, B., et al. (2009). HIV Evolution in Early Infection: Selection Pressures, Patterns of Insertion and Deletion, and the Impact of APOBEC. *Plos Pathog.* 5 (5), e1000414. doi:10.1371/journal.ppat.1000414

Yu, G. (2020). Using Ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinformatics* 69 (1), e96. doi:10.1002/cpbi.96

Zhang, Q., Wang, P., Kim, Y., Haste-Andersen, P., Beaver, J., Bourne, P. E., et al. (2008). Immune Epitope Database Analysis Resource (IEDB-AR). *Nucleic Acids Res.* 36 (Suppl. l_2), W513–W518. doi:10.1093/nar/gkn254

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin. *Nature* 579 (7798), 270–273. doi:10.1038/s41586-020-2012-7

Zwart, M. P., Willemsen, A., Daròs, J.-A., and Elena, S. F. (2014). Experimental Evolution of Pseudogenization and Gene Loss in a Plant RNA Virus. *Mol. Biol. Evol.* 31 (1), 121–134. doi:10.1093/molbev/mst175

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.