



Evolutionary Dynamics of Indels in SARS-CoV-2 Spike Glycoprotein

R Shyama Prasad Rao¹ , Nagib Ahsan^{2,3}, Chunhui Xu⁴, Lingtao Su⁴, Jacob Verburg⁵, Luca Fornelli^{2,6}, Daisuke Kihara^{5,7} and Dong Xu⁴

¹Biostatistics and Bioinformatics Division, Yenepoya Research Center, Yenepoya University, Mangaluru, Karnataka, India. ²Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK, USA. ³Mass Spectrometry, Proteomics and Metabolomics Core Facility, Stephenson Life Sciences Research Center, University of Oklahoma, Norman, OK, USA. ⁴Department of Electrical Engineering and Computer Science, Informatics Institute, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, USA. ⁵Department of Biological Sciences, Purdue University, West Lafayette, IN, USA. ⁶Department of Biology, University of Oklahoma, Norman, OK, USA. ⁷Department of Computer Science, Purdue University, West Lafayette, IN, USA.

Evolutionary Bioinformatics
Volume 17: 1–10
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11769343211064616


ABSTRACT: SARS-CoV-2, responsible for the current COVID-19 pandemic that claimed over 5.0 million lives, belongs to a class of enveloped viruses that undergo quick evolutionary adjustments under selection pressure. Numerous variants have emerged in SARS-CoV-2, posing a serious challenge to the global vaccination effort and COVID-19 management. The evolutionary dynamics of this virus are only beginning to be explored. In this work, we have analysed 1.79 million spike glycoprotein sequences of SARS-CoV-2 and found that the virus is fine-tuning the spike with numerous amino acid insertions and deletions (indels). Indels seem to have a selective advantage as the proportions of sequences with indels steadily increased over time, currently at over 89%, with similar trends across countries/variants. There were as many as 420 unique indel positions and 447 unique combinations of indels. Despite their high frequency, indels resulted in only minimal alteration of N-glycosylation sites, including both gain and loss. As indels and point mutations are positively correlated and sequences with indels have significantly more point mutations, they have implications in the evolutionary dynamics of the SARS-CoV-2 spike glycoprotein.

KEYWORDS: Computational proteomics, COVID-19, indels, molecular evolution, N-glycosylation sites, SARS-CoV-2, selection, sequence analysis

RECEIVED: August 7, 2021. **ACCEPTED:** November 12, 2021.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work did not receive any specific funding. DX acknowledges support by the National Institutes of Health (R35-GM126985). DK acknowledges supports by the National Institutes of Health (R01GM133840, R01GM123055) and the National Science Foundation (DBI2003635, CMMI1825941 and MCB1925643). JV is supported by NIGMS-funded predoctoral

fellowship (T32 GM132024). NA acknowledges the initial funding support from the OU VPRP Office for the establishment of the Proteomics Core Facility.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: R Shyama Prasad Rao, Biostatistics and Bioinformatics Division, Yenepoya Research Center, Yenepoya University, Mangaluru, Karnataka 575018, India. Email: drsprao@gmail.com

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), responsible for the currently ongoing pandemic of coronavirus disease 2019 (COVID-19),¹ has infected more than 246 million people and killed over 5.0 million.² Related coronaviruses – SARS-CoV-1 and Middle East respiratory syndrome coronavirus (MERS-CoV) – have also caused pandemics in the recent past.

SARS-CoV-2 belongs to the general class of enveloped viruses (which include influenza and human immunodeficiency viruses, among others) that show great plasticity and immune evasiveness due to a protective lipid bilayer and embedded glycoproteins that are heavily N-glycosylated and used as a ‘glycan shield’.^{3–5} The lipid bilayer envelope of these viruses is particularly sensitive to desiccation, heat and detergents.⁶ However, the envelope glycoproteins of many enveloped viruses are known to be particularly variable and found to evolve quickly under selection pressure. As a result, the patterns and drivers of envelope glycoprotein variations in these

viruses have been studied keenly.^{4,7–9} Yet, as the ‘enveloped viruses’ is a broad category, the key features that lead to differences in infectivity and antigenicity among different members, for example HIV-1 versus SARS-CoV-2, is of particular interest.⁵

Given the nature of the pandemic, the genomic architecture and its evolutionary dynamics are being keenly explored in coronaviruses^{10,11} and SARS-CoV-2 in particular.^{12–15} Pachetti et al¹⁶ have shown the emergence of mutations in the SARS-CoV-2 genome and RNA-dependent-RNA polymerase as a mutational hotspot. Mercatelli and Giorgi¹⁷ have analysed 48 635 SARS-CoV-2 complete genomes and found 7.23 mutations per sample on average. Given its importance as a key interactor of angiotensin-converting enzyme 2 (ACE2) for host cell entry and as a target for neutralising antibodies, spike glycoprotein in the envelope of SARS-CoV-2 is special and therefore its variants are keenly watched (<https://www.gisaid.org/hcov19-variants/>). D614G was found to be a prevalent spike mutation. However, its precise effect is unclear as it was



known to increase infectivity¹⁸ as well as increase susceptibility to neutralisation.¹⁹ Numerous other mutations in the spike glycoprotein have also been documented.²⁰

Studies on variants of SARS-CoV-2 mainly focus on point mutations. This is because there is a massive prevalence of single-nucleotide polymorphisms (SNPs) compared to short indels, which only account for 0.8% of mutations.¹⁷ A key reason for indels being less common is that they are more deleterious due to frame-shifting than SNPs.^{21–23} As SARS-CoV-2 has been shown to accumulate indels,⁵ we are only beginning to explore them and appreciate their myriad roles. For example, Chrisman et al²⁴ have looked at indels in the SARS-CoV-2 genome and mapped it to regions of discontinuous transcription breakpoints. Lee et al²⁵ have shown a novel indel in nucleocapsid (N) gene leading to negative results for N gene-based RT-PCR that was approved by US/FDA and EU/CE-IVD. Their work also emphasised the genetic variability and rapid evolution of SARS-CoV-2.²⁵

While indels have been explored predominantly at the genomic level,^{12,17} they were less emphatically examined at the proteomic level and even less so in the spike glycoprotein. Despite their rarity, indels accelerate protein evolution,²⁶ and could be especially interesting and important in the spike glycoprotein. For example, indels can be beneficial as recurrent deletions in SARS-CoV-2 spike glycoprotein are shown to drive antibody escape and accelerate antigenic evolution.²⁷ Garry and Gallaher²⁸ have explored naturally occurring indels in multiple coronavirus spike proteins and provided evidence against a laboratory origin of SARS-CoV-2. Garry et al²⁹ have also shown that the mutations in ‘variants of concern’ (VOC) commonly occur near indels. Despite these revelations, given an accumulating wealth of SARS-CoV-2 genomic data, large-scale patterns of indels in spike glycoproteins have not been fully explored and appreciated.

With this background, we sought to answer some of the open questions: (1) What is the prevalence and pattern of indels in the spike protein? (2) What are the evolutionary dynamics of sequences with indels? (3) Is there any relationship with point mutations? and (4) What is the effect of indels on N-glycosylation sites? We analysed a large set of 1.79 million SARS-CoV-2 spike protein sequences and showed that over 50% contained 1 or more indels. The proportion of sequences with indels has risen sharply, and currently over 76% of unique sequence variants and 89% of total sequences have indels. Indels and point mutations are positively correlated and sequences with indels seem to have more point mutations overall. Further, indels had minimal effect on N-glycosylation sites. We discuss these findings in the context of the evolutionary dynamics of the viral protein.

Materials and Methods

Spike sequences and metadata

The SARS-CoV-2 spike protein sequences and related metadata were obtained from the GISAID website (<https://www.gisaid.org/>; accessed on June 3, 2021).³⁰ The spike protein sequences were based on the translation of the genome after alignment to the reference hCoV-19/Wuhan/WIV04/2019 (EPI_ISL_402124) and were in the fasta format. The associated tsv metadata included date of sample collection, location/country of origin, and clade/lineage information of the virus, among other details.

Sequence analyses

There were a total of 1 790 224 sequences in the database at the time of access. However, as there were numerous quality issues with the data,³¹ many sequences were filtered out. For example, 465 419 sequences containing X (on average of 82.4 X per sequence) that arose from the translation of low-quality regions and/or ambiguous bases in the genome were excluded. Incomplete sequences based on missing N-terminal and/or C-terminal codons were ignored. As our interest was to look for the pattern of short indels,²² disproportionately short sequences (eg, 3744 sequences were very short – less than 1000 residues in length) that were missing internal parts possibly due to sequencing/annotation issues were ignored. Finally, sequences with incomplete metadata on the date of sample collection were also excluded.

Sequence analyses

In the final set of 1 311 545 spike protein sequences, there were a total of 49 118 unique sequences based on 100% identity cut-off³² that included all possible variants. Average identity (based on CD-HIT, <http://weizhong-lab.ucsd.edu/cd-hit/>) of sequences was 99.47% (median = 99.53%). For each sequence, a pair-wise alignment with the reference (EPI_ISL_402124) was done using Biopython.³³ The BLOSUM62 was used as the substitution matrix, and gap open and extension penalties were set at 11 and 1, respectively.³⁴ Any gap in the query sequence was considered a deletion and a gap in the reference was considered an insertion.²¹ Finally, a mismatched residue was considered a point mutation (or substitution).

To see the temporal dynamics, the proportions of sequences with indels were plotted against the date of sample collection (month-wise). Based on the sequence metadata, country and clade/lineage-specific dynamics of indels were also plotted. For each alignment, the number of indels was enumerated, and the positions of indels were mapped to the reference sequence. Further, potential N-glycosylation sites were identified based on the regular pattern of tripeptide sequons NXS or NXT where N is asparagine, S is serine, T is threonine and X is any amino acid residue, and compared with potential N-glycosylation sites found in the reference sequence (Table 1 and Supplemental Table S1).^{4,9,35} Multiple sequence alignments, where required, were done using Clustal Omega (<https://www.ebi.ac.uk/Tools/msa/clustalo/>). All sequence analysis and data handling, where specifically not mentioned, were performed in Python; and visualisation/graphs were created in Microsoft Excel.

Table 1. List of N-glycosylation site variants due to indels in SARS-CoV-2 spike protein.

ACCESSION ID	LEN	SITE	ALIGNMENT ^a	COUNTRY	DATE ^b	CLADE	LINEAGE	FREQ ^c
EPI_ISL_959416	1274	N149	145-YHKNNNKSWMES 145-YHKNN-KSWMES	Bangladesh	20-01-2021	GR	B.1.1.25	2
EPI_ISL_1134689	1274	N149	145-YHKNNNKSWMES 145-YHKNN-KSWMES	USA	13-02-2021	GH	B.1.2	3
EPI_ISL_1164478	1271	N. . .#	242-LALHN--LTPGD 242-LALHRSYLTPGD	Turkey	15-02-2021	GR	B.1.1.189	2
EPI_ISL_1167125	1273	N149 N438	145-YHKNNKKSWMES 145-YHKNNK-SWMES 433-VIAWNSTNNLD 433-VIAWNS-NNLD	Senegal	27-01-2021	O	B.1	1
EPI_ISL_1167127	1273	N149	145-YHKNNKKSWMES 145-YHKNNK-SWMES	Senegal	02-02-2021	O	B.1	1
EPI_ISL_1272919	1274	N149	145-YHKNNNKSWMES 145-YHKNN-KSWMES	USA	25-02-2021	GH	B.1.2	1
EPI_ISL_1359859	1268	N. . .#	239-LALHN--LTPGD 242-LALHRSYLTPGD	France	27-01-2021	GRY	B.1.1.7 VOC Alpha	1
EPI_ISL_1591625	1274	N1134	1130-IGIVNNNTVYDP 1130-IGIVNN-TVYDP	Mexico	25-03-2021	GR	B.1.1.222	2
EPI_ISL_1791465	1278	N. . .#	157-FRVI NT CYSS 157-FRV-----YSS	USA	18-01-2021	GH	B.1.2	2
EPI_ISL_1843929	1269	N17	13-SQCVN TT RTQ 13-SQCVNLTTRTQ	Germany	10-04-2021	GRY	B.1.1.7 VOC Alpha	1
EPI_ISL_1922075	1278	N. . .#	867-DEMI NFT ISAQ 867-DEMI-----AQ	Turkey	12-03-2021	GH	B.1.469	1
EPI_ISL_2013547	1274	N61	57-PFFSN SVT WFHA 57-PFFSN-VTWFHA	USA	08-04-2021	GH	B.1.429 VOI Epsilon	1
EPI_ISL_2091257	1272	N1074	1070-AQEK N -STAPA 1070-AQEK NFT TAPA	Mexico	12-04-2021	S	A.2.5	1
EPI_ISL_2102041	1270	N. . .#	283-TDDGN IT DAAL 286-TD--AVDCAL	Turkey	09-03-2021	GRY	B.1.1.7 VOC Alpha	1
EPI_ISL_2102044	1270	N17 N. . .#	13-SQCVN L --TQ 13-SQCVNLTTRTQ 23-PAYT NYT NSFT 26-PA---YTNSFT	Turkey	09-03-2021	G	B.1.36.17	1
EPI_ISL_2262279	1267	N17	13-SQCVN F --TQ 13-SQCVNLTTRTQ	Germany	10-05-2021	GH	B.1.351 VOC Beta	1
EPI_ISL_2269445	1271	N1134	1127-IGIVNNNTVYDP 1130-IGIVNN-TVYDP	USA	10-05-2021	GRY	B.1.1.7 VOC Alpha	1

– New site not present in the reference sequence. Representative Accession IDs (based on the earliest date of sample collection) are arranged in ascending order. An additional list of loss of sites due to deletions is given in Supplemental Table S1.

^aGain of sites is highlighted in green, loss of sites in blue, altered sites in orange, original sites in yellow and non-sites in grey.

^bDate of sample collection.

^cFrequency out of 1 311 545 sequences.

Structural analyses

The positions of indels and N-glycosylation sites were visualised on the 3D structure of SARS-CoV-2 spike glycoprotein (PDB ID: 6VXX or 6XR8) using the Visual Molecular Dynamics (VMD) programme (<https://www.ks.uiuc.edu/Research/vmd/>). To see if indels have any preference for sequence/structural/functional features such as surface-exposed

regions, solvent accessible surface area (SASA) information was obtained using the DSSP programme, which calculates an accessibility score (ranged from 0 to 277) from the 3D structure (<http://swift.cmbi.ru.nl/gv/dssp/>).³⁶ Protein disorder was calculated (values ranged from 0 to 0.41) using DISpro (<http://scratch.proteomics.ics.uci.edu/>), and shorter disorder regions known as molecular recognition features (MoRF) were quantified (values ranged from 0.21 to 0.8) using MoRFChiBi_Web

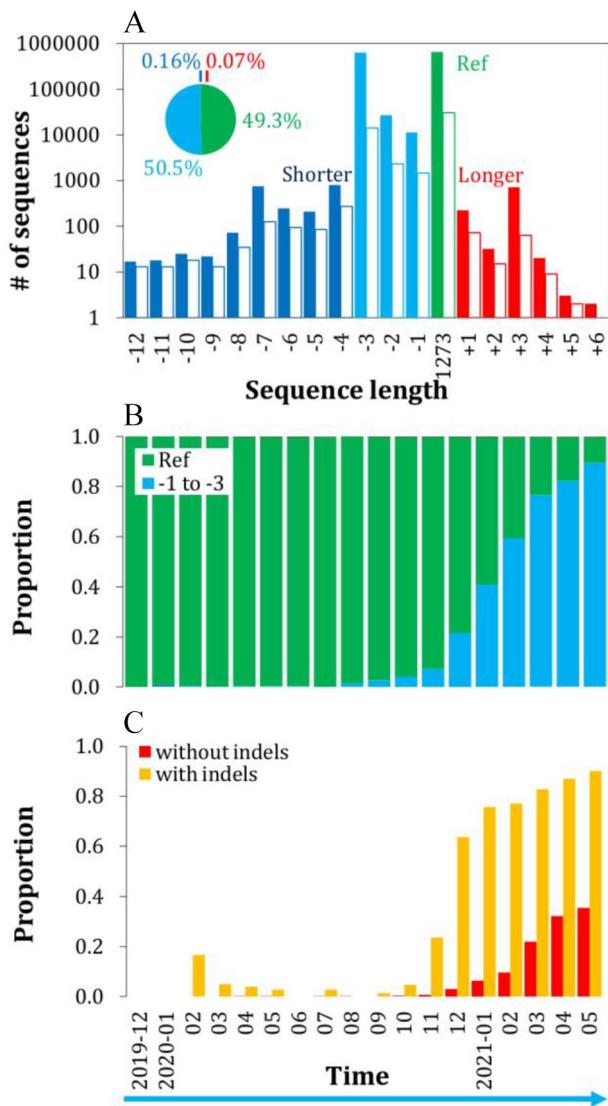


Figure 1. (A) Distribution of sequences with indels in SARS-CoV-2 spike glycoprotein ($n=1\,311\,545$). Over 50% of sequences (inset pie chart) have at least 1 indel, with sequences containing 3 deleted residues being very frequent. The pattern is similar (open bars) even if only unique sequences ($n=49\,118$) are considered. (B) Proteins/sequences with indels clearly seem to have a selective advantage as their proportion has risen sharply over time and currently (May 2021) represents 89.3% of all sequences. (C) Month-wise proportions of variants of concern/interest coming from sequences with indels compared to that of without indels.

(<https://morf.msl.ubc.ca/index.xhtml>).³⁷ Finally, information on different functional domains of SARS-CoV-2 spike glycoprotein was obtained from the literature/UniProt (<https://www.uniprot.org/>) and residue overlap coefficient was enumerated (Supplemental Table S2).³⁸

Statistical analyses

Where required, a 1-proportion Z-test was used to check if the observed proportion was significantly different from the expected. A chi-square test for independence was performed to check whether (multiple) sample proportions were significantly

different.³⁹ Correlations between indels and other variables (such as point mutations, accessibility scores, etc.) were measured using a more robust Kendall τ coefficient. The significance of correlation coefficient was tested using `cor.test()`, which is based on t -distribution or approximation. A t -test was used to compare the means of 2 groups (eg, mutations in sequences with or without indels). Where required, the P -values were corrected for multiple comparisons using Benjamini-Hochberg (BH) method.⁴⁰ All statistical tests were done using R.

Results

Distribution of sequences with indels

The SARS-CoV-2 spike glycoprotein reference sequence (EPI_ISL_402124) contains 1273 amino acid residues. The distribution of sequences with short indels was plotted based on the number of sequences in each length category and shown in Figure 1A as a bar diagram. Overall, the distribution is similar for all sequences ($n=1\,311\,545$, filled bars) and unique sequences ($n=49\,118$, open bars). Over 50% of all sequences (inset pie chart) had at least 1 indel. Sequences with deletions (50.5%) were far more common than sequences with insertions (0.07%). Further, sequences containing 3-residue deletions were very frequent. A small number of sequences (0.16%) had deletions of more than 3 residues. However, these proportions (36.4%, 0.33% and 1.35%) are significantly different ($\chi^2=6991.5$, $P\approx 0$) if only unique sequences were considered.

Dynamics of sequences with indels

The proportions of sequences with indels over time (month-wise) are given in Figure 1B, which shows an increasing trend. For example, in August 2020 the proportion of sequences with 1 to 3 deletions were about 1.4% which increased to 89.3% in May 2021. The proportions of sequences with insertions, and deletions of more than 3 residues also increased (Supplemental Figure S1A). A similar increasing trend is seen even if only unique sequences were considered (Supplemental Figure S1B). In May 2021, 76.3% of unique sequences have 1 to 3 deletions, and 1.8% of unique sequences with deletions of more than 3 residues. This increasing trend of sequences with indels holds true across countries (Supplemental Figure S1C). Almost all sequence variants currently present in the United Kingdom and South Africa have indels. On the other hand, only a small proportion of sequences from Brazil currently have indels. It should be noted that some patterns were a bit noisy due to small sample sizes (Supplemental Figure S2A), for example, in early months and/or country-wise trends. The proportions of sequence contributions from many countries that have reported variants of concern such as India, Brazil, South Africa and Nigeria were very small (Supplemental Figure S2B). In particular, just 0.74% of sequences were from India. However, it becomes 2.7% if only unique sequences were considered (Supplemental Figure S2C). The distribution of proportions of



Figure 2. (A) Map of indels (insertion in green and deletion in red) in SARS-CoV-2 spike glycoprotein. Incidence of indels along the sequence. The first panel shows the frequency (scale at right indicates the number of unique sequence variants) and the second panel shows the occurrence of indels. As many as 420 indel positions (142 insertion and 358 deletion positions) are present. Three-residue deletion of 69, 70 and 144 is the most common combination, but there are as many as 447 unique combinations of indels. (B) Multiple sequence alignment (using Clustal Omega, <https://www.ebi.ac.uk/Tools/msa/clustalo/>) shows 17 combinations of deletions present in Delta (B.1.617.2) and Kappa (B.1.617.1) variants (representative sequences based on the earliest date of sampling).
 *Previously 'Delta', but 'None' in the latest update of Pango v.3.1.14 2021-09-28. See also Supplemental Figure S3.

sequences with indels over time (Figure 1B and Supplemental Figure S1B) is heavily influenced by the dominant variant Alpha (Supplemental Figure S1D). However, the proportions of sequences with indels are also increasing in all other variants (Supplemental Figure S1D – lower panel).

It is interesting to note that a far higher proportion of variants of concern/interest (that have been sharply increasing in the past 6 months) come from sequences with indels compared to sequences without indels (0.793 vs 0.093, $P \approx 0$, 2-proportion Z-test). A month-wise trend is shown in Figure 1C. Overall, 84.1% of unique sequences from VOC/I have 1 or more indels.

Incidence of indels along the spike sequence

Figure 2A shows the map of indels along the spike glycoprotein sequence. Supplemental Figure S3A-C individually show the maps of indels for sequences with deletions (n = 18 560), zero net indels (sequences that contained equal numbers of 1 or more insertions and deletions, n = 92) and insertions (n = 161). While the average number of insertions (2.1, n = 161 unique sequences with insertions) and deletions (2.8, n = 18 560 unique sequences with deletions) were small, there were as

many as 420 unique indel positions (142 insertion and 358 deletion positions, 80 common positions with insertion or deletion) (Supplemental Table S2). However, it may be noted that indels were far less common (odds=0.47) in the C-terminal half of the spike protein sequence. Further, indels were frequent only in a few residue-positions. For example, there were just 14 residue-positions (69-70, 140-144, 156-157, 241-243 and 246-247) with 100 or more instances of deletion and only 2 residue-positions (216 and 217) with 100 or more instances of insertion. Nevertheless, there were as many as 447 unique combinations of indels (ranging from 1 to 15 indels) – for example – insertions at 6, 144, 214-216, etc., and deletions of 69-70, 144, 69-70, 144, 156-157, 241-243, etc. Among 3-residue deletions, deletion of 69 (amino acid H), 70 (V) and 144 (Y) was the most common combination that has emerged very early in the major lineage B.1.1.7 (VOC Alpha). However, as seen in multiple sequence alignment (Supplemental Figure S3D), there is variability in the position of 3-residue deletion leading to the emergence of new deletion variants. Figure 2B shows the multiple sequence alignment of Delta (B.1.617.2) and Kappa (B.1.617.1) variants that contain as many as 17 combinations of deletions (deletion of residues 156 (E) and 157 (F) being the most common). It may be noted

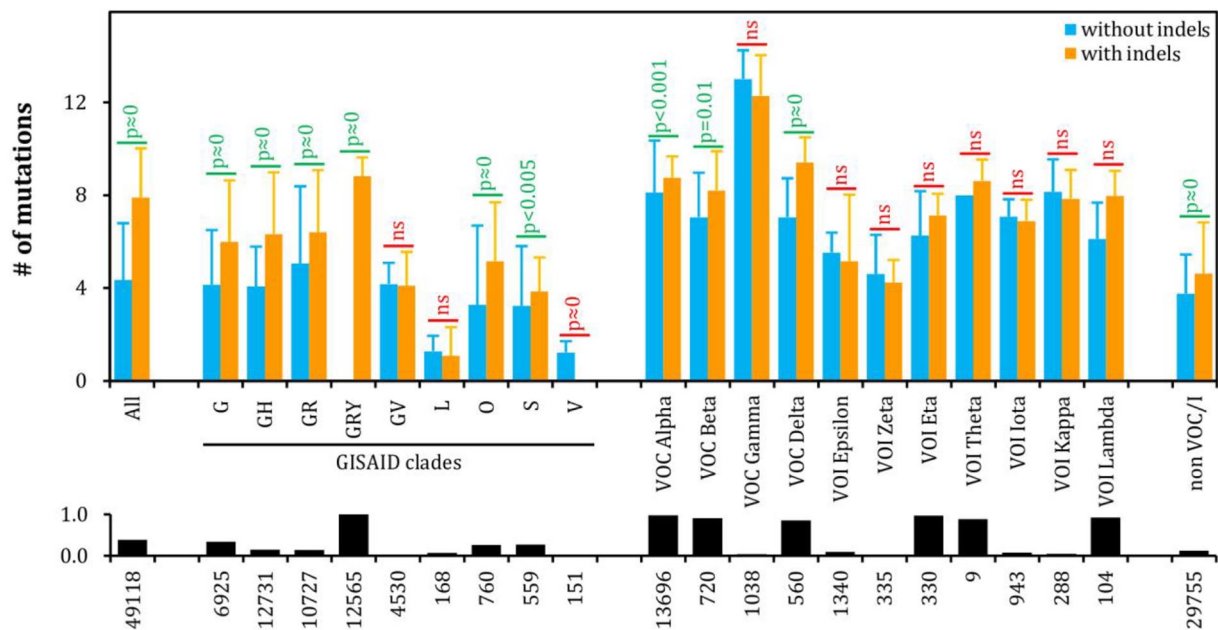


Figure 3. Indels versus point mutations. Bar plot shows the average number of point mutations (mean \pm SD) in sequences with indels compared to sequences without indels. Point mutations are significantly more (*t*-test with BH correction) in sequences with indels across clades and lineages of variants of concern/interest. They are not significant (ns, or opposite as in clade V) when the sample size and/or proportion of sequences with indels are too small (lower panel of bar plot and sample size).

that 79% of sequences from Delta and Kappa variants of concern/interest ($n=9133$, currently represent 0.7% of the total) contain 1 or more deletions.

Based on the correlation, indels showed a low (but significant) preference for surface-exposed regions ($\tau=0.054$ and $P=.027$ for insertions, and $\tau=0.104$ and $P=1.2E-5$ for deletions). Correlation with overall protein disorder was not significant ($\tau=0.155$ and $P=.56$ for insertions, and $\tau=0.091$ and $P=.66$ for deletions), possibly because long disordered regions were very few and far apart. Correlation with shorter disordered regions (MoRF) was low but significant ($\tau=0.122$ and $P=6.7E-8$ for insertions, and $\tau=0.107$ and $P=7.1E-7$ for deletions). On the 3D structure of SARS-CoV-2 spike glycoprotein (Supplemental Figure S4), indels were prevalent in much of the outer side of the N-terminal domain (NTD). This was reflected in the domain analysis wherein NTD and terminal regions showed a high overlap coefficient (Supplemental Table S2). Deletions, in particular, were also more frequent at the flanks of the receptor-binding domain (RBD), but were far less common in the S2 subunit region and were almost absent at the inner side of the subunits (Supplemental Figure S4).

Indels versus point mutations

The spike sequences with indels had more (over 1.81 times; 7.9 ± 2.1 vs 4.3 ± 2.5 , mean \pm standard deviation) point mutations than sequences without indels. Similar patterns were seen even when different GISAID clades or lineages were taken separately (Figure 3). However, they were not significant (*t*-test with BH correction) in some groups when the proportion of

sequences with indels was very small or due to a small sample size. Overall, VOC/I had more point mutations compared to non-VOC/I, but in both groups sequences with indels had significantly more point mutations.

The distribution of point mutations along the sequence is shown in Supplemental Figure S5. Sequences with indels had, apart from D614G, 6 more prevalent mutations. There were 96 positions in sequences with indels ($n=18813$) and 101 in sequences without indels ($n=30305$, scaled to the number of sequences with indels) that had 100 or more instances of point mutations. Overall, the N-terminal region had longer stretches of residues with more than 100 occurrences of point mutations. There was a small but significant positive correlation ($\tau=0.224$, $P=4.0E-25$) between the distribution of indels and point mutations along the primary sequence ($\tau=0.136$, $P=1.9E-9$ for insertions; $\tau=0.22$, $P=5.5E-24$ for deletions). Many point mutations in VOC Delta were differentially abundant (2-proportion Z-test with BH correction) in sequences with indels compared to sequences without indels (Supplemental Figure S5C) and were more common in the N-terminal half where indels are also present (Figure 2B).

Effect of indels on N-glycosylation sites

Based on the occurrence of sequons, there are 22 potential N-glycosylation sites (7 NXS and 15 NXT) in the SARS-CoV-2 spike glycoprotein reference sequence (EPI_ISL_402124). Despite indels being present in over 50% of the total 1311545 sequences, there was remarkably minimal effect on N-glycosylation sites due to indels. The list of 67

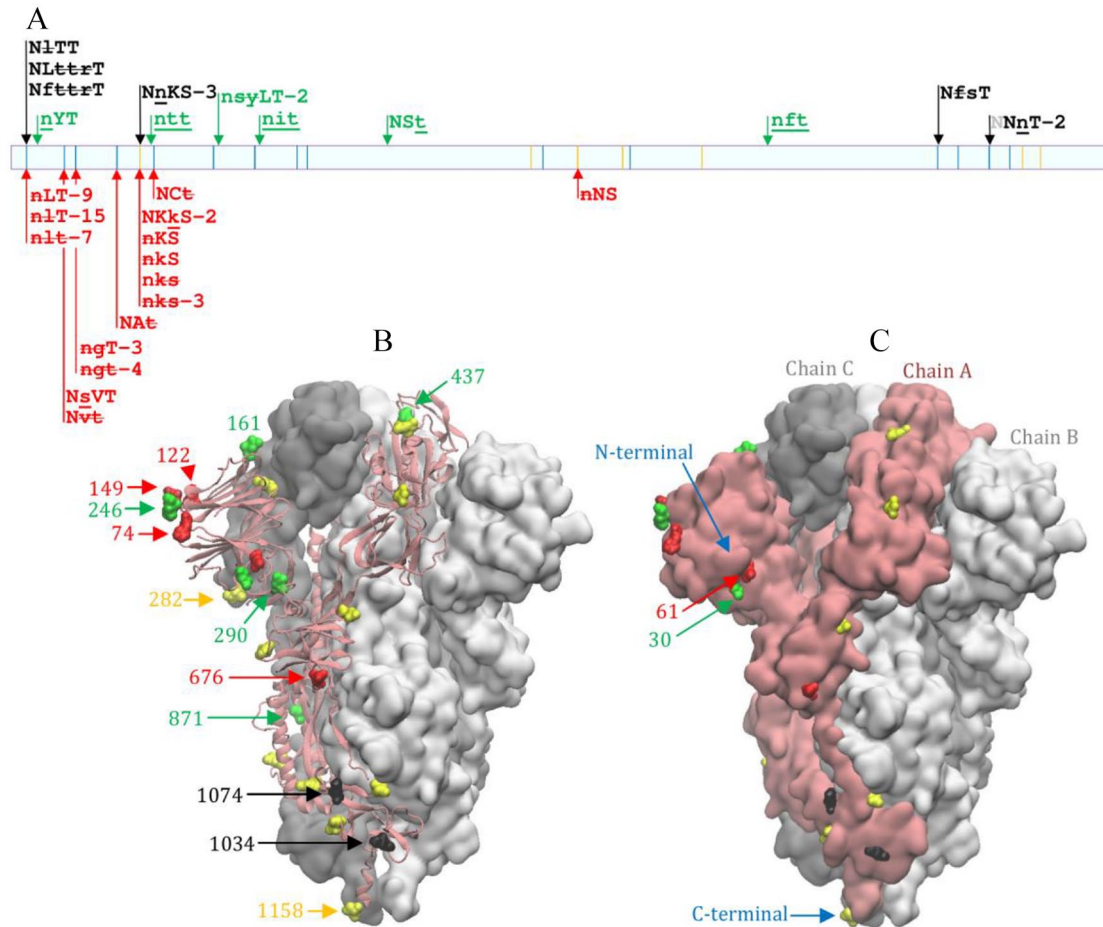


Figure 4. (A) Alteration of N-glycosylation sites due to indels. The panel shows potential N-glycosylation sites (NXS sequons in orange and NXT in blue) along the sequence and the effect of indels (gain of site in green, loss of site in red, and altered site in black). Altered residues (compared to reference sequence) are shown in lower case letters with insertions underscored, and deletions struck through. Multiple instances of the same type of change are numbered after hyphen. The gains of sites were more scattered, while the losses of sites were mostly at the N-terminal part of the sequence. (B) Cartoon and (C) space-fill structures of SARS-CoV-2 spike glycoprotein showing the positions of N-glycosylation sites. Some key sites are indicated by arrows and numbers. Of the 6 gains of sites, 3 sites (at 290, 437 and 871) are completely buried in the 3D structure.

instances of N-glycosylation sites that have been altered by the indels (in 65 unique sequences) is given in Table 1 (Supplemental Table S1), and their positions are shown in Figure 4A to C. There were 7 instances of gain of sites (Table 1, green) – due to insertions (eg, near position 27, A---YT to AYTNYT), or deletions with substitution (eg, near position 246, RSYLT to N---LT). There were also many more instances of loss of sites (Table 1, blue) – mostly due to deletions (eg, at position 17, VNLT to V--T), but also due to insertions (eg, at position 61, N-VT to NSVT). While the loss of sites occurred mostly at the N-terminal part of the spike, the gain of sites was a bit more scattered. It is important to note that all gains of sites were of NXT types. However, 3 sites (at around 290, 437 and 871) were completely buried in the 3-D structure (Figure 4B and C). There were also a few alterations of existing sites due to insertions or deletions (Table 1, orange). It may be noted that these indel-based N-glycosylation site alternations occurred in sequences belonging to many clades/lineages – many of them were of variants of concern (Table 1 and Supplemental Table S1).

Discussion

There was great interest and urgency to unravel the architecture of the SARS-CoV-2 genome.¹ Given the severity of the pandemic, there is currently an explosion of viral sequencing bringing along the concerns of data accessibility and ownership,^{41,42} data integrity/quality,³¹ and inequality of sequencing effort/data collection among countries.^{43,44} Nonetheless, the availability of a vast amount of sequences has allowed the scientific community to track the changes in the SARS-CoV-2 genome as the pandemic is progressing. Numerous studies have shown the emergence and dynamics of new variants, although the main emphasis was on non-synonymous substitutions at the genomic level.^{12,14,16,18,20,29}

Indel variants were underexplored and unappreciated due to their relative rarity,^{12,17,23} but it is becoming evident that they play key roles in the SARS-CoV-2 genome.^{24,25} Indels are even less explored at the proteomic level because they are primarily found in untranslated regions.¹² In this work, we showed that there is an incursion of short indels at numerous positions in the SARS-CoV-2 spike glycoprotein. Of these, 2 very common

deletions (Δ H69/ Δ V70, primarily found in UK variants) were well known and shown to have recurrent emergence and transmission.⁴⁵ While deletions facilitate antibody escape, it was found that BNT162b2 vaccine-elicited sera can still neutralise 69/70 deletion variant.⁴⁶ One reason could be concurrent substitutions that offset this effect. For example, D614G substitution, also found in the deletion variant (clade G, prevalent in Europe), was shown to increase SARS-CoV-2 susceptibility to neutralisation.¹⁹ Some independent deletions of 5 to 7 residues were known to occur in and near the furin-like cleavage site (around residue position 681). It was hypothesised that those deletions might be involved in viral infection.⁴⁷ However, at present, the functional implications of numerous other indels are completely unexplored/unknown.

The proportion of proteins/sequences with indels has risen sharply over time. Currently, 78.4% of the viral variants have indels; and while Δ 69-70 and Δ 144 were present in the majority of the variants, there also seems to be an increasing trend for longer indels. Thus, indels seem to have a selective advantage, although random drift cannot be ignored. Recurrent deletions in the SARS-CoV-2 spike glycoprotein are known to drive antibody escape.²⁷ For example, recurrent deletions (Δ 141-144 and Δ 146, and Δ 243-244) in spike N-terminal domain (NTD) abolished its binding with neutralising antibody 4A8, and Δ 140 caused a 4-fold reduction in neutralisation titre.⁴⁸ The emergence of novel indels leading to variants of concern could be a challenge to vaccines and COVID-19 management.⁴⁹⁻⁵¹ This could be further exacerbated as sequencing efforts in many (developing) countries were minimal,^{43,44} but there seem to be disproportionately more variants, for example, in India. The viral diversity/variants may only be fully appreciated if there is better sequencing effort in these countries, many of which are reporting the emergence of variants of concern. For example, Resende et al⁵² have found convergent indels in the NTD of spike/SARS-CoV-2 lineages with mutations of concern circulating in Brazil, while Tegally et al⁵³ found Δ 242-244 in the SARS-CoV-2 variant of concern in South Africa. The recurrent emergence of insertions (between R214 and D215) in the NTD, and their progressive increase in multiple lineages, including VOC have been recently documented.⁵⁴ It is important to note that many indel positions are highly variable due to independent/multiple origins of indels^{27,53} and/or nearby substitutions that affect alignment. Indels have special relevance as they can fine-tune the 3D structure beyond point mutations and are known to occur in surface-exposed loops.²⁶ As the SARS-CoV-2 will be with us forever,⁵⁵ there is a need, equitably across countries to monitor the dynamics of variants, including indels.

Point mutations (or substitutions) tend to accumulate near indels.^{56,57} In fact, indels are the driving forces as heterozygosity of indels was proposed as mutagenic to surrounding sequences.⁵⁶ As indels are less constrained and have higher structural influence than substitutions,⁵⁸ they are frequently

under positive selection, for example, in cancer.⁵⁷ While the spike mutations in 'variants of concern' (VOC) were known to occur near indels,²⁹ here, we showed a large-scale relationship between indels and point mutations. GISAID clades are based on a statistical distribution of genome distances in phylogenetic clusters.⁵⁹ They have the evolutionary relationship of $S > L$ (and O, V) $> G > GH > GR$ (and GV) $> GRY$.⁶⁰ Advanced clades seem to have comparatively more mutations. Overall, mutations were more frequent in sequences with indels. This relation holds true even in variants of concern that already have extensive mutations.⁶¹ It is interesting and important to note that sequences with indels have several differentially abundant point mutations in VOC Delta, posing a global challenge. Mutations in RBD, in particular, were shown to affect ACE2 interaction. For example, deep mutational scanning of RBD found that most of the 3804 individual mutations were deleterious for ACE2 binding.⁶² However, RBD with mutations such as V367F, Y453F, and N501Y showed stronger interaction – faster association and slower dissociation rate – with ACE2^{63,64} and possibly had minimal effect on antibody neutralisation.⁶⁵ Shah et al.⁶⁶ showed that insertion of Gly at 482 hinders antibody neutralisation. However, the effect of indels on RBD and ACE2 interaction is not explored.

Despite numerous instances of indels, there were only a few instances of alterations of N-glycosylation sites in the spike protein. While some existing sites were modified, there were a few more instances of gain and loss of sites. Interestingly, all gains of sites in spike were of NXT type, which were known to be preferred by viral glycoproteins.^{4,35,67} However, given that some gains of sites were buried in the 3-D structure, they are unlikely to get selected/fixed. Proteins of other enveloped viruses, for example, haemagglutinin (HA) of influenza virus A/H1N1 (since 1918), A/H3N2 (since 1968), and recent A/H5N1 are all accumulating more N-glycosylation sites⁴ and/or modifying the existing sites.^{7,35} It is important to watch the dynamics of N-glycosylation sites in spike as SARS-CoV-2 transforms the vulnerabilities of its glycan shield.⁶⁸ For instance, the spike protein has 25% glycans by weight, which shield approximately 40% of the surface⁶⁹ as against 50% glycans by weight which shield 71% to 97% in gp120 of HIV-1, countering vaccine development and/or neutralisation by antibody.⁷⁰ On the other hand, loss of N-glycosylation sites has a selective disadvantage as removal of N331 and N343 drastically reduced infectivity, revealing the importance of glycosylation for viral infectivity.²⁰ While the SARS-CoV-2 spike utilises a glycan shield, it also modulates conformational dynamics of the receptor-binding domain by glycosylation. For example, deletion of sites by N165A and N234A mutations reduces spike binding to its receptor ACE2.³

To mention a key limitation of this study, we looked at the patterns of indels only in spike protein. Further, we do not give any explicit mechanistic insights into the emergence/dynamics of indels and their relationship with substitutions.

In conclusion, we show that SARS-CoV-2 is fine-tuning the spike with numerous indels. There seems to be a selective advantage as the proportions of indel variants steadily increased over time with similar trends across countries/variants. As many as 420 unique indel positions and 447 unique combinations of indels were present. Indels and point mutations are positively correlated and sequences with indels had significantly more point mutations. Despite their frequency, indels resulted in only minimal alteration of N-glycosylation sites.

Author Contributions

RSPR initiated the work and wrote the paper. All authors contributed and were involved in the revision.

Statement of Ethics

The work is in compliance with ethical standards. No ethical clearance was necessary.

ORCID iD

R Shyama Prasad Rao  <https://orcid.org/0000-0002-2285-6788>

Data Availability

The SARS-CoV-2 sequences and metadata used in this work are available upon registration, as per the terms of the Database Access Agreement, at GISAID (<https://www.gisaid.org/>).

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

- Zhou P, Yang X-L, Wang X-G, et al. Addendum: a pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;588:E6.
- worldometer. Worldometers.info. Accessed June 16, 2021. <https://www.worldometers.info/coronavirus/>
- Casalino L, Gaieb Z, Goldsmith JA, et al. Beyond shielding: the roles of glycans in the SARS-CoV-2 spike protein. *ACS Cent Sci*. 2020;6:1722–1734.
- Cui J, Smith T, Robbins PW, Samuelson J. Darwinian selection for sites of Asn-linked glycosylation in phylogenetically disparate eukaryotes and viruses. *Proc Natl Acad Sci USA*. 2009;106:13421–13426.
- Fischer W, Giorgi EE, Chakraborty S, et al. HIV-1 and SARS-CoV-2: patterns in the evolution of two pandemic pathogens. *Cell Host Microbe*. 2021;29:1093–1110.
- Firquet S, Beaujard S, Lobert P-E, et al. Survival of enveloped and non-enveloped viruses on inanimate surfaces. *Microbes Environ*. 2015;30:140–144.
- Rao RS, Wollenweber B. Subtle evolutionary changes in the distribution of N-glycosylation sequons in the HIV-1 envelope glycoprotein 120. *Int J Biol Sci*. 2010;6:407–418.
- Rao RSP, Foley BT, Xu D, et al. Evolutionary dynamics of N-glycosylation sites in hemorrhagic fever viral envelope proteins. *J Proteins Proteomics*. 2015;6:40.
- Zhang M, Gaschen B, Blay W, et al. Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. *Glycobiology*. 2004;14:1229–1246.
- Pavlović-Lazetić GM, Mitić NS, Tomović AM, Pavlović MD, Beljanski MV. SARS-CoV genome polymorphism: a bioinformatics study. *Genom Proteom Bioinform*. 2005;3:18–35.
- Woo PC, Huang Y, Lau SK, Yuen K-Y. Coronavirus genomics and bioinformatics analysis. *Viruses*. 2010;2:1804–1820.
- Badua CLDC, Baldo KAT, Medina PMB. Genomic and proteomic mutation landscapes of SARS-CoV-2. *J Med Virol*. 2021;93:1702–1721.
- Li X, Giorgi EE, Marichannegowda MH, et al. Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci Adv*. 2020;6:eabb9153.
- Lokman SM, Rasheduzzaman M, Salauddin A, et al. Exploring the genomic and proteomic variations of SARS-CoV-2 spike glycoprotein: a computational biology approach. *Infect Genet Evol*. 2020;84:104389.
- Peacock TP, Penrice-Randal R, Hiscox JA, Barclay WS. SARS-CoV-2 one year on: evidence for ongoing viral adaptation. *J Gen Virol*. 2021;102:001584.
- Pachetti M, Marini B, Benedetti F, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J Transl Med*. 2020;18:18–179.
- Mercatelli D, Giorgi FM. Geographic and genomic distribution of SARS-CoV-2 mutations. *Front Microbiol*. 2020;11:1800.
- Korber B, Fischer WM, Gnanakaran S, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*. 2020;182:812–827.
- Weissman D, Alameh M-G, de Silva T, et al. D614G spike mutation increases SARS CoV-2 susceptibility to neutralization. *Cell Host Microbe*. 2021;29:23–31.
- Li Q, Wu J, Nie J, et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell*. 2020;182:1284–1294.
- Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS One*. 2012;7:e46688.
- Lin M, Whitmire S, Chen J, Farrel A, Shi X, Guo JT. Effects of short indels on protein structure and function in human genomes. *Sci Rep*. 2017;7:9313.
- Mills RE, Luttig CT, Larkins CE, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res*. 2006;16:1182–1190.
- Chrisman BS, Paskov K, Stockham N, et al. Indels in SARS-CoV-2 occur at template-switching hotspots. *BioData Min*. 2021;14:20.
- Lee S, Won D, Kim C-K, et al. Novel indel mutation in the N gene of SARS-CoV-2 clinical samples that were diagnosed positive in a commercial RT-PCR assay. *Virus Res*. 2021;297:198398.
- Light S, Sagit R, Sachenkova O, Ekman D, Elofsson A. Protein expansion is primarily due to indels in intrinsically disordered regions. *Mol Biol Evol*. 2013;30:2645–2653.
- McCarthy KR, Rennick LJ, Nambulli S, et al. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science*. 2021;371:1139–1142.
- Garry RF, Gallaher WR. Naturally occurring indels in multiple coronavirus spikes. *Virological*. 2021. Accessed October 29, 2021. <https://virological.org/t/naturally-occurring-indels-in-multiple-coronavirus-spikes/560>
- Garry RF, Andersen KG, Gallaher WR, et al. Spike protein mutations in novel SARS-CoV-2 'variants of concern' commonly occur in or near indels. *Virological*. 2021. Accessed October 29, 2021. <https://virological.org/t/spike-protein-mutations-in-novel-sars-cov-2-variants-of-concern-commonly-occur-in-or-near-indels/605>
- Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data – from vision to reality. *EuroSurveillance*. 2017;22:30494.
- Maior ND, Walker C, Borges R, et al. Issues with SARS-CoV-2 sequencing data. *Virological*. 2020. Accessed October 29, 2021. <https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473>
- Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics*. 2010;26:680–682.
- Cock PJ, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25:1422–1423.
- McGinnis S, Madden TL. BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*. 2004;32:W20–W25.
- Rao RS, Bernd W. Do N-glycoproteins have preference for specific sequons? *Bioinformatics*. 2010;5:208–212.
- Touw WG, Baakman C, Black J, et al. A series of PDB-related databanks for everyday needs. *Nucleic Acids Res*. 2015;43:D364–D368.
- Malhis N, Jacobson M, Gsponer J. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. *Nucleic Acids Res*. 2016;44:W488–W493.
- Vijaymeena MK, Kavitha K. A survey on similarity measures in text mining. *Mach Learn Appl*. 2016;3:19–28.
- Agresti A. *An Introduction to Categorical Data Analysis*. 2nd ed. John Wiley & Sons; 2007:394.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc*. 1995;57:289–300.
- Maxmen A. Why some researchers oppose unrestricted sharing of coronavirus genome data. *Nature*. 2021;593:176–177.
- Van Noorden R. Scientists call for fully open sharing of coronavirus genome data. *Nature*. 2021;590:195–196.
- Colson P, Raoult D. Global discrepancies between numbers of available SARS-CoV-2 genomes and human development indexes at country scales. *Viruses*. 2021;13:775.
- Kaur B. Are all nations doing enough on SARS-CoV-2 sequencing? Clearly not. *Down to Earth*. 2021. Accessed October 29, 2021. <https://www.downtoearth.org.in/news/health/are-all-nations-doing-enough-on-sars-cov-2-sequencing-clearly-not-75064>

45. Meng B, Kemp SA, Papa G, et al. Recurrent emergence of SARS-CoV-2 spike deletion Δ H69/ Δ V70 and its role in the alpha variant B.1.1.7. *Cell Rep.* 2021;35:109292.
46. Xie X, Liu Y, Liu J, et al. Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K and N501Y variants by BNT162b2 vaccine-elicited sera. *Nat Med.* 2021;27:620–621.
47. Liu Z, Zheng H, Lin H, et al. Identification of common deletions in the spike protein of severe acute respiratory syndrome coronavirus 2. *J Virol.* 2020;94:e00790.
48. Harvey WT, Carabelli AM, Jackson B, et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol.* 2021;19:409–424.
49. Bian L, Gao F, Zhang J, et al. Effects of SARS-CoV-2 variants on vaccine efficacy and response strategies. *Expert Rev Vaccines.* 2021;20:365–373.
50. Gupta RK. Will SARS-CoV-2 variants of concern affect the promise of vaccines? *Nat Rev Immunol.* 2021;21:340–341.
51. Zhou D, Dejnirattisai W, Supasa P, et al. Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell.* 2021;184:2348–2361.e6.
52. Resende PC, Naveca FG, Lins RD, et al. The ongoing evolution of variants of concern and interest of SARS-CoV-2 in Brazil revealed by convergent indels in the amino (N)-terminal domain of the spike protein. *Virus Evol.* 2021;7:veab069.
53. Tegally H, Wilkinson E, Giovanetti M, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature.* 2021;592:438–443.
54. Gerdol M, Dishnica K, Giorgetti A. Emergence of a recurrent insertion in the N-terminal domain of the SARS-CoV-2 spike glycoprotein. *bioRxiv.* 2021. Accessed October 29, 2021. <https://www.biorxiv.org/content/10.1101/2021.04.17.440288v2.full.pdf>
55. Phillips N. The coronavirus is here to stay – here's what that means. *Nature.* 2021;590:382–384.
56. Tian D, Wang Q, Zhang P, et al. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature.* 2008;455:105–108.
57. Yang H, Zhong Y, Peng C, Chen JQ, Tian D. Important role of indels in somatic mutations of human cancer genes. *BMC Med Genet.* 2010;11:128.
58. Zhang Z, Wang Y, Wang L, Gao P. The combined effects of amino acid substitutions and indels on the evolution of structure within protein families. *PLoS One.* 2010;5:e14316.
59. Han AX, Parker E, Scholer F, Maurer-Stroh S, Russell CA. Phylogenetic clustering by linear integer programming (PhyCLIP). *Mol Biol Evol.* 2019;36:1580–1595.
60. Alm E, Broberg EK, Connor T, et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *EuroSurveillance.* 2020;25:2001410.
61. Wang P, Nair MS, Liu L, et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature.* 2021;593:130–135.
62. Starr TN, Greaney AJ, Hilton SK, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell.* 2020;182:1295–1310.e20.
63. Ou J, Zhou Z, Dai R, et al. V367F mutation in SARS-CoV-2 spike RBD emerging during the early transmission phase enhances viral infectivity through increased human ACE2 receptor binding affinity. *J Virol.* 2021;95:e0061721.
64. Tian F, Tong B, Sun L, et al. N501Y mutation of spike protein in SARS-CoV-2 strengthens its binding to receptor ACE2. *eLife.* 2021;10:e69091.
65. Bayarri-Olmos R, Rosbjerg A, Johnsen LB, et al. The SARS-CoV-2 Y453F mink variant displays a pronounced increase in ACE-2 affinity but does not challenge antibody neutralization. *J Biol Chem.* 2021;296:100536.
66. Shah M, Ahmad B, Choi S, Woo HG. Mutations in the SARS-CoV-2 spike RBD are responsible for stronger ACE2 binding and poor anti-SARS-CoV mAbs cross-neutralization. *Comput Struct Biotechnol J.* 2020;18:3402–3414.
67. Rao RS, Buus OT, Wollenweber B. Distribution of N-glycosylation sequons in proteins: how apart are they? *Comput Biol Chem.* 2011;35:57–61.
68. Watanabe Y, Berndsen ZT, Raghvani J, et al. Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nat Commun.* 2020;11:2688.
69. Grant OC, Montgomery D, Ito K, Woods RJ. Analysis of the SARS-CoV-2 spike protein glycan shield reveals implications for immune recognition. *Sci Rep.* 2020;10:14991.
70. Pancera M, Zhou T, Druz A, et al. Structure and immune recognition of trimeric prefusion HIV-1 env. *Nature.* 2014;514:455–461.