Evolutionary dynamics of indels in SARS-CoV-2 spike glycoprotein

R. Shyama Prasad Rao\textsuperscript{1,*}, Nagib Ahsan\textsuperscript{2,3}, Chunhui Xu\textsuperscript{4}, Lingtao Su\textsuperscript{4}, Jacob Verburgt\textsuperscript{5}, Luca Fornelli\textsuperscript{2,6}, Daisuke Kihara\textsuperscript{5,7}, Dong Xu\textsuperscript{4}

\textsuperscript{1}Biostatistics and Bioinformatics Division, Yenepoya Research Center, Yenepoya University, Mangaluru 575018, India
\textsuperscript{2}Department of Chemistry and Biochemistry, University of Oklahoma, Norman, Oklahoma 73019, USA
\textsuperscript{3}Mass Spectrometry, Proteomics and Metabolomic Core Facility, Stephenson Life Sciences Research Center, University of Oklahoma, Norman, Oklahoma 73019, USA
\textsuperscript{4}Department of Electrical Engineering and Computer Science, Informatics Institute, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, Missouri 65211, USA
\textsuperscript{5}Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907, USA
\textsuperscript{6}Department of Biology, University of Oklahoma, Norman, Oklahoma 73019, USA
\textsuperscript{7}Department of Computer Science, Purdue University, West Lafayette, Indiana 47907, USA

*Corresponding Author
E-mail: drrsprao@gmail.com
Abstract

SARS-CoV-2, responsible for the current COVID-19 pandemic that claimed over 4.2 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 that are currently 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 were steadily increasing 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, including both gain and loss, of N-glycosylation sites. As indels and point mutations are positively correlated and sequences with indels have significantly more point mutations, they have implications in the context of evolutionary dynamics of the SARS-CoV-2 spike glycoprotein.

Keywords: Computational proteomics, COVID-19, Molecular evolution, N-glycosylation sites, SARS-CoV-2, Selection, Sequence analysis
Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), responsible for the currently ongoing pandemic of coronavirus disease 2019 (COVID-19) (Zhou et al., 2020), has infected more than 198 million people and killed over 4.2 million (Anonymous, 2021). 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" (Casalino et al., 2020; Cui et al., 2009; Watanabe et al., 2019). Several genomic features such as point mutations, insertions and deletions (indels), and recombinations impart high diversity within enveloped viruses and profoundly increase their adaptability (Fischer et al., 2021). In fact, the envelop glycoproteins of these viruses are particularly variable and found to evolve quickly under selection pressure. As a result, the patterns and drivers of envelop glycoprotein variations in these viruses have been studied extensively (Cui et al., 2009; Rao and Wollenweber, 2010b; Rao et al., 2015; Zhang et al., 2004).

Given the nature of the pandemic, the genomic architecture and its evolutionary dynamics are being keenly explored in coronaviruses (Pavlovic-Lazetic et al., 2005; Woo et al., 2010) and in SARS-CoV-2 in particular (Badua et al., 2021; Li et al., 2020b; Lokman et al., 2020; Peacock et al., 2021). Pachetti et al. (2020) have shown the emergence of mutations in SARS-CoV-2 genome and RNA-dependent-RNA polymerase as a mutational hotspot. Mercatelli and Giorgi (2020) have analysed 48635 SARS-CoV-2 complete genomes and found on average 7.23 mutations per sample. Given its importance as a key interactor of angiotensin-converting enzyme 2 (ACE2) for host cell entry and as a target for neutralizing antibodies, spike glycoprotein in the envelop 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 infectivity (Korber et al., 2020) as well as increase susceptibility to neutralization (Weissman et al., 2021). Numerous other mutations in the spike glycoprotein have also been documented (Li et al., 2020a).

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 (Mercatelli and Giorgi, 2020). A key reason for indels being less common is that they are more deleterious due to frame-shifting compared to SNPs (Choi et al., 2012; Lin et al., 2017; Mills et al., 2006). As SARS-CoV-2 has been shown to accumulate indels (Fischer et al., 2021), we are only beginning to explore them and appreciate their myriad roles. For example, Chrisman et al. (2021) have looked at indels in the SARS-CoV-2 genome and mapped it to regions of discontinuous transcription breakpoints. Lee et al. (2021) 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 emphasized the genetic variability and rapid evolution of SARS-CoV-2 (Lee et al., 2021).

While indels have been explored predominantly at the genomic level (Badua et al., 2021; Mercatelli and Giorgi, 2020), they were less emphatically examined at the proteomic level and even less so in the spike glycoprotein. Despite their rarity, indels accelerate protein evolution (Light et al., 2013), 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 (McCarthy et al., 2021).
Garry and Gallaher (2021) have explored naturally occurring indels in multiple coronavirus spike proteins and provided evidence against a laboratory origin of SARS-CoV-2. Garry et al. (2021) 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 show that over 50% contained one 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 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 03, 2021) (Shu and McCauley, 2017). 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 1790224 sequences in the database at the time of access. However, as there were numerous quality issues with the data (Maio et al., 2020) many sequences were filtered out. For example, there were 465419 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 (Lin et al., 2017), disproportionately short sequences (for example, 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.

In the final set of 1311545 spike protein sequences, there were a total of 49118 unique sequences based on 100% identity cut-off (Huang et al., 2010) that included all possible variants. For each sequence, a pair-wise alignment with the reference (EPI_ISL_402124) was done using Biopython (Cock et al., 2009). The BLOSUM62 was used as the substitution matrix, and gap open and extension penalties were set at 11 and 1 respectively (McGinnis and Madden, 2004). Any gap in the query sequence was considered as a deletion and a gap in the reference was considered as an insertion (Choi et al., 2012). Finally, a mismatched residue was considered as 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 S1) (Cui et al., 2009; Rao and Wollenweber, 2010a; Zhang et al., 2004). 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 visualization/graphs were created in Microsoft Excel.

Structural analyses
The positions of indels and N-glycosylation sites were visualized on the 3D-structure of SARS-CoV-2 spike glycoprotein (PDB ID: 6VXX or 6XR8) using the Visual Molecular Dynamics (VMD) program (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 program which calculates an accessibility score (ranged from 0 to 277) from the 3D-structure (http://swift.cmbi.ru.nl/gv/dssp/; Touw et al., 2015). 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 (https://morf.msl.ubc.ca/index.xhtml) (Malhis et al., 2016). 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 (Table S2) (Vijaymeena and Kavitha, 2016).

Statistical analyses
Where required, a one-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 (Agresti, 2007). Correlations between indels and other variables (such as point mutations, accessibility scores, etc.) were measured using 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 two groups (for example, mutations in sequences with or without indels), and where required, the p-values were corrected for multiple comparison using Benjamini-Hochberg (BH) method (Benjamini and Hochberg, 1995). 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 Fig. 1A as a bar diagram. Overall, the distribution is similar for all sequences (n=1311545, filled bars) and unique sequences (n=49118, open bars). Over 50% of all sequences (inset pie chart) had at least one net indel. It is obvious that sequences with net deletions (50.5%) were far more common compared to sequences with net insertions (0.07%). Further, sequences containing three-residue net deletions were very frequent. A small number of sequences (0.16%) had net deletions of more than three residues. However, these proportions (36.4%, 0.33%, and 1.35%) are significantly different ($\chi^2=6991.5$, p=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 Fig. 1B, which shows an increasing trend. For example, in August 2020 the proportion of sequences with one to three deletions were about 1.4% which increased to 89.3% in May 2021. The proportions of sequences with insertions, and deletions of more than three residues were also increasing (Fig. S1A). A similar increasing trend is seen even if only unique sequences were considered (Fig. S1B). In May 2021, 76.3% of unique sequences have one to three deletions, and 1.8% of unique sequences with deletions of more than three residues. This increasing trend of sequences with indels holds true across countries (Fig. 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 (Fig. 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 (Fig. S2B).

In particular, just 0.74% of sequences were from India. However, it becomes 2.7% if only unique sequences were considered (Fig. S2C).

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

Incidence of indels along the spike sequence

Figure 2A shows the map of indels along the spike glycoprotein sequence. Figures S3A-C individually show the maps of indels for sequences with net deletions (n=18560), zero net indels (n=92), and net insertions (n=161). While the average number of insertions (2.1, n=161 unique sequences with net insertions) and deletions (2.8, n=18560 unique sequences with net 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) (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 two 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, 156-157, 241-243, etc. Among three-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 (Fig. S3D), there is variability in the position of three-residue deletion leading to the emergence of new deletion variants. Fig. 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 that 79% of sequences from Delta and Kappa variants of concern/interest (n=9133, currently represent 0.7% of the total) contain one or more deletions.

Based on the correlation, indels showed a low (but significant) preference for surface-exposed regions (τ=0.054 and p=0.027 for insertions, and τ=0.104 and p=1.2E-5 for deletions). Correlation with overall protein disorder was not significant (τ= 0.155 and p=0.56 for insertions, and τ=0.091 and p=0.66 for deletions) possibly because long disordered regions were very few and far apart. Correlation with shorter disordered regions (MoRF) was low but quite significant (τ= 0.122 and p=6.7E-8 for insertions, and τ=0.107 and p=7.1E-7 for deletions). On the 3D-
structure of SARS-CoV-2 spike glycoprotein (Fig. 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 (Table S2). Deletions, in particular, were also more frequent at the flanks of 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 (Fig. 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 ± standard deviation) point mutations compared to sequences without indels. Similar patterns were seen even when different GISAID clades or lineages were taken separately (Fig. 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 Fig. S5. Sequences with indels had, apart from D614G, six more very common mutations. There were 96 positions in sequences with indels (n=18813) and 101 in sequences without indels (n=30305, normalized 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 (τ=0.224, p=4.0E-25) between the distribution of indels and point mutations along the primary sequence (τ=0.136, p=1.9E-9 for insertions; τ=0.22, p=5.5E-24 for deletions). Many point mutations in VOC Delta were differentially abundant (two-proportion Z-test with BH correction) in sequences with indels compared to sequences without indels (Fig. S5C) and were more common in the N-terminal half where indels are also present (Fig. 2B).

**Effect of indels on N-glycosylation sites**

Based on the occurrence of sequons, there are 22 potential N-glycosylation sites (seven 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 instances of N-glycosylation sites that have been altered by the indels (in 65 unique sequences) is given in Table 1 (and S1) and their positions are shown in Fig. 3A-C. There were seven instances of gain of sites (Table 1, green) - due to insertions (for example near position 27, A---YT to AYTNYT), or deletions with substitution (for example near position 246, RSYLT to N--LT). There were also many more instances of loss of sites (Table 1, blue) - mostly due to deletions (for example at position 17, VNLT to V--T), but also due to insertions (for example 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, three sites (at around 290, 437, and 871) were completely buried in the 3-D structure (Fig. 3B-C). There were also a few instances of 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 (Tables 1 and S1).

**Discussion**

There was great interest and urgency to unravel the architecture of the SARS-CoV-2 genome (Zhou et al., 2020). Given the severity of the pandemic, there is currently an explosion of viral
sequencing bringing along the concerns of data accessibility and ownership (Maxmen, 2021; Noorden, 2021), data integrity/quality (Maio et al., 2020), and inequality of sequencing effort/data collection among countries (Colson and Raoult, 2021; Kaur, 2021). 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 (Badua et al., 2021; Garry et al., 2021; Korber et al., 2021; Li et al., 2020a; Lokman et al., 2020; Pachetti et al., 2020).

Indel variants were underexplored and unappreciated due to their relative rarity (Badua et al., 2021; Mercatelli and Giorgi, 2020; Mills et al., 2006), but it is becoming evident that they play key roles in the SARS-CoV-2 genome (Chrisman et al., 2021; Lee et al., 2021). One reason for indels being even less explored at the proteomic level was because they are primarily found in untranslated regions (Badua et al., 2021). 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, two very common deletions (ΔH69/ΔV70, primarily found in UK variants) were well known and shown to have recurrent emergence and transmission (Kemp et al., 2020). While deletions facilitate antibody escape, it was found that BNT162b2 vaccine-elicited sera can still neutralize 69/70 deletion variant (Xie et al., 2021). 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 neutralization (Weissman et al., 2021). Some independent deletions of five to seven residues were known to occur in and near the furin-like cleavage site (around residue position 681), and it was hypothesized that those deletions might be involved in viral infection (Liu et al., 2020). 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 Δ69-70 and Δ144 were present in the majority of the variants, there also seems to be an increasing trend for longer indels. Clearly, indels seem to have a selective advantage. Recurrent deletions in the SARS-CoV-2 spike glycoprotein are known to drive antibody escape (McCarthy et al., 2021). For example, recurrent deletions (Δ141-144 and Δ146, and Δ243-244) in spike N-terminal domain (NTD) abolished its binding with neutralizing antibody 4A8, and Δ140 caused four-fold reduction in neutralization titre (Harvey et al., 2021). The emergence of novel indels leading to variants of concern could be a challenge to vaccines and COVID-19 management (Bian et al., 2021; Gupta, 2021; Zhou et al., 2021). This could be further exacerbated as sequencing efforts in many (developing) countries were minimal (Colson and Raoult, 2021; Kaur, 2021), 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 courtiers, many of which are reporting the emergence of variants of concern. For example, Resende et al. (2021) have found convergent indels in the NTD of spike/SARS-CoV-2 lineages with mutations of concern circulating in Brazil, while Tegally et al. (2021) found Δ242-244 in the SARS-CoV-2 variant of concern in South Africa. Recurrent emergence of insertions (between R214 and D215) in the NTD, and their progressive increase in multiple lineages including VOC have been recently documented (Gerdol, 2021). It is important to note that many indel positions are highly variable due to independent/multiple origins of indels (McCarthy et al., 2021; Tegally et al., 2021) 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 (Light et al., 2013). As the SARS-CoV-2 will be with us forever (Phillips, 2021), there is a need, equitably across countries, to monitor the dynamics of variants, including indels.
Point mutations (or substitutions) tend to accumulate near the indels (Tian et al., 2008; Yang et al., 2010). In fact, indels are the driver forces as heterozygosity of indels was proposed as mutagenic to surrounding sequences (Tian et al., 2008). As indels are less constrained and have higher structural influence compared to substitutions (Zhang et al., 2010), they are frequently under positive selection, for example, in cancer (Yang et al., 2010). While the spike mutations in ‘variants of concern’ (VOC) were known to occur near indels (Garry et al., 2021), here we showed a large-scale relationship between indels and point mutations. Overall, mutations were more frequent in sequences with indels. This relation holds true even in variants of concern that already have extensive mutations (Wang et al., 2021). It is interesting and important to note that sequences with indels have several differentially abundant point mutations in VOC Delta, which is currently posing a global challenge.

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 (Cui et al., 2009; Rao and Wollenweber, 2010a; Rao et al., 2011). 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, hemagglutinin (HA) of influenza virus A/H1N1 (since 1918), A/H3N2 (since 1968), and recent A/H5N1 are all accumulating more N-glycosylation sites (Cui et al., 2009) and/or modifying the existing sites (Rao and Wollenweber, 2010a; 2010b). It is important to watch the dynamics of N-glycosylation sites in spike as SARS-CoV-2 transforms the vulnerabilities of its glycan shield (Watanabe et al., 2020). For instance, the spike protein has 25% glycans by weight which shield approximately 40% of the surface (Grant et al., 2020) as against 50% glycans by weight which shield 71-97% in gp120 of HIV-1 countering vaccine development and/or neutralization by antibody (Pancera et al., 2014). On the other hand, loss of N-glycosylation sites has selective disadvantage as removal of N331 and N343 drastically reduced infectivity, revealing the importance of glycosylation for viral infectivity (Li et al., 2020a). While the SARS-CoV-2 spike utilizes 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 (Casalino et al., 2020).

In conclusion, we show that SARS-CoV-2 is fine-tuning the spike with numerous indels. There seems to have a selective advantage as the proportions of indel variants were steadily increasing 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.

**Acknowledgments and 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.

**Conflicts of Interest**

The authors declare that there is no conflict of interest.
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/).

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

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

ORCID ID
R. Shyama Prasad Rao https://orcid.org/0000-0002-2285-6788
Nagib Ahsan https://orcid.org/0000-0003-2752-3795
Daisuke Kihara https://orcid.org/0000-0003-4091-6614
Dong Xu https://orcid.org/0000-0002-4809-0514

Supplemental information
Supplemental information for this article is available online.
References

Agresti A (2007). An introduction to categorical data analysis, 2nd ed. New York: John Wiley & Sons. Page 38.

Anonymous (2021). Worldometers.info (https://www.worldometers.info/coronavirus/) last accessed on 16-06-2021.

Badua CLDC, Baldo KAT, Medina PMB (2021). Genomic and proteomic mutation landscapes of SARS-CoV-2. Journal of Medical Virology 93:1702-1721.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society 57:289-300.

Bian L, Gao F, Zhang J, et al. (2021). Effects of SARS-CoV-2 variants on vaccine efficacy and response strategies. Expert Review of Vaccines 20:365-373.

Casalino L, Gaieb Z, Goldsmith JA, et al. (2020). Beyond shielding: The roles of glycans in the SARS-CoV-2 spike protein. ACS Central Science 6:1722-1734.

Choi Y, Sims GE, Murphy S, et al. (2012). Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 7:e46688.

Chrisman BS, Paskov K, Stockham N, et al. (2021). Indels in SARS-CoV-2 occur at template-switching hotspots. BioData Mining 14:20.

Cock PJA, Antao T, Chang JT (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422-1423.

Colson P, Raoult D (2021). Global discrepancies between numbers of available SARS-CoV-2 genomes and human development indexes at country scales. Viruses 13:775.

Cui J, Smith T, Robbins PW, Samuelson J (2009). Darwinian selection for sites of Asn-linked glycosylation in phylogenetically disparate eukaryotes and viruses. Proceedings of the National Academy of Sciences USA 106:13421-13426.

Fischer W, Giorgi EE, Chakraborty S, et al. (2021). HIV-1 and SARS-CoV-2: Patterns in the evolution of two pandemic pathogens. Cell Host and Microbe (https://doi.org/10.1016/j.chom.2021.05.012).

Garry RF, Andersen KG, Gallaher WR, et al. (2021). Spike protein mutations in novel SARS-CoV-2 ‘variants of concern’ commonly occur in or near indels. Virological (https://virological.org/t/spike-protein-mutations-in-novel-sars-cov-2-variants-of-concern-commonly-occur-in-or-near-indels/605).

Garry RF, Gallaher WR (2021). Naturally occurring indels in multiple coronavirus spikes. Virological (https://virological.org/t/naturally-occurring-indels-in-multiple-coronavirus-spikes/560).
Gerdol M (2021). Emergence of a recurrent insertion in the N-terminal domain of the SARS-CoV-2 spike glycoprotein (https://www.biorxiv.org/content/10.1101/2021.04.17.440288v2.full.pdf).

Grant OC, Montgomery D, Ito K, Woods RJ (2020). Analysis of the SARS-CoV-2 spike protein glycan shield reveals implications for immune recognition. Scientific Reports 10:14991.

Gupta RK (2021). Will SARS-CoV-2 variants of concern affect the promise of vaccines? Nature Reviews Immunology 21:340-341.

Harvey WT, Carabelli AM, Jackson B, et al. (2021). SARS-CoV-2 variants, spike mutations and immune escape. Nature Reviews Microbiology 19:409-424.

Huang Y, Niu B, Gao Y, et al. (2010). CD-HIT suite: A web server for clustering and comparing biological sequences. Bioinformatics 26:680-682.

Kaur B (2021). Are all nations doing enough on SARS-CoV-2 sequencing? Clearly not. Down to Earth (https://www.downtoearth.org.in/news/health/are-all-nations-doing-enough-on-sars-cov-2-sequencing-clearly-not-75064).

Kemp SA, Datir RP, Collier DA, et al. (2020). Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion ΔH69/ΔV70 (https://www.biorxiv.org/content/biorxiv/early/2020/12/15/2020.12.14.422555.full.pdf).

Korber B, Fischer WM, Gnanakaran S, et al. (2020). Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. Cell 182:812-827.

Lee S, Won D, Kim C-K, et al. (2021). Novel indel mutation in the N gene of SARS-CoV-2 clinical samples that were diagnosed positive in a commercial RT-PCR assay. Virus Research 297:198398.

Li Q, Wu J, Nie J, et al. (2020a). The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell 182:1284-1294.

Li X, Giorgi EE, Marichannegowda MH, et al. (2020b). Emergence of SARS-CoV-2 through recombination and strong purifying selection. Science Advances 6:eabb9153.

Light S, Sagit R, Sachenkova O, et al. (2013). Protein expansion is primarily due to indels in intrinsically disordered regions. Molecular Biology and Evolution 30:2645-2653.

Lin M, Whitmire S, Chen J, et al. (2017). Effects of short indels on protein structure and function in human genomes. Scientific Reports 7:9313.

Liu Z, Zheng H, Lin H, et al. (2020). Identification of common deletions in the spike protein of severe acute respiratory syndrome coronavirus 2. Journal of Virology 94:e00790-20.

Lokman SM, Rasheduzzaman M, Salauddin A, et al. (2020). Exploring the genomic and proteomic variations of SARS-CoV-2 spike glycoprotein: A computational biology approach. Infection, Genetics and Evolution 84:104389.

Maio ND, Walker C, Borges R, et al. (2020). Issues with SARS-CoV-2 sequencing data. Virological (https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473).
Malhis N, Jacobson M, Gsponer J (2016). MoRFchibi SYSTEM: Software tools for the identification of MoRFs in protein sequences. Nucleic Acids Research 44:W488-W493.

Maxmen A (2021). Why some researchers oppose unrestricted sharing of coronavirus genome data. Nature 593:176-177.

McCarthy KR, Rennick LJ, Nambulli S, et al. (2021). Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. Science 371:1139-1142.

McGinnis S, Madden TL (2004). BLAST: At the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Research 32:W20-W25.

Mercatelli D, Giorgi FM (2020). Geographic and genomic distribution of SARS-CoV-2 mutations. Frontiers in Microbiology 11:1800.

Mills RE, Luttig CT, Larkins CE, et al. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Research 16:1182-1190.

Noorden RV (2021). Scientists call for fully open sharing of coronavirus genome data. Nature 590:195-196.

Pachetti M, Marini B, Benedetti F, et al. (2020). Emerging SARS-CoV-2 mutation hotspots include a novel RNA-dependent-RNA polymerase variant. Journal of Translational Medicine 18:179.

Pancera M, Zhou T, Druz A, et al. (2014). Structure and immune recognition of trimeric prefusion HIV-1 Env. Nature 514:455-461.

Pavlovic-Lazetic GM, Mitic NS, Tomovic AM, et al. (2005). SARS-CoV genome polymorphism: A bioinformatics study. Genomics, Proteomics & Bioinformatics 3:18-35.

Peacock TP, Penrice-Randal R, Hiscox JA, Barclay WS (2021). SARS-CoV-2 one year on: Evidence for ongoing viral adaptation. Journal of General Virology 102:001584.

Phillips N (2021). The coronavirus is here to stay - Here’s what that means. Nature 590:382-384.

Rao RSP, Buus OT, Wollenweber B (2011). Distribution of N-glycosylation sequons in proteins: How apart are they? Computational Biology and Chemistry 35:57-61.

Rao RSP, Foley BT, Xu D, et al. (2015). Evolutionary dynamics of N-glycosylation sites in hemorrhagic fever viral envelope proteins. Journal of Proteins and Proteomics 6:40.

Rao RSP, Wollenweber B (2010a). Do N-glycoproteins have preference for specific sequons? Bioinformation 5:208-212.

Rao RSP, Wollenweber B (2010b). Subtle evolutionary changes in the distribution of N-glycosylation sequons in the HIV-1 envelope glycoprotein 120. International Journal of Biological Sciences 6:407-418.

Resende PC, Naveca FG, Lins RD, et al. (2021). 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
(https://www.medrxiv.org/content/10.1101/2021.03.19.21253946v1).

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

Tegally H, Wilkinson E, Giovanetti M, et al. (2021). Detection of a SARS-CoV-2 variant of concern in South Africa. Nature 592:438-443.

Tian D, Wang Q, Zhang P, et al. (2008). Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature 455:105-108.

Touw WG, Baakman C, Black J, et al. (2015) A series of PDB-related databanks for everyday needs. Nucleic Acids Research 43:D364-D368.

Vijaymeena MK, Kavitha K (2016). A survey on similarity measures in text mining. Machine Learning and Applications 3:19-28.

Wang P, Nai MS, Liu L, et al. (2021). Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. Nature 593:130-135.

Watanabe Y, Berndsen ZT, Raghwani J, et al. (2020). Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. Nature Communications 11:2688.

Watanabe Y, Bowden TA, Wilson IA, Crispin M (2019). Exploitation of glycosylation in enveloped virus pathobiology. Biochimica et Biophysica Acta 1863:1480-1497.

Weissman D, Alameh M-G, de Silva T, et al. (2021). D614G spike mutation increases SARS-CoV-2 susceptibility to neutralization. Cell Host & Microbe 29:23-31.

Woo PCY, Huang Y, Lau SKP, Yuen K-Y (2010). Coronavirus genomics and bioinformatics analysis. Viruses 2:1804-1820.

Xie X, Liu Y, Liu J, et al. (2021). Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K and N501Y variants by BNT162b2 vaccine-elicited sera. Nature Medicine 27:620-621.

Yang H, Zhong Y, Peng C, et al. (2010). Important role of indels in somatic mutations of human cancer genes. BMC Medical Genetics 11:128.

Zhang M, Gaschen B, Blay W, et al. (2004). Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. Glycobiology 14:1229-1246.

Zhang Z, Wang Y, Wang L, Gao P (2010). The combined effects of amino acid substitutions and indels on the evolution of structure within protein families. PLoS ONE 5:e14316.

Zhou D, Dejnirattisai W, Supasa P, et al. (2021) Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. Cell 184:2348-2361.e6.

Zhou P, Yang X-L, Wang X-G, et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270-273.
**Figure legends**

**Fig. 1.** (A) Distribution of sequences with indels in SARS-CoV-2 spike glycoprotein \((n=1311545)\). Over 50\% of sequences (inset pie chart) have at least one indel, with sequences containing three net deleted residues being very frequent. The pattern is similar (open bars) even if only unique sequences \((n=49118)\) 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.

**Fig. 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). See also Fig. S3.

**Fig. 3.** Indels versus point mutations. Bar plot shows the average number of point mutations (mean±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 sample size and/or proportion of sequences with indels are too small (lower panel of bar plot and sample size).

**Fig. 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 six gains of sites, three sites (at 290, 437, and 871) are completely buried in the 3-D structure.
Table 1. List of N-glycosylation site variants due to indels in SARS-CoV-2 spike protein.

| Accession ID | Len Site | Alignment¹ | Country     | Date⁵ | Clade | Lineage | Freq² |
|--------------|---------|------------|-------------|-------|-------|---------|-------|
| EPI_ISL_959416 1274 N149 | 145-YHKNNKKSWMES | Bangladesh | 20-01-2021 | GR    | B.1.1.25 | 2      |
| EPI_ISL_1134689 1274 N149 | 145-YHKNNKKSWMES | USA       | 13-02-2021 | GH    | B.1.2  | 3      |
| EPI_ISL_1164478 1271 N...# | 242-LALHN---LPGD | Turkey    | 15-02-2021 | GR    | B.1.1.189 | 2      |
| EPI_ISL_1167125 1273 N149 | 145-YHKNNKKSWMES | Senegal   | 27-01-2021 | O     | B.1    | 1      |
| EPI_ISL_1167127 1273 N149 | 145-YHKNNKKSWMES | Senegal   | 02-02-2021 | O     | B.1    | 1      |
| EPI_ISL_1272919 1274 N149 | 145-YHKNNKKSWMES | USA       | 25-02-2021 | GH    | B.1.2  | 1      |
| EPI_ISL_1359859 1268 N...# | 239-LALHN---LPGD | France    | 27-01-2021 | GRY   | B.1.1.7 | 1      |
| EPI_ISL_1591625 1274 N1134 | 1130-IGIVNNNTVYDP | Senegal | 25-03-2021 | GR    | B.1.1.222 | 2      |
| EPI_ISL_1791456 1278 N...# | 157-FRVIV-----YSS | USA       | 18-01-2021 | GH    | B.1.2  | 2      |
| EPI_ISL_1843929 1269 N17 | 13-SQCVN-7TRTO | Germany   | 10-04-2021 | GRY   | B.1.1.7 | 1      |
| EPI_ISL_1922075 1278 N...# | 867-DEMI-----AQ | Turkey    | 12-03-2021 | GH    | B.1.469 | 1      |
| EPI_ISL_2013547 1274 N61 | 57-PFFSNT---NFHA | USA       | 08-04-2021 | GH    | B.1.429 | 1      |
| EPI_ISL_2091257 1272 N1074 | 1070-AQKRE-SAPA | Mexico    | 12-04-2021 | S     | A.2.5  | 1      |
| EPI_ISL_2102041 1270 N...# | 262-TD-----YSLAL | Turkey    | 09-03-2021 | GRY   | B.1.1.7 | 1      |
| EPI_ISL_2102044 1270 N17 | 13-SQCVN-----TQ | Turkey    | 09-03-2021 | G     | B.1.36.17 | 1      |
| EPI_ISL_2262279 1267 N17 | 13-SQCVV-----TRO | Germany   | 10-05-2021 | GH    | B.1.351 | 1      |
| EPI_ISL_2269445 1271 N1134 | 1127-IGIVNNNTVYDP | USA       | 10-05-2021 | GRY   | B.1.1.7 | 1      |

¹Gain of sites is highlighted in green, loss of sites in blue, altered sites in orange, original sites in yellow, and non-sites in grey. ²Date of sample collection. ³Frequency out of 1311545 sequences. # - 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 Table S1.
Fig. 1.

A

B

C

- Ref
- -1 to -3

- without indels
- with indels

# of sequences

Proportion

Proportion

Sequence length

Time

Fig. 1.
Fig. 2.
Fig. 3.
Fig. 4.