Insertions in SARS-CoV-2 genome caused by template switch and duplications give rise to new variants of potential concern

Sofya K. Garushyants, Igor B. Rogozin, Eugene V. Koonin*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

*For correspondence: koonin@ncbi.nlm.nih.gov

Abstract

The appearance of multiple new SARS-CoV-2 variants during the winter of 2020-2021 is a matter of grave concern. Some of these new variants, such as B.1.351 and B.1.1.17, manifest higher infectivity and virulence than the earlier SARS-CoV-2 variants, with potential dramatic effects on the course of the COVID-19 pandemic. So far, analysis of new SARS-CoV-2 variants focused primarily on point nucleotide substitutions and short deletions that are readily identifiable by comparison to consensus genome sequences. In contrast, insertions have largely escaped the attention of researchers although the furin site insert in the spike protein is thought to be a determinant of SARS-CoV-2 virulence and other inserts might have contributed to coronavirus pathogenicity as well. Here, we investigate insertions in SARS-CoV-2 genomes and identify 141 unique inserts of different lengths. We present evidence that these inserts reflect actual virus variance rather than sequencing errors. Two principal mechanisms appear to account for the inserts in the SARS-CoV-2 genomes, polymerase slippage and template switch that might be associated with the synthesis of subgenomic RNAs. We show that inserts in the Spike glycoprotein can affect its antigenic properties and thus have to be monitored. At least, two inserts in the N-terminal domain of the Spike (ins246DSWG and ins15ATLRI) that were first detected in January 2021 are predicted to lead to escape from neutralizing antibodies whereas other inserts might result in escape from T-cell immunity.

Main text

The first SARS-CoV-2 genome was sequenced in January 2020. Since then, hundreds of thousands of virus genomes have been collected and sequenced. Comparative analysis of SARS-CoV-2 variants has provided for the identification of the routes of virus transmission [1–4], the selective pressure on different genes [5], and the discovery of new variants associated with higher infectivity [6–8]. In many cases, genome analysis only included search for point mutations, but some deletions also have been identified, such as del69-70 one of the characteristic mutations of B.1.1.7 and Cluster 5 [2,3]. Moreover, recently, recurrent deletions have been shown to drive antibody escape [9]. However, insertion sequences are mostly ignored, both during variant calling step and in the downstream analysis.

Although insufficiently studied, insertions appear to be crucial for beta-coronavirus evolution. Three insertions in the spike (S) glycoprotein and in the nucleoprotein (N) have been shown to differentiate highly pathogenic beta-coronaviruses (SARS-CoV-1, SARS-CoV-2 and MERS) from mildly pathogenic and non-pathogenic strains and suggested to be the key determinants of human coronaviruses pathogenicity [10]. The best characterized insert in SARS-CoV-2 is the PRRA tetrapeptide that so far is unique to SARS-CoV-2 and introduces a polybasic furin cleavage site into the S protein, enhancing its binding to the receptor [11,12]. Furthermore, the entire receptor-binding motif (RBM) domain of the S protein, most likely, was introduced into the SARS-CoV-2 genome via homologous recombination with coronaviruses from pangolins, which could have been a critical step in the evolution of SARS-CoV-2's ability to infect humans [13–15]. Similar frequent homologous recombination events among coronaviruses, and in particular in the sarbecovirus lineage, suggest that homologous recombination events is a common evolutionary mechanism that might have produced new coronavirus strains with changed properties on multiple occasions [15,16]. In contrast, non-homologous recombination in RNA viruses appears to be rarely detected, and its molecular mechanisms remains poorly understood [17].

In infected cells, beta-coronaviruses produce 5 to 8 major subgenomic RNAs (sgRNAs) [18,19]. Eight canonical sgRNAs are required for the expression of all encoded proteins of SARS-CoV-2. These sgRNAs are produced by joining the transcript of the 5' end of the genome (TRS site)

with the beginning of the transcripts of the respective open reading frames (ORFs) [20]. In addition, SARS-CoV-2 has been reported to produce multiple noncanonical sgRNAs, some of which include the TRS at 5' end, whereas others are TRS-independent [21,22].

Inserts in the SARS-CoV-2 genome are categorized in the CoV-GLUE database [23], and the preliminary results on systematic characterization of the structural variance and inserts in particular have been reported [24]. Forty structural variants including three inserts, three nucleotides long each, were discovered and shown to occur in specific regions of the SARS-CoV-2 genome. These variants were further demonstrated to be enriched near the 5' and 3' breakpoints of the TRS-independent transcriptome. Additionally, indels have been shown to occur in arms of the folded SARS-CoV-2 genomic RNA [24]. However, longer inserts that might have been introduced into the virus genome during SARS-CoV-2 evolution, to our knowledge, have not been systematically analyzed.

Here we report the comprehensive census of the inserts that during the evolution of SARS-CoV-2 over the course of the pandemic and show that at least some of these result from the virus evolution and not from experimental errors. These inserts are not randomly distributed along the genome, most being located in the 3'terminal half of the genome and co-localizing with 3' breakpoints of non-canonical (nc) sgRNAs. We show that the long insertions occur either as a result of the formation of nc-sgRNAs or by duplication of adjacent sequences. We analyze in detail the inserts in the S glycoprotein and show that at least two of these are located in a close proximity to the antibody-binding site in the N-terminal domain (NTD), whereas others are also located in NTD loops and might lead to antibody escape, and/or T cell evasion.

Identification of inserts in SARS-CoV-2 genomes

To compile a reliable catalogue of inserts in SARS-CoV-2 genome, we analyzed all the 498224 sequences present in the GISAID multiple genome alignment (compiled on February 23, 2021). From this alignment, we extracted all sequences that contained insertions in comparison with the reference genome. After this initial filtering, insertions were identified in 4468 genomes, with 296 unique events detected in total.

To eliminate insertions resulting from sequencing errors, we performed several additional filtering steps. First, we retained for further analysis only those insertions that were multiples of three, and thus would not lead to frameshifts, resulting in the reduction of the dataset to 157 unique events in 1030 genomes ranging in length from 3 to 195 nucleotides (Supplementary Table 1).

We then screened the Sequence Read Archive (SRA) database for the corresponding raw read data. We were able to obtain raw reads for 48 inserts (Supplementary Table 1), and verified the insertions in 32 cases. All insertions except one that we were unable to validate with the raw data analysis were of the length 3 or 6 nucleotides. We removed those unconfirmed events from our dataset that resulted in 141 events. Among these inserts, 65 were three nucleotides in length and 22 were of length 6, whereas the rest were longer (Figure 1a). We observed that inserts of lengths 3 and 6 had a distinct nucleotide composition with a substantial excess of uracil, at about 45%, whereas the composition of the longer inserts was similar to that of the SARS-CoV-2 genome average, with about 30% U (Figure 1b). The similar trend is observed for inserts verified by read data, although the available data is insufficient to demonstrate the significance of this trend for the 6 nucleotide inserts (Supplementary Figure 1). Thus, we split the collection of inserts into two categories, the short inserts of length 3 and 6 nucleotides, and the long inserts, which we analyzed separately.

We then checked whether inserts that were present in multiple genome sequences were monophyletic, that is, whether the genomes containing the same insertion formed a clade in the large phylogenetic tree containing more than 300,000 SARS-CoV-2 genomes (see Materials and Methods). Of the 37 short inserts identified in multiple genomes, 11 were found to be

monophyletic, and thus, apparently, originating from the single evolutionary event (Supplementary Table 2, Supplementary Figure 2). In 9 cases, identical insertions were observed in genomes submitted from the same laboratory, and mostly, on the same date, which implies that the genomes were sequenced and analyzed together, and makes it difficult to rule out a sequencing error. Interestingly, all 14 cases that can be confirmed by read data were not monophyletic. However, among the 18 long inserts that were found in multiple genomes, 13 were monophyletic, and only in five of these cases, sequences were from the same laboratory. What is more, all 4 long inserts present in multiple genomes and confirmed by read data were monophyletic (Supplementary Table 2, Supplementary Figures 3).

As the result of all these checks, the inserts detected in SARS-CoV-2 genomes fell into the following categories: 87 short inserts, among which 21 were confirmed by read data; and 54 long (at least, 9 nucleotides) inserts. We additionally classified the long inserts into four groups, in the order of increasing confidence: 29 singletons, 5 non-monophyletic inserts observed in multiple genomes, 9 monophyletic inserts observed in multiple genomes, and 11 inserts (7 singletons and 4 monophyletic ones), for which the insertions were confirmed by the raw sequence data analysis. We thus concluded that the 21 short inserts confirmed by read data and 25 long inserts that were detected in multiple genomes (monophyletic and not) and/or confirmed by raw sequencing data represented the most reliable insertion events that are currently observable throughout the evolution of SARS-CoV-2 (Supplementary Table 3).

Insertions are non-uniformly distributed along the SARS-CoV-2 genome

We found that the insertions were not randomly distributed along the genome, with most occurring in the 3'-terminal third of the genome (Figure 1c). Two, not necessarily mutually exclusive main hypothesis have been proposed on the origin of the short inserts (structural variants in the coronavirus genomes, namely, that they are associated with loops in the virus RNA structure or occur in the hotspots of template switch, at the breakpoints of TRS-independent transcripts [24]. To distinguish between these two mechanisms, we compared the distribution of 141 inserts along the SARS-CoV-2 genome with the distributions of structured

regions [25] and of template switch hotspots [22](Figure 1d). We detected a strong association of the insertions with the template switch hotspots ($r = 0.37$, p-value = 2.3×10^{-11}). Almost 30% of the inserts occurred within 5 nucleotides of a template switch hotspot, whereas less than 10% are expected by chance (Figure 1e). The observed pattern of inserts occurring in stems is the same as expected at random, indicating that inserts were not overrepresented in loops (Figure 1f). Both these observations held when we included in the analysis not all the 141 inserts, but only the 46 highly confident ones (Supplementary Figure 4). Thus, many inserts in the CoV-2 genomes are associated with template switch hotspots.

Short insertions in SARS-CoV-2 are generated by template sliding

The notable difference in nucleotide composition and different phyletic patterns of short and long inserts imply that the two types of insertions occur via different mechanisms. As pointed out above, the short insertions are rarely monophyletic, indicating that short U-rich sequences are inserted in the same position in the SARS-CoV-2 genome on multiple, independent occasions during virus evolution. Taken together, these observations suggest that such short insertions occur via template sliding (polymerase stuttering) on short runs of As or Us in the template (negative strand or positive strand, respectively) RNA [26–28] (Supplementary figure 5a). This could be either a biological phenomenon occurring during SARS-CoV-2 evolution, in case the errors are produced by stuttering of the coronavirus RdRP, or an artifact if the errors come from the reverse transcriptase or DNA polymerase that is used for RNA sequencing. It cannot be completely ruled out that these short inserts are a mix of biological and experimental polymerase errors. However, for the 19 inserts of length 3 that were confirmed by sequencing data analysis, we also detected the U enrichment. Those inserts were observed at high allele frequencies in the data (Supplementary Table 1), and thus, are unlikely to be experimental errors. Additionally, short inserts appear to be represented with the same frequency in SARS-CoV-2 genomes sequenced with different technologies, including Illumina MiSeq, NovoSeq and NextSeq and even Oxford Nanopore or IonTorrent (Supplementary Table 1). Furthermore, elevated rate of thymine insertion has not been reported as a common error of either Illumina or Oxford

159 Nanopore technology [29–32]. In contrast, production of longer transcripts and slow processing on
160 polyU tracts has been demonstrated for nsp12 (RdRP) of SARS-CoV-1 [33]. Additionally, the
161 RdRp complex of SARS-CoV lacking the proof-reading domain has been shown to
162 misincorporate more nucleotides compared with other viral polymerases [34]. Thus, a substantial
163 contribution of sequencing errors to the origin of short inserts in SARS-CoV-2 genomes appears
164 unlikely.

165

166 Long insertions in SARS-CoV-2 are caused by template switching and local
167 duplications

168 For in-depth analysis of the long inserts, we selected only the 25 high-confidence ones (see
169 above), which included 117 genomes and ranged in size from 9 to 27 nucleotides (Figure 2,
170 Supplementary Table 4).

171 Insertions were mostly observed in genome sequences from Europe (82) and US (25), and
172 originated from different laboratories that employed different protocols. Furthermore, these
173 events started to accumulate in early November 2020, and the median collection date of the
174 genomes containing the long inserts is January, 9 2021. Seven of the 25 reliable long insertions
175 are located in the S gene, which is significantly higher than expected by chance (Fisher exact test
176 p-value = 0.0165). The excess of inserts in the S gene suggests that their spread in the virus
177 population could be driven by positive selection for enhancement of the interaction of SARS-
178 CoV-2 with the host cells that could be conferred by the inserts.

179 The length of these high-confidence inserts allowed us to search for matching sequences both in
180 SARS-CoV-2 genomes and in other viruses. For 13 cases, we were unable to identify the
181 probable origin of the insertion. For four inserts, we detected a local duplication that most likely
182 gave rise to the insertion (Supplementary Table 4; Supplementary Figure 5b). Three out of these
183 four were found in multiple genomes and two of them were monophyletic although there was no
184 raw read data for any of these genomes. In one more case, the insertion was a singleton, but was
185 supported by raw data.

In 8 more cases, we detected significant matches in the SARS-CoV-2 genome, 6 in the coding strand and two in the complementary strand (Figure 2a; Supplementary Table 4). Among these 8 insertions, five were monophyletic (2 confirmed by raw data), and two more were singletons supported by raw data. The apparent origin of inserts from distant parts of the SARS-CoV-2 genomes implies template switch (Supplementary Figure 5c). We hypothesized that template switching occurs during the formation of the nc sgRNAs. To test this possibility, we compared the insert locations and the sites of the likely origin of the inserts with the available experimental data on the SARS-CoV-2 transcriptome [22]. Hotspots of template switching are characterized by polymerase "jumping" from one location on the genome to another, which yields shorter sgRNAs. As mentioned above, inserts tend to occur close to template switch hotspots, so for the inserts with a traceable origin, we additionally checked whether their sites of origin occurred close to the site of RdRp "jumping". Although the information on the SARS-CoV-2 transcriptome is limited, among the 8 cases we found that two insert sites were located within one end of the junction, whereas their corresponding sites of origin were within 100 nucleotides of the other side of the same junction (Figure 2a). To assess the significance of this finding, we performed two permutation tests (see Material and Methods), in one of which the real insertion positions were matched with start sites chosen randomly, whereas in the second one, both types of sites were selected at random. Both tests showed that the co-localization of the inserts with template switch junctions was significant (Figure 2 b,c).

Thus, high-confidence long inserts in the SARS-CoV-2 genome apparently originated either by local duplication or by template switch which, at least in some cases, seemed to be associated with nc sgRNA synthesis. Notably, the PRRA insert, the furin cleavage site that is one of characteristic features of SARS-CoV-2, resembles the long inserts analyzed here. Although this insert has a high GC-content compared to the genomic average of SARS-CoV-2, it falls within the GC-content range of the long inserts (Supplementary Figure 1b). Furthermore, this insert is located within 20 nucleotides of a template switch hotspot at position 22,582 [22]. Although we were unable to identify a statistically significant match that would allow us to map the origin of this insert to a particular location within the SARS-CoV-2 genome, it appears likely that this

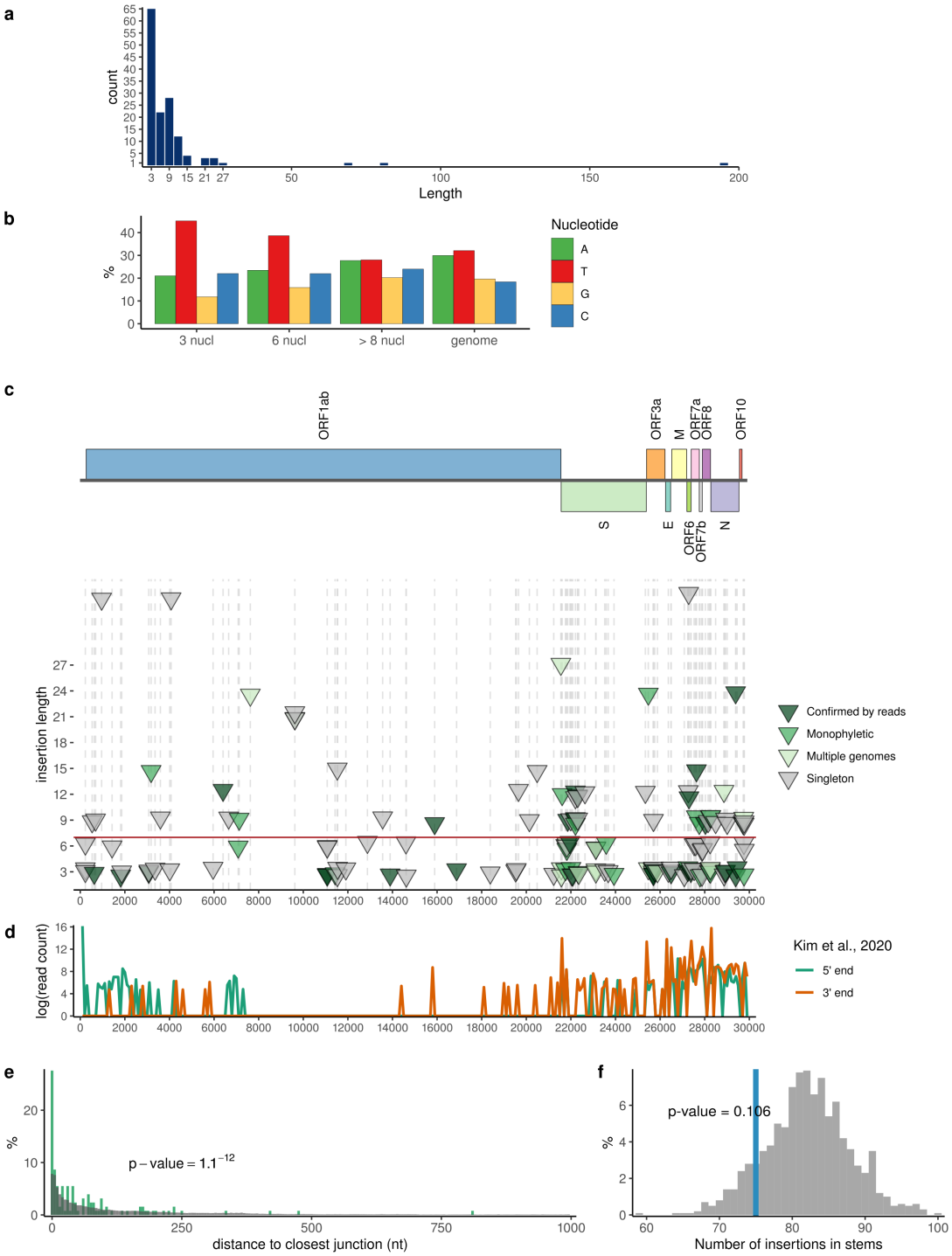
insert also originated by template switch, with subsequent substitutions erasing the similarity to the origin sequence.

Insertions in the S protein produce putative antibody escape variants

As indicated above, insertions are non-uniformly distributed along the SARS-CoV-2 genome (Figure 1c). In particular, among the 25 long inserts identified with high confidence, 7 were located in the S protein, suggesting that these inserts could persist due to their adaptive value to the virus. Three of the 7 inserts in S were observed in multiple genomes that formed compact clades in the phylogenetic tree, and ins214TDR in position 22,204 was strongly supported by raw sequencing data. In four more cases, the inserts were found in single genomes, but again, were strongly supported by raw data, and reached allele frequency close to one in the raw sequences, so these are highly unlikely to be artifacts (Supplementary Table 1).

All 7 long inserts in the S protein were located in the N-terminal domain (NTD), and four of these occurred in the same genome position, 22,004 (Figure 3). Compared to the receptor binding domain, the NTD initially attracted much less attention. Subsequently, however, multiple substitutions associated with variants of concern and observed in immunocompromised individuals with extended COVID-19 disease were identified in the NTD [2,35,36]. To evaluate potential functional effects of the inserts in the NTD, we mapped them onto the protein structure. All these inserts occurred on the protein surface (Figure 3), and two, ins15ATLRI and ins246DSWG, were located in an epitope that is recognized by antibodies obtained from convalescent plasma of recent COVID-19 patients [37]. Furthermore, ins246DSWG is located in the loop that is responsible for the interaction with the 4A8 antibody and potentially other antibodies (Figure 3a). Thus, at least these two insertions might be associated with the escape of SARS-CoV-2 variants from immune antibodies. The presence of multiple insertions in the same site, 22,004, suggests an important role of portion of the NTD in SARS-CoV-2 infection, especially, given that multiple deletion variants have been reported in the same region, 21971-22005 [9]. These insertions and ins98KAE are located in the neighboring loops, and given that the central region of the NTD has been shown to be essential for the virus interaction with CD4+

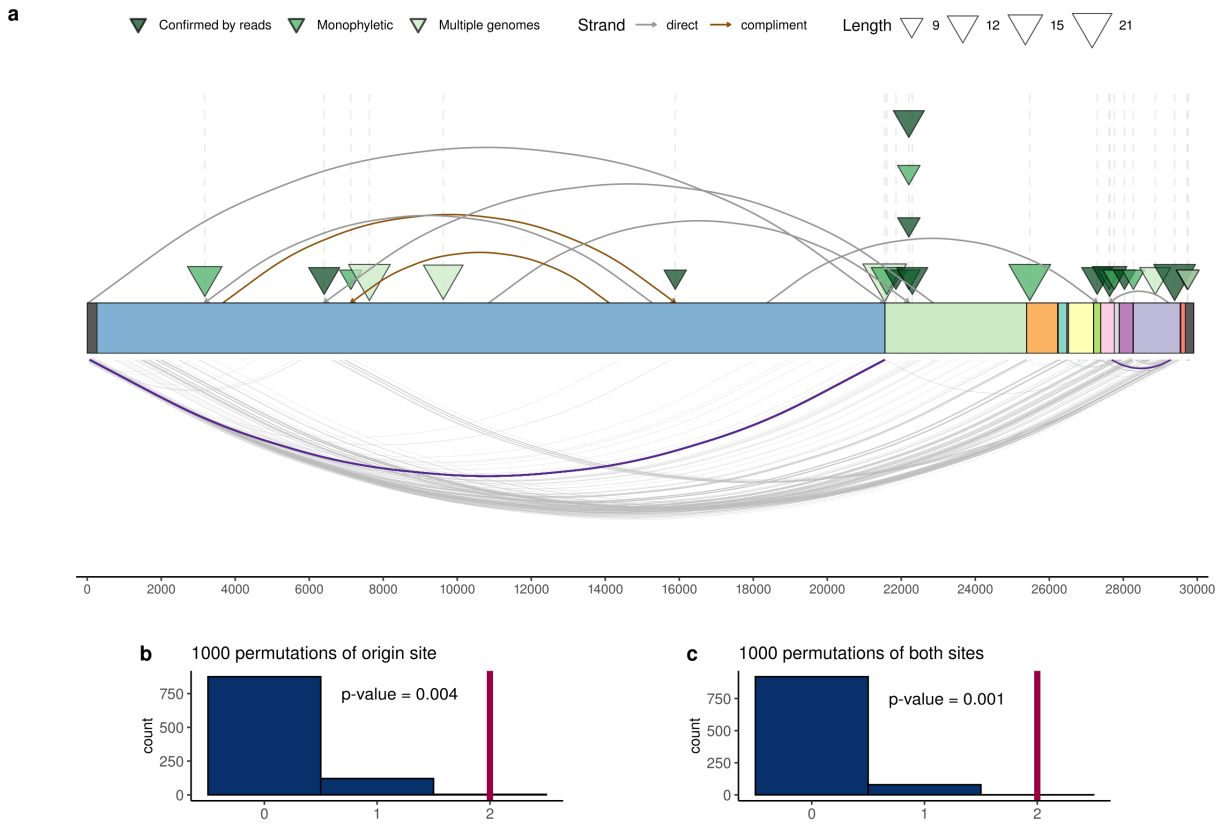
242 cells [38], could be associated with the escape from the T-cell immunity. Furthermore, recent
243 evidence suggests that this region contains an additional epitope for antibody binding [39]. Because
244 these insertions were detected only in recent samples, it appears that the respective variants have
245 to be further monitored.



246 **Figure 1. Inserions in SARS-CoV-2 genome.** (a) Distribution of insert lengths. (b) Nucleotide
247 composition of inserts of different lengths and full SARS-CoV-2 genome. (c) Distribution of
248 inserts along the genome. Each triangle represents one insertion event. The level of confidence in

249 each variant is represented by color: dark green, confirmed by sequencing read analysis; green,
250 monophyletic in the tree, no read data available; light green, observed multiple times, but not
251 monophyletic; grey, singletons (Supplementary Table 3). (d) Experimental data on SARS-CoV-2
252 transcriptome [22] showing template switch hotspots during the formation of sgRNAs. Lines
253 represent the coverage of junction sites by reads; green, 5' end of the junction; brown, 3' end of
254 the junction. (d) Distance from inserts to closest template switch hotspot site (green) compared
255 with random expectation (grey). Wilcoxon rank sum test p-value is provided. (e) The number of
256 inserts that occur in structured regions of SARS-CoV-2 genomic RNA (blue) compared with
257 random expectation (grey). Permutation test p-value is provided. The data on SARS-CoV-2
258 structure was obtained from [25].

259



260 **Figure 2. Long insertions possibly occur as a result of template switch and formation of nc**
261 **sgRNAs.** (a) Each triangle shows an independent insertion event, colored as in Fig. 1. Curves on
262 the upper side of the plot connect the insertion origin site and insertion position, brown color

indicates that the origin sequence is on the same strand, and grey color shows that the origin sequence is on complementary strand, Curves at the bottom of the plot represent the experimental data on sgRNAs from Kim et al. [22]. Curves highlighted in violet correspond to the two cases when insert and corresponding origin site co-occur with sgRNA junctions. The SARS-CoV-2 genes are colored as in Fig.1.

(b) permutation test, in which only the positions of the origins were randomly sampled 1000 times from the genome.

(c) permutation test, in which both ends were randomly sampled.

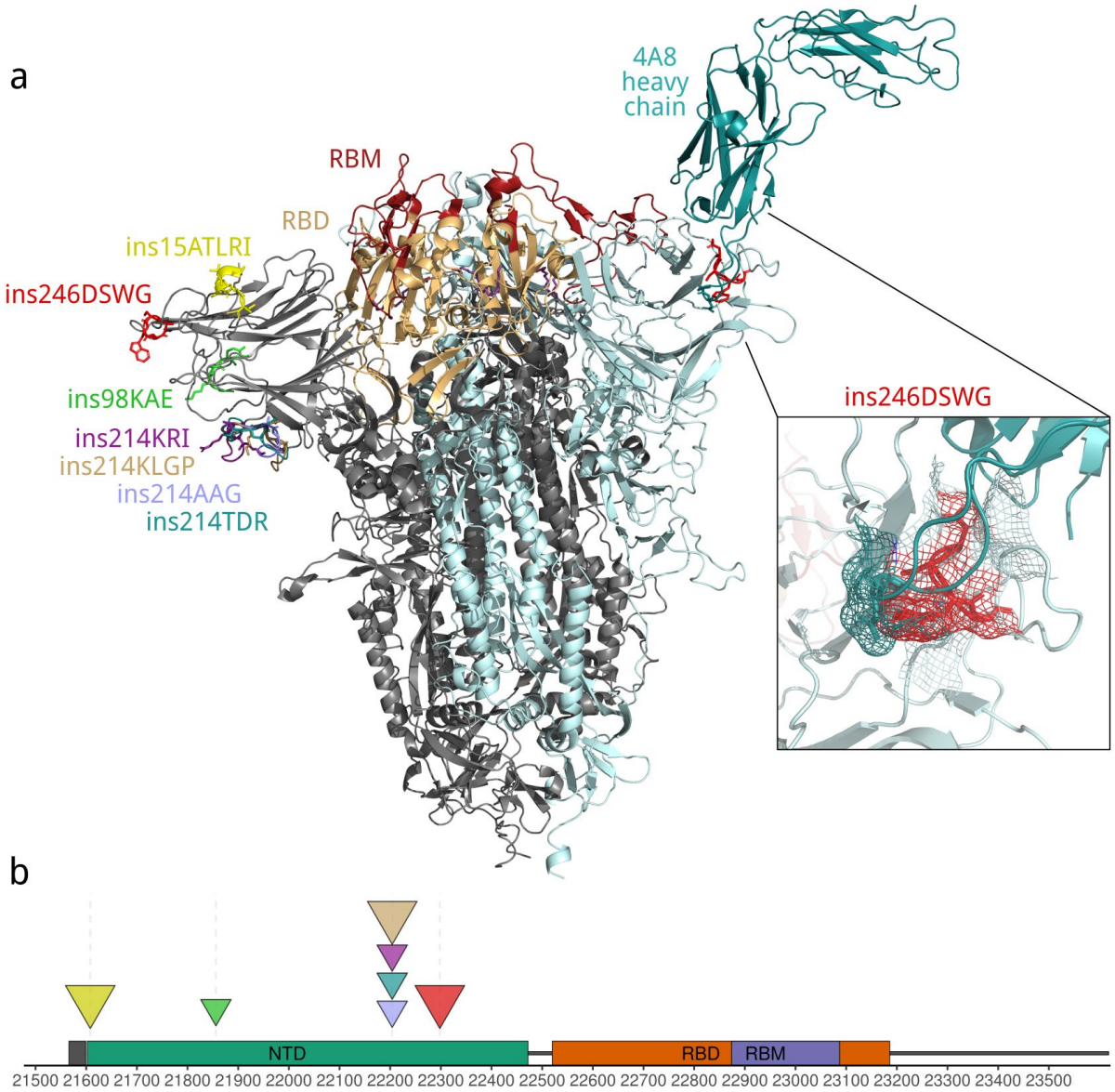


Figure 3. Location of insertion sites in SARS-CoV-2 S protein.

Two superimposed S protein structures are shown in grey (PDB ID: 7cn8) and in light blue (PDB ID: 7cl2). Wheat, receptor-binding domain (RBD), dark red, receptor binding motif (RBM), cyan, heavy chain of the 4A8 antibody (PDB ID: 7cl2). Each insertion is shown in a distinct color and in the sticks representation. The models were generated with the SWISS-model web server. (b) Location of insertions in the genome of SARS-CoV-2. Full description of insertions is provided in the Supplementary Tables 3 and 4.

Discussion

Although structural variation is an important driver of betacoronaviruses evolution, in the genome analysis during the current pandemics, part of the structural variations, namely, long insertions, to our knowledge, have not been systematically analyzed. This is a glaring omission given that insertions in the S and N protein appear to contribute to the betacoronavirus pathogenicity. In particular, the furin cleavage site inserted into the S protein seems to be crucial for SARS-CoV-2 pathogenicity [40,41]. Furthermore, betacoronaviruses are known to produce transcripts longer then their genomes [20], suggesting that insertions are a natural part of the life cycle of these viruses. Here we attempted a comprehensive identification and analysis of insertions in the SARS-CoV-2 protein-coding sequences that originated in the course of the current pandemic.

We found that short and longer insertions substantially differed with respect to their nucleotide compositions and mapping to the phylogenetic tree, suggesting that different mechanisms were at play. The short inserts were strongly enriched in U and in most cases occurred independently on the phylogenetic tree. It appears likely that these inserts occur as a results of RdRP slippage on short runs of A or U. Indeed, the observed excess of U in these inserts resembles the error profile of SARS-CoV-1 RdRp [33]. In contrast, the composition of the long inserts was close to that of the virus genome, and many of these insertions were found to be monophyletic, that is, these appear to be rare events that did not occur at nucleotide runs. Sequence analysis of the SARS-CoV-2 genomes indicates that these insertions occur either through polymerase slippage resulting in tandem duplication or more commonly, seem to have been triggered by illegitimate template switching associated with the formation of nc sgRNAs. For approximately half of the long insertions, we were unable to pinpoint the source of the inserted sequence and thus could not rule out that a third mechanism is involved. The PRRA insert that comprises the furin cleavage site in the S protein resembled the younger long inserts and likely originated by template switching as well, with the similarity to the origin sequence eliminated by subsequent point mutations, possibly, driven by positive selection.

Remarkably, long inserts are overrepresented in the S glycoprotein, particularly, in the NTD. Examination of the locations of these inserts on S protein structure strongly suggests that at least some of the inserts in the NTD result in the escape of the respective variants from neutralizing antibodies and, possibly, also from the T-cell response. The excess of insertions in the S protein is compatible with this protein being the principal area of virus adaptation. However, the location of most of the inserts in the NTD, as opposed to the RBD, is unexpected. Considering that all the detected inserts appeared at a relatively late stage of the pandemic, it seems likely that the structure of the RBD was already largely optimized for receptor binding at the onset of the pandemic such that most insertions would have a deleterious effect. In contrast, insertions into the NTD might allow the virus to escape immunity without compromising the interaction with the host cells. Thus, the insertion variants appear to merit monitoring, especially, at a time when vaccination might select for escape variants.

Materials and methods

GISAID data

The full multiple alignment of 498,224 complete SARS-CoV-2 genomes (version 0223) was downloaded from GISAID (https://www.gisaid.org/). From this alignment, we extracted all positions of insertions. An insertion was defined as addition of any number of columns compared to the SARS-CoV-2 reference genome (hCoV-19/Wuhan/Hu-1/2019 (NC_045512.2)). All insertions detected in the first and last 100 positions of the reference sequence were discarded as potentially erroneous. The alignment around the potential insertions was manually inspected. All the sequences that had more than two insertions were discarded, in order to avoid genomes with multiple sequencing errors. Information on the laboratory of origin, sequencing platform and consensus assembly methods (where available) was extracted from GISAID metadata.

Insertion validation from raw read data

Raw reads were downloaded from SRA database (https://www.ncbi.nlm.nih.gov/sra) with SRA Toolkit (Supplementary Table 1). The reads were mapped to the SARS-CoV-2 reference genome (NC_045512.2) with bowtie2 version 2.2.1 [42], either in pair mode of single read mode, depending on the type of data deposited to the SRA. The variants in each genome were called with LoFreq version 2.1.5 [43] as described in Galaxy (https://github.com/galaxyproject/SARS-CoV-2/blob/master/genomics/4-Variation/variation_analysis.ipynb). All insertions identified with LoFreq were visualized with the IGV software and manually inspected. An insertion was considered a real biological event if it had an allele frequency in reads of at least 60%, was located in the middle of the amplification fragment, and was covered by at least 100 reads.

Search for origins of long insertions

Search for putative duplications/template switch events with and without mismatches was performed against various datasets, for example, SARS-CoV-2 and closely related SARS-CoV genomes from bats and pangolin. Each insertion sequence was compared to all subsequences from a target sequence. All sequences with either the perfect match or with mismatches was retrieved (putative insertion source, PIS). If a PIS was located immediately upstream or downstream of an insertion sequence, it was annotated as duplication. If the PIS was located in any other positions, the template switch model was accepted as the best explanation of the observed insertion sequence.

To assess the significance of putative duplications and template switch events, we designed a sampling procedure to test a hypothesis that an insertion is not the result of spurious matches between an insertion sequence and corresponding PIS. Each insertion sequence was shuffled and scanned against datasets. We used the number of mismatches between an insertion sequence (observed or shuffled) and PIS as a weight W. A distribution of weights $W_{shuffled}$ was calculated for 1,000 shuffled insertion sequences. This distribution was used to calculate the probability $P(W_{observed} \geq W_{shuffled})$. This probability is equal to the number of shuffled insertion sequences with $W_{shuffled}$ equal to or smaller than $W_{observed}$. Small probability values ($P(W_{observed} \geq W_{shuffled}) \leq$

0.05) indicate statistical support for the hypothesis that the analyzed insertion sequence results from a duplication or a template switch.

Analysis of transcriptome data and genomic RNA structure

To compare insert locations with RNA secondary structure, we utilized the data from Huston et al., 2021 uploaded to github: https://github.com/pylelab/SARS-CoV-2_SHAPE_MaP_structure. For our analysis we used the data from full-length secondary structure map (.ct file). We considered all paired bases to be in stems, whereas those that are not paired were considered to be located in the loops. Thus, an insert was assigned to the stem if it appeared in a position that is known to be paired with another residue.

The data on the SARS-CoV-2 transcriptome was extracted from Kim et al., 2020 [22]. Pearson correlation coefficient between insertion locations and template switch hotspots was calculated for bins of size 100 nucleotides with cor.test() function in R version 3.6.3.

To calculate the random distributions for the analyses of distances to the closest junction and appearance of insertions in stems, we performed 1000 permutations, where each time the same number of genome positions was randomly selected from the genome as in the inserts dataset (141 for the analysis of all inserts, and 46 for the analysis of highly confident inserts). To compare the distributions of distances for the real data and random control, the Wilcoxon sum rank test was performed. In the case of inserts in stems, the p-value is the portion of cases in our simulation that had the same or smaller number of junctions as the real data.

To analyze whether long insertions coincide with template switch hotspots, we utilized the data on 5' and 3' ends of junctions from [22]. The junction end have to be located within 100 nucleotides from the insertion site and insertion source positions. To verify significance of these findings we performed two simulations. In first scenario the positions of inserts were fixed to the real positions from the data, but the locations of source sequences were randomly sampled 1000 times from the genome, in second scenario both source and insertion site positions were randomly sampled 1000 times. The p-value is the portion of cases in our simulation that have the same or larger number of junctions as the real data.

Phylogenetic analysis

The locations of the SARS-CoV-2 genomes selected for analysis on the phylogenetic tree of 302425 sequences from Genbank, COG-UK and CNCB (2021-02-10) that is available at UCSC was determined by UShER [44]. An insert was defined as monophyletic if it was observed in at least two genomes, and those genomes formed a stable clade on the phylogenetic tree or were located in the same stem cluster. The clades containing the genomes of interest were extracted and vizualized with ETE 3 package for Python [45].

Models of the spike protein and visualization

Models were build with SWISS-model [46], with the default parameters. The models shown on Figure 3 are based on two different initial PDB structures: Cryo-EM structure of PCoV_GX spike glycoprotein (PDB ID: 7cn8), and complex of SARS-CoV-2 spike glycoprotein with 4A8 antibody (PDB ID: 7cl2). The first structure was selected because it was the structure with the highest amino acid identity to the consensus sequence that cover most of the S protein. The obtained protein models were visualized with Open-Source PyMOL version 2.4.

Data availability

GISAID data used for this research are subject to GISAID's Terms and Conditions. SARS-CoV-2 genome sequences and metadata are available for download from GISAID EpiCoV™. The acknowledgements to all Originating and Submitting laboratories are provided in the Supplementary Table 5.

Custom R and Python scripts utilized for data analysis and visualization are available on github: https://github.com/garushyants/covid_insertions_paper

Acknowledgements

The authors are grateful to Koonin group members for useful discussions. The authors thank Elena Nabieva for suggestions about variant calling pipelines. This study was supported by the Intramural Research Program of the U.S. National Library of Medicine at the National Institutes of Health.

Authors contributions

IBR and EVK initiated the study. EVK designed and supervised the project. GSK and IBR collected the data. GSK extracted and verified the inserts, analyzed the data and built protein models. IBR and GSK analyzed the insertion mechanisms and the origins of inserts. GSK and EVK wrote the manuscript that was edited and approved by all authors.

416 # References

1. Candido, D. S. *et al.* Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* **369**, 1255–1260 (2020).
2. du Plessis, L. *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021).
3. Munnink, B. B. O. *et al.* Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* (2020) doi:10.1126/science.abe5901.
4. Komissarov, A. B. *et al.* Genomic epidemiology of the early stages of the SARS-CoV-2 outbreak in Russia. *Nat. Commun.* **12**, 649 (2021).
5. Martin, D. P. *et al.* The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. *medRxiv* 2021.02.23.21252268 (2021) doi:10.1101/2021.02.23.21252268.
6. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* (2021) doi:10.1126/science.abg3055.
7. Tegally, H. *et al.* Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv* 2020.12.21.20248640 (2020) doi:10.1101/2020.12.21.20248640.
8. Sabino, E. C. *et al.* Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. *The Lancet* **397**, 452–455 (2021).
9. McCarthy, K. R. *et al.* Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* **371**, 1139–1142 (2021).
10. Gussow, A. B. *et al.* Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proc. Natl. Acad. Sci.* **117**, 15193–15199 (2020).
11. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450–452 (2020).

12. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).
13. Graham, R. L. & Baric, R. S. Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission. *J. Virol.* **84**, 3134–3146 (2010).
14. Xiao, K. *et al.* Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* **583**, 286–289 (2020).
15. Li, X. *et al.* Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci. Adv.* **6**, (2020).
16. Boni, M. F. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* **5**, 1408–1417 (2020).
17. Simon-Loriere, E. & Holmes, E. C. Why do RNA viruses recombine? *Nat. Rev. Microbiol.* **9**, 617–626 (2011).
18. Sethna, P. B., Hung, S. L. & Brian, D. A. Coronavirus subgenomic minus-strand RNAs and the potential for mRNA replicons. *Proc. Natl. Acad. Sci.* **86**, 5626–5630 (1989).
19. Sawicki, S. G., Sawicki, D. L. & Siddell, S. G. A Contemporary View of Coronavirus Transcription. *J. Virol.* **81**, 20–29 (2007).
20. V'kovski, P., Kratzel, A., Steiner, S., Stalder, H. & Thiel, V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nat. Rev. Microbiol.* **19**, 155–170 (2021).
21. Nomburg, J., Meyerson, M. & DeCaprio, J. A. Pervasive generation of non-canonical subgenomic RNAs by SARS-CoV-2. *Genome Med.* **12**, 108 (2020).
22. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914-921.e10 (2020).
23. Singer, J., Gifford, R., Cotten, M. & Robertson, D. CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. (2020) doi:10.20944/preprints202006.0225.v1.

24. Chrisman, B. *et al.* Structural Variants in SARS-CoV-2 Occur at Template-Switching Hotspots. *bioRxiv* 2020.09.01.278952 (2020) doi:10.1101/2020.09.01.278952.
25. Huston, N. C. *et al.* Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. *Mol. Cell* **81**, 584-598.e5 (2021).
26. Kondrashov, A. S. & Rogozin, I. B. Context of deletions and insertions in human coding sequences. *Hum. Mutat.* **23**, 177–185 (2004).
27. Hausmann, S., Garcin, D., Delenda, C. & Kolakofsky, D. The versatility of paramyxovirus RNA polymerase stuttering. *J. Virol.* **73**, 5568–5576 (1999).
28. Zheng, H., Lee, H. A., Palese, P. & García-Sastre, A. Influenza A Virus RNA Polymerase Has the Ability To Stutter at the Polyadenylation Site of a Viral RNA Template during RNA Replication. *J. Virol.* **73**, 5240–5243 (1999).
29. Pfeiffer, F. *et al.* Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* **8**, 10950 (2018).
30. Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**, 90 (2018).
31. Ma, X. *et al.* Analysis of error profiles in deep next-generation sequencing data. *Genome Biol.* **20**, 50 (2019).
32. Dohm, J. C., Peters, P., Stralis-Pavese, N. & Himmelbauer, H. Benchmarking of long-read correction methods. *NAR Genomics Bioinforma.* **2**, (2020).
33. te Velthuis, A. J. W., Arnold, J. J., Cameron, C. E., van den Worm, S. H. E. & Snijder, E. J. The RNA polymerase activity of SARS-coronavirus nsp12 is primer dependent. *Nucleic Acids Res.* **38**, 203–214 (2010).
34. Ferron, F. *et al.* Structural and molecular basis of mismatch correction and ribavirin excision from coronavirus RNA. *Proc. Natl. Acad. Sci.* **115**, E162–E171 (2018).

35. Kemp, S. A. *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* 1–10 (2021) doi:10.1038/s41586-021-03291-y.

36. Sepulcri, C. *et al.* The longest persistence of viable SARS-CoV-2 with recurrence of viremia and relapsing symptomatic COVID-19 in an immunocompromised patient – a case study. *medRxiv* 2021.01.23.21249554 (2021) doi:10.1101/2021.01.23.21249554.

37. Cerutti, G. *et al.* Potent SARS-CoV-2 Neutralizing Antibodies Directed Against Spike N-Terminal Domain Target a Single Supersite. *bioRxiv* 2021.01.10.426120 (2021) doi:10.1101/2021.01.10.426120.

38. Tarke, A. *et al.* Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. *bioRxiv* 2020.12.08.416750 (2020) doi:10.1101/2020.12.08.416750.

39. Rosa, A. *et al.* SARS-CoV-2 can recruit a haem metabolite to evade antibody immunity. *Sci. Adv.* eabg7607 (2021) doi:10.1126/sciadv.abg7607.

40. Johnson, B. A. *et al.* Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. *Nature* 1–7 (2021) doi:10.1038/s41586-021-03237-4.

41. Papa, G. *et al.* Furin cleavage of SARS-CoV-2 Spike promotes but is not essential for infection and cell-cell fusion. *PLOS Pathog.* **17**, e1009246 (2021).

42. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

43. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).

44. Turakhia, Y. *et al.* Ultrafast Sample Placement on Existing Trees (UShER) Empowers Real-Time Phylogenetics for the SARS-CoV-2 Pandemic. *bioRxiv* 2020.09.26.314971 (2020) doi:10.1101/2020.09.26.314971.

45. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

46. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).