AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Spectrum

RESEARCH ARTICLE

Insertion-and-Deletion Mutations between the Genomes of SARS-CoV, SARS-CoV-2, and Bat Coronavirus RaTG13

Tetsuya Akaishi[a,b,c]

[a]Division of General Medicine, Tohoku University, Sendai, Japan
[b]Department of Education and Support for Regional Medicine, Tohoku University, Sendai, Japan
[c]COVID-19 Screening Test Center, Tohoku University, Sendai, Japan

ABSTRACT The evolutional process of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) development remains inconclusive. This study compared the genome sequences of severe acute respiratory syndrome coronavirus (SARS-CoV), bat coronavirus RaTG13, and SARS-CoV-2. In total, the genomes of SARS-CoV-2 and RaTG13 were 77.9% and 77.7% identical to the genome of SARS-CoV, respectively. A total of 3.6% (1,068 bases) of the SARS-CoV-2 genome was derived from insertion and/or deletion (indel) mutations, and 18.6% (5,548 bases) was from point mutations from the genome of SARS-CoV. At least 35 indel sites were confirmed in the genome of SARS-CoV-2, in which 17 were with ≥10 consecutive bases long. Ten of these relatively long indels were located in the spike (S) gene, five in nonstructural protein 3 (Nsp3) gene of open reading frame (ORF) 1a, and one in ORF8 and noncoding region. Seventeen (48.6%) of the 35 indels were based on insertion-and-deletion mutations with exchanged gene sequences of 7–325 consecutive bases. Almost the complete ORF8 gene was replaced by a single 325 consecutive base-long indel. The distribution of these indels was roughly in accordance with the distribution of the rate of point mutation rate around the indels. The genome sequence of SARS-CoV-2 was 96.0% identical to that of RaTG13. There was no long insertion-and-deletion mutation between the genomes of RaTG13 and SARS-CoV-2. The findings of the uneven distribution of multiple indels and the presence of multiple long insertion-and-deletion mutations with exchanged consecutive base sequences in the viral genome may provide insights into SARS-CoV-2 development.

IMPORTANCE The developmental mechanism of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) remains inconclusive. This study compared the base sequence one-by-one between severe acute respiratory syndrome coronavirus (SARS-CoV) or bat coronavirus RaTG13 and SARS-CoV-2. The genomes of SARS-CoV-2 and RaTG13 were 77.9% and 77.7% identical to the genome of SARS-CoV, respectively. Seventeen of the 35 sites with insertion and/or deletion mutations between SARS-CoV-2 and SARS-CoV were based on insertion-and-deletion mutations with the replacement of 7–325 consecutive bases. Most of these long insertion-and-deletion sites were concentrated in the nonstructural protein 3 (Nsp3) gene of open reading frame (ORF) 1a, S1 domain of the spike protein, and ORF8 genes. Such long insertion-and-deletion mutations were not observed between the genomes of RaTG13 and SARS-CoV-2. The presence of multiple long insertion-and-deletion mutations in the genome of SARS-CoV-2 and their uneven distributions may provide further insights into the development of the virus.

KEYWORDS bat coronavirus RaTG13, coronavirus disease 2019 (COVID-19), mutation, insertion-and-deletion mutation, severe acute respiratory syndrome coronavirus (SARS-CoV), severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)

Since the emergence of the coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the evolutionary process

Editor Maria Grazia Cusi, University of Siena

Address correspondence to t-akaishi@med.tohoku.ac.jp.
The authors declare no conflict of interest.
Received 1 March 2022
Accepted 6 May 2022
Published 6 June 2022

of the virus has been rigorously discussed (1–5). Elucidating the mechanism of its emergence may be important for not only effectively dealing with the current pandemic with intermittent appearance of consequential variant strains but also preventing the occurrence of future outbreaks of different emerging infectious diseases in the future. Currently, COVID-19 is considered a zoonotic disease, and its progenitor is considered to have emerged and maintained in bats as natural reservoir hosts (6), as the same with severe acute respiratory syndrome coronavirus (SARS-CoV) in 2002–2003 (7). This hypothesis is supported by the more than 95% similarity between the genome sequences of SARS-CoV-2 and bat coronavirus RaTG13, extracted from *Rhinolophus affinis* (8). The supposed progenitors of SARS-CoV-2 were also seen in Malayan pangolins, which also could have played some roles in maintaining the progenitor viruses in natural environments (9). Sometime between 2013 and 2019, the progenitor virus gained a functional polybasic furin cleavage site at the boundary region between S1 and S2 domains of the spike (S) gene through the insertion of four amino acid residues of "-PRRA-" (10). This produced the furin cleavage motif with "-RRAR-" at the S1/S2 boundary area, after which the furin or related proteases are believed to efficiently cleave the protein after entering the host cells (11). This insertion at the S1/S2 boundary has not been confirmed in the potential progenitor viruses including SARS-CoV, RaTG13, or pangolin coronaviruses (such as GD/P1L and GD/P2S), which is a strong rationale for the critical role of the acquired polybasic cleavage site in adapting to sustained human-to-human transmission of SARS-CoV-2 (12, 13). Furthermore, another plausible hypothesis for the emergence of SARS-CoV-2 is a highly variable genome sequence in the receptor-binding domain (RBD) of the S1 gene (14, 15), which could have contributed to the increased binding of S protein to the human receptor angiotensin-converting enzyme 2 (ACE2) and the enhanced immune evasion of human immunity to SARS-CoV-2 (16–18). Although the acquisition of these additional characteristics of the virus has been proposed as a promising scenario for SARS-CoV-2 development, the types and incidence of mutations across the whole viral genome behind the evolution of the virus have not been fully evaluated. This study aimed to gain a deep insight into the evolutionary process of lineage B SARS betacoronavirus by comparing the reference genome sequences and mutations occurring in both the coding and noncoding regions of SARS-CoV, RaTG13, and SARS-CoV-2.

RESULTS

Overall mutations between the genome of SARS-CoV-2 and SARS-CoV. Details of the insertion and/or deletion (indel) mutations and point mutations (substitutions) in the whole genome of SARS-CoV-2, compared to the genome sequence of SARS-CoV, are summarized in Table 1. In the 29,903 bases of the genomes of SARS-CoV-2, 5,548 (18.6%) bases were mutated based on point mutations and 1,068 (3.6%) bases were mutated based on indels. In total, the genome sequence of SARS-CoV-2 was 77.9% (23,287 of the 29,903 bases) identical to that of SARS-CoV. The rate of point mutation in the coding regions was relatively equal between the listed genes, but it was significantly higher than that in the noncoding regions ($P < 0.0001$, chi-square test). In contrast to the relatively even distribution of point substitution across the genome sequence, indels were disproportionally distributed across the coding regions. The rate of indels in each gene was the lowest with 0.0% in the envelope (E), open reading frame (ORF) 6, ORF7, nucleocapsid (N), and ORF10 genes, whereas it was the highest with 90.7% in ORF8 gene. At least 35 indel sites were confirmed in the genome of SARS-CoV-2, in which 17 indels were with $\geq$10 consecutive bases long. Ten of these relatively long indels were located in the S gene, five in nonstructural protein (Nsp) 3 gene of ORF1a, and one in ORF8 and noncoding region. Seventeen (48.6%) of the 35 indels were based on insertion-and-deletion mutations (that is, insertion and deletion mutations were simultaneously occurred at exactly the same position) with exchanged gene sequences of 7–325 consecutive bases. Furthermore, point mutation patterns between the whole genomes of SARS-CoV and SARS-CoV-2 were evaluated (Table 2).

TABLE 1 Composition of the indel mutations and point substitutions of the genome of SARS-CoV-2, compared to that of SARS-CoV[a]

Gene regions	Total no. of bases (Wuhan-Hu-1)	Base counts with indels	Rate of indel mutations	No. of bases after excluding indels	Base counts with point substitutions	Point substitution rate[b]
ORF1ab	21,290	366	0.0172	20,924	4,120	0.1970
S gene	3,822	315	0.0824	3,507	809	0.2307
ORF3	828	8	0.0097	820	196	0.2390
E gene	228	0	0	228	12	0.0526
M gene	669	3	0.0045	666	98	0.1471
ORF6	186	0	0	186	43	0.2312
ORF7	494	0	0	494	83	0.1680
ORF8	366	332	0.9071	34	4	0.1176
N gene	1,260	0	0	1,260	142	0.1127
ORF10	117	0	0	117	8	0.0684
All noncoding regions	643	44	0.0684	599	33	0.0551
Total	29,903	1,068	0.0357	28,835	5,548	0.1924

[a]The shown numbers of bases are for the genome sequence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) with the whole genome sequence of 29,903 bases. Different from the point substitutions in the coding regions, indels were uneven distributed across the whole genome sequences. SARS-CoV, severe acute respiratory syndrome coronavirus; ORF, open reading frame.
[b]Denominator of the rate of point substitution was the number of bases after excluding the insertion and/or deletion (indel) mutations in each of the gene regions.

C > T (18.7%) substitution was more frequent than other types of substitution, while G > C (2.2%) and C > G (2.2%) substitution frequencies were lower than those of the others. The base compositions within the indel mutations in the genomes of SARS-CoV and SARS-CoV-2 are summarized at the bottom of Table 2. The base composition within the indels was largely preserved between the two viruses.

Overall mutations between the genome of RaTG13 and SARS-CoV. Next, the whole genome sequences between bat coronavirus RaTG13 and SARS-CoV were compared. The genome sequence of RaTG13 was 77.7% (23,204 bases of the 29,855 bases) identical to the genome of SARS-CoV. In the 6,651 bases of RaTG13 with mutations from SARS-CoV, 1,019 bases (15.3%) were derived from indels and 5,632 (84.7%) were derived from point mutations. Most of the indels between the genomes of SARS-CoV-2 and SARS-CoV, as described in the previous section, were also confirmed between the genomes of RaTG13 and SARS-CoV.

Overall mutations between the genome of SARS-CoV-2 and RaTG13. Next, the whole genome sequences between SARS-CoV-2 and RaTG13 were compared. The genome sequence of SARS-CoV-2 was 96.0% (28,720 of the 29,903 bases) identical to that

TABLE 2 Mutation profiles between the genome sequence of SARS-CoV and SARS-CoV-2[a]

Coronavirus species	Adenine (A)	Thymine (T)	Guanine (G)	Cytosine (C)	Total
SARS-CoV (2002–2003) genome base composition					
Base count	8,476 (28.5%)	9,135 (30.7%)	6,186 (20.8%)	5,939 (20.0%)	29,736 (100.0%)
SARS-CoV-2 (2019) genome base composition					
Base count	8,954 (29.9%)	9,594 (32.1%)	5,863 (19.6%)	5,492 (18.4%)	29,903 (100.0%)
Single base substitution (vertical > horizontal bases) between SARS-CoV and SARS-CoV-2, base count (%)					
			(SARS-CoV-2)		
A (SARS-CoV)		617 (11.1%)	473 (8.5%)	251 (4.5%)	1,341 (24.2%)
T (SARS-CoV)	649 (11.7%)		195 (3.5%)	710 (12.8%)	1,554 (28.0%)
G (SARS-CoV)	722 (13.0%)	289 (5.2%)		124 (2.2%)	1,135 (20.5%)
C (SARS-CoV)	359 (6.5%)	1,039 (18.7%)	120 (2.2%)		1,518 (27.4%)
Total	1,730 (31.2%)	1,945 (35.1%)	788 (14.2%)	1,085 (19.6%)	5,548 (100.0%)
Base composition within the insertion and/or deletion mutation sites, base count (%)					
SARS-CoV	259 (28.9%)	261 (29.1%)	199 (22.2%)	177 (19.8%)	896 (100.0%)
SARS-CoV-2	349 (32.7%)	331 (31.0%)	185 (17.3%)	203 (19.0%)	1,068 (100.0%)

[a]The compositions of the 5,548 point mutations between the whole genome sequences of SARS-CoV in 2002–2003 and the subsequent SARS-CoV-2 in 2019 are shown. C > T (18.7%) substitution was more frequent than other types of substitution, whereas G > C (2.2%) and C > G (2.2%) substitutions were less frequent than others. SARS-CoV, severe acute respiratory syndrome coronavirus; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

FIG 1 Spike gene sequence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and mutation types in severe acute respiratory syndrome coronavirus (SARS-CoV). A comparison of S gene sequence between SARS-CoV and SARS-CoV-2 revealed that at least seven 14–245 base-long insertion-and-deletion mutations were concentrated in the N-terminal domain (NTD) of the S1 gene. The shown point substitution rates (26.5% for S1 and 18.9% for S2) are for the sequences excluding the insertion and/or deletion sites. If these excluded sites are included in the count, the mutation rate in the S1 gene increases up to 45.0%. Turquoise blue, yellow, blue, pink, and gray colors indicate reserved bases, substituted bases by point mutations, mutated bases with insertions, mutated bases based on insertion-and-deletion mutation with preserved base size, and mutated bases based on insertion-and-deletion mutations with changed base size, respectively.

of RaTG13. The mutated 1,183 bases were comprised of 1157 (97.8%) with point mutations and 26 (2.2%) with indels. Importantly, long insertion-and-deletion mutations, which were present between the genomes of SARS-CoV-2 and SARS-CoV, were absent between SARS-CoV-2 and RaTG13 genomes.

Mutations in SARS-CoV-2 spike gene. The confirmed mutation sites and mutation types in the S gene between SARS-CoV and SARS-CoV-2 are shown in Fig. 1. All indels in the S gene of SARS-CoV-2 were also confirmed in RaTG13. A total of 11 indels were confirmed in the S gene; 9 (81.8%) were based on insertion-and-deletion mutations and 2 (18.2%) were based on insertions. As can be seen, most of the indels were

FIG 2 Examples of insertion-and-deletion mutations in the N-terminal domain of spike gene. In the present study, gene mutations were categorized into the following four general subtypes: point mutation, insertion, deletion, and insertion-and-deletion mutation. With an insertion-and-deletion mutation, consecutive bases were exchanged by totally different sequences with the same or different base size. Most of the observed insertion-and-deletion mutations in spike N-terminal domain involved ≥10 consecutive bases and resulted in changed base sizes.

concentrated in the N-terminal domain (NTD) of S1 gene. The actual insertion-and-deletion mutations in the NTDs of SARS-CoV S and SARS-CoV-2 S are shown in Fig. 2. As can be seen, most of the insertion-and-deletion mutations were accompanied by changed base sizes. The substitution status of amino acids in SARS-CoV-2 spike RBD, compared with the amino acids in SARS-CoV spike RBD, is shown in Fig. 3. The amino acids of the aforementioned insertion-and-deletion mutation sites were replaced by totally different amino acid sequences (gray color). In the 292 amino acids in the spike NTD of SARS-CoV-2, 152 (52.1%) were substituted (76 based on point mutations and 76 based on indels) from those in SARS-CoV spike NTD. In the 197 amino acids in the spike RBD, 53 (26.9%) were substituted (36 based on point mutation and 17 based on indels). In the 588 amino acids in S2 domain, 60 (10.2%; all based on point mutations) were substituted. The list of the types and position of amino acids that are suggested to be critical for binding to ACE2 receptors in the RBD of SARS-CoV and SARS-CoV-2 spike RBD is shown in Table 3. Among the 19 listed amino acids in SARS-CoV-2 spike RBD, 11 were substituted from those in SARS-CoV spike RBD. The binding surface of ACE2 is known to be negatively charged, and substitutions to positively charged amino acids are generally considered to stabilize the RBD-ACE2 binding (19, 20). Three-dimensional molecular structures of the S protein in SARS-CoV and those in SARS-CoV-2 with closed ("down") and open ("up") conformations are shown in Fig. 4. Conformational changes in SARS-CoV-2 spike NTD (blue color) and RBD (yellow color), compared with those in SARS-CoV, can be seen. The RBD of SARS-CoV-2 are more centralized to the central pore in the axial view than those of SARS-CoV. The indel sites in SARS-CoV-2 S protein are shown as the consecutive amino acids in the red color. With the insertion-and-deletion mutations, the number of amino acid in the spike NTD was increased from 279 to 292 amino acids, and this could have partially contributed to the conformational change in the spike NTD in SARS-CoV-2.

Mutations in other structural genes of SARS-CoV-2. Mutations in other structural genes, such as E, membrane (M), and N, are shown in Fig. 5. Compared to the mutation status in the S gene, the prevalence of point substitution and indels in these three

FIG 3 Amino acid substitution status in SARS-CoV-2 spike protein compared with SARS-CoV. The substitution status of amino acids (AA) in SARS-CoV-2 spike protein compared with those in SARS-CoV spike protein is shown. As similar to the base substitution status, amino acid substitutions were also concentrated in the S1 domain, especially in the N-terminal domain. The insertion-and-deletion mutations in S1 gene resulted in totally different amino acid sequences with the replacements of consecutive amino acids.

genes was significantly lower between SARS-CoV and SARS-CoV-2. The overall mutation rates were 8.9%, 15.1%, and 11.6% for E, M, and N genes, respectively. The point substitution rates after excluding indel sites were 4.8%, 13.9%, and 11.6% for E, M, and N genes, respectively. The overall prevalence of mutations in each of the E, M, and N

TABLE 3 Amino acids critical for binding to ACE2 receptors in the RBD of SARS-CoV and SARS-CoV-2 spike RBD[a]

2002–2003 SARS-CoV		Corresponding AA in SARS-CoV-2		
SARS-CoV S (AA)	Characteristics of the R group	SARS-CoV-2 S (AA)	Characteristics of the R group	AA substitution
V404	Val: nonpolar, aliphatic	K417	Lys: positively charged	Yes
T433	Thr: polar, uncharged	G446	Gly: nonpolar, aliphatic	Yes
Y436	Tyr: nonpolar, aromatic	Y449	Tyr: nonpolar, aromatic	No
K439	Lys: positively charged	L452	Leu: nonpolar, aliphatic	Yes
Y442	Tyr: nonpolar, aromatic	L455	Leu: nonpolar, aliphatic	Yes
Y440	Tyr: nonpolar, aromatic	Y453	Tyr: nonpolar, aromatic	No
P462	Pro: polar, uncharged	A475	Ala: nonpolar, aliphatic	Yes
P470	Pro: polar, uncharged	E484	Glu: negatively charged	Yes
L472	Leu: nonpolar, aliphatic	F486	Phe: nonpolar, aromatic	Yes
N473	Asn: polar, uncharged	N487	Asn: polar, uncharged	No
Y475	Tyr: nonpolar, aromatic	Y489	Tyr: nonpolar, aromatic	No
N479	Asn: polar, uncharged	Q493	Gln: polar, uncharged	Yes
D480	Asp: negatively charged	S494	Ser: polar, uncharged	Yes
G482	Gly: nonpolar, aliphatic	G496	Gly: nonpolar, aliphatic	No
Y484	Tyr: nonpolar, aromatic	Q498	Gln: polar, uncharged	Yes
T486	Thr: polar, uncharged	T500	Thr: polar, uncharged	No
T487	Thr: polar, uncharged	N501	Asn: polar, uncharged	Yes
G488	Gly: nonpolar, aliphatic	G502	Gly: nonpolar, aliphatic	No
Y491	Tyr: nonpolar, aromatic	Y505	Tyr: nonpolar, aromatic	No

[a]Amino acids type and position for those suggested to be critical for binding to ACE2 receptor in the amino acid sequences of SARS-CoV and SARS-CoV-2 spike RBD are shown. Among the listed 19 amino acids in SARS-CoV-2 spike RBD, 11 were substituted from those in SARS-CoV. Because the binding surface of ACE2 is known to be negatively charged, substitution to positively charged amino acids may contribute to stabilize the RBD-ACE2 interaction. AA, amino acid; ACE2, angiotensin-converting enzyme 2; R group, side chain; RBD, receptor-binding domain.

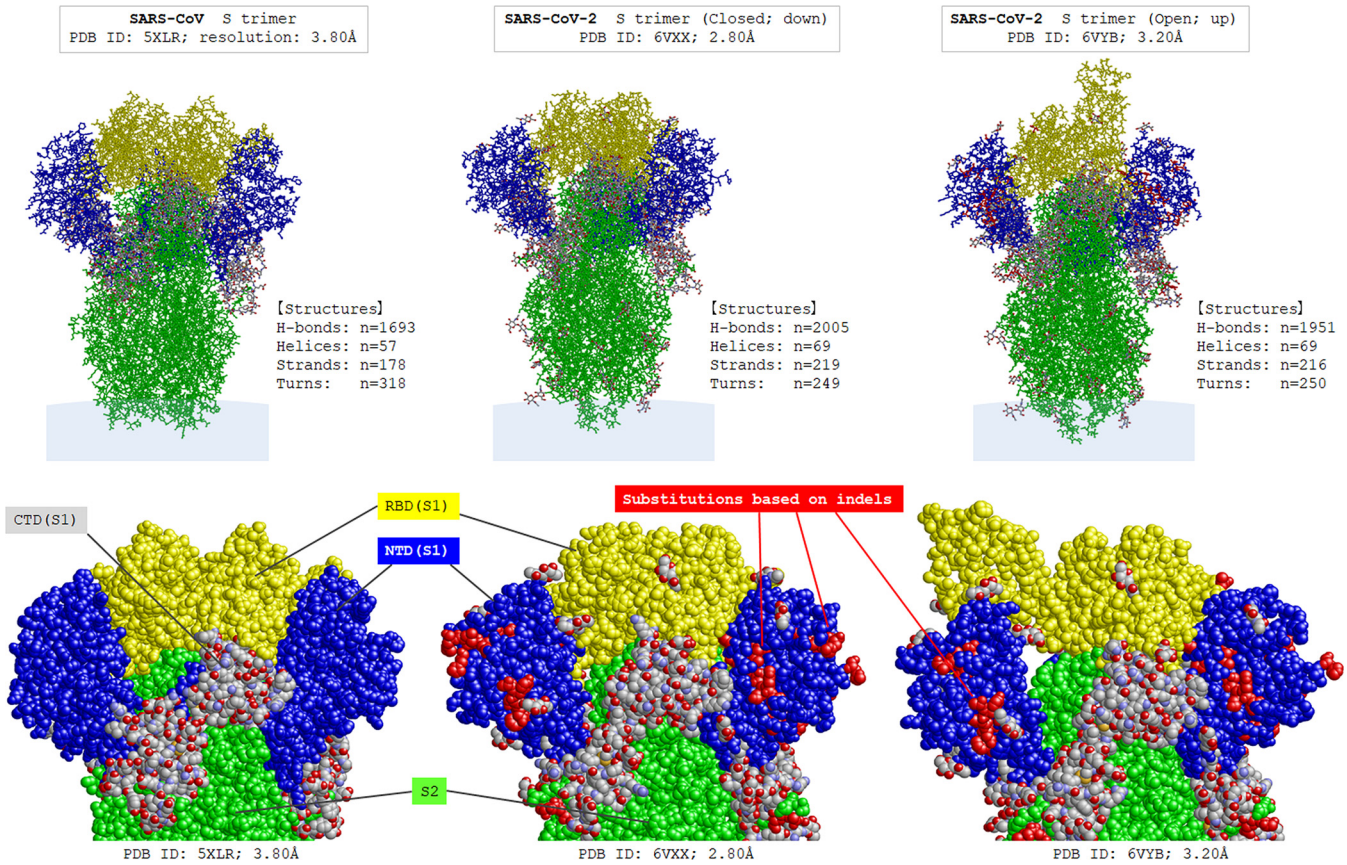

FIG 4 Molecular structures of the spike protein in SARS-CoV and SARS-CoV-2. Three-dimensional molecular structures of the S protein (closed state) in SARS-CoV and those in SARS-CoV-2 with closed and open states are shown. (Top) overall pictures of these proteins; (bottom) enlarged views of their S1 domains. The S1 NTD is shown in blue, receptor-binding domain (RBD) is in yellow, S2 domain is in green, and other subdomains including S1 CTD are in gray. The indel sites in SARS-CoV-2 S are shown as the consecutive amino acids colored in red. Conformational changes in SARS-CoV-2 spike NTD and RBD, compared with those in SARS-CoV, can be seen, and the RBD of SARS-CoV-2 are more centralized to the central pore than those of SARS-CoV. CTD, C-terminal domain; H-bonds, hydrogen bonds; indel: insertion and/or deletion; PDB, Protein Data Bank; S, spike.

genes was significantly lower than that in the S gene ($P < 0.0001$, for all three genes, chi-square test).

Mutations in nonstructural genes of SARS-CoV-2. Mutations in nonstructural genes between SARS-CoV and SARS-CoV-2 are shown in Fig. 6. Mutations in the ORF1ab genes were not shown in the figure, as the gene sizes were too large to show. Multiple long indels with 7–136 consecutive bases were concentrated in the Nsp3 gene of the ORF1a. Furthermore, almost the complete ORF8 gene was mutated via a 325 consecutive base long insertion–and-deletion mutation. This was the largest indel site in the entire genome of SARS-CoV-2.

Mutations in the noncoding regions. Mutations in the noncoding region upstream of each coding gene were studied and compared between the three betacoronaviruses. In the noncoding region upstream of the ORF1a gene in SARS-CoV-2 (that is, 1–265 bases), one short deletion site with the loss of three consecutive bases was confirmed. Other mutations in this region were all with point mutations. The point substitution rate in this noncoding region was 8.8%. The sequences in other noncoding regions in the three types of betacoronaviruses are shown in Fig. 7. Indels were confirmed in 5 of the 10 evaluated noncoding regions (upstream of ORF1a, S, M, ORF7a, and ORF10). Two indels occurred in the Kozak sequences (upstream of S and M genes), one of which gained the ideal Kozak consensus sequence motif of "-gcc-" in RaTG13 and SARS-CoV-2.

Details of the abnormally long indels. The indels with the involved base size of ≥10 consecutive bases in the whole genome, excluding the indels at the both ends of the genome (24 bases), are listed in Table 4. At least 16 indels with 10–325 consecutive

2019 SARS-CoV-2 (Wuhan-Hu-1) E gene (26245 - 26472 base; NCBI RefSeq: NC_045512.2) Substitution rate: 4.8%

atgtactcattcgtttcggaagagacaggtacgttaatagttaatagcgtacttctttttcttgctttcgtggtattcttgctagttacactagccatccttactgcgc
ttcgattgtgtgcgtactgctgcaatattgttaacgtgagtcttgtaaaaccttcttttacgtttactctcgtgttaaaaatctgaattcttctagagttcctgatct
tctggtctaa

2019 SARS-CoV-2 (Wuhan-Hu-1) M gene (26523 - 27191 base; NCBI RefSeq: NC_045512.2) Substitution rate: 14.7%

atggcagattccaacggtactattaccgttgaagagcttaaaagctccttgaacaatggaacctagtaataggtttcctattccttacatggatttgtcttctacaat
ttgcctatgccaacaggaataggtttttgtatataattaagttaattttcctctggctgttatggccagtaactttagcttgttttgtgcttgctgctgtttacagaat
aaattggatcaccggtggaattgctatcgcaatggcttgtcttgtaggcttgatgtggctcagctacttcattgcttctttcagactgtttgcgcgtacgcgttccatg
tggtcattcaatccagaaactaacattcttctcaacgtgccactccatggcactattctgaccagaccgcttctagaaagtgaactcgtaatcggagctgtgatccttc
gtggacatcttcgtattgctggacaccatctaggacgctgtgacatcaaggacctgcctaaagaaatcactgttgctacatcacgaacgctttcttattacaaattggg
agcttcgcagcgtgtagcaggtgactcaggttttgctgcatacagtcgctacaggattggcaactataaattaaacacagaccattccagtagcagtgacaatattgct
ttgcttgtacagtaa

2019 SARS-CoV-2 (Wuhan-Hu-1) N gene (28274 - 29533 base; NCBI RefSeq: NC 045512.2) Substitution rate: 11.6%

atgtctgataatggaccccaaaatcagcgaaatgcaccccgcattacgtttggtggaccctcagattcaactggcagtaaccagaatggagaacgcagtggggcgcgat
caaaacaacgtcggccccaaggtttacccaataatactgcgtcttggttcaccgctctcactcaacatggcaaggaagacccttaaattccctcgaggacaaggcgttcc
aattaacaccaatagcagtccagatgaccaaattggctactaccgaagagctaccagacgaattcgtggtggtgacggtaaaatgaaagatctcagtccaagatggtat
ttctactacctaggaactgggccagaagctggacttccctatggtgctaacaaagacggcatcatatgggttgcaactgagggagccttgaatacaccaaaagatcaca
ttggcacccgcaatcctgctaacaatgctgcaatcgtgctacaacttcctcaaggaacaacattgccaaaaggcttctacgcagaagggagcagaggcggcagtcaagc
ctcttctcgttcctcatcacgtagtcgcaacagttcaagaaattcaactccaggcagcagtaggggaacttctcctgctagaatggctggcaatggcggtgatgctgct
cttgctttgctgctgcttgacagattgaaccagcttgagagcaaaatgtctggtaaaggccaacaacaacaaggccaaactgtcactaagaaatctgctgctgaggctt
ctaagaagcctcggcaaaaacgtactgccactaaagcatacaatgtaacacaagcttcggcagacgtggtccagaacaaacccaaggaaatttttggggaccaggaact
aatcagacaaggaactgattacaaacattggccgcaaattgcacaatttgcccccagcgcttcagcgttcttcggaatgtcgcgcattggcatggaagtcacaccttcg
ggaacgtggttgacctacacaggtgccatcaaattggatgacaaagatccaaatttcaaagatcaagtcattttgctgaataagcatattgacgcatacaaaacattcc
caccaacagagcctaaaaaggacaaaaagaagaaggctgatgaaactcaagccttaccgcagagacagaagaaacagcaaactgtgactcttcttcctgctgcagattt
ggatgatttctccaaacaattgcaacaatccatgagcagtgctgactcaactcaggcctaa

FIG 5 Sequences of structural proteins in SARS-CoV-2 and the mutations in SARS-CoV. The shown point substitution rates are for the sequences excluding the gray-colored insertion–deletion sites (sequences with preserved bases or point substitutions). Turquoise blue, yellow, and blue colors indicate reserved bases, bases substituted by point mutations, and bases with insertions, respectively.

bases were confirmed; 14 sites with insertion-and-deletion and 2 sites with insertions. Five sites were located in the Nsp3 gene of ORF1a gene, nine in the S1 gene, one in the S1/S2 boundary area, and another in the ORF8 gene. The possibility of inversion or duplication for the sequence of the insertion-and-deletion sites was checked, but it was not likely to explain the observed mutations. The distribution of the indel sites in the whole genome of SARS-CoV-2 is shown in Fig. 8a. Below the panel, line graphs of the moving average ($\pm$50 bases) for point substitution rates in each base position of SARS-CoV-2 genome are superimposed, when compared to SARS-CoV (Fig. 8b, top) or RaTG13 (Fig. 8b, bottom). The indels were concentrated in the gene regions with high point substitution rates and disproportionally distributed across the genome, suggesting the presence of evolutionary selection pressure behind the uneven distributions of the indels across the genome of SARS-CoV-2. Furthermore, the line graph of point mutation rate and the distribution of indels between the genomes of RaTG13 and SARS-CoV are shown in Fig. 8c, which were largely the same with those between the genomes of SARS-CoV-2 and SARS-CoV.

DISCUSSION

Generally, the incidence of replication error during RNA virus proliferation increases up to 10^{-5} to 10^{-4} errors per base replication (21, 22). This is because most virus genome-encoded RNA-dependent RNA polymerases cannot repair the errors themselves. The SARS-CoV-2 genome contains genes encoding nonstructural protein Nsp12 and Nsp14, which collaboratively function to repair replication errors; thus, the estimated

FIG 6 Sequences of ORF3a–ORF10 in SARS-CoV-2 and the mutations in SARS-CoV. The mutation rate in the ORF10 gene was significantly lower than that in other ORF family genes. Almost the complete ORF8 was substituted by a 333 consecutive base-long insertion-and-deletion mutation. Turquoise blue, yellow, blue, and gray colors indicate reserved bases, bases substituted by point mutations, bases with insertion, and mutated bases based on insertion-and-deletion mutations with changed base size, respectively. ORF, open reading frame.

mutation rate of SARS-CoV-2 is currently approximately 10^{-6} to 10^{-5} errors per base replication (23–25). The current estimate of mutation rate in SARS-CoV-2 after adjusting for the effects of evolutionary selection was approximately 10^{-3} mutations per site per year (24, 26, 27), which was comparable to that in SARS-CoV (28). Based on this estimation, the observed mutation rate in this study across the coding regions of structural genes and noncoding regions between SARS-CoV and SARS-CoV-2 or RaTG13 seems to be significantly high as a natural evolutionary process, if we simply regard that SARS-CoV is a progenitor of SARS-CoV-2. This study revealed that at least 17 of the 35 indel sites were based on insertion-and-deletion mutations and were concentrated in the NTD of the S1, Nsp3 gene of the ORF1a, and ORF8 genes. The comparison of the three-dimensional molecular structures between the NTD of the S1 gene in SARS-CoV and that in SARS-CoV-2 suggested that the conformation of the domain was largely changed by the indels, possibly further affecting the conformation of the adjacent RBD. Although the exact molecular effect of this conformational change to the binding capacity of the RBD is uncertain, the conformational change could have affected the binding capacity of S protein against ACE2 in SARS-CoV-2, compared to that in SARS-CoV (29, 30). The strength of RBD-ACE2 binding, evaluated by the dissociation constant, showed discrepancies between different experimental approaches (surface

FIG 7 Evolution of mutations in noncoding regions in the three betacoronaviruses. Insertions or deletions occurred in four of the nine noncoding regions between coding genes, among which two of the three coincidentally occurred in the Kozak sequence-related positions (−3 to −1 positions from the start codon "aug"). One mutation upstream of the M gene realized the ideal Kozak motif of "gcc" in RaTG13 and SARS-CoV-2. Interestingly, noncoding regions upstream of the S gene and ORF7a gene contained different insertions or deletions at exactly the same position in RaTG13 and SARS-CoV-2. Turquoise blue, yellow, and gray colors indicate reserved bases, bases substituted by point mutations, and mutated bases based on insertion and/or deletion mutations with changed base size, respectively.

plasmon resonance versus biolayer interferometry binding) (31). Another recent study utilizing molecular dynamic simulation suggested that the surface of SARS-CoV-2 spike RBD is more positively charged and more attractive to the negatively charged surface of ACE2 than that of SARS-CoV, but the simulated total electrostatic forces between spike RBD and ACE2 were stronger in SARS-CoV than in SARS-CoV-2 (32). In spite of

TABLE 4 Insertion and/or deletion mutations with ≥10 consecutive bases and with changed base size between SARS-CoV and SARS-CoV-2[a]

Virus and base position	Genome sequences with consecutive gene mutations
SARS-CoV-2 (3,064–3,070 b)	5′-aggtgat-3′ (Nsp3 gene of ORF1a)
SARS-CoV (3,051–3,060 b)	5′-cgatgcagag-3′
SARS-CoV-2 (3,216–3,351 b)	5′-gtcaacaaactgttggtcaacaagacggcagtgaggacaatcagacaactactattcaaacaattgttgaggttcaacctcaattagagatggaacttaca ccagttgttcagactattgaagtgaatagttttag-3′ (Nsp3 gene of ORF1a)
SARS-CoV (3,206–3,346 b)	5′-ctgagcaatcagagattgagccagaaccagaacctacacctgaagaaccagttaatcagtttactggttatttaaaacttactgacaatgttgccattaaatgt gttgacatcgttaaggaggcacaaagtgctaatccta-3′
SARS-CoV-2 (3,365–3,441 b)	5′-cttactgacaatgtatacattaaaaatgcagacattgtggaagaagctaaaaaggtaaaaccaacagtggttgttaa-3′ (Nsp3 gene of ORF1a; insertion only with no corresponding sequence in SARS-CoV)
SARS-CoV-2 (3,826–3,955 b)	5′-ttcaagctttttggaaatgaagagtgaaaagcaagttgaacaaaagatcgctgagattcctaaagaggaagttaagccatttataactgaaagtaaaccttc agttgaacagagaaaacaagatgataag-3′ (Nsp3 gene of ORF1a)
SARS-CoV (3,744–3,867 b)	5′-catggattatcttgataacctgaagcctagagtggaagcacctaaacaagaggagccaccaaacacagaagattccaaaactgaggagaaatccgtc gtacagaagcctgtcgatgtgaagcca-3′
SARS-CoV-2 (4,880–4,894 b)	5′-agtaatcctaccaca-3′ (Nps3 gene of ORF1a)
SARS-CoV (4,792–4,809 b)	5′-ctggagagccccgtcgag-3′
SARS-CoV-2 (21,574–21,587 b)	5′-tcttgttttattgc-3′ (5′ terminal region, S1 gene)
SARS-CoV (21,488–21,505 b)	5′-cttattatttcttactct-3′
SARS-CoV-2 (21,595–21,647 b)	5′-ctctagtcagtgtgttaatcttacaaccagaactcaattacccccctgcataca-3′ (NTD, S1 gene)
SARS-CoV (21,513–21,558 b)	5′-ggtagtgaccttgaccggtgcaccacttttgatgatgttcaagctc-3′
SARS-CoV-2 (21,654–21,662 b)	5′-ctttcacac-3′ (NTD, S1 gene)
SARS-CoV (21,565–21,588 b)	5′-acactcaacatacttcatctatga-3′
SARS-CoV-2 (21,766–21,786 b)	5′-tacatgtctctgggaccaatgg-3′ (NTD, S1 gene; insertion only with no corresponding sequence in SARS-CoV)
SARS-CoV-2 (22,002–22,044 b)	5′-aaaacaacaaaagttggatggaaagtgagttcagagtttattc-3′(NTD, S1 gene)
SARS-CoV (21,922–21,952 b)	5′-tgggtacacagacacatactatgatattcga-3′
SARS-CoV-2 (22,159–22,186 b)	5′-ttattttaaaatatattctaagcacacg-3′(NTD, S1 gene)
SARS-CoV (22,067–22,094 b)	5′-gtttctctatgtttataagggctatcaa-3′
SARS-CoV-2 (22,270–22,345 b)	5′-taggtttcaaactttacttgctttacatagaagttatttgactcctggtgattcttcttcaggttggacagctggt-3′ (NTD, S1 gene)
SARS-CoV (22,163–22,220 b)	5′-aaattttagagccattcttacagccttttcacctgctcaagacatttggggcacgtca-3′
SARS-CoV-2 (22,974–22,991 b)	5′-aaatctatcaggccggta-3′ (RBD, S1 gene)
SARS-CoV (22,849–22,866 b)	5′-tgcctttctcccctgatg-3′
SARS-CoV-2 (23,004–23,014 b)	5′-atggtgttgaa-3′ (RBD, S1 gene)
SARS-CoV (22,879–22,886 b)	5′-ccccacct-3′
SARS-CoV-2 (23,590–23,630 b)	5′-tcagactaattctcctcggcgggcacgtagtgtagctagtc-3′ (S1/S2 furin cleavage site)
SARS-CoV (23,462–23,490 b)	5′-agtttctttattacgtagtactagccaaa-3′
SARS-CoV-2 (27,905–28,229 b)	5′-tgttttcttaggaatcatcacaactgtagctgcatttcaccaagaatgtagtttacagtcatgtactcaacatcaaccatatgtagttgatgacccgtgtcctatt cacttctattctaaatggtatattagagtaggagctagaaaatcagcacctttaattgaattgtgcgtggatgaggctggttctaaatcacccattcagt acatcgatatcggtaattatacagtttcctgtttacctttacaattaattgccaggaacctaaattgggtagtcttgtagtgcgttgttc gttctatgaagacttttagagtatcat-3′ (ORF8 gene)
SARS-CoV (27,775–28,067 b)	5′-cattgttttgacttgtatttctctatgcagttgcatatgcactgtagtacagcgctgtgcatctaataaacctcatgtgcttgaagatccttgtaaggtacaacac taggggtaatacttatagcactgcttggctttgtgctctaggaaaggttttacctttttcatagatggcacactatggttcaaacatgcacacctaatgtt actatcaactgtcaagatccagctggtggtgcgcttatagctaggtgttggtaccttcatgaaggtcaccaaactgctgcatttaga-3′

[a]Mutations listed in the table are those with insertion and/or deletion mutations involving ≥10 consecutive bases. Most of these mutations were accompanied by changed base sizes. These mutations in ORF1 were concentrated in the Nsp3 gene region. In the 13 listed long insertion and/or deletion mutations in the S gene, 9 occurred in the N-terminal domain (NTD), 3 occurred in the receptor-binding domain, and 1 occurred in the S1/S2 boundary region that produced the furin cleavage site in SARS-CoV-2. Repeating sequences that may promote a large length of insertions were not apparent in the listed mutation sites. Nsp3, nonstructural protein 3; b, base.

these inconsistent experimental results, the binding of RBD to ACE2 is still thought to be easier in SARS-CoV-2 than in SARS-CoV, as the sequence change in the hinge between RBD and other S domains could have made the RBD more flexible and easier to open for binding ACE2 (32). Furthermore, if the two viruses were derived from reservoir animals like bats or pangolins, these conformational changes in RaTG13 or SARS-CoV-2 S protein may have further granted immune evasion to the viruses.

This study also revealed that almost the complete ORF8 gene was substituted, based on an approximately 325 consecutive base-long insertion-and-deletion mutation. Although the genome sequences of ORF8 differ significantly between different coronaviruses and exact function of ORF8 remain unknown, the persistence of ORF8 in different lineages is proposed to suggest that it may play some unknown roles in SARS-CoV-2 replication (33). Meanwhile, the finding of this study that almost all of the genome sequence of ORF8 in SARS-CoV-2 was mutated from those in SARS-CoV may support that ORF8 is possibly dispensable for SARS-CoV-2 survival. In general, indels are less frequent than point substitutions in the natural environment, and when indels

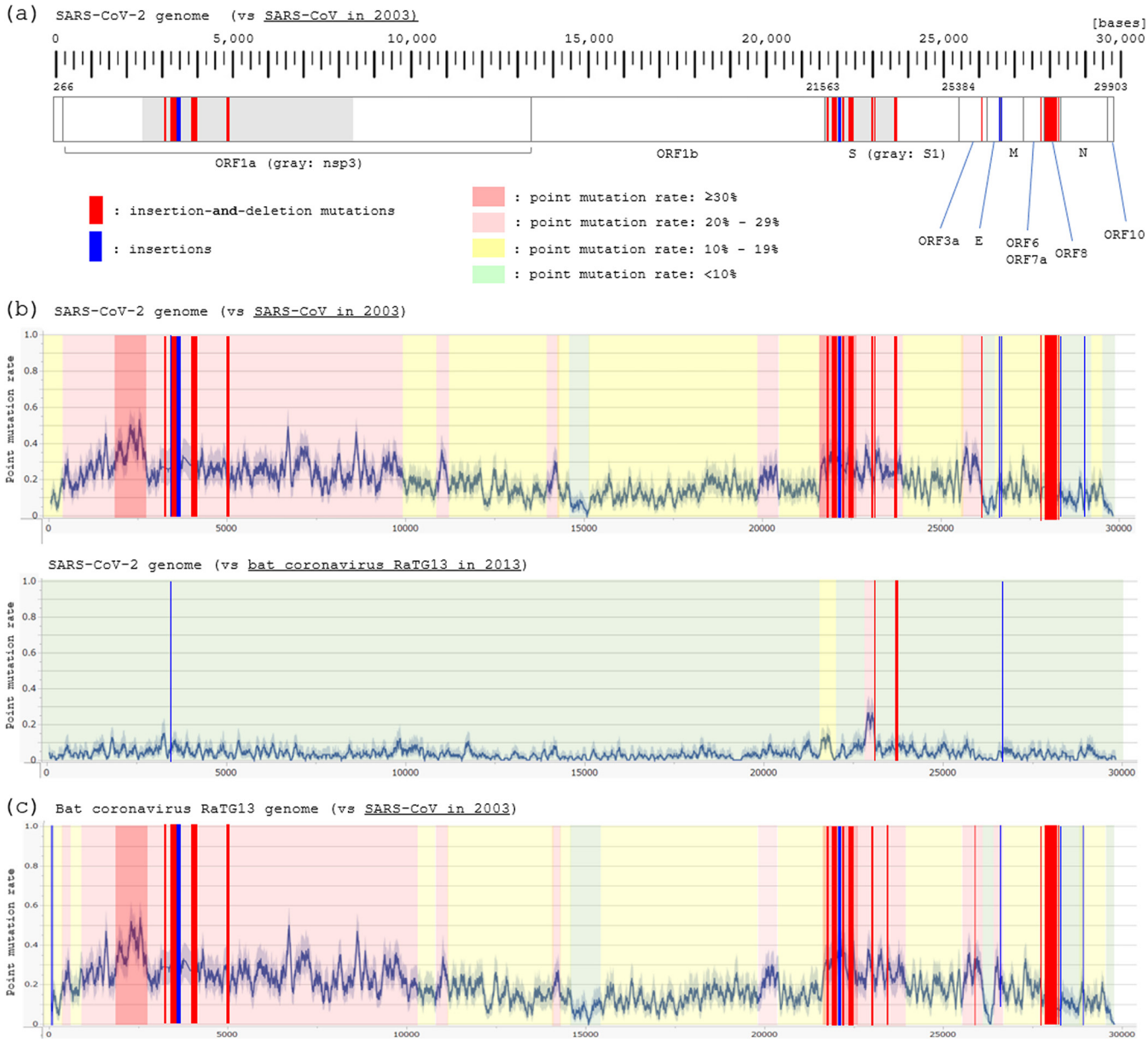

FIG 8 Distribution of insertion and/or deletion mutations on the genomes of SARS-CoV-2 and RaTG13. (a) Medium- to large-sized insertion and/or deletion mutations with approximately 10–300 consecutive base long were concentrated in the Nsp3 gene of the ORF1a, S1 domain of S gene, and ORF8 gene. (b) The line graphs show the simple moving average (±50 bases) for the point mutation rate across the genome of SARS-CoV-2, compared to the genome of SARS-CoV (top) or RaTG13 (bottom). The distribution of the indels (red and blue bars) was roughly matched to the distribution of the point mutation rate. (c) The line graph shows the simple moving average (±50 bases) for the point mutation rate across the genome of RaTG13, compared to that of SARS-CoV. The pattern of the point mutation rate and the distribution of indels roughly matched to those between SARS-CoV and SARS-CoV-2.

occur, most of the mutations are with relatively short length of consecutive bases. One of the reasons for this is that an indel would result in a frameshift in most cases, which usually causes a fatal change to the subsequent amino acid sequence, unless the change in base size is a multiple of three. Certainly, an up to 10,000 base-long indel can be observed in some eukaryotes as a natural evolutionary process. The presence of multiple indels in the genome can be a rationale for determining species and building phylogenetic trees (34–36). Some indels may have affected the evolvability of the involved species (37). The exact mechanisms of the occurrence of long indels in the natural environments remains uncertain, but the recently introduced genetic engineering technique using clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated proteins (Cas) may partially explain the phenomenon (38, 39).

The CRISPR technology was originally developed from the bacterial CRISPR-Cas9 antiviral immune system, which had already existed in almost all archaea and many bacteria in the natural environment (40–43). The CRISPR-Cas system is useful in RNA and DNA editing that realized relatively precise genome modifications, including an induction of indel to the DNA sequences (44, 45). The CRISPR toolkit is further expected to shorten the time for developing an effective live attenuated vaccine for some viruses or realizing viral load reduction after viral infections (45, 46). However, whether the CRISPR-Cas13 system (which targets RNA and not DNA) can realize the long insertion-and-deletion mutations in RNA molecules that were observed in the present study remains unknown. Conceivable hypotheses to explain the observed concentrated indels in the spike NTD of RaTG13 and SARS-CoV-2 include a natural evolutionary process. Generally, most indels are more deleterious than point substitution and vulnerable to evolutionary selection pressure (47). If indels occur in the coding regions with less selective tolerance, such as the structural core proteins and other proteins that are essential for survival, most of such deleterious mutations would surely be selected out (48). This may explain why indels were so concentrated in the NTD of the S1 gene in RaTG13 and SARS-CoV-2.

Another notable finding of this study was that insertion and/or deletion was confirmed in five of the 10 evaluated noncoding regions upstream of coding regions, 2 of which occurred in the Kozak sequences just before the start codon. Mutations upstream of S gene changed the Kozak sequence in −3 to −1 positions from "-aac-" (SARS-CoV) to "-acc-" (RaTG13) and that of M gene from "-atc-" (SARS-CoV) to "-gcc-" (RaTG13). These sequences are exactly the same with the "strong" Kozak consensus sequence motifs known to promote protein translation initiation by facilitating the assembly and translation start of the ribosome. Kozak sequences significantly regulate translation efficiency in eukaryotes (49–51). As the viruses depend on the host cells for their translation machinery, it is reasonable to expect that the translation efficiency of SARS-CoV-2 is also significantly influenced by the Kozak sequences upstream of coding regions. Furthermore, two insertion and/or deletion mutations occurred at the same position between RaTG13 and Wuhan-Hu-1 (noncoding regions upstream of the S gene and ORF7a gene). Whether such coincidence can occur in the natural environment remains uncertain, but this finding may be a clue to consider the possible evolutionary mechanism of RaTG13 and SARS-CoV-2.

This study has some limitations. First, the exact numbers of the inducted insertion-and-deletion mutation in the S, Nsp3, and ORF8 genes may not be definite, as some of the long insertion-and-deletion mutations may comprise several different mutations in the same positions. Furthermore, the exact roles of Nsp3 and ORF8 have not been fully elucidated. As a result, the exact biological and physiological significance of the observed concentrated insertion-and-deletion mutations in these genes in relation to SARS-CoV-2 development remains uncertain. As another limitation, this study only evaluated the genome sequence of three betacoronaviruses. Recently, genome sequences of the betacoronavirus sampled from *Rhinolophus shameli* bats in Cambodia in 2010 (RshSTT182 and RshSTT200) were reported to share 92.6% identity with SARS-CoV-2, implying that the progenitors of SARS-CoV-2 could have wider geographic distribution than previously expected, extending from Southeast Asia to southern China (52). To further elucidate the evolutionary process of SARS-CoV-2, genome sequence analysis with further betacoronaviruses that may lie between SARS-CoV and SARS-CoV-2 in the phylogenetic tree, sampled from diverse rhinolophid bats species in different countries, will be essential (53). Lastly, the coronavirus gene expression cannot be judged simply by the gene sequence or translation efficiency. There are many other viral factors that may affect gene expression levels, such as the functions of proteases coded in the viral RNA, viral subgenomic RNA amount, and RNA secondary structures (54–57). These unevaluated factors should also be compared between SARS-CoV and RaTG13 or SARS-CoV-2 to conclude the effect of the observed gene mutations on the changes in gene expression levels between the three betacoronaviruses.

In conclusion, mutations between SARS-CoV and SARS-CoV-2 were unevenly distributed across the virus genome. In the S1 gene, at least nine indels with greater or equal to seven consecutive bases long, including eight indels based on insertion-and-deletion mutations, were concentrated. Replacement of relatively long consecutive bases with insertion-and-deletion mutations were also confirmed in ORF1a (Nsp3) and ORF8 genes. The uneven distribution of medium- to large-sized indels based on insertion-and-deletion mutations in the SARS-CoV-2 genome may provide further insights into the evolutionary process of the betacoronavirus in the last decades. Future studies comparing SARS-CoV-2 genome with viral genomes of various betacoronaviruses, collected from wide range of animals, and evaluating the accumulation process of the observed insertion-and-deletion mutations in SARS-CoV-2 genome may offer us further insights into the process of SARS-CoV-2 development.

MATERIALS AND METHODS

Genome sequence references. The genome sequence of the SARS-CoV-2 isolate Wuhan-Hu-1, the National Institute of Health (NIH) genetic sequence database (GenBank) was referenced (reference sequence: NC_045512.2) (6, 58). The GenBank database for bat coronavirus RaTG13 was also referenced (reference: MN996532.2) (8, 59). The SARS-CoV genome sequence was obtained from a previous report by the Centers for Disease Control and Prevention and GenBank database (reference: AY345986.1) (60–62).

Evaluated genome regions. This study investigated all genome sequences and compared the sequences of SARS-CoV isolate CUHK-AG01, RaTG13, and SARS-CoV-2 isolate Wuhan-Hu-1. The evaluated specific genes were as follows: ORF1ab (266–21,555 bases in Wuhan-Hu-1), S (S1 and S2 subunits; 21,563–25,384 bases), ORF3a (25,393–26,220 bases), E (26,245–26,472 bases), M (26,523–27,191 bases), ORF6 (27,202–27,387 bases), ORF7a (27,394–27,759 bases), ORF8 (27,894–28,259 bases), N (28,274–29,533 bases), and ORF10 (29,558–29,674 bases). In addition to these coding regions, the sequences of the noncoding region upstream of ORF1a, S, ORF3, E, M, ORF6, ORF7a, ORF8, N, and ORF10 were further compared.

Types and rate of mutations. The mutations in the sequences of SARS betacoronaviruses were classified into indels and point mutations. Indels were further classified into insertion, deletion, and insertion-and-deletion mutations. Most of the insertion-and-deletion mutations resulted in changed base sizes, but some of them did not. Other popular mutations, such as duplication, inversion, or repeat expansion, were not apparent in the genome sequences evaluated in this study. Mutations in greater than or equal to seven consecutive bases in the coding regions were considered not to be a coincidental repetition of single base substitution, based on the knowledge that approximately 80% genome sequences were identical between SARS-CoV and SARS-CoV-2. Therefore, the probability of coincidental repetition of seven consecutive bases based on coincidental repetitive single base substitution is $(1/5)^7$ = 1/78,125, which is sufficiently low for the size of each evaluated coding gene in this study. Substitutions in less than or equal to six consecutive bases with preserved base size were regarded as coincidental repetition of point mutations, as the probability is not too low to reject the hypothesis of coincidental repetition.

Point substitution rate by genes. In each of the genes, point substitution rate in different gene positions were evaluated. In the genes with concentrated indels inside (i.e., ORF1ab and S genes), line graphs for the point substitution rates by the positions across the genes were depicted. In these two genes, the line graphs of the point substitution rate were obtained by calculating the simple moving average of the 50 bases before and after a specific base (i.e., ±50 bases for each base). Then, the simple moving average of point mutation rate at the base position of k (SMA_k) can be described as below by using the point mutation status (0 or 1) at the base position of k (M_k):

$$SMA_k = \frac{1}{100}\sum_{i=-50}^{49} M_{k+i}$$

The levels of the moving average of point substitution rates were categorized into the following four groups: very high (≥30%), high (20% to 29%), moderate (10% to 19%), and low (<10%).

Three-dimensional molecular structures of the S protein. The three-dimensional molecular structures were built and compared between the betacoronaviruses by using RasMol Software (http://www.openrasmol.org/). The structures for the S proteins of SARS-CoV and SARS-CoV-2 were obtained from RCSB Protein Data Bank with the Protein Data Bank file format (https://www.rcsb.org/) (16, 63). General appearance of the secondary, tertially, and quaternary structures of the S protein, in relation to the sites of the observed insertion-and-deletion mutations, was then visually evaluated and compared between that in SARS-CoV and in SARS-CoV-2.

Statistical analyses. The frequency of base substitution in a specific gene between different types of coronaviruses was compared by the chi-square test, using R Statistical Software (version 4.0.5; R Foundation, Vienna, Austria). Statistical significance was set at $P < 0.05$.

Data availability. This study does not use original sequencing data. All used sequencing data can be obtained from the NIH GenBank homepage. Genome sequence of the SARS-CoV-2 (Wuhan-Hu-1) is available at https://www.ncbi.nlm.nih.gov/nuccore/1798174254. Genome sequence of the bat

coronavirus RaTG13 is available at https://www.ncbi.nlm.nih.gov/nuccore/MN996532. Genome sequence of the SARS-CoV (CUHK-AG01) is available at https://www.ncbi.nlm.nih.gov/nuccore/AY345986.

ACKNOWLEDGMENT

The author declares no conflict to be disclosed for this study.

REFERENCES

1. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J, Lu J. 2020. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev 7:1012–1023. https://doi.org/10.1093/nsr/nwaa036.
2. Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry BW, Castoe TA, Rambaut A, Robertson DL. 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat Microbiol 5:1408–1417. https://doi.org/10.1038/s41564-020-0771-4.
3. Cui J, Li F, Shi ZL. 2019. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 17:181–192. https://doi.org/10.1038/s41579-018-0118-9.
4. Segreto R, Deigin Y. 2021. The genetic structure of SARS-CoV-2 does not rule out a laboratory origin: SARS-COV-2 chimeric structure and furin cleavage site might be the result of genetic manipulation. Bioessays 43:e2000240. https://doi.org/10.1002/bies.202000240.
5. Sirotkin K, Sirotkin D. 2020. Might SARS-CoV-2 have arisen via serial passage through an animal host or cell culture?: a potential explanation for much of the novel coronavirus' distinctive genome. Bioessays 42:e2000091. https://doi.org/10.1002/bies.202000091.
6. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y, Yuan M-L, Zhang Y-L, Dai F-H, Liu Y, Wang Q-M, Zheng J-J, Xu L, Holmes EC, Zhang Y-Z. 2020. A new coronavirus associated with human respiratory disease in China. Nature 579:265–269. https://doi.org/10.1038/s41586-020-2008-3.
7. Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, Wang H, Crameri G, Hu Z, Zhang H, Zhang J, McEachern J, Field H, Daszak P, Eaton BT, Zhang S, Wang L-F. 2005. Bats are natural reservoirs of SARS-like coronaviruses. Science 310:676–679. https://doi.org/10.1126/science.1118391.
8. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L, Chen H-D, Chen J, Luo Y, Guo H, Jiang R-D, Liu M-Q, Chen Y, Shen X-R, Wang X, Zheng X-S, Zhao K, Chen Q-J, Deng F, Liu L-L, Yan B, Zhan F-X, Wang Y-Y, Xiao G-F, Shi Z-L. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270–273. https://doi.org/10.1038/s41586-020-2012-7.
9. Zhang T, Wu Q, Zhang Z. 2020. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. Curr Biol 30:1346–1351.e1342. https://doi.org/10.1016/j.cub.2020.03.022.
10. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of SARS-CoV-2. Nat Med 26:450–452. https://doi.org/10.1038/s41591-020-0820-9.
11. Sasaki M, Uemura K, Sato A, Toba S, Sanaki T, Maenaka K, Hall WW, Orba Y, Sawa H. 2021. SARS-CoV-2 variants with mutations at the S1/S2 cleavage site are generated in vitro during propagation in TMPRSS2-deficient cells. PLoS Pathog 17:e1009233. https://doi.org/10.1371/journal.ppat.1009233.
12. Lam TT-Y, Jia N, Zhang Y-W, Shum MH-H, Jiang J-F, Zhu H-C, Tong Y-G, Shi Y-X, Ni X-B, Liao Y-S, Li W-J, Jiang B-G, Wei W, Yuan T-T, Zheng K, Cui X-M, Li J, Pei G-Q, Qiang X, Cheung WY-M, Li L-F, Sun F-F, Qin S, Huang J-C, Leung GM, Holmes EC, Hu Y-L, Guan Y, Cao W-C. 2020. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature 583:282–285. https://doi.org/10.1038/s41586-020-2169-0.
13. Winstone H, Lista MJ, Reid AC, Bouton C, Pickering S, Galao RP, Kerridge C, Doores KJ, Swanson CM, Neil SJD. 2021. The polybasic cleavage site in SARS-CoV-2 spike modulates viral sensitivity to type i interferon and IFITM2. J Virol 95:e02422-20. https://doi.org/10.1128/JVI.02422-20.
14. Chen J, Wang R, Wang M, Wei GW. 2020. Mutations strengthened SARS-CoV-2 infectivity. J Mol Biol 432:5212–5226. https://doi.org/10.1016/j.jmb.2020.07.009.
15. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, Ludden C, Reeve R, Rambaut A, Peacock SJ, Robertson DL, COVID-19 Genomics UK (COG-UK) Consortium. 2021. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol 19:409–424. https://doi.org/10.1038/s41579-021-00573-0.
16. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. 2020. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 181:281–292.e286. https://doi.org/10.1016/j.cell.2020.02.058.
17. Laffeber C, de Koning K, Kanaar R, Lebbink JHG. 2021. Experimental evidence for enhanced receptor binding by rapidly spreading SARS-CoV-2 variants. J Mol Biol 433:167058. https://doi.org/10.1016/j.jmb.2021.167058.
18. Cai Y, Zhang J, Xiao T, Lavine CL, Rawson S, Peng H, Zhu H, Anand K, Tong P, Gautam A, Lu S, Sterling SM, Walsh RM, Rits-Volloch S, Lu J, Wesemann DR, Yang W, Seaman MS, Chen B. 2021. Structural basis for enhanced infectivity and immune evasion of SARS-CoV-2 variants. Science 373:642–648. https://doi.org/10.1126/science.abi9745.
19. Prabakaran P, Xiao X, Dimitrov DS. 2004. A model of the ACE2 structure and function as a SARS-CoV receptor. Biochem Biophys Res Commun 314:235–241. https://doi.org/10.1016/j.bbrc.2003.12.081.
20. Hassanzadeh K, Perez Pena H, Dragotto J, Buccarello L, Iorio F, Pieraccini S, Sancini G, Feligioni M. 2020. Considerations around the SARS-CoV-2 spike protein with particular attention to COVID-19 brain infection and neurological symptoms. ACS Chem Neurosci 11:2361–2369. https://doi.org/10.1021/acschemneuro.0c00373.
21. Wells VR, Plotch SJ, DeStefano JJ. 2001. Determination of the mutation rate of poliovirus RNA-dependent RNA polymerase. Virus Res 74:119–132. https://doi.org/10.1016/S0168-1702(00)00256-2.
22. García-Villada L, Drake JW. 2012. The three faces of riboviral spontaneous mutation: spectrum, mode of genome replication, and mutation rate. PLoS Genet 8:e1002832. https://doi.org/10.1371/journal.pgen.1002832.
23. Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X, Scherbakova S, Graham RL, Baric RS, Stockwell TB, Spiro DJ, Denison MR. 2010. Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. PLoS Pathog 6:e1000896. https://doi.org/10.1371/journal.ppat.1000896.
24. Bar-On YM, Flamholz A, Phillips R, Milo R. 2020. SARS-CoV-2 (COVID-19) by the numbers. Elife 9:e57309. https://doi.org/10.7554/eLife.57309.
25. Otto SP, Day T, Arino J, Colijn C, Dushoff J, Li M, Mechai S, Van Domselaar G, Wu J, Earn DJD, Ogden NH. 2021. The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. Curr Biol 31:R918–R929. https://doi.org/10.1016/j.cub.2021.06.049.
26. De Maio N, Walker CR, Turakhi Y, Lanfear R, Corbett-Detig R, Goldman N. 2021. Mutation rates and selection on synonymous mutations in SARS-CoV-2. Genome Biol Evol 13:evab087. https://doi.org/10.1093/gbe/evab087.
27. Banerjee A, Mossman K, Grandvaux N. 2021. Molecular determinants of SARS-CoV-2 variants. Trends Microbiol 29:871–873. https://doi.org/10.1016/j.tim.2021.07.002.
28. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang Y-P, Boerwinkle E, Fu Y-X. 2004. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC Evol Biol 4:21. https://doi.org/10.1186/1471-2148-4-21.
29. Lu J, Sun PD. 2020. High affinity binding of SARS-CoV-2 spike protein enhances ACE2 carboxypeptidase activity. J Biol Chem 295:18579–18588. https://doi.org/10.1074/jbc.RA120.015303.
30. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. 2020. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science 367:1444–1448. https://doi.org/10.1126/science.abb2762.
31. Zamorano Cuervo N, Grandvaux N. 2020. ACE2: evidence of role as entry receptor for SARS-CoV-2 and implications in comorbidities. Elife 9:e61390. https://doi.org/10.7554/eLife.61390.
32. Xie Y, Karki CB, Du D, Li H, Wang J, Sobitan A, Teng S, Tang Q, Li L. 2020. Spike Proteins of SARS-CoV and SARS-CoV-2 utilize different mechanisms to bind with human ACE2. Front Mol Biosci 7:591873–591873. https://doi.org/10.3389/fmolb.2020.591873.
33. Pereira F. 2020. Evolutionary dynamics of the SARS-CoV-2 ORF8 accessory gene. Infect Genet Evol 85:104525–104525. https://doi.org/10.1016/j.meegid.2020.104525.
34. Pereira F, Carneiro J, Matthiesen R, van Asch B, Pinto N, Gusmão L, Amorim A. 2010. Identification of species by multiplex analysis of variable-length sequences. Nucleic Acids Res 38:e203. https://doi.org/10.1093/nar/gkq865.
35. Nakamura H, Muro T, Imamura S, Yuasa I. 2009. Forensic species identification based on size variation of mitochondrial DNA hypervariable regions. Int J Legal Med 123:177–184. https://doi.org/10.1007/s00414-008-0306-7.

36. Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermat T, Corthier G, Brochmann C, Willerslev E. 2007. Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. Nucleic Acids Res 35:e14. https://doi.org/10.1093/nar/gkl938.
37. Consuegra J, Gaffé J, Lenski RE, Hindré T, Barrick JE, Tenaillon O, Schneider D. 2021. Insertion-sequence-mediated mutations both promote and constrain evolvability during a long-term experiment with bacteria. Nat Commun 12:980. https://doi.org/10.1038/s41467-021-21210-7.
38. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. 2013. Genome engineering using the CRISPR-Cas9 system. Nat Protoc 8:2281–2308. https://doi.org/10.1038/nprot.2013.143.
39. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337:816–821. https://doi.org/10.1126/science.1225829.
40. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science 315:1709–1712. https://doi.org/10.1126/science.1138140.
41. Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, Dickman MJ, Makarova KS, Koonin EV, van der Oost J. 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321:960–964. https://doi.org/10.1126/science.1159689.
42. Marraffini LA, Sontheimer EJ. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science 322:1843–1845. https://doi.org/10.1126/science.1165771.
43. Takeuchi N, Wolf YI, Makarova KS, Koonin EV. 2012. Nature and intensity of selection pressure on CRISPR-associated genes. J Bacteriol 194:1216–1225. https://doi.org/10.1128/JB.06521-11.
44. Matsoukas IG. 2018. Commentary: RNA editing with CRISPR-Cas13. Front Genet 9:134. https://doi.org/10.3389/fgene.2018.00134.
45. Jamehdor S, Naserian S, Teimoori A. 2022. Enhanced high mutation rate and natural selection to produce attenuated viral vaccine with CRISPR toolkit in RNA viruses especially SARS-CoV-2. Infect Genet Evol 97:105188. https://doi.org/10.1016/j.meegid.2021.105188.
46. Jamehdor S, Pajouhanfar S, Saba S, Uzan G, Teimoori A, Naserian S. 2022. Principles and applications of CRISPR toolkit in virus manipulation, diagnosis, and virus-host interactions. Cells 11:999. https://doi.org/10.3390/cells11060999.
47. Sung W, Ackerman MS, Dillon MM, Platt TG, Fuqua C, Cooper VS, Lynch M. 2016. Evolution of the insertion-deletion mutation rate across the tree of life. G3 (Bethesda) 6:2583–2591. https://doi.org/10.1534/g3.116.030890.
48. Taylor MS, Ponting CP, Copley RR. 2004. Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. Genome Res 14:555–566. https://doi.org/10.1101/gr.1977804.
49. Kozak M. 1984. Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin in vivo. Nature 308:241–246. https://doi.org/10.1038/308241a0.
50. Kozak M. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 44:283–292. https://doi.org/10.1016/0092-8674(86)90762-2.
51. Kozak M. 1987. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res 15:8125–8148. https://doi.org/10.1093/nar/15.20.8125.
52. Delaune D, Hul V, Karlsson EA, Hassanin A, Ou TP, Baidaliuk A, Gámbaro F, Prot M, Tu VT, Chea S, Keatts L, Mazet J, Johnson CK, Buchy P, Dussart P, Goldstein T, Simon-Lorière E, Duong V. 2021. A novel SARS-CoV-2 related coronavirus in bats from Cambodia. Nat Commun 12:6563. https://doi.org/10.1038/s41467-021-26809-4.
53. Zhou H, Ji J, Chen X, Bi Y, Li J, Wang Q, Hu T, Song H, Zhao R, Chen Y, Cui M, Zhang Y, Hughes AC, Holmes EC, Shi W. 2021. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. Cell 184:4380–4391.e4314. https://doi.org/10.1016/j.cell.2021.06.008.
54. Miller WA, Koev G. 2000. Synthesis of subgenomic RNAs by positive-strand RNA viruses. Virology 273:1–8. https://doi.org/10.1006/viro.2000.0421.
55. van den Born E, Posthuma CC, Gultyaev AP, Snijder EJ. 2005. Discontinuous subgenomic RNA synthesis in arteriviruses is guided by an RNA hairpin structure located in the genomic leader region. J Virol 79:6312–6324. https://doi.org/10.1128/JVI.79.10.6312-6324.2005.
56. Dougherty WG, Semler BL. 1993. Expression of virus-encoded proteinases: functional and structural similarities with cellular enzymes. Microbiol Rev 57:781–822. https://doi.org/10.1128/mr.57.4.781-822.1993.
57. Huston NC, Wan H, Strine MS, de Cesaris Araujo Tavares R, Wilen CB, Pyle AM. 2021. Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. Mol Cell 81:584–598.e585. https://doi.org/10.1016/j.molcel.2020.12.041.
58. National Center for Biotechnology Information (NCBI). 2020. Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome. NCBI Reference Sequence: NC_045512.2. https://www.ncbi.nlm.nih.gov/nuccore/1798174254. Accessed February 3, 2022.
59. National Center for Biotechnology Information (NCBI). 2020. Bat coronavirus RaTG13, complete genome. GenBank: MN996532.2. https://www.ncbi.nlm.nih.gov/nuccore/MN996532. Accessed February 3, 2022.
60. Centers for Disease Control and Prevention (CDC). 2003. SARS-associated coronavirus (SARS-CoV) Sequencing. https://www.cdc.gov/sars/lab/downloads/nucleoseq.pdf. Accessed February 3, 2022.
61. National Center for Biotechnology Information (NCBI). 2003. SARS coronavirus CUHK-AG01, complete genome (VRL 29-NOV-2003). GenBank: AY345986.1. https://www.ncbi.nlm.nih.gov/nuccore/AY345986.1. Accessed February 3, 2022.
62. Chim SSC, Tsui SKW, Chan KCA, Au TCC, Hung ECW, Tong YK, Chiu RWK, Ng EKO, Chan PKS, Chu CM, Sung JJY, Tam JS, Fung KP, Waye MMY, Lee CY, Yuen KY, Lo YMD. 2003. Genomic characterisation of the severe acute respiratory syndrome coronavirus of Amoy Gardens outbreak in Hong Kong. Lancet 362:1807–1808. https://doi.org/10.1016/S0140-6736(03)14901-X.
63. Gui M, Song W, Zhou H, Xu J, Chen S, Xiang Y, Wang X. 2017. Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. Cell Res 27:119–129. https://doi.org/10.1038/cr.2016.152.