Trinucleotide Substitutions at Two Locations in the SARS-CoV-2 Nucleocapsid (*N*) Gene

Tetsuya Akaishi,[1,2] Kei Fujiwara[3] and Tadashi Ishii[1,2]

[1]Department of Education and Support for Regional Medicine, Tohoku University Hospital, Sendai, Miyagi, Japan
[2]COVID-19 Testing Center, Tohoku University, Sendai, Miyagi, Japan
[3]Department of Gastroenterology and Metabolism, Nagoya City University, Nagoya, Aichi, Japan

The genomes of sarbecoviruses, including severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), incorporate mutations with short sequence exchanges based on unknown processes. Currently, the presence of such short-sequence exchanges among the genomes of different SARS-CoV-2 lineages remains uncertain. In the present study, multiple SARS-CoV-2 genome sequences from different clades or sublineages were collected from an international mass sequence database and compared to identify the presence of short sequence exchanges. Initial screening with multiple sequence alignments identified two locations with trinucleotide substitutions, both in the nucleocapsid (*N*) gene. The first exchange from 5'-GAT-3' to 5'-CTA-3' at nucleotide positions 28,280-28,282 resulted in a change in the amino acid from aspartic acid (D) to leucine (L), which was predominant in clade GRY (Alpha). The second exchange from 5'-GGG-3' to 5'-AAC-3' at nucleotide positions 28,881-28,883 resulted in an amino acid change from arginine and glycine (RG) to lysine and arginine (KR), which was predominant in GR (Gamma), GRY (Alpha), and GRA (Omicron). Both trinucleotide substitutions occurred before June 2020. The sequence identity rate between these lineages suggests that coincidental succession of single-nucleotide substitutions is unlikely. Basic local alignment search tool sequence search revealed the absence of intermediating mutations based on single-base substitutions or overlapping indels before the emergence of these trinucleotide substitutions. These findings suggest that trinucleotide substitutions could have developed via an en bloc exchange. In summary, trinucleotide substitutions at two locations in the SARS-CoV-2 *N* gene were identified. This mutation may provide insights into the evolution of SARS-CoV-2.

Keywords: GISAID; nucleocapsid protein; severe acute respiratory syndrome coronavirus 2; trinucleotide exchange
Tohoku J. Exp. Med., 2023 May, **260** (1), 21-27.
doi: 10.1620/tjem.2023.J010

Introduction

The pandemic of coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is still categorized among the primary public health concerns globally in 2023 (Johns Hopkins University 2023). In early 2022, the largest outbreak surge was reported worldwide owing to the emergence of the Omicron variant (Karim and Karim 2021; Araf et al. 2022). During the pandemic, the intermittent emergence of multiple variants of concern (VOC) was observed, which played a primary role in maintaining the pandemic across various countries worldwide (Mlcochova et al. 2021; Tao et al. 2021). The mutation profiles in previous VOCs have been primarily evaluated and categorized based on single nucleotide substitutions (i.e., point mutations), especially in the receptor-binding domain (RBD) of the SARS-CoV-2 spike (*S*) gene S1 subunit (SeyedAlinaghi et al. 2021; Hajizadeh et al. 2022). Recently, hotspots of insertions/deletions (indels) have been reported in the open reading frame 1a (*ORF1a*) polyprotein-encoding gene and N-terminal domain (NTD) of the *S1* gene in various SARS-related coronaviruses (Akaishi 2022a), suggesting the importance of mutations in these genomic sites that are different from S1 RBD in the subgenus Sarbecovirus. Moreover, highly polymorphic indel sites have been identified in SARS-

Received January 22, 2023; revised and accepted February 7, 2023; J-STAGE Advance online publication February 16, 2023

Correspondence: Tetsuya Akaishi, Department of Education and Support for Regional Medicine, Tohoku University Hospital, 1-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi 980-8574, Japan.
e-mail: t-akaishi@med.tohoku.ac.jp

CoV-2 S1-NTD sampled from humans (Akaishi et al. 2022a). In addition to these traditional mutations with single nucleotide substitutions or indels, thorough exchange of short sequence with unknown mechanisms has been suggested to exist in the genomes of SARS-related coronaviruses (Akaishi et al. 2022b). However, the exact developmental processes of such short-sequence exchanges or the presence of sequence exchanges among different SARS-CoV-2 lineages remains uncertain. To search for traits and developmental processes of such a thorough exchange of consecutive nucleotides in SARS-CoV-2 genomes, the present study compared multiple whole genome sequences of SARS-CoV-2 belonging to different lineages and examined for such an exchange of consecutive bases across the genomes that cannot be simply explained by the combination of conventional mutation types, such as point mutations, insertions, and deletions.

Methods

Evaluation of SARS-CoV-2 genome sequences

In the present study, 49 SARS-CoV-2 genome sequences sampled from humans were collected and compared to search for the presence of exchanges in consecutive bases that cannot be attributed to the combinations of point mutations, insertions, and deletions. The genome sequence of the original Wuhan-Hu-1 strain was obtained from the NCBI GenBank database (Bethesda, MD, USA) with accession ID MN908947.3. Other sequences from subsequent lineages were obtained from the Global Initiative on Sharing All Influenza Data (GISAID) database (Munich, Germany) (Elbe and Buckland-Merrett 2017; Shu and McCauley 2017; Khare et al. 2021; GISAID 2023), which were registered and available by December 22nd, 2022. Four sequences were randomly selected from the database for each of the following clades or lineages: L, GH (Beta), GR (Gamma), G, GRY (Alpha), GK (Delta), and GRA (Omicron), with sublineages BA.1, BA.2, BA.5, BQ.1, BQ.1.1, and XBB. A list of evaluated 49 genome sequences is shown in Table 1.

Multiple sequence alignments and sequence identity analysis

With 49 whole genome sequences (nt 1-29,903), multiple sequence alignments were performed using Molecular Evolutionary Genetics Analysis Version 11 (MEGA11) software (Tamura et al. 2021). Regarding the alignment parameters, the gap opening penalty score was set to –400, and the gap extension penalty score was set to 0. Multiple sequence alignments were performed after dividing the collected sequences into two sets, each with the original Wuhan-Hu-1 sequence and 24 sequences from subsequent lineages (two sequences from each clade or sublineage). Using the aligned sequences, base positions with exchanges of ≥ 3 consecutive nucleotides were examined across the whole genomes. Furthermore, for sequences with exchanges of consecutive nucleotides, the sequence identity rate (%) in comparison with the original Wuhan-Hu-1 genome sequence was calculated using Multiple Alignment using Fast Fourier Transform (MAFFT) software, offered by the European Molecular Biology Laboratory (EMBL) (European Molecular Biology Laboratory 2023).

Ethics and data availability

The present study was approved by the institutional review board of Tohoku University Graduate School of Medicine (approval number: 2022-1-720). The findings of this study are based on metadata associated with 14,329,052 sequences, which were sampled from humans and available on GISAID up to December 22nd, 2022, via EPI_

Table 1. List of the evaluated 49 genome sequences of SARS-CoV-2 sampled from humans.

Clade (lineages)	GISAID Accession ID
Original strain	Wuhan-Hu-1 (GenBank Accession: MN908947.3)
L (B)	EPI_ISL_16138010, EPI_ISL_413015, EPI_ISL_416739, EPI_ISL_420520
G (B.1)	EPI_ISL_15755020, EPI_ISL_14379838, EPI_ISL_766913, EPI_ISL_770906
GH (Beta; B.1.)	EPI_ISL_16119477, EPI_ISL_647686, EPI_ISL_419585, EPI_ISL_671465
GR (Gamma)	EPI_ISL_9754628, EPI_ISL_16225655, EPI_ISL_423654, EPI_ISL_563257
GRY (Alpha)	EPI_ISL_916362, EPI_ISL_3486562, EPI_ISL_659373, EPI_ISL_939571
GK (Delta)	EPI_ISL_16215541, EPI_ISL_11233238, EPI_ISL_1360317, EPI_ISL_2516080
GRA (Omicron; BA.1)	EPI_ISL_16216152, EPI_ISL_11425679, EPI_ISL_8271120, EPI_ISL_8593285
GRA (Omicron; BA.2)	EPI_ISL_14335247, EPI_ISL_16175961, EPI_ISL_8419126, EPI_ISL_9871917
GRA (Omicron; BA.5)	EPI_ISL_16145048, EPI_ISL_12919999, EPI_ISL_12893688, EPI_ISL_15336666
GRA (Omicron; BQ.1)	EPI_ISL_16222509, EPI_ISL_14850768, EPI_ISL_14773615, EPI_ISL_15888367
GRA (Omicron; BQ.1.1)	EPI_ISL_16222134, EPI_ISL_15029320, EPI_ISL_15001867, EPI_ISL_15881942
GRA (Omicron; XBB)	EPI_ISL_15826137, EPI_ISL_15276133, EPI_ISL_15276133, EPI_ISL_15853550

The genome sequence of the original Wuhan-Hu-1 was obtained from the NCBI GenBank database. The other 48 sequences, 4 from each clade or sublineage, were randomly selected from the GISAID database from the overall sequences that were registered and available by December 22nd, 2022.

GISAID, Global Initiative on Sharing All Influenza Data.

SET_230112cr.

Results

Sites with an exchange of ≥ 3 consecutive nucleotides

Two sets of multiple sequence alignments with different sets of 25 sequences identified two sites with base exchanges in three consecutive nucleotides, both in the *N* gene. The first site was located at the N-terminal domain of the coding region of the *N* gene corresponding to the nucleotide positions of 28,280-28,282 nt in the SARS-CoV-2 genome sequence, and a change in the nucleotide sequence from "GAT" to "CTA" was observed. The coded amino acid was changed from aspartic acid (D) to leucine (L). This first trinucleotide exchange site was identified in both evaluated sequences from clade GRY (Alpha). The second site was located at the nucleotide positions of 28,881-28,883 nt in the middle of the SARS-CoV-2 *N* gene, and a change in the nucleotide sequence from "GGG" to "AAC" was observed. The coded amino acids were changed from the succession of arginine and glycine (203-204 amino acids: RG) to lysine and arginine (KR). This second trinucleotide exchange site was identified in the clades GR (Gamma), GRY (Alpha), and GRA (Omicron), including all evaluated sublineages of Omicron (BA.1, BA.2, BA.5, BQ.1, BQ.1.1, and XBB). The results of multiple sequence alignments with the evaluated 25 sequences at these two sites of trinucleotide exchanges are shown in Fig. 1.

Probability of successive point mutations

Next, to exclude the possibility that the observed trinucleotide exchanges in the SARS-CoV-2 *N* gene were developed by the gradual accumulation of point nucleotide substitutions coincidentally at three successive nucleotide positions, the expected probability of observation of three successive point mutations was estimated based on the overall mutation rates between the original Wuhan-Hu-1 and one of the evaluated SARS-CoV-2 genomes from clade GRY (EPI_ISL_916362). Based on EMBL MAFFT sequence identity analysis of the two sequences, the sequence identity was 99.91%, after excluding nucleotide positions with indels. Based on the nucleotide mutation rate of 0.09% (9 in 10,000 nucleotides), the expected probability of observing point mutations in three successive nucleotides across 29,903 bases (Wuhan-Hu-1 whole genome) was 29,903 × (9E-4)^3 ≈ 2.180E-05 (i.e., 2.180E-03%). As we observed two sites with a three-base exchange, the probability was (2.180E-05)^2 ≈ 4.752E-10 (i.e., 4.752E-8%). This value was sufficiently low; hence, we could conclude that the two sites of trinucleotide exchange may not have developed by a coincidental succession of single base substitutions in three consecutive nucleotides.

Number of sequences with each amino acid substitution in the GISAID database

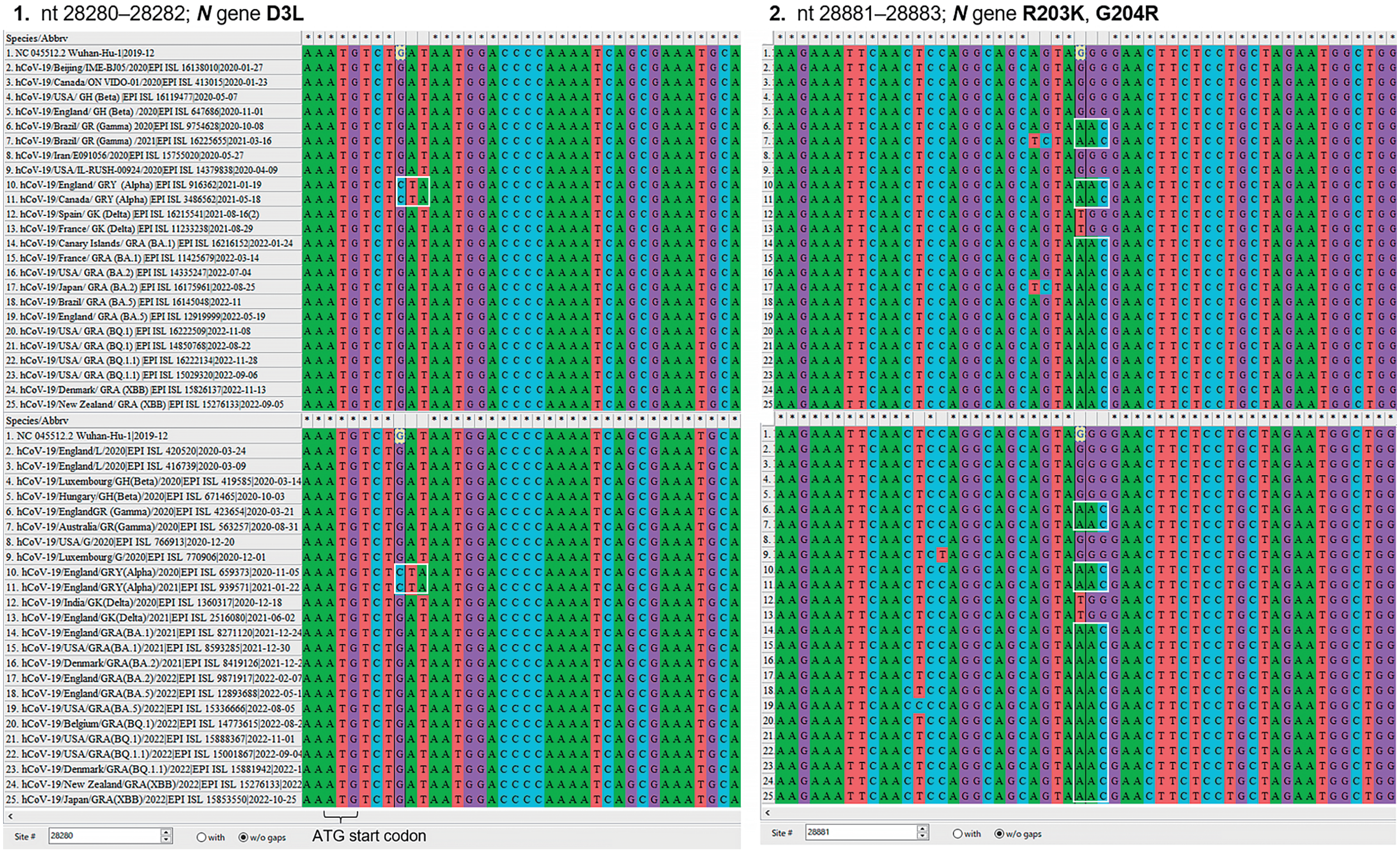
Next, the overall number of registered sequences in the GISAID database with each of the observed three amino acid substitutions was investigated. These values were further evaluated using different GISAID clades and sublineages. The results are presented in Table 2. N_D3L substitution was observed in 98.80% of sequences from the clade GRY (Alpha); however, it was also noted in a small number of sequences from the clade G (5.45%) and GR (Gamma; 16.72%). Meanwhile, the successive N_R203K and N_G204R substitutions were almost exclusively observed in GR (Gamma; 97.08%), GRY (Alpha; 92.68%), and GRA (Omicron; 95.93%).

Intermediating mutations based on single-base substitutions

To further rule out the possibility of gradual accumulation of single nucleotide substitutions in the first three-base exchange site at 28,280-28,282 nt, the numbers of registered sequences in the GISAID database with conceivable intermediating types of single nucleotide substitution linking 5'-GAT-3' to 5'-CTA-3' were evaluated. Three possible single-point substitutions are possible: 5'-GAT-3' (amino acid: D) to 5'-CAT-3' (amino acid: H), 5'-GTT-3' (amino acid: V), and 5'-GAA-3' (amino acid: E). Among the registered overall 14,329,052 sequences, the number of registered sequences with N_D3H substitution was 91 (0.0006%), that with N_D3V substitution was 259 (0.002%), and that with N_D3E substitution was 913 (0.006%) sequences. These very low frequencies of conceivable single nucleotide substitutions linking 5'-GAT-3' to 5'-CTA-3' at nt 28,280-28,282 supported that the observed trinucleotide exchange could have occurred simultaneously as an en bloc sequence exchange. As this three-base exchange started to gradually increase in clade G, the prevalence of sequences with each of the three possible intermediating substitutions was further evaluated among the 354,434 sequences belonging to clade G. The number of sequences with N_D3H was 21 (0.006%), that with N_D3V was 12 (0.003%), and that with N_D3E was 155 (0.04%). Again, the results supported that the three-base exchange occurred en bloc at once and not via gradual accumulation of single base substitutions in successive base positions.

Intermediating mutations based on overlapping indels

Next, to exclude the possibility of two-step indel process (i.e., a three-base insertion after a three-base deletion or a three-base deletion after a three-base insertion) for the first three-base exchange (amino acid: D > L) at the N-terminus of *N* gene, a GISAID database search with N_ins3L, N_ins4L, and N_D3del was performed. None of the three indel patterns were identified among the overall 14,329,052 sequences registered in the GISAID database (n = 0/14,329,052; 0.0%, for all of the three types). This finding suggests that a multistep developmental process of three-base exchange based on overlapped three-base insertion and three-base deletion is unlikely.


1. nt 28280–28282; *N* gene **D3L**
2. nt 28881–28883; *N* gene **R203K**, **G204R**
ATG start codon
Site # 28280
○with ●w/o gaps
Site # 28881
○with ●w/o gaps

```
Species/Abbrv
1. NC 045512.2 Wuhan-Hu-1|2019-12                                AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
2. hCoV-19/Beijing/IME-BJ05/2020|EPI ISL 16138010|2020-01-27     AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
3. hCoV-19/Canada/ON VIDO-01/2020|EPI ISL 413015|2020-01-23      AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
4. hCoV-19/USA/ GH (Beta) |EPI ISL 16119477|2020-05-07           AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
5. hCoV-19/England/ GH (Beta) /2020|EPI ISL 647686|2020-11-01    AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
6. hCoV-19/Brazil/ GR (Gamma) 2020|EPI ISL 9754628|2020-10-08    AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
7. hCoV-19/Brazil/ GR (Gamma) /2021|EPI ISL 16225655|2021-03-16  AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
8. hCoV-19/Iran/E091056/2020|EPI ISL 15755020|2020-05-27         AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
9. hCoV-19/USA/IL-RUSH-00924/2020|EPI ISL 14379838|2020-04-09    AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
10. hCoV-19/England/ GRY (Alpha) |EPI ISL 916362|2021-01-19      AAATGTCTCTAAATGGACCCCAAAATCAGCGAAATGCA
11. hCoV-19/Canada/ GRY (Alpha) |EPI ISL 3486562|2021-05-18      AAATGTCTCTAAATGGACCCCAAAATCAGCGAAATGCA
12. hCoV-19/Spain/ GK (Delta) |EPI ISL 16215541|2021-08-16(2)    AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
13. hCoV-19/France/ GK (Delta) |EPI ISL 11233238|2021-08-29      AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
14. hCoV-19/Canary Islands/ GRA (BA.1) |EPI ISL 16216152|2022-01-24 AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
15. hCoV-19/France/ GRA (BA.1) |EPI ISL 11425679|2022-03-14      AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
16. hCoV-19/USA/ GRA (BA.2) |EPI ISL 14335247|2022-07-04         AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
17. hCoV-19/Japan/ GRA (BA.2) |EPI ISL 16175961|2022-08-25       AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
18. hCoV-19/Brazil/ GRA (BA.5) |EPI ISL 16145048|2022-11         AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
19. hCoV-19/England/ GRA (BA.5) |EPI ISL 12919999|2022-05-19     AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
20. hCoV-19/USA/ GRA (BQ.1) |EPI ISL 16222509|2022-11-08         AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
21. hCoV-19/USA/ GRA (BQ.1) |EPI ISL 14850768|2022-08-22         AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
22. hCoV-19/USA/ GRA (BQ.1.1) |EPI ISL 16222134|2022-11-28       AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
23. hCoV-19/USA/ GRA (BQ.1.1) |EPI ISL 15029320|2022-09-06       AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
24. hCoV-19/Denmark/ GRA (XBB) |EPI ISL 15826137|2022-11-13      AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
25. hCoV-19/New Zealand/ GRA (XBB) |EPI ISL 15276133|2022-09-05  AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
Species/Abbrv
1. NC 045512.2 Wuhan-Hu-1|2019-12                                AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
2. hCoV-19/England/L/2020|EPI ISL 420520|2020-03-24              AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
3. hCoV-19/England/L/2020|EPI ISL 416739|2020-03-09              AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
4. hCoV-19/Luxembourg/GH(Beta)/2020|EPI ISL 419585|2020-03-14    AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
5. hCoV-19/Hungary/GH(Beta)/2020|EPI ISL 671465|2020-10-03       AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
6. hCoV-19/EnglandGR (Gamma)/2020|EPI ISL 423654|2020-03-21      AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
7. hCoV-19/Australia/GR(Gamma)/2020|EPI ISL 563257|2020-08-31    AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
8. hCoV-19/USA/G/2020|EPI ISL 766913|2020-12-20                  AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
9. hCoV-19/Luxembourg/G/2020|EPI ISL 770906|2020-12-01           AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
10. hCoV-19/England/GRY(Alpha)/2020|EPI ISL 659373|2020-11-05    AAATGTCTCTAAATGGACCCCAAAATCAGCGAAATGCA
11. hCoV-19/England/GRY(Alpha)/2021|EPI ISL 939571|2021-01-22    AAATGTCTCTAAATGGACCCCAAAATCAGCGAAATGCA
12. hCoV-19/India/GK(Delta)/2020|EPI ISL 1360317|2020-12-18      AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
13. hCoV-19/England/GK(Delta)/2021|EPI ISL 2516080|2021-06-02    AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
14. hCoV-19/England/GRA(BA.1)/2021|EPI ISL 8271120|2021-12-24    AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
15. hCoV-19/USA/GRA(BA.1)/2021|EPI ISL 8593285|2021-12-30        AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
16. hCoV-19/Denmark/GRA(BA.2)/2021|EPI ISL 8419126|2021-12-2     AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
17. hCoV-19/England/GRA(BA.2)/2022|EPI ISL 9871917|2022-02-07    AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
18. hCoV-19/England/GRA(BA.5)/2022|EPI ISL 12893688|2022-05-1    AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
19. hCoV-19/USA/GRA(BA.5)/2022|EPI ISL 15336666|2022-08-05       AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
20. hCoV-19/Belgium/GRA(BQ.1)/2022|EPI ISL 14773615|2022-08-2    AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
21. hCoV-19/USA/GRA(BQ.1)/2022|EPI ISL 15888367|2022-11-01       AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
22. hCoV-19/USA/GRA(BQ.1.1)/2022|EPI ISL 15001867|2022-09-04     AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
23. hCoV-19/Denmark/GRA(BQ.1.1)/2022|EPI ISL 15881942|2022-1     AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
24. hCoV-19/New Zealand/GRA(XBB)/2022|EPI ISL 15276133|2022      AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
25. hCoV-19/Japan/GRA(XBB)/2022|EPI ISL 15853550|2022-10-25      AAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCA
```

```
1.  AAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGG
2.  AAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGG
3.  AAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGG
4.  AAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGG
5.  AAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGG
6.  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
7.  AAGAAATTCAACTCCAGGCAGCTCTAAACGAACTTCTCCTGCTAGAATGGCTGG
8.  AAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGG
9.  AAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGG
10  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
11  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
12  AAGAAATTCAACTCCAGGCAGCAGTATGGGAACTTCTCCTGCTAGAATGGCTGG
13  AAGAAATTCAACTCCAGGCAGCAGTATGGGAACTTCTCCTGCTAGAATGGCTGG
14  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
15  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
16  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
17  AAGAAATTCAACTCCAGGCAGCTCTAAACGAACTTCTCCTGCTAGAATGGCTGG
18  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
19  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
20  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
21  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
22  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
23  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
24  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
25  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG

1.  AAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGG
2.  AAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGG
3.  AAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGG
4.  AAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGG
5.  AAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGG
6.  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
7.  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
8.  AAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGG
9.  AAGAAATTCAACTCTAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGG
10  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
11  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
12  AAGAAATTCAACTCCAGGCAGCAGTATGGGAACTTCTCCTGCTAGAATGGCTGG
13  AAGAAATTCAACTCCAGGCAGCAGTATGGGAACTTCTCCTGCTAGAATGGCTGG
14  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
15  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
16  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
17  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
18  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
19  AAGAAATTCAACCCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
20  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
21  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
22  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
23  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
24  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
25  AAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCTCCTGCTAGAATGGCTGG
```

Fig. 1. Identification of two sites of trinucleotide exchange in the SARS-CoV-2 *N* gene.

The panels show the results of two sets of multiple sequence alignments with 25 evaluated sequences (one with the original Wuhan-Hu-1 and 24 from subsequent lineages) in each set. Sequence alignments identified two nucleotide exchange sites in three consecutive nucleotides. Both sites were located in the *N* gene. Considering that the sequence identity between the original Wuhan-Hu-1 and the evaluated sequence from the clade GRY (Alpha) was 99.91%, the observation of two sites of three-base exchange based on a gradual accumulation of single nucleotide substitutions is not probable. Rather, a supposition of a new mutation type, an en bloc exchange of short consecutive nucleotides, is needed to explain the observed mutations.

Examination via BLAST for conceivable intermediating mutations

Finally, to estimate the period of each conceivable intermediating sequence, with substitutions in two of the three nucleotide positions for the two locations in the *N* gene, a sequence search was performed using the Basic Local Alignment Search Tool (BLAST) from the NCBI. The obtained results, including the confirmed first date and the location of each intermediate sequence, are summarized in Table 3. In the first location with a trinucleotide substitution at 28,280-28,282 nt, the confirmed oldest mutant with 5'-CTA-3' dated back to May 24th, 2020 (GenBank Accession: ON299968.1), whereas the earliest intermediating sequences with two nucleotide substitutions dated back to December 10th, 2020, which was much later than the emergence of the mutant with the trinucleotide substitution. In the second location with a trinucleotide substitution at 28,881-28,883 nt, the confirmed oldest mutant with

Table 2. Numbers of registered sequences with each amino acid substitution in the GISAID database.

GISAID clades	N_D3L	N_R203K only	N_G204R only	R203K+G204R	Total
Overall, n (%)	1,180,110 (8.24%)	97,097 (0.68%)	20,484 (0.14%)	8,207,171 (57.28%)	14,329,052
L, n (%)	6 (0.09%)	20 (0.30%)	0 (0.0%)	0 (0.0%)	6,605
G, n (%)	19,311 (5.45%)	3,456 (0.98%)	606 (0.17%)	1,366 (0.39%)	354,434
GH (Beta), n (%)	1,236 (0.21%)	848 (0.14%)	62 (0.01%)	3,984 (0.67%)	598,959
GR (Gamma), n (%)	93,775 (16.72%)	11,698 (2.09%)	4,504 (0.80%)	544,472 (97.08%)	560,870
GRY (Alpha), n (%)	1,064,373 (98.80%)	75,058 (6.97%)	503 (0.05%)	998,460 (92.68%)	1,077,272
GK (Delta), n (%)	236 (0.005%)	279 (0.006%)	1,101 (0.02%)	7 (0.0001%)	4,550,807
GRA (Omicron), n (%)	166 (0.002%)	5,330 (0.08%)	13,660 (0.20%)	6,654,537 (95.93%)	6,936,976
sublineage BA.1	23 (0.005%)	303 (0.06%)	893 (0.18%)	471,576 (96.91%)	486,590
sublineage BA.2	24 (0.002%)	1,382 (0.11%)	2,619 (0.21%)	1,213,421 (97.15%)	1,249,032
sublineage BA.5	0 (0.0%)	13 (0.04%)	64 (0.18%)	34,428 (98.48%)	34,961
sublineage BQ.1	0 (0.0%)	5 (0.03%)	11 (0.06%)	18,415 (98.96%)	18,609
sublineage BQ.1.1	0 (0.0%)	15 (0.03%)	60 (0.14%)	43,261 (98.70%)	43,830
sublineage XBB	0 (0.0%)	0 (0.0%)	1 (0.05%)	1,953 (89.71%)	2,177

The number (n) and prevalence (%) of registered SARS-CoV-2 genome sequences collected from humans with each of the three types of amino acid substitutions are listed. The number and rate were obtained within each of the following GISAID clades and sublineages: L, GH (Beta), GR (Gamma), G, GRY (Alpha), GK (Delta), and GRA (Omicron) sublineages BA.1, BA.2, BA.5, BQ.1, BQ.1.1, and XBB.
GISAID, Global Initiative on Sharing All Influenza Data; N, nucleocapsid.

Table 3. The Basic Local Alignment Search Tool (BLAST) search results for the earliest strains with sequences linking the original sequences and mutants with trinucleotide substitution.

Sequence	Confirmed earliest sequence (GenBank Accession ID)	Collection date	Location
The first location at 28,280-28,282 nt, with a change from 5'-GAT-3' to 5'-CTA-3'			
Original (5'-GAT-3')	Wunan-Hu-1 (NC_045512.2)	Dec. 2019	China
5'-CTT-3'	ESP/hCoV-19_Spain_CT-HUVH-Bellvitge60554_2021/2021 (MW769733.1)	Jan. 19, 2021	Spain
5'-CAA-3'	HKG/HKPU-00084/2020 (MZ266433.1)	Dec. 10, 2020	Hong Kong
Mutant (5'-CTA-3')	FRA/IHUCOVID-005143-N1/2020 (ON299968.1)	May 24, 2020	France
The second location at 28,881-28,883 nt, with a change from 5'-GGG-3' to 5'-AAC-3'			
Original (5'-GGG-3')	Wunan-Hu-1 (NC_045512.2)	Dec. 2019	China
5'-AAG-3'	USA/CA-CZB-13501/2020 (MW483620.1)	Oct. 28, 2020	USA
5'-AGC-3'	ESP/hCoV-19_Spain_CT-HUVH-Bellvitge56331_2021/2021 (MW769712.1)	Jan. 19, 2021	Spain
5'-GAC-3'	USA/CA-CDPH-UC407/2020 (MW973159.1)	Nov. 17, 2020	USA
Mutant (5'-AAC-3')	PER/covper071/2020 (MW030211.1)	Mar. 14, 2020	Peru

A BLAST sequence search was performed on February 2nd, 2023. In the first location at 28,280-28,282 nt, the oldest confirmed mutant with trinucleotide substitution dated back to May 24th, 2020, whereas the earliest intermediating sequences with two nucleotide substitutions dated back to December 10th, 2020. In the second location at 28,881-28,883 nt, the confirmed oldest mutant with trinucleotide substitution dated back to March 14th, 2020, whereas the earliest intermediating sequences with two nucleotide substitutions dated back to November 17th, 2020. In both locations, the mutants with a trinucleotide substitution were collected much earlier than other mutants with intermediating sequences, suggesting an en bloc development of the trinucleotide substitutions.

5'-AAC-3' dated back to March 14th, 2020 (GenBank Accession: MW030211.1), whereas the earliest intermediating sequences with two nucleotide substitutions dated back to November 17th, 2020, which was also much later than the emergence of the mutant with the trinucleotide substitution.

Discussion

In the present study, the presence of a possible new type of gene mutation, an en bloc exchange of short consecutive bases, was reported at two sites in the SARS-CoV-2 *N* gene. The possibility of coincidental accumulation of single nucleotide substitutions or overlapping indels was ruled out in the present study by performing a comprehensive BLAST sequence search and GISAID database search for each conceivable intermediating sequence linking the original strain and mutants with trinucleotide substitutions. Consequently, the observed trinucleotide substitutions in the SARS-CoV-2 *N* gene may imply a novel type of mutation, which is different from previously known traditional mutations, such as point mutations, insertions, deletions, inversions, duplications, translocations, or recombinations (Gu et al. 2008; Lee et al. 2012). Currently, the exact developmental mechanisms of sequence exchanges involving dozens of consecutive nucleotides in sarbecoviruses remain uncertain (Akaishi 2022a; Akaishi et al. 2022b); however, the results of the present study imply that en bloc sequence substitutions may have played a role in sarbecovirus evolution. Further studies are warranted to elucidate the presence and roles of en bloc sequence substitutions in viruses.

This study further showed that the prevalence of each trinucleotide substitution remarkably differed between different clades of SARS-CoV-2, suggesting that trinucleotide substitutions may have played a role in the spread of the virus. Currently, profiles of mutations between different variants of concern are primarily compared based on mutations in the S1 RBD. However, mutations in other gene locations outside the S1 RBD, such as the S1 NTD or *N* gene, may also need to be considered. Generally, the *N* gene is highly conserved, and the frequency of mutation occurrence is much lower than that of the *S* gene (Thakur et al. 2022). A previous study that examined a recombinant SARS-CoV-2 alpha variant with cloning techniques suggested that the R203K + G204R mutation could increase viral replication and enhance the pathogenesis (Johnson et al. 2022). More specifically, the R203K + G204R mutation is located in the serin-rich domain of the *N* gene, and the phosphorylation level of this domain is considered to regulate nucleocapsid function via the liquid-liquid phase separation (Carlson et al. 2020). By modulating the phosphorylation level of the nucleocapsid, trinucleotide substitution may increase viral fitness with enhanced adaptation in humans.

Currently available genetic engineering technologies, including the clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated system (Jinek et al. 2012; Ran et al. 2013), seem to be unable to realize en bloc exchange of short consecutive nucleotides in a single-stranded RNA genome without performing additional complicated processes, such as preparing a double-stranded DNA sequence and exchanging a specific position. Some unknown mechanisms that realize short-sequence exchanges in single-stranded RNA molecules may exist in natural environments, including host cells. Moreover, the mutation profiles of sarbecoviruses differ significantly from those of other virus species, such as adenoviruses or influenza viruses, with different frequencies and lengths of indels (Akaishi 2022b). This fact suggests another possibility that short sequence exchange is a phenomenon specific to some viruses, and the virus genomes may encode molecules that realize such mutations. SARS-CoV-2 RNA-dependent RNA polymerases (RdRp) play a major role in the replication machinery, and replication fidelity is remarkably influenced by mutations in some non-structural proteins (Eckerle et al. 2010; Pachetti et al. 2020). Further studies are needed to determine whether the viral replication machinery is a key player in the development of trinucleotide substitutions in the *N* gene.

The present study has several limitations. First, we only identified exchanges of three consecutive bases. Future studies are needed to determine whether there are en bloc sequence exchanges involving > 3 consecutive nucleotides in SARS-related coronaviruses or other organisms. Another limitation was that the present study only evaluated the genome sequences of samples derived from humans, and whether the observed three-base exchange occurred only in humans or also in other animal hosts could not be determined. Furthermore, whether the developmental mechanisms of such en bloc exchange mutations are coded in the viral genome itself or are realized by the transcription machinery of the host cell remains unknown. Studies are required to determine the exact mechanisms associated with en bloc short-sequence exchange in SARS-related coronaviruses. Lastly, a research has shown how large populations, especially those with high mutation rates, can seem to fix multiple mutations simultaneously (Weinreich and Chao 2005), which is typically observed to avoid low-fitness intermediates. Similarly, studies have examined polymerase errors, and it is not uncommon for the polymerase to make mistakes with sites located close together (Drake 2007). These facts render it difficult to conclude that the observed trinucleotide substitutions in the SARS-CoV-2 *N* gene truly developed from an en bloc sequence exchange at once.

In summary, the present study identified two locations of trinucleotide substitutions in the SARS-CoV-2 *N* gene, which were difficult to explain using traditional single nucleotide substitutions and/or indels. Further studies are warranted to determine the exact mechanisms underlying the substitution of continuous nucleotides.

Acknowledgments

We gratefully acknowledge all data contributors, i.e., the authors and their originating laboratories responsible for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

Author Contributions

T.A. conceived, performed analyses, and drafted the manuscript. K.F. performed analyses, verified the results, and critically reviewed and revised the manuscript. T.I. supervised the study, and critically reviewed and revised the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

References

Akaishi, T. (2022a) Insertion-and-deletion mutations between the genomes of SARS-CoV, SARS-CoV-2, and bat coronavirus RaTG13. *Microbiol. Spectr.*, **10**, e0071622.

Akaishi, T. (2022b) Comparison of insertion, deletion, and point mutations in the genomes of human adenovirus HAdvC-2 and SARS-CoV-2. *Tohoku J. Exp. Med.*, **258**, 23-27.

Akaishi, T., Fujiwara, K. & Ishii, T. (2022a) Variable number tandem repeats of a 9-base insertion in the N-terminal domain of severe acute respiratory syndrome coronavirus 2 spike gene. *Front. Microbiol.*, **13**, 1089399.

Akaishi, T., Horii, A. & Ishii, T. (2022b) Sequence exchange involving dozens of consecutive bases with external origin in SARS-related coronaviruses. *J. Virol.*, **96**, e0100222.

Araf, Y., Akter, F., Tang, Y.D., Fatemi, R., Parvez, M.S.A., Zheng, C. & Hossain, M.G. (2022) Omicron variant of SARS-CoV-2: genomics, transmissibility, and responses to current COVID-19 vaccines. *J. Med. Virol.*, **94**, 1825-1832.

Carlson, C.R., Asfaha, J.B., Ghent, C.M., Howard, C.J., Hartooni, N., Safari, M., Frankel, A.D. & Morgan, D.O. (2020) Phosphoregulation of phase separation by the SARS-CoV-2 N protein suggests a biophysical basis for its dual functions. *Mol. Cell*, **80**, 1092-1103.e1094.

Drake, J.W. (2007) Mutations in clusters and showers. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 8203-8204.

Eckerle, L.D., Becker, M.M., Halpin, R.A., Li, K., Venter, E., Lu, X., Scherbakova, S., Graham, R.L., Baric, R.S., Stockwell, T.B., Spiro, D.J. & Denison, M.R. (2010) Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. *PLoS Pathog.*, **6**, e1000896.

Elbe, S. & Buckland-Merrett, G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.*, **1**, 33-46.

European Molecular Biology Laboratory (2023) Multiple Sequence Alignment. (MAFFT).
https://www.ebi.ac.uk/Tools/msa/mafft/
[*Accessed*: January 11, 2023].

GISAID (2023) Tracking of hCoV-19 Variants.
https://gisaid.org/hcov19-variants/
[*Accessed*: January 11, 2023].

Gu, W., Zhang, F. & Lupski, J.R. (2008) Mechanisms for human genomic rearrangements. *Pathogenetics*, **1**, 4.

Hajizadeh, F., Khanizadeh, S., Khodadadi, H., Mokhayeri, Y., Ajorloo, M., Malekshahi, A. & Heydari, E. (2022) SARS-COV-2 RBD (receptor binding domain) mutations and variants (a sectional-analytical study). *Microb. Pathog.*, **168**, 105595.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. & Charpentier, E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816-821.

Johns Hopkins University (2023) COVID-19 Dashboard.
https://coronavirus.jhu.edu/map.html
[*Accessed*: January 11, 2023].

Johnson, B.A., Zhou, Y., Lokugamage, K.G., Vu, M.N., Bopp, N., Crocquet-Valdes, P.A., Kalveram, B., Schindewolf, C., Liu, Y., Scharton, D., Plante, J.A., Xie, X., Aguilar, P., Weaver, S.C., Shi, P.Y., et al. (2022) Nucleocapsid mutations in SARS-CoV-2 augment replication and pathogenesis. *PLoS Pathog.*, **18**, e1010627.

Karim, S.S.A. & Karim, Q.A. (2021) Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. *Lancet*, **398**, 2126-2128.

Khare, S., Gurry, C., Freitas, L., Schultz, M.B., Bach, G., Diallo, A., Akite, N., Ho, J., Lee, R.T., Yeo, W. & Maurer-Stroh, S.; Gisaid Core Curation Team (2021) GISAID's role in pandemic response. *China CDC Wkly.*, **3**, 1049-1051.

Lee, H.J., Kweon, J., Kim, E., Kim, S. & Kim, J.S. (2012) Targeted chromosomal duplications and inversions in the human genome using zinc finger nucleases. *Genome Res.*, **22**, 539-548.

Mlcochova, P., Kemp, S.A., Dhar, M.S., Papa, G., Meng, B., Ferreira, I., Datir, R., Collier, D.A., Albecka, A., Singh, S., Pandey, R., Brown, J., Zhou, J., Goonawardane, N., Mishra, S., et al. (2021) SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature*, **599**, 114-119.

Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., Masciovecchio, C., Angeletti, S., Ciccozzi, M., Gallo, R.C., Zella, D. & Ippodrino, R. (2020) Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.*, **18**, 179.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A. & Zhang, F. (2013) Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.*, **8**, 2281-2308.

SeyedAlinaghi, S., Mirzapour, P., Dadras, O., Pashaei, Z., Karimi, A., MohsseniPour, M., Soleymanzadeh, M., Barzegary, A., Afsahi, A.M., Vahedi, F., Shamsabadi, A., Behnezhad, F., Saeidi, S., Mehraeen, E. & Shayesteh, J. (2021) Characterization of SARS-CoV-2 different variants and related morbidity and mortality: a systematic review. *Eur. J. Med. Res.*, **26**, 51.

Shu, Y. & McCauley, J. (2017) GISAID: global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.*, **22**, 30494.

Tamura, K., Stecher, G. & Kumar, S. (2021) MEGA11: molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.*, **38**, 3022-3027.

Tao, K., Tzou, P.L., Nouhin, J., Gupta, R.K., de Oliveira, T., Kosakovsky Pond, S.L., Fera, D. & Shafer, R.W. (2021) The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. Rev. Genet.*, **22**, 757-773.

Thakur, S., Sasi, S., Pillai, S.G., Nag, A., Shukla, D., Singhal, R., Phalke, S. & Velu, G.S.K. (2022) SARS-CoV-2 mutations and their impact on diagnostics, therapeutics and vaccines. *Front. Med.* (*Lausanne*), **9**, 815389.

Weinreich, D.M. & Chao, L. (2005) Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. *Evolution*, **59**, 1175-1182.