Archives of Microbiology (2023) 205:154
https://doi.org/10.1007/s00203-023-03493-0

ORIGINAL PAPER

Insertion and deletion mutations preserved in SARS-CoV-2 variants

Tetsuya Akaishi[1,2] · **Kei Fujiwara**[3]

Received: 24 January 2023 / Revised: 16 March 2023 / Accepted: 18 March 2023 / Published online: 31 March 2023

Abstract
The insertion/deletion (indel) mutation profiles of SARS-CoV-2 variants, including Omicron, remain unclear. We compared whole-genome sequences from various lineages and used preserved indels to infer the ancestral relationships between different lineages. Thirteen indel patterns from twelve sites were seen in ≥ 2 sequences; six of these sites were located in the N-terminal domain of the viral spike gene. Preserved indels in the coding regions were also identified in the non-structural protein 3 (*Nsp3*), *Nsp6*, and nucleocapsid genes. Seven of the thirteen indel patterns were specific to the Omicron variants, four of which were observed in BA.1, making it the most mutated variant. Other preserved indels observed in the Omicron variants were also seen in Alpha and/or Gamma, but not Delta, suggesting that Omicron is phylogenetically more proximal to Alpha. We demonstrated distinct profiles of preserved indels among SARS-CoV-2 variants and sublineages, suggesting the importance of indels in viral evolution.

Keywords Insertion\deletion (indel) · Origin · Omicron variant · Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) · Virus reservoirs

Introduction

The world is still struggling against COVID-19, which is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The largest outbreak wave was caused by the emergence of the Omicron variant in early 2022 (Johns Hopkins University 2023), and the pandemic was maintained by the intermittent emergence of various sublineages in the subsequent months. The mutation profiles in the genome of the Omicron variants are remarkably different from those of previous variants, such as Alpha, Beta, Gamma, or Delta (He et al. 2020; Kandeel et al. 2022; Konishi 2022; Shrestha et al. 2022), and the origin of Omicron has been vigorously discussed since the emergence of the variant, which has not concluded yet (Wei et al. 2021; Du et al. 2022; Sun et al. 2022). The genome sequence of SARS-CoV-2 has been reported to incorporate a large number of point mutations and structural variants with insertions/deletions (indels) compared to its supposed predecessor SARS-related coronaviruses, such as SARS-CoV in 2002–2003 or bat coronavirus RaTG-13 detected in 2013 (Akaishi 2022b). The ratio between point mutations and indels has been reported to differ significantly among different viral types (Akaishi 2022a). Furthermore, indels in SARS-related coronaviruses have been reported to be concentrated in specific genomic positions, including the N-terminal domain (NTD) of the spike (*S*) gene, thoroughly exchanging dozens of bases at the involved sites (Akaishi et al. 2022). In the early stages of the COVID-19 pandemic, convergent indels have been suggested in the NTD of the SARS-CoV-2 S1 subdomain (S1-NTD), which could have altered viral antigenicity and promoted immune escape with enhanced resistance against neutralizing antibodies (Resende et al. 2021). However, the distribution and profiles of indels among different SARS-CoV-2 variants during the COVID-19 pandemic remain largely unknown. Therefore, the present study comprehensively evaluated the insertion/deletion mutation profiles of different SARS-CoV-2 lineages, including the latest

Communicated by Yusuf Akhter.

✉ Tetsuya Akaishi
t-akaishi@med.tohoku.ac.jp

[1] Department of Education and Support for Regional Medicine, Tohoku University, Seiryo-Machi 1-1, Aoba-Ku, Sendai, Miyagi 980-8574, Japan

[2] COVID-19 Testing Center, Tohoku University, Sendai, Japan

[3] Department of Gastroenterology and Metabolism, Nagoya City University, Nagoya, Japan

Omicron sublineages and recombinant, using a large worldwide sequence database.

Methods

Initial evaluation of 25 sequences for screening indels

A total of 14,329,052 human genome sequences, which were registered and available in the Global Initiative on Sharing All Influenza Data (GISAID) database up to December 22, 2022, were evaluated in the present study (Elbe and Buckland-Merrett 2017; Shu and McCauley 2017; Khare et al. 2021). The associated EPI_SET ID is specified in the subsequent data availability statement. From these available SARS-CoV-2 sequences, two sequences from each of the following clades were randomly selected: clades L (PANGO lineage: B), G (B.1), GH (B.1.; Beta), GR (Gamma), GRY (Alpha), GK (Delta), and GRA (Omicron). Two sequences from each of the following lineages were selected from the GRA clade: BA.1, BA.2, BA.5, BQ.1, BQ1.1, and recombinant XBB.

Initial multiple sequence alignments for screening indels

Using the initially collected 25 viral genome sequences (1 original Wuhan-Hu-1 sequence and 2 sequences from each clade of different SARS-CoV-2 lineages), multiple sequence alignment was performed using the Molecular Evolutionary Genetics Analysis Version 11 (MEGA11) software (Tamura et al. 2021). Multiple Sequence Comparison by Log-Expectation (MUSCLE) was performed to align the whole-genome sequences. For the alignment parameters, the gap opening penalty score was set to − 400 and the gap extension penalty score was set to 0. Based on the aligned sequences, indel sites across the whole genome were selected based on the presence of indels in two or more selected sequences. Furthermore, phylogenetic analysis was performed using multiple aligned sequences to estimate the ancestral relationship between the 25 SARS-CoV-2 strains evaluated. A phylogenetic tree was constructed with 100 bootstrap replicates.

GISAID database search for each indel

To determine whether the indels identified in the initial screening alignments matched the peak of indel distributions in the overall registered sequences, the line graphs for the amino acid position-specific number of indels in the SARS-CoV-2 S1-NTD are depicted. We evaluated whether the peaks of the line graphs matched the positions of the identified indels. The distribution of these indel peaks was compared with that of previously reported common point mutations to infer the relationship between structural variants and point mutations in the SARS-CoV-2 S1-NTD.

Results

Evaluation of 25 SARS-CoV-2 sequences

First, the genome sequences of 25 SARS-CoV-2 variants, including the original Wuhan-Hu-1 strain, were selected. The genome sequence of Wuhan-Hu-1 was obtained from the NCBI GenBank database (Wu et al. 2020), whereas the other 24 sequences were obtained from the GISAID database; 2 sequences were randomly selected from each of the GISAID clades L, GH (Beta), GR (Gamma), G, GRY (Alpha), and GK (Delta). Moreover, two sequences were randomly selected from each of the following PANGO lineages in clade GRA (Omicron): BA.1, BA.2, BA.5, BQ.1, BQ.1.1, and recombinant XBB. The 25 sequences evaluated are listed in Table 1.

Phylogenetic reconstruction of the 25 SARS-CoV-2 sequences evaluated

First, to investigate the ancestral state among the 25 sequences from different SARS-CoV-2 variants, phylogenetic tree reconstruction was performed using MEGA11 software. The reconstructed phylogenetic tree is shown in Fig. 1. The results indicated that the SARS-CoV-2 variants belonging to clades L, GH (Beta), GR (Gamma), G, and GRY (Alpha) were phylogenetically more proximal to the original Wuhan-Hu-1 genome than the viruses from clades GK (Delta) and GRA (Omicron). The viruses from clades GK (Delta) and GRA (Omicron) were the most phylogenetically distant from each other, suggesting a divergence of the Delta variant predecessors in the early stages of the pandemic, supposedly before December 2020 (Lopez Bernal et al. 2021; Cascella et al. 2022).

Identification of preserved indels

The indels identified across the genomes of the evaluated 25 sequences, which were identified in more than 2 of the 25 sequences, are listed in Table 2. A total of 12 indel sites with 13 indel patterns were identified: 6 from *S1-NTD*, 3 from non-coding areas, and 1 each from the non-structural protein 3 (*Nsp3*), *Nsp6*, and nucleocapsid (*N*). Seven of the

Table 1 List of SARS-CoV-2 strains from different clades and lineages used in screening multiple alignments

No	Location	Collection date	Clade (PANGO lineages)	GISAID accession ID
1	Wuhan, China	2019-12	Original	(Wuhan-Hu-1; GenBank NC_045512)
2	Beijing, China	2020-01-27	L (B)	EPI_ISL_16138010
3	Ontario, Canada	2020-01-23	L (B)	EPI_ISL_413015
4	Illinois, USA	2020-05-07	GH (Beta; B.1.)	EPI_ISL_16119477
5	England, UK	2020-11-01	GH (Beta; B.1.)	EPI_ISL_647686
6	Rio de Janeiro, Brazil	2020-10-08	GR (Gamma)	EPI_ISL_9754628
7	Sao Paulo, Brazil	2021-03-16	GR (Gamma)	EPI_ISL_16225655
8	Lorestan, Iran	2020-05-27	G (B.1)	EPI_ISL_15755020
9	Illinois, USA	2020-04-09	G (B.1)	EPI_ISL_14379838
10	England, UK	2021-01-19	GRY (Alpha)	EPI_ISL_916362
11	British Columbia, Canada	2021-05-18	GRY (Alpha)	EPI_ISL_3486562
12	Gran Canaria, Canary Islands	2021-08-16	GK (Delta)	EPI_ISL_16215541
13	Hauts-de-France, France	2021-08-29	GK (Delta)	EPI_ISL_11233238
14	Gran Canaria, Canary Islands	2022-01-24	GRA (BA.1)	EPI_ISL_16216152
15	Nouvelle-Aquitaine, France	2022-03-14	GRA (BA.1)	EPI_ISL_11425679
16	Utah, USA	2022-07-04	GRA (BA.2)	EPI_ISL_14335247
17	Fukuoka, Japan	2022-08-25	GRA (BA.2)	EPI_ISL_16175961
18	Maranhao, Brazil	2022-11	GRA (BA.5)	EPI_ISL_16145048
19	England, UK	2022-05-19	GRA (BA.5)	EPI_ISL_12919999
20	California, USA	2022-11-08	GRA (BQ.1)	EPI_ISL_16222509
21	New York, USA	2022-08-22	GRA (BQ.1)	EPI_ISL_14850768
22	California, USA	2022-11-28	GRA (BQ.1.1)	EPI_ISL_16222134
23	New York, USA	2022-09-06	GRA (BQ.1.1)	EPI_ISL_15029320
24	Sjaelland, Denmark	2022-11-13	GRA (XBB)	EPI_ISL_15826137
25	New Zealand	2022-09-05	GRA (XBB)	EPI_ISL_15276133

thirteen observed indel types (53.85%) were specific to the Omicron sublineages, suggesting the presence of unknown viral reservoirs of Omicron predecessors between 2021 and 2022, possibly linking the remote phylogenetic branches between the Omicron and other known previous variants. Especially, four (30.77%) were only observed in the first Omicron sublineage, BA.1, suggesting a distinct ancestral origin for BA.1, which could have been diverged from predecessors of other subsequent Omicron sublineages.

Characteristic indels that could be useful for inferring phylogenetic structures among the evaluated variants are shown in Fig. 2. In *Nsp6* (nt 11,228–11,236; 9-base deletion) and *S1-NTD* (nt 21,765–21,770; 6-base deletion), indels highly specific to the Alpha (and Gamma) variants were preserved in many Omicron variant sequences. In particular, the six-base deletion in *S1-NTD* was preserved in the Omicron sublineages BA.1, BA.5, BQ.1, and BQ.1.1, but not in BA.2, or recombinant XBB. The three-base deletion in nucleotides 21,987–21,989 of *S1-NTD* was observed in both Alpha and Omicron sublineage recombinant XBB. However, these could have occurred independently at exactly the same position based on the point mutation (G > A) at four nucleotides upstream of the indel. At this indel site, a nine-base deletion was observed in both sequences from the Omicron sublineage BA.1, which could have developed in the very early stage after the divergence of BA.1 from the reservoirs of the Omicron predecessors.

The nine-base deletion in *Nsp6* corresponds to either (A) NSP6_S106_F108del (i.e., S106del, G107del, and F108del) or (B) NSP6_L105_G107del (i.e., L105del, S106del, and G107del), and this nine-base deletion was observed in only one of the two sequences from the Gamma variants. Based on this finding, the total number of registered sequences with this nine-base deletion in the GISAID database (until December 22, 2022) was evaluated for each of the following clades: GR (Gamma), GRY (Alpha), GRA (Omicron) BA.1, GRA BA.2, and GRA XBB. As a result, the nine-base deletion was incorporated in 40.58% (n = [A] 226,577 + [B] 1035/560,870 sequences) of the GR clade, 98.49% (n = [A] 1,060,915 + [B] 142/1,077,272 sequences) of the GRY clade, 94.89% (n = [A] 832 + [B] 460,913/486,590 sequences) of the Omicron sublineage BA.1, 96.03%

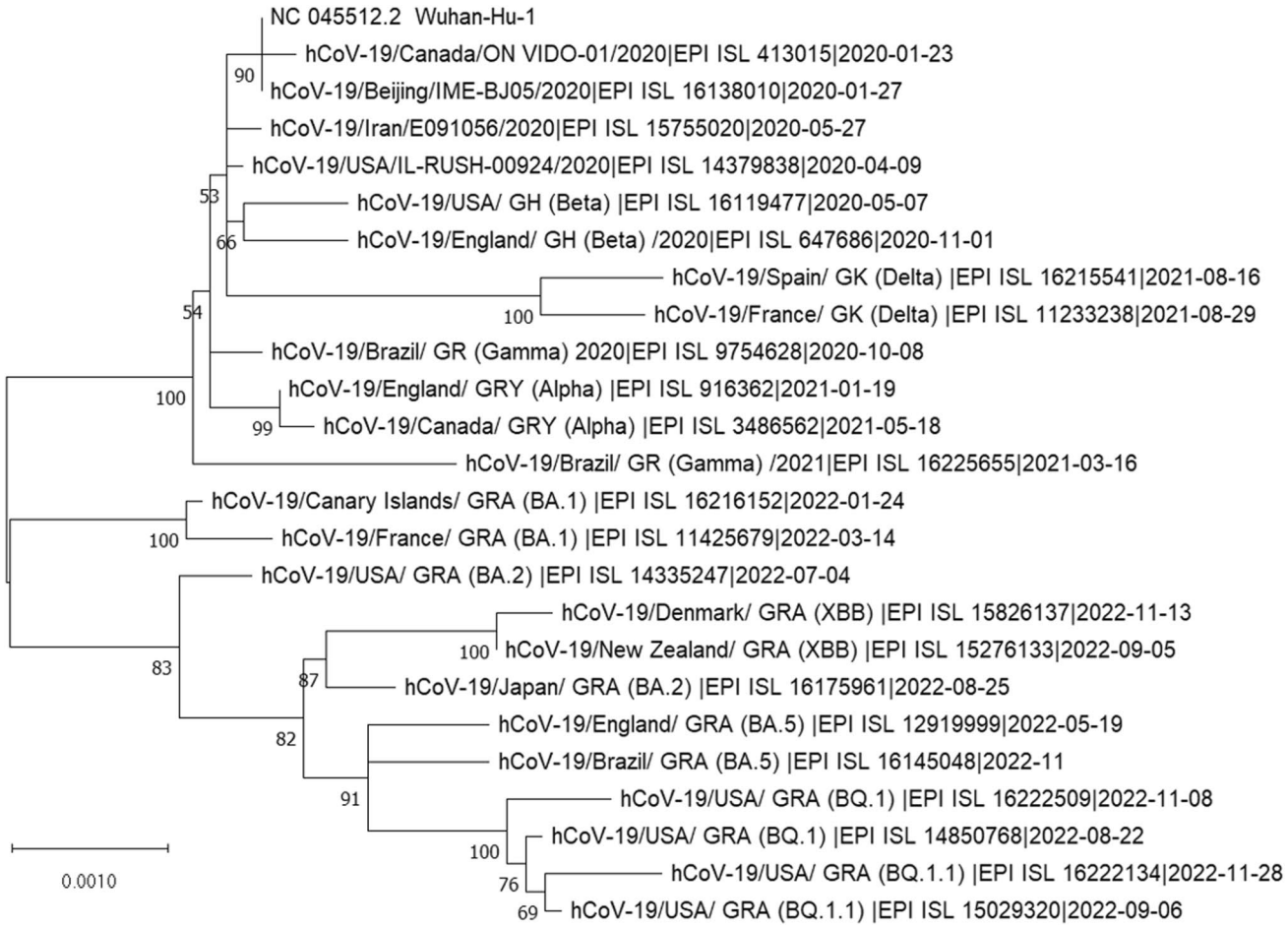

Fig. 1 Phylogenetic analysis of 25 SARS-CoV-2 strains. The phylogenetic maximum-likelihood tree for the full genomes of the 25 viruses used in this study was generated using MEGA11 software after multiple sequence alignments were obtained from 100 bootstrap replicates. Bootstrap values are shown to the left of major nodes. The branch length represents the genetic distance measured by the number of nucleotide substitutions per site. The results of this phylogenetic reconstruction suggest a relatively close ancestral proximity between the original Wuhan-Hu-1, Alpha, Beta, and Gamma variant strains. Both the Delta and Omicron variant strains were phylogenetically distant from these conventional viruses. The Delta and Omicron variants were phylogenetically the most distant, suggesting that the Delta and Omicron predecessors created different viral reservoirs with low levels of genetic recombination

(n = [A] 1,197,346 + [B] 2090/1,249,032 sequences) of the Omicron sublineage BA.2, and 81.81% (n = [A] 1,781 + [B] 0/2,177 sequences) of the Omicron sublineage recombinant XBB. Among the sequences in clade GR (Gamma), 58.91% (n = 330,434/560,870) of the registered sequences did without this nine-base deletion.

Amino acid position-specific numbers of indels in S1-NTD

Finally, to confirm whether the common indel sites identified in the present study matched the indel hotspots among the overall registered SARS-CoV-2 genome sequences, the number of registered sequences in the GISAID database was checked for each amino acid position in S1-NTD between amino acid positions 120–230, in which the preserved indels were most densely observed. The obtained numbers are shown as line graphs in Fig. 3, together with the positions of common point mutation sites in the same domain. In the present study, the identified preserved indels in S1-NTD exactly matched the distribution peaks of the site-specific numbers of registered sequences among the overall registered sequences. Furthermore, the distribution of indels approximately matched the distribution of previously reported common point mutations, suggesting a common developmental process of point mutations and indels in this domain. Although the preserved indels were concentrated in limited locations of the genome sequence, non-preserved

Table 2 List of indels observed in ≥ 2 evaluated sequences

No	Gene	Nucleotide position in Wuhan-Hu-1 genome [nt]	AA substitutions	Observed lineages
1	Nsp3	nt 6513–6515 (3-base deletion)	NSP3_S12265del	GRA (BA.1)
2	Nsp6	nt 11,288–11,296 (9-base deletion)	NSP6_S106_F108del	GR, GRY, GRA (BA.1, BA.2, BA.5, BQ.1, BQ.1.1, XBB)*
3	S1-NTD	nt 21,633–21,641 (9-base deletion)	Spike_L24_P26del	GRA (BA.2, BA.5, BQ.1, BQ.1.1, XBB)
4	S1-NTD	nt 21,765–21,770 (6-base deletion)	Spike_H69_V70del	GRY, GRA (BA.1, BA.5, BQ.1, BQ.1.1)[a]
5	S1-NTD	nt 21,987–21,995 (polymorphic)[a]	3-base deletion: Spike_Y144del 9-base deletion: Spike_V143_Y145del	3-base deletion: GRY, GRA (XBB) 9-base deletion: GRA (BA.1)
6	S1-NTD	nt 22,029–22,034 (6-base deletion)	Spikle_E156_F157del	GK
7	S1-NTD	nt 22,194–22,196 (3-base deletion)	Spike_N211del	GRA (BA.1)
8	S1-NTD	nt 22,204/22205 (9-base insertion)	Spike_ins_214EPE	GRA (BA.1)
9	Non-coding	nt 28,248–28,253 (6-base deletion)	Non-coding (3′-terminal of ORF8)	GK
10	Non-coding	nt 28,271 (1-base deletion)	Non-coding (5′-upstream of N)	GRY, GK
11	N	nt 28,362–28,370 (9-base deletion)	N_E31_S33del	GRA (BA.1, BA.2, BA.5, BQ.1, BQ.1.1, XBB)
12	Non-coding	nt 29,734–29,759 (26-base deletion)	Non-coded 3′-terminal region	GRA (BA.2, BA.5, BQ.1, BQ.1.1, XBB)[b]

*This deletion in GR (Gamma) was observed in one of the two sequences. This mutation is associated with interference in the interferon pathway, allowing the virus to evade the innate immune response

[a]Not seen in clades BA.2 and recombinant XBB

[b]This deletion was also observed in one of two sequences from the lineage BA.2

indels were extensively observed across the whole genome. This finding suggests that indels occur frequently and randomly across the genome sequence, and only a limited number survive and spread in the population.

Discussion

In this study, the distribution, type, and prevalence of preserved indels in SARS-CoV-2 variants across the genome were comprehensively evaluated using the GISAID database. Approximately half of the identified preserved indels were specific to the Omicron variants and were not observed in the present variants, whereas other indels in the Omicron variants were also observed in previous Alpha and/or Gamma variants. Meanwhile, we could not identify common indel traits shared between the Delta and Omicron variants, suggesting a remote phylogenetic origin between the Alpha/Gamma/Omicron and Delta variants. Moreover, among the Omicron sublineages, the most mutated variant in view of indel profiles was BA.1, compared to subsequent sublineages BA.2, BA.5, BQ.1, BQ.1.1, and recombinant XBB. The suggested ancestral states among the evaluated SARS-CoV-2 variant lineages based on the observed indel characteristics of each variant, utilizing the concept of virus reservoirs comprising the predecessors of each variant, are summarized in Fig. 4. The Beta clade was estimated to have branched from the main reservoirs before the divergence of the Alpha/Gamma and Delta predecessors, which is estimated to have occurred before December 2020 (Lopez Bernal et al. 2021; Cascella et al. 2022). As there were no characteristic indels between the sequences from the Delta clade and those from other clades in the present study, it is reasonable to separate the virus reservoirs for the Alpha/Gamma/Omicron and Delta variants. However, the presence of unknown virus reservoirs with Omicron predecessors that can link Omicron with Alpha/Gamma, which remained unidentified until late 2021, is needed. Identifying the presence and hosts of such unidentified viral reservoirs is important for preventing the emergence of further variants of concern (VOC). Possible hosts of such unknown viral reservoirs may include animals in natural environments, remote isolated human communities where virus detection is rarely performed, and immunocompromised humans, such as those infected with the human immunodeficiency virus (HIV) (Burki 2022; Du et al. 2022; Mallapaty 2022; Tarcsai et al. 2022). Previous studies have found that mutation profiles in the SARS-CoV-2 genomes sampled from immunocompromised patients with HIV who were chronically infected with COVID-19 resembled those found in the viral genomes of the Omicron variants (Hoffman et al. 2021; Cele et al. 2022; Sonnleitner et al. 2022). Animal hosts are also promising hypotheses, as SARS-CoV-2 has been reported to infect humans and a wide range of animals, including dogs, cats, mice, hamsters, and primates (Mahdy et al. 2020; Sun et al. 2022). Such cross-species transmission has been suggested to facilitate

Fig. 2 Characteristic indel sites suggesting ancestral relationship between SARS-CoV-2 variants. The multiple sequence alignments were performed using MEGA11 software, with a gap opening penalty score of -400 and gap extension penalty score of 0. The aligned sequences indicated a close ancestral proximity between the Alpha or Gamma predecessors and the Omicron predecessors from the viewpoint of insertion/deletion (indel) mutation patterns. In contrast, the indel pattern traits in the Delta variant strains could not be identified in the Omicron variant strains. The three-base deletion in *S1-NTD* of the XBB lineages could have been independently generated from the three-base deletion in the same amino acid position of the Alpha variant strains, considering the point mutation pattern (G > A) at four nucleotides upstream from this deletion site. Blue boxes show sequences for Gamma (top), Alpha (middle), and Omicron (bottom) variant strains

the rapid adaptation of the virus to new host species and the emergence of novel variant strains (Bashor et al. 2021). As the present study demonstrated, indels occur frequently and randomly at every nucleotide position in the SARS-CoV-2 variant genomes, with a comparable rate of point mutations, and VOCs may further emerge in the future as long as unidentified viral reservoirs survive in the environment. Further studies with animal hosts in the natural environment or immunocompromised humans with chronic COVID-19 infection are needed to identify the presence of unidentified viral reservoirs. In these studies, comparing the profiles of the characteristic indels, as performed in the current study, may be beneficial for estimating the ancestral state and phylogenetic structures between newly identified viruses and those already sequenced among humans.

Another notable finding of the present study was that the preserved indels identified in the SARS-CoV-2 variants evaluated were located in specific coding regions, including the *S1-NTD* and the *N* genes, both of which were identified as indel hotspots in SARS-related coronaviruses sampled before the COVID-19 pandemic (Akaishi 2022b). These genomic regions are also hotspots for point mutations in SARS-CoV-2 variants sampled from humans and other animals, suggesting the presence of positive selection pressures in viruses with mutations in these genome regions (Bashor et al. 2021). Theoretically, the probability of insertions or deletions in the viral genome is equal for all nucleotide positions. However, because many indels are fatal for the virus or vulnerable to negative selective forces, only a limited number of indels are preserved in the descendant viral lineages. A possible theory may include

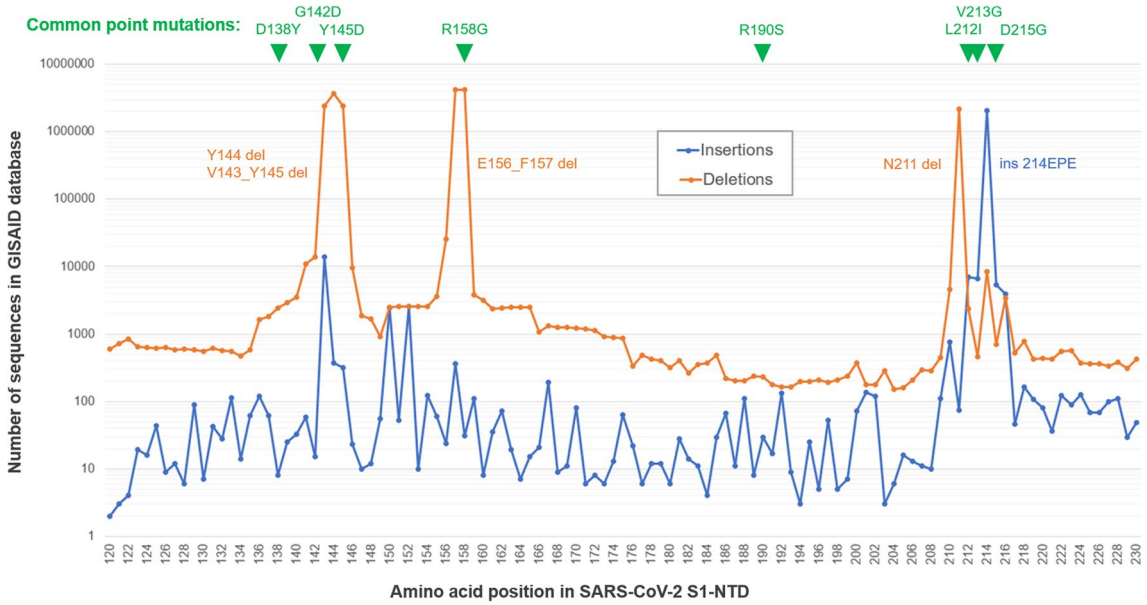

Fig. 3 Position-specific numbers of insertions/deletions in S1-NTD, registered in GISAID database. The sequences were registered and available up to December 22, 2022. All identified insertions/deletions (indels) created sharp peaks in the line graphs, suggesting the reliability of the sampling process of genome sequences in this study. The peaks of the indels approximately matched the distributions of the previous common point mutations, which are shown in green at the top of this figure. This finding suggests that the spread and survival of indels share common mechanisms with the spread and survival of point mutations, such as immune escape or changes in receptor connectivity

the different rates of error in proofreading between different genes. Other theories include the effects of positive selection pressure, possibly based on enhanced viral infectivity or immune escape. A previous study using molecular dynamics simulation implied that mutations in S1-NTD may play an initial role in the kinetics of virus adhesion to host lipid raft gangliosides (Fantini et al. 2021). Currently, either theory based on enhanced viral adhesion to the host cell surface and immune escape is promising to explain why indels are frequently preserved in *S1-NTD* and *N* genes.

The present study has some limitations; it only included 25 strains from various SARS-CoV-2 lineages. Complete genome sequencing data are required to obtain detailed genome-wide indel profiles. Hence, the sequence data used here could be biased by different levels of complete genome sequencing from different areas and countries and may not fully represent the overview of the indel profiles of SARS-CoV-2 prevailing worldwide. Certainly, although we tried to select the initial 25 sequences as randomly as possible in time and location, most of them were the sequences from the developed countries. Further studies analyzing sequences collected in other developing countries are needed to conclude the finding that structural variants with indels are likely to be preserved in several specific SARS-CoV-2 genome regions, such as S1-NTD and *N* gene.

In summary, the present study demonstrated distinct profiles of preserved indels among different SARS-CoV-2 variants. SARS-related coronaviruses and SARS-CoV-2 variants incorporate indel hotspots in similar specific genome positions, including the *S1-NTD* and *N* genes. Together with point mutations, indels in *S1-NTD* in SARS-related coronaviruses may alter their antigenicity and enhance their immune escape. Future studies are warranted to elucidate the potential role of the frequent occurrence of indels in these genes in the evolution of SARS-related coronaviruses and SARS-CoV-2.

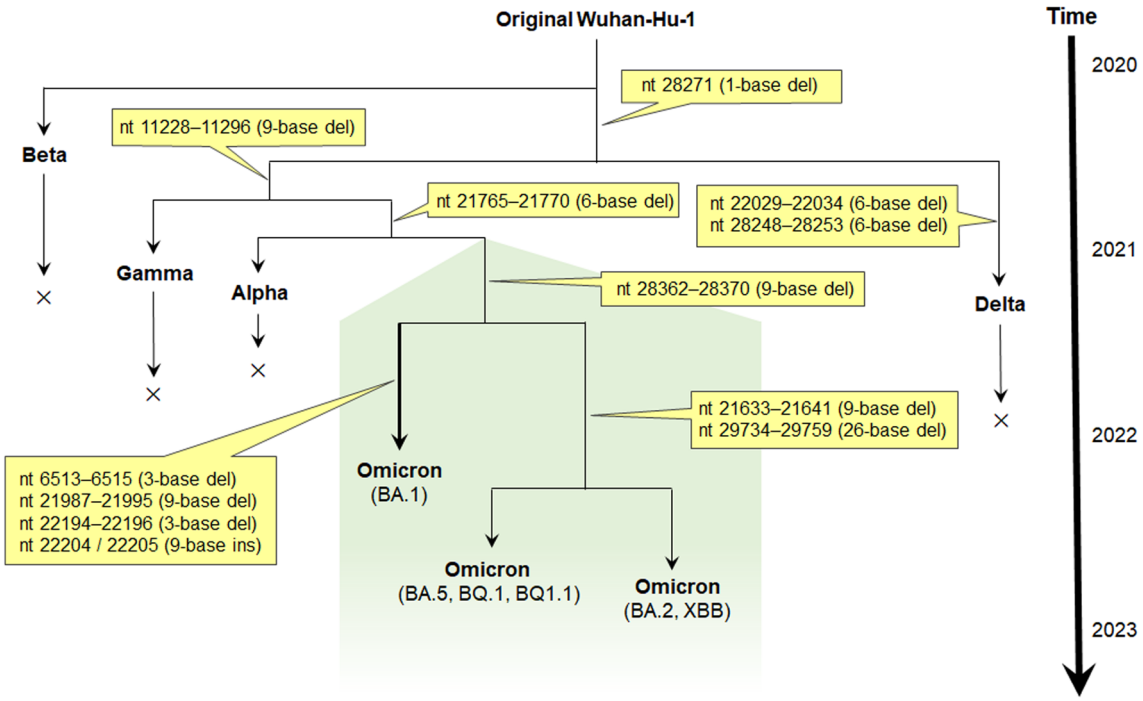

Fig. 4 Estimated ancestral states in SARS-CoV-2 variants from observed variant-specific indels. The results of the present study, based on indel profile analyses, suggest common features between the Alpha, Gamma, and Omicron predecessors. The estimated phylogenetic structures of these three variants based on indel profiles are shown. Traits of the characteristic indels specific to the Delta variant were not identified in these three variants. Among the Omicron variant sublineages, BA.1 was the most mutated in view of indels, suggesting the presence of some accelerating factors for the development of indels in the Omicron BA.1 predecessors. Current promising hypotheses for undiscovered viral reservoirs with Omicron predecessors include immunocompromised hosts with chronic COVID-19 infection, isolated human groups, and animals in natural environments

Acknowledgements 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 TA and KF contributed to the conception, data collection, data purification, and data analysis. TA drafted the manuscript. KF critically reviewed and revised the manuscript.

Funding The present study was not funded.

Data availability The findings of this study are based on metadata associated with 14,329,052 sequences that were sampled from humans and available up to December 22, 2022, on GISAID at gisaid.org/EPI_SET_230112cr.

Declarations

Conflict of interest The authors declare no conflict of interest to be disclosed for the present study.

Ethics approval This study was approved by the institutional review board of the Tohoku University Graduate School of Medicine (approval number: 2022-1-720).

References

Akaishi T (2022a) 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. https://doi.org/10.1620/tjem.2022.J049

Akaishi T (2022b) Insertion-and-deletion mutations between the genomes of SARS-CoV, SARS-CoV-2, and bat coronavirus RaTG13. Microbiol Spectr 10:e0071622. https://doi.org/10.1128/spectrum.00716-22

Akaishi T, Horii A, Ishii T (2022) Sequence exchange involving dozens of consecutive bases with external origin in SARS-related Coronaviruses. J Virol 96:e0100222. https://doi.org/10.1128/jvi.01002-22

Bashor L, Gagne RB, Bosco-Lauth AM, Bowen RA, Stenglein M, VandeWoude S (2021) SARS-CoV-2 evolution in animals suggests mechanisms for rapid variant selection. Proc Natl Acad Sci U S A. https://doi.org/10.1073/pnas.2105253118

Burki T (2022) The origin of SARS-CoV-2 variants of concern. Lancet Infect Dis 22:174–175. https://doi.org/10.1016/s1473-3099(22)00015-9

Cascella M, Rajnik M, Aleem A, Dulebohn SC, Di Napoli R (2022) Features, Evaluation, and Treatment of Coronavirus (COVID-19). StatPearls. StatPearls Publishing LLC, Treasure Island (FL)

Cele S et al (2022) SARS-CoV-2 prolonged infection during advanced HIV disease evolves extensive immune escape. Cell Host Microbe 30:154-162.e155. https://doi.org/10.1016/j.chom.2022.01.005

Du P, Gao GF, Wang Q (2022) The mysterious origins of the Omicron variant of SARS-CoV-2. Innovation (camb) 3:100206. https://doi.org/10.1016/j.xinn.2022.100206

Elbe S, Buckland-Merrett G (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall 1:33–46. https://doi.org/10.1002/gch2.1018

Fantini J, Yahi N, Azzaz F, Chahinian H (2021) Structural dynamics of SARS-CoV-2 variants: A health monitoring strategy for anticipating Covid-19 outbreaks. J Infect 83:197–206. https://doi.org/10.1016/j.jinf.2021.06.001

He X, Hong W, Pan X, Lu G, Wei X (2020) Wei X (2021) SARS-CoV-2 Omicron variant: characteristics and prevention. MedComm 2:838–845. https://doi.org/10.1002/mco2.110

Hoffman SA et al (2021) SARS-CoV-2 neutralization resistance mutations in patient with HIV/AIDS, California, USA. Emerg Infect Dis 27:2720–2723. https://doi.org/10.3201/eid2710.211461

Johns Hopkins University (2023) COVID-19 Dashboard.

Kandeel M, Mohamed MEM, Abd El-Lateef HM, Venugopala KN, El-Beltagi HS (2022) Omicron variant genome evolution and phylogenetics. J Med Virol 94:1627–1632. https://doi.org/10.1002/jmv.27515

Khare S et al (2021) GISAID's role in pandemic response. China CDC Wkly 3:1049–1051. https://doi.org/10.46234/ccdcw2021.255

Konishi T (2022) Mutations in SARS-CoV-2 are on the increase against the acquired immunity. PLoS ONE 17:e0271305. https://doi.org/10.1371/journal.pone.0271305

Lopez Bernal J et al (2021) Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. N Engl J Med 385:585–594. https://doi.org/10.1056/NEJMoa2108891

Mahdy MAA, Younis W, Ewaida Z (2020) An overview of SARS-CoV-2 and animal infection. Front Vet Sci 7:596391. https://doi.org/10.3389/fvets.2020.596391

Mallapaty S (2022) Where did Omicron come from? Three key theories. Nature 602:26–28. https://doi.org/10.1038/d41586-022-00215-2

Resende PC et al (2021) The ongoing evolution of variants of concern and interest of SARS-CoV-2 in Brazil revealed by convergent indels in the amino (N)-terminal domain of the spike protein. Virus Evol 7:veab069. https://doi.org/10.1093/ve/veab069

Shrestha LB, Foster C, Rawlinson W, Tedla N, Bull RA (2022) Evolution of the SARS-CoV-2 omicron variants BA.1 to BA.5: Implications for immune escape and transmission. Rev Med Virol 32:e2381. https://doi.org/10.1002/rmv.2381

Shu Y, McCauley J (2017) GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. https://doi.org/10.2807/1560-7917.Es.2017.22.13.30494

Sonnleitner ST et al (2022) Cumulative SARS-CoV-2 mutations and corresponding changes in immunity in an immunocompromised patient indicate viral evolution within the host. Nat Commun 13:2560. https://doi.org/10.1038/s41467-022-30163-4

Sun Y, Lin W, Dong W, Xu J (2022) Origin and evolutionary analysis of the SARS-CoV-2 Omicron variant. J Biosaf Biosecur 4:33–37. https://doi.org/10.1016/j.jobb.2021.12.001

Tamura K, Stecher G, Kumar S (2021) MEGA11: molecular evolutionary genetics analysis version 11. Mol Biol Evol 38:3022–3027. https://doi.org/10.1093/molbev/msab120

Tarcsai KR, Corolciuc O, Tordai A, Ongrádi J (2022) SARS-CoV-2 infection in HIV-infected patients: potential role in the high mutational load of the Omicron variant emerging in South Africa. Geroscience 44:2337–2345. https://doi.org/10.1007/s11357-022-00603-6

Wei C, Shan KJ, Wang W, Zhang S, Huan Q, Qian W (2021) Evidence for a mouse origin of the SARS-CoV-2 Omicron variant. J Genet Genomics 48:1111–1121. https://doi.org/10.1016/j.jgg.2021.12.003

Wu F et al (2020) A new coronavirus associated with human respiratory disease in China. Nature 579:265–269. https://doi.org/10.1038/s41586-020-2008-3