Tohoku J. Exp. Med., 2023, **259**, 17-26

Genetic Recombination Sites Away from the Insertion/Deletion Hotspots in SARS-Related Coronaviruses

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

[1]Department of Education and Support for Regional Medicine, Tohoku University, 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 genome sequences of severe acute respiratory syndrome (SARS)-related coronaviruses (sarbecoviruses) have been reported to include many long and complex insertions/deletions (indels) in specific genomic regions, including open reading frame 1a (*ORF1a*), S1 domain of the spike, and *ORF8* genes. These indel hotspots incorporate various non-classical, long, and complex indels with uncertain developmental processes. A possible explanation for these complex and diversified indels at the hotspots is genetic recombination. To determine the possible association between recombination events and development of indel hotspots, this study investigated the genome sequences of many sarbecoviruses from different countries and hosts and compared the distributions of the indel hotspots and recombination sites by performing multiple sequence alignments and recombination analyses. The genomes of 19 SARS-related coronaviruses (15 coronaviruses that infect bats, two that infect humans, one that infects pangolins, and one that infects civets), including human-infecting SARS-CoV and SARS-CoV-2, were evaluated. Hotspots of complex indels with diverse RNA sequences around gaps were clustered in non-structural protein 2 (Nsp2) and Nsp3 of *ORF1a*, *S1*, and *ORF8*. Phylogenetic reconstructions revealed different structures of the inferred phylogenetic trees between genomic regions, and recombination analyses identified multiple recombination sites across *ORF1ab* and *S* genes. However, the nucleotide positions of the indel hotspots were not identical with the identified recombination sites in the recombinant viruses, suggesting the involvement of different developmental processes of indel hotspots and genetic recombination. Further research is required to elucidate the developmental mechanisms underpinning clustered complex indels and recombination events in the evolutionary history of sarbecoviruses.

Keywords: genetic recombination; multiple sequence alignment; phylogenetic reconstruction; SARS-CoV-2; SARS-related coronaviruses
Tohoku J. Exp. Med., 2023 January, **259** (1), 17-26.
doi: 10.1620/tjem.2022.J093

Introduction

The coronavirus disease 2019 (COVID-19) pandemic, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), remains one of the largest public health concerns in the world in 2022 (Biancolella et al. 2022; Callaway 2022; Murray 2022). The intermittent emergence of the concerned variants facilitated the long-term continuation of the pandemic, possibly through mechanisms including immune evasion by the virus (Lazarevic et al. 2021; Beyer and Forero 2022; Sun et al. 2022). Most variants of concern during the COVID-19 pandemic in humans are considered to be derived from point mutations in the S1 domain of the spike (*S*) gene (Perez-Gomez 2021; Cosar et al. 2022). Additionally, the genomes of severe acute respiratory syndrome (SARS)-related coronaviruses, belonging to the subgenus Sarbecovirus, have been reported to incorporate a large number of non-classical long and complex insertions/deletions (indels) with heavily changed RNA sequences around the gaps of aligned sequences in specific genomic regions, such as non-structural protein 2 (Nsp2), Nsp3, S1 domain of the *S* gene, and open reading frame 8 (*ORF8*) (Akaishi 2022a, b). Such a novel and complex indel cannot be explained by conventional classical muta-

Received September 25, 2022; revised and accepted October 26, 2022; J-STAGE Advance online publication November 10, 2022
Correspondence: Tetsuya Akaishi, Department of Education and Support for Regional Medicine, Tohoku University, 1-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi 980-8574, Japan.
e-mail: t-akaishi@med.tohoku.ac.jp

tion types, such as point mutations and sole insertions and deletions (Akaishi et al. 2022). Therefore, the developmental process of these non-classical complex indels in the genomes of SARS-related coronaviruses remains unknown. Moreover, the species jump of SARS-CoV-2 herald to the pandemic has been suggested to be caused by a 12-nucleotide in-frame insertion at the furin cleavage site between the S1 and S2 domains of the *S* gene (Andersen et al. 2020; Johnson et al. 2021). Elucidating the developmental process and mechanisms of the long and complex indels in viruses will offer deeper insights into the evolutionary history of SARS-CoV-2 and the role of clustered complex indels in the evolution of SARS-related coronaviruses in different host animals. Currently, one of the most promising theories explaining the development of such non-classical complex indels is the replicative or non-replicative genetic recombination of the unsegmented viral genome from two or more coinfected SARS-related coronaviruses, such as copy-choice recombination (Simon-Loriere and Holmes 2011; Pérez-Losada et al. 2015; Muslin et al. 2019). To estimate the possible association between genetic recombination and development of indel hotspots in sarbecovirus genomes, we performed multiple sequence alignments, phylogenetic reconstructions, and recombination analyses with many sarbecoviruses from different countries and host animals. The topological characteristics of the inferred phylogenetic trees were compared between the different genomic regions, in addition to comparing the distributions of the identified recombination sites with indel hotspots.

Materials and Methods

Enrolled viruses and genomes

The present study enrolled a total of 19 SARS-related coronaviruses and performed multiple sequence alignment and phylogenetic studies of the genomes. SARS-related coronaviruses (sampled country, year; GenBank accession ID) are summarized in Table 1. (Chim et al. 2003; Li et al. 2005; Wang et al. 2005; Drexler et al. 2010; Lau et al. 2010; Wu et al. 2016; Hu et al. 2017; Lin et al. 2017; Wang et al. 2017; Hu et al. 2018; Han et al. 2019; Tao and Tong 2019; Lam et al. 2020; Murakami et al. 2020; Wu et al. 2020; Zhou et al. 2020). The geographic distributions of the 15 SARS-related coronaviruses from China are shown on a map of mainland China (Fig. 1) using Map Chart Software (https://www.mapchart.net).

Multiple sequence alignment of coronavirus genomes

Multiple sequence alignment for the specific coding regions of the 19 SARS-related coronaviruses was performed using Molecular Evolutionary Genetics Analysis Version 11 (MEGA11) software (Tamura et al. 2021). The MUSCLE program was run to align the sequences using the following set of parameters: gap opening penalty score of –400 and gap extension penalty score of 0. Adjacent sequences before and after each gap upon sequence alignment were reviewed to determine whether the observed gaps were derived from classical short in-frame indels or other non-classical complex indels.

Table 1. List of the evaluated 19 SARS-related coronavirus species.

GenBank ID	Author (publication year)	Country	Sampling year	Host	Virus species
MN908947	Wu et al. (2020)	China	2019	Human	SARS-CoV-2 Wuhan-Hu-1
MN996532	Zhou et al. (2020)	China	2013	Bat	RaTG13
MG772933	Hu et al. (2018)	China	2015-2017	Bat	bat-SL-CoVZC45
LC556375	Murakami et al. (2020)	Japan	2013	Bat	Rc-o319
MT040333	Lam et al. (2020)	China	2017-2018	Pangolin*	PCoV_GX-P4L
AY345986	Chim et al. (2003)	China	2003	Human	SARS-CoV CUHK-AG01
AY572034	Wang et al. (2005)	China	2004	Civet	Civet CoV civet007
DQ412043	Li et al. (2005)	China	2004	Bat	Bat CoV Rm1
KY938558	Son (2021)†	Korea	2017	Bat	Bat CoV 16BO133
MK211374	Han et al. (2019)	China	2016-2017	Bat	BtRI-BetaCoV/SC2018
KY417142	Hu et al. (2017)	China	2011-2015	Bat	Bat CoV As6526
KU973692	Wang et al. (2017)	China	2012	Bat	Bat CoV F46
KJ473816	Wu et al. (2016)	China	2013	Bat	BtRs-BetaCoV/YN2013
KY417144	Hu et al. (2017)	China	2011-2015	Bat	Bat CoV Rs4084
GQ153548	Lau et al. (2010)	China	2004-2008	Bat	Bat CoV HKU3-13
DQ071615	Li et al. (2005)	China	2004	Bat	Bat CoV Rp3
KF294457	Lin et al. (2017)	China	2012-2015	Bat	Bat CoV Longquan-140
KY352407	Tao and Tong (2019)	Kenya	2006-2010	Bat	Bat CoV BtKY72
NC_014470	Drexler et al. (2010)	Bulgaria	2008	Bat	BM48-31/BGR/2008

*Malayan pangolin seized during anti-smuggling operation.
†Unpublished.

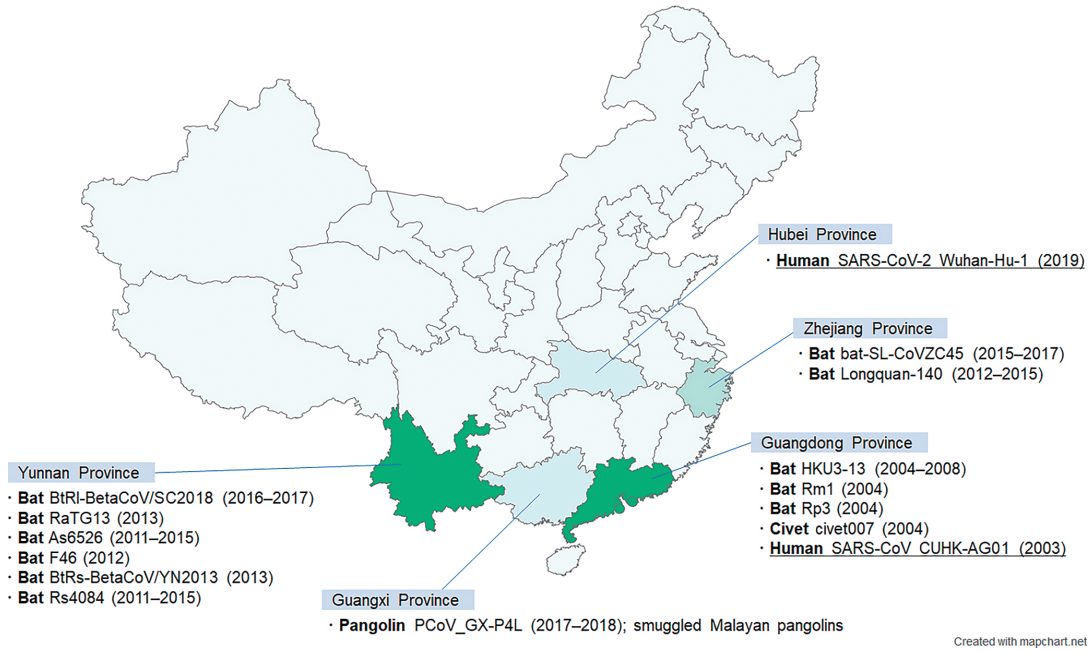

Fig. 1. Geographic distributions of the evaluated SARS-related coronaviruses from China.
A total of 19 SARS-related coronaviruses were evaluated in the present study, 15 of which have been sequenced in human or animal samples from China in the last two decades. Of the 15 virus species from China, 11 were from the species that infect bats, two from those that infect humans (SARS-CoV and SARS-CoV-2), one that infects civets, and one that infects pangolins. The pangolin coronavirus GX-P4L from Guangxi Province was sampled from pangolins that were supposedly smuggled from Southeast Asia. The other four evaluated virus species from countries outside China were Bulgaria (bat BM48-31/BGR/2008), Korea (bat 16BO133), Kenya (bat BtKY72), and Japan (bat Rc-o319). The figure was constructed using Map Chart software (available at https://www.mapchart.net/).

Inference of phylogenetic maximum likelihood (ML) trees in specific genomic regions

To estimate ancestral states among the five SARS-related coronaviruses, gene region-specific phylogenetic ML trees were inferred for different subdomains in *ORF1ab* (Nsp1-2, indel hotspot domain and subsequent regions of Nsp3, and Nsp4-16) and spike genes (*N*-terminal domain, receptor-binding domain, and S2) using the Tamura-Nei model (Tamura and Nei 1993). Phylogenetic reconstruction was performed for the membrane (*M*), *ORF8*, and nucleocapsid (*N*) genes. Phylogenetic trees were constructed using the MEGA11 software with 100 bootstraps resamplings. The nearest-neighbor interchange method was used for the ML heuristic search process. Branch lengths were the genetic distances measured by the number of substitutions per site.

Recombination analysis

To detect potential recombination sites in the *ORF1ab* and *S* genes, recombination analyses were performed using the Recombination Detection Program Version 5 (RDP5) (Martin et al. 2021). Each of the detected potential recombination events was conservatively accepted when the recombination signal was detected by at least five of the following six methods implemented in RDP5: RDP, GENECONV, MaxChi, Bootscan, SiScan, and 3Seq, according to previous studies (Delaune et al. 2021). The recombination signals with strong *P* values, as suggested by $P < 1.0E\text{-}20$, were considered possible recombination events. MaxChi breakpoint matrices were depicted for these recombinants to determine the optimal locations of the breakpoints.

Ethics approval

This study did not use original data from human or animal subjects and approval from the institutional review board was not applicable to the present study.

Results

Multiple alignment and phylogenetic reconstructions in Nsp3

Multiple sequence alignment using MEGA11 was performed for the *ORF1ab* gene in the whole genome of the 19 SARS-related coronaviruses. At least three indel hotspots in *ORF1ab*; one in Nsp2 (2,500-2,600 base positions) and two in Nsp3 (3,000-3,300 and 3,800-3,900 base positions), were found. The actual aligned sequences before the first indel hotspot in Nsp3 (a), those at the first indel hotspot in Nsp3 (b), and those after the hotspot (c) are shown in Fig. 2. In the sequences outside the indel hotspots, indels were only sparsely observed, most of which were comprised of short 3-base in-frame indels, and sequences around the gaps were largely conserved. Meanwhile, in the indel hotspots, different types of indels clustered in similar regions, and

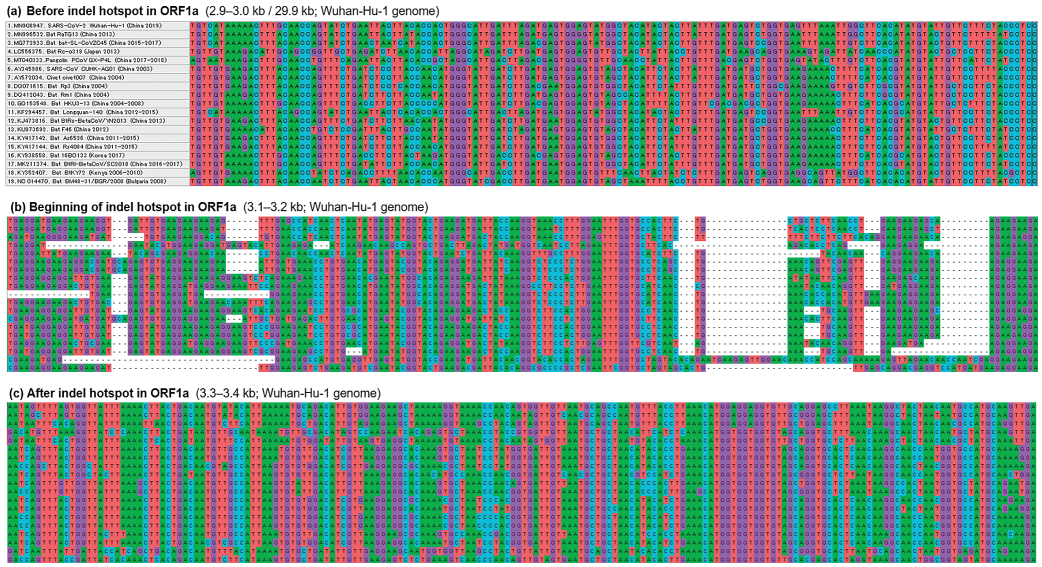

Fig. 2. Multiple sequence alignment showing insertion/deletion hotspots in *ORF1a*.
The results of multiple sequence alignment with MEGA11 software for the following three different domains of Nsp3 in the *ORF1a* gene are shown: (a) before the first indel hotspot in Nsp3, (b) the beginning of the first indel hotspot in Nsp3 located at 3,000-3,300 base positions, and (c) after the first indel hotspot in Nsp3. Relatively long and structurally varied indels were frequently observed in the indel hotspots, and the sequences around the gaps upon sequence alignment were heavily mutated in some indel spots, suggesting non-classical and complex developmental processes of these indels in the hotspots.
Nsp3, non-structural protein 3; ORF1a, open reading frame 1a; SARS-CoV, severe acute respiratory syndrome coronavirus.

many of the indels involved relatively long segments with 10 or more consecutive nucleotides. Furthermore, the sequences around the gaps upon sequence alignment were not conserved between the viruses, suggesting a complex developmental process involving these non-classical indels. The inferred phylogenetic ML trees for the four subdomains of *ORF1ab* are shown in Fig. 3. In total, Wuhan-Hu-1, RaTG13, bat-SL-CoVZC45, Rc-o319, and PCoV_GX-P4L belonged to the COVID clade (colored in yellow), while CUHK-AG01, civet007, Rm1, 16BO133, BtRI-BetaCoV/SC2018, As6526, F46, BtRs-BetaCoV/YN2013, Rs4084, HKU3-13, Rp3, and Longquan-140 belonged to the SARS clade (colored in green), and the other two from Kenya and Bulgaria belonged to another clade (colored in red). In Nsp1-2, the structure of the inferred phylogenetic tree was reliable with decent bootstrapping values in each node; however, Longquan-140, which is the closest species to HKU3-13 in other genetic regions, was aligned in the COVID clade. In Nsp4-16, bat-SL-CoVZC45, which is closest to Wuhan-Hu-1 and RaTG13 in other genetic regions, was aligned in the SARS clade.

Multiple alignment and phylogenetic reconstructions in S gene

Multiple alignments were performed for the three subdomains of the *S* gene (Fig. 4). An indel hotspot was identified in the *N*-terminal domain of the *S* gene (21,500-22,500 bp). The actual sequences in the indel hotspot in the *S* gene (Fig. 4a) showed diverse patterns of complex indels between viruses, similar to the indel hotspot in Nsp3, and the sequences around the gaps were not largely conserved. In contrast, indels outside the indel hotspot were sparse, and most comprised 3-base in-frame indels with conserved sequences around the indels (Fig. 4b, c). The inferred phylogenetic ML trees for the three subdomains of the *S* gene are shown in Fig. 5a-c. The phylogenetic reconstruction was unstable in the *N*-terminal domain with relatively low bootstrapping values, whereas the reconstructions were more stable in the receptor-binding domain and *S2* gene. The phylogenetic tree for the *S2* gene clearly separated these three clades. In the receptor-binding domain, bat-SL-CoVZC45 belonged to the SARS clade and aligned closest to 16BO133.

Phylogenetic reconstructions in M, ORF8, and N genes

Phylogenetic reconstructions were performed for the *M*, *ORF8*, and *N* genes (Fig. 5d-f). In the *M* and *N* genes, the three clades of different colors were clearly separated. In the *ORF8* gene, another indel hotspot, the inferred phylogenetic trees were unstable with lower bootstrapping values. 16BO133 was aligned in the COVID clade and Rc-o319 was aligned in the SARS clade. Although the inferred trees for *ORF8* had low bootstrapping values, the *ORF8* gene of Rc-o319 suggestively had a completely different developmental history from that of the other four viruses belonging to the COVID clade.

Recombination analyses in ORF1ab and S genes

The potential recombination sites in the *ORF1ab* and *S* genes were determined using RDP5 with *P* values less than

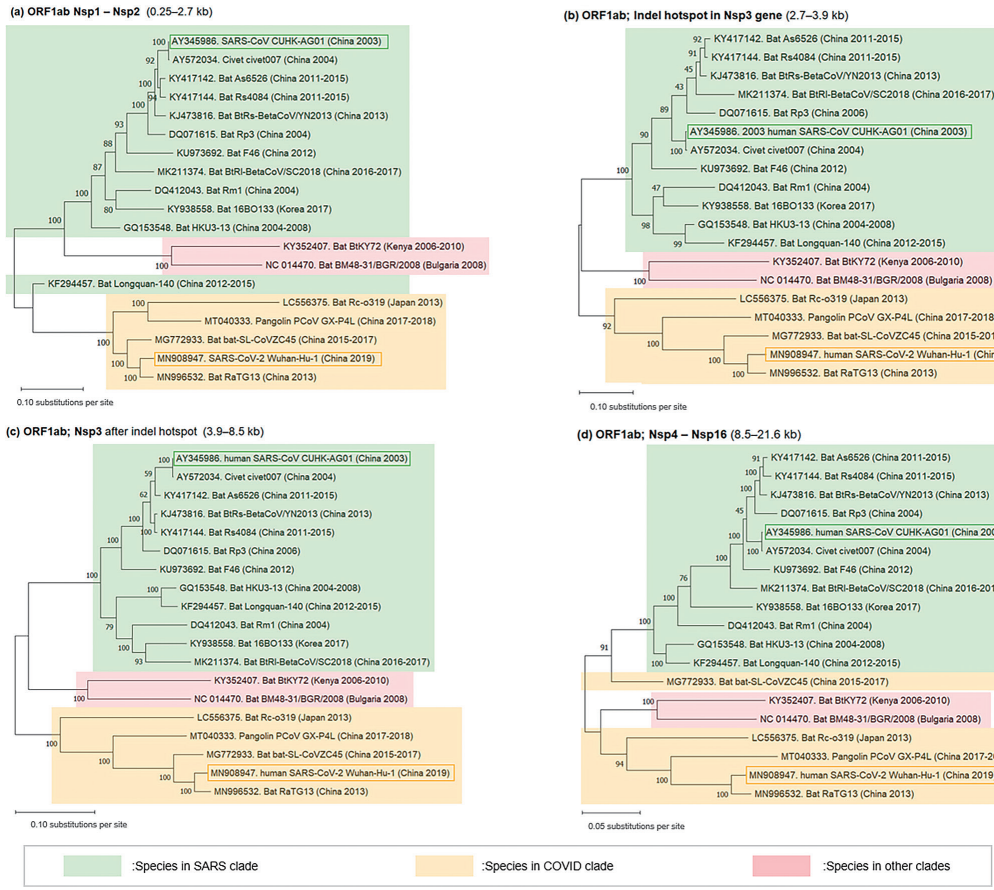

Fig. 3. Phylogenetic maximum likelihood trees in specific regions of *ORF1ab*.
The results of phylogenetic reconstructions of the four subdomains of *ORF1ab* are shown. The five virus species belonging to the COVID clade are colored in yellow, 12 viruses belonging to the SARS clade are colored in green, and the other clade with the remaining two viruses is colored in red. In the first subdomain with Nsp1 and Nsp2, the bat SARS-like coronavirus Longquan-140 was aligned separately from all three clades, suggesting a different developmental process of this subdomain in Longquan-140 from the other 11 viruses belonging to the SARS clade, such as genetic recombination somewhere in the subdomain.
COVID, coronavirus disease 2019; Nsp, non-structural protein; ORF1ab, open reading frame 1ab; SARS, severe acute respiratory syndrome coronavirus.

1.0E-20 in the present study. The suggested recombinants (recombination breakpoints) in *ORF1ab* were bat-SL-CoVZC45 (11,600 and 20,300 base positions, P = 6.66E-275), bat Longquan-140 (1,400 and 2,770 base positions, P = 2.08E-188), and two different recombination events in bat BtRl-BetaCoV/SC2018 (2,690 and 3,500 base positions, P = 1.79E-23; 8,990 and 18,820 base positions, P = 3.97E-67). Those in the *S* gene were as follows: bat F46 (22,580 base position, P = 5.61E-51) and bat As6526 (23,310 base position, P = 3.24E-42). Pairwise identity line graphs for the first two recombinants in the *ORF1ab* gene are shown in Fig. 6, together with MaxChi breakpoint matrices. The identified recombination sites did not necessarily fall within the range of the identified indel hotspots in the *ORF1ab* gene (at 2,500-2,600, 3,000-3,300, and 3,800-3,900 base positions). Pairwise identity graphs and MaxChi breakpoint matrices of the *S* gene are shown in Fig. 7. None of the recombination sites fell within the range of the identified indel hotspot in the *S* gene (21,500-22,500 base positions).

Discussion

Sequence alignment in the present study demonstrated the presence of indel hotspots in specific regions of the genomes of SARS-related coronaviruses, such as Nsp2, Nsp3, *S1*, and *ORF8*. Furthermore, diverse patterns of complex indels with heavily mutated sequences around the gaps upon sequence alignment were clustered in these indel hotspots. The structures of the reconstructed phylogenetic ML trees and viral components of each clade were dynamically mutated by the genetic regions. Indel hotspots may reportedly correspond to the location of RNA-dependent RNA polymerase (RdRp) template-switching hotspots with genetic recombination, causing so-called copy-choice genetic recombination (Chrisman et al. 2021). This copy-choice genetic recombination theory could explain the developmental process of the observed complex indels in the genomes of SARS-related coronaviruses, which are dif-

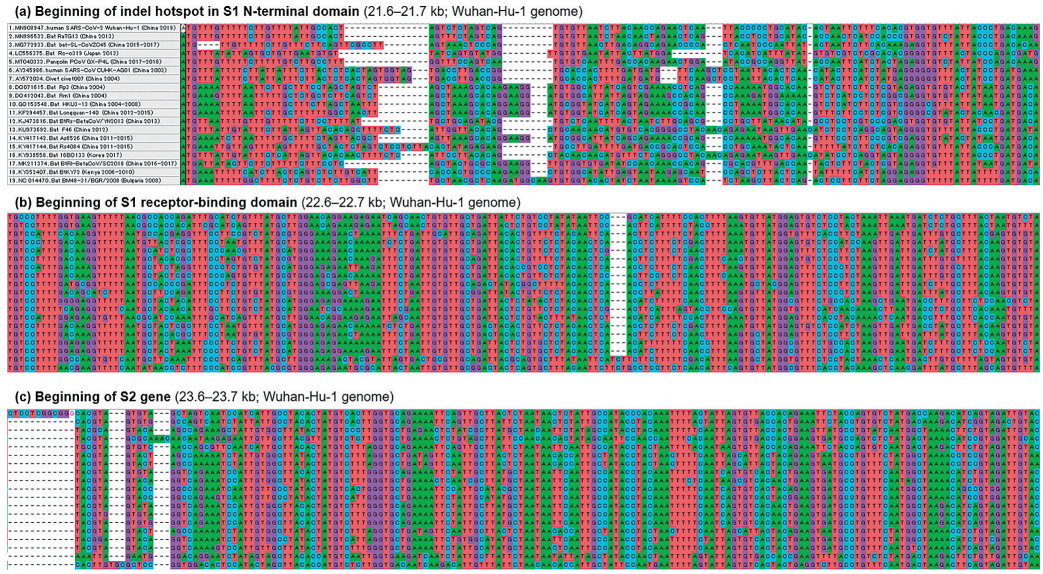

Fig. 4. Multiple sequence alignment in the spike gene.
The results of the multiple sequence alignment with MEGA11 software for the following three different domains of the S gene are shown: the beginning of the *N*-terminal domain of the *S* gene (a), the beginning of the receptor-binding domain (b), and the beginning of the *S2* gene (c). Similar to the indels in *ORF1a*, indels in the indel hotspot of the *N*-terminal domain comprised relatively long and structurally varied indels. These non-classical indels were accompanied by heavily mutated sequences in the adjacent sequences around the gaps, suggesting non-classical complex processes in the development of these indels. In other domains outside the indel hotspots, indels were sparse, and most were 3-base in-frame indels with conserved sequences around the gaps. For reference, the gap at the head of the *S2* gene is the furin cleavage site with a 12-base in-frame insertion, which is one of the essential mutations that could have caused the emergence of SARS-CoV-2 in 2019.

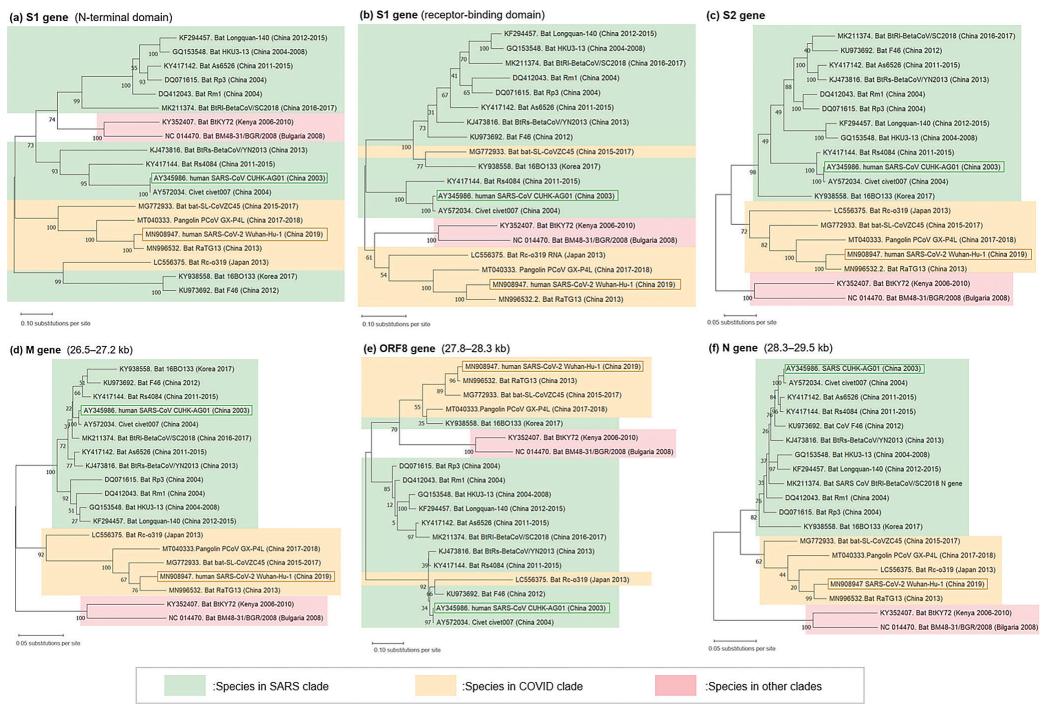

Fig. 5. Phylogenetic maximum likelihood trees in *S*, *M*, *ORF8*, and *N* genes.
The results of the phylogenetic reconstructions for the three subdomains of the *S* (a-c), *M* (d), *ORF8* (e), and *N* genes (f) are shown. Similar to the phylogenetic trees in *ORF1ab*, the structures of the inferred phylogenetic trees differed by the genetic regions. For example, bat-SL-CoVZC45 was aligned in the SARS clade with receptor-binding domain sequences, and 16BO133 was aligned with *ORF8* gene sequences in the COVID clade.
COVID, coronavirus disease 2019; Nsp, non-structural protein; ORF1ab, open reading frame 1ab; SARS, severe acute respiratory syndrome coronavirus.

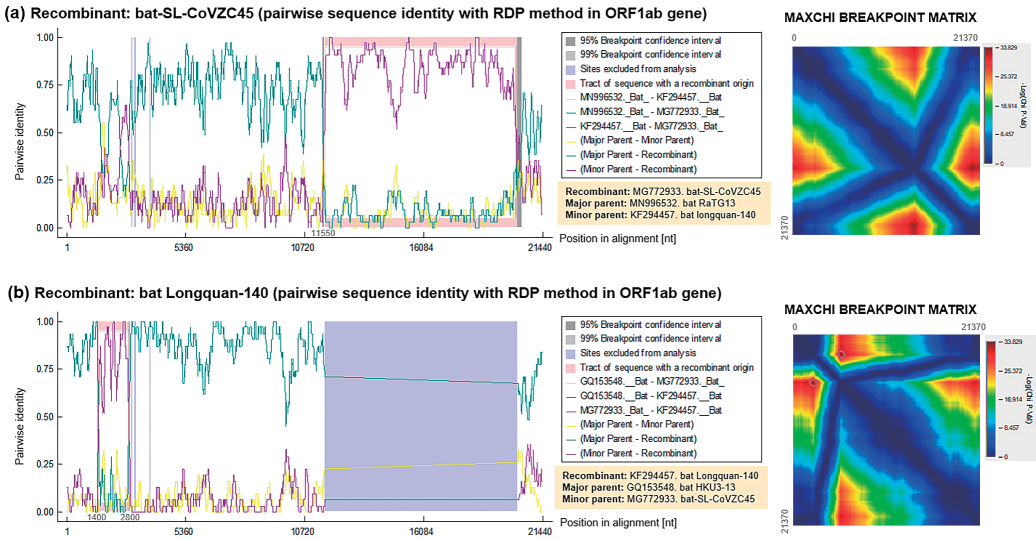

Fig. 6. Identification of recombination events in *ORF1ab* gene of sarbecoviruses using Recombination Detection Program Version 5 (RDP5).
Two recombination events in the *ORF1ab* gene, as identified in bat-SL-CoVZC45 (a) and Longquan-140 (b), fulfilled the statistical significance of P <1.0E-20 for accepting potential recombination events. MaxChi breakpoint matrices were built to determine the optimal locations of the recombination sites. Some other potential recombination sites were found; however, these recombination sites did not necessarily fall within the ranges of the identified indel hotspots, suggesting different mechanisms between recombination events and the development of indel hotspots.
ORF, open reading frame.

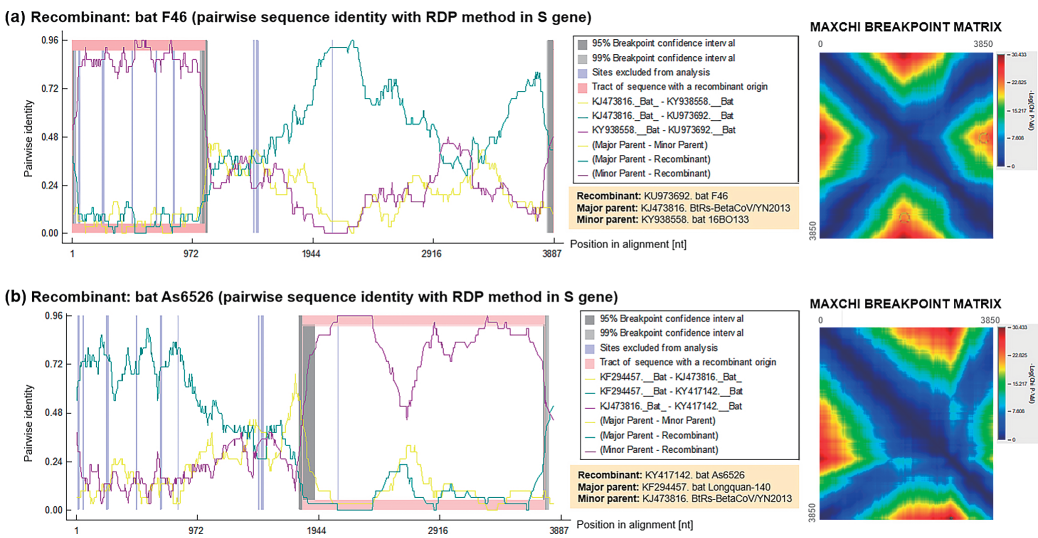

Fig.7. Identification of recombination events in *S* gene of sarbecoviruses using Recombination Detection Program Version 5 (RDP5).
Both recombination events in the spike gene (*S*), as identified in bat F46 (a) and As6526 (b), fulfilled the statistical significance of P < 1.0E-20 for accepting the potential recombination events. MaxChi breakpoint matrices were built to determine the optimal locations of the recombination sites. These recombination sites did not fall within the indel hotspots in the *S* gene, similar to *ORF1ab*, suggesting different mechanisms between recombination events and the development of indel hotspots.

ficult to be explained with conventional ordinary mutations such as point mutations or short in-frame indels. Moreover, the theory could explain the different structures of inferred phylogenetic ML trees between genomic regions of SARS-related coronaviruses, which are single-stranded RNA viruses with a single segment comprising approximately 29,900 nucleotides. Unlike viruses with multiple genome segments, such as influenza viruses which often undergo genetic reassortment, SARS-related coronaviruses are unsegmented viruses. Therefore, the different structures of inferred phylogenetic trees between genomic regions of the sarbecoviruses suggest replicative or non-replicative genetic recombination within the unsegmented genomes. The possible occurrence of recombination events in the evolution-

ary process of sarbecoviruses has been previously demonstrated (Lin et al. 2017; Boni et al. 2020; Li et al. 2020). Another notable finding of the present study was that the identified recombination sites in *ORF1ab* and *S* were independent of the indel hotspots in these genes. This finding implies that the developmental process of the indel hotspots, often with highly polymorphic sequence patterns, in sarbecoviruses is different from recombination events. Although clues for determining the exact developmental processes of the indel hotspots and genetic recombination in sarbecoviruses remain scarce, both types of genetic mutation events have evidently and continuously contributed to the survival and evolution of the viruses.

For recombination to occur between the viruses, coexistence of the genomes from different SARS-related coronavirus species in the same host cell would be required, suggesting the occurrence of coinfection with two or more different SARS-related coronaviruses in one host animal. A large variety of SARS-related coronaviruses from many different host animals has been previously demonstrated to be broadly distributed across southern China and Southeast Asia, making this possible problematic mechanism difficult to deal with (Delaune et al. 2021; Zhou et al. 2021). To counteract the emergence of newer and more dangerous SARS-CoV-2 variant strains and other sarbecoviruses from uncertain mutation mechanisms, further studies elucidating the developmental process of highly polymorphic indels at indel hotspots, and genetic recombination is needed. Continued field surveillance for sampling from animals is important to deal with the emergence of novel recombinant SARS-related coronaviruses in the future, achieve an overview of the geographic distributions, and identify the overlapping areas and natural environments of different SARS-related coronaviruses. Attempts to suppress the risks and occasions of the coinfection of different viruses in the natural environments would be one of the promising strategies to counteract the emergence of novel SARS-related coronavirus species in the future.

The present study has several limitations. First, the structures of the inferred phylogenetic ML trees in some genomic regions were unstable, with relatively low reproducibility in 100 bootstrap resamplings, especially for the *S1* and peri-*ORF8* genes, both of which are indel hotspots with clustered long and complex indels. In these regions, virus species belonging to the SARS-clade and COVID-19-clade could not be clearly distinguished. Extra caution is required when interpreting the inferred phylogenetic trees in these regions, which may not be suitable for multiple sequence alignment or phylogenetic reconstructions. Second, this study could not determine the molecular mechanisms of genetic recombination, such as whether the recombination events were based on replicative or non-replicative mechanisms. The former replicative mechanisms include the copy-choice recombination based on RdRp template switching with homologous templates from co-infected different viral species (Chrisman et al. 2021; Francisco Junior et al. 2022); however, most of the mechanisms and processes underpinning copy-choice recombination remain unclear. Future studies are needed to elucidate the possible role of genetic recombination in the evolutionary history of SARS-related coronaviruses.

In summary, the present study dealt with wide variety of SARS-related coronaviruses across the countries worldwide from different host animals and revealed the presence of indel hotspots with clustered non-classical complex indels in the Nsp2, Nsp3, *S1*, and *ORF8* genes. Phylogenetic reconstructions in different genomic regions suggested different structures of inferred phylogenetic trees between genomic regions, suggesting the occurrence of recombination events. However, the distributions of recombination sites and indel hotspots were not identical in most recombinant viruses. Further studies are needed to delineate the roles and mechanisms of genetic recombination and clustered complex indels in the evolutionary history of SARS-related coronaviruses.

Author Contributions

T.A. performed the analyses and drafted the manuscript. K.F. re-performed and verified the analyses, 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) 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. (2022b) Insertion-and-deletion mutations between the genomes of SARS-CoV, SARS-CoV-2, and bat coronavirus RaTG13. *Microbiol. Spectr.*, **10**, e0071622.

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.

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C. & Garry, R.F. (2020) The proximal origin of SARS-CoV-2. *Nat. Med.*, **26**, 450-452.

Beyer, D.K. & Forero, A. (2022) Mechanisms of antiviral immune evasion of SARS-CoV-2. *J. Mol. Biol.*, **434**, 167265.

Biancolella, M., Colona, V.L., Mehrian-Shai, R., Watt, J.L., Luzzatto, L., Novelli, G. & Reichardt, J.K.V. (2022) COVID-19 2022 update: transition of the pandemic to the endemic phase. *Hum. Genomics*, **16**, 19.

Boni, M.F., Lemey, P., Jiang, X., Lam, T.T., Perry, B.W., Castoe, T.A., Rambaut, A. & Robertson, D.L. (2020) Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.*, **5**, 1408-1417.

Callaway, E. (2022) Are COVID surges becoming more predictable? New Omicron variants offer a hint. *Nature*, **605**, 204-206.

Chim, S.S., Tsui, S.K., Chan, K.C., Au, T.C., Hung, E.C., Tong, Y.K., Chiu, R.W., Ng, E.K., Chan, P.K., Chu, C.M., Sung, J.J., Tam, J.S., Fung, K.P., Waye, M.M., Lee, C.Y., et al. (2003) Genomic characterisation of the severe acute respiratory syndrome coronavirus of Amoy Gardens outbreak in Hong

Kong. *Lancet*, **362**, 1807-1808.
Chrisman, B.S., Paskov, K., Stockham, N., Tabatabaei, K., Jung, J.Y., Washington, P., Varma, M., Sun, M.W., Maleki, S. & Wall, D.P. (2021) Indels in SARS-CoV-2 occur at template-switching hotspots. *BioData Min.*, **14**, 20.
Cosar, B., Karagulleoglu, Z.Y., Unal, S., Ince, A.T., Uncuoglu, D.B., Tuncer, G., Kilinc, B.R., Ozkan, Y.E., Ozkoc, H.C., Demir, I.N., Eker, A., Karagoz, F., Simsek, S.Y., Yasar, B., Pala, M., et al. (2022) SARS-CoV-2 mutations and their viral variants. *Cytokine Growth Factor Rev.*, **63**, 10-22.
Delaune, D., Hul, V., Karlsson, E.A., Hassanin, A., Ou, T.P., Baidaliuk, A., Gámbaro, F., Prot, M., Tu, V.T., Chea, S., Keatts, L., Mazet, J., Johnson, C.K., Buchy, P., Dussart, P., et al. (2021) A novel SARS-CoV-2 related coronavirus in bats from Cambodia. *Nat. Commun.*, **12**, 6563.
Drexler, J.F., Gloza-Rausch, F., Glende, J., Corman, V.M., Muth, D., Goettsche, M., Seebens, A., Niedrig, M., Pfefferle, S., Yordanov, S., Zhelyazkov, L., Hermanns, U., Vallo, P., Lukashev, A., Müller, M.A., et al. (2010) Genomic characterization of severe acute respiratory syndrome-related coronavirus in European bats and classification of coronaviruses based on partial RNA-dependent RNA polymerase gene sequences. *J. Virol.*, **84**, 11336-11349.
Francisco Junior, R.D.S., de Almeida, L.G.P., Lamarca, A.P., Cavalcante, L., Martins, Y., Gerber, A.L., Guimarães, A.P.C., Salviano, R.B., Dos Santos, F.L., de Oliveira, T.H., de Souza, I.V., de Carvalho, E.M., Ribeiro, M.S., Carvalho, S., da Silva, F.D., et al. (2022) Emergence of within-host SARS-CoV-2 recombinant genome after coinfection by Gamma and Delta variants: a case report. *Front. Public Health*, **10**, 849978.
Han, Y., Du, J., Su, H., Zhang, J., Zhu, G., Zhang, S., Wu, Z. & Jin, Q. (2019) Identification of diverse bat alphacoronaviruses and betacoronaviruses in China provides new insights into the evolution and origin of coronavirus-related diseases. *Front. Microbiol.*, **10**, 1900.
Hu, B., Zeng, L.P., Yang, X.L., Ge, X.Y., Zhang, W., Li, B., Xie, J.Z., Shen, X.R., Zhang, Y.Z., Wang, N., Luo, D.S., Zheng, X.S., Wang, M.N., Daszak, P., Wang, L.F., et al. (2017) Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog.*, **13**, e1006698.
Hu, D., Zhu, C., Ai, L., He, T., Wang, Y., Ye, F., Yang, L., Ding, C., Zhu, X., Lv, R., Zhu, J., Hassan, B., Feng, Y., Tan, W. & Wang, C. (2018) Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. *Emerg. Microbes Infect.*, **7**, 154.
Johnson, B.A., Xie, X., Bailey, A.L., Kalveram, B., Lokugamage, K.G., Muruato, A., Zou, J., Zhang, X., Juelich, T., Smith, J.K., Zhang, L., Bopp, N., Schindewolf, C., Vu, M., Vanderheiden, A., et al. (2021) Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. *Nature*, **591**, 293-299.
Lam, T.T., Jia, N., Zhang, Y.W., Shum, M.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., et al. (2020) Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature*, **583**, 282-285.
Lau, S.K., Li, K.S., Huang, Y., Shek, C.T., Tse, H., Wang, M., Choi, G.K., Xu, H., Lam, C.S., Guo, R., Chan, K.H., Zheng, B.J., Woo, P.C. & Yuen, K.Y. (2010) Ecoepidemiology and complete genome comparison of different strains of severe acute respiratory syndrome-related Rhinolophus bat coronavirus in China reveal bats as a reservoir for acute, self-limiting infection that allows recombination events. *J. Virol.*, **84**, 2808-2819.
Lazarevic, I., Pravica, V., Miljanovic, D. & Cupic, M. (2021) Immune evasion of SARS-CoV-2 emerging variants: what have we learnt so far? *Viruses*, **13**, 1192.
Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J.H., Wang, H., Crameri, G., Hu, Z., Zhang, H., Zhang, J., McEachern, J., Field, H., Daszak, P., Eaton, B.T., et al. (2005) Bats are natural reservoirs of SARS-like coronaviruses. *Science*, **310**, 676-679.
Li, X., Giorgi, E.E., Marichannegowda, M.H., Foley, B., Xiao, C., Kong, X.P., Chen, Y., Gnanakaran, S., Korber, B. & Gao, F. (2020) Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci. Adv.*, **6**, eabb9153.
Lin, X.D., Wang, W., Hao, Z.Y., Wang, Z.X., Guo, W.P., Guan, X.Q., Wang, M.R., Wang, H.W., Zhou, R.H., Li, M.H., Tang, G.P., Wu, J., Holmes, E.C. & Zhang, Y.Z. (2017) Extensive diversity of coronaviruses in bats from China. *Virology*, **507**, 1-10.
Martin, D.P., Varsani, A., Roumagnac, P., Botha, G., Maslamoney, S., Schwab, T., Kelz, Z., Kumar, V. & Murrell, B. (2021) RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol.*, **7**, veaa087.
Murakami, S., Kitamura, T., Suzuki, J., Sato, R., Aoi, T., Fujii, M., Matsugo, H., Kamiki, H., Ishida, H., Takenaka-Uema, A., Shimojima, M. & Horimoto, T. (2020) Detection and characterization of bat sarbecovirus phylogenetically related to SARS-CoV-2, Japan. *Emerg. Infect. Dis.*, **26**, 3025-3029.
Murray, C.J.L. (2022) COVID-19 will continue but the end of the pandemic is near. *Lancet*, **399**, 417-419.
Muslin, C., Mac Kain, A., Bessaud, M., Blondel, B. & Delpeyroux, F. (2019) Recombination in enteroviruses, a multi-step modular evolutionary process. *Viruses*, **11**, 859.
Perez-Gomez, R. (2021) The development of SARS-CoV-2 variants: the gene makes the disease. *J. Dev. Biol.*, **9**, 58.
Pérez-Losada, M., Arenas, M., Galán, J.C., Palero, F. & González-Candelas, F. (2015) Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. *Infect. Genet. Evol.*, **30**, 296-307.
Simon-Loriere, E. & Holmes, E.C. (2011) Why do RNA viruses recombine? *Nat. Rev. Microbiol.*, **9**, 617-626.
Sun, C., Xie, C., Bu, G.L., Zhong, L.Y. & Zeng, M.S. (2022) Molecular characteristics, immune evasion, and impact of SARS-CoV-2 variants. *Signal Transduct. Target. Ther.*, **7**, 202.
Tamura, K. & Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, **10**, 512-526.
Tamura, K., Stecher, G. & Kumar, S. (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.*, **38**, 3022-3027.
Tao, Y. & Tong, S. (2019) Complete genome sequence of a severe acute respiratory syndrome-related coronavirus from Kenyan bats. *Microbiol. Resour. Announc.*, **8**, e00548-19.
Wang, L., Fu, S., Cao, Y., Zhang, H., Feng, Y., Yang, W., Nie, K., Ma, X. & Liang, G. (2017) Discovery and genetic analysis of novel coronaviruses in least horseshoe bats in southwestern China. *Emerg. Microbes Infect.*, **6**, e14.
Wang, M., Yan, M., Xu, H., Liang, W., Kan, B., Zheng, B., Chen, H., Zheng, H., Xu, Y., Zhang, E., Wang, H., Ye, J., Li, G., Li, M., Cui, Z., et al. (2005) SARS-CoV infection in a restaurant from palm civet. *Emerg. Infect. Dis.*, **11**, 1860-1865.
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., et al. (2020) A new coronavirus associated with human respiratory disease in China. *Nature*, **579**, 265-269.
Wu, Z., Yang, L., Ren, X., Zhang, J., Yang, F., Zhang, S. & Jin, Q. (2016) ORF8-related genetic evidence for Chinese horseshoe bats as the source of human severe acute respiratory syndrome coronavirus. *J. Infect. Dis.*, **213**, 579-583.
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, A.C., Holmes, E.C. & 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.
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., et al. (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, **579**, 270-273.