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

Tetsuya Akaishi,1,2 Kei Fujiwara3 and Tadashi Ishii1,2

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

Correspondence: T. 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
Genetic recombination sites away from the insertion/deletion hotspots in SARS-related coronaviruses

Tetsuya Akaishi 1,2, Kei Fujiwara 3, Tadashi Ishii 1,2

1. Department of Education and Support for Regional Medicine, Tohoku University, Sendai, Japan

2. COVID-19 Testing Center, Tohoku University, Sendai, Japan

3. Department of Gastroenterology and Metabolism, Nagoya City University, Nagoya, Japan

Correspondence: T. Akaishi

Department of Education and Support for Regional Medicine, Tohoku University

Seiryo-machi 1-1, Aoba-ku, Sendai, Miyagi, 980-8574, Japan

Email: t-akaishi@med.tohoku.ac.jp

Running title: Recombination and indels in SARS-related coronaviruses
ABSTRACT

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 sarbecovirus species 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 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-related
coronaviruses; SARS-CoV-2
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 2022b; Akaishi 2022a). Such a novel and complex indel cannot be explained by conventional classical mutation 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 (Figure 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.

**Inference of phylogenetic ML trees in specific genomic regions**
To estimate ancestral states among the five SARS-related coronaviruses, gene region-specific phylogenetic maximum likelihood (ML) trees were inferred for different subdomains in \textit{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 (\textit{M}), \textit{ORF8}, and nucleocapsid (\textit{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.

\textbf{Recombination analysis}

To detect potential recombination sites in the \textit{ORF1ab} and \textit{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 \( P < 1.0E^{-20} \), 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 (2500–2600 base positions) and two in Nsp3 (3000–3300 and 3800–3900 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 Figure 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 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 Figure 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 (Figure 4). An indel hotspot was identified in the N-terminal domain of the S gene (21500–22500 bp). The actual sequences in the indel hotspot in the S gene (panel a) 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 (panels b and c). The inferred phylogenetic ML trees for the
three subdomains of the S gene are shown in Figure 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 (Figure 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 Re-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 1.0E–20 in the present study. The suggested recombinants (recombination
breakpoints) in ORF1ab were bat-SL-CoVZC45 (11600 and 20300 base positions, \( P = 6.66 \times 10^{-275} \)), bat Longquan-140 (1400 and 2770 base positions, \( P = 2.08 \times 10^{-188} \)), and two different recombination events in bat BtRI-BetaCoV/SC2018 (2690 and 3500 base positions, \( P = 1.79 \times 10^{-23} \); 8990 and 18820 base positions, \( P = 3.97 \times 10^{-67} \)). Those in the S gene were as follows: bat F46 (22580 base position, \( P = 5.61 \times 10^{-51} \)) and bat As6526 (23310 base position, \( P = 3.24 \times 10^{-42} \)). Pairwise identity line graphs for the first two recombinants in the ORF1ab gene are shown in Figure 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 2500–2600, 3000–3300, and 3800–3900 base positions). Pairwise identity graphs and MaxChi breakpoint matrices of the S gene are shown in Figure 7. None of the recombination sites fell within the range of the identified indel hotspot in the S gene (21500–22500 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 difficult 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 29900 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 evolutionary 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 sarbecovirus species 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


Tao, Y. & Tong, S. (2019) Complete Genome Sequence of a Severe Acute Respiratory Syndrome-Related Coronavirus from Kenyan Bats. *Microbiol Resour Announc*, **8**.


18


FIGURE LEGENDS

Figure 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-0319). The figure was constructed using MapChart software (available at https://www.mapchart.net/).

Figure 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 3000–3300 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

---

**Figure 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
Figure 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.

Figure 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

\textit{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

\textbf{Figure 6. Identification of recombination events in \textit{ORF1ab} gene of sarbecoviruses using RDP5.}

Two recombination events in the \textit{ORF1ab} gene, as identified in bat-SL-CoVZC45 (a) and Longquan-140 (b), fulfilled the statistical significance of $P < 1.0 \cdot 10^{-20}$ 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; RDP, Recombination Detection Program

\textbf{Figure 7. Identification of recombination events in \textit{S} gene of sarbecoviruses using Recombination Detection Program Version 5.}
Both recombination events in the spike gene (S), as identified in bat F46 (a) and As6526 (b), fulfilled the statistical significance of $P < 1.0 \times 10^{-20}$ 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.
Table 1. List of the evaluated 19 SARS-related coronavirus species

<table>
<thead>
<tr>
<th>GenBank ID</th>
<th>Author (publication year)</th>
<th>Country</th>
<th>Sampling year</th>
<th>Host</th>
<th>Virus species</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN908947</td>
<td>Wu (2020)</td>
<td>China</td>
<td>2019</td>
<td>Human</td>
<td>SARS-CoV-2 Wuhan-Hu′1</td>
</tr>
<tr>
<td>MN996532</td>
<td>Zhou (2020)</td>
<td>China</td>
<td>2013</td>
<td>Bat</td>
<td>RaTG13</td>
</tr>
<tr>
<td>LC556375</td>
<td>Murakami (2020)</td>
<td>Japan</td>
<td>2013</td>
<td>Bat</td>
<td>Re′o319</td>
</tr>
<tr>
<td>MT040333</td>
<td>Lam (2020)</td>
<td>China</td>
<td>2017–2018</td>
<td>Pangolin*</td>
<td>PCoV_GX-P4L</td>
</tr>
<tr>
<td>DQ412043</td>
<td>Li (2005)</td>
<td>China</td>
<td>2004</td>
<td>Bat</td>
<td>Bat CoV Rm1</td>
</tr>
<tr>
<td>KY938558</td>
<td>Son (2021)†</td>
<td>Korea</td>
<td>2017</td>
<td>Bat</td>
<td>Bat CoV 16BO133</td>
</tr>
<tr>
<td>MK211374</td>
<td>Han (2019)</td>
<td>China</td>
<td>2016–2017</td>
<td>Bat</td>
<td>BtRI-BetaCoV/SC2018</td>
</tr>
<tr>
<td>DQ071615</td>
<td>Li (2005)</td>
<td>China</td>
<td>2004</td>
<td>Bat</td>
<td>Bat CoV Rp3</td>
</tr>
</tbody>
</table>

* Malayan pangolin seized during anti-smuggling operation
† Unpublished
Fig. 1

- **Hubei Province**
  - *Human SARS-CoV-2 Wuhan-Hu-1 (2019)*

- **Zhejiang Province**
  - *Bat Longquan-140 (2012–2015)*

- **Guangdong Province**
  - *Bat HKU3-13 (2004–2008)*
  - *Bat Rm1 (2004)*
  - *Bat Rp3 (2004)*
  - *Civet civet007 (2004)*
  - *Human SARS-CoV_CUHK-AG01 (2003)*

- **Yunnan Province**
  - *Bat RaTG13 (2013)*
  - *Bat As5526 (2011–2015)*
  - *Bat F46 (2012)*
  - *Bat BtRs-BetaCoV/YN2013 (2013)*
  - *Bat Rs4084 (2011–2015)*

- **Guangxi Province**
  - *Pangolin PCoV_GX-P4L (2017–2018); smuggled Malayan pangolins*
Fig. 2

(a) Before indel hotspot in ORF1a (2.9–3.0 kb / 29.9 kb; Wuhan-Hu-1 genome)

(b) Beginning of indel hotspot in ORF1a (3.1–3.2 kb; Wuhan-Hu-1 genome)

(c) After indel hotspot in ORF1a (3.3–3.4 kb; Wuhan-Hu-1 genome)
Fig. 3

(a) ORF1ab Nsp1 – Nsp2 (0.25–2.7 kb)

(b) ORF1ab; Indel hotspot in Nsp3 gene (2.7–3.9 kb)

(c) ORF1ab; Nsp3 after indel hotspot (3.5–8.5 kb)

(d) ORF1ab; Nsp4 – Nsp16 (8.5–21.6 kb)
(a) Beginning of indel hotspot in S1 N-terminal domain (21.6–21.7 kb; Wuhan-Hu-1 genome)

(b) Beginning of S1 receptor-binding domain (22.6–22.7 kb; Wuhan-Hu-1 genome)

(c) Beginning of S2 gene (23.6–23.7 kb; Wuhan-Hu-1 genome)
Fig. 5
Fig. 6

(a) Recombinant: bat-SL-CoVZC45 (pairwise sequence identity with RDP method in ORF1ab gene)

(b) Recombinant: bat Longquan-140 (pairwise sequence identity with RDP method in ORF1ab gene)
Fig. 7

(a) Recombinant: bat F46 (pairwise sequence identity with RDP method in S gene)

(b) Recombinant: bat As6526 (pairwise sequence identity with RDP method in S gene)