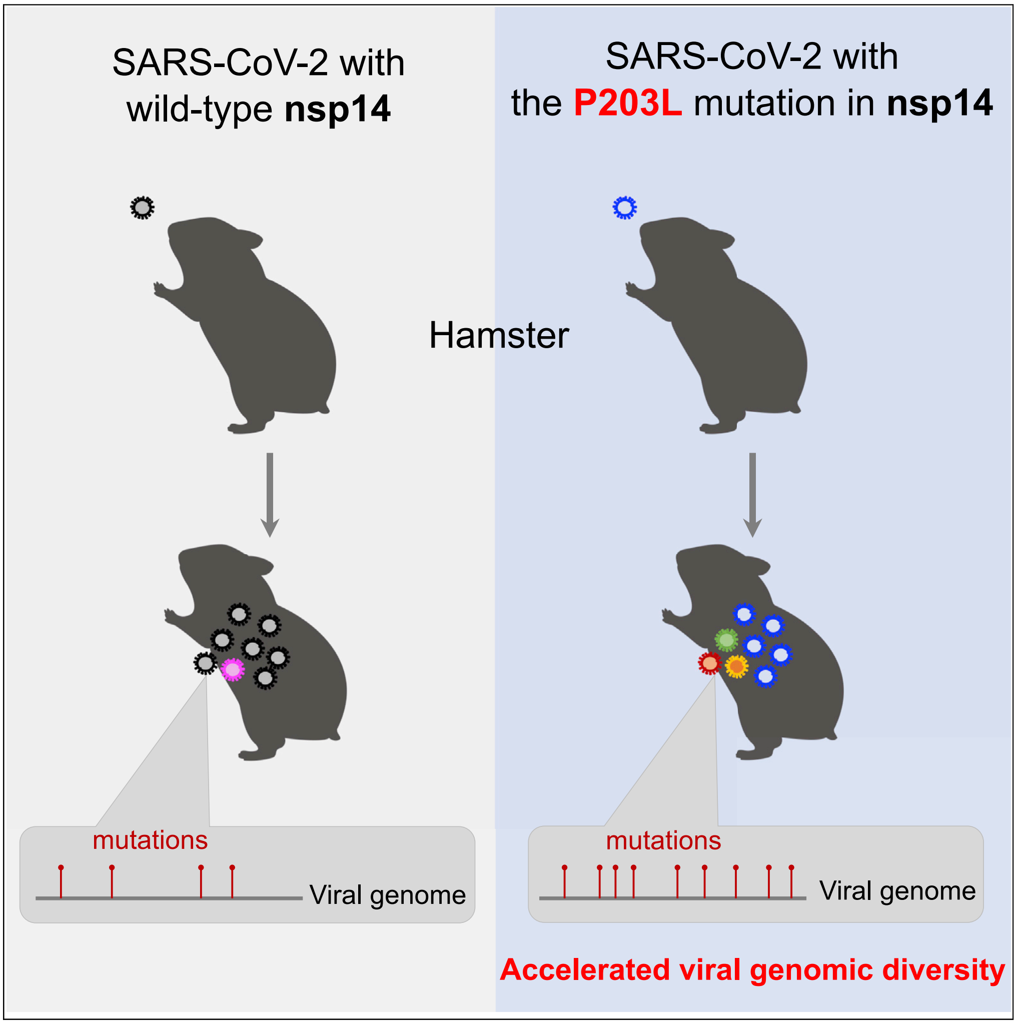
iScience

CellPress
OPEN ACCESS

Article

Genomic diversity of SARS-CoV-2 can be accelerated by mutations in the nsp14 gene



Kosuke Takada, Mahoko Takahashi Ueda, Shintaro Shichinohe, ..., Yoshiharu Matsuura, Tokiko Watanabe, So Nakagawa

tokikow@biken.osaka-u.ac.jp (T.W.)
so@tokai.ac.jp (S.N.)

Highlights

Amino acids of nonstructural protein 14 (nsp14) are well conserved in coronaviruses

P203L in nsp14 was not detected among coronaviruses but observed in SARS-CoV-2

Genome analysis suggested P203L nsp14 variants have a higher evolutionary rate

In vivo studies confirmed that the P203L nsp14 accelerates genomic diversity

Takada et al., iScience *26*, 106210
March 17, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.isci.2023.106210

Article

Genomic diversity of SARS-CoV-2 can be accelerated by mutations in the nsp14 gene

Kosuke Takada,[1,2] Mahoko Takahashi Ueda,[3] Shintaro Shichinohe,[1] Yurie Kida,[1] Chikako Ono,[4,5] Yoshiharu Matsuura,[4,5] Tokiko Watanabe,[1,5,6,*] and So Nakagawa[2,7,8,*]

SUMMARY

Coronaviruses, including severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), encode a proofreading exonuclease, nonstructural protein 14 (nsp14), that helps ensure replication competence at a low evolutionary rate compared with other RNA viruses. In the current pandemic, SARS-CoV-2 has accumulated diverse genomic mutations including in nsp14. Here, to clarify whether amino acid substitutions in nsp14 affect the genomic diversity and evolution of SARS-CoV-2, we searched for amino acid substitutions in nature that may interfere with nsp14 function. We found that viruses carrying a proline-to-leucine change at position 203 (P203L) have a high evolutionary rate and that a recombinant SARS-CoV-2 virus with the P203L mutation acquired more diverse genomic mutations than wild-type virus during its replication in hamsters. Our findings suggest that substitutions, such as P203L, in nsp14 may accelerate the genomic diversity of SARS-CoV-2, contributing to virus evolution during the pandemic.

INTRODUCTION

Virus mutates rapidly, resulting in the increased genetic diversity that drives virus evolution and facilitates virus adaptation. Generally, RNA viruses have a high mutation rate because they replicate their genome by using an error-prone RNA-dependent RNA polymerase (RdRp) so that nucleotide incorporation errors are not corrected.[1–3] In contrast, coronaviruses have lower mutation rates because they possess nonstructural protein 14 (nsp14), which contains an exoribonuclease (ExoN) domain that provide a proofreading function.[4,5] Nonetheless, in the case of severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), which emerged in China at the end of 2019 and caused the COVID-19 pandemic, numerous variants have emerged by acquiring multiple mutations in a short period of time. Some of these variants have been a cause for concern because of their increased infectivity, transmissibility, pathogenicity, and/or reduced sensitivity to vaccines and therapeutic agents. These variants, classified as 'variants of concern (VOCs)' by the World Health Organization (WHO), include Alpha (PANGO ID B.1.1.7), Beta (PANGO ID B.1.351), Gamma (PANGO ID P.1), Delta (PANGO ID B.1.617.2), and most recently Omicron (PANGO ID B.1.1.529). To control COVID-19 and bring the pandemic to an end, it is important to understand the mechanism of the emergence of these diverse variants and the genomic evolution of SARS-CoV-2.

Coronaviruses possess a positive-sense, single-stranded RNA genome that is one of the largest among known RNA viruses, ranging from 26 to 32 kb.[2] There are two large open reading frames (ORFs) at the 5′ proximal end, ORF1a and ORF1b, which comprise approximately two-thirds of the genome, and encode two large polyproteins, pp1a and pp1ab, that are cleaved into about 16 nonstructural proteins (nsps). Some of the processed nsps assemble to form the replication-transcription complex, which contains an RdRp (nsp12), the helicase (nsp13), processivity factors (nsp7 and nsp8), and the proofreading ExoN complex (nsp14 and nsp10).[6–8] During genome replication, nsp14 ExoN activity plays a role in the RNA proofreading machinery by efficiently removing mismatched 3′-endnucleotides.[9] Experiments in which the nsp14 of SARS-CoV was functionally inactivated revealed 10–20 times higher mutation rates,[5,10,11] indicating that nsp14 may be essential for maintaining the virus genome.

In this study, to clarify whether amino acid substitutions in nsp14 affect the genomic diversity and evolution of SARS-CoV-2, we searched for amino acid substitutions in nature that may interfere with the function of nsp14 in SARS-CoV-2. First, to obtain functionally important sites in nsp14, we examined 62 representative coronaviruses belonging to the family Coronaviridae and found that 99 of the 527 amino acid sites of nsp14

[1] Department of Molecular Virology, Research Institute for Microbial Diseases, Osaka University, Suita, Osaka 565-0871, Japan

[2] Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Kanagawa 259-1193, Japan

[3] Department of Genomic Function and Diversity, Medical Research Institute, Tokyo Medical and Dental University, Bunkyo, Tokyo 113-8510, Japan

[4] Laboratory of Virus Control, Research Institute for Microbial Diseases, Osaka University, Suita, Osaka 565-0871, Japan

[5] Center for Infectious Disease Education and Research, Osaka University, Suita, Osaka 565-0871, Japan

[6] Center for Advanced Modalities and DDS, Osaka University, Suita, Osaka 565-0871, Japan

[7] Bioinformation and DDBJ Center, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan

[8] Lead contact

*Correspondence: tokikow@biken.osaka-u.ac.jp (T.W.), so@tokai.ac.jp (S.N.)

https://doi.org/10.1016/j.isci.2023.106210

were evolutionarily conserved. We then examined nsp14 sequences obtained from 28,082 SARS-CoV-2 genomes available in the GISAID EpiCoV database, which includes not only sequencing data but also sampling date, country, and institution information,[12] and identified an additional 6 amino acid changes in nsp14 mutants that were not detected in the 62 representative coronaviruses. We examined the genome mutation rates of these mutants and found that the amino acid replacement of proline (P) for leucine (L) at position 203 of nsp14 could be involved in increasing the nucleotide mutation rate of SARS-CoV-2 genomes. To investigate the biological significance of the nsp14 P203L substitution, we generated recombinant SARS-CoV-2 viruses possessing wild-type nsp14 or mutant nsp14 with P203L and characterized them in a hamster model. We found that the nsp14-P203L mutant grown in the lungs of hamsters acquired significantly more diverse genomic mutations than did the wild-type virus. These results indicate that substitutions in nsp14, such as P203L, could accelerate the genomic diversity of SARS-CoV-2.

RESULTS

Conserved regions of nsp14 proteins among coronaviruses

We first generated a phylogenetic tree of the amino acid sequences of the nsp14 gene obtained from 62 representative coronaviruses (the viruses used in this analysis are summarized in Table S1). The phylogenetic tree clearly showed the relationship among the four genera of coronaviruses (Figure 1A), consistent with results obtained by using the partial amino acid sequences of ORF1ab described in a previous report.[13] Our results suggest that the nsp14 gene is broadly conserved among coronaviruses, and that the conserved amino acid sites could be important for its molecular function.

Next, we examined the conserved amino acid sites of the nsp14 gene among the 62 coronaviruses (Figure 1B). The amino acid positions of nsp14 were based on the SARS-CoV three-dimensional structure (PDB ID: 5C8S); SARS-CoV and SARS-CoV-2 both belong to the subgenus *Sarbecovirus* in the genus *Betacoronavirus*. For nsp14, the number of identical amino acids was 482/527 (91.5%) in 8 representative coronaviruses belonging to the subgenus *Sarbecovirus*, and 359/527 (68.1%) if we included the nsp14 of the Bat Hp-betacoronavirus/Zhejiang2013 (GenBank ID: NC_025217.1) belonging to the subgenus *Hibecovirus*, an outgroup of *Sarbecovirus*. Moreover, 200/527 (38.0%) positions were conserved among 29 viruses of the genus *Betacoronavirus*, and 108/527 (20.5%) positions were conserved among the 61 coronaviruses (Figure 1B). If we included an unclassified coronavirus found in *Tropidophorus sinicus* (Chinese waterside skink),[15] the conserved amino acid positions were 99/527 (18.8%). These results suggest that these amino acid sites of the nsp14 gene are conserved among diverse coronaviruses because they may be functionally important. We further validated this hypothesis by calculating non-synonymous (dN) and synonymous (dS) substitution rates on a per-site basis for the nsp14 sequences of the 62 coronaviruses by using the SLAC program (see STAR Methods).[16] For the dN/dS analysis, we removed 23 amino acid sites that contained insertions and deletions out of the total 527 amino acid sites in the SARS-CoV-2 nsp14 of 61 coronaviruses (excluding the Chinese waterside skink coronavirus that contains a mixture of nucleotides at these positions). Thus, 504 amino acid sites of the 61 coronavirus nsp14 multiple alignment were calculated for the dN/dS ratios; no positive selection sites were found in nsp14, whereas 379 of the 504 codon sites were under negative selection ($p < 0.01$, Figure S1). In total, 91 amino acids were unchanged and under negative selection in the nsp14 of the 62 representative coronaviruses.

We then investigated whether the amino acids that are known to be important for the error-correcting function of nsp14 are conserved. Coronavirus nsp14 possesses both ExoN and N7-MTase activities. The ExoN domain of coronavirus nsp14 was originally identified based on its sequence similarity to distant cell homologs[9] and has been assigned to the DEDD exonuclease superfamily, which includes the proofreading domains of many DNA polymerases and other eukaryotic and prokaryotic exonucleases. The nsp14 proteins have five conserved active site residues distributed in three standard motifs of the primary structure: residues 90D/92E (motif I), 191E (motif II), and 268H/273D (motif III).[17–19] We found that the five active sites of nsp14 were conserved among the 62 coronaviruses examined in this study (Figure 1C). Moreover, there were three zinc fingers that were essential for nsp14 function: zinc finger 1 (ZF1) comprising 207C, 210C, 226C, and 229H; zinc finger 2 (ZF2) comprising 257H, 261C, 264H, and 279C; and zinc finger 3 (ZF3) comprising 452C, 473C, 484C, and 487H. S-adenosyl methionine (SAM)-binding motif I includes residues 331D, 333G, 335P, and 337A (or 337G), which are involved in the (N7 guanine)-methyl transferase reaction by nsp14. We found that all these residues were conserved among the nsp14 sequences of the 62 coronaviruses examined (Figure S2). These results suggest that the amino acids of nsp14 that are functionally important are conserved among the diversified coronaviruses.

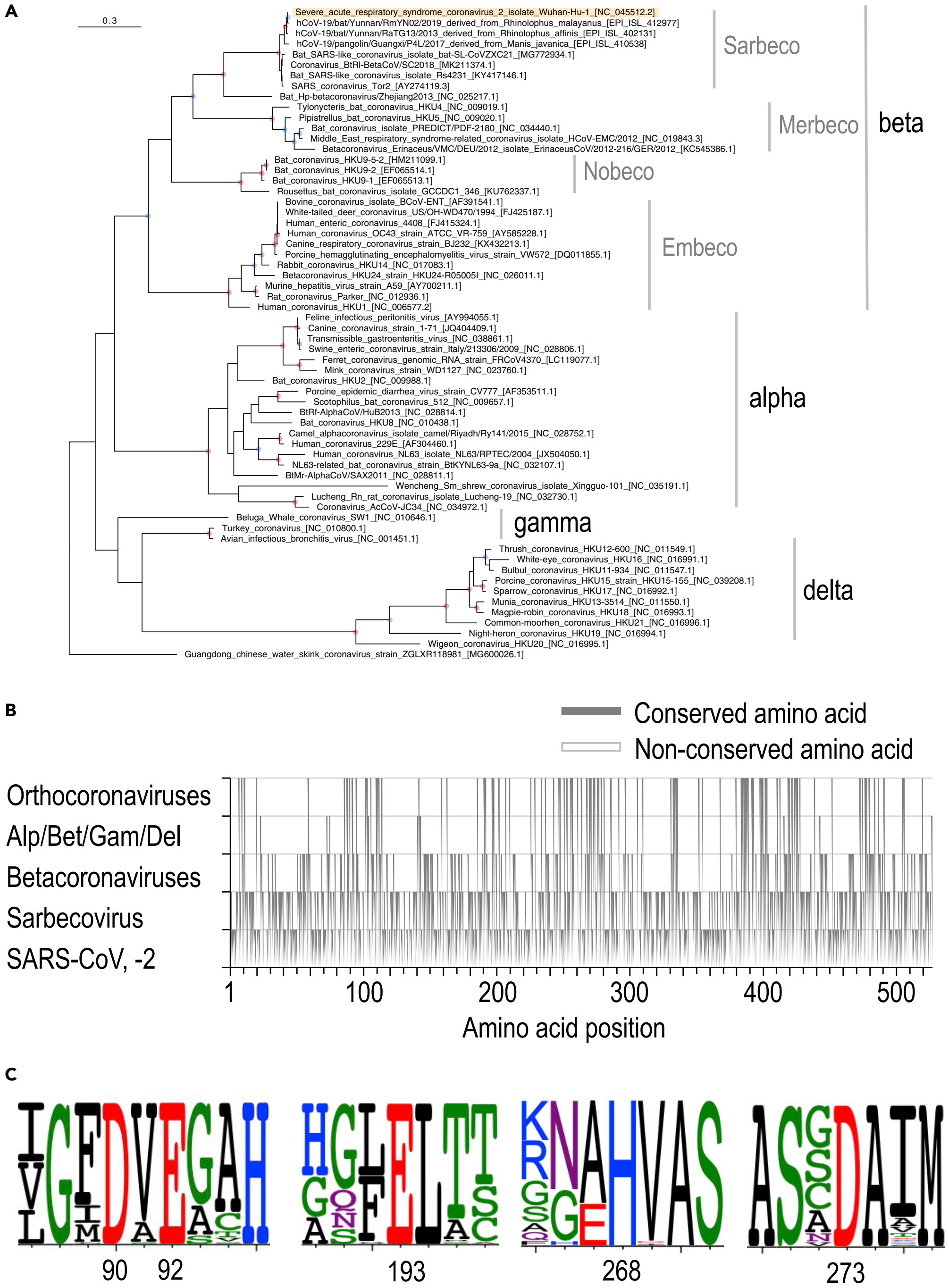


Figure 1. Phylogenetic tree and amino acid sequence alignment of nsp14 derived from representative coronaviruses

(A) Maximum likelihood (ML)-based phylogenetic tree of the 62 representative coronaviruses. Amino acid sequences of nsp14 in ORF1ab were used for this analysis. A red or blue circle in an internal node corresponds to bootstrap values ≥95% or ≥80%, respectively. The virus located at the top is SARS-CoV-2 (hCoV-19/Wuhan/Hu-1/2019).

(B) Conserved amino acid residues among each group. The amino acid sequence of the nsp14 of the 62 representative coronaviruses is shown according to the relative amino acid positions of SARS-CoV-2 nsp14 (note that the gaps in the SARS-CoV-2 sequence are not shown). The conserved amino acid positions are indicated in gray.

(C) Five conserved active site residues involved in proofreading. The amino acid proportions for each site were calculated based on the sequence alignment of the nsp14 of the 62 representative coronaviruses visualized by using WebLogo.[14] The amino acids are colored according to their chemical properties. See also Figure S2 for the entire alignment.

Table 1. Amino acid sequence of nsp14 in SARS-CoV-2 epidemic strains

Appearance frequency ranking	Substitution in nsp14	Number of sequences	Frequency of appearance (%)
1	–	25,913	92.27
2	A320V	192	0.68
3	F233L	144	0.51
4	S374A	123	0.44
5	**L177F**	**116**	**0.41**
6	T250I	61	0.22
7	**F377L**	**51**	**0.18**
8	A138V	39	0.14
9	V120A	35	0.12
10	T16I	34	0.12
11	**S369F**	**33**	**0.12**
12	G481S	31	0.11
13	**M501I**	**28**	**0.10**
13	D345G	28	0.10
13	N129D	28	0.10
16	**P203L**	**26**	**0.09**
16	**L177F/T372I**	**26**	**0.09**
18	**P297S**	**25**	**0.09**
18	T31I	25	0.09
20	D324Y	24	0.09

28,082 SARS-CoV-2 genome sequences that were annotated as Human and free of undetermined or mixed nucleotides were used in the analysis. Amino acid substitutions were defined as a change relative to the hCoV-19/Wuhan/Hu-1/2019 reference sequence. Boldface characters indicate that the amino acid replacement was not observed in the 62 coronaviruses examined in this analysis. —, no mutation was found compared with the reference sequence.

Diversity of nsp14 in SARS-CoV-2

Amino acid comparisons of nsp14 in the 62 representative coronaviruses revealed that several sites are strongly conserved and that some substitutions may have deleterious effects on the survival of SARS-CoV-2. SARS-CoV-2 was first detected in humans in 2019 and has spread rapidly all over the world in a short period of time. Therefore, some nsp14 mutations that impair its error-correcting function may remain in the SARS-CoV-2 population. Accordingly, we attempted to identify such deleterious mutations in the current SARS-CoV-2 population.

We downloaded the aligned nucleotide sequences of SARS-CoV-2 (87,625 sequences) from the GISAID database as of September 7, 2020. The SARS-CoV-2 genomes isolated from humans were extracted (87,304 sequences), and the sequences containing undetermined and/or mixed nucleotides were removed; the remaining 28,082 sequences were used in subsequent analyses. The coding region of nsp14 was translated and compared. The results showed that 25,913 nsp14 amino acid sequences (92.27%) were identical to that of the reference sequence, whereas 2169 sequences contained at least one amino acid substitution in nsp14 (Table S2). The top 20 variants possessing nsp14 with amino acids that were different from the major sequence (shown in Figure S2) are summarized in Table 1. Note that frameshift mutants were excluded from the table because they may affect downstream genes as well. Notably, among the 20 amino acid nsp14 mutants, six amino acids (177F, 377L, 369F, 501I, 203L and 297S) were not found in the 62 nsp14 proteins of the representative coronaviruses (Figure S2). We further analyzed the evolutionary selection pressures of these 6 amino acid sites by using 73 coronaviruses belonging to the genus *Betacoronaviruses* that includes SARS-CoV-2 (Table S3). We confirmed that the six amino acid variants were not found in any of the 73 Beta-coronaviruses (Table S3). The six conserved positions (177, 203, 297, 369, 377, and 501) were also found to be under negative selection ($p < 0.01$, Table S4). Overall, these results indicate that the conserved amino acid sites of nsp14 are also evolutionarily selected in the genus *Betacoronaviruses*.

Table 2. Nucleotide mutation rates of SARS-CoV-2 with different nsp14 variants

Ranking of nsp14	Substitution in nsp14	Number of comparison sequences	Substitution rate per year in the whole genome
1	–	25,471	19.8
5	L177F	114	27.3
7	F377L	51	13.0
11	S369F	33	26.2
13	M501I	28	21.9
16	P203L	24	35.9
17	L177F/T372I	26	59.8
18	P297S	25	17.1

The number of mismatched nucleotides in the alignment of the reference sequence with the comparison sequence of each group was counted, and the average value was calculated. hCoV-19/Wuhan/Hu-1/2019 was selected as the reference sequence. Note that the estimated substitution rates were not phylogenetically corrected. —, no mutation was found compared with the hCoV-19/Wuhan/Hu-1/2019 sequence.

Correlation between the amino acid sequences of nsp14 and nucleotide mutation rates in the viral genome

To examine whether any of the six amino acid substitutions in the nsp14 of SARS-CoV-2 variants could affect its proofreading function, we looked for a correlation between the amino acid variants of nsp14 and the genetic diversities of SARS-CoV-2 genomes. For this purpose, we roughly estimated evolutionary rates of SARS-CoV-2 possessing mutated nsp14, compared to that of SARS-CoV-2 with wild-type nsp14. Sequences for which sampling date information (year, month, and day) was not available were excluded from this analysis. Nucleotide mutations per year were calculated based on the regression coefficient using hCoV-19/Wuhan/Hu-1/2019 (GISAID ID: EPI_ISL_402125) as the reference sequence (see STAR Methods). Note that the estimated evolutionary rates of each nsp14 mutant were not phylogenetically corrected; the estimated rates were used only to obtain nsp14 mutation candidate(s) that may affect evolutionary rates.

We found that SARS-CoV-2 containing wild-type nsp14 had 19.8 nucleotide mutations in its whole genome per year (Table 2; Figure S3 for the regression coefficient). Conversely, seven SARS-CoV-2 variants containing nsp14 with amino acid substitutions showed 27.3 (L177F), 59.8 [L177F/T372I (*i.e.*, L177F and T372I)], 35.9 (P203L), 17.1 (P297S), 26.2 (S369F), 13.0 (F377L), or 21.9 (M501I) nucleotide mutations per year (Table 2; Figure S3). Although the substitution rates of the L177F/T372I and F377L variants were found to be unreliable because of limited time ranges, those of the other variants were correlated (Figure S3). Therefore, SARS-CoV-2 variants possessing nsp14 with the L177F, P203L, S369F, or M501I substitution may have a higher genomic mutation rate than that of SARS-CoV-2 carrying wild-type nsp14. In particular, the P203L variant of nsp14 (nsp14-P203L), which showed the highest mutation rate (35.7 nucleotide mutations per year), is a strong candidate that enhances the diversity of SARS-CoV-2 genomes.

We then verified the substitution rates of the nsp14-P203L variants for each single cluster in which the maximum number was observed: 6 genomes for the nsp14-203L mutant and 194 genomes for the nsp14-203P wild-type in the B.1.1.33 lineage (Figures 2A and S4 with GISAID IDs). We found that nsp14-P203L variants in the B.1.1.33 lineage forming a single clade showed a high nucleotide evolutionary rate (58.8 nucleotide mutations/year) compared with those with the nsp14-203P (i.e., wild-type) strains in the same lineage (16.5 nucleotide mutations/year) (Figure 2B and Table 3). Although the number of variants was limited (6), a correlation was clearly observed (Figure 2B). In addition, six other nsp-P203L variants in a different cluster (B.1) also showed a relatively high nucleotide mutation rate (28.0 nucleotide mutations/year), assuming that variants were in the same clade if their PANGO IDs[20] and GISAID clades were identical (Table 3; Figure S5 for regression coefficients). With the same assumption, we did the same calculation for the other nsp14 mutants in the lineages (Table 3; Figure S5) and found that nsp14-P203L in the B.1.1.33 lineage exhibited the highest substitution rate. These results indicate that SARS-CoV-2 variants containing nsp14-P203L are strong candidates for variants that readily accumulate mutations in the viral population.

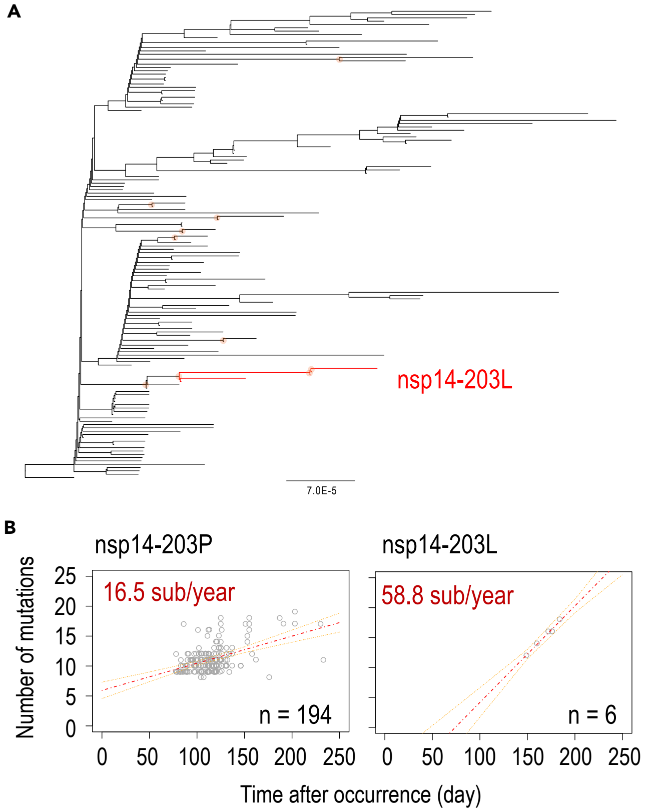


Figure 2. Phylogeny and nucleotide mutation rate of nsp14-P203L variants in the PANGO lineage B.1.1.33
(A) ML tree of SARS-CoV-2 genomes in the B.1.1.33-lineage. The nsp14-P203L variant is highlighted in red. An orange circle corresponds to bootstrap values ≥70%. See also Figure S4 for the tree with GISAID ID.
(B) Nucleotide mutation rates of nsp14-203P or nsp14-203L variants belonging to the B.1.1.33-lineage. All sample dots are displayed. Day-0 was set to when the Wuhan-Hu-1 reference strain was sampled (December 26, 2019). Mutation rates per year in the genome are shown in red letters. The red and orange dotted lines correspond to the regression line and 95% confidence intervals, respectively.

The frequency of nucleotide mutations of coronaviruses may be affected not only by the nsp14 proofreading mechanism but also by the characteristics of other proteins, such as the fidelity of the RNA-dependent RNA polymerase (nonstructural protein 12; nsp12).[21] It is possible that amino acid substitutions in other proteins involved in genome replication affect the nucleotide mutation rates of SARS-CoV-2. Therefore, we examined amino acid substitutions among nsp14-203L variants involving the following genes that participate in gene replication: nsp7, nsp8, nsp9, nsp10, nsp12, nsp13, nsp15, and nsp16. The results showed that a proline-to-leucine substitution at position 323 (P323L) of nsp12 frequently occurred (20/25) in sequences of SARS-CoV-2 possessing nsp14-203L (Table S5). However, the P323L mutation in nsp12 (nsp12-P323L) is known to accompany an aspartate-to-glycine substitution at position 614 (D614G) of S protein.[22,23] The S-D614G mutant was first reported at the end of January 2020 in China and Germany, and by March 2020, this mutant was a major variant in various regions all over the world.[22] The P323L mutation in nsp12 was found in nsp14-203L variants as well. Indeed, the genomic mutation rate of SARS-CoV-2 with the P323L amino acid substitution in nsp12, was found to be 17.7 nucleotide mutations per year (Figure S6). Therefore, it is unlikely that P323L in nsp12 directly affects the genomic mutation rate of SARS-CoV-2 viruses carrying the P203L amino acid substitution in nsp14.

The nsp14-P203L variant in the SARS-CoV-2 population

The nsp14-P203L variants have been isolated from 26 patients mainly in Europe and North America, with the first being reported on March 5, 2020 in The Netherlands (GISAID ID: EPI_ISL_454756, Table S6). Considering their PANGO lineages[20] and GISAID clades (shown in Table S6), the nsp14-P203L variants

Table 3. Nucleotide mutation rates of the seven nsp14 variants for each cluster in which the maximum number was observed

Ranking of nsp14	Substitution in nsp14	Lineage (PANGO ID)	Number of sequences containing the nsp14 substitution	Substitution rate per year in the whole genome
5	L177F	B.1.1.25	54	7.8
7	F377L	B	50	8.5
11	S369F	B.1.1	17	13.4
13	M501I	B.1	11	4.4
16	P203L	B.1.1.33	6	58.8
		B.1	6	28.0
17	L177F/T372I	B.1.113	20	50.1
18	P297S	B.4	11	16.1

The number of mismatched nucleotides in the alignment of the reference sequence with the comparison sequence of each group was counted, and the average value was calculated. hCoV-19/Wuhan/Hu-1/2019 was selected as the reference sequence. A scatterplot of these nsp14 is shown in Figures 2B and S5.

appeared at least 10 times. In fact, the nsp14-P203L substitution was observed in both Spike D614 strains and G614 variants, which have enhanced infection efficiency and have spread worldwide.[22] However, we also found that the prevalence of nsp14-P203L variants did not increase steadily over time; the largest number of nsp14-P203L variants found in the B.1.1.33 and B.1 lineages was only six for each (Tables 3 and S6).

The nsp14-P203L mutation effect on protein structure

To assess the effect of the P203L mutation of nps14, we first investigated the location of residue 203 by using a tertiary structure model of SARS-CoV-2 nsp14 (PDB ID: 7EGQ, Figures 3A and S7A).[24] The residue 203 is exposed on the surface of the protein and is in the loop stretching from residues 200 to 214, which is in the interstices of the zinc and magnesium binding sites (positions 207 and 191). We found that the loop is located at the edge of an interaction surface for binding to the co-factor SARS-CoV-2 nsp10 (Figures 3B and S7A). The interaction with nsp10 has been shown to greatly increase the nucleolytic activity of SARS-CoV nsp14.[25] Therefore, to investigate the possibility that residues at position 203 play an important role in the nsp10-nsp14 interaction, we calculated the change in binding free energy (ΔΔGbind) induced by nsp14-P203L by using the BeAtMuSiC program. The predicted result showed that this substitution

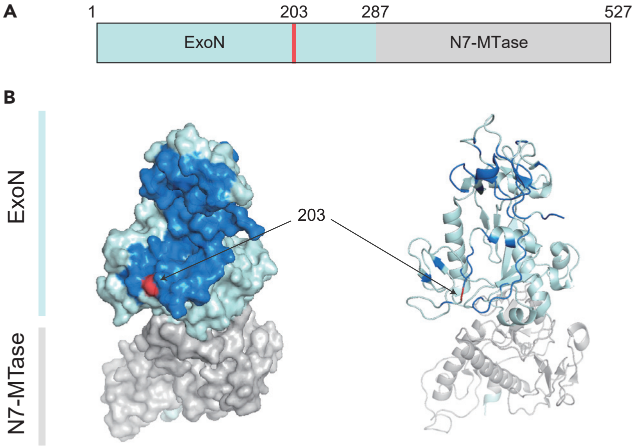


Figure 3. Three-dimensional position of the amino acid at position 203 in nsp14

(A) Domain structure of SARS-CoV-2 nsp14 with residue numbers. The ExoN and N7-Mtase domains are shown in light blue and gray, respectively. The amino acid at position 203 in nsp14 is shown in magenta.

(B) Schematic representation of nsp14. The coloring is the same as in (A). The site of interaction with nsp10 is shown in dark blue.

See also Figure S7.

decreases binding affinity (ΔΔGbind = 0.43 kcal/mol), which could be related to increasing the hydrophobicity by the P203L amino acid replacement (Figure S7B). These results suggested that the P203L of nsp14 may reduce binding affinity to nsp10, possibly resulting in relatively low fidelity of RNA replication.

Characterization of a recombinant SARS-CoV-2 virus possessing nsp14-P203L *in vitro*

Genome data analysis suggested that SARS-CoV-2 variants containing nsp14-P203L could accumulate mutations in the viral population. Does the nsp14-P203L amino acid substitution affect viral properties (e.g., increasing the diversity of the viral genome)? To attempt to answer this question, we generated recombinant SARS-CoV-2 (based on hCoV-19/Japan/TY-WK-521/2020) viruses possessing wild-type nsp14 or P203L nsp14 by using a recently established circular polymerase extension reaction (CPER) method.[26] We then we examined the effect of the nsp14-P203L amino acid substitution on the biological properties of SARS-CoV-2.

To determine whether the nsp14-P203L amino acid substitution affects the viral growth efficiency in cultured cells, we examined viral replication efficiency in VeroE6/TMPRSS2 cells. We found that both wild-type virus and nsp14-P203L viruses grew well, with max titers (mean titers = 7.35 log unit and 7.01 log unit, respectively) at 36 h post-infection (Figure 4A), although the titer of the wild-type virus at 24 h post-infection was slightly higher than that of the nsp14-P203L virus (Figure 4A).

Next, we analyzed the genomic diversity of the recombinant virus grown in the cultured cells. Three plaque-purified wild-type or nsp14-P203L virus clones were used to infect VeroE6/TMPRSS2 cells, and the genomes of the viruses in the supernatant were collected 48 h later and analyzed by RNA sequencing. Mutations with a read depth of at least 1000 times and an allele frequency of at least 0.5% were compared between the wild-type and nsp14-P203L mutant viruses. The nsp14-P203L mutant grown in VeroE6/TMPRSS2 had 51.33 (±10.96) mutations and the wild-type virus had 46.33 (±9.46) mutations, which was not significantly different ($p = 0.65$) (Figure 4B). These results suggest that nsp14-P203L allows efficient virus replication and may not affect viral genome mutagenesis in culture cells.

Biological properties of the recombinant nsp14-P203L virus *in vivo*

Next, to examine whether the nsp14-P203L amino acid substitution affects viral properties *in vivo* (e.g., increasing viral replication efficiency and/or viral genome diversity), we investigated the biological properties of the nsp14-P203L mutant in a hamster model (Figure 4C). We inoculated the hamsters intranasally with 10^3 PFU of wild-type virus or virus possessing nsp14-P203L and collected lung and nasal turbinate samples at 3 and 6 days post-infection (dpi) to analyze viral titers (Figure 4D). We found that the nsp14-P203L virus replicated well in both the lungs and nasal turbinates of the infected hamsters at 3 and 6 dpi, and the viral titers in the lungs of the hamsters infected with the nsp14-P203L virus at 3 dpi were similar to those in the lungs of the hamsters infected with the wild-type virus (mean titers were 8.38 log unit or 8.33 log unit, respectively). We also observed clinical symptoms, including weight loss for 6 days post-infection. We found that hamsters infected with the nsp14-P203L virus showed similar weight loss to hamsters infected with the wild-type virus (Figure 4E). The nsp14-P203L mutant virus thus grew as efficiently as the wild-type virus in hamster lungs and nasal turbinate and caused similar weight loss.

To further investigate the effect of nsp14-P203L on viral genome diversity *in vivo*, we RNA-sequenced viral genomes extracted from viruses isolated from the lungs of hamsters infected with the nsp14-P203L or wild-type virus at 6 dpi. We found that the number of mutations in the nsp14-P203L group was significantly higher than those detected in the wild-type group (63.25 ± 12.54 mutations versus 32.75 ± 3.34 mutations, respectively; $p = 0.0066$) (Figure 4F), and that these mutations were widely distributed in the viral genomes (Figure 4G). These results indicate that the nsp14-P203L virus replicates well in the respiratory tracts of hamsters and readily acquires more diverse mutations.

DISCUSSION

The mutation rate of coronaviruses, including SARS-CoV-2, is relatively low compared to other RNA viruses because coronaviruses possess an error-correcting exonuclease, nsp14. In this study, we examined mutations that occurred in nsp14, and found that SARS-CoV-2 possessing nsp14 with a proline to leucine substitution at position 203 showed a relatively higher substitution rate than SARS-CoV-2 possessing wild-type nsp14 (Tables 2 and 3; Figure 2). Protein structure prediction analysis of SARS-CoV-2 nsp14 suggested that

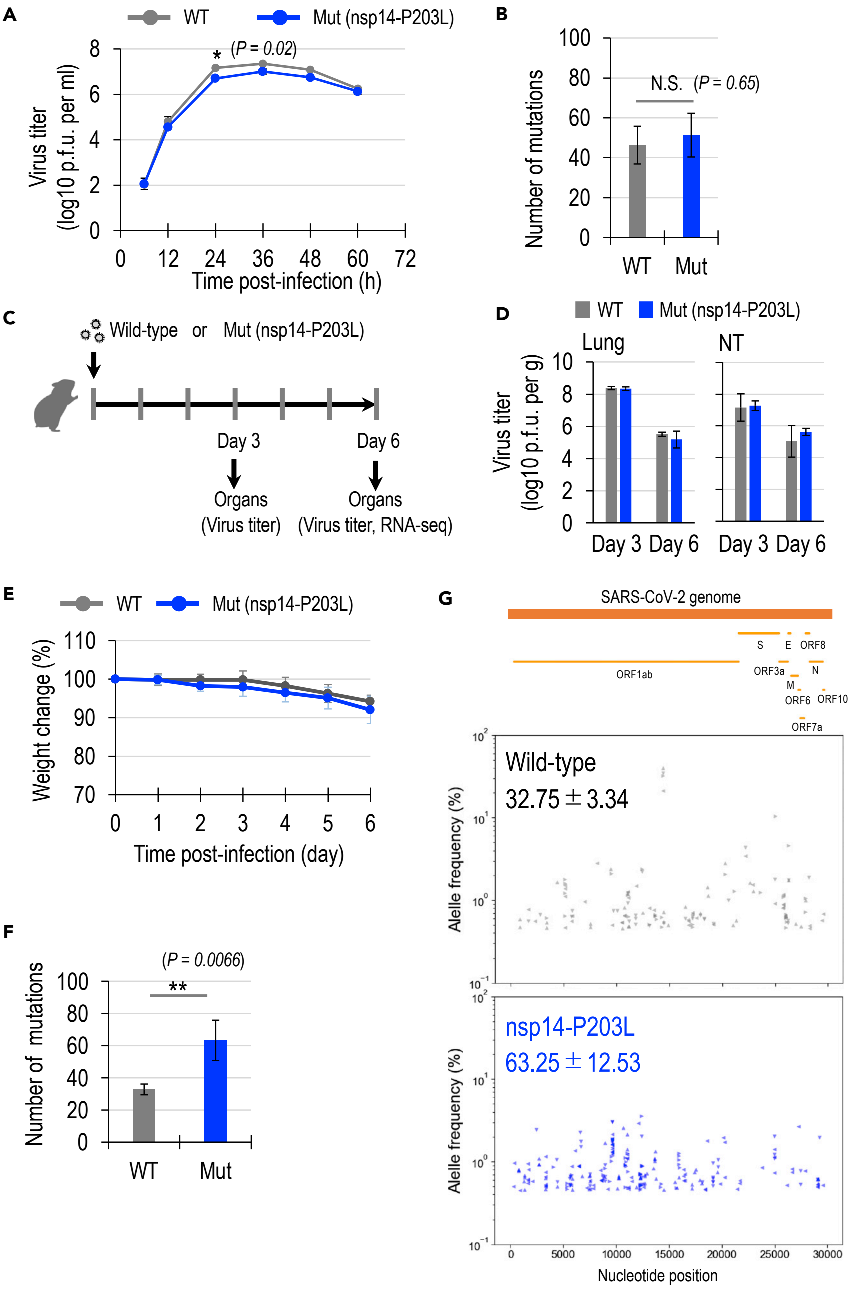


Figure 4. Characterization of the recombinant nsp14-P203L virus *in vitro* and *in vivo*

(A and B) Growth efficiency and genomic mutations of nsp14-P203L variants in cultured cells. (A) Growth kinetics of recombinant viruses in VeroE6/TMPRSS2 cells. VeroE6/TMPRSS2 cells were infected with viruses at a multiplicity of infection of 0.001. The supernatants of the infected cells were harvested at the indicated times, and virus titers were determined by the using plaque assays in VeroE6/TMPRSS2 cells. Data are shown as the mean of three independent experiments done in triplicate. Error bars indicate the standard deviation. Mean values were compared by using a two-way ANOVA, followed by Dunnett's test (*p <0.05). Asterisks indicate statistically significant differences between WT and Mut viruses. (B) Number of genomic mutation types detected in viruses grown in VeroE6/TMPRSS2 cells. Three wild-type or mutant viruses were each plaque-purified, and each viral clone was infected into VeroE6/TMPRSS2 cells; supernatants were collected after 48 h. RNA was extracted from the virus in the supernatant and RNA sequencing was performed. Mutations with a read depth of at least 1000x and an allele frequency of at least 0.5% were compared between the wild-type and mutant viruses. Mean values were compared by using an unpaired Student's t test (N.D.; p >0.05).

(C–G) Growth efficiency and genomic mutations of nsp14-P203L variants in Syrian hamsters. (C) Schematic overview of the animal experiment. Syrian hamsters were inoculated with 10^3 PFU of wild-type or nsp14-P203L virus via the intranasal

Figure 4. *Continued*

(30 μL) route. (D) Virus replication in infected Syrian hamsters. Four Syrian hamsters per group were killed on days 3 and 6 post-infection for virus titration. Virus titers in lung or nasal turbinate (NT) organs were determined by using plaque assays in VeroE6/TMPRSS2 cells. Error bars indicate the SD. Mean values were compared by using a two-way ANOVA, followed by Dunnett's test. (E) Body weight changes in Syrian hamsters after viral infection. Body weights of virus-infected (n = 8) animals were monitored daily for 3 days. Body weights of virus-infected (n = 4) animals were monitored daily for 6 days. Data are presented as the mean percentages of the starting weight (±SD). (F) Number of genomic mutation types detected in viruses grown in Syrian hamster lung. RNA was extracted from the virus in lung homogenates from the Day 6 post-infection Syrian hamsters, and RNA sequencing was performed. Mutations with a read depth of at least 1000x and an allele frequency of at least 0.5% were compared between the wild-type and mutant viruses. Mean values compared by using an unpaired Student's *t* test (**p <0.01). (G) SARS-CoV-2 genome organization and genomic locations of synonymous or nonsynonymous substitutions detected in the genomes of viruses grown in hamster lungs. The different directions of the triangles indicate each of the four individual hamsters. The number of nucleotide changes in viruses grown in Syrian hamster lung. The number of nucleotide changes, detected as synonymous and nonsynonymous substitutions, at a read depth of at least 1000x and an allele frequency of >0.5% were compared between the wild-type and nsp14-P203L mutant viruses.

the P203L replacement in nsp14 decreases the binding affinity of nsp10, possibly reducing the proofreading activity (Figures 3 and S7). Moreover, we demonstrated that a recombinant SARS-CoV-2 possessing this P203L substitution in nsp14 acquired significantly more diverse genomic mutations than the wild-type virus did during its replication in a hamster model, suggesting that mutations in nsp14, such as P203L, could accelerate the genomic diversity of SARS-CoV-2 (Figure 4).

The nsp14-P203L mutant virus grown in VeroE6/TMPRSS2 cells did not have significantly diverse genomic mutations compared to the wild-type virus, whereas the nsp14-P203L mutant virus grown in hamster lungs had significantly diverse mutations in its genome compared to the wild-type virus. One reason for this difference may be the number of genome copies of each virus in the experiments. Viruses grown in VeroE6/TMPRSS2 cells were sampled to examine genomic mutations at 2 days post-infection, whereas viruses grown in hamster lungs were assessed at 6 days post-infection. Therefore, the viruses may have replicated more in the hamsters' lungs due to the longer incubation time. In addition, in VeroE6/TMPRSS2 cells, the nsp14-P203L mutant virus showed slightly reduced growth efficiency compared to the wild-type virus (Figure 4A). Therefore, the wild-type virus may replicate more frequently in VeroE6/TMPRSS2 cells compared to the nsp14-P203L mutant virus, resulting accumulation of genomic mutations. In contrast, in hamster lungs, the replication efficiency of the nsp14-P203L mutant virus was similar to that of the wild-type virus (Figure 4D). Therefore, because the genome replication frequency is almost the same between the wild-type and the nsp14-P203L mutant viruses, the reduced proofreading activity owing to nsp14-P203L can be correctly detected. Overall, our findings suggest that amino acid substitutions in nsp14-P203L might affect viral properties such as genome mutation proofreading and genome replication efficiency.

Alanine substitution of an amino acid that is important for the proofreading activity of the nsp14 of SARS-CoV has been shown to increase viral genomic mutations in cultured cells[5,11] and *in vivo*.[10] In addition, Graham et al. reported that SARS-CoV is also attenuated by such a substitution.[10] The effects of amino acid substitutions in nsp14 on coronavirus infections in nature have not been investigated. In our study, candidates that might affect nsp14 function were sought by focusing on amino acid substitutions in the nsp14 of epidemic viruses and comparing genomic mutation rates.

Limitations of the study

The molecular mechanism of how amino acid mutations in nsp14 are involved in proofreading activity is not clear. In addition, the number of SARS-CoV-2 genomes, including nsp14-203L variants, was limited in our study, which limited our statistical analysis. Because the number of SARS-CoV-2 genomes deposited in the GISAID has increased (14,535,971 genomes, as of January 11, 2023), further molecular evolutionary analyses with phylogenetic correction may reveal more nsp14 mutations that affect its proofreading activity.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY

  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- ● EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Animal experiments and approvals
  - ○ Cells
  - ○ Reverse genetics
  - ○ Deep sequencing analysis
  - ○ Experimental infection of Syrian hamsters
- ● METHOD DETAILS
  - ○ Coronavirus sequence data
  - ○ Molecular evolutionary phylogenetic analysis
  - ○ Substitution rate analysis
  - ○ Measurement of natural selection pressure
  - ○ *In silico* structural analysis
- ● QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106210.

ACKNOWLEDGMENTS

We thank Dr. Susan Watson for scientific editing. We acknowledge the NGS Core laboratory, Genome Information Research Center, Institute for Microbial Diseases, Osaka University, for assistance with the RNA sequence. We thank the GISAID EpiCoV database (https://www.gisaid.org), which shares sequence data of SARS-CoV-2 and related viral species, and all researchers who deposited SARS-CoV-2 genomes and related information in GISAID (https://epicov.org/epi3/epi_set/230110sz). This study was supported by JSPS KAKENHI Grants-in-Aid for Scientific Research on Innovative Areas JP16H06429 (to Y.M., T.W., and S.N.), JP16K21723 (to Y.M., T.W., and S.N.), JP16K21723 (to Y.M.), JP16H06432 (to Y.M.), JP16H06434 (to T.W), and JP19H04843 (to S.N.); Grants-in-Aid for Scientific Research (B) JP22H02521 (to T.W.) and JP21H02736 (to C.O); Grant-in-Aid for Early-Career Scientists JP22K15469 (to K.T.); Grant-in-Aid for JSPS Fellows JP21J01036 (to K.T.); AMED Research Program on Emerging and Re-emerging Infectious Diseases JP20fk0108281 (to Y.M.), JP19fk0108113 (to Y.M. and T.W.), JP20fk0108401 (to C.O.), JP21fk0108493 (to C.O.), and JP19fk0108171 (to S.N.); AMED Cyclic Innovation for Clinical Empowerment (CiCLE) JP20pc0101047 (to Y. M.); AMED under Grant Number JP223fa627002 (to T.W.); AMED Advanced Research and Development Programs for Medical Innovation (AMED-CREST) JP22gm1610010 (to T.W.); 2020 Tokai University School of Medicine Research Aid (to S.N.); Japan Science and Technology Agency (JST) CREST JPMJCR20H6 (to S.N.); JST Moonshot R&D JPMJMS2025 (to Y.M.); the Tokyo Biochemical Research Foundation (to T.W.); the Takeda Science Foundation (to T.W.); and RIKAKEN HOLDINGS CO. Young Researcher Support Grant-in-aid (to K.T.). The super-computing resources were provided by the Human Genome Center (the Univ. of Tokyo) and the NIG supercomputer at ROIS National Institute of Genetics.

AUTHOR CONTRIBUTIONS

K.T., T.W., and S.N. conceived the study. K.T., M.T.U., and S.N analyzed data. C.O. and Y.M. established 293-hACE2/hTMPRSS2 cells. K.T. generated recombinant viruses. K.T. and Y.K. performed *in vitro* experiments. K.T., S.S., and T.W. performed animal experiments. K.T, M.T.U., T.W., and S.N. wrote the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors have no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: November 18, 2022
Revised: January 23, 2023
Accepted: February 13, 2023
Published: February 16, 2023

REFERENCES

1. Sjaarda, C.P., Guthrie, J.L., Mubareka, S., Simpson, J.T., Hamelin, B., Wong, H., Mortimer, L., Slinger, R., McArthur, A.G., Desjardins, M., et al. (2021). Temporal dynamics and evolution of SARS-CoV-2 demonstrate the necessity of ongoing viral genome sequencing in Ontario, Canada. mSphere *6*, e00011-21. https://doi.org/10.1128/mSphere.00011-21.

2. Gorbalenya, A.E., Enjuanes, L., Ziebuhr, J., and Snijder, E.J. (2006). Nidovirales: evolving the largest RNA virus genome. Virus Res. *117*, 17–37. https://doi.org/10.1016/j.virusres.2006.01.017.

3. Robson, F., Khan, K.S., Le, T.K., Paris, C., Demirbag, S., Barfuss, P., Rocchi, P., and Ng, W.L. (2020). Coronavirus RNA proofreading: molecular basis and therapeutic targeting. Mol. Cell *79*, 710–727. https://doi.org/10.1016/j.molcel.2020.11.048.

4. Eckerle, L.D., Lu, X., Sperry, S.M., Choi, L., and Denison, M.R. (2007). High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. J. Virol. *81*, 12135–12144. https://doi.org/10.1128/JVI.01296-07.

5. Eckerle, L.D., Becker, M.M., Halpin, R.A., Li, K., Venter, E., Lu, X., Scherbakova, S., Graham, R.L., Baric, R.S., Stockwell, T.B., et al. (2010). Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. PLoS Pathog. *6*, e1000896. https://doi.org/10.1371/journal.ppat.1000896.

6. Sola, I., Almazán, F., Zúñiga, S., and Enjuanes, L. (2015). Continuous and discontinuous RNA Synthesis in coronaviruses. Annu. Rev. Virol. *2*, 265–288. https://doi.org/10.1146/annurev-virology-100114-055218.

7. te Velthuis, A.J.W., van den Worm, S.H.E., and Snijder, E.J. (2012). The SARS-coronavirus nsp7+nsp8 complex is a unique multimeric RNA polymerase capable of both de novo initiation and primer extension. Nucleic Acids Res. *40*, 1737–1747. https://doi.org/10.1093/nar/gkr893.

8. Ahn, D.G., Choi, J.K., Taylor, D.R., and Oh, J.W. (2012). Biochemical characterization of a recombinant SARS coronavirus nsp12 RNA-dependent RNA polymerase capable of copying viral RNA templates. Arch. Virol. *157*, 2095–2104. https://doi.org/10.1007/s00705-012-1404-x.

9. Snijder, E.J., Bredenbeek, P.J., Dobbe, J.C., Thiel, V., Ziebuhr, J., Poon, L.L.M., Guan, Y., Rozanov, M., Spaan, W.J.M., and Gorbalenya, A.E. (2003). Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. J. Mol. Biol. *331*, 991–1004. https://doi.org/10.1016/s0022-2836(03)00865-9.

10. Graham, R.L., Becker, M.M., Eckerle, L.D., Bolles, M., Denison, M.R., and Baric, R.S. (2012). A live, impaired-fidelity coronavirus vaccine protects in an aged, immunocompromised mouse model of lethal disease. Nat. Med. *18*, 1820–1826. https://doi.org/10.1038/nm.2972.

11. Smith, E.C., Blanc, H., Surdel, M.C., Vignuzzi, M., and Denison, M.R. (2013). Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. PLoS Pathog. *9*, e1003565. https://doi.org/10.1371/journal.ppat.1003565.

12. Shu, Y., and McCauley, J. (2017). GISAID: global initiative on sharing all influenza data - from vision to reality. Euro Surveill. *22*, 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

13. Nakagawa, S., and Miyazawa, T. (2020). Genome evolution of SARS-CoV-2 and its virological characteristics. Inflamm. Regen. *40*, 17. https://doi.org/10.1186/s41232-020-00126-7.

14. Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190. https://doi.org/10.1101/gr.849004.

15. Shi, M., Lin, X.D., Chen, X., Tian, J.H., Chen, L.J., Li, K., Wang, W., Eden, J.S., Shen, J.J., Liu, L., et al. (2018). The evolutionary history of vertebrate RNA viruses. Nature *556*, 197–202. https://doi.org/10.1038/s41586-018-0012-7.

16. Kosakovsky Pond, S.L., and Frost, S.D.W. (2005). Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. *22*, 1208–1222. https://doi.org/10.1093/molbev/msi105.

17. Ogando, N.S., Ferron, F., Decroly, E., Canard, B., Posthuma, C.C., and Snijder, E.J. (2019). The curious case of the nidovirus exoribonuclease: its role in RNA Synthesis and replication fidelity. Front. Microbiol. *10*, 1813. https://doi.org/10.3389/fmicb.2019.01813.

18. Romano, M., Ruggiero, A., Squeglia, F., Maga, G., and Berisio, R. (2020). A structural view of SARS-CoV-2 RNA replication machinery: RNA Synthesis, proofreading and final capping. Cells *9*, 1267. https://doi.org/10.3390/cells9051267.

19. Munaweera, R., and Hu, Y.S. (2020). Computational characterizations of the interactions between the pontacyl violet 6R and exoribonuclease as a potential drug target against SARS-CoV-2. Front. Chem. *8*, 627340. https://doi.org/10.3389/fchem.2020.627340.

20. Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., and Pybus, O.G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. *5*, 1403–1407. https://doi.org/10.1038/s41564-020-0770-5.

21. Graepel, K.W., Lu, X., Case, J.B., Sexton, N.R., Smith, E.C., and Denison, M.R. (2017). Proofreading-deficient coronaviruses adapt for increased fitness over long-term passage without reversion of exoribonuclease-inactivating mutations. mBio *8*, e01503-17. https://doi.org/10.1128/mBio.01503-17.

22. Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B., et al. (2020). Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell *182*, 812–827.e19. https://doi.org/10.1016/j.cell.2020.06.043.

23. Hou, Y.J., Chiba, S., Halfmann, P., Ehre, C., Kuroda, M., Dinnon, K.H., 3rd, Leist, S.R., Schäfer, A., Nakajima, N., Takahashi, K., et al. (2020). SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. Science *370*, 1464–1468. https://doi.org/10.1126/science.abe8499.

24. Yan, L., Yang, Y., Li, M., Zhang, Y., Zheng, L., Ge, J., Huang, Y.C., Liu, Z., Wang, T., Gao, S., et al. (2021). Coupling of N7-methyltransferase and 3′-5′ exoribonuclease with SARS-CoV-2 polymerase reveals mechanisms for capping and proofreading. Cell *184*, 3474–3485.e11. https://doi.org/10.1016/j.cell.2021.05.033.

25. Bouvet, M., Imbert, I., Subissi, L., Gluais, L., Canard, B., and Decroly, E. (2012). RNA 3′-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex. Proc. Natl. Acad. Sci. USA *109*, 9372–9377. https://doi.org/10.1073/pnas.1201130109.

26. Torii, S., Ono, C., Suzuki, R., Morioka, Y., Anzai, I., Fauzyah, Y., Maeda, Y., Kamitani, W., Fukuhara, T., and Matsuura, Y. (2021). Establishment of a reverse genetics system for SARS-CoV-2 using circular polymerase extension reaction. Cell Rep. *35*, 109014. https://doi.org/10.1016/j.celrep.2021.109014.

27. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780. https://doi.org/10.1093/molbev/mst010.

28. Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics *27*, 1164–1165. https://doi.org/10.1093/bioinformatics/btr088.

29. Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., Stamatakis, A., and Wren, J. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics *35*, 4453–4455. https://doi.org/10.1093/bioinformatics/btz305.

30. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

31. Darriba, D., Posada, D., Kozlov, A.M., Stamatakis, A., Morel, B., and Flouri, T. (2020). ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. Mol. Biol. Evol. *37*, 291–294. https://doi.org/10.1093/molbev/msz189.

32. Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). Mega X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. *35*, 1547–1549. https://doi.org/10.1093/molbev/msy096.

33. Abascal, F., Zardoya, R., and Telford, M.J. (2010). TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. *38*, W7–W13. https://doi.org/10.1093/nar/gkq291.

34. Weaver, S., Shank, S.D., Spielman, S.J., Li, M., Muse, S.V., and Kosakovsky Pond, S.L. (2018). Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. Mol. Biol. Evol. *35*, 773–777. https://doi.org/10.1093/molbev/msx335.

35. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. j. *17*, 10. https://doi.org/10.14806/ej.17.1.200.

36. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

37. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

38. Dehouck, Y., Kwasigroch, J.M., Rooman, M., and Gilis, D. (2013). BeAtMuSiC: prediction of changes in protein-protein binding affinity on mutations. Nucleic Acids Res. *41*, W333–W339. https://doi.org/10.1093/nar/gkt450.

39. Lanave, C., Preparata, G., Saccone, C., and Serio, G. (1984). A new method for calculating evolutionary substitution rates. J. Mol. Evol. *20*, 86–93. https://doi.org/10.1007/BF02101990.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Recombinant SARS-CoV-2	This study	N/A
Recombinant SARS-CoV-2 (nsp14-P203L)	This study	N/A
Escherichia coli DH5-a	New England Biolabs	Cat#C2987
Chemicals, peptides, and recombinant proteins		
TransIT-LT1 Transfection Reagent	Takara	Cat# MIR2300
Fetal bovine serum	Gibco	Cat# 42Q7361K
Penicillin-streptomycin	Nacalai	Cat# 09367-34
DMEM (high glucose)	Nacalai Tesque	Cat# 08459-64
DMEM high glucose powder	Sigma	Cat# D5648
HEPES(1M)	Gibco	Cat# 15630080
Amphotericin B solution (250 ug/mL)	Sigma	Cat# A2942
Gentamicin Sulfate Solution (50mg/ml)	nacalai	Cat# 11980-14
7.5% NaHCO3	Sigma	Cat# S8761-100ML
SeaPlaque Agarose	Lonza	Cat# 50100
SeaKem Agarose	Lonza	Cat# 50074
G418	Nacalai Tesque	Cat# 09380-44
Critical commercial assays		
QIAamp viral RNA mini kit	Qiagen	Cat# 52906
PrimeSTAR GXL DNA polymerase	Takara	Cat# R050A
ProtoScript II Reverse Transcriptase	New England Biolabs	Cat# M0368S
NEBNext® Ultra™ II Non-Directional RNA Second Strand Synthesis Module	New England Biolabs	Cat# E6111S/L
SureSelectXT HS and XT Low input enzymatic fragmentation kit	Agilent	Cat# 5191-4080
NovaSeq 6000 S4 Reagent Kit v1.5 (200 cycles)	Illumina	Cat# 20028313
PrimeScript 1st strand cDNA Synthesis Kit	Takara	Cat# 6110A
Custom probe	Agilent	https://www.agilent.com/cs/library/applications/application-sars-cov-2-sequencing-5994-4538en-agilent.pdf
Deposited data		
Viral genome sequencing data by RNA-seq	DDBJ Sequence Read Archive	Accession number: DRR411740 - DRR411753
Experimental models: Cell lines		
Human: HEK293-hACE2/hTMPRSS2 cells	This study	N/A
African green monkey (*Chlorocebus sabaeus*): VeroE6/TMPRSS2 cells	JCRB Cell Bank	JCRB1819
Oligonucleotides		
Linker/F1-F: CTATATAAGCAGAGCTCGTTTAGTGAACCGTattaaaggtttataccttcccaggtaac	Torii et al.[26]	N/A
F1/F2-R: cagattcaacttgcatggcattgttagtagccttatttaaggctcctgc	Torii et al.[26]	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
F1/F2-F: gcaggagccttaaataaggctacta acaatgccatgcaagttgaatctg	Torii et al.[26]	N/A
F2/F3-R: ggtaggattttccactacttcttcaga gactggttttagatcttcgcaggc	Torii et al.[26]	N/A
F2/F3-F: gcctgcgaagatctaaaaccagtctct gaagaagtagtggaaaatcctacc	Torii et al.[26]	N/A
F3/F4-R: ggtgcacagcgca gcttcttcaaaagtactaaagg	Torii et al.[26]	N/A
F3/F4-F: caccactaattcaacctattggtgctttg gacatatcagcatctatagtagctggtgg	Torii et al.[26]	N/A
F4/F5-R: gtttaaaaacgattgtgcatcagctgactg	Torii et al.[26]	N/A
F4/F5-F: cacagtctgtaccgtctgcggtatgtgga aaggttatggctgtagttgtgatc	Torii et al.[26]	N/A
F5/F6-R: gcggtgtgtacatagcctcataaaactca ggttcccaataccttgaagtg	Torii et al.[26]	N/A
F5/F6-F: cacttcaaggtattgggaacctgagttttatg aggctatgtacacaccgc	Torii et al.[26]	N/A
F6/F7-R: catacaaactgccaccatcacaaccagg caagttaaggttagatagcactctag	Torii et al.[26]	N/A
F6/F7-F: ctagagtgctatctaaccttaacttgcctggt tgtgatggtggcagtttgtatg	Torii et al.[26]	N/A
F7/F8-R: ctagagactagtggcaataaaacaagaaaa acaaacattgttcgtttagttgttaac	Torii et al.[26]	N/A
F7/F8-F: gttaacaactaaacgaacaatgtttgtttttcttg ttttattgccactagtctctag	Torii et al.[26]	N/A
F8/F9-R: gcagcaggatccacaagaacaacagccctt gagacaactacagcaactgg	Torii et al.[26]	N/A
F8/F9-F: ccagttgctgtagttgtctcaagggctgttgtt cttgtggatcctgctgc	Torii et al.[26]	N/A
F9/Linker-R: GGAGATGCCATGCCGACCC tttttttttttttttttttttttttgtcattctcctaag	Torii et al.[26]	N/A
F9/Linker-F: cttaggagaatgacaaaaaaaaaaaaaaa aaaaaaaaaaGGGTCGGCATGGCATCTCC	Torii et al.[26]	N/A
Linker/F1-R: gttacctgggaaggtataaacctttaatAC GGTTCACTAAACGAGCTCTGCTTATATAG	Torii et al.[26]	N/A
Recombinant DNA		
Plasmid: pCSII-sars-cov-2 F1	Torii et al.[26]	N/A
Plasmid: pCSII-sars-cov-2 F2	Torii et al.[26]	N/A
Plasmid: pCSII-sars-cov-2 F3	Torii et al.[26]	N/A
Plasmid: pCSII-sars-cov-2 F4	Torii et al.[26]	N/A
Plasmid: pcDNA3.1.-sars-cov-2 F5	Torii et al.[26]	N/A
Plasmid: pcDNA3.1.-sars-cov-2 F6	Torii et al.[26]	N/A
Plasmid: pcDNA3.1.-sars-cov-2 F6 (nsp14-P203L)	This study	N/A
Plasmid: pCSII-sars-cov-2 F7	Torii et al.[26]	N/A
Plasmid: pCSII-sars-cov-2 F8	Torii et al.[26]	N/A
Plasmid: pCSII-sars-cov-2 F9	Torii et al.[26]	N/A
Plasmid: pMW118 CoV2-CMVlinker	Torii et al.[26]	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
MAFFT version v7.453	Katoh and Standley[27]	https://mafft.cbrc.jp/alignment/software
WebLogo 3	Crooks et al.[14]	https://weblogo.threeplusone.com
ProtTest3 version 3.4.2	Darriba et al.[28]	https://github.com/ddarriba/prottest3
RAxML-NG version 1.0.0	Kozlov et al.[29]	https://github.com/amkozlov/raxml-ng
CD-HIT-EST version 4.8.1	Li et al.[30]	https://sites.google.com/view/cd-hit
ModelTest-NG	Darriba et al.[31]	https://github.com/ddarriba/modeltest
MEGA X	Kumar et al.[32]	https://www.megasoftware.net
TranslatorX	Abascal et al.[33]	http://translatorx.co.uk
Datamonkey software	Weaver et al.[34]	https://www.datamonkey.org
cutadapt version 3.2	Martin[35]	https://cutadapt.readthedocs.io/en/v3.2/
BWA version 0.7,17-r1188	Li and Durbin[36]	http://bio-bwa.sourceforge.net
MuTect2 in GATK suite version 4.2	McKenna et al.[37]	https://github.com/broadinstitute/gatk/releases
PyMOL molecular graphics system v2.4.0	The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC	https://pymol.org/
BeAtMuSiC program version 1.0	Dehouck et al.[38]	http://babylone.3bio.ulb.ac.be/beatmusic/
GraphPad Prism software version 9.4.0	GraphPad Software	https://www.graphpad.com/scientific-software/prism/
FigTree v1.4.4	http://tree.bio.ed.ac.uk/software/figtree/	http://tree.bio.ed.ac.uk/software/figtree/
R v4.2.1	The R Foundation	https://www.r-project.org/
Python v3.7	Python Software Foundation	https://www.python.org
Other		
GISAID EpiCoV database	Shu et al.[12]	https://epicov.org/epi3/epi_set/230110sz

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to and will be fulfilled by the lead contact, So Nakagawa (so@tokai.ac.jp).

Materials availability

All unique reagents generated in this study are listed in the key resources table and available from the lead contact with a completed Materials Transfer Agreement.

Data and code availability

- RNA-seq data have been deposited at DDBJ Sequence Read Archive (DRA) with the following accession numbers: DRR411740 - DRR411753. The GISAID accession numbers used in this study are listed in STAR Methods.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Animal experiments and approvals

All animal experiments were approved by the Animal Research Committee of the Research Institute for Microbial Diseases, Osaka University (assurance number, R03-04-0). All animals were housed under specific pathogen-free conditions in a temperature and humidity control environment with a 12 h:12 h light dark cycle and *ad libitum* access to water and standard laboratory chow. Virus inoculations were performed under anesthesia, and all efforts were made to minimize animal suffering. These *in vivo* studies were not blinded, and animals were randomly assigned to infection groups. No sample-size calculations were performed to power each study. Instead, sample size was determined based on prior *in vivo* virus challenge experiments.

Cells

VeroE6/TMPRSS2 (JCRB 1819) cells were maintained in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% FBS, 1% Penicillin-Streptomycin Solution, and 1 mg/ml geneticin (G418; Nacalai Tesque, Cat# 09380-44). HEK293-hACE2/hTMPRSS2 cells were maintained in DMEM supplemented with 10% FBS and 1% Penicillin-Streptomycin Solution. VeroE6/TMPRSS2 and HEK293-hACE2/hTMPRSS2 cells were maintained at 37°C with 5% CO_2.

Reverse genetics

Recombinant SARS-CoV-2 was generated by using circular polymerase extension reaction (CPER) as previously described.[26] Briefly, nine DNA fragments encoding the partial genome of SARS-CoV-2 (hCoV-19/Japan/TY-WK-521/2020) were amplified by PCR using PrimeSTAR GXL DNA polymerase (Takara, Cat# R050A). A linker fragment encoding the hepatitis delta virus ribozyme, the bovine growth hormone poly A signal, and the cytomegalovirus promoter was also amplified by PCR. The ten DNA fragments were mixed and used for CPER.[26]

To produce recombinant SARS-CoV-2 (seed virus), the CPER products were transfected into HEK293-hACE2/hTMPRSS2 cells by using TransIT-LT1 (Takara, Cat# MIR2305). At one day post-transfection, the culture medium was replaced with DMEM (high glucose) (Nacalai Tesque, Cat# 08459-64) containing 2% FBS and 1% Penicillin-Streptomycin Solution. At 4–10 days post-transfection, the culture medium was collected and stock at −80°C. The viruses were used to inoculate VeroE6/TMPRSS2 cells to produce stock viruses. No mutations were found in the wild-type viruses, whereas the nsp14-P203L viruses were found to contain mixed unexpected mutations such as C12119T/C (nsp8-P10S/P) and C17143T/C (nsp13-R303C/R). To obtain viruses with the single mutation of nsp14-P203L, a plaque-purified virus clone was grown in VeroE6/TMPRSS2 cells. All experiments with transfectants generated by reverse genetics were performed in an enhanced biosafety level 3 (BSL3) containment laboratory approved for such used by Osaka university.

Deep sequencing analysis

Viral RNA was extracted from samples by using the QIAamp Viral RNA Mini kit (Qiagen) according to the manufacturer's instructions. The extracted RNA was reverse transcribed with ProtoScript II (NEB) using N6 primers. The cDNA was converted to double-stranded DNA by NEBNext Second Strand Synthesis (NEB). In addition, double-stranded DNA was fragmented by the SureSelect Fragmentation Enzyme. Fragmented DNA was made into an NGS library with the SureSelect XT Low Input Kit. In addition, hybridization was performed using a SARS-CoV-2 custom probe, and sequence-ready libraries were sequenced in a paired-end run using by MiSeq (Illumina).

The raw FASTQ format data files that were obtained from the sequencing samples were filtered using cutadapt version 3.2[35] to remove low-quality reads and adapters. Then, the filtered sequences were mapped into the reference SARS-CoV-2 genome (Wuhan-Hu-1, NC_045512.2) by using BWA version 0.7,17-r1188.[36] Variant calling was done using MuTect2 in GATK suite version 4.2,[37] and nucleotide changes were counted by using in-house python scripts.

Experimental infection of Syrian hamsters

Seven-week-old male wild-type Syrian hamsters (Japan SLC Inc., Shizuoka, Japan) were used in this study. Baseline body weights were measured before infection. Under isoflurane anesthesia, four hamsters per group were intranasally inoculated with 10^3 PFU in 30 μl of recombinant virus. Body weight was monitored daily for 6 days after infection. For virological examinations, four hamster per group were intranasally infected with recombinant viruses; 3 and 6 dpi, the animals were euthanized, and nasal turbinate and lungs were collected. The virus titers in the nasal turbinate, trachea and lungs were determined by performing plaque assays on VeroE6/TMPRSS2 cells.

METHOD DETAILS

Coronavirus sequence data

The amino acid sequences of open reading frame 1ab (ORF1ab) of the 62 representative coronaviruses (summarized in Table S1) were obtained from NCBI (https://www.ncbi.nlm.nih.gov/) and the GISAID EpiCoV database (https://www.gisaid.org) as reported previously.[13] Aligned nucleotide sequences of SARS-CoV-2 and their metadata were obtained from the GISAID EpiCoV database as of September 7, 2020 (87,625 sequences). The coronavirus genome sequences annotated as having "Human" hosts were extracted, and sequences containing undetermined and/or mixed nucleotides were removed. The 28,082 remaining genomes were used for the analyses. The open reading frame (ORF) of each sequence was determined from the alignment sequence based on the ORF information of hCoV-19/Wuhan/Hu-1/2019 (GISAID ID: EPI_ISL_402125, GenBank ID: NC_045512.2), which was used as the reference sequence. Gap sites were removed, and the nucleotide sequences were translated. To detect amino acid substitutions for each site, the amino acid sequence after the stop codon was removed and amino acid sequences were aligned by using the L-INS-i program in MAFFT version v7.453.[27] From the sequence alignment of the nsp14 of the representative coronaviruses, the amino acids aligned at the amino acid positions of SARS-CoV-2 are shown using WebLogo 3.[14] The color of each amino acids is indicated according to its chemical properties within the default option. Note that, in the alignment sequence, gap sites were removed based on the SARS-CoV-2 nsp14 sequence.

Molecular evolutionary phylogenetic analysis

To compute the phylogenetic tree of the nsp14 of the 62 representative coronaviruses, their nsp14 amino acid sequences were aligned by using the L-INS-i program in MAFFT version 7.453.[27] To estimate the appropriate amino acid substitution model, we used ProtTest3 version 3.4.2[28] based on the Bayesian information criterion (BIC) values. We then computed a phylogenetic tree by using RAxML-NG version 1.0.0[29] with rapid 1000 bootstrap replicates.[39]

We also generated a phylogenetic tree of SARS-CoV-2 genomes belonging to the PANGO lineage B.1.1.33 including nsp14-P203L variants. Using CD-HIT-EST version 4.8.1,[30] identical genomic sequences were removed: from 200 to 142 sequences. The 142 nucleotide sequences were aligned by using the L-INS-i program in MAFFT version 7.453.[27] To estimate the appropriate nucleotide mutation model, we used ModelTest-NG[31] based on the Akaike information criterion (AIC) values, and GTR+I+G4 was selected. We then computed a phylogenetic tree by using RAxML-NG version 1.0.0[29] with rapid 1000 bootstrap replicates.[39]

Substitution rate analysis

A multiple sequence alignment of SARS-CoV-2 genomes obtained from the GISAID database was used to analyze the genomic evolutionary rate. The number of nucleotide mutations per site from between sequences was counted. Analyses were conducted using the Kimura 2-parameter model. Codon positions included 1st+2nd+3rd+Noncoding. All gaps were removed for each sequence pair with a pairwise deletion option. The number of nucleotide mutations was counted in MEGA X.[32] The hCoV-19/Wuhan/Hu-1/2019 (GISAID ID: EPI_ISL_402125, GenBank ID: NC_045512.2) was selected as the reference sequence. Sequences with information that included the collection date were used for the analysis. Note that we omitted sequences that were collected before December 26, 2019, the date on which the reference sequences were collected. The approximate straight line was calculated by using the least-squares method and in-house python scripts.

Measurement of natural selection pressure

TranslatorX[33] was used to align the nucleotide sequences based on the aligned amino acid sequences of the nsp14 of the representative coronaviruses. To determine the evolutionary selection pressures on the nsp14 gene in coronaviruses, we computed the ratio of nonsynonymous substitution rates (dN) and synonymous substitution rates (dS) per site (dN/dS) by using Datamonkey software[16,34] through MEGA X.[32]

In silico structural analysis

A PDB file (7EGQ) of the SARS-CoV-2 replication-transcription complex (RTC) was obtained from the PDB database (https://www.rcsb.org). Using the nsp14-nsp10 complex (chains H and K) of the structure, further analyses were performed. The residues at the interface of the SARS-CoV-2 nsp10-nsp14 docked complex were determined using the "InterfaceResidues" PyMol script (http://www.pymolwiki.org/index.php/InterfaceResidues/). We used the SARS-CoV-2 nsp10-nsp14 docked complex to predict the effect of the P203L mutation on protein-protein interactions by using the BeAtMuSiC program version 1.0 (http://babylone.3bio.ulb.ac.be/beatmusic/query.php).[38] P203L mutant generation, hydrophobicity calculation, and visualization were performed using PyMol version 2.4.0.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis

All data were analyzed in GraphPad Prism software (version 9.4.0). Statistical analysis included ANOVA with multiple corrections, post-tests, or an unpaired Student's t test. Differences among groups were considered significant for *P* values < 0.05.