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]*.

[1]*Department of Molecular Virology, Research Institute for Microbial Diseases, Osaka University, Suita, Osaka, Japan.*

[2]*Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Kanagawa, Japan.*

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

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

[5]*Center for Infectious Disease Education and Research, Osaka University, Suita, Osaka, Japan.*

[6]*Center for Advanced Modalities and DDS, Osaka University, Suita, Osaka, Japan.*

[7]*Bioinformation and DDBJ Center, National Institute of Genetics, Mishima, Shizuoka, Japan.*

*Corresponding author

E-mail: tokikow@biken.osaka-u.ac.jp (Tokiko Watanabe) and so@tokai.ac.jp (So Nakagawa)

ORCID: 0000-0001-6461-5379 (Tokiko Watanabe) and 0000-0003-1760-3839 (So Nakagawa)

ABSTRACT

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.

Keywords

coronavirus, mutation, non-structural protein 14, SARS-CoV-2

36 **Introduction**

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

52 Coronaviruses possess a positive-sense, single-stranded RNA genome that is one of the
53 largest among known RNA viruses, ranging from 26 to 32 kb (2). There are two large open reading
54 frames (ORFs) at the 5′ proximal end, ORF1a and ORF1b, which comprise approximately two-thirds
55 of the genome, and encode two large polyproteins, pp1a and pp1ab, that are cleaved into about 16
56 nonstructural proteins (nsps). Some of the processed nsps assemble to form the replication-
57 transcription complex, which contains an RdRp (nsp12), the helicase (nsp13), processivity factors
58 (nsp7 and nsp8), and the proofreading ExoN complex (nsp14 and nsp10) (6-8). During genome
59 replication, nsp14 ExoN activity plays a role in the RNA proofreading machinery by efficiently
60 removing mismatched 3'-endnucleotides (9). Experiments in which the nsp14 of SARS-CoV was

61 functionally inactivated revealed 10–20 times higher mutation rates (5, 10, 11), indicating that nsp14
62 may be essential for maintaining the virus genome.

63 In this study, to clarify whether amino acid substitutions in nsp14 affect the genomic diversity
64 and evolution of SARS-CoV-2, we searched for amino acid substitutions in nature that may interfere
65 with the function of nsp14 in SARS-CoV-2. First, to obtain functionally important sites in nsp14, we
66 examined 62 representative coronaviruses belonging to the family *Coronaviridae* and found that 99 of
67 the 527 amino acid sites of nsp14 were evolutionarily conserved. We then examined nsp14 sequences
68 obtained from 28,082 SARS-CoV-2 genomes available in the GISAID EpiCoV database, which
69 includes not only sequencing data but also sampling date, country, and institution information (12),
70 and identified an additional 6 amino acid changes in nsp14 mutants that were not detected in the 62
71 representative coronaviruses. We examined the genome mutation rates of these mutants and found that
72 the amino acid replacement of proline (P) for leucine (L) at position 203 of nsp14 could be involved
73 in increasing the nucleotide mutation rate of SARS-CoV-2 genomes. To investigate the biological
74 significance of the nsp14 P203L substitution, we generated recombinant SARS-CoV-2 viruses
75 possessing wild-type nsp14 or mutant nsp14 with P203L and characterized them in a hamster model.
76 We found that the nsp14-P203L mutant grown in the lungs of hamsters acquired significantly more
77 diverse genomic mutations than did the wild-type virus. These results indicate that substitutions in
78 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 Supplementary Table 1). The phylogenetic tree clearly showed the relationship among the four genera of coronaviruses (Fig. 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 (Fig. 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 (Fig. 1B). If we included an unclassified coronavirus found in *Tropidophorus sinicus* (Chinese waterside skink) (14), 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 nonsynonymous (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 Materials and Methods) (15). 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

104 nsp14 of 61 coronaviruses (excluding the Chinese waterside skink coronavirus that contains a mixture
105 of nucleotides at these positions). Thus, 504 amino acid sites of the 61 coronavirus nsp14 multiple
106 alignment were calculated for the dN/dS ratios; no positive selection sites were found in nsp14,
107 whereas 379 of the 504 codon sites were under negative selection ($P < 0.01$, Supplementary Fig. 1).
108 In total, 91 amino acids were unchanged and under negative selection in the nsp14 of the 62
109 representative coronaviruses.

110 We then investigated whether the amino acids that are known to be important for the error-
111 correcting function of nsp14 are conserved. Coronavirus nsp14 possesses both ExoN and N7-MTase
112 activities. The ExoN domain of coronavirus nsp14 was originally identified based on its sequence
113 similarity to distant cell homologs (9) and has been assigned to the DEDD exonuclease superfamily,
114 which includes the proofreading domains of many DNA polymerases and other eukaryotic and
115 prokaryotic exonucleases. The nsp14 proteins have four conserved active site residues distributed in
116 three standard motifs of the primary structure: residues 90D/92E (motif I), 191E (motif II), and
117 268H/273D (motif III) (16-18). We found that the five active sites of nsp14 were conserved among
118 the 62 coronaviruses examined in this study (Fig. 1C). Moreover, there were three zinc fingers that
119 were essential for nsp14 function: zinc finger 1 (ZF1) comprising 207C, 210C, 226C, and 229H; zinc
120 finger 2 (ZF2) comprising 257H, 261C, 264H, and 279C; and zinc finger 3 (ZF3) comprising 452C,
121 473C, 484C, and 487H. S-adenosyl methionine (SAM)-binding motif I includes residues 331D, 333G,
122 335P, and 337A (or 337G), which are involved in the (N7 guanine)-methyl transferase reaction by
123 nsp14. We found that all these residues were conserved among the nsp14 sequences of the 62
124 coronaviruses examined (Supplementary Fig. 2). These results suggest that the amino acids of nsp14
125 that are functionally important are conserved among the diversified coronaviruses.
126

127 **Diversity of nsp14 in SARS-CoV-2**

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

134 We downloaded the aligned nucleotide sequences of SARS-CoV-2 (87,625 sequences) from
135 the GISAID database as of September 7, 2020. The SARS-CoV-2 genomes isolated from humans were
136 extracted (87,304 sequences), and the sequences containing undetermined and/or mixed nucleotides
137 were removed; the remaining 28,082 sequences were used in subsequent analyses. The coding region
138 of nsp14 was translated and compared. The results showed that 25,913 nsp14 amino acid sequences
139 (92.27%) were identical to that of the reference sequence, whereas 2169 sequences contained at least
140 one amino acid substitution in nsp14 (Supplementary Table 2). The top 20 variants possessing nsp14
141 with amino acids that were different from the major sequence (shown in Supplementary Fig. 2) are
142 summarized in Table 1. Note that frameshift mutants were excluded from the table because they may
143 affect downstream genes as well. Notably, among the 20 amino acid nsp14 mutants, six amino acids
144 (177F, 377L, 369F, 501I, 203L and 297S) were not found in the 62 nsp14 proteins of the representative
145 coronaviruses (Supplementary Fig. 2). We further analyzed the evolutionary selection pressures of
146 these 6 amino acid sites by using 73 coronaviruses belonging to the genus *Betacoronaviruses* that
147 includes SARS-CoV-2 (Supplementary Table 3). We confirmed that the six amino acid variants were
148 not found in any of the 73 Beta-coronaviruses (Supplementary Table 3). The six conserved positions
149 (177, 203, 297, 369, 377, and 501) were also found to be under negative selection ($P < 0.01$,
150 Supplementary Table 4). Overall, these results indicate that the conserved amino acid sites of nsp14
151 are also evolutionarily selected in the genus *Betacoronaviruses*.
152

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

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

165 We found that SARS-CoV-2 containing wild-type nsp14 had 19.8 nucleotide mutations in its
166 whole genome per year (Table 2; Supplementary Fig. 3 for the regression coefficient). Conversely,
167 seven SARS-CoV-2 variants containing nsp14 with amino acid substitutions showed 27.3 (L177F),
168 59.8 [L177F/T372I (*i.e.*, L177F and T372I)], 35.9 (P203L), 17.1 (P297S), 26.2 (S369F), 13.0 (F377L),
169 or 21.9 (M501I) nucleotide mutations per year (Table 2; Supplementary Fig. 3). Although the
170 substitution rates of the L177F/T372I and F377L variants were found to be unreliable because of
171 limited time ranges, those of the other variants were correlated (Supplementary Fig. 3). Therefore,
172 SARS-CoV-2 variants possessing nsp14 with the L177F, P203L, S369F, or M501I substitution may
173 have a higher genomic mutation rate than that of SARS-CoV-2 carrying wild-type nsp14. In particular,
174 the P203L variant of nsp14 (nsp14-P203L), which showed the highest mutation rate (35.7 nucleotide
175 mutations per year), is a strong candidate that enhances the diversity of SARS-CoV-2 genomes.

176 We then verified the substitution rates of the nsp14-P203L variants for each single cluster in
177 which the maximum number was observed: 6 genomes for the nsp14-203L mutant and 194 genomes

178 for the nsp14-203P wildtype in the B.1.1.33 lineage (Fig. 2A; Supplementary Fig. 4 with GISAID IDs).
179 We found that nsp14-P203L variants in the B.1.1.33 lineage forming a single clade showed a high
180 nucleotide evolutionary rate (58.8 nucleotide mutations/year) compared with those with the nsp14-
181 203P (*i.e.*, wildtype) strains in the same lineage (16.5 nucleotide mutations/year) (Fig. 2B and Table
182 3). Although the number of variants was limited (6), a correlation was clearly observed (Fig. 2B). In
183 addition, six other nsp-P203L variants in a different cluster (B.1) also showed a relatively high
184 nucleotide mutation rate (28.0 nucleotide mutations/year), assuming that variants were in the same
185 clade if their PANGO IDs (19) and GISAID clades were identical (Table 3; Supplementary Fig. 5 for
186 regression coefficients). With the same assumption, we did the same calculation for the other nsp14
187 mutants in the lineages (Table 3; Supplementary Fig. 5) and found that nsp14-P203L in the B.1.1.33
188 lineage exhibited the highest substitution rate. These results indicate that SARS-CoV-2 variants
189 containing nsp14-P203L are strong candidates for variants that readily accumulate mutations in the
190 viral population.

191 The frequency of nucleotide mutations of coronaviruses may be affected not only by the nsp14
192 proofreading mechanism but also by the characteristics of other proteins, such as the fidelity of the
193 RNA-dependent RNA polymerase (nonstructural protein 12; nsp12) (20). It is possible that amino acid
194 substitutions in other proteins involved in genome replication affect the nucleotide mutation rates of
195 SARS-CoV-2. Therefore, we examined amino acid substitutions among nsp14-203L variants
196 involving the following genes that participate in gene replication: nsp7, nsp8, nsp9, nsp10, nsp12,
197 nsp13, nsp15, and nsp16. The results showed that a proline-to-leucine substitution at position 323
198 (P323L) of nsp12 frequently occurred (20/25) in sequences of SARS-CoV-2 possessing nsp14-203L
199 (Supplementary Table 5). However, the P323L mutation in nsp12 (nsp12-P323L) is known to
200 accompany an aspartate-to-glycine substitution at position 614 (D614G) of S protein (21, 22). The S-
201 D614G mutant was first reported at the end of January 2020 in China and Germany, and by March
202 2020, this mutant was a major variant in various regions all over the world (21). The P323L mutation

203 in nsp12 was found in nsp14-203L variants as well. Indeed, the genomic mutation rate of SARS-CoV-
204 2 with the P323L amino acid substitution in nsp12, was found to be 17.7 nucleotide mutations per year
205 (Supplementary Fig. 6). Therefore, it is unlikely that P323L in nsp12 directly affects the genomic
206 mutation rate of SARS-CoV-2 viruses carrying the P203L amino acid substitution in nsp14.
207

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

209 The nsp14-P203L variants have been isolated from 26 patients mainly in Europe and North
210 America, with the first being reported on March 5, 2020 in The Netherlands (GISAID ID:
211 EPI_ISL_454756, Supplementary Table 6). Considering their PANGO lineages (19) and GISAID
212 clades (shown in Supplementary Table 6), the nsp14-P203L variants appeared at least 10 times. In fact,
213 the nsp14-P203L substitution was observed in both Spike D614 strains and G614 variants, which have
214 enhanced infection efficiency and have spread worldwide (21). However, we also found that the
215 prevalence of nsp14-P203L variants did not increase steadily over time; the largest number of nsp14-
216 P203L variants found in the B.1.1.33 and B.1 lineages was only six for each (Supplementary Table 6).
217

218 **The nsp14-P203L mutation effect on protein structure**

219 To assess the effect of the P203L mutation of nps14, we first investigated the location of
220 residue 203 by using a tertiary structure model of SARS-CoV-2 nsp14 (PDB ID: 7EGQ) (23). The
221 residue 203 is exposed on the surface of the protein and is in the loop stretching from residues 200 to
222 214, which is in the interstices of the zinc and magnesium binding sites (positions 207 and 191). We
223 found that the loop is located at the edge of an interaction surface for binding to the co-factor SARS-
224 CoV-2 nsp10 (Fig. 3B and Supplementary Fig. 7A). The interaction with nsp10 has been shown to
225 greatly increase the nucleolytic activity of SARS-CoV nsp14 (24). Therefore, to investigate the
226 possibility that residues at or around position 203 play an important role in the nsp10-nsp14 interaction,
227 we calculated the change in binding free energy (ΔΔGbind) induced by nsp14-P203L by using the

228 BeQtMuSicC program. The predicted result showed that this substitution decreases binding affinity
229 (ΔΔGbind 0.43 kcal/mol, Supplementary Fig. 7B), suggesting that the residue at position 203 is
230 important to establish the interaction interface for binding to nsp10.
231
232 **Characterization of a recombinant SARS-CoV-2 virus possessing nsp14-P203L *in vitro*.**
233 Genome data analysis suggested that SARS-CoV-2 variants containing nsp14-P203L could
234 accumulate mutations in the viral population. Does the nsp14-P203L amino acid substitution affect
235 viral properties (e.g., increasing the diversity of the viral genome)? To attempt to answer this question,
236 we generated recombinant SARS-CoV-2 (based on hCoV-19/Japan/TY-WK-521/2020) viruses
237 possessing wild-type nsp14 or P203L nsp14 by using a recently established circular polymerase
238 extension reaction (CPER) method (25). We then we examined the effect of the nsp14-P203L amino
239 acid substitution on the biological properties of SARS-CoV-2.
240 To determine whether the nsp14-P203L amino acid substitution affects the viral growth
241 efficiency in cultured cells, we examined viral replication efficiency in VeroE6/TMPRSS2 cells. We
242 found that both wild-type virus and nsp14-P203L viruses grew well, with max titers (mean titers =
243 7.35 log unit and 7.01 log unit, respectively) at 36 h post-infection (Fig. 4A), although the titer of the
244 wild-type virus at 24 h post-infection was slightly higher than that of the nsp14-P203L virus (Fig. 4A).
245 Next, we analyzed the genomic diversity of the recombinant virus grown in the cultured cells.
246 Three plaque-purified wild-type or nsp14-P203L virus clones were used to infect VeroE6/TMPRSS2
247 cells, and the genomes of the viruses in the supernatant were collected 48 h later and analyzed by RNA
248 sequencing. Mutations with a read depth of at least 1000 times and an allele frequency of at least 0.5%
249 were compared between the wild-type and nsp14-P203L mutant viruses. The nsp14-P203L mutant
250 grown in VeroE6/TMPRSS2 had 51.33 (±10.96) mutations and the wild-type virus had 46.33 (±9.46)
251 mutations, which was not significantly different (P = 0.65) (Fig. 4B). These results suggest that nsp14-
252 P203L allows efficient virus replication and may not affect viral genome mutagenesis in culture cells.

253

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

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

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

275 **Discussion**

276 The mutation rate of coronaviruses, including SARS-CoV-2, is relatively low compared to
277 other RNA viruses because coronaviruses possess an error-correcting exonuclease, nsp14. In this study,
278 we examined mutations that occurred in nsp14, and found that SARS-CoV-2 possessing nsp14 with a
279 proline to leucine substitution at position 203 showed a relatively higher substitution rate than SARS-
280 CoV-2 possessing wild-type nsp14 (Tables 2 and 3; Fig. 2). Protein structure prediction analysis of
281 SARS-CoV-2 nsp14 suggested that the P203L replacement in nsp14 decreases the binding affinity of
282 nsp10, possibly reducing the proofreading activity (Fig. 3; Supplementary Fig. 7). Moreover, we
283 demonstrated that a recombinant SARS-CoV-2 possessing this P203L substitution in nsp14 acquired
284 significantly more diverse genomic mutations than the wild-type virus did during its replication in a
285 hamster model, suggesting that mutations in nsp14, such as P203L, could accelerate the genomic
286 diversity of SARS-CoV-2 (Figure 4).

287 The nsp14-P203L mutant virus grown in VeroE6/TMPRSS2 cells did not have significantly
288 diverse genomic mutations compared to the wild-type virus, whereas the nsp14-P203L mutant virus
289 grown in hamster lungs had significantly diverse mutations in its genome compared to the wild-type
290 virus. One reason for this difference may be the number of genome copies of each virus in the
291 experiments. Viruses grown in VeroE6/TMPRSS2 cells were sampled to examine genomic mutations
292 at 2 days post-infection, whereas viruses grown in hamster lungs were assessed at 6 days post-infection.
293 Therefore, the viruses may have replicated more in the hamsters' lungs due to the longer incubation
294 time. In addition, in VeroE6/TMPRSS2 cells, the nsp14-P203L mutant virus showed slightly reduced
295 growth efficiency compared to the wild-type virus (Fig. 4A). Therefore, the wild-type virus may
296 replicate more frequently in VeroE6/TMPRSS2 cells compared to the nsp14-P203L mutant virus,
297 resulting accumulation of genomic mutations. In contrast, in hamster lungs, the replication efficiency
298 of the nsp14-P203L mutant virus was similar to that of the wild-type virus (Fig. 4D). Therefore, since
299 the genome replication frequency is almost the same between the wild-type and the nsp14-P203L

300 mutant viruses, the reduced proofreading activity due to nsp14-P203L can be correctly detected.
301 Overall, our findings suggest that amino acid substitutions in nsp14-P203L might affect viral
302 properties such as genome mutation proofreading and genome replication efficiency.

303 Alanine substitution of an amino acid that is important for the proofreading activity of the
304 nsp14 of SARS-CoV has been shown to increase viral genomic mutations in cultured cells (5, 11) and
305 *in vivo* (10). In addition, Graham et al. reported that SARS-CoV is also attenuated by such a
306 substitution (10). The effects of amino acid substitutions in nsp14 on coronavirus infections in nature
307 have not been investigated. In our study, candidates that might affect nsp14 function were sought by
308 focusing on amino acid substitutions in the nsp14 of epidemic viruses and comparing genomic
309 mutation rates. However, the molecular mechanism of how amino acid mutations in nsp14 are involved
310 in proofreading activity is not clear. In addition, the number of SARS-CoV-2 genomes, including
311 nsp14-203L variants, was limited in our study, which limited our statistical analysis. Since the number
312 of SARS-CoV-2 genomes deposited in the GISAID has increased (14,535,971 genomes, as of January
313 11, 2023), further molecular evolutionary analyses with phylogenetic correction may reveal more
314 nsp14 mutations that affect its proofreading activity. Nevertheless, our findings demonstrate that rare
315 amino acid substitutions (particularly leucine at amino acid 203) in coronavirus nsp14 may increase
316 the rate of genomic mutations.

317 **References**

318 1. Sjaarda CP, Guthrie JL, Mubareka S, Simpson JT, Hamelin B, Wong H, et al. Temporal
319 Dynamics and Evolution of SARS-CoV-2 Demonstrate the Necessity of Ongoing Viral Genome
320 Sequencing in Ontario, Canada. mSphere. 2021;6(3).
321 2. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ. Nidovirales: evolving the largest RNA
322 virus genome. Virus Res. 2006;117(1):17-37.
323 3. Robson F, Khan KS, Le TK, Paris C, Demirbag S, Barfuss P, et al. Coronavirus RNA
324 Proofreading: Molecular Basis and Therapeutic Targeting. Mol Cell. 2020;79(5):710-27.
325 4. Eckerle LD, Lu X, Sperry SM, Choi L, Denison MR. High fidelity of murine hepatitis virus
326 replication is decreased in nsp14 exoribonuclease mutants. J Virol. 2007;81(22):12135-44.
327 5. Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X, et al. Infidelity of SARS-CoV
328 Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. PLoS
329 Pathog. 2010;6(5):e1000896.
330 6. Sola I, Almazan F, Zuniga S, Enjuanes L. Continuous and Discontinuous RNA Synthesis in
331 Coronaviruses. Annu Rev Virol. 2015;2(1):265-88.
332 7. te Velthuis AJ, van den Worm SH, Snijder EJ. The SARS-coronavirus nsp7+nsp8 complex
333 is a unique multimeric RNA polymerase capable of both de novo initiation and primer extension.
334 Nucleic Acids Res. 2012;40(4):1737-47.
335 8. Ahn DG, Choi JK, Taylor DR, Oh JW. Biochemical characterization of a recombinant
336 SARS coronavirus nsp12 RNA-dependent RNA polymerase capable of copying viral RNA
337 templates. Arch Virol. 2012;157(11):2095-104.
338 9. Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LLM, et al. Unique and
339 Conserved Features of Genome and Proteome of SARS-coronavirus, an Early Split-off From the
340 Coronavirus Group 2 Lineage. Journal of Molecular Biology. 2003;331(5):991-1004.
341 10. Graham RL, Becker MM, Eckerle LD, Bolles M, Denison MR, Baric RS. A live, impaired-
342 fidelity coronavirus vaccine protects in an aged, immunocompromised mouse model of lethal
343 disease. Nat Med. 2012;18(12):1820-6.
344 11. Smith EC, Blanc H, Surdel MC, Vignuzzi M, Denison MR. Coronaviruses lacking
345 exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and
346 potential therapeutics. PLoS Pathog. 2013;9(8):e1003565.
347 12. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to
348 reality. Euro Surveill. 2017;22(13).
349 13. Nakagawa S, Miyazawa T. Genome evolution of SARS-CoV-2 and its virological
350 characteristics. Inflamm Regen. 2020;40:17.
351 14. Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, et al. The evolutionary history of
352 vertebrate RNA viruses. Nature. 2018;556(7700):197-202.

15. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005;22(5):1208-22.
16. Ogando NS, Ferron F, Decroly E, Canard B, Posthuma CC, Snijder EJ. The Curious Case of the Nidovirus Exoribonuclease: Its Role in RNA Synthesis and Replication Fidelity. Front Microbiol. 2019;10:1813.
17. Romano M, Ruggiero A, Squeglia F, Maga G, Berisio R. A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping. Cells. 2020;9(5).
18. Munaweera R, Hu YS. Computational Characterizations of the Interactions Between the Pontacyl Violet 6R and Exoribonuclease as a Potential Drug Target Against SARS-CoV-2. Front Chem. 2020;8:627340.
19. Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020;5(11):1403-7.
20. Graepel KW, Lu X, Case JB, Sexton NR, Smith EC, Denison MR. Proofreading-Deficient Coronaviruses Adapt for Increased Fitness over Long-Term Passage without Reversion of Exoribonuclease-Inactivating Mutations. mBio. 2017;8(6).
21. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 2020;182(4):812-27 e19.
22. Hou YJ, Chiba S, Halfmann P, Ehre C, Kuroda M, Dinnon KH, 3rd, et al. SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. Science. 2020;370(6523):1464-8.
23. Yan L, Yang Y, Li M, Zhang Y, Zheng L, Ge J, et al. Coupling of N7-methyltransferase and 3'-5' exoribonuclease with SARS-CoV-2 polymerase reveals mechanisms for capping and proofreading. Cell. 2021;184(13):3474-85 e11.
24. Bouvet M, Imbert I, Subissi L, Gluais L, Canard B, Decroly E. RNA 3'-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex. Proc Natl Acad Sci U S A. 2012;109(24):9372-7.
25. Torii S, Ono C, Suzuki R, Morioka Y, Anzai I, Fauzyah Y, et al. Establishment of a reverse genetics system for SARS-CoV-2 using circular polymerase extension reaction. Cell Rep. 2021;35(3):109014.
26. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772-80.
27. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188-90.
28. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011;27(8):1164-5.

390 29. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A, Wren J. RAxML-NG: a fast,
391 scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics.
392 2019;35(21):4453-5.
393 30. Lanave C, Preparata G, Saccone C, Serio G. A new method for calculating evolutionary
394 substitution rates. J Mol Evol. 1984;20(1):86-93.
395 31. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or
396 nucleotide sequences. Bioinformatics. 2006;22(13):1658-9.
397 32. Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. ModelTest-NG: A
398 New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. Mol Biol Evol.
399 2020;37(1):291-4.
400 33. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary
401 Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018;35(6):1547-9.
402 34. Abascal F, Zardoya R, Telford MJ. TranslatorX: multiple alignment of nucleotide sequences
403 guided by amino acid translations. Nucleic Acids Res. 2010;38(Web Server issue):W7-13.
404 35. Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. Datamonkey
405 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. Mol
406 Biol Evol. 2018;35(3):773-7.
407 36. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
408 EMBnetjournal. 2011;17(1):3.
409 37. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
410 Bioinformatics. 2009;25(14):1754-60.
411 38. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The
412 Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing
413 data. Genome Res. 2010;20(9):1297-303.
414 39. Dehouck Y, Kwasigroch JM, Rooman M, Gilis D. BeAtMuSiC: Prediction of changes in
415 protein-protein binding affinity on mutations. Nucleic Acids Res. 2013;41(Web Server issue):W333-
416 9.
417

418 **Methods**

419 **Coronavirus sequence data**

420 The amino acid sequences of open reading frame 1ab (ORF1ab) of the 62 representative
421 coronaviruses (summarized in Supplementary Table 1) were obtained from NCBI
422 (https://www.ncbi.nlm.nih.gov/) and the GISAID EpiCoV database (https://www.gisaid.org) as
423 reported previously (13). Aligned nucleotide sequences of SARS-CoV-2 and their metadata were
424 obtained from the GISAID EpiCoV database as of September 7, 2020 (87,625 sequences). The
425 coronavirus genome sequences annotated as having "Human" hosts were extracted, and sequences
426 containing undetermined and/or mixed nucleotides were removed. The 28,082 remaining genomes
427 were used for the analyses. The open reading frame (ORF) of each sequence was determined from the
428 alignment sequence based on the ORF information of hCoV-19/Wuhan/Hu-1/2019 (GISAID ID:
429 EPI_ISL_402125, GenBank ID: NC_045512.2), which was used as the reference sequence. Gap sites
430 were removed, and the nucleotide sequences were translated. To detect amino acid substitutions for
431 each site, the amino acid sequence after the stop codon was removed and amino acid sequences were
432 aligned by using the L-INS-i program in MAFFT version v7.453 (26). From the sequence alignment
433 of the nsp14 of the representative coronaviruses, the amino acids aligned at the amino acid positions
434 of SARS-CoV-2 are shown using WebLogo 3 (27). The color of each amino acids is indicated
435 according to its chemical properties within the default option. Note that, in the alignment sequence,
436 gap sites were removed based on the SARS-CoV-2 nsp14 sequence.

437

438 **Molecular evolutionary phylogenetic analysis**

439 To compute the phylogenetic tree of the nsp14 of the 62 representative coronaviruses, their
440 nsp14 amino acid sequences were aligned by using the L-INS-i program in MAFFT version 7.453 (26).
441 To estimate the appropriate amino acid substitution model, we used ProtTest3 version 3.4.2 (28) based

442 on the Bayesian information criterion (BIC) values. We then computed a phylogenetic tree by using
443 RAxML-NG version 1.0.0 (29) with rapid 1000 bootstrap replicates (30).
444 We also generated a phylogenetic tree of SARS-CoV-2 genomes belonging to the PANGO
445 lineage B.1.1.33 including nsp14-P203L variants. Using CD-HIT-EST version 4.8.1 (31), identical
446 genomic sequences were removed: from 200 to 142 sequences. The 142 nucleotide sequences were
447 aligned by using the L-INS-i program in MAFFT version 7.453 (26). To estimate the appropriate
448 nucleotide mutation model, we used ModelTest-NG (32) based on the Akaike information criterion
449 (AIC) values, and GTR+I+G4 was selected. We then computed a phylogenetic tree by using RAxML-
450 NG version 1.0.0 (29) with rapid 1000 bootstrap replicates (30).
451
452 **Substitution rate analysis**
453 A multiple sequence alignment of SARS-CoV-2 genomes obtained from the GISAID
454 database was used to analyze the genomic evolutionary rate. The number of nucleotide mutations per
455 site from between sequences was counted. Analyses were conducted using the Kimura 2-parameter
456 model. Codon positions included 1st+2nd+3rd+Noncoding. All gaps were removed for each sequence
457 pair with a pairwise deletion option. The number of nucleotide mutations was counted in MEGA X
458 (33). The hCoV-19/Wuhan/Hu-1/2019 (GISAID ID: EPI_ISL_402125, GenBank ID: NC_045512.2)
459 was selected as the reference sequence. Sequences with information that included the collection date
460 were used for the analysis. Note that we omitted sequences that were collected before December 26,
461 2019, the date on which the reference sequences were collected. The approximate straight line was
462 calculated by using the least-squares method and in-house python scripts.
463
464 **Measurement of natural selection pressure**
465 TranslatorX (34) was used to align the nucleotide sequences based on the aligned amino acid
466 sequences of the nsp14 of the representative coronaviruses. To determine the evolutionary selection

467 pressures on the nsp14 gene in coronaviruses, we computed the ratio of nonsynonymous substitution
468 rates (dN) and synonymous substitution rates (dS) per site (dN/dS) by using Datamonkey software (15,
469 35) through MEGA X (33).
470
471 **Cells**
472 VeroE6/TMPRSS2 (JCRB 1819) cells were maintained in Dulbecco's Modified Eagle
473 Medium (DMEM) supplemented with 10% FBS, 1% Penicillin-Streptomycin Solution, and 1 mg/ml
474 geneticin (G418; Nacalai Tesque, Cat# 09380-44). HEK293-hACE2/hTMPRSS2 cells were
475 maintained in DMEM supplemented with 10% FBS and 1% Penicillin-Streptomycin Solution.
476 VeroE6/TMPRSS2 and HEK293-hACE2/hTMPRSS2 cells were maintained at 37°C with 5% CO_2.
477
478 **Reverse genetics**
479 Recombinant SARS-CoV-2 was generated by using circular polymerase extension reaction
480 (CPER) as previously described (25). Briefly, nine DNA fragments encoding the partial genome of
481 SARS-CoV-2 (hCoV-19/Japan/TY-WK-521/2020) were amplified by PCR using PrimeSTAR GXL
482 DNA polymerase (Takara, Cat# R050A). A linker fragment encoding the hepatitis delta virus
483 ribozyme, the bovine growth hormone poly A signal, and the cytomegalovirus promoter was also
484 amplified by PCR. The ten DNA fragments were mixed and used for CPER (25).
485 To produce recombinant SARS-CoV-2 (seed virus), the CPER products were transfected into
486 HEK293-hACE2/hTMPRSS2 cells by using TransIT-LT1 (Takara, Cat# MIR2305). At one day post-
487 transfection, the culture medium was replaced with DMEM (high glucose) (Nacalai Tesque, Cat#
488 08459-64) containing 2% FBS and 1% Penicillin-Streptomycin Solution. At 4–10 days post-
489 transfection, the culture medium was collected and stock at -80°C. The viruses were used to inoculate
490 VeroE6/TMPRSS2 cells to produce stock viruses. No mutations were found in the wild-type viruses,
491 whereas the nsp14-P203L viruses were found to contain mixed unexpected mutations such as

492 C12119T/C (nsp8-P10S/P) and C17143T/C (nsp13-R303C/R). To obtain viruses with the single
493 mutation of nsp14-P203L, a plaque-purified virus clone was grown in VeroE6/TMPRSS2 cells. All
494 experiments with transfectants generated by reverse genetics were performed in an enhanced biosafety
495 level 3 (BSL3) containment laboratory approved for such used by Osaka university.
496
497 **Deep sequencing analysis**
498 Viral RNA was extracted from samples by using the QIAamp Viral RNA Mini kit (Qiagen)
499 according to the manufacturer's instructions. The extracted RNA was reverse transcribed with
500 ProtoScript II (NEB) using N6 primers. The cDNA was converted to double-stranded DNA by
501 NEBNext Second Strand Synthesis (NEB). In addition, double-stranded DNA was fragmented by the
502 SureSelect Fragmentation Enzyme. Fragmented DNA was made into an NGS library with the
503 SureSelect XT Low Input Kit. In addition, hybridization was performed using a SARS-CoV-2 custom
504 probe, and sequence-ready libraries were sequenced in a paired-end run using by MiSeq (Illumina).
505 The raw FASTQ format data files that were obtained from the sequencing samples were
506 filtered using cutadapt version 3.2 (36) to remove low-quality reads and adapters. Then, the filtered
507 sequences were mapped into the reference SARS-CoV-2 genome (Wuhan-Hu-1, NC_045512.2) by
508 using BWA version 0.7,17-r1188(37). Variant calling was done using MuTect2 in GATK suite version
509 4.2 (38), and nucleotide changes were counted by using in-house python scripts.
510
511 **Animal experiments and approvals**
512 All animal experiments were approved by the Animal Research Committee of the Research
513 Institute for Microbial Diseases, Osaka University (assurance number, R03-04-0). All animals were
514 housed under specific pathogen-free conditions in a temperature and humidity control environment
515 with a 12 h:12 h light dark cycle and ad libitum access to water and standard laboratory chow. Virus
516 inoculations were performed under anesthesia, and all efforts were made to minimize animal suffering.

517 These *in vivo* studies were not blinded, and animals were randomly assigned to infection groups. No
518 sample-size calculations were performed to power each study. Instead, sample size was determined
519 based on prior *in vivo* virus challenge experiments.
520
521 **Experimental infection of Syrian hamsters**
522 Seven-week-old male wild-type Syrian hamsters (Japan SLC Inc., Shizuoka, Japan) were used
523 in this study. Baseline body weights were measured before infection. Under isoflurane anesthesia, four
524 hamsters per group were intranasally inoculated with 10^3 PFU in 30 µl of recombinant virus. Body
525 weight was monitored daily for 6 days after infection. For virological examinations, four hamster per
526 group were intranasally infected with recombinant viruses; 3 and 6 dpi, the animals were euthanized,
527 and nasal turbinate and lungs were collected. The virus titers in the nasal turbinate, trachea and lungs
528 were determined by performing plaque assays on VeroE6/TMPRSS2 cells.
529
530 ***In silico* structural analysis**
531 A PDB file (7EGQ) of the SARS-CoV-2 replication-transcription complex (RTC) was
532 obtained from the PDB database (https://www.rcsb.org). Using the nsp14-nsp10 complex (chains H
533 and K) of the structure, further analyses were performed. The residues at the interface of the SARS-
534 CoV-2 nsp10-nsp14 docked complex were determined using the "InterfaceResidues" PyMol script
535 (http://www.pymolwiki.org/index.php/InterfaceResidues/). We used the SARS-CoV-2 nsp10-nsp14
536 docked complex to predict the effect of the P203L mutation on protein-protein interactions by using
537 the BeAtMuSiC program version 1.0 (http://babylone.3bio.ulb.ac.be/beatmusic/query.php) (39).
538 P203L mutant generation, hydrophobicity calculation, and visualization were performed using PyMol
539 version 2.4.0.
540
541 **Statistical analysis**

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

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 KAKENHI Grants-in-Aid for Scientific Research on Innovative Areas 16H06429 (to Y.M., T.W., and S.N.), 16K21723 (to Y.M., T.W., and S.N.), JP16K21723 (to Y.M.), JP16H06432 (to Y.M.), 16H06434 (to T.W), and 19H04843 (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 22K15469 (to K.T.); Grant-in-Aid for JSPS Fellows 21J01036 (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 19fk0108171 (to S.N.); AMED Cyclic Innovation for Clinical Empowerment (CiCLE) JP20pc0101047 (to Y. M.); AMED under Grant Number JP223fa627002h (to T.W.); AMED Advanced Research and Development Programs for Medical Innovation (AMED-CREST) 22gm1610010h0001 (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 information

570 **Affiliations**

571 Department of Molecular Virology, Research Institute for Microbial Diseases, Osaka University,
572 Osaka, Japan.
573 Kosuke Takada, Shintaro Shichinohe, Yurie Kida and Tokiko Watanabe.
574
575 Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Kanagawa,
576 Japan.
577 Kosuke Takada and So Nakagawa.
578
579 Department of Genomic Function and Diversity, Medical Research Institute, Tokyo Medical and
580 Dental University, Bunkyo, Tokyo, Japan
581 Mahoko Takahashi Ueda.
582
583 Laboratory of Virus Control, Research Institute for Microbial Diseases, Osaka University, Suita,
584 Osaka, Japan
585 Chikako Ono and Yoshiharu Matsuura
586
587 Center for Infectious Disease Education and Research, Osaka University, Suita, Osaka, Japan.
588 Chikako Ono, Yoshiharu Matsuura and Tokiko Watanabe.
589
590 Center for Advanced Modalities and DDS, Osaka University, Osaka, Japan.
591 Tokiko Watanabe.
592

593 **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.

Competing interests

The authors have no competing interests.

Data Availability Statements

The data underlying this article will be shared upon reasonable request to the corresponding author. All RNA-seq data sequenced in this study were deposited in DDBJ Sequence Read Archive (DRA) with the following accession numbers: DRR411740 - DRR411753

Corresponding authors

Correspondence to Tokiko Watanabe <tokikow@biken.osaka-u.ac.jp> and So Nakagawa <so@tokai.ac.jp>

610 **Figure legends**

611

612 **Fig. 1: Phylogenetic tree and amino acid sequence alignment of nsp14 derived from**
613 **representative coronaviruses**

614 (A) Maximum likelihood (ML)-based phylogenetic tree of the 62 representative coronaviruses.
615 Amino acid sequences of nsp14 in ORF1ab were used for this analysis. A red or blue circle
616 in an internal node corresponds to bootstrap values ≥ 95% or ≥ 80 %, respectively. The virus
617 located at the top is SARS-CoV-2 (hCoV-19/Wuhan/Hu-1/2019). (B) Conserved amino acid
618 residues among each group. The amino acid sequence of the nsp14 of the 62 representative
619 coronaviruses is shown according to the relative amino acid positions of SARS-CoV-2 nsp14
620 (note that the gaps in the SARS-CoV-2 sequence are not shown). The conserved amino acid
621 positions are indicated in gray. (C) Five conserved active site residues involved in
622 proofreading. The amino acid proportions for each site were calculated based on the
623 sequence alignment of the nsp14 of the 62 representative coronaviruses visualized by using
624 WebLogo (27). The amino acids are colored according to their chemical properties.

625

626 **Fig. 2: Phylogeny and nucleotide mutation rate of nsp14-P203L variants in the PANGO**
627 **lineage B.1.1.33**

628 (A) ML tree of SARS-CoV-2 genomes in the B.1.1.33-lineage. The nsp14-P203L variant is
629 highlighted in red. An orange circle corresponds to bootstrap values ≥ 70%. (B) Nucleotide
630 mutation rates of nsp14-203P or nsp14-203L variants belonging to the B.1.1.33-lineage. All
631 sample dots are displayed. Mutation rates per year in the genome are shown in red letters.
632 The red and orange dotted lines correspond to the regression line and 95% confidence
633 intervals, respectively.

634

635 **Fig. 3: Three-dimensional position of the amino acid at position 203 in nsp14**

636 (A) Domain structure of nsp14 with residue numbers. The ExoN and N7-Mtase domains are
637 shown in light blue and gray, respectively. The amino acid at position 203 in nsp14 is shown
638 in magenta. (B) Schematic representation of nsp14. The coloring is the same as in (A). The
639 site of interaction with nsp10 is shown in dark blue.

640

641 **Fig. 4: Characterization of the recombinant nsp14-P203L virus *in vitro* and *in vivo***

642 (A and B) Growth efficiency and genomic mutations of nsp14-P203L variants in cultured cells.
643 (A) Growth kinetics of recombinant viruses in VeroE6/TMPRSS2 cells. VeroE6/TMPRSS2
644 cells were infected with viruses at a multiplicity of infection of 0.001. The supernatants of the
645 infected cells were harvested at the indicated times, and virus titers were determined by the
646 using plaque assays in VeroE6/TMPRSS2 cells. Data are shown as the mean of three
647 independent experiments done in triplicate. Error bars indicate the standard deviation. Mean
648 values were compared by using a two-way ANOVA, followed by Dunnett's test (*P<0.05).
649 Asterisks indicate statistically significant differences between WT and Mut viruses. (B)
650 Number of genomic mutation types detected in viruses grown in VeroE6/TMPRSS2 cells.
651 Three wild-type or mutant viruses were each plaque-purified, and each viral clone was
652 infected into VeroE6/TMPRSS2 cells; supernatants were collected after 48 hours. RNA was
653 extracted from the virus in the supernatant and RNA sequencing was performed. Mutations
654 with a read depth of at least 1000x and an allele frequency of at least 0.5% were compared
655 between the wild-type and mutant viruses. Mean values were compared by using an unpaired
656 Student's t test (N.D.; P>0.05). (C, D, E, F and G) Growth efficiency and genomic mutations
657 of nsp14-P203L variants in Syrian hamsters. (C) Schematic overview of the animal
658 experiment. Syrian hamsters were inoculated with 103 PFU of wild-type or nsp14-P203L virus
659 via the intranasal (30 μL) route. (D) Virus replication in infected Syrian hamsters. Four Syrian
660 hamsters per group were killed on days 3 and 6 post-infection for virus titration. Virus titers
661 in lung or nasal turbinate (NT) organs were determined by using plaque assays in
662 VeroE6/TMPRSS2 cells. Error bars indicate the standard deviation. Mean values were
663 compared by using a two-way ANOVA, followed by Dunnett's test. (E) Body weight changes
664 in Syrian hamsters after viral infection. Body weights of virus-infected (n = 8) animals were
665 monitored daily for 3 days. Body weights of virus-infected (n = 4) animals were monitored
666 daily for 6 days. Data are presented as the mean percentages of the starting weight (±SD).
667 (F) Number of genomic mutation types detected in viruses grown in Syrian hamster lung.
668 RNA was extracted from the virus in lung homogenates from the Day 6 post-infection Syrian
669 hamsters, and RNA sequencing was performed. Mutations with a read depth of at least 1000x
670 and an allele frequency of at least 0.5% were compared between the wild-type and mutant
671 viruses. Mean values compared by using an unpaired Student's t test (**P<0.01). (G) SARS-
672 CoV-2 genome organization and genomic locations of synonymous or nonsynonymous
673 substitutions detected in the genomes of viruses grown in hamster lungs. The different
674 directions of the triangles indicate each of the four individual hamsters. The number of

675 nucleotide changes in viruses grown in Syrian hamster lung. The number of nucleotide
676 changes, detected as synonymous and nonsynonymous substitutions, at a read depth of at
677 least 1000x and an allele frequency of >0.5% were compared between the wild-type and
678 nsp14-P203L mutant viruses.

679 **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	–	25913	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

680 28,082 SARS-CoV-2 genome sequences that were annotated as Human and free of undetermined or mixed
681 nucleotides were used in the analysis. Amino acid substitutions were defined as a change relative to the hCoV-
682 19/Wuhan/Hu-1/2019 reference sequence. Boldface characters indicate that the amino acid replacement was not
683 observed in the 62 coronaviruses examined in this analysis. —, no mutation was found compared with the
684 reference sequence.

685 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	–	25471	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

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

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

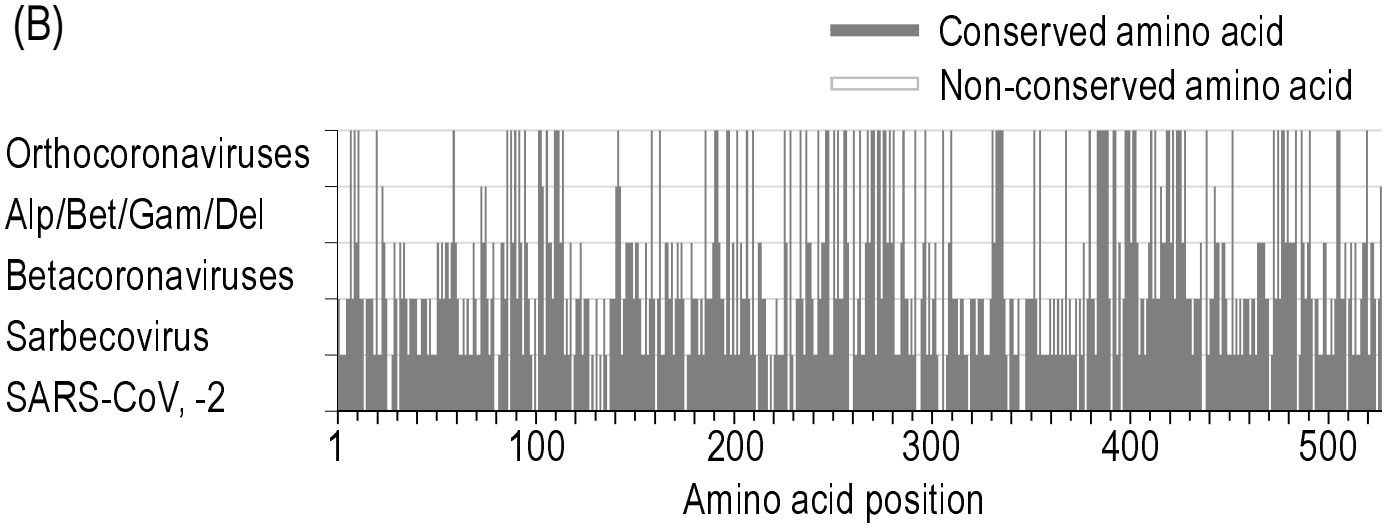
Ranking of nsp14	Substitution in nsp14	Lineage (PANGO ID)	Number of sequences containing the nsp14 substitution	Substitution rate per year in the whole genome
5	L177F	B.1.1.25	54	7.8
7	F377L	B	50	8.5
11	S369F	B.1.1	17	13.4
13	M501I	B.1	11	4.4
16	P203L	B.1.1.33	6	58.8
		B.1	6	28.0
17	L177F/T372I	B.1.113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

692 The number of mismatched nucleotides in the alignment of the reference sequence with the comparison
693 sequence of each group was counted, and the average value was calculated. hCoV-19/Wuhan/Hu-1/2019 was
694 selected as the reference sequence. A scatter plot of these nsp14 is shown in Fig 2B and Supplementary Fig 5.

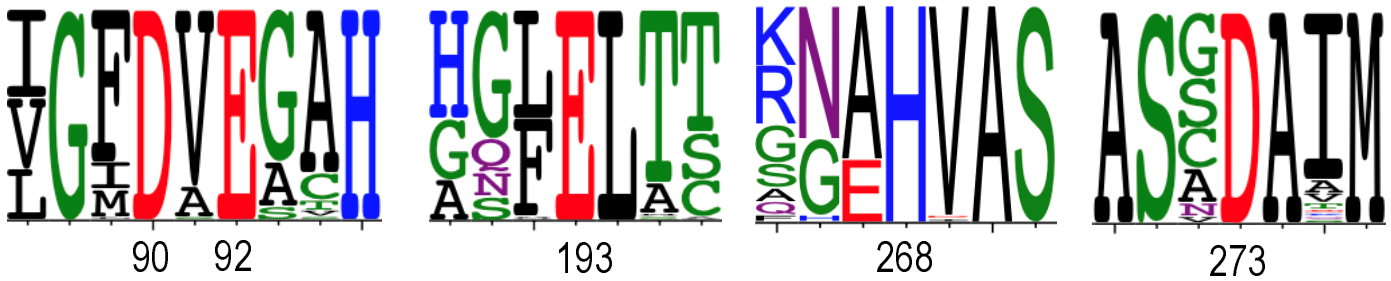
(A)

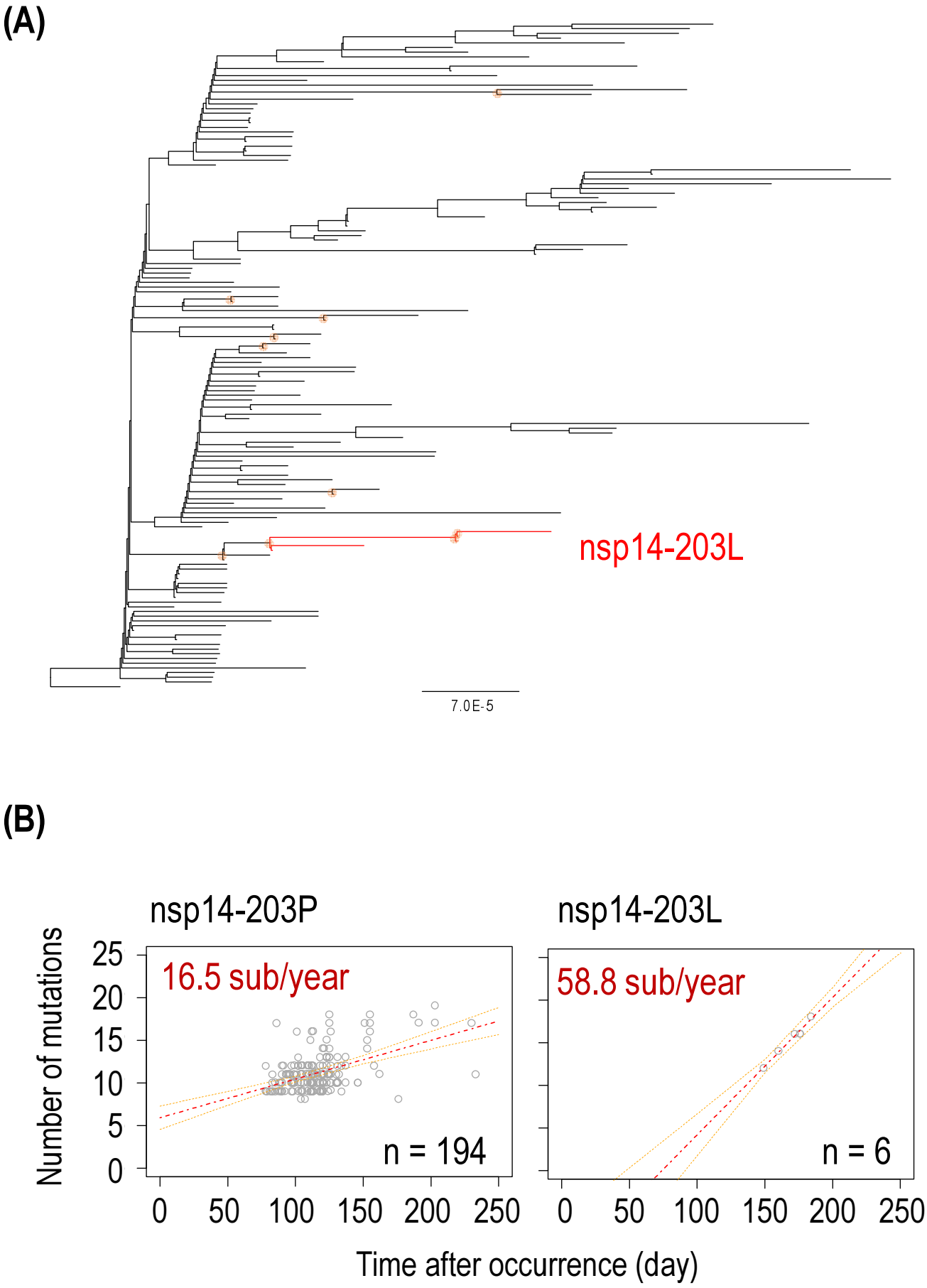
0.3
Severe_acute_respiratory_syndrome_coronavirus_2_isolate_Wuhan-Hu-1_[NC_045512.2]
hCoV-19/bat/Yunnan/RmYN02/2019_derived_from_Rhinolophus_malayanus_[EPI_ISL_412977]
hCoV-19/bat/Yunnan/RaTG13/2013_derived_from_Rhinolophus_affinis_[EPI_ISL_402131]
hCoV-19/pangolin/Guangxi/P4L/2017_derived_from_Manis_javanica_[EPI_ISL_410538]
Bat_SARS-like_coronavirus_isolate_bat-SL-CoVZXC21_[MG772934.1]
Coronavirus_BtRf-BetaCoV/SC2018_[MK211374.1]
Bat_SARS-like_coronavirus_isolate_Rs4231_[KY417146.1]
SARS_coronavirus_Tor2_[AY274119.3]
Bat_Hp-betacoronavirus/Zhejiang2013_[NC_025217.1]
Tylonycteris_bat_coronavirus_HKU4_[NC_009019.1]
Pipistrellus_bat_coronavirus_HKU5_[NC_009020.1]
Bat_coronavirus_isolate_PREDICT/PDF-2180_[NC_034440.1]
Middle_East_respiratory_syndrome-related_coronavirus_isolate_HCoV-EMC/2012_[NC_019843.3]
Betacoronavirus_Erinaceus/VMC/DEU/2012_isolate_ErinaceusCoV/2012-216/GER/2012_[KC545386.1]
Bat_coronavirus_HKU9-5-2_[HM211099.1]
Bat_coronavirus_HKU9-2_[EF065514.1]
Bat_coronavirus_HKU9-1_[EF065513.1]
Rousettus_bat_coronavirus_isolate_GCCDC1_346_[KU762337.1]
Bovine_coronavirus_isolate_BCoV-ENT_[AF391541.1]
White-tailed_deer_coronavirus_US/OH-WD470/1994_[FJ425187.1]
Human_enteric_coronavirus_4408_[FJ415324.1]
Human_coronavirus_OC43_strain_ATCC_VR-759_[AY585228.1]
Canine_respiratory_coronavirus_strain_BJ232_[KX432213.1]
Porcine_hemagglutinating_encephalomyelitis_virus_strain_VW572_[DQ011855.1]
Rabbit_coronavirus_HKU14_[NC_017083.1]
Betacoronavirus_HKU24_strain_HKU24-R05005I_[NC_026011.1]
Murine_hepatitis_virus_strain_A59_[AY700211.1]
Rat_coronavirus_Parker_[NC_012936.1]
Human_coronavirus_HKU1_[NC_006577.2]
Feline_infectious_peritonitis_virus_[AY994055.1]
Canine_coronavirus_strain_1-71_[JQ404409.1]
Transmissible_gastroenteritis_virus_[NC_038861.1]
Swine_enteric_coronavirus_strain_Italy/213306/2009_[NC_028806.1]
Ferret_coronavirus_genomic_RNA_strain_FRCoV4370_[LC119077.1]
Mink_coronavirus_strain_WD1127_[NC_023760.1]
Bat_coronavirus_HKU2_[NC_009988.1]
Porcine_epidemic_diarrhea_virus_strain_CV777_[AF353511.1]
Scotophilus_bat_coronavirus_512_[NC_009657.1]
BtRf-AlphaCoV/HuB2013_[NC_028814.1]
Bat_coronavirus_HKU8_[NC_010438.1]
Camel_alphacoronavirus_isolate_camel/Riyadh/Ry141/2015_[NC_028752.1]
Human_coronavirus_229E_[AF304460.1]
Human_coronavirus_NL63_isolate_NL63/RPTEC/2004_[JX504050.1]
NL63-related_bat_coronavirus_strain_BtKYNL63-9a_[NC_032107.1]
BtMr-AlphaCoV/SAX2011_[NC_028811.1]
Wencheng_Sm_shrew_coronavirus_isolate_Xingguo-101_[NC_035191.1]
Lucheng_Rn_rat_coronavirus_isolate_Lucheng-19_[NC_032730.1]
Coronavirus_AcCoV-JC34_[NC_034972.1]
Beluga_Whale_coronavirus_SW1_[NC_010646.1]
Turkey_coronavirus_[NC_010800.1]
Avian_infectious_bronchitis_virus_[NC_001451.1]
Thrush_coronavirus_HKU12-600_[NC_011549.1]
White-eye_coronavirus_HKU16_[NC_016991.1]
Bulbul_coronavirus_HKU11-934_[NC_011547.1]
Porcine_coronavirus_HKU15_strain_HKU15-155_[NC_039208.1]
Sparrow_coronavirus_HKU17_[NC_016992.1]
Munia_coronavirus_HKU13-3514_[NC_011550.1]
Magpie-robin_coronavirus_HKU18_[NC_016993.1]
Common-moorhen_coronavirus_HKU21_[NC_016996.1]
Night-heron_coronavirus_HKU19_[NC_016994.1]
Wigeon_coronavirus_HKU20_[NC_016995.1]
Guangdong_chinese_water_skink_coronavirus_strain_ZGLXR118981_[MG600026.1]
Sarbeco
Merbeco
beta
Nobeco
Embeco
alpha
gamma
delta

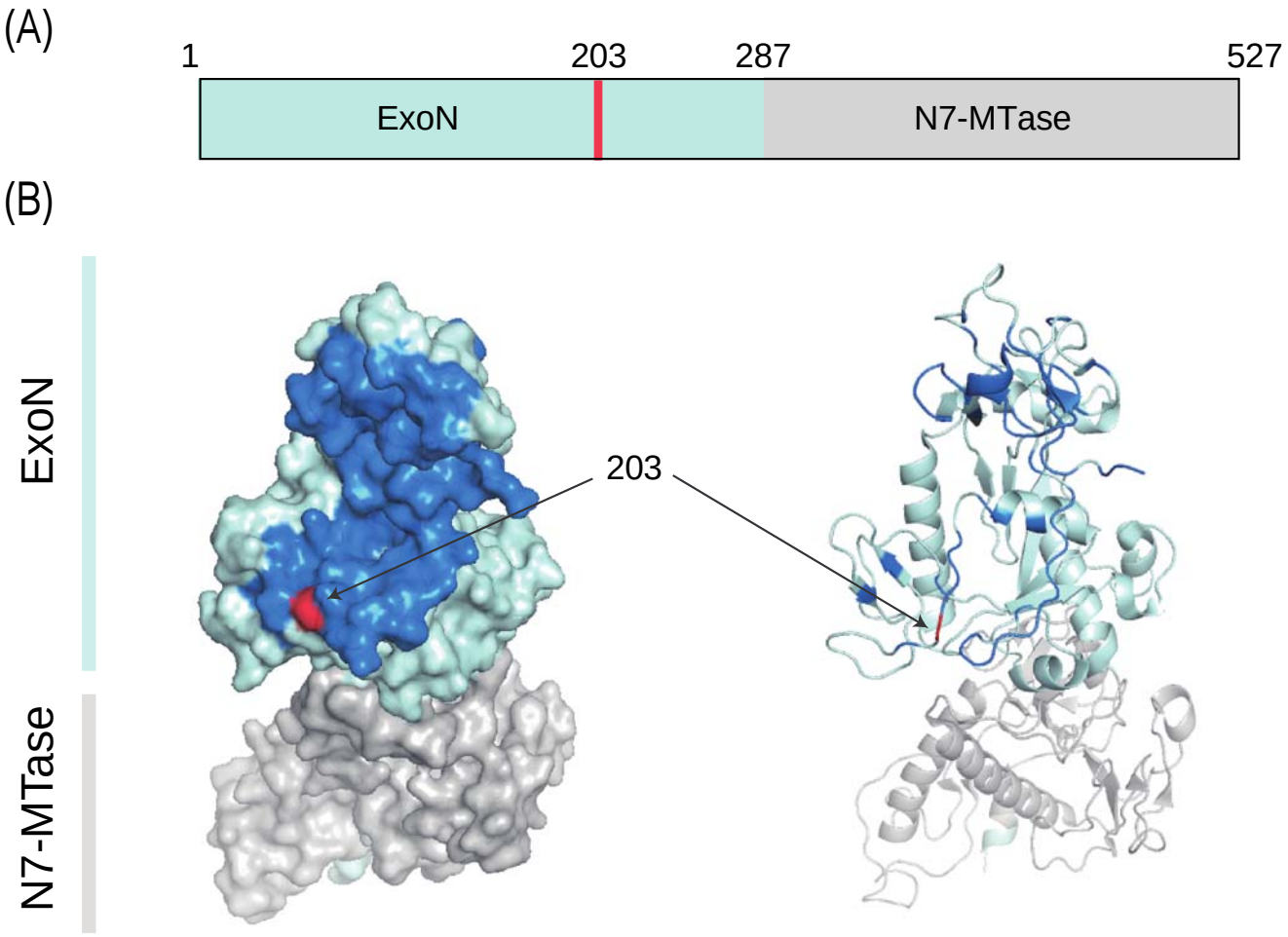
(B)

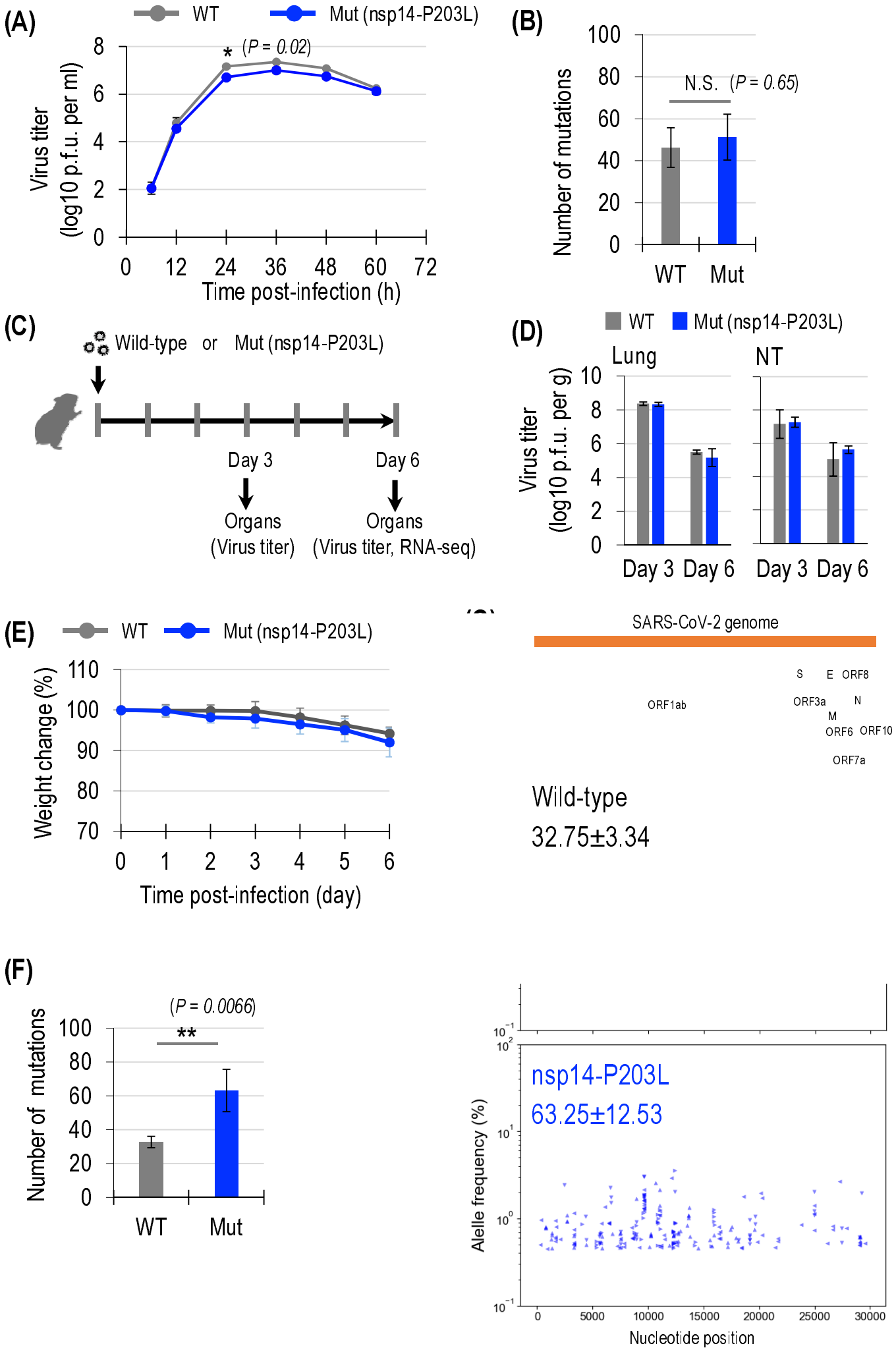

Conserved amino acid
Non-conserved amino acid
Orthocoronaviruses
Alp/Bet/Gam/Del
Betacoronaviruses
Sarbecovirus
SARS-CoV, -2
1
100
200
300
400
500
Amino acid position

(C)


90
92
193
268
273


(A)
nsp14-203L
7.0E-5
(B)
nsp14-203P
nsp14-203L
16.5 sub/year
58.8 sub/year
n = 194
n = 6
Number of mutations
0
5
10
15
20
25
0
50
100
150
200
250
Time after occurrence (day)


(A)
1
203
287
527
ExoN
N7-MTase
(B)
ExoN
N7-MTase
203


(A)
WT
Mut (nsp14-P203L)
* (P = 0.02)
Virus titer (log10 p.f.u. per ml)
Time post-infection (h)
(B)
N.S. (P = 0.65)
Number of mutations
WT
Mut
(C)
Wild-type or Mut (nsp14-P203L)
Day 3
Day 6
Organs (Virus titer)
Organs (Virus titer, RNA-seq)
(D)
WT
Mut (nsp14-P203L)
Lung
NT
Virus titer (log10 p.f.u. per g)
Day 3
Day 6
(E)
WT
Mut (nsp14-P203L)
Weight change (%)
Time post-infection (day)
SARS-CoV-2 genome
S
E
ORF8
ORF1ab
ORF3a
N
M
ORF6
ORF10
ORF7a
Wild-type
32.75±3.34
(F)
(P = 0.0066)
**
Number of mutations
WT
Mut
nsp14-P203L
63.25±12.53
Alelle frequency (%)
Nucleotide position