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

Kosuke Takada[1,2], Mahoko Takahashi Ueda[3], Tokiko Watanabe[1] and So Nakagawa[2]*.

[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*

*Corresponding author

E-mail: so@tokai.ac.jp (So Nakagawa)

Abstract

Nucleotide substitution rate of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is relatively low compared to the other RNA viruses because coronaviruses including SARS-CoV-2 encode non-structural protein 14 (nsp14) that is an error-correcting exonuclease protein. In this study, to understand genome evolution of SARS-CoV-2 in the current pandemic, we examined mutations of SARS-CoV-2 nsp14 which could inhibit its error-correcting function. First, to obtain functionally important sites of nsp14, we examined 62 representative coronaviruses belonging to alpha, beta, gamma, delta, and unclassified coronaviruses. As a result, 99 out of 527 amino acid sites of nsp14 were evolutionarily conserved. We then examined nsp14 sequences obtained from 28,082 SARS-CoV-2 genomes and identified 6 amino acid changes in nsp14 mutants that were not detected in the 62 representative coronaviruses. We examined genome substitution rates of these mutants and found that an nsp14 mutant with a proline to leucine change at position 203 (P203L) showed a higher substitution rate (35.9 substitutions/year) than SARS-CoV-2 possessing wild-type nsp14 (19.8 substitutions/year). We confirmed that the substitution rate of the P203L is significantly higher than those of other variants containing mutations in structural proteins. Although the number of SARS-CoV-2 variants containing P203L mutation of nsp14 is limited (26), these mutants appeared at least 10 times independently in the current pandemic. These results indicated that the molecular function of nsp14 is important for survival of various coronaviruses including SARS-CoV-2 and that some mutations such as P203L of nsp14 inhibiting its error-correcting function are removed rapidly due to their deleterious effects.

(250 words, Max 250 words)

36 **Introduction**

37 Coronavirus disease of 2019 (COVID-19) became a pandemic in 2020, infecting more than
38 60 million people worldwide as of the end of November 2020 with no end in sight. Severe acute
39 respiratory syndrome (SARS) coronavirus-2 (SARS-CoV-2) is the causing agent of this symptom (Lu,
40 et al. 2020; Wu, et al. 2020; Zhou, et al. 2020; Zhu, et al. 2020). SARS-CoV-2 is a positive-sense
41 single-stranded RNA (+ssRNA) that belongs to the coronavirus family (*Coronaviridae*), which is a
42 member of the order *Nidovirus* (Lu, et al. 2020; Wu, et al. 2020; Zhou, et al. 2020; Zhu, et al. 2020).
43 According to the classification of the International Committee on Taxonomy of Viruses (ICTV),
44 *Coronaviridae* are divided into two subfamilies, *Letovirinae* and *Orthocoronavirinae*.
45 *Orthocoronavirinae* is composed of four genera (*Alpha*-, *Beta*-, *Gamma*-, and *Delta*-) of coronaviruses
46 (CoVs). *Betacoronavirus* including SARS-CoV-2 as well as SARS-CoV are mainly composed of 5
47 subgenera: *Salbecovirus*, *Hibecovirus*, *Nobecovirus, Merbecovirus*, and *Embecovirus* (Woo, et al.
48 2010; Nakagawa and Miyazawa 2020). SARS-CoV-2 and SARS-CoV belongs to the subgenus
49 *Salbecovirus* (Lu, et al. 2020; Wu, et al. 2020; Zhu, et al. 2020).

50 One of the unique genomic characteristics of CoVs is their lower mutation rates compared
51 with other RNA viruses (Eckerle, et al. 2007; Eckerle, et al. 2010). RNA viruses usually have a high
52 mutation rate, which is partially caused by the relatively low fidelity of RNA-dependent RNA
53 polymerase (RdRp) and the lack of correction of nucleotide incorporation errors. In CoVs, however,
54 nonstructural protein 14 (nsp14) in an open reading frame 1ab (ORF1ab) cleaved by internal proteases
55 (present in nsp3 and nsp5) (Hegyi and Ziebuhr 2002; Thiel, et al. 2003) contains an exoribonuclease
56 (ExoN) domain that performs proofreading during replication steps (Snijder, et al. 2003). Functional
57 inactivation experiments of nsp14 of SARS-CoV demonstrated 10-20 times higher mutation rates
58 (Eckerle, et al. 2010; Graham, et al. 2012; Smith, et al. 2013). It is known that CoVs have the largest
59 RNA genome of all RNA viruses (Gorbalenya, et al. 2006), and, therefore, nsp14 could be important
60 for maintaining their genomes.

To understand genome evolution of SARS-CoV-2 during the current pandemic, we examined SARS-CoV-2 nsp14 mutants that may impair its error-correcting function. First, to examine the evolution and diversity of nsp14 in CoVs, we generated a phylogenetic tree of the nsp14 gene obtained from a variety of 62 CoVs belonging to the four genera as well as an unclassified one (Supplementary table 1). We then evaluated evolutionary conserved amino acid sites to infer functionally important sites. We examined 28,082 SARS-CoV-2 genomes that are available in the GISAID database, which includes not only sequencing data but also sampling date, country, and institution information (Shu and McCauley 2017). We examined various nsp14 SARS-CoV-2 mutants that had spread in the viral population. Among them, highly frequent nsp14 mutants that were not observed among the diversified 62 CoVs were identified. Based on those results, we found that an amino acid replacement from proline (P) to leucine (L) at position 203 of nsp14 increased a nucleotide substitution rate of SARS-CoV-2 genomes. We validated the higher mutation rate compared with those of structural proteins. These results indicate that some of variations in nsp14 may increase diversity of SARS-CoV-2 genomes.

Results

Conserved region of nsp14 proteins in various coronaviruses

We first generated a phylogenetic tree of amino acid sequences of the nsp14 gene obtained from representative 62 coronaviruses (CoVs); the viruses used in the analysis was summarized in Supplementary table 1. The phylogenetic tree based on the 62 amino acid sequences of nsp14 clearly showed the relationship of the four genera of CoVs (Fig. 1), which is almost identical to the results obtained by using the partial amino acid sequences of ORF1ab described in a previous report (Nakagawa and Miyazawa 2020). Our results suggest that the nsp14 gene is under negative selection due to its molecular function, and that the conserved amino acid sites in particular could be important.

Subsequently, we examined the conserved amino acid sites of the nsp14 gene among the 62 CoVs (Fig 2A). The amino acid positions of nsp14 are based on SARS-CoV, three-dimensional structure of which is available (PDB ID: 5C8S). SARS-CoV is a coronavirus belonging to the subgenus *Sarbecovirus* in the genus *Betacoronavirus*, which is the same as SARS-CoV-2. For nsp14, the number of identical amino acids was 482/527 (91.5%) in 8 representative CoVs belonging to the subgenus *Sarbecovirus*, and 359/527 (68.1%) if we added nsp14 of the Bat Hp-betacoronavirus/Zhejiang2013 (GenBank ID: NC_025217.1) belonging to the subgenus *Hibecovirus* which is known as the outgroup of *Sarbecovirus.* Moreover, 200/527 (38.0%) positions were preserved among 29 viruses of the genus *Betacoronavirus*, and 108/527 (20.5%) positions were preserved among the 61 CoVs (Supplementary table 1). If we include an unclassified CoV found in *Tropidophorus sinicus* (Chinese waterside skink) (Shi, et al. 2018), the conserved amino acid positions were 99/527 (18.8%). These results suggest that various amino acid sites of nsp14 gene are conserved among diverse CoVs 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 62 CoVs using SLAC program (see Materials and Methods) (Kosakovsky Pond and Frost 2005). Note that codon position 504 of nsp14 was calculated using 61 CoVs excluding the Chinese waterside skink CoV because it has a mixed nucleotide at that position. As a result, no positive

102 selection sites were found in nsp14, whereas 379 out of 527 sites were under negative selection ($P <$
103 0.01, Supplementary fig 1). In total, 91 amino acids were unchanged as well as under negative selection
104 in nsp14 of the 62 representative CoVs.

105 We investigated whether the amino acids that are known to be important for the error-
106 correcting function of nsp14 are conserved. CoV nsp14 is an enzyme that possesses both ExoN and
107 N7-MTase activities. The ExoN domain of CoV nsp14 was originally identified based on sequence
108 similarity to distant cell homologs (Snijder, et al. 2003) and has been assigned to the DEDD
109 exonuclease superfamily, which includes the proofreading domains of many DNA polymerases and
110 other eukaryotic and prokaryotic exonucleases. The nsp14 proteins with four conserved active site
111 residues were distributed in three standard motifs of the primary structure: residues 90D/92E (motif I),
112 191E (motif II), and 273D (motif III) (Ogando, et al. 2019; Romano, et al. 2020). We found that the
113 four active sites of nsp14 were conserved among the 62 CoVs examined in this analysis (Fig 2B).
114 Moreover, there were three zinc fingers that were essential for nsp14 function: the first zinc finger
115 (ZF1): 207C, 210C, 226C, and 229H; the second zinc finger (ZF2): 257H, 261C, 264H, and 279C; and
116 the third zinc finger (ZF3): 452C, 473C, 484C, and 487H. S-adenosyl methionine (SAM)-binding
117 motif I include residues 331D, 333G, 335P and 337A (or 337G), which were involved in the (N7
118 guanine)-methyl transferase reaction by nsp14. We found that all of these residues were conserved
119 among the nsp14 sequences of the 62 CoVs examined in this analysis (Supplementary fig 2). These
120 results suggest that the amino acids of the nsp14 that are functionally important are conserved among
121 diversified CoVs.

122

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

124 Amino acid comparisons of nsp14 in the 62 representative CoVs revealed that several sites
125 are strongly conserved, and that some substitutions may have deleterious effects for the survive of
126 SARS-CoV-2. SARS-CoV-2 was first detected in humans in 2019 and has spread rapidly all over the
127 world in a short period of time. Therefore, some mutations of nsp14 that impair its error-correcting

128 function may remain in the SARS-CoV-2 population. Therefore, we attempted to identify such
129 deleterious mutations in the current SARS-CoV-2 population.

130 We downloaded the aligned nucleotide sequence of SARS-CoV2 (87,625 sequences) from
131 the GISAID database as of September 7, 2020. The SARS-CoV2 genomes isolated from humans were
132 extracted (87,304 sequences), and the sequences containing undetermined and/or mixed nucleotides
133 were removed; the remaining 28,082 sequences were used in subsequent analyses. The coding region
134 of nsp14 is translated and compared. As a result, 25,913 nsp14 amino acid sequences (92.27%) were
135 identical to that of the reference sequence, while 2169 sequences contain at least one amino acid
136 replacement in nsp14 (Supplementary table 2). Among them, the top 20 variants possessing nsp14
137 with amino acids that were different from the major sequence are summarized in Table 1. Note that
138 frameshift mutants were excluded in Table 1 because they may affect downstream genes as well. It
139 was revealed that SARS-CoV-2, which acquired various mutations in nsp14, was present, although
140 the number of examined sequences is limited. Notably, among the 20 amino acid nsp14 mutants, 6
141 amino acids (177F, 377L, 369F, 501I, 203L and 297S) were not found in the 62 nsp14 of the
142 representative CoVs (Supplementary fig 2).

143 We further analyzed the evolutionary selection pressures of these 6 amino acid sites using 73
144 CoVs belonging to the genus *Betacoronaviruses* that includes SARS-CoV-2 (Supplementary table 3).
145 We confirmed the 6 amino acid variants were not found in the 73 Beta-CoVs (Supplementary table 3).
146 The 6 conserved positions (177, 203, 297, 369, 377, and 501) were also found to be under negative
147 selection ($P < 0.01$, Supplementary table 4). Overall, the results indicated that the conserved amino
148 acid sites of nsp14 are evolutionarily selected in the genus *Betacoronaviruses* as well.

149

150 **Correlation between amino acid sequences of nsp14 and nucleotide substitution rates in the viral**
151 **genome**

152 To examine whether the 6 amino acid replacements observed in the nsp14 gene of SARS-
153 CoV-2 variants affect its proofreading function, we investigated a correlation between the amino acid

154 variants of nsp14 and the genetic diversities of SARS-CoV-2 genomes. The nucleotide substitution
155 rates of SARS-CoV-2 possessing mutated nsp14 variants were compared to that of SARS-CoV-2 with
156 wild-type nsp14. Note that sequences for which sampling date information (year, month, and day) was
157 not available were excluded from this analysis. Nucleotide substitutions per year were calculated based
158 on the regression coefficient using hCoV-19/Wuhan/Hu-1/2019 (GISAID ID: EPI_ISL_402125) as a
159 reference sequence (see Materials and Methods).

160 We found that SARS-CoV-2 containing the wild type nsp14 had 19.8 nucleotide substitutions
161 in whole genome per year (Table 2). Conversely, 7 SARS-CoV-2 variants containing nsp14 with
162 amino acid replacements showed 27.3 (L177F), 59.8 (L177F/T372I), 35.9 (P203L), 17.1 (P297S), 26.2
163 (S369F), 13.0 (F377L), or 21.9 (M501I) nucleotide substitutions per year (Fig 3). Although the
164 substitution rates of the L177F/T372I and F377L variants were found to be not reliable because of
165 limited time ranges, those of the other variants were clearly correlated (Fig 3). Therefore, SARS-CoV-
166 2 variants possessing nsp14 with the L177F, P203L, S369F, or M501I replacement may have a
167 particularly higher genomic mutation rate than do SARS-CoV-2 variants with wild-type nsp14.

168 We further verified the elevated substitution rates found in various nsp14 mutants. We used
169 all sequences that did not contain any ambiguous nucleotides (shown in Supplementary table 5) to
170 calculate the mutation rates; however, this could cause an overestimation because of sampling biases.
171 Therefore, we extracted representative SARS-CoV-2 genomes by selecting the sequence showing the
172 oldest sampling date if the genome sequences were identical. Using the representative dataset, we
173 analyzed the nucleotide substitutions per year of the amino acid replacements for each site of nsp14 if
174 amino acid replacements were present in 20 or more sequences. Note that amino acid replacements
175 caused by frameshifts were removed. As a result, 16 amino acid replacements in nsp14 were evaluated,
176 and the mean value of nucleotide substitution rate was 20.7 (Supplementary fig 3A). Among them, we
177 found that the P203L variant of nsp14 (nsp14-P203L) showed the highest mutation rate (35.7
178 nucleotide mutations per year).

179 We also evaluated the substitution rates of the 7 nsp14 variants for each single cluster in which
180 the maximum number was observed. We assumed that variants were in the same clade if their Pangolin
181 lineages (Rambaut, et al. 2020) and GISAID clades were identical. We found that nsp14-P203L
182 variants in the Pangolin lineage B.1.1.33 showed the largest nucleotide substitution rate (58.8
183 nucleotide substitutions/year) (Fig 4A, B and Table 3). Although the number of variants was limited
184 (6), a correlation was clearly observed (Fig 4B and Table 3). In addition, other 6 nsp-P203L variants
185 in a different cluster (B.1) also showed a relatively high nucleotide substitution rate (28.0 nucleotide
186 substitutions/year) (Table3 and Supplementary fig 4). These results clearly indicate that SARS-CoV-
187 2 variants containing nsp14-P203L could accumulate mutations in the viral population.
188
189 **Mutations in structural genes of SARS-CoV-2 and validation of substitution rates**
190 It was reported that nucleotide mutations in the M and E structural genes have been used to
191 calculate genome-wide mutation load of SARS-CoV-2, the mutations of which may not affect RNA
192 replication (Eskier, Karakulah, et al. 2020; Eskier, Suner, et al. 2020). Therefore, we also analyzed
193 nucleotide mutations of the structural genes that do not directly interact with viral RNAs: the S, M,
194 and E genes. As the number of nucleotide mutations in the E or M gene was only 0 or 1 in most
195 sequences, we summed up the number of mutations in those three genes. We found that viruses
196 containing nsp14-L177F or nsp14-P203L had higher rates of nucleotide substitutions in the S, M and
197 E genes per year than did viruses with the wild-type nsp14 (Table 2). These results suggest that viruses
198 carrying nsp14-L177F and nsp14-P203L viruses have an increased substitution rate.
199 To examine whether the substitution rates observed in the nsp14 mutants are high, we
200 calculated the nucleotide substitution rates of the mutants of each structural gene if they were present
201 in 20 or more sequences. 40 mutants containing 22, 3 and 5 amino acid replacements in sites of the S,
202 E and M genes, respectively, were remained. The mean rate of nucleotide substitutions per year was
203 18.4, and its distribution is shown in Supplementary fig 3B. We then investigated a possibility that the
204 increase in nucleotide substitution rates observed in the nsp14 gene occurred by chance. We found that

205 the nucleotide substitution rate of the nsp14-P203L mutant was over 5% in the distribution, although
206 the largest rate was found in the S-V6F mutant (Fig 5). However, the elevated nucleotide substitution
207 rate of V6F was found to be due to the effects of an outlier sample shown in the black dot in
208 Supplementary fig 5. In fact, if we removed the outlier sample, the nucleotide substitution rate was
209 reduced to 26.2. Conversely, the amino acid replacement rate of nsp14 did not show any extreme bias
210 in the sample. These results suggest that P203L of nsp14 has a higher rate of genomic mutation
211 compared with other amino acid replacements.
212
213 **nsp14-P203L variant in the SARS-CoV-2 population**
214 The nsp14-P203L variants have been isolated 26 patients mainly in Europe and North
215 America, with the first being reported on March 5, 2020 in Netherlands (GISAID ID: EPI_ISL_454756,
216 Supplementary table 5). Considering their Pangolin lineages (Rambaut, et al. 2020) as well as GISAID
217 clades (shown in Supplementary table 5), the nsp14-P203L variant appeared at least 10 times. In fact,
218 the nsp14-P203L variants were observed in both of the Spike D614 and G614 variants, which changes
219 the efficiency of infection and have spread worldwide (Korber, et al. 2020). However, we also found
220 that the nsp14-P203L variants did not increase steadily over time. As we examined, the largest number
221 of the nsp14-P203L variants found in the B.1.1.33 and B.1 lineages was only 6 for each
222 (Supplementary table 5).
223
224 **nsp14-P203L mutation effect on protein structure.**
225 To estimate an effect of the P203L mutation of nps14, we first investigated the location of the
226 residue 203 using a tertiary structure model of SARS-CoV nsp14 (PDB ID: 5C8S) (Ma, et al. 2015).
227 As the protein structure of SARS-CoV-2 nsp14 is currently not available in the PDB database, we used
228 the protein structure of SARS-CoV nsp14. Note that the amino acid residues around position 203 of
229 nsp14 SARS-CoV and SARS-CoV-2 are well conserved (Supplementary fig 6A). Residue 203 is
230 exposed on the surface of the protein and is located in the loop stretching from residues 200 to 214

231 where is in the interstices of zinc and magnesium binding sites (positions 207 and 191, respectively,
232 Fig 6). We found that the loop is close to the amino acids 1-58, which constitutes an interaction
233 surface for binding a co-factor, SARS-CoV nsp10 (Supplementary fig 6B) (Ogando, et al. 2020). It
234 was reported that the interaction with nsp10 greatly increases nucleolytic activity of SARS-CoV nsp14
235 (Bouvet, et al. 2012). Therefore, we investigated the possibility that residues at or around position 203
236 play an important role in nsp10-nsp14 interaction. We searched the interaction interface using SARS-
237 CoV nsp14-nsp10 complex structure (PDB ID: 5C8S) (Ma, et al. 2015) by focusing the residues
238 around position 203. Residues 199-203 (VKIGP) of the nsp14 were identified as a part of interaction
239 interface. However, the detailed analysis of the protein complex showed that the residues 200 and 201
240 alone bound directly to the residues 9 and 21 of the nsp10 which are also well conserved between
241 SARS-CoV and SARS-CoV-2 (Supplementary fig 6A). These results suggest that the residue 203 may
242 support binding to nsp10 indirectly by constituting the interaction interface around residues 200 and
243 201. We further predicted the impact of the P203L mutation on protein-protein binding affinity using
244 BeAtMuSiC program (Dehouck, et al. 2013), and found that the mutation decreases the binding
245 affinity of the protein (0.38 kcal/mol). The mutation of P to L increases hydrophobicity at the position
246 203 (Supplementary fig 6C and 6D), which might alter the local structure afforded by the residues
247 around position 203, and lead to a change in the affinity between nsp14 and nsp10.
248

249 **Discussion**

250 In this study, we found that nsp14 was well conserved in highly diversified CoVs (Fig 2), and
251 that 379 out of the 527 amino acid sites of nsp14 of the 62 representative CoVs were under negative
252 selection (Supplementary fig 1). For SARS-CoV-2, we found that 92.27% of the viruses had an nsp14
253 amino acid sequence that was identical to the reference sequence in all of the genome sequences
254 examined in this study (Table 1). We found that the nsp14-P203L variants showing a higher nucleotide
255 substitution rate did not increase steadily over time (0.09% in the current population, Table 1),
256 although they appeared multiple times. In fact, the nsp14-nsp10 complex of SARS-CoV suggests that
257 the amino acid residue at position 203 of nsp14 could be important to interact with nsp10, around
258 which amino acids are well conserved between SARS-CoV-2 and SARS-CoV (Supplementary fig 6A).
259 The nsp14-P203L mutants were suggest to weakly interact with nsp10 (Supplementary fig 6C and 6D),
260 possibly resulting in reduction of the proofreading activity of nsp14. Overall, these results suggest the
261 biological importance of nsp14 in various CoVs including SARS-CoV-2, and that nsp14 mutations
262 may have deleterious effects on the survival of these viruses. In RNA viruses, rapid mutation and
263 substitution rates are sometimes beneficial because certain mutations allow them to survive in various
264 situations (Mansky and Cunningham 2000; Vignuzzi, et al. 2006; Holmes 2009; Duffy 2018).
265 However, in CoVs, high mutation rates may not be preferred, which may be beneficial for the
266 development of vaccines and antiviral drugs.

267 The frequency of nucleotide substitutions of CoVs may be affected not only by the nsp14
268 proofreading mechanism but also by the characteristics of other proteins, such as fidelity of RNA-
269 dependent RNA polymerase (nonstructural protein 12; nsp12) (Graepel, et al. 2017). In addition, the
270 proofreading activity of nsp14 is also affected by the amino acid substitution of nsp10, which
271 contributes to its interaction with nsp14 (Bouvet, et al. 2012). It is possible that amino acid
272 replacements in other proteins involved in genome replication, rather than nsp14-203L, actually
273 affected the nucleotide substitution rates of SARS-CoV-2. Therefore, amino acid replacements were
274 examined among nsp14-203L variants with the following genes being involved in gene replication:

275 nsp7, nsp8, nsp9, nsp10, nsp12, nsp13, nsp15, and nsp16. As result, proline-to-lysine replacement at
276 position 323 (P323L) of nsp12 was frequently found (20/25) in sequences of SARS-CoV-2 possessing
277 nsp14-203L (Supplemental table 6). However, the P323L mutation in nsp12 (nsp12-P323L) is known
278 to accompany an aspartate-to-glycine replacement at position 614 (D614G) of S protein (Hou, et al.
279 2020; Korber, et al. 2020). The S-D614G mutant was first reported in the end of January 2020 in China
280 and Germany, and after March 2020, the mutant is major variant in various regions in all over the
281 world (Korber, et al. 2020). Therefore, P323L mutation nsp12 was found in nsp14-203L variants as
282 well. Indeed, the genomic mutation rate of SARS-CoV-2 with the P323L amino acid replacement in
283 nsp12, was found to be 17.7 nucleotide mutations per year (Supplemental fig 7). Therefore, it is
284 unlikely that the P323L in nsp12 directly affects the genomic mutation rate of SARS-CoV-2 with the
285 P203L amino acid replacement in nsp14.

286 According to previous reports, SARS-CoV-2 base substitutions are mostly C to U nucleotide
287 substitutions (Simmonds 2020). Our analysis also revealed that SARS-CoV-2 has many C to U
288 nucleotide substitutions (Supplemental table 7). Furthermore, it was clarified that there are many
289 nucleotide substitutions from C to U in the nucleotide substitution in the population possessing nsp14-
290 P203L (Supplemental table 7). These results suggest that the proofreading activities of nsp14-203P
291 and nsp14-203L may not have nucleotide preferences.

292 It is known that alanine substitution of an amino acid that is important for the proofreading
293 activity of nsp14 of SARS-CoV increases viral genomic mutations in cultured cell (Eckerle, et al.
294 2010; Smith, et al. 2013) and *in vivo* (Graham, et al. 2012). In addition, Graham et al. reported that the
295 SARS-CoV virus is attenuated by alanine replacement of an amino acid that is important for the
296 proofreading activity of nsp14 (Graham, et al. 2012). The effects of amino acid replacements in nsp14
297 on coronavirus infections, which have been found in nature, have not been investigated to date. In our
298 study, the candidates that might affect nsp14 function were searched by focusing on the amino acid
299 replacements of nsp14 in epidemic viruses and comparing genomic mutation rates; subsequently, we
300 proposed candidates for amino acid replacements that may affect viral properties. However, the

301 molecular mechanisms underlying the amino acid replacements in SARS-CoV-2 suggested by our
302 study are unclear. Further studies are required to clarify the effects of the candidate amino acids on the
303 properties of CoVs. In conclusion, we propose that rare amino acid replacements (particularly leucine
304 at amino acid site 203) in coronavirus nsp14 may increase the rate of genomic mutations.
305

306 **References**

307 Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences
308 guided by amino acid translations. Nucleic Acids Res 38:W7-13.
309 Bouvet M, Imbert I, Subissi L, Gluais L, Canard B, Decroly E. 2012. RNA 3'-end mismatch excision
310 by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14
311 exoribonuclease complex. Proc Natl Acad Sci U S A 109:9372-9377.
312 Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator.
313 Genome Res 14:1188-1190.
314 Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. 2020. ModelTest-NG: A New
315 and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. Mol Biol Evol
316 37:291-294.
317 Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of
318 protein evolution. Bioinformatics 27:1164-1165.
319 Dehouck Y, Kwasigroch JM, Rooman M, Gilis D. 2013. BeAtMuSiC: Prediction of changes in
320 protein-protein binding affinity on mutations. Nucleic Acids Res 41:W333-339.
321 Duffy S. 2018. Why are RNA virus mutation rates so damn high? PLoS Biol 16:e3000003.
322 Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X, Scherbakova S, Graham RL, Baric RS,
323 Stockwell TB, et al. 2010. Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is
324 revealed by complete genome sequencing. PLoS Pathog 6:e1000896.
325 Eckerle LD, Lu X, Sperry SM, Choi L, Denison MR. 2007. High fidelity of murine hepatitis virus
326 replication is decreased in nsp14 exoribonuclease mutants. J Virol 81:12135-12144.
327 Eskier D, Karakulah G, Suner A, Oktay Y. 2020. RdRp mutations are associated with SARS-CoV-2
328 genome evolution. PeerJ 8:e9587.
329 Eskier D, Suner A, Oktay Y, Karakulah G. 2020. Mutations of SARS-CoV-2 nsp14 exhibit strong
330 association with increased genome-wide mutation load. PeerJ 8:e10181.
331 Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ. 2006. Nidovirales: evolving the largest RNA
332 virus genome. Virus Res 117:17-37.
333 Graepel KW, Lu X, Case JB, Sexton NR, Smith EC, Denison MR. 2017. Proofreading-Deficient
334 Coronaviruses Adapt for Increased Fitness over Long-Term Passage without Reversion of
335 Exoribonuclease-Inactivating Mutations. mBio 8.
336 Graham RL, Becker MM, Eckerle LD, Bolles M, Denison MR, Baric RS. 2012. A live, impaired-
337 fidelity coronavirus vaccine protects in an aged, immunocompromised mouse model of lethal
338 disease. Nat Med 18:1820-1826.
339 Hegyi A, Ziebuhr J. 2002. Conservation of substrate specificities among coronavirus main proteases.
340 Journal of General Virology 83:595-599.
341 Holmes EC. 2009. The Evolutionary Genetics of Emerging Viruses. Annual Review of Ecology,
342 Evolution, and Systematics 40:353-372.

Hou YJ, Chiba S, Halfmann P, Ehre C, Kuroda M, Dinnon KH, Leist SR, Schafer A, Nakajima N, Takahashi K, et al. 2020. SARS-CoV-2 D614G Variant Exhibits Enhanced Replication ex vivo and Earlier Transmission in vivo. bioRxiv.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780.

Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, 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 e819.

Kosakovsky Pond SL, Frost SD. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22:1208-1222.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A, Wren J. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35:4453-4455.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35:1547-1549.

Lanave C, Preparata G, Saccone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. J Mol Evol 20:86-93.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-1659.

Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, et al. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet 395:565-574.

Ma Y, Wu L, Shaw N, Gao Y, Wang J, Sun Y, Lou Z, Yan L, Zhang R, Rao Z. 2015. Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. Proc Natl Acad Sci U S A 112:9436-9441.

Mansky LM, Cunningham KS. 2000. Virus mutators and antimutators. Trends in Genetics 16:512-517.

Nakagawa S, Miyazawa T. 2020. Genome evolution of SARS-CoV-2 and its virological characteristics. Inflamm Regen 40:17.

Ogando NS, Ferron F, Decroly E, Canard B, Posthuma CC, Snijder EJ. 2019. The Curious Case of the Nidovirus Exoribonuclease: Its Role in RNA Synthesis and Replication Fidelity. Front Microbiol 10:1813.

Ogando NS, Zevenhoven-Dobbe JC, van der Meer Y, Bredenbeek PJ, Posthuma CC, Snijder EJ. 2020. The Enzymatic Activity of the nsp14 Exoribonuclease Is Critical for Replication of MERS-CoV and SARS-CoV-2. J Virol 94.

Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C, du Plessis L, Pybus OG. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 5:1403-1407.

Romano M, Ruggiero A, Squeglia F, Maga G, Berisio R. 2020. A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping. Cells 9.

Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, Wang W, Eden JS, Shen JJ, Liu L, et al. 2018. The evolutionary history of vertebrate RNA viruses. Nature 556:197-202.

Shu Y, McCauley J. 2017. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill 22.

Simmonds P. 2020. Rampant C-->U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. mSphere 5.

Smith EC, Blanc H, Surdel MC, Vignuzzi M, Denison MR. 2013. Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. PLoS Pathog 9:e1003565.

Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LLM, Guan Y, Rozanov M, Spaan WJM, Gorbalenya AE. 2003. Unique and Conserved Features of Genome and Proteome of SARS-coronavirus, an Early Split-off From the Coronavirus Group 2 Lineage. Journal of Molecular Biology 331:991-1004.

Thiel V, Ivanov KA, Putics A, Hertzig T, Schelle B, Bayer S, Weissbrich B, Snijder EJ, Rabenau H, Doerr HW, et al. 2003. Mechanisms and enzymes involved in SARS coronavirus genome expression. J Gen Virol 84:2305-2315.

Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R. 2006. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439:344-348.

Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. 2018. Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. Mol Biol Evol 35:773-777.

Woo PC, Huang Y, Lau SK, Yuen KY. 2010. Coronavirus genomics and bioinformatics analysis. Viruses 2:1804-1820.

Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, et al. 2020. A new coronavirus associated with human respiratory disease in China. Nature 579:265-269.

Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270-273.

Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, et al. 2020. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med 382:727-733.

416 **Methods**

417 **Coronavirus sequence data**

418 Amino acid sequences of open reading frame 1ab (ORF1ab) of the 62 representative CoVs
419 (summarized in Supplementary Table 1) were obtained from NCBI (https://www.ncbi.nlm.nih.gov/)
420 and the GISAID EpiCoV database (https://www.gisaid.org) as reported previously (Nakagawa and
421 Miyazawa 2020). Aligned nucleotide sequences of SARS-CoV-2 and their metadata were obtained
422 from the GISAID EpiCoV database as of September 7, 2020 (87,625 sequences). The CoV genome
423 sequences of which host were annotated as "Human" were extracted, and sequences containing
424 undetermined and/or mixed nucleotides were removed. The 28,082 remaining genomes were used for
425 the analyses. The open reading frame (ORF) of each sequence was determined from the alignment
426 sequence based on the ORF information of hCoV-19/Wuhan/Hu-1/2019 (GISAID ID:
427 EPI_ISL_402125, GenBank ID: NC_045512.2), which was used as the reference sequence. We then
428 removed the gaps from the alignment and translated the nucleotide sequences. To detect amino acid
429 replacements for each site, the amino acid sequence after the stop codon was removed and amino acid
430 sequences were aligned using the L-INS-i program in MAFFT version v7.453 (Katoh and Standley
431 2013). From the sequence alignment of nsp14 of the representative CoVs, the amino acids aligned at
432 the amino acid positions of SARS-CoV are shown using WebLogo 3 (Crooks, et al. 2004). Note that,
433 in the alignment sequence, gap sites were removed based on the SARS-CoV nsp14 sequence.

434

435 **Molecular evolutionary phylogenetic analysis**

436 To compute phylogenetic tree of nsp14 of the 62 representative CoVs, their nsp14 amino acid
437 sequences were aligned using the L-INS-i program in MAFFT version 7.453 (Katoh and Standley
438 2013). To estimate the appropriate amino acid replacement model, we used ProtTest3 version 3.4.2
439 (Darriba, et al. 2011) based on the Bayesian information criterion (BIC) values. We then computed a
440 phylogenetic tree using RAxML-NG version 1.0.0 (Kozlov, et al. 2019) with rapid 1000 bootstrap
441 replicates (Lanave, et al. 1984).

We also generated a phylogenetic tree of SARS-CoV-2 genomes belonging to the Pangolin lineages B.1.1.33 including nsp14-P203L variants. Using CD-HIT-EST version 4.8.1 (Li and Godzik 2006), the identical genomic sequences were removed: from 200 to 142 sequences. The 142 nucleotide sequences were aligned by the L-INS-i program in MAFFT version 7.453 (Katoh and Standley 2013). To estimate the appropriate nucleotide substitution model, we used ModelTest-NG (Darriba, et al. 2020) based on the Akaike information criterion (AIC) values, and GTR+I+G4 was selected. We then computed a phylogenetic tree using RAxML-NG version 1.0.0 (Kozlov, et al. 2019) with rapid 1000 bootstrap replicates (Lanave, et al. 1984).

Mutation rate analysis

A multiple sequence alignment of SARS-CoV-2 genomes obtained from the GISAID database was used to analyze the genomic mutation rate. The number of nucleotide substitutions per site from between sequences was counted. Analyses were conducted using the Kimura 2-parameter model. Codon positions included were 1st+2nd+3rd+Noncoding. All gaps were removed for each sequence pair with a pairwise deletion option. Counting of the number of nucleotide substitution was conducted in MEGA X (Kumar, et al. 2018). 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 the information on 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 using the least-squares method using in-house python scripts.

Measurement of natural selection pressure

TranslatorX (Abascal, et al. 2010) was used to align the nucleotide sequence based on the aligned amino acid sequence of nsp14 of the representative CoVs. To determine evolutionary selection

468 pressures operating the nsp14 gene in CoVs, we computed the ratio of nonsynonymous substitution
469 rates (dN) and synonymous substitution rates (dS) per site (dN/dS) using the Datamonkey software
470 (Kosakovsky Pond and Frost 2005; Weaver, et al. 2018) through MEGA X (Kumar, et al. 2018).
471
472 **Structural analysis**
473 A PDB file (5C8S) of the SARS-CoV nsp14-nsp10 complex was obtained from the PDB database
474 (https://www.rcsb.org). Using chains C and D of the structure, further analyses were performed. The
475 residues at the interface of the SARS-CoV nsp10-nsp14 docked complex were determined using the
476 "InterfaceResidues" PyMol script (http://www.pymolwiki.org/index.php/InterfaceResidues/). We
477 used the SARS-CoV nsp10-nsp14 docked complex for the prediction of the effect of the P203L
478 mutation on protein-protein interactions using the BeAtMuSiC program version 1.0
479 (http://babylone.3bio.ulb.ac.be/beatmusic/query.php) (Dehouck, et al. 2013). P203L mutant
480 generation, hydrophobicity calculation, and visualization were performed using the Discovery Studio
481 Visualizer software.
482

Acknowledgments

We thank the GISAID database (https://www.gisaid.org), which shares sequence data of SARS-CoV-2 and related viral species. This study was supported by KAKENHI Grant-in-Aid for Scientific Research on Innovative Areas 16H06429 (to T.W. and S.N.), 16K21723 (to T.W. and S.N.), 16H06434 (to T.W), and 19H04843 (to S.N.); AMED Research Program on Emerging and Re-emerging Infectious Diseases JP19fk018113 (to T.W.) and 19fk0108171 (to S.N.); 2020 Tokai University School of Medicine Research Aid (to S.N.) The super-computing resource was provided by Human Genome Center (the Univ. of Tokyo) and the NIG supercomputer at ROIS National Institute of Genetics.

Author information

Affiliations

Department of Molecular Virology, Research Institute for Microbial Diseases, Osaka University, Osaka, Japan.

Kosuke Takada and Tokiko Watanabe.

Department of Genomic Function and Diversity, Medical Research Institute, Tokyo Medical and Dental University, Bunkyo, Tokyo, Japan

Mahoko Takahashi Ueda.

Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Kanagawa, Japan.

Kosuke Takada and So Nakagawa.

Contributions

K.T., T.W. and S.N. conceived the concept of this study. K.T., M.T.U and S.N analyzed data. K.T, M.T.U and S.N wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

K.T., M.T.U., T.W. and S.N. have no competing interests.

Corresponding author

Correspondence to So Nakagawa <so@tokai.ac.jp>

520 **Figure Legends**

521

522 **Figure 1. Phylogenetic tree of nsp14**

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

527

528 **Figure 2. Amino acid sequence alignment of selected nsp14 derived from representative CoVs**

529 (A) Conserved amino acid residues among each group. The amino acid sequence of nsp14 of the 62
530 representative CoVs is shown according to the amino acid position of SARS-CoV-2, where the gapped
531 positions in the SARS-CoV-2 sequence were not shown. The conserved amino acid positions are
532 indicated in gray. (B) Four conserved active site residues involved in proofreading. The amino acid
533 fractions for each site were calculated based on the sequence alignment of nsp14 of the 62
534 representative CoVs visualized by using WebLogo (Crooks, et al. 2004).

535

536 **Figure 3. Scatter plot of nucleotide substitutions in SARS-CoV-2 genomes**

537 Genome diversity of SARS-CoV-2 containing wild-type (WT) nsp14 and 7 nsp14 mutants is shown.
538 The X-axis indicates the sampling date and the Y-axis indicates the number of nucleotide substitutions
539 in the genomes. The dotted line corresponds to the regression line.

540

541 **Figure 4. Phylogeny and nucleotide substitution rate of nsp14-P203L variants in the Pangolin**
542 **lineages B.1.1.33**

543 (A) ML tree of SARS-CoV-2 genomes in B.1.1.33-lineage. The nsp14-P203L variant is highlighted
544 in orange. A gray circle corresponds to bootstrap values ≥ 70%. (B) Nucleotide substitution rates of

545 nsp14-P203L variants belonging to the B.1.1.33-lineages. All dots of samples are displayed. The
546 dotted line corresponds to the regression line.
547
548 **Figure 5. Frequency of nucleotide substitution rate of SARS-CoV-2 genome sequences in nsp14**
549 **or structural gene variants**
550 The nucleotide substitution rates of SARS-CoV-2 genome sequences were calculated for each variant
551 containing amino acid mutations in the nsp14 or in the S, M, and E genes. We used genomes with 20
552 or more nucleotide substitutions. Frequency of the substitution rate of genome sequences with amino
553 acid replacements for each site in nsp14 or in S, M, and E are shown in dark gray or light gray,
554 respectively. The thin dotted line indicates the median value for each.
555
556 **Figure 6. The three-dimensional position of amino acid position 203 in nsp14**
557 (A) Domain structure of nsp14 with residue numbers. The interaction surface with nsp10 is indicated
558 by the hatched box. The ExoN and N7-Mtase domains are shown in light blue and gray, respectively.
559 The loop from positions 200 to 214 is marked by a dark gray bar. The amino acid at position 203 in
560 nsp14 is shown in magenta. The amino acid positions of magnesium binding (191), nsp10 binding
561 (200 and 201), and zinc binding (207 and 210) in nsp14 are shown in purple, light pink and orange,
562 respectively. (B) Schematic representation of nsp14. The coloring is the same as in (A).
563

564 Table 1. Frequent amino acid sequence of nsp14 in SARS-CoV-2 epidemic strains.

Ranking of appearance frequency	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

565 28,082 sequences that were annotated as Human and free of undetermined or mixed nucleotides, were used in
566 the analysis. Amino acid replacements were defined with hCoV-19/Wuhan/Hu-1/2019 as the reference sequence.
567 —, no mutation was found compared with the reference sequence.

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

Ranking of nsp14	Substitution in nsp14	Number of comparison sequences	Substitution rate per year in 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

569 The number of mismatched nucleotides in the alignment sequences of the reference sequence and the
570 comparison sequence of each group was counted, and the average value was calculated. The hCoV-
571 19/Wuhan/Hu-1/2019 was selected as the reference sequence. —, no mutation was found compared with the
572 sequences from the hCoV-19/Wuhan/Hu-1/2019.

573 Table 3. Nucleotide substitution rates of the 7 nsp14 variants for each single cluster in which the maximum number
574 was observed

Ranking of nsp14	Substitution in nsp14	Lineage (Pangolin ID)	Number of comparison sequences	Substitution
5	L177F	B.1.1.25	54	
7	F377L	B	50	
11	S369F	B.1.1	17	
13	M501I	B.1	11	
16	P203L	B.1.1.33	6	
		B.1	6	
17	L177F, T372I	B.1.113	20	
18	P297S	B.4	11	

575 The number of mismatched nucleotides in the alignment sequences of the reference sequence and the
576 comparison sequence of each group was counted, and the average value was calculated. The hCoV-
577 19/Wuhan/Hu-1/2019 was selected as the reference sequence. A scatter plot of these nsp14 is shown in Fig 4B
578 or Supplementary figure 4.

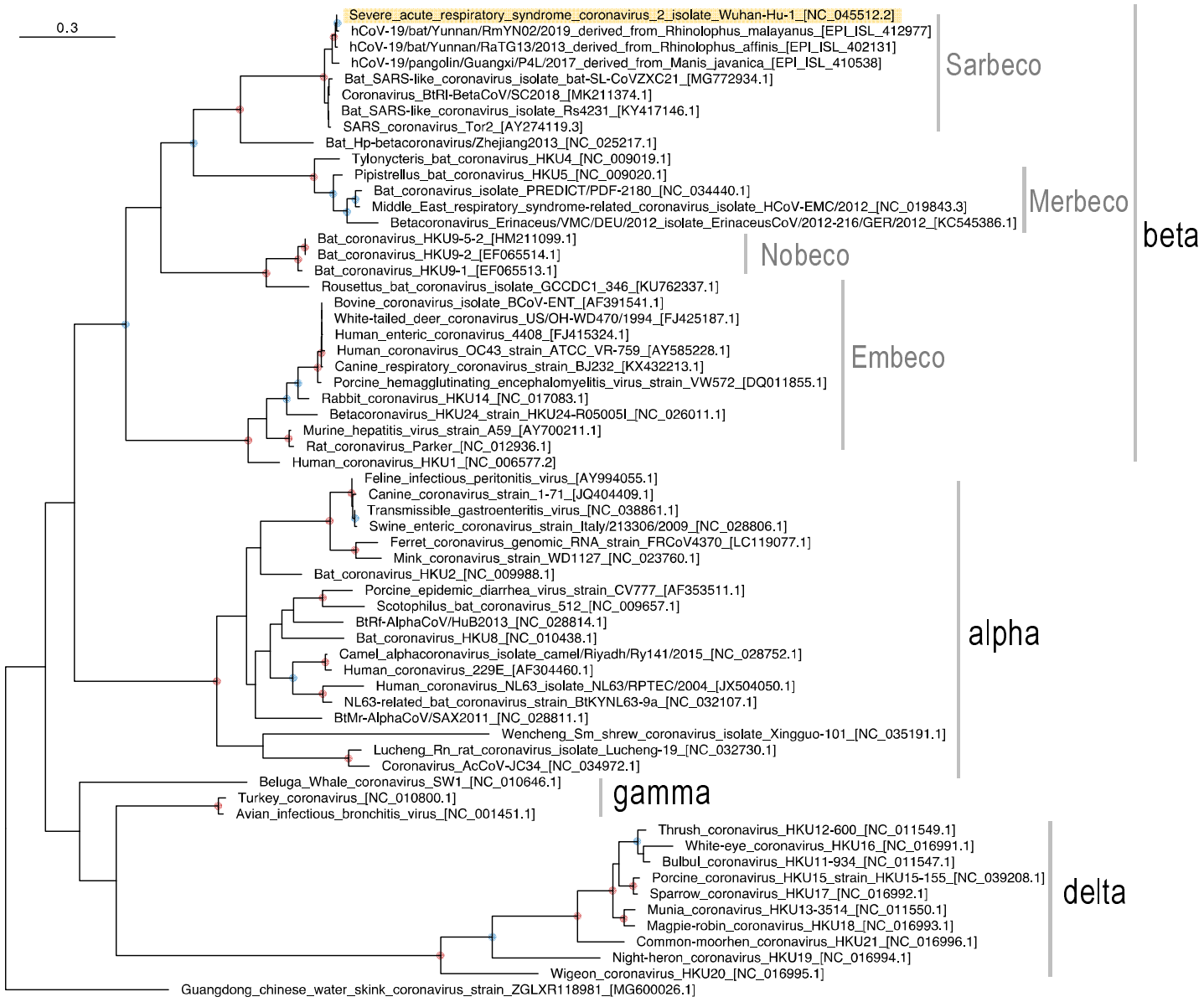
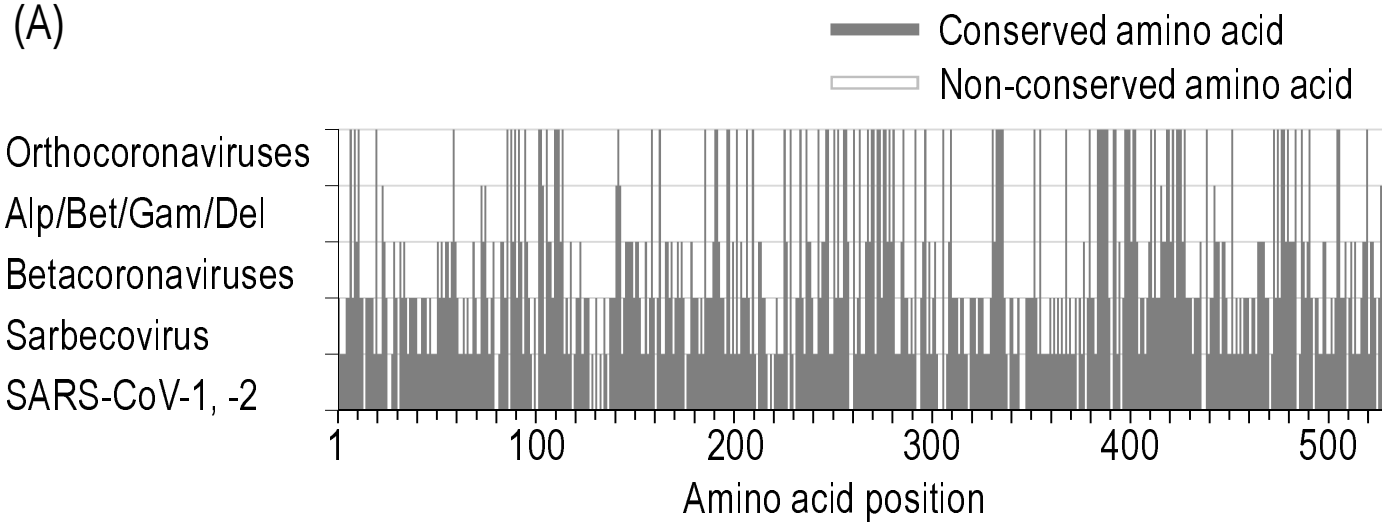

0.3
Severe_acute_respiratory_syndrome_coronavirus_2_isolate_Wuhan-Hu-1_[NC_045512.2]
hCoV-19/bat/Yunnan/RmYN02/2019_derived_from_Rhinolophus_malayanus_[EPI_ISL_412977]
hCoV-19/bat/Yunnan/RaTG13/2013_derived_from_Rhinolophus_affinis_[EPI_ISL_402131]
hCoV-19/pangolin/Guangxi/P4L/2017_derived_from_Manis_javanica_[EPI_ISL_410538]
Bat_SARS-like_coronavirus_isolate_bat-SL-CoVZXC21_[MG772934.1]
Coronavirus_BtRl-BetaCoV/SC2018_[MK211374.1]
Bat_SARS-like_coronavirus_isolate_Rs4231_[KY417146.1]
SARS_coronavirus_Tor2_[AY274119.3]
Sarbeco
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]
Merbeco
Bat_coronavirus_HKU9-5-2_[HM211099.1]
Bat_coronavirus_HKU9-2_[EF065514.1]
Bat_coronavirus_HKU9-1_[EF065513.1]
Nobeco
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]
Embeco
beta
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]
alpha
Beluga_Whale_coronavirus_SW1_[NC_010646.1]
Turkey_coronavirus_[NC_010800.1]
Avian_infectious_bronchitis_virus_[NC_001451.1]
gamma
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]
delta
Guangdong_chinese_water_skink_coronavirus_strain_ZGLXR118981_[MG600026.1]

Figure 1

(A)


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

(B)

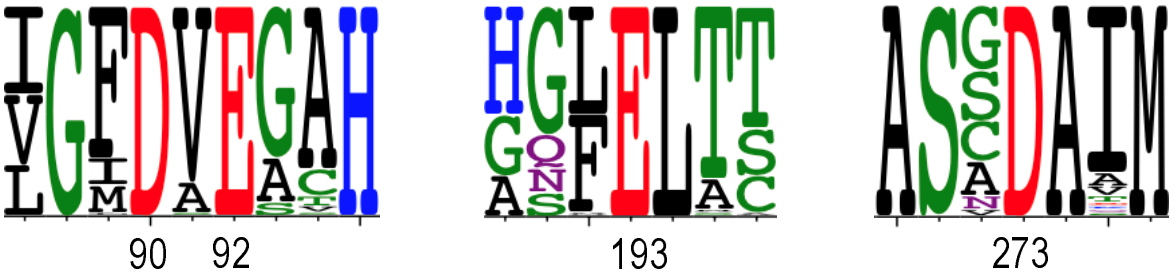

90
92
193
273

Figure 2

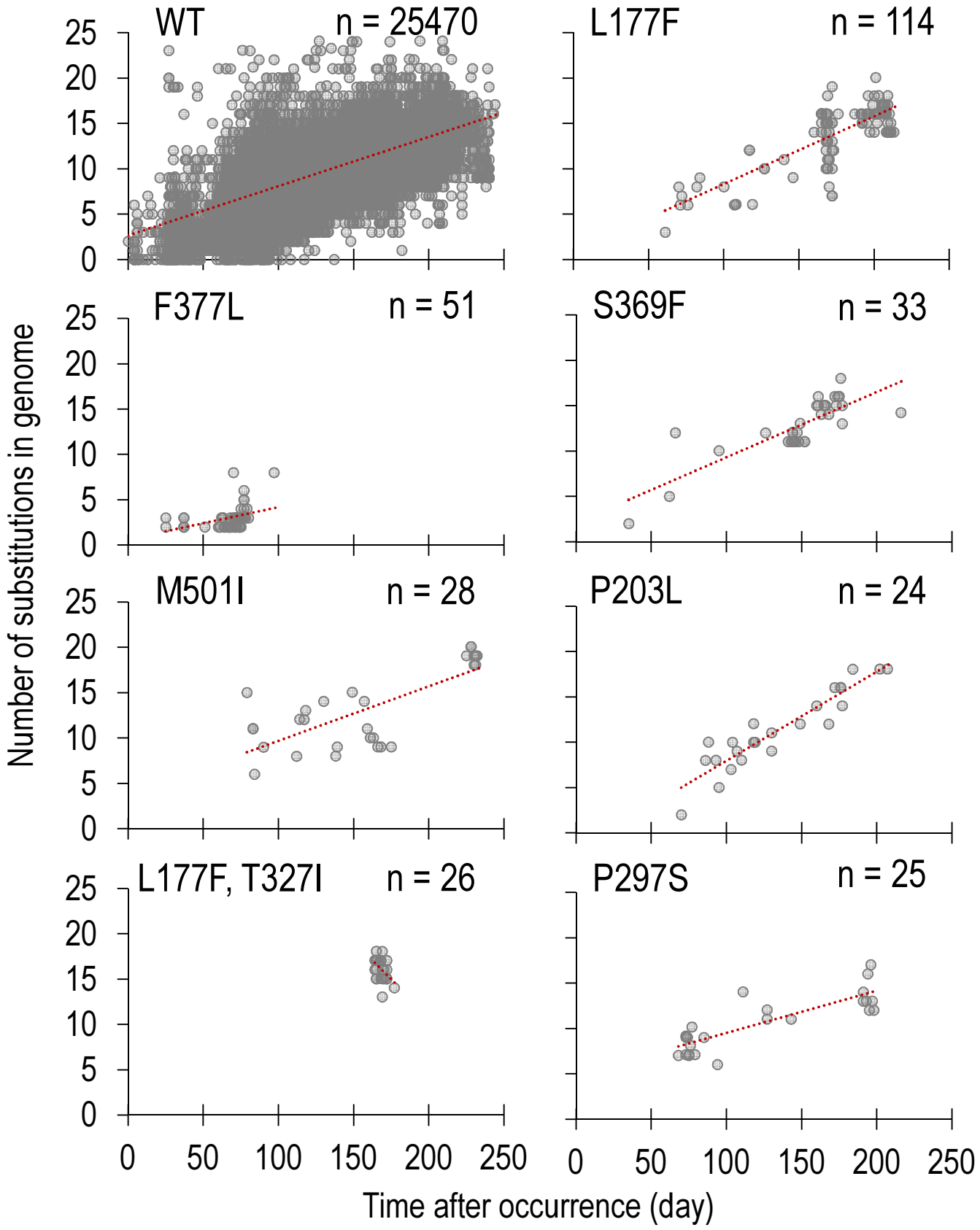

WT
n = 25470
L177F
n = 114
F377L
n = 51
S369F
n = 33
M501I
n = 28
P203L
n = 24
L177F, T327I
n = 26
P297S
n = 25
Number of substitutions in genome
Time after occurrence (day)
0
5
10
15
20
25
0
50
100
150
200
250

Figure 3

Figure 4 continue

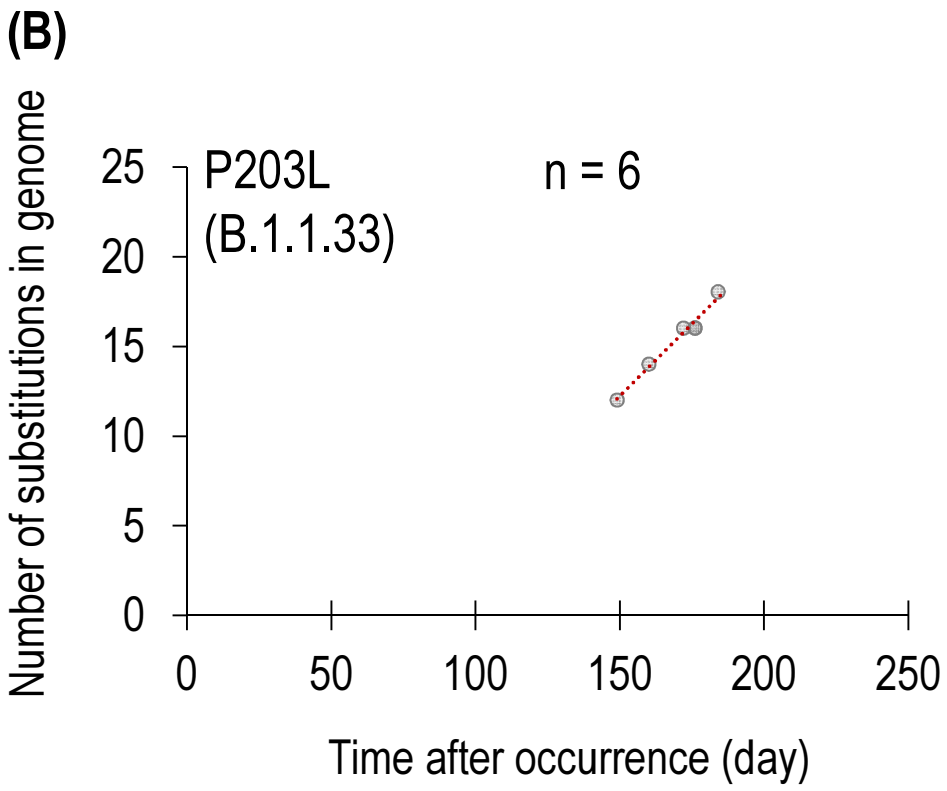

(B)
P203L
(B.1.1.33)
n = 6
Number of substitutions in genome
25
20
15
10
5
0
0
50
100
150
200
250
Time after occurrence (day)

Figure 4

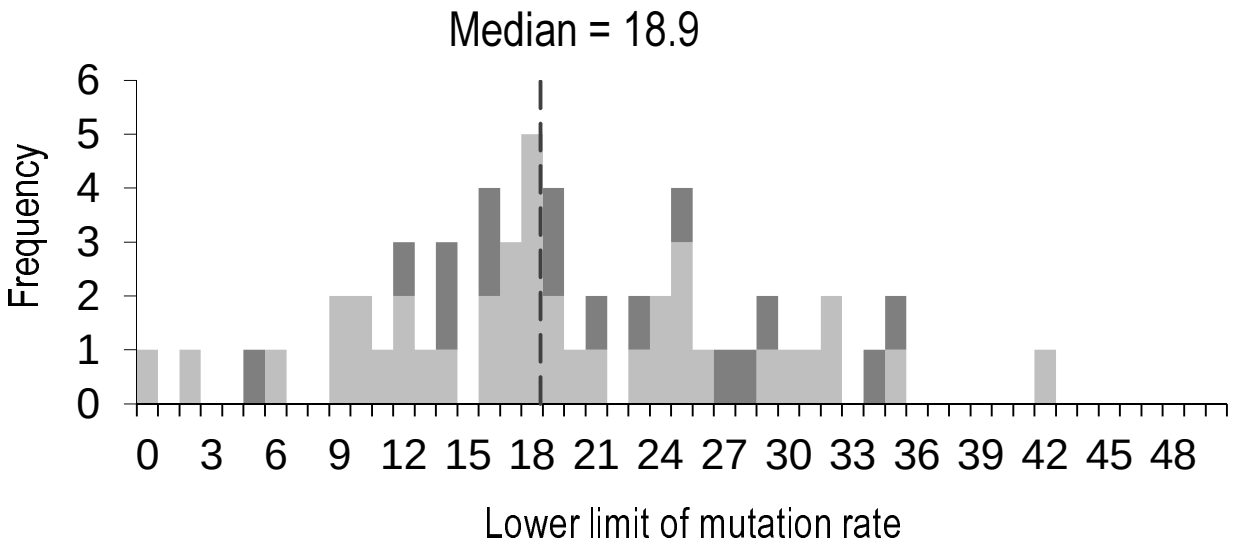
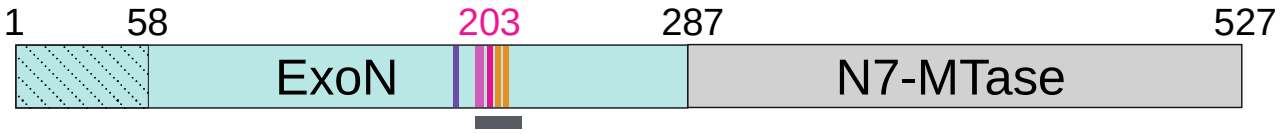

Median = 18.9
Frequency
0
1
2
3
4
5
6
0 3 6 9 12 15 18 21 24 27 30 33 36 39 42 45 48
Lower limit of mutation rate

Figure 5

(A)



(B)

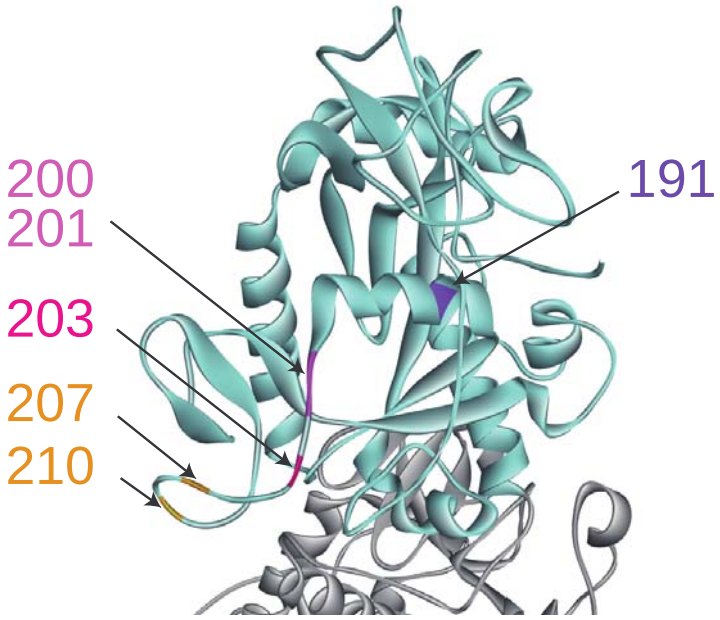


Figure 6