Mutations of SARS-CoV-2 nsp14 exhibit strong association with increased genome-wide mutation load

Doğa Eskier [Equal first author, 1, 2] , **Aslı Suner** [Equal first author, 3] , **Yavuz Oktay** [Corresp., 1, 2, 4] , **Gökhan Karakülah** [Corresp. 1, 2]

[1] Izmir Biomedicine and Genome Center, Izmir, Turkey

[2] Izmir International Biomedicine and Genome Institute, Dokuz Eylül University, Izmir, Turkey

[3] Department of Biostatistics and Medical Informatics, Faculty of Medicine, Ege University, Izmir, Turkey

[4] Faculty of Medicine, Department of Medical Biology, Dokuz Eylül University, Izmir, Turkey

Corresponding Authors: Yavuz Oktay, Gökhan Karakülah
Email address: yavuz.oktay@ibg.edu.tr, gokhan.karakulah@deu.edu.tr

SARS-CoV-2 is a betacoronavirus responsible for human cases of COVID-19, a pandemic with global impact that first emerged in late 2019. Since then, the viral genome has shown considerable variance as the disease spread across the world, in part due to the zoonotic origins of the virus and the human host adaptation process. As a virus with an RNA genome that codes for its own genomic replication proteins, mutations in these proteins can significantly impact the variance rate of the genome, affecting both the survival and infection rate of the virus, and attempts at combating the disease. In this study, we analyzed the mutation densities of viral isolates carrying frequently observed mutations for four proteins in the RNA synthesis complex over time in comparison to wildtype isolates. Our observations suggest mutations in nsp14, an error-correcting exonuclease protein, have the strongest association with increased mutation load in both regions without selective pressure and across the genome, compared to nsp7, 8, and 12, which form the core polymerase complex. We propose nsp14 as a priority research target for understanding genomic variance rate in SARS-CoV-2 isolates, and nsp14 mutations as potential predictors for high mutability strains.

1 **Mutations of SARS-CoV-2 nsp14 exhibit strong association with increased genome-wide**

2 **mutation load**

4 Doğa Eskier[1, 2, *], Aslı Suner[3, *], Yavuz Oktay[1, 2, 4, †], Gökhan Karakülah[1, 2, †]

6 [1]Izmir Biomedicine and Genome Center (IBG), Izmir, Turkey

7 [2]Izmir International Biomedicine and Genome Institute, Dokuz Eylül University, Izmir, Turkey

8 [3]Department of Biostatistics and Medical Informatics, Faculty of Medicine, Ege University,

9 Izmir, Turkey

10 [4]Faculty of Medicine, Department of Medical Biology, Dokuz Eylül University, Izmir, Turkey

11 [*]These authors contributed equally to the work presented here and should therefore be regarded as

12 first authors.

14 [†]Co-corresponding authors:

15 Gökhan Karakülah, E-mail: gokhan.karakulah@deu.edu.tr

16 Yavuz Oktay, E-mail: yavuz.oktay@ibg.edu.tr

Abstract

SARS-CoV-2 is a betacoronavirus responsible for human cases of COVID-19, a pandemic with global impact that first emerged in late 2019. Since then, the viral genome has shown considerable variance as the disease spread across the world, in part due to the zoonotic origins of the virus and the human host adaptation process. As a virus with an RNA genome that codes for its own genomic replication proteins, mutations in these proteins can significantly impact the variance rate of the genome, affecting both the survival and infection rate of the virus, and attempts at combating the disease. In this study, we analyzed the mutation densities of viral isolates carrying frequently observed mutations for four proteins in the RNA synthesis complex over time in comparison to wildtype isolates. Our observations suggest mutations in nsp14, an error-correcting exonuclease protein, have the strongest association with increased mutation load in both regions without selective pressure and across the genome, compared to nsp7, 8, and 12, which form the core polymerase complex. We propose nsp14 as a priority research target for understanding genomic variance rate in SARS-CoV-2 isolates, and nsp14 mutations as potential predictors for high mutability strains.

Keywords: SARS-CoV-2, COVID-19, nsp14, 3'-5' exonuclease, RNA-dependent RNA polymerase, RdRp, mutation rate

Introduction

COVID-19 is an ongoing global pandemic characterized by long-term respiratory system damage in patients, and caused by the SARS-CoV-2 betacoronavirus. It is likely of zoonotic origin, but capable of human-to-human transmission, and since the first observed cases in the Wuhan province of China (Chan et al., 2020; Riou & Althaus, 2020), it has infected over 14 million people, with 612,054 recorded deaths (as of 22 July 2020). In addition to its immediate

41 effects on the respiratory system, its long term effects are still being researched, including

42 symptoms such as neuroinvasion (Li, Bai & Hashikawa, 2020; Wu et al., 2020), cardiovascular

43 complications (Kochi et al., 2020; Zhu et al., 2020), and gastrointestinal and liver damage (Lee,

44 Huo & Huang, 2020; Xu et al., 2020). Due to its high transmissibility, and capacity for

45 asymptomatic transmission (Wong et al., 2020), study of COVID-19 and its underlying pathogen

46 remain a high priority. As a result, the high amount of frequently updated data on viral genomes

47 on databases such as GISAID (Elbe & Buckland-Merrett, 2017) and NextStrain (Hadfield et al.,

48 2018) provides researchers with invaluable resources to track the evolution of the virus as it

49 spreads across the world.

50 SARS-CoV-2 has a linear, single-stranded RNA genome, and does not depend on host proteins

51 for genomic replication, instead using an RNA synthesis complex formed from nonstructural

52 proteins (nsp) coded by its own genome. Four of the key proteins involved in the complex are

53 nsp7, nsp8, nsp12, and nsp14, all of which are formed from cleavage of the polyprotein Orf1ab

54 into mature peptides. Nsp12, also known as RdRp (RNA-dependent RNA polymerase), is

55 responsible for synthesizing new strands of RNA using the viral genome as a template. Nsp7 and

56 nsp8 act as essential co-factors for the polymerase unit, together creating the core polymerase

57 complex (Kirchdoerfer & Ward, 2019; Peng et al., 2020), while nsp14 is an exonuclease which

58 provides error-correcting capability to the RNA synthesis complex, therefore allowing the

59 SARS-CoV-2 to maintain its large size genome (Subissi et al., 2014; Ma et al., 2015; Ogando et

60 al., 2019; Romano et al., 2020). Owing to their role in maintaining replication fidelity and

61 directly affecting the mutation-selection equilibrium of RNA viruses, these proteins are key

62 targets of study in understanding the mutation accumulation and adaptive evolution of the virus

63 (Eckerle et al., 2010; Peng et al., 2020).

64 In our previous study, we examined the top 10 most frequent mutations in the SARS-CoV-2

65 nsp12, and identified that four of them are associated with an increase in mutation density in two

66 genes, the membrane glycoprotein (M) and the envelope glycoprotein (E) (the combination of

67 which is hereafter referred to as MoE, as we previously described), which are not under selective

68 pressure, and mutations in these genes are potential markers of reduced replication fidelity

69 (Eskier et al., 2020a). In this study, we follow up on our previous findings and analyze the

70 mutations in nsps 7, 8, and 14, in addition to nsp12, to identify whether the mutations are

71 associated with a nonselective increase in mutation load or not. We then examine whole genome

72 mutation densities in mutant isolates in comparison to wildtype isolates using linear regression

73 models, in order to understand whether the mutations are associated with potential functional

74 impact. Our findings indicate that mutations in nsp14 are most likely to be predictors of

75 accelerated mutation load increase.

Materials and Methods

76

Genome sequence filtering, retrieval, and preprocessing

77

78 As previously described (Eskier et al., 2020a), SARS-CoV-2 isolate genome sequences and the

79 corresponding metadata were obtained from the GISAID EpiCoV database (date of accession: 17

80 June 2020). We applied further quality filters, including selecting only isolates obtained from

81 human hosts (excluding environmental samples and animal hosts), those sequenced for the full

82 length of the genome (sequence size of 29 kb or greater), and those with high coverage for the

83 reference genome (< 1% N content, < 0.05% unique mutations, no unverified indel mutations).

84 To ensure alignment accuracy, all nonstandard unverified nucleotide masking was changed to N

85 due to the specifications of the alignment software, using the Linux *sed* command, and the

86 isolates were aligned against the SARS-CoV-2 reference genome (NCBI Reference Sequence

87   NC_045512.2, available at https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2) using the

88   MAFFT (v7.450) alignment software (Katoh et al., 2002), using the parameters outlined in the

89   software manual for aligning closely related viral genomes (available at

90   https://mafft.cbrc.jp/alignment/software/closelyrelatedviralgenomes.html). Variant sites in the

91   isolates were annotated using snp-sites (2.5.1), bcftools (1.10.2), and ANNOVAR (release date

92   24 October 2019) software (Wang, Li & Hakonarson, 2010; Page et al., 2016), to identify

93   whether a given mutation was synonymous or nonsynonymous. In addition, the 5' untranslated

94   region of the genome (bases 1-265) and the 100 nucleotides at the 3' end were removed from the

95   alignment and annotation files due to a high number of gaps and unidentified nucleotides. We

96   further removed any sequences with incomplete sequencing location or date data in order to

97   avoid complications in downstream analyses. Following the filters, 29,600 genomes were used

98   for the analyses.

99

100   **Mutation density calculation**

101   Variants were categorized as synonymous and nonsynonymous following annotation by

102   ANNOVAR, with intergenic or terminal mutations being considered synonymous. Gene

103   mutation densities were calculated separately for synonymous and nonsynonymous mutations, as

104   well as the total of SNVs, for each isolate, using a non-reference nucleotides per kilobase of

105   region metric. Mutation densities were calculated for the combined membrane glycoprotein (M)

106   and envelope glycoprotein (E) genes (MoE), the surface glycoprotein gene (S), and the whole

107   genome.

108 **Statistical Analysis**

109 Descriptive statistics for continuous variable days were calculated with mean, standard deviation,

110 median, and interquartile range. Kolmogorov–Smirnov test was used to check the normality

111 assumption of the continuous variables. In cases of non-normally distributed data, the Wilcoxon

112 rank-sum (Mann-Whitney U) test was performed to determine whether the difference between

113 the two MoE status groups was statistically significant. The Fisher's exact test and the Pearson

114 chi-square test were used for the analysis of categorical variables. The univariate logistic

115 regression method was utilized to assess the mutations associated with MoE status in single

116 variables, and then multiple logistic regression method was performed. The final multiple

117 logistic regression model was executed with the backward stepwise method. The relationship

118 between mutation density and time in isolates with mutations of interest, as well as in the group

119 comprising all isolates, was examined via non-polynomial linear regression model and

120 Spearman's rank correlation. A p-value of less than 0.05 was considered statistically significant.

121 All statistical analyses were performed using IBM SPSS version 25.0 (Chicago, IL, USA).

122 **Results and Discussion**

123 **Increases in the mutation load of SARS-CoV-2 are unevenly distributed across its genome**

124 To identify the trends in SARS-CoV-2 mutation load over time, we calculated the average

125 mutation density per day for all isolates for whole genome, S gene, and MoE regions, capping

126 outliers at the 95th and 5th percentile values to minimize the potential effects of sequencing errors

127 (Fig. 1). Our results show that both at the genome level and the S gene, a very strong positive

128 correlation between average mutation density and time. In comparison, MoE has a weak positive

129 correlation, with a wider spread of mean density in early and late periods compared to the

130 genome and the S gene. This is consistent with reduced selective pressure on the M and E genes,

131   as has previously been described (Dilucca et al., 2020b). The top nonsynonymous mutation is

132   23403A>G (in 22271 isolates), responsible for the D614G substitution in the spike protein,

133   followed by the 14408C>T mutation (in 22226 isolates) in the nsp12 region of the Orf1ab gene,

134   causing P323L substitution in the RdRp protein, and the 28144C>T mutation (in 3081 isolates),

135   responsible for the L84S substitution in the Orf8 protein. The most common synonymous

136   mutation is the 8782C>T mutation (in 3047 isolates), and is found on the nsp4 coding region of

137   the Orf1ab gene. For the S gene, the most frequent synonymous mutation is the 23731C>T

138   mutation (in 622 isolates), and the second most common nonsynonymous mutation, after the

139   aforementioned D614G mutation, is 25350C>T (in 215 isolates), responsible for the P1263L

140   substitution. For MoE, the most common synonymous and nonsynonymous mutations are

141   26735C>T (in 341 isolates) and 27046C>T (in 530 isolates), respectively, both of which are

142   found in the M gene, and the latter of which causes T175M amino acid substitution. Other than

143   the D614G mutation, all of the mentioned mutations are C>T substitutions, the prevalence of

144   which in T- or A-rich regions of the SARS-CoV-2 genome have been previously documented

145   (Simmonds, 2020).

146   **Mutations in RNA synthesis complex proteins are associated with higher mutation load**

147   After identifying the increase in mutation load over time, which was more prominent in genes

148   with high functional impact (S, Orf1ab) compared to other structural genes (M, E, N), as seen in

149   Figure 1 and Supplementary Figures 1 and 2, we sought to examine possible associations of

150   variants in proteins involved in SARS-CoV-2 genome replication with the increase. We first

151   identified the five most frequently observed mutations for nsps 7, 8, 12 (also known as RdRp)

152   and 14, four of the proteins cleaved from the Orf1ab polyprotein and are involved in the RNA

153   polymerization, followed by analyzing the association of each mutation with the presence of

154   MoE mutations (hereafter referred to as MoE status) using the chi-square test. 12 out of the 20

155   mutations were found to have a significant association with MoE status (p-value < 0.05) (Table

156   1). Compared to our previous findings on the top 10 nsp12 mutations (Eskier et al. 2020), which

157   was based on an analysis of 11,208 samples as of 5 May 2020, 13536C>T and 13862C>T have

158   increased in rank of appearance, from 6th and 7th to 4th and 5th, respectively, and decreased in p-

159   value to show statistically significant associations. In addition, the 13730C>T mutation have

160   increased in rank of appearance from 4th to 3rd. Out of the other nsps tested, nsp14 was found to

161   have four significant mutations, while nsp7 had two and nsp8 had one.

162   **Effects of geographical location on MoE status**

163   In addition to time and genotype, we also examined the potential association between the

164   location of isolates and MoE status as a possible confounding factor. We first examined whether

165   there is a significant association between location, defined here as continent the isolate was

166   originally obtained, and MoE status. Our results indicate that there is a strong association

167   between location and MoE status, with the highest percentage of MoE present isolates in Asia

168   (14.5%), and the percentage ratio in South America (6.5%) (p-value <0.001). In comparison to

169   our previous findings, South America had a dramatic decrease in MoE present isolate

170   percentage, likely as a result of the increased sequencing efforts (from 118 isolates to 416)

171   removing potential sampling biases or localized founder effects. Africa, Asia, and North

172   America had an increase in MoE present proportion, while Europe, Oceania, and South America

173   showed lowered percentages (Table 2).

174   After observing the potential confounding effect of location on MoE status, we sought to

175   understand whether a location is more or less likely to predict MoE status, using a logistic

176   regression model (Table 3). Comparing each individual region (1) to the other five (0), we found

177   that Asia, Europe, and North and South America are all possible predictors of MoE status (p-

178   value < 0.05), with Asia and Europe 1.697 and 1.184 times as likely to be MoE present as the

179   other regions, and North and South America 0.589 and 0.650 times as likely, respectively.

180   Using these findings, we created different logistic regression models to identify which of the 12

181   mutations are likely to be independent predictors of MoE status (Table 4). In the single variable

182   model, all 12 mutations we previously identified and location were found to be potential

183   predictors (p-value < 0.05). Forming final models including the 12 mutations (Final Model A)

184   and the mutations as well as locations (Final Model B), we observed that the predictor effect of

185   two of the mutations nsp8 12478G>A and nsp14 18998C>T do not appear to be sufficiently

186   independent of the other mutations in Final Model A. After adding the location variable to the

187   Final Model A, location remains a significant predictor, with all five non-reference locations less

188   likely to predict MoE than Asia, the reference location, and nsp12 14805C>T is found to not

189   have a predictor effect independent of location (p-value = 0.073). Following Final Model B, nine

190   mutations appear to have a significant association with MoE status, independent of other

191   variables: 11916C>T, 12073C>T, 13536C>T, 13730C>T, 13862C>T, 14408C>T, 18060C>T,

192   18736T>C, and 18877C>T (p-value < 0.05).

193   **Nsp14 mutations have significant impact on increased genomic mutation density**

194   We then examined the effects of each mutation on genomic mutation density to see whether the

195   relationship between the mutations and MoE status are indicative of a genome-wide trend. Due

196   to selection potentially effecting nonsynonymous mutations differentially, we separated the

197   mutations in the two categories and calculated mutation density separately for each category. Our

198   results show that nsp14 mutations show the most consistent association with mutations between

199   MoE and the whole genome. All three nsp14 mutations (18060C>T, 18736T>C, and 18877C>T)

200 which have a significant association with MoE status also show a similar relationship with

201 genomic mutation density (Fig. 2). 18060C>T (L7L) has the lowest odds ratio for MoE status

202 (Table 4), and while it shows a slower increase in synonymous mutation density compared to

203 wildtype isolates (Fig. 2A), it has a significant impact on faster mutation density increase in

204 nonsynonymous mutations (Fig. 2B). In comparison, 18877C>T (L270L) (Fig. 2C-D) and

205 18736T>C (F233L) (Fig. 2E-F) both show a high prediction capacity for MoE and an increased

206 mutation density. In comparison, mutations in nsp7 (Supp. Figs. 3-4) and nsp12 (Supp. Figs. 5-8)

207 show much lower impact on altered mutation density increase rate. 12073C>T, an nsp7 mutation,

208 displays high divergence from wildtype isolate patterns; however, its low sample size (n = 16)

209 creates a skewed distribution of isolates across time, complicating any potential inference.

210

211

Conclusions

213 Our previous work identified RdRp mutations as contributors to the evolution of the SARS-CoV-

214 2 genome and this study confirmed those findings. Furthermore, we hypothesized that mutations

215 of the other critical components of the viral replication and transcription machinery may have

216 similar effects. Our results implicate nsp14 as a source of increased mutation rate in SARS-CoV-

217 2 genomes. Three of the five most common nsp14 mutations, namely 18060C>T, 18736T>C and

218 18877C>T are associated with increases in both genome-wide mutational load, as well as MoE

219 status, an alternative indicator of mutational rate and virus evolution. Interestingly all three are

220 located within the ExoN domain, which is responsible for the proofreading activity of nsp14;

221 however, only 18736T>C mutation is non-synonymous (F233L), while 18060C>T and

222 18877C>T are synonymous mutations and therefore, only after functional studies it will be

223 possible to understand their effects on viral replication processes.

224 The origins and fates of the three nsp14 mutations are also quite different: Being present in the

225 first case detected in the Washington state of the US in mid-January, 18060C>T mutation has

226 been almost completely confined to the US, as 1,657 of 2,007 isolates (82.6%) originating from

227 the US (https://bigd.big.ac.cn/ncov/variation/annotation/variant/18060, accessed 6 September

228 2020). On the other hand, 18877C>T mutation arising around at the end of January likely in

229 Saudi Arabia and being detected in much less cases (n=893), is still present in many isolates,

230 most frequently in Saudi Arabia (54.1%) and Turkey (37.4%). 18736T>C mutation was first

231 detected in the US at the beginning of March and like the 18060C>T mutation, has almost

232 completely been limited to the US (281/362 or 77.6%). Unlike the other two, this mutation has

233 been detected in only two isolates since 27 May, and not after 1 July 2020. However, it should be

234 noted that 18877C>T mutation arose within the dominant 23403A>G / 14408C>T lineage, while

235 the other two nsp14 mutations are in different lineages. Therefore, dominance or disappearance

236 of different nsp14 mutations may have less to do with these particular mutations and more with

237 the co-mutations. Yet, we cannot rule out possible effects of these nsp14 mutations on the fitness

238 of SARS-CoV-2.

239 Previous studies on alphacoronavirus nsp14 protein had shown that nsp14, via its exonuclease

240 activity, can modulate host-virus interactions, degrading double-stranded RNA produced during

241 genome replication to suppress immune response, thus increasing viral viability (Becares et al.,

242 2016). SARS-CoV-2 nsp14, due to similar exonuclease activity, is therefore a potential

243 modulator of host interactions, independent of its link to increased mutation load. However, the

244 exact effect of the mutations we identified, two of which are synonymous and may only

245   indirectly affect protein structure, have to be studied experimentally to show any possible

246   changes in viral property that they might affect. Of note, a recent study where codon usage of

247   SARS-CoV-2 was analyzed in terms of temporal evolution of the virus genome revealed that

248   nsp14 is one of three genes (together with S and N genes) that display the highest Codon

249   Adaptation Index (CAI) values (Dilucca et al., 2020a). CAI is a measure of optimal codon usage

250   and indicates how well codons adapt to the host. Based on higher CAI values in nsp14, one could

251   speculate that such mutations have been accumulating preferentially to reach the optimal

252   mutation rate that allows the most advantageous mutation-selection equilibrium for SARS-CoV-

253   2. Indeed, our previous results (Eskier et al., 2020b) indicated that the mutation densities of

254   SARS-CoV-2 genomes are closely related to the pandemic stage and population dynamics

255   directly affects the average mutational load of the viral genome. During the rapid growth stages,

256   such as those observed in March in the UK and the US, replication fidelity can be traded off to

257   gain higher replication rates and broader mutational diversity. However, mutations in the

258   replication machinery that result in too high mutation rates would likely be detrimental and

259   eliminated. On the other hand, a small percentage of the resulting mutations could possibly be

260   advantageous, including those that could confer resistance to antiviral drugs. So far, we or others

261   have not been able to detect such mutations advantageous for the virus, however, higher

262   mutation rates make appearance of such a mutation more likely.

263   We believe that the mutations discussed in this study can be of help to future studies, in both

264   fighting the COVID-19 pandemic, and better understanding of how mutations in coronavirus

265   replication proteins can affect viral viability and replication fidelity in hosts. Also, it is yet to be

266   determined whether COVID-19 cases infected with SARS-CoV-2 that has mutation(s) that are

267 associated with higher mutation rate respond better to nucleoside analogs, such as remdesivir or

268 ribavirin.

269

270 **Additional Information and Declarations**

271 **Acknowledgement**

272 The authors would like to thank Mr. Alirıza Arıbaş from Izmir Biomedicine and Genome Center

273 for his technical assistance. The authors also would like to extend their thanks to Izmir

274 Biomedicine and Genome Center (IBG) COVID19 platform IBG-COVID19 for their support in

275 implementing the study and the Scientific and Technological Research Council of Turkey

276 (TUBITAK) for their financial support of IBG-COVID19.

277 **Funding**

278 Yavuz Oktay is supported by the Turkish Academy of Sciences Young Investigator Program

279 (TÜBA- GEBİP). The funders had no role in study design, data collection and analysis, decision

280 to publish, or preparation of the manuscript.

281 **Grant Disclosures**

282 The following grant information was disclosed by the authors:

283 Turkish Academy of Sciences Young Investigator Program (TÜBA- GEBİP).

284 **Competing Interests**

285 Aslı Suner and Gökhan Karakülah are Academic Editors at PeerJ.

286 **Author Contributions**

287 Doğa Eskier, Aslı Suner, Gökhan Karakülah and Yavuz Oktay conceived and designed the

288 experiments, performed the experiments, analyzed the data, prepared figures and/or tables,

289 authored or reviewed drafts of the paper, and approved the final draft.

290 **Data Availability**

291 The data is available at Mendeley: Eskier, Doğa; Suner, Aslı; Oktay, Yavuz; Karakülah, Gökhan

292 (2020), "SARS-CoV-2 GISAID isolates (2020-06-17) genotyping VCF", Mendeley Data, v1.

293 http://dx.doi.org/10.17632/63t5c7xb4c.1

294 **Supplemental Information**

295 Supplemental materials are included with this research.

296 **References**

297 Becares M, Pascual-Iglesias A, Nogales A, Sola I, Enjuanes L, Zuñiga S. 2016. Mutagenesis of

298   Coronavirus nsp14 Reveals Its Potential Role in Modulation of the Innate Immune

299   Response. *Journal of virology* 90:5399–5414. DOI: 10.1128/JVI.03259-15.

300 Chan JF-W, Yuan S, Kok K-H, To KK-W, Chu H, Yang J, Xing F, Liu J, Yip CC-Y, Poon RW-

301   S, Tsoi H-W, Lo SK-F, Chan K-H, Poon VK-M, Chan W-M, Ip JD, Cai J-P, Cheng VC-

302   C, Chen H, Hui CK-M, Yuen K-Y. 2020. A familial cluster of pneumonia associated with

303   the 2019 novel coronavirus indicating person-to-person transmission: a study of a family

304   cluster. *Lancet (London, England)* 395:514–523. DOI: 10.1016/S0140-6736(20)30154-9.

305 Dilucca M, Forcelloni S, Georgakilas AG, Giansanti A, Pavlopoulou A. 2020a. Temporal

306   evolution and adaptation of SARS-COV 2 codon usage. *bioRxiv*:2020.05.29.123976.

307   DOI: 10.1101/2020.05.29.123976.

308   Dilucca M, Forcelloni S, Georgakilas AG, Giansanti A, Pavlopoulou A. 2020b. Codon Usage

309   and Phenotypic Divergences of SARS-CoV-2 Genes. *Viruses* 12:498. DOI:

310   10.3390/v12050498.

311   Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X, Scherbakova S, Graham RL, Baric

312   RS, Stockwell TB, Spiro DJ, Denison MR. 2010. Infidelity of SARS-CoV Nsp14-

313   exonuclease mutant virus replication is revealed by complete genome sequencing. *PLoS*

314   *pathogens* 6:e1000896. DOI: 10.1371/journal.ppat.1000896.

315   Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID's innovative

316   contribution to global health. *Global Challenges* 1:33–46. DOI: 10.1002/gch2.1018.

317   Eskier D, Karakülah G, Suner A, Oktay Y. 2020a. RdRp mutations are associated with SARS-

318   CoV-2 genome evolution. *PeerJ* 8:e9587. DOI: 10.7717/peerj.9587.

319   Eskier D, Suner A, Karakülah G, Oktay Y. 2020b. Mutation density changes in SARS-CoV-2 are

320   related to the pandemic stage but to a lesser extent in the dominant strain with mutations

321   in spike and RdRp. *PeerJ* 8:e9703. DOI: 10.7717/peerj.9703.

322   Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T,

323   Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*

324   34:4121–4123. DOI: 10.1093/bioinformatics/bty407.

325   Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple

326   sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30:3059–

327   3066. DOI: 10.1093/nar/gkf436.

328   Kirchdoerfer RN, Ward AB. 2019. Structure of the SARS-CoV nsp12 polymerase bound to nsp7

329   and nsp8 co-factors. *Nature Communications* 10:2342. DOI: 10.1038/s41467-019-10280-

330   3.

331   Kochi AN, Tagliari AP, Forleo GB, Fassini GM, Tondo C. 2020. Cardiac and arrhythmic

332   complications in patients with COVID-19. *Journal of Cardiovascular Electrophysiology*

333   31:1003–1008. DOI: 10.1111/jce.14479.

334   Lee I-C, Huo T-I, Huang Y-H. 2020. Gastrointestinal and liver manifestations in patients with

335   COVID-19. *Journal of the Chinese Medical Association: JCMA* 83:521–523. DOI:

336   10.1097/JCMA.0000000000000319.

337   Li Y-C, Bai W-Z, Hashikawa T. 2020. The neuroinvasive potential of SARS-CoV2 may play a

338   role in the respiratory failure of COVID-19 patients. *Journal of Medical Virology*

339   92:552–555. DOI: 10.1002/jmv.25728.

340   Ma Y, Wu L, Shaw N, Gao Y, Wang J, Sun Y, Lou Z, Yan L, Zhang R, Rao Z. 2015. Structural

341   basis and functional analysis of the SARS coronavirus nsp14–nsp10 complex.

342   *Proceedings of the National Academy of Sciences of the United States of America*

343   112:9436–9441. DOI: 10.1073/pnas.1508686112.

344   Ogando NS, Ferron F, Decroly E, Canard B, Posthuma CC, Snijder EJ. 2019. The Curious Case

345   of the Nidovirus Exoribonuclease: Its Role in RNA Synthesis and Replication Fidelity.

346   *Frontiers in Microbiology* 10. DOI: 10.3389/fmicb.2019.01813.

347   Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites:

348   rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial genomics*

349   2:e000056–e000056. DOI: 10.1099/mgen.0.000056.

350   Peng Q, Peng R, Yuan B, Zhao J, Wang M, Wang X, Wang Q, Sun Y, Fan Z, Qi J, Gao GF, Shi

351   Y. 2020. Structural and Biochemical Characterization of the nsp12-nsp7-nsp8 Core

352   Polymerase Complex from SARS-CoV-2. *Cell Reports* 31:107774. DOI:

353   10.1016/j.celrep.2020.107774.

354   Riou J, Althaus CL. 2020. Pattern of early human-to-human transmission of Wuhan 2019 novel

355   coronavirus (2019-nCoV), December 2019 to January 2020. *Eurosurveillance* 25. DOI:

356   10.2807/1560-7917.ES.2020.25.4.2000058.

357   Romano M, Ruggiero A, Squeglia F, Maga G, Berisio R. 2020. A Structural View of SARS-

358   CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping.

359   *Cells* 9. DOI: 10.3390/cells9051267.

360   Simmonds P. 2020. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other

361   Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary

362   Trajectories. *mSphere* 5. DOI: 10.1128/mSphere.00408-20.

363   Subissi L, Posthuma CC, Collet A, Zevenhoven-Dobbe JC, Gorbalenya AE, Decroly E, Snijder

364   EJ, Canard B, Imbert I. 2014. One severe acute respiratory syndrome coronavirus protein

365   complex integrates processive RNA polymerase and exonuclease activities. *Proceedings*

366   *of the National Academy of Sciences* 111:E3900–E3909. DOI:

367   10.1073/pnas.1323705111.

368   Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants

369   from high-throughput sequencing data. *Nucleic Acids Research* 38:e164–e164. DOI:

370   10.1093/nar/gkq603.

371   Wong J, Jamaludin SA, Alikhan MF, Chaw L. 2020. Asymptomatic transmission of SARS-CoV-

372   2 and implications for mass gatherings. *Influenza and Other Respiratory Viruses*. DOI:

373   10.1111/irv.12767.

374   Wu Y, Xu X, Chen Z, Duan J, Hashimoto K, Yang L, Liu C, Yang C. 2020. Nervous system

375   involvement after infection with COVID-19 and other coronaviruses. *Brain, Behavior,*

376   *and Immunity* 87:18–22. DOI: 10.1016/j.bbi.2020.03.031.

377   Xu L, Liu J, Lu M, Yang D, Zheng X. 2020. Liver injury during highly pathogenic human

378   coronavirus infections. *Liver International: Official Journal of the International*

379   *Association for the Study of the Liver* 40:998–1004. DOI: 10.1111/liv.14435.

380   Zhu H, Rhee J-W, Cheng P, Waliany S, Chang A, Witteles RM, Maecker H, Davis MM, Nguyen

381   PK, Wu SM. 2020. Cardiovascular Complications in Patients with COVID-19:

382   Consequences of Viral Toxicities and Host Immune Response. *Current Cardiology*

383   *Reports* 22:32. DOI: 10.1007/s11886-020-01292-3.

384

385   **Figure Legends**

386   **Figure 1. The average mutation density per day for genome, S gene, and M and E genes.**

387   (A) The mutation density vs. time for the whole SARS-CoV-2 genome. (B) The mutation density

388   vs. time for the S gene. (C) The combined mutation density vs. time for the M and E genes.

389   Values in y-axis represent the average number of SNVs in the corresponding day, normalized by

390   kilobase of region of interest. SNV counts of genomes are normalized by capping at the 25- and

391   75-percentile values to minimize the effects of potential sequencing or assembly artifacts.

392   Correlation scores are calculated using Spearman rank correlation.

393

394   **Figure 2. The distribution of synonymous and nonsynonymous mutations in isolates**

395   **carrying nsp14 mutations compared to wildtype isolates.** (A-B) Isolates carrying the

396   synonymous 18060C>T mutation (n = 1585). (C-D) Isolates carrying the synonymous

397   18877C>T mutation (n = 893). (E-F) Isolates carrying the nonsynonymous 18736T>C mutation

398   (n=236). Wildtype isolates in all graphs carry the reference nucleotide for the nine positions of

399   interest (11916, 12073, 13536, 13730, 13862, 14408, 18060, 18736, 18877) (n = 5910).

400   Correlation scores are calculated using Spearman rank correlation.

Comparisons of MoE and nsp mutations

Table 1. Comparisons of MoE and nsp mutations.

NSP	Mutations	Values	MoE Absent		MoE Present		Total		p
			n	%	n	%	n	%	
nsp7	11916C>T S3884L	Absent Present	26326 433	98.4 1.6	2833 8	99.7 0.3	29159 441	98.5 1.5	<0.001*
	12076C>T N3937N	Absent Present	26735 24	99.9 0.1	2837 4	99.9 0.1	29572 28	99.9 0.1	0.339
	11919C>T S3885F	Absent Present	26738 21	99.9 0.1	2840 1	100.0 -	29578 22	99.9 0.1	0.717
	12073C>T D3936D	Absent Present	26750 9	100.0 -	2834 7	99.8 0.2	29584 16	99.9 0.1	<0.001*
	11962C>T L3899L	Absent Present	26746 13	100.0 -	2840 1	100.0 -	29586 14	100.0 -	1.000
nsp8	12478G>A M4071I	Absent Present	26757 2	100.0 -	2750 91	96.8 3.2	29507 93	99.7 0.3	<0.001*
	12550G>A L4095L	Absent Present	26697 62	99.8 0.2	2841 -	100.0 -	29538 62	99.8 0.2	-
	12415C>T N4050N	Absent Present	26725 34	99.9 0.1	2841 -	100.0 -	29566 34	99.9 0.1	-
	12557A>G I4098V	Absent Present	26729 30	99.9 0.1	2841 -	100.0 -	29566 30	99.9 0.1	-
	12400C>T L4045L	Absent Present	26734 25	99.9 0.1	2840 1	100.0 -	29574 26	99.9 0.1	0.508
nsp12	14408C>T P4715L	Absent Present	7498 19261	28.0 72.0	702 2139	24.7 75.3	8200 21400	27.7 72.3	<0.001*
	14805C>T Y4847Y	Absent Present	24397 2362	91.2 8.8	2704 137	95.2 4.8	27101 2499	91.6 8.4	<0.001*
	13730C>T A4489V	Absent Present	26238 521	98.1 1.9	2820 21	99.3 0.7	29058 542	98.2 1.8	<0.001*
	13536C>T Y4424Y	Absent Present	26469 290	98.9 1.1	2823 18	99.4 0.6	29292 308	99.0 1.0	0.025*
	13862C>T T4533I	Absent Present	26535 224	99.2 0.8	2833 8	99.7 0.3	29368 232	99.2 0.8	0.001*
nsp14	18060C>T L5932L	Absent Present	25247 1512	94.3 5.7	2768 73	97.4 2.6	28015 1585	94.6 5.4	0.001*
	18877C>T L6205L	Absent Present	26185 574	97.9 2.1	2522 319	88.8 11.2	28707 893	97.0 3.0	0.001*
	18998C>T A6245V	Absent Present	26454 305	98.9 1.1	2836 5	99.8 0.2	29290 310	99.0 1.0	0.001*
	18736T>C F6158L	Absent Present	26751 8	100.0 -	2613 228	92.0 8.0	29364 236	99.2 0.8	0.001*
	19524C>T L6420L	Absent Present	26530 229	99.1 0.9	2825 16	99.4 0.6	29355 245	99.2 0.8	0.102
	Total		26759	100.0	2840	100.0	29600	100.0	

2 *p-value<0.05 was statistically significant.

Distribution of MoE across geographical locations.

*p-value<0.05 was statistically significant.

1   **Table 2.** Distribution of MoE across geographical locations.

Locations	MoE absent		MoE present		Total		p
	n	**%**	**n**	**%**	**n**	**%**	
Asia	2319	85.5	394	14.5	2713	100.0	
Africa	297	90.3	32	9.7	329	100.0	
South America	389	93.5	27	6.5	416	100.0	<0.001*
Europe	14879	89.8	1697	10.2	16576	100.0	
North America	7401	93.4	522	6.6	7923	100.0	
Oceania	1474	89.7	169	10.3	1643	100.0	
Total	26759	90.4	2841	9.6	29600	100.0	

2   *p-value<0.05 was statistically significant.

Logistic regression model of MoE and location on single variables.

Each location was represented as itself (1) and others (0). OR: Odds-Ratio; C.I.: confidence interval, *p-value<0.05 was statistically significant.

1 **Table 3.** Logistic regression model of MoE and location on single variables. Each location was

2 represented as itself (1) and others (0).

Locations	p	OR	95% C.I.
Asia	<0.001*	1.697	1.513 to 1.903
Africa	0.937	1.015	0.703 to 1.465
South America	0.032*	0.650	0.439 to 0.963
Europe	<0.001*	1.184	1.095 to 1.281
North America	<0.001*	0.589	0.533 to 0.650
Oceania	0.330	1.085	0.921 to 1.278

5 OR, Odds-Ratio; C.I.: confidence interval, *p-value<0.05 was statistically significant.

Logistic regression model of MoE on single variables and a final model.

(Final Model A) Logistic regression model of ten mutations on final model. (Final Model B) Logistic regression model of four mutations and location on final model. OR: Odds-Ratio; C.I.: confidence interval; Multiple logistic regression final model was executed on all these statistically significant variables, included together in the model, and selected with backward stepwise method; *p-value<0.05 was statistically significant.

1 **Table 4.** Logistic regression model of MoE on single variables and a final model. (Final Model A) Logistic regression model of ten
2 mutations on final model. (Final Model B) Logistic regression model of four mutations and location on final model.

	Single Variables			Final Model A			Final Model B		
Mutations	**p**	**OR**	**95% C.I.**	**p**	**OR**	**95% C.I.**	**p**	**OR**	**95% C.I.**
Nsp7.11916	<0.001*	0.172	0.085 to 0.346	<0.001*	0.180	0.089 to 0.363	0.001*	0.314	0.154 to 0.641
Nsp7.12076	0.403	1.571	0.545 to 4.530	-	-	-	-	-	-
Nsp7.11919	0.433	0.448	0.060 to 3.334	-	-	-	-	-	-
Nsp7.12073	<0.001*	7.341	2.732 to 19.728	<0.001*	8.108	3.009 to 21.847	<0.001*	9.164	3.311 to 25.361
Nsp7.11962	0.756	0.724	0.095 to 5.540	-	-	-	-	-	-
Nsp8.12478	<0.001*	442.707	108.996 to 1798.139	-	-	-	-	-	-
Nsp8.12550	0.997	-	-	-	-	-	-	-	-
Nsp8.12415	0.998	-	-	-	-	-	-	-	-
Nsp8.12557	0.998	-	-	-	-	-	-	-	-
Nsp8.12400	0.388	0.377	0.051 to 2.780	-	-	-	-	-	-
Nsp12.14408	<0.001*	1.186	1.085 to 1.297	<0.001*	1.310	1.144 to 1.500	<0.001*	1.662	1.435 to 1.926
Nsp12.14805	<0.001*	0.523	0.439 to 0.625	0.007*	0.746	0.603 to 0.923	0.073	0.817	0.655 to 1.019
Nsp12.13730	<0.001*	0.375	0.242 to 0.581	0.002*	0.497	0.317 to 0.778	<0.001*	0.393	0.250 to 0.619
Nsp12.13536	0.026*	0.582	0.361 to 0.938	0.044*	0.611	0.379 to 0.987	0.009*	0.528	0.327 to 0.855
Nsp12.13862	0.002*	0.335	0.165 to 0.678	0.004*	0.355	0.175 to 0.720	0.001*	0.293	0.144 to 0.594
Nsp14.18060	<0.001*	0.440	0.347 to 0.559	0.001*	0.625	0.479 to 0.816	0.001*	1.658	1.244 to 2.209
Nsp14.18877	<0.001*	5.770	5.002 to 6.656	<0.001*	5.543	4.793 to 6.409	<0.001*	6.437	5.483 to 7.557
Nsp14.18998	<0.001*	0.153	0.063 to 0.370	-	-	-	-	-	-
Nsp14.18736	<0.001*	291.773	144.002 to 591.182	<0.001*	368.884	180.195 to 755.153	<0.001*	970.884	469.324 to 2008.453
Nsp14.19524	0.104	0.656	0.395 to 1.091	-	-	-	-	-	-
Location	<0.001*	-	-				<0.001*		
Africa	0.019*	0.634	0.434-0.927	-	-	-	0.017*	0.580	0.391-0.860
South America	<0.001*	0.409	0.273-0.612	-	-	-	<0.001*	0.302	0.198-0.461
Europe	<0.001*	0.671	0.597-0.755	-	-	-	<0.001*	0.681	0.591-0.785
North America	<0.001*	0.415	0.361-0.477	-	-	-	<0.001*	0.228	0.192-0.271
Oceania	<0.001*	0.675	0.557-0.817	-	-	-	<0.001*	0.536	0.428-0.670

3   OR, Odds-Ratio; C.I.: confidence interval; Multiple logistic regression final model was executed on all these statistically significant variables, included together in the model, and

4   selected with backward stepwise method; *p-value<0.05 was statistically significant.

Figure 1

The average mutation density per day for genome, S gene, and M and E genes.

(A) The mutation density vs. time for the whole SARS-CoV-2 genome. (B) The mutation density vs. time for the S gene. (C) The combined mutation density vs. time for the M and E genes. Values in y-axis represent the average number of SNVs in the corresponding day, normalized by kilobase of region of interest. SNV counts of genomes are normalized by capping at the 25- and 75-percentile values to minimize the effects of potential sequencing or assembly artifacts. Correlation scores are calculated using Spearman rank correlation.
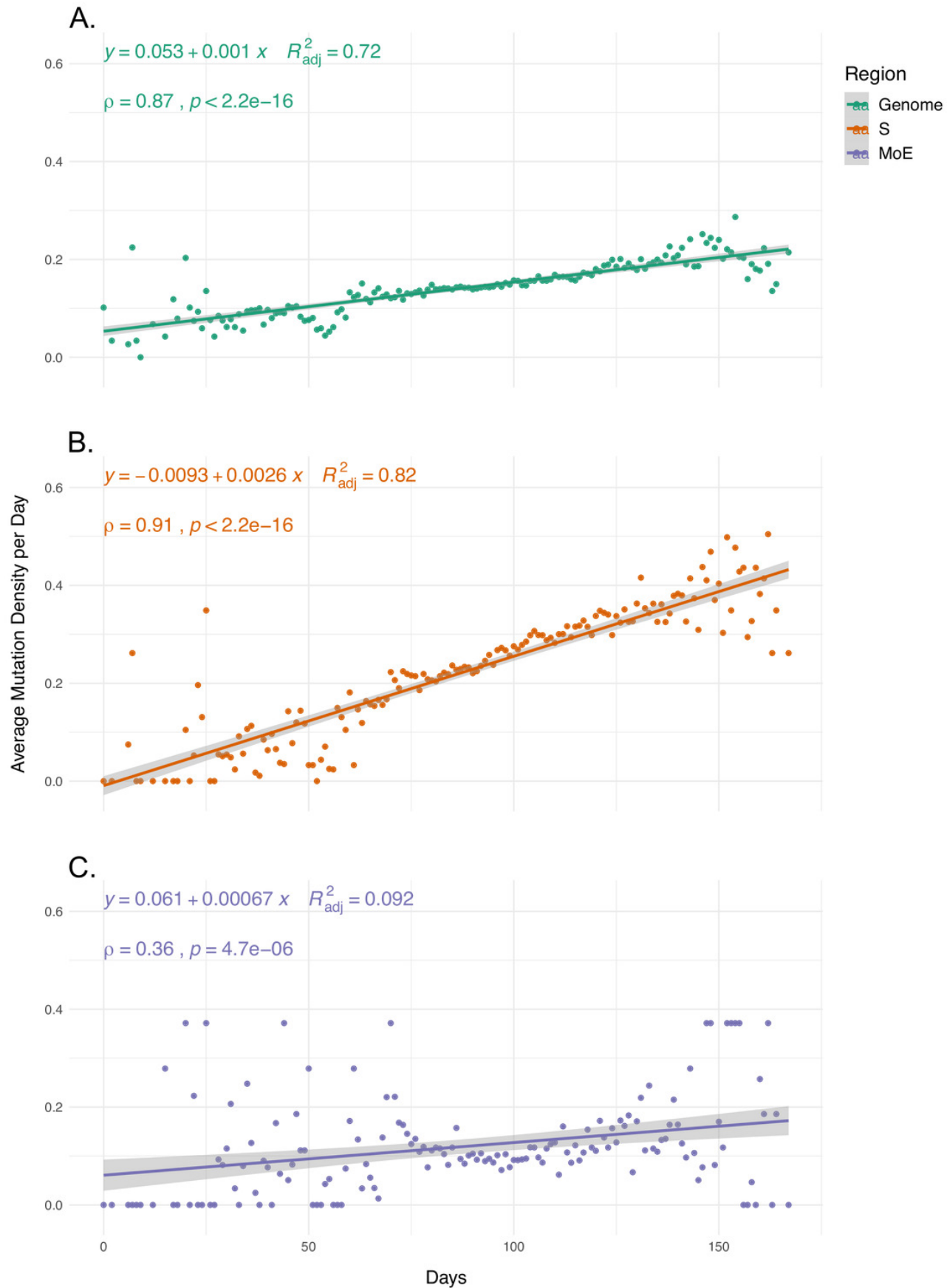
Figure 2

The distribution of synonymous and nonsynonymous mutations in isolates carrying nsp14 mutations compared to wildtype isolates.

(A-B) Isolates carrying the synonymous 18060C>T mutation (n = 1585). (C-D) Isolates carrying the synonymous 18877C>T mutation (n = 893). (E-F) Isolates carrying the nonsynonymous 18736T>C mutation (n=236). Wildtype isolates in all graphs carry the reference nucleotide for the nine positions of interest (11916, 12073, 13536, 13730, 13862, 14408, 18060, 18736, 18877) (n = 5910). Correlation scores are calculated using Spearman rank correlation.

Synonymous

Nonsynonymous

A.

$y = 0.069 + 0.00029\ x$ $R^2_{adj} = 0.31$ $\rho = 0.58$, $p = 3.4e{-}09$

$y = 0.022 + 5 \times 10^{-4}\ x$ $R^2_{adj} = 0.6$ $\rho = 0.79$, $p < 2.2e{-}16$

B.

$y = 0.019 + 0.0012\ x$ $R^2_{adj} = 0.79$ $\rho = 0.89$, $p < 2.2e{-}16$

$y = 0.03 + 0.00055\ x$ $R^2_{ad} = 0.55$ $\rho = 0.77$, $p < 2.2e{-}16$

Genotype
— mutant
— wildtype

C.

$y = -0.033 + 0.001\ x$ $R^2_{adj} = 0.59$ $\rho = 0.76$, $p < 2.2e{-}16$

$y = 0.022 + 5 \times 10^{-4}\ x$ $R^2_{adj} = 0.6$ $\rho = 0.79$, $p < 2.2e{-}16$

D.

$y = 0.018 + 0.00083\ x$ $R^2_{adj} = 0.6$ $\rho = 0.82$, $p < 2.2e{-}16$

$y = 0.03 + 0.00055\ x$ $R^2_{adj} = 0.55$ $\rho = 0.77$, $p < 2.2e{-}16$

E.

$y = 0.031 + 0.00072\ x$ $R^2_{adj} = 0.25$ $\rho = 0.55$, $p = 0.00011$

$y = 0.022 + 5 \times 10^{-4}\ x$ $R^2_{adj} = 0.6$ $\rho = 0.79$, $p < 2.2e{-}16$

F.

$y = 0.029 + 0.0019\ x$ $R^2_{adj} = 0.56$ $\rho = 0.76$, $p = 3.3e{-}09$

$y = 0.03 + 0.00055\ x$ $R^2_{adj} = 0.55$ $\rho = 0.77$, $p < 2.2e{-}16$

Average Mutation Density per Day

Days