Title: RdRp mutations are associated with SARS-CoV-2 genome evolution

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

[1]İzmir Biomedicine and Genome Center (IBG), 35340, İnciraltı, İzmir, Turkey

[2]İzmir International Biomedicine and Genome Institute (iBG-İzmir), Dokuz Eylül University, 35340, İnciraltı, İzmir, Turkey

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

[4]Faculty of Medicine, Department of Medical Biology, Dokuz Eylül University, 35340, İnciraltı, İzmir, Turkey

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

[†]Corresponding author: Yavuz Oktay, E-mail: yavuz.oktay@ibg.edu.tr

Abstract

COVID-19, caused by the novel SARS-CoV-2 virus, started in China in late 2019, and soon became a global pandemic. With the help of thousands of viral genome sequences that have been accumulating, it has become possible to track the evolution of viral genome over time as it spread across the world. An important question that still needs to be answered is whether any of the common mutations affect the viral properties, and therefore the disease characteristics. Therefore, we sought to understand the effects of mutations in RNA-dependent RNA polymerase (RdRp), particularly the common 14408C>T mutation, on mutation rate and viral spread. By focusing on mutations in the slowly evolving M or E genes, we aimed to minimize the effects of selective pressure. Our results indicate that 14408C>T mutation increases the mutation rate, while the third-most common RdRp mutation, 15324C>T, has the opposite effect. It is possible that 14408C>T mutation may have contributed to the dominance of its co-mutations in Europe and elsewhere.

Keywords: SARS-CoV-2, COVID-19, RdRp, mutation rate, RNA-dependent RNA polymerase

Introduction

SARS-CoV-2 is a novel betacoronavirus originally identified in December 2019, and given the official name on 11 February 2020. It is responsible for the ongoing COVID-19 pandemic, with the earliest known patients located potentially as early as November 2019, in the Hubei province of China. Human to human transmission of the virus was confirmed on 20 January 2020[1], with 6 deaths and 282 confirmed cases on 21 January 2020, which, as of 13 May 2020, had respectively increased to over 283 thousand deaths and 4.09 million cases, with a projected mortality rate of < 7%, and an R0 number estimation of 1.4 – 3.8[2]. Due to its high transmission rate, and worldwide distribution of known cases, SARS-CoV-2 is a high priority for medical research despite the low mortality rate. Furthermore, new COVID-19 symptoms continue to be discovered in even recovering patients[3,4], making it difficult to fully understand the global impact of the disease. To date, there is no known targeted treatment or vaccine for SARS-CoV-2.

SARS-CoV-2 has a 29903-nucleotide long single stranded sense RNA genome, which codes for 12 peptides, including two closely related polyproteins, Orf1a and Orf1ab, which are further cleaved into 26 mature peptides, and the main structural proteins, such as the surface glycoprotein, nucleocapsid phosphoprotein, membrane glycoprotein, and envelope glycoprotein (genes S, N, M, and E, respectively)[5]. The primary binding target of SARS-CoV-2 surface glycoprotein for entry into the human cell is the ACE2 protein[6], localized in the cell membrane in a number of tissues[7]. It is predicted that the virus is zoonotic in origin, and a mutation in the surface glycoprotein structure enabled transmission to human hosts. As a result of the origins of the disease, and due to the targeted nature of vaccine and drug discovery efforts, identifying the replication-related mutation rate and the global mutatome of SARS-CoV-2 is crucial to efforts in combating the disease. Despite having proof-reading capability, analyses of SARS-CoV-2 genomes indicated nucleotide substitution rates comparable to other RNA viruses that lack such capability[8] It is difficult to pinpoint the underlying causes without functional studies; however, one plausible explanation would be reduced fidelity of the main RNA polymerase, namely RNA-dependent RNA polymerase (RdRp, also known as nsp12), due to mutations.

SARS-CoV-2 does not depend on host polymerases to replicate its genome, instead using the RdRp and associated proteins (i.e. nsp7, nsp8, and nsp14, the latter an exonuclease with error correction capabilities), all of which are encoded by its own genome[9,10].. Among the SARS-CoV-2 isolates from all over the world, several widespread mutations on the RdRp coding region of the Orf1ab polyprotein gene have been identified. Key among them is the 14408C>T transition, identified in over 7000 isolates across multiple continents. In an early analysis of 137 SARS-CoV-2 genomes from North America and Europe, Pachetti et al. suggested that this particular mutation is associated with higher number of mutations, through a mechanism not yet fully unknown[11]. The proline to leucine substitution (P323L) caused by the 14408C>T mutation has been suggested to rigidify the RdRp protein structure, which may exert its effects through altered interaction with other components of the replication / transcription machinery or with the RNA template, and thereby resulting in an altered mutation rate[11,12]. However, further studies are needed to test these hypotheses.

Although the higher number of mutations in genomes with RdRp 14408C>T mutation was suggested to be caused by lower fidelity of the mutant enzyme, it could also be due to many other epidemiological factors. Importantly, natural selection acting on different levels and ways in different environments could potentially lead to faulty estimates of any change in the replicational mutation rate. Therefore, in order to minimize the effects of natural selection, we focused our attention on parts of the genome that may be under lower selective pressure.

A recent study by Dilucca et al. showed that different SARS-CoV-2 genes are under varying levels of selective pressure, with M and E integral proteins being subject to relatively low natural selection and a low, non-selective mutation rate, largely as a result of accumulation of replication errors. On the other hand, key proteins for virulence and transmissibility, such as the S protein, seem to be under high selective pressure, possibly as a result of novel host adaptation[13]. To identify how the RdRp mutations affect the mutation rate of the SARS-CoV-2 genome, we examined the relationships of the RdRp mutations with mutations found in M or E proteins (hereafter referred to as MoE), in terms of both time and location. In particular, we focused on the 10 most common mutations in the RdRp region, with the goal of identifying whether each variant is associated with increased or decreased non-selective mutation rates, and whether the geographical distribution of these variants might suggest the presence of multiple forms of virus with various mutability across the globe.

Materials and Methods

Genome sequence filtering, retrieval, and preprocessing

SARS-CoV-2 isolate genome sequences were obtained from the GISAID EpiCoV database ([14]). The genomes were filtered for those obtained from human hosts, a sequence length of at least 29000 nucleotides, and high coverage (< 1% undefined nucleotides, < 0.05% mutations not seen in any other isolate, and no indel mutations that were not verified by the submitter). The filters resulted in a total of 11901 remaining genomes (as of May 5th, 2020). The genomes were aligned against the reference genome sequence obtained from the NCBI Nucleotide database in the FASTA format, under the locus ID NC_045512.2 (https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2), after nonstandard unresolved base calls (sequence characters which are not one of A, C, G, T, N, or -) were changed into the standard unresolved sequence character N via the Unix *sed* command. The alignment was performed with the MAFFT multiple sequence alignment program, using the "--auto --keeplength --addfragments isolate_genomes.fa reference_genome.fa > alignment.fa" parameters. The sites differing from the reference sequence were extracted using snp-sites (https://github.com/sanger-pathogens/snp-sites), with the "-v -o variants.vcf alignment.fa" options. The resulting VCF file was modified for compatibility with the following steps using text editing and bcftools (http://www.htslib.org/download/), replacing the first column, indicating reference sequence name, with NC_045512v2, and separating different variants at the same nucleotide to individual lines, using the VCF processing guide available in the ANNOVAR documentation (https://doc-openbio.readthedocs.io/projects/annovar/en/latest/articles/VCF/). The final VCF file was converted into an avinput file, using convert2annovar.pl found under ANNOVAR, with the parameters "-format vcf4old variants.vcf > variants.avinput". The custom ANNOVAR gene annotations for SARS-CoV-2 were obtained from ANNOVAR resources, decompressed, and placed in the sarscov2db directory. The variants were then annotated in terms of their relationships to gene loci and products, using the table_annovar.pl function of ANNOVAR, with the parameters "-buildver NC_045512v2 variants.avinput sarscov2db/ -protocol avGene -operation g".

Following the alignment and annotation, the 5' untranslated region of the genome (bases 1-265) and the 100 nucleotides at the 3' end were removed from analysis due to lack of quality sequencing in a majority of isolates. To ensure a vigorous examination of the association of both time and location and the mutations, we have further filtered out isolate genomes without well-defined time of sequencing metadata (year – month – day), and an undefined geographical location, for a final count of 11208 genomes.

Statistical Analysis

Descriptive statistics for continuous variable days were calculated with mean, standard deviation, median, and interquartile range. Shapiro-Wilk test was used to check the normality assumption of the continuous variable. In cases of non-normally distributed data, the Wilcoxon rank-sum (Mann-Whitney U) test was performed to determine whether the difference between the two MoE groups was statistically significant. The Fisher's exact test and the Pearson chi-square test were used for the analysis of categorical variables. The univariate logistic regression method was utilized to assess the mutations associated with MoE in single variables, and then multiple logistic regression method was performed. The final multiple logistic regression model was executed with the backward stepwise method. A p-value of less than 0.05 was considered statistically significant. All statistical analyses were performed using IBM SPSS version 25.0 (Chicago, IL, USA).

Results and Discussion

Mutation profile of SARS-CoV-2 genome as of 5 May 2020

After the low quality filters were applied, 5658 nucleotides, making up 18.9% of the SARS-CoV-2 genome, were found to carry a mutation in at least one isolate, with 2668 of these sites being mutated in multiple isolates. The sites mutated in at least two isolates had a mean of 26.25 mutated isolates, and a median of 3, To identify the distribution of common mutations by the number of isolates they are found in, we identified the top 50 mutated sites and the number of isolates with a non-reference resolved base in those sites (Fig. 1). Three nucleotides, 3037, 14408, and 23403, were found to be mutated in over 7000 isolates. Out of these three, 14408C>T was previously established as a mutation of interest for the RdRp gene. 23403A>G is a nonsynonymous mutation in the surface glycoprotein, while 3037 is a synonymous mutation in nsp3, a replication scaffolding protein[15]. Notably, only 26 of the top 50 sites were found in the Orf1ab coding region, despite it comprising 71% of the SARS-CoV-2 genome, with the percentage being 57%, 61%, and 65.3% when we consider top 100, top 200, and all mutated sites, respectively (data not shown). Other than 14408C>T, two of the top 50 mutations were also in the RdRP coding sequence, although both of them are synonymous mutations that do not presumably affect the protein structure. Two of the sites were found in the membrane glycoprotein coding region, with 26729T>C being a synonymous mutation, and 27046C>T being a nonsynonymous mutation causing a T175M mutation in the peptide sequence. None of the top 50 mutated sites were found in the envelope glycoprotein region, which has only 23 sites mutated in multiple isolates.

We then examined the distribution of mutant RdRp, envelope, and membrane protein genes by geographical location (Africa, Asia, Europe, North America, Oceania, and South America), in order to see what percentage of isolates from each region carried a mutation in these coding regions. South America had the highest percentage of RdRp mutants, 93.22%, while Asia had only 32.71%, the lowest among the regions. South America also had the highest number of mutant isolates for the M gene, and the second highest for the E gene, at 11.02% and 2.54% respectively (Fig. 2).

Associations between MoE and top ten frequently observed mutations in RdRp

To examine how the most common mutations in RdRp affect mutation rate of the SARS-CoV-2 genome, we identified the 10 most frequently mutated nucleotides in the RdRp region. Table 1 summarizes the frequencies and comparisons of MoE between RdRp mutants. There are statistically significant associations between the mutations at nucleotides 14408, 14805, 15324, 13730 and MoE (p<0.05). However, our statistical analysis indicates that there are no significant associations between the mutations 14786, 13536, 13862, 13627, 14877, 15540 and MoE (p>0.05).

Table 1. Comparisons of MoE and RdRp mutations.

Mutations	**Values**	**MoE Absent**		**MoE Present**		**Total**		**p**
		n	**%**	**n**	**%**	**n**	**%**	
14408	**Absent**	3905	38.4	288	27.7	4193	37.4	<0.001*
	Present	6262	61.6	753	72.3	7015	62.6	
	Total	10167	100.0	1041	100.0	11208	100.0	
14805	**Absent**	9113	89.6	982	94.3	10095	90.1	<0.001*
	Present	1054	10.4	59	5.7	1113	9.9	
	Total	10167	100.0	1041	100.0	11208	100.0	
15324	**Absent**	9866	97.0	1037	99.6	10903	97.3	<0.001*
	Present	301	3.0	4	0.4	305	2.7	
	Total	10167	100.0	1041	100.0	11208	100.0	
13730	**Absent**	10074	99.1	1039	99.8	11113	99.2	0.011*
	Present	93	0.9	2	0.2	95	0.8	
	Total	10167	100.0	1041	100.0	11208	100.0	
14786	**Absent**	10089	99.2	1038	99.7	11127	99.3	0.085
	Present	78	0.8	3	0.3	81	0.7	
	Total	10167	100.0	1041	100.0	11208	100.0	
13536	**Absent**	10112	99.5	1040	99.9	11152	99.5	0.060
	Present	55	0.5	1	0.1	56	0.5	
	Total	10167	100.0	1041	100.0	11208	100.0	
13862	**Absent**	10128	99.6	1035	99.4	11163	99.6	0.349
	Present	39	0.4	6	0.6	45	0.4	
	Total	10167	100.0	1041	100.0	11208	100.0	
13627	**Absent**	10130	99.6	1040	99.9	11170	99.7	0.256
	Present	37	0.4	1	0.1	38	0.3	
	Total	10167	100.0	1041	100.0	11208	100.0	
14877	**Absent**	10129	99.6	1041	100.0	11170	99.7	-
	Present	38	0.4	-	-	38	0.3	
	Total	10167	100.0	1041	100.0	11208	100.0	
15540	**Absent**	10130	99.6	1040	99.9	11170	99.7	0.256
	Present	37	0.4	1	0.1	38	0.3	
	Total	10167	100.0	1041	100.0	11208	100.0	

*p-value<0.05 was statistically significant.

We also tested for possible associations between MoE and time (in days). Mean days in the MoE absent and present groups were 85.80±15.56, and 85.77±15.61, respectively. The days variable in the MoE absent group was not statistically significantly higher than the MoE present group (Median (IQR): 88.00 (15) versus 86.00 (18); p = 0.095). Because it was not statistically significant, we did not include 'days' in the logistic regression models.

Associations between MoE and geographic locations

Distribution of SARS-CoV-2 mutations show variability among geographical locations, mainly due to founder effects, as well as various other epidemiological factors. In order to compare the distribution of MoE among different geographic locations, Table 2 shows that there are statistically significant associations between the locations and MoE (p<0.001). The most frequently observed location for the MoE is Europe (n=658), however, it is largely due to higher representation of European viral genomes in the GISAID database. The highest proportion of MoE is seen in South America (12.7), whereas North America has the lowest (5.2%).

Table 2. Distribution of MoE across geographical locations.

Locations	**Moe Absent**		**Moe Present**		**Total**		**p**
	n	**%**	**n**	**%**	**n**	**%**	
Africa	65	91.5	6	8.5	71	100.0	<0.001*
Asia	790	92.0	69	8.0	859	100.0	
Europe	5111	88.6	658	11.4	5769	100.0	
North America	3194	94.8	176	5.2	3370	100.0	
Oceania	904	88.5	117	11.5	1021	100.0	
South America	103	87.3	15	12.7	118	100.0	
Total	10167	90.7	1041	9.3	11208	100.0	

*p-value<0.05 was statistically significant.

Logistic regression models of the MoE

Next, we evaluated location in the univariate logistic regression models of the MoE (absent (0) and present (1)) for each location (itself (1) and others (0)) (Table 3). Europe, North America and Oceania were found statistically significant to predict MoE (p<0.05). In these locations, while the odds ratio for Europe was 1.700 (95 % CI, 1.490-1.939; p<0.001), odds ratios for North America and the Oceania were 0.444 (95 % CI, 0.376-0.525; p<0.001) and 1.297 (95 % CI, 1.058-1.591; p=0.012) for the presence of MoE. Thus, our results suggest that SARS-CoV-2 genomes in Europe are 1.7 times, and genomes in the Oceania are almost

1.3 times more likely, while those in North America are >2.2 times less likely, to have MoE than other locations.

Table 3. Logistic regression model of MoE and location on single variables. Each location was represented as itself (1) and others (0).

Locations	p	OR	95% C.I.
Africa	0.807	0.901	0.389 to 2.084
Asia	0.188	0.843	0.653 to 1.087
Europe	<0.001*	1.700	1.490 to 1.939
North America	<0.001*	0.444	0.376 to 0.525
Oceania	0.012*	1.297	1.058 to 1.591
South America	0.200	1.428	0.828 to 2.465

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

In the univariate logistic regression models of the MoE (absent (0) and present (1)), when the ten mutations were included separately in the models, 14408, 14805, 15324, and 13730 were found statistically significant to predict MoE (p<0.05) (Table 4). In the final model (Final Model A), significant associations were also detected between MoE and these four mutations (p<0.05). In the final model of the four mutations, the odds ratio for 14408 was 1.522 (95 % CI, 1.305-1.776; p<0.001) for the MoE. Thus, our results suggest that SARS-CoV-2 genomes with the 14408C>T mutation are 1.5 times more likely to have MoE. We also evaluated ‘location’ in the univariate logistic regression models of the MoE, and found that it was statistically significant to predict MoE (p>0.001). Therefore, the final model of logistic regression analysis for independent variables 14408, 14805, 15324, 13730 and ‘location’ (Asia is the reference group) was then built to evaluate their associations with MoE (Final Model B). This final analysis revealed that the same four mutations (14408, 14805, 15324, 13730) and ‘location’ were significantly associated with MoE (p<0.05). For the location, Europe, North America and Oceania were found to be statistically significant on single variable model for the location to predict MoE (p=0.003, p=0.002, and p=0.014, respectively), similarly same locations were also statistically significant in the final model (p=0.036, p<0.001, and p=0.026, respectively). In the final model of four mutations and location (Final Model B), our results indicated that viral genomes in Europe are 1.35 times, genomes in Oceania are 1.45 more likely to have MoE than genomes in Asia. These results indicate that both RdRp mutations and location independently predict MoE status.

Table 4. Logistic regression model of MoE on single variables and a final model. (Final Model A) Logistic regression model of four 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.**
n14408	<0.001*	1.630	1.415 to 1.878	<0.001*	1.522	1.305 to 1.776	0.004*	1.282	1.082 to 1.519
n14805	<0.001*	0.519	0.396 to 0.681	0.008*	0.673	0.502 to 0.903	<0.001*	0.478	0.352 to 0.648
n15324	<0.001*	0.126	0.047 to 0.340	<0.001*	0.108	0.040 to 0.290	<0.001*	0.089	0.033 to 0.240
n13730	0.028*	0.209	0.051 to 0.847	0.049*	0.243	0.060 to 0.992	0.025*	0.201	0.049 to 0.820
n14786	0.095	0.374	0.118 to 1.186	-	-	-	-	-	-
n13536	0.086	0.177	0.024 to 1.279	-	-	-	-	-	-
n13862	0.352	1.505	0.636 to 3.564	-	-	-	-	-	-
n13627	0.188	0.263	0.036 to 1.921	-	-	-	-	-	-
n14877	0.998	0.000	0.000 to -	-	-	-	-	-	-
n15540	0.188	2.263	0.036 to 1.921	-	-	-	-	-	-
Location	<0.001*	-	-				<0.001*		
Africa	0.901	1.057	0.442-2.527	-	-	-	0.574	1.292	0.528-3.160
South America	0.092	1.667	0.920-3.023	-	-	-	0.263	1.416	0.770-2.605
Europe	0.003*	1.474	1.138-1.910	-	-	-	0.036*	1.353	1.020-1.794
North America	0.002*	0.631	0.473-0.842	-	-	-	<0.001*	0.537	0.398-0.725
Oceania	0.014*	1.482	1.084-2.025	-	-	-	0.026*	1.448	1.045-2.008

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.

Whereas the 14408C>T mutation predicted higher risk of MoE, the other three significant mutations in RdRp predicted a lower risk, particularly the 15324C>T mutation, which predicted about 10-fold reduced risk of MoE. Although location was another predictor of MoE, as expected, multivariate logistic regression analysis indicated that the association between RdRp mutations and mutational status of M or E genes was independent from location.

Two of the four significant RdRp mutations were first detected around the same time, 15324 on January 22, and 14408 on January 24, both in Asia. 14805 was first detected in a European genome on February 9, and the most recent of the four, 3730 was detected in an Asian genome on March 4. Despite arising within a few days interval (based on first genomes which have been detected in so far), 14408 and 15324 display >20-fold difference in their spread: 14408 (n=7015 genomes) vs. 15324 (n=305 genomes). Although this observation may be explained by better adaptation of the viruses with a mutation that cause increased mutation rate to changing environments, it could as well be explained by founder effects, genetic drift, and other epidemiological factors. More data and particularly functional studies where mutant viruses can be compared side by side will be required to test this hypothesis.

Our observation that the two different mutants of RdRp result in ~14-fold difference in the likelihood of having a mutation in parts of the genome that evolves relatively slow and under less selective pressure (M and E genes) supports the hypothesis that mutations of RdRp contribute significantly to the SARS-CoV-2 genome evolution. A mutant RdRp that is more error-prone would be expected to increase viral genetic diversity and allow the virus spread under different selective pressures, such as spreading to different populations. As lower-fidelity is also associated with higher speed, such mutations may also allow higher titers of virus within host cells. On the other hand, a higher-fidelity polymerase would be suitable where optimal conditions are reached and errors in replication would be costly. Although preliminary studies suggest that 14408C>T (P323L in RdRp) could lower replication fidelity, it is less clear how the synonymous 15324C>T mutation could lead to lower mutation rates. It should be noted that 288 of 305 (94.4%) genomes worldwide with the 15324C>T mutation also have the 14408C>T mutation, and MoE rate is 1.39% (4/288) among double mutants, whereas it is 11.13% (749/6727) for 14408C>T-only mutants. It is possible that 15324C>T modulates the interaction of viral genome with host factors and indirectly affect the 14408C>T mutation through such factors; as there are currently only 305 genome sequences available with this mutation, this question may be better answered as more viral genome sequences accumulate and functional studies are performed.

Three other mutations that co-evolved and seen together with 14408C>T are 23403A>G (D614G in S protein) and 3037C>T (F106F in NSP3). The first 14408C>T mutation dates to a patient whose sample was collected on January 24 in China, but sequenced and submitted to GISAID on April 10. However, it took 27 days for the second case with the same mutation to appear, and interestingly, not in China, but in Italy. Two days later, on February 22, the first case with 14408C>T was reported in Australia, 31 days after the first SARS-CoV-2 case in the country. Following its introduction to Europe, it took another 10 days for the emergence of the second case in Asia, which can be possibly attributed to strict measures taken by the authorities that led to a steep decline in viral spread particularly in China. In contrast to Europe, North and South America, where 14408C>T became the dominant form together with its co-mutations (23403A>G and 3037C>T), 14408C>T and its co-mutations remained as the minor form in Asia, 14408C>T being present in only 15.9% (137/859) of viral genomes. Emergence of 14408C>T in South America, North America and Africa was 5, 7 and 8 days following the first European mutant virus, and again became the dominant form, as 81.3%, 59.4% and 80.3% of viral genomes carry the mutation, respectively.

A recent study postulated that the one of the co-mutations of 14408C>T, namely 23403A>G that causes D614G mutation in the S protein may result in a more transmissible form of SARS-CoV-2 (16). This claim was based mainly on the observation that D614G mutant virus became the dominant form in more than one geographical location upon its introduction, as summarized above for its co-mutation 14408C>T. However, other explanations based on stochastic factors are equally possible, unless mechanistic insight arises that could explain how this particular mutation could lead to higher transmissibility. A study by Bhattacharyya suggested that D614G mutation creates an additional protease cleavage site near the S1-S2 junction, which may increase the success of viral integration with the host cell, and linked its dominance in Europe to certain human variants that control expression of TMPRSS2[17].

On the other hand, it is intriguing that between the first appearance of three co-mutations (on 14408, 23403, 3037) on January 24 in a Chinese isolate (EPI_ISL_422425) and their second co-appearance on February 20 in an Italian isolate, there are at least 6 and possibly 8 different viral genomes where three of the four co-mutations exist, with the exception of 14408C>T: on January 28 in Germany (EPI_ISL_406862), on February 5, 6, 7 and 8 in China (EPI_ISL_429080, EPI_ISL_429081, EPI_ISL_416334, EPI_ISL_412982, and EPI_ISL_429089); and two more Chinese viral sequences that failed our quality control standards and therefore eliminated from the overall analysis. Despite weeks of existence, unless completely confined and eliminated, only after the appearance of the first Italian case with all four mutations on February 20, 23403A>G became the dominant form. If this form of SARS-CoV-2 is really more transmissible, the next question that needs to be answered is whether it is due to any one of the three mutations alone, or whether a combination of two or three are needed. Based on the lack of successful spreading of the virus in its absence and our results showing increased mutability in its presence, we speculate that 14408C>T could be cooperating with the other two mutations. Alternatively, altered mutation rate may be a byproduct and the RdRp mutations may act through speeding up or slowing down the replication process,which would in turn affect the viral load and virulence. Also, mutant RdRp may become more resistant to anti-viral drugs, such as the commonly used remdesivir. Such implications make RdRp mutations attractive targets for epidemiological and functional studies with direct therapeutic implications.

Conclusions

Effects of different mutations on SARS-CoV-2 phenotypes (i.e. mutation rate, transmissibility, virulence, immune evasion etc.) are hot topics of research as there is an intense race worldwide to develop therapies and understand the viral biology. Some of these studies suggested that RdRp and Spike protein mutations could significantly affect the virus behavior and therefore the human health. Our study sheds light on the effects RdRp mutations, particularly 14408C>T mutation, on the mutability and possibly transmissibility of SARS-CoV-2. Further functional studies are required to test our findings.

Additional Information and Declarations

Acknowledgement

The authors would like to thank Mr. Alirıza Arıbaş and Mr. Ahmet Bursalı from Izmir Biomedicine and Genome Center for their technical assistance. The authors also would like to extend their thanks to Izmir Biomedicine and Genome Center (IBG) COVID19 platform IBG-COVID19 for their support in implementing the study and the Scientific and Technological Research Council of Turkey (TUBITAK) for their financial support of IBG-COVID19.

Funding

YO is supported by the Turkish Academy of Sciences Young Investigator Program (TÜBA-GEBİP)

Grant Disclosures

Not applicable

Competing Interests

The authors declare no competing interest.

Author Contributions

- DE conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.

- GK conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.

- AS conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.

- YO conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Supplemental Information

Not applicable

References

1. Chan, J. F.-W. *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet Lond. Engl.* **395**, 514–523 (2020).
2. Riou, J. & Althaus, C. L. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull.* **25**, (2020).
3. Li, J. *et al.* Emerging evidence for neuropsycho-consequences of COVID-19. *Curr. Neuropharmacol.* (2020) doi:10.2174/1570159X18666200507085335.
4. Wilson, M. P., Katlariwala, P. & Low, G. Potential implications of novel coronavirus disease (COVID-19) related gastrointestinal symptoms for abdominal imaging. *Radiogr. Lond. Engl. 1995* (2020) doi:10.1016/j.radi.2020.04.016.
5. Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome. (2020).
6. Tian, X. *et al.* Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. *Emerg. Microbes Infect.* **9**, 382–385 (2020).
7. The protein expression profile of ACE2 in human tissues | bioRxiv. https://www.biorxiv.org/content/10.1101/2020.03.31.016048v1.
8. Zhang, Y.-Z. & Holmes, E. C. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. *Cell* **181**, 223–227 (2020).
9. Ma, Y. *et al.* Structural basis and functional analysis of the SARS coronavirus nsp14–nsp10 complex. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 9436–9441 (2015).

10. Subissi, L. *et al.* One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. *Proc. Natl. Acad. Sci.* **111**, E3900–E3909 (2014).

11. Pachetti, M. *et al.* Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.* **18**, 179 (2020).

12. Begum, F. *et al. Specific mutations in SARS-CoV2 RNA dependent RNA polymerase and helicase alter protein structure, dynamics and thus function: Effect on viral RNA replication*. http://biorxiv.org/lookup/doi/10.1101/2020.04.26.063024 (2020) doi:10.1101/2020.04.26.063024.

13. Dilucca, M., Forcelloni, S., Georgakilas, A. G., Giansanti, A. & Pavlopoulou, A. Codon Usage and Phenotypic Divergences of SARS-CoV-2 Genes. *Viruses* **12**, 498 (2020).

14. GISAID Initiative. https://www.epicov.org/epi3/frontend#272e13.

15. Yin, C. Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics* (2020) doi:10.1016/j.ygeno.2020.04.016.

16. Korber, B. *et al.* Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv* 2020.04.29.069054 (2020) doi:10.1101/2020.04.29.069054.

17. Bhattacharyya, C. *et al. Global Spread of SARS-CoV-2 Subtype with Spike Protein Mutation D614G is Shaped by Human Genomic Variations that Regulate Expression of* TMPRSS2 *and* MX1 *Genes*. http://biorxiv.org/lookup/doi/10.1101/2020.05.04.075911 (2020) doi:10.1101/2020.05.04.075911.

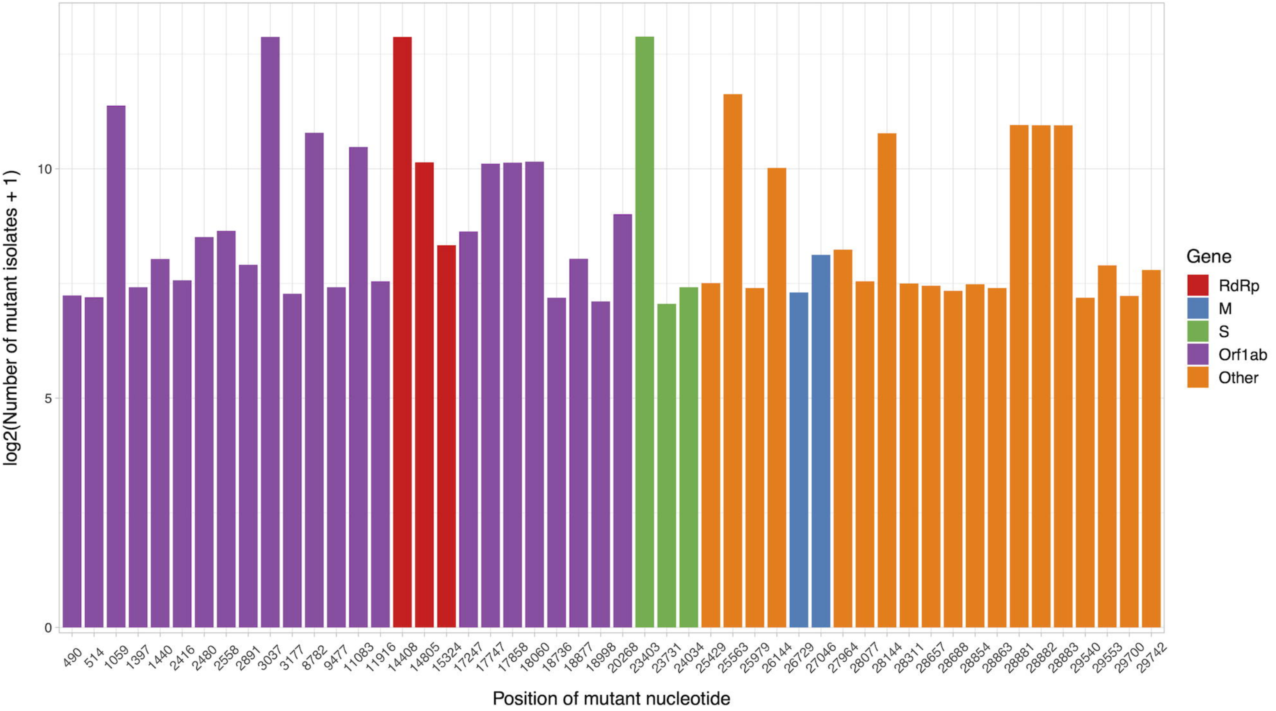
Figure Legends

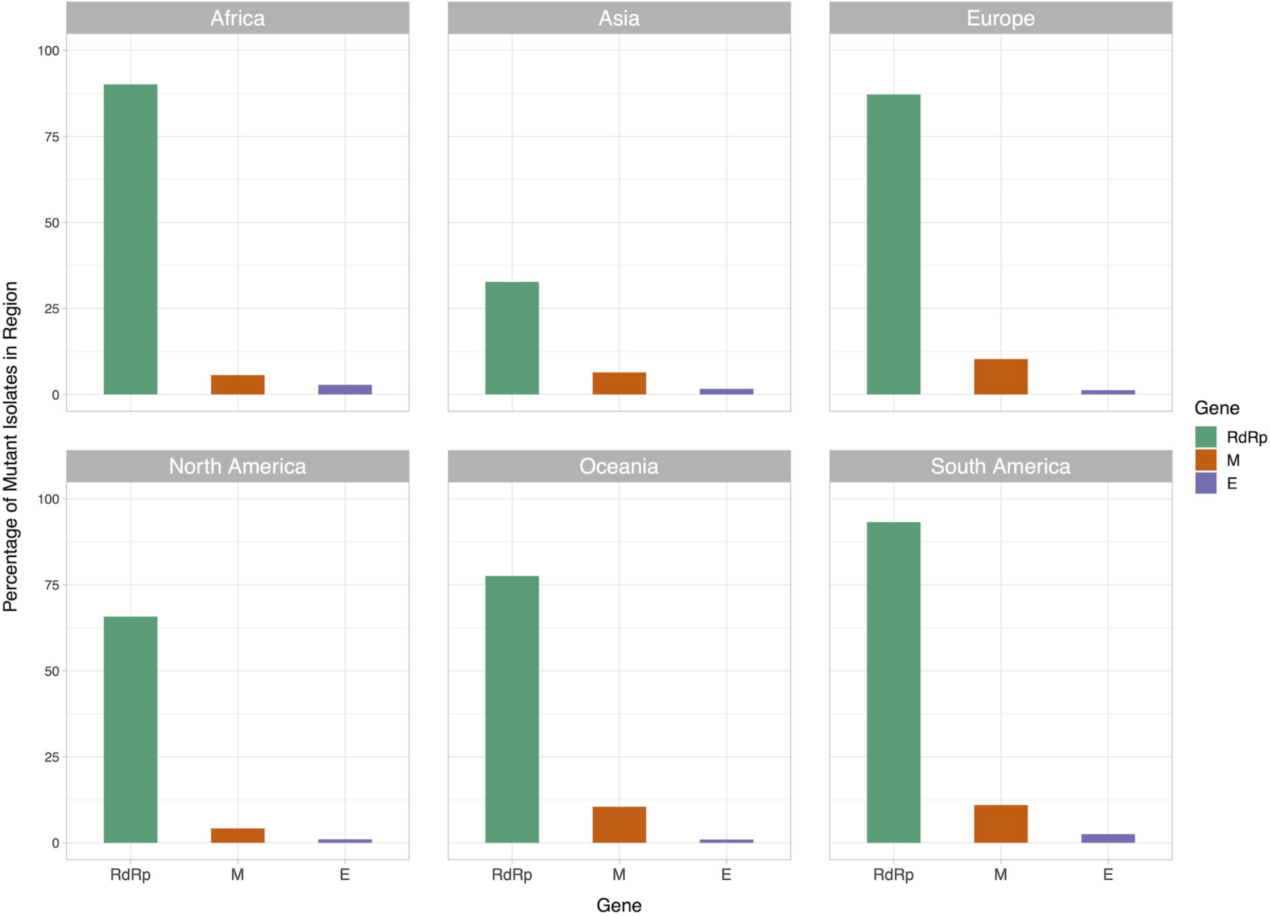
Figure 1.

Bar graph of top 50 most mutated nucleotides vs. log2-transformed number of samples with non-reference nucleotide at position. The x-axis represents the position of the nucleotide in the reference genome, the y-axis represents log 2 of number of isolates with disagreeing nucleotide aligning to the position in sequence plus 1. Unresolved sequence calls during library sequencing or gaps are not included in the number of isolates. Colors of the bars indicate the gene locus or mature peptide region where the nucleotide is in, with the RdRp mature peptide being considered separately from the remainder of the Orf1ab region. The 5' untranslated region and the 3'-most 100 nucleotides are not included in the graph.

Figure 2.

Bar graph of percent of isolates per region containing non-reference coding sequences for the proteins envelope glycoprotein (E), membrane glycoprotein (M), and RdRp protein. violet bars represent percent of isolates with mutant envelope glycoprotein, while orange bars represent the same percentage for membrane glycoprotein, and green bars represent the same for RdRp.


log2(Number of mutant isolates + 1)
0
5
10
490
514
1059
1397
1440
2416
2480
2558
2891
3037
3177
8782
9477
11083
11916
14408
14805
15324
17247
17747
17858
18060
18736
18877
18998
20268
23403
23731
24034
25429
25563
25979
26144
26729
27046
27964
28077
28144
28311
28657
28688
28854
28863
28881
28882
28883
29540
29553
29700
29742
Position of mutant nucleotide
Gene
RdRp
M
S
Orf1ab
Other


Africa
Asia
Europe
North America
Oceania
South America
100
75
50
25
0
Percentage of Mutant Isolates in Region
RdRp
M
E
Gene
Gene
RdRp
M
E