Genetics & Applications

Institute for Genetic Engineering and Biotechnology University of Sarajevo

Research article Open access

Comparison of SARS-CoV-2 Virus Variant Genomes Detected in China and the USA

Mehmet Cihan Sahingil[1*], Yakup Ozkazanc[2]

[1]*The Scientific and Technical Research Council of Turkey, Tubitak Bilgem/Iltaren, Umitkoy 06800 Ankara, Turkey*
[2]*Department of Electrical and Electronics Engineering, Hacettepe University, Beytepe 06800 Ankara, Turkey*

DOI: 10.31383/ga.vol4iss2pp10-26

***Correspondence**
E-mail: mehmet.sahingil@tubitak.gov.tr

Received
September, 2020
Accepted
November, 2020
Published
December, 2020

Copyright: ©2020 Genetics & Applications, The Official Publication of the Institute for Genetic Engineering and Biotechnology, University of Sarajevo

Keywords

COVID-19, SARS-CoV-2, Mutation, Genomic Variants, Multi-Alignment

Abstract

In the spreading period of the SARS-CoV-2 virus which is the cause of COVID-19 in the world, it is seen that the genome of the virus mutates and this mutation processes create new SARS-CoV-2 variants. In this study, we examined two variant genome groups. The first group includes the genomes of variants detected in China, while the second includes those detected in the USA. We applied the multiple sequence alignment process on these publicly available SARS-CoV-2 virus genomes and reported the obtained results. There are 87 genomes for China variants and 200 for USA variants in the used data. The analyses were made for each domain of the genomes. The results of the analysis show that the variant genomes in the two SARS-CoV-2 groups examined have some similar mutation features as well as different characteristics. It is also one of the results obtained in this study that the domains where the most mutations are detected on the genome are regions that have properties that affect the interaction of the virus with the host.

Introduction

The new coronavirus, which is also denoted as SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2), was discovered in December 2019 (Wu et al., 2020). According to the report of the World Health Organization (WHO, 2020), this name was chosen because the virus is genetically related to the coronavirus responsible for the SARS outbreak of 2003. While related, the two viruses differ from each other. The name of the disease which is caused by the virus is COVID-19 (Coronavirus Disease-19). The virus was discovered firstly in Wuhan city in China. Because the infectiousness of the virus is very high, the virus spread quickly in China and then in many countries globally. The first confirmed case of COVID-19

infection in the USA was reported on January 20, 2020 (Holshue et al., 2020). The virus is affecting 213 countries and territories around the world and two international conveyances with 4,609,337 cases, including 307,496 deaths as of May 15, 2020 (Worldometer, 2020).

As the virus is transmitted from person to person via droplet (Li et al., 2020) easily, more patients are infected and as a result, the virus is expected to accumulate more variants. Thus, while the virus is spreading globally, it appears that the genome of the virus mutates and these mutations create new SARS-CoV-2 variants (Koyama et al., 2020). Mutations are divided into various subtypes. Some mutation types are synonymous which means the mutation of the corresponding nucleotide does not cause a change in the coded amino acid. On the other hand, some mutation types are missense which means the mutation causes a change in the coded amino acid. Although there are other types of mutations as well (Antonarakis and Nomenclature Working Group, 1998; Den Dunnen and Antonarakis, 2001), in this study we mostly concentrated on these two types of mutations in coding domains of the variant genomes. SARS-CoV-2 viruses are RNA viruses with a positive-stranded genome of approximately 30 kb. There are some specific coding domains on the genome. The 5' two-thirds of the genome represents a large open reading frame (ORF) (Rozhgar et al., 2020). The 3' end of the coronavirus genome includes several structural and accessory protein genes. Figure 1 shows the gene structure of the SARS-CoV-2 genome. ORF1ab region contains 16 non-structural proteins and S contains two sub spike protein genes, S1 and S2. The position and size information of the domains in the genome are listed in Table 1 with brief descriptions (USCS Genome Browser, 2020).

The frequency of mutation occurrence in these domains varies. While mutations are rarely seen in some domains, the possibility of mutation in some may be high. Therefore, to be used in further studies, it should be revealed that in which domains mutations are seen more frequently, even for which nucleotides the probability of mutation is higher.

There are some early works that consider the SARS-CoV-2 variant analysis for the variants detected in a specific location, especially in China (Koyama et al., 2020; Rozhgar et al., 2020). In this study, we compared the variation characteristics of two variant genome groups which are gathered from two distinct locations of the world, China and the USA. The variant genomes in both groups were aligned, some statistical characteristics for each of the domains of the SARS-CoV-2 genome were obtained and the obtained comparison results were reported.

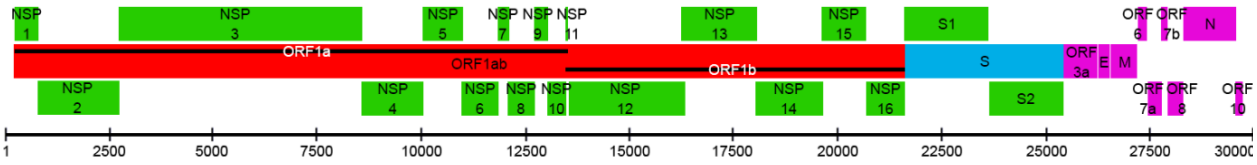

Figure 1. Gene Structure of the SARS-CoV-2 Genome

Material and methods

Coronaviruses are RNA viruses. The coronavirus has a positive-stranded genome of approximately 30 kb. As shown in Figure 1, the 5' two-thirds of the genome represents a large ORF which is denoted as ORF1ab (Rozhgar et al., 2020). This large ORF consists of two large replicase sub-ORFs, ORF1a, and ORF 1b. The 3' end of the coronavirus genome includes several structural and accessory protein genes. These domains are a spike (S) glycoprotein gene, an envelope (E) protein gene, a membrane (M) glycoprotein gene, a nucleocapsid (N) phosphoprotein gene, and several ORFs that encode putative non-structural proteins (NSP) (Vijgen et al., 2005). To examine the nucleotide sequences in these domains, we gathered the genomes detected in China (the group 1 genome set) and the genomes detected in the USA (the group 2 genome set) as the first step of our study. The links to the genomic data of the reference genome and its variants are publicly available on a web page of China's National Genomics Data Center (NGDC) which is a COVID-19 dedicated page (NGDC, 2020).

Table 1. The Details Of The SARS-CoV-2 Genome Domains (USCS Genome Browser, 2020)

Domain Name	Position	Size	Brief Description
NSP1 (Non-structural protein 1)	266-805	540	Interacts with the 40S ribosomal subunit within the host cell and prevents the host's translation.
NSP2	806-2719	1914	PHB1 (prohibitin1) and PHB2 (prohibitin2) proteins are known to play a role in protecting cells from various stresses and ensuring the functional integrity of the mitochondrial organelle of the cell. NSP2 may play a role in the cell survival signaling pathway modulation of the host cell by interacting with these proteins.
NSP3	2720-8554	5835	NSP3 and NSP4 play a critical role in viral replication.
NSP4	8555-10054	1500	
NSP5 (3CL-PRO) (3C-like proteinase)	10055-10972	918	NSP5 cleaves the C-terminus of replicase polyprotein.
NSP6	10973-11842	870	NSP6 is involved in the initial induction of autophagosomes from the reticulum endoplasmic of the host cell.
NSP7	11843-12091	249	NSP7 and NSP8 may form a hexadecamer which participates in the viral replication process. NSP7 and NSP8 do this by acting as a primase.
NSP8	12092-12685	594	
NSP9	12686-13024	339	NSP9 may participate in viral replication. This is done by the action of NSP9 as an ssRNA binding protein.
NSP10	13025-13441	417	NSP10 is known to play an important role in cap methylation of viral mRNAs.
NSP11	13442-13480	39	No additional description.
NSP12 (Pol) (RNA-directed RNA polymerase)	13442-16236	2795	NSP12 is involved in the transcription and replication of the viral RNA genome.
NSP13 (Hel) (Helicase)	16237-18039	1803	NSP13 has a zinc-binding domain in N-terminus.
NSP14 (ExoN) (Proofreading exoribonuclease)	18040-19620	1581	NSP14 enzyme possesses an exoribonuclease activity acting on both ssRNA and dsRNA. Besides, it has also an N7-guanine methyltransferase activity.
NSP15 (Uridylate-specific endoribonuclease)	19621-20658	1038	No additional description.
NSP16 (2'-O-methyltransferase)	20659-21552	894	No additional description.
S glycoprotein (Spike glycoprotein)	21599-25381	3783	S1 (21599-23617) initiates the infection by interacting with the host receptor and binding the virion to the cell membrane. S2 (24008-25381) functions as an unmasked viral fusion peptide.
ORF3a (Protein 3a)	25393-26217	825	ORF3a has a multi-functional structure. It forms ion channels which are homotetrameric potassium sensitive. It may also modulate virus release. Upregulates the expression of fibrinogen genes (alpha (FGA), beta (FGB), and gamma (FGG)) in host lung epithelial cells. Additionally, it triggers cell culture apoptosis.
ORF4 (E protein) (Envelope small membrane protein)	26245-26469	225	ORF4 plays a critical role in virus morphogenesis and assembly.
M protein (Membrane protein)	26523-27188	666	The M protein is a viral envelope component. It interacts with other viral proteins and thus plays a role in virus morphogenesis and assembly.
ORF6 (ns6) (Non-structural protein 6)	27202-27384	183	ORF6 may stimulate cellular DNA synthesis.
ORF7a (Protein 7a)	27439-27756	318	ORF7a is a non-structural protein needed in virus replication in cell culture.
ORF7b (ns7b) (Non-structural protein 7b)	27756-27887	129	No additional description.
ORF8	27894-28259	366	ORF 8 may play a role in host-virus interaction.
N (Nucleoprotein) (NC)	28274-29533	1260	N has an important role in viral replication. It packages the viral genome into a coiled ribonucleocapsid. Besides, it plays a critical role in virion assembly by interacting with the M protein and the genome of the virus.
ORF10	29558-29674	117	No additional description

The numbers of variant genomes in group 1 and group 2 are 87 and 200, respectively. Because most of the analyzes in this study are percentage based, the difference between the numbers of variant genomes in the groups does not affect the results dramatically.

As the second step, we applied multiple sequence alignment on the variant sequences by using the Bioinformatics toolbox of MATLAB software. The gap penalty and extension penalty are accepted as 10 and 0.5, respectively. In the third step of the study, we used NC_045512 (USCS Genome Browser, 2020) as the reference genome and compared the multi-aligned variant genomes with NC_045512 genome. The main reason for using this genome as the reference genome is to make the results comparable with the previous studies (Koyama et al., 2020; Lu et al., 2020; Rozhgar et al. 2020) where this genome is also used as the reference genome. As a result of this comparison process, we obtained some important statistical results.

Results and Discussion

We firstly investigated the non-coding parts, which are also denoted as untranslated regions (UTR), of the SARS-CoV-2 genomes for all the variant genomes. The first coding domain of the genome is NSP-1 (Figure 1) and the starting nucleotide position of NSP-1 is 266. Therefore the first 265 nucleotides are taken as the non-coding region (5' UTR). The most common nucleotide mutation types and the corresponding mutation locations are shown in Figure 2. To make this analysis we compared the reference sequence with 287 variant genomes (87 for China, 200 for the USA) of the SARS-CoV-2 virus detected in China and the USA.

The solid line shows the most common mutation percentage values in the corresponding nucleotide position of the genome. The nucleotide positions of the genome are shown on the x-axis of the graph. On the other hand, the points in the figure show the most common nucleotide mutation types in the corresponding nucleotide position. The nucleotide mutation types are shown on the right side of the figure. There are 17 different cases: 1 "No Mutation" case, 12 "mutation between nucleotide types" cases, and 4 "deletion" cases. The percentage value for the "No-Mutation" type shows the percentage of variants that have the same nucleotide type with the reference genome for the investigated genome position. The percentage values within the brackets are indicating the percentage of the corresponding most common mutation types among all the mutation types in the investigated genome region. For example, 22% of the most commonly detected mutations are C>T mutations in the 5' UTR region and the corresponding mutation positions are

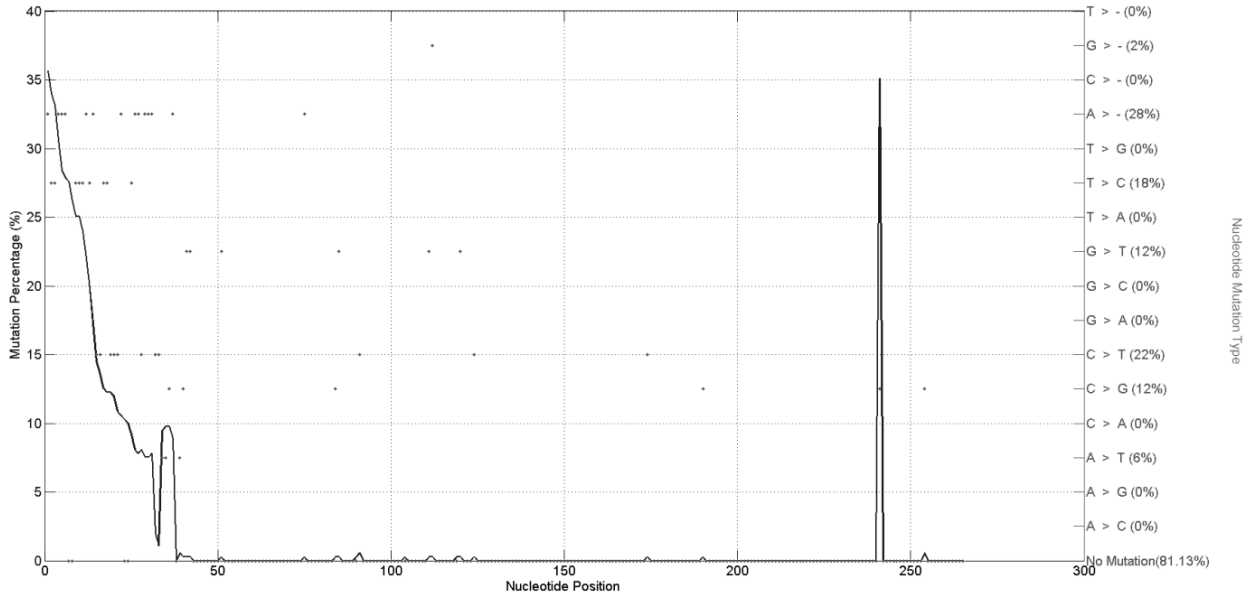

Figure 2. The Most Common Nucleotide Mutations in 5' UTR of the SARS-CoV-2 Genome

shown with points in the figure. In the following parts of this paper, this graphical display method will be used for expressing mutation percentages.

Figure 2 shows that the mutations mostly occur in the anterior part (~40 nucleotides) of 5'UTR. The most common mutation types are A>- (deletion) (28%), C>T mutation (22%), and T>C mutation (18%). Moreover, a mutation in position 241 must be also considered. As seen in the figure, the most common mutation type in this position is C>G mutation and this mutation type is observed in 35% of all the investigated variants. The percentage of mutations observed at nucleotide positions other than the anterior part of the 5'UTR and position 241 is less than 5%.

Additionally, the last domain (ORF10) finishes in 29674th nucleotide. The remaining part of the genome after this position is taken as the non-coding region (3 'UTR). The most common nucleotide mutation types and the corresponding mutation locations are shown in Figure 3. To make this analysis we used the same reference genome and variant genomes which were used in the5' UTR analysis. The mutations occur mostly in the last 40 nucleotides of the non-coding region. The most occurred mutation types are T>C mutation (35.61%), A>- deletion (32.87%), and C>T mutation (19.17%). It should be noted that these mutation types are the same as those with the highest probabilities detected in the 5 'UTR analysis. The mutation percentages in other nucleotide positions are below 10%.

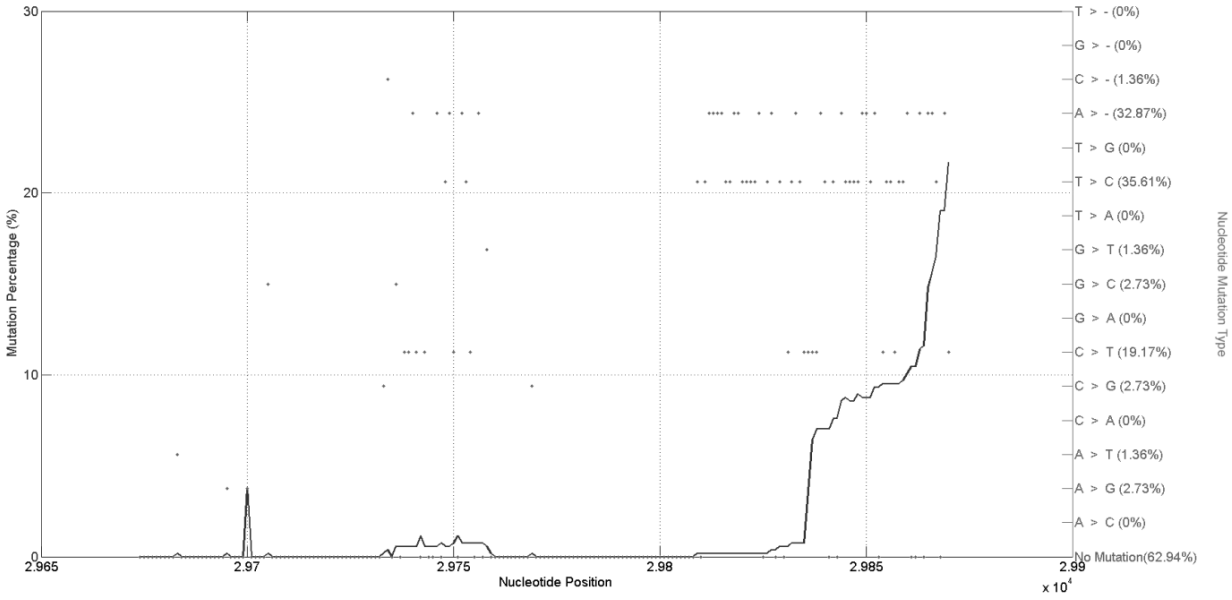

Figure 3. The Most Common Nucleotide Mutations in 3' UTR of the SARS-CoV-2 Genome

Analysis of Mutations in SARS-CoV-2 China Variants

After investigating the non-coding regions of the SARS-CoV-2 genome we investigated the coding regions of the genome. As the first step, we considered only the variants gathered from China. The total number of variants in the China group is 87. The most common nucleotide mutation percentages, the mutation types, the corresponding mutation locations, and the domain boundaries are shown in Figure 4.

The dashed green vertical lines are showing the boundaries of the reference SARS-CoV-2 genome domains. The names of the genomic domains are shown at the top of the chart. The percentage values next to the domain names are the percentages of the total mutation number in the corresponding domain to the total mutation number in all variant genomes. The percentage values are calculated by using the formula shown in (1).

$$Domain\ Mutation\ \%_i = \frac{(Total\ number\ of\ mutations\ in\ domain\ i)}{\sum_{j=1}^{N}(Total\ number\ of\ mutations\ in\ domain\ j)} \cdot 100 \qquad (1)$$

where N is the total number of domains.

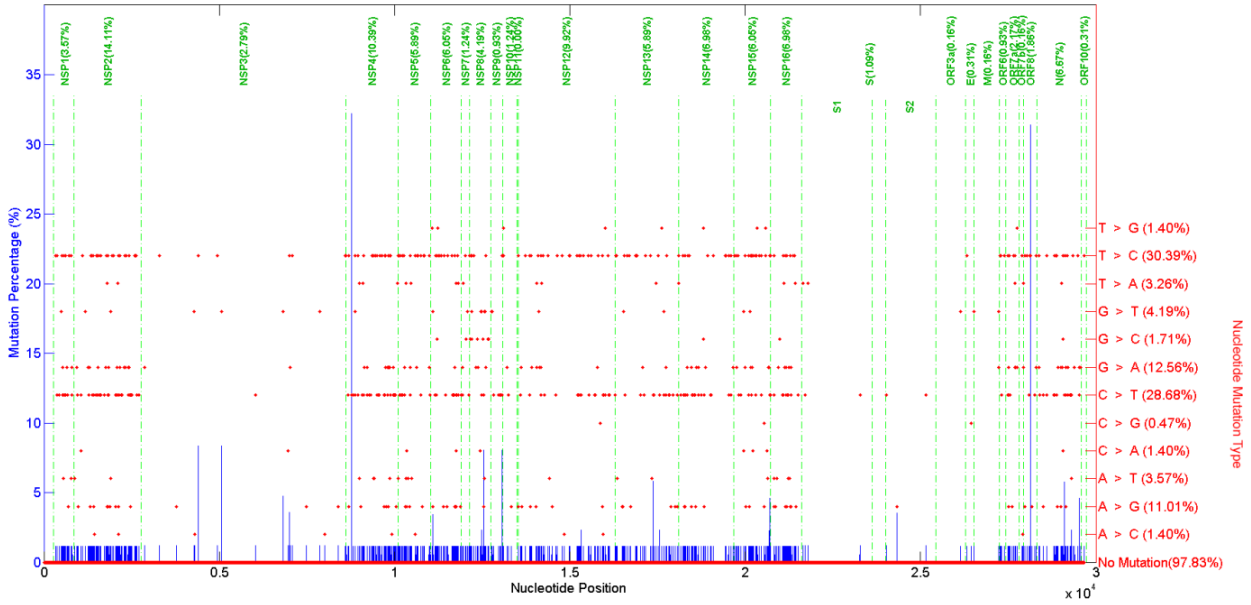

Figure 4. The Most Common Nucleotide Mutations in the Coding Region of the SARS-CoV-2 Genome Variants Detected in China

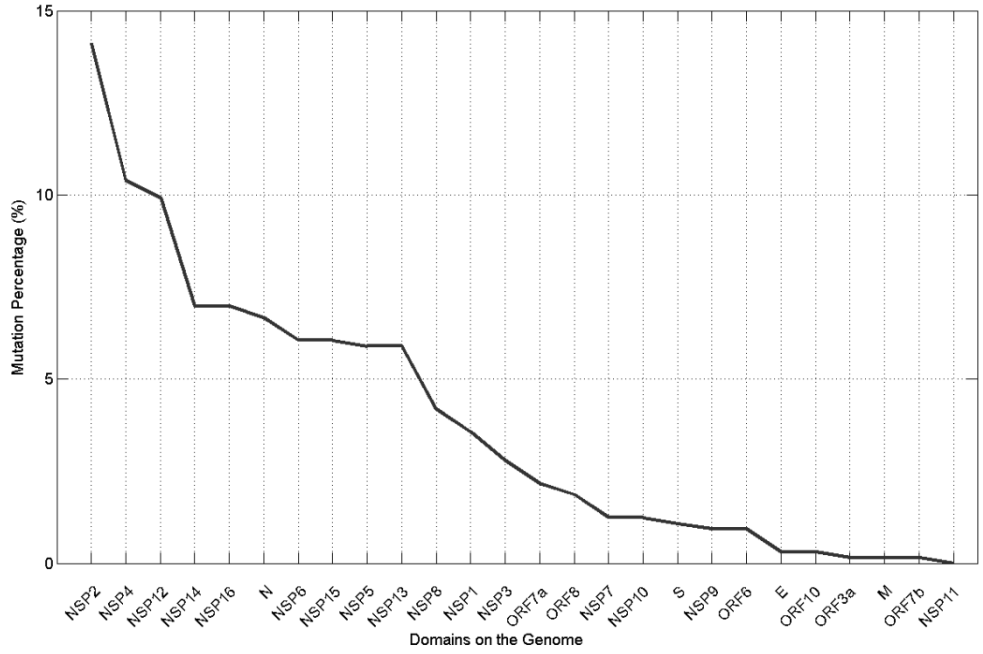

Figure 5. The Mutation Percentages in Domains of the SARS-CoV-2 Genome Variants Detected in China

Examining Figure 4, it can be easily seen that mutations at two nucleotide positions dominate the results. These points are 8782 (C>T mutation) and 28144 (T>C mutation). This result is consistent with the previous studies (Koyama et al., 2020; Lu et al., 2020; Rozhgar et al. 2020). The percentage values of all other mutations at specific nucleotide positions are less than 10%. The most common nucleotide mutation types are T>C (30.39%) and C>T (28.68%). Considering the nucleotide positions of the mutations, it is seen that the total mutation number for some genome domains are very low. The

domains where mutation percentages are less than 2% are NSP7, NSP9-11, S, ORF3a, E, M, ORF6, ORF7b, ORF8, and ORF10. The domain mutation percentages which are calculated by the formula given in (1) are shown in Figure 5.

The nucleotide numbers of domains differ from each other. Therefore, mutation density can also be mentioned for each domain. In this step, the nucleotide mutation densities for each domain are calculated by using the formula (2) and the obtained results are shown in Figure 6. For variants detected in China, the domain with the highest mutation density is NSP16, while the least dense domain is NSP11.

As is known, not all nucleotide mutations cause an amino acid change in the translation step of the gene regulation process. In the synonymous mutation types, the nucleotide mutates but the changed codon encodes the same amino acid type as the unchanged codon. On the other hand, some nucleotide mutations cause changes in amino acid type. These types of mutations which are denoted as missense mutations cause the resulting protein to change its structure. This is why missense mutations are more critical. Figure 7 shows the most common types of mutations in the coding region of SARS-CoV-2 genome variants detected in China. The red dots indicate the most common type of mutations at specific nucleotide positions. The red dots on the "No Mutation" line are the nucleotide positions where no mutation was detected among all variants. The percentage value next to the "No Mutation" label indicates the percentage of unmutated nucleotides in the genome to the total number of nucleotides. Similarly, the red dots on the "Missense" and "Synonymous" lines indicate the nucleotide positions where missense and synonymous mutations were detected, respectively. The percentage values next to the "Missense" and "Synonymous" labels show the percentage of missense and synonymous mutations to the total number of mutations in the genome, respectively.

Figure 7 shows that the percentage of synonymous mutations is 73.9% of all mutations, while the percentage of missense mutations is 26.1% of all mutations. The number of missense and synonymous mutations varies from domain to domain in the genome. Considering the percentages of synonym and missed mutations to all mutations for each domain, the graph given in Figure 8 is obtained.

$$Mutation\ Density\ \%_i = \frac{(Total\ number\ of\ mutations\ in\ domain\ i)}{(Total\ number\ of\ nucleotides\ in\ domain\ i)} \cdot 100 \qquad (2)$$

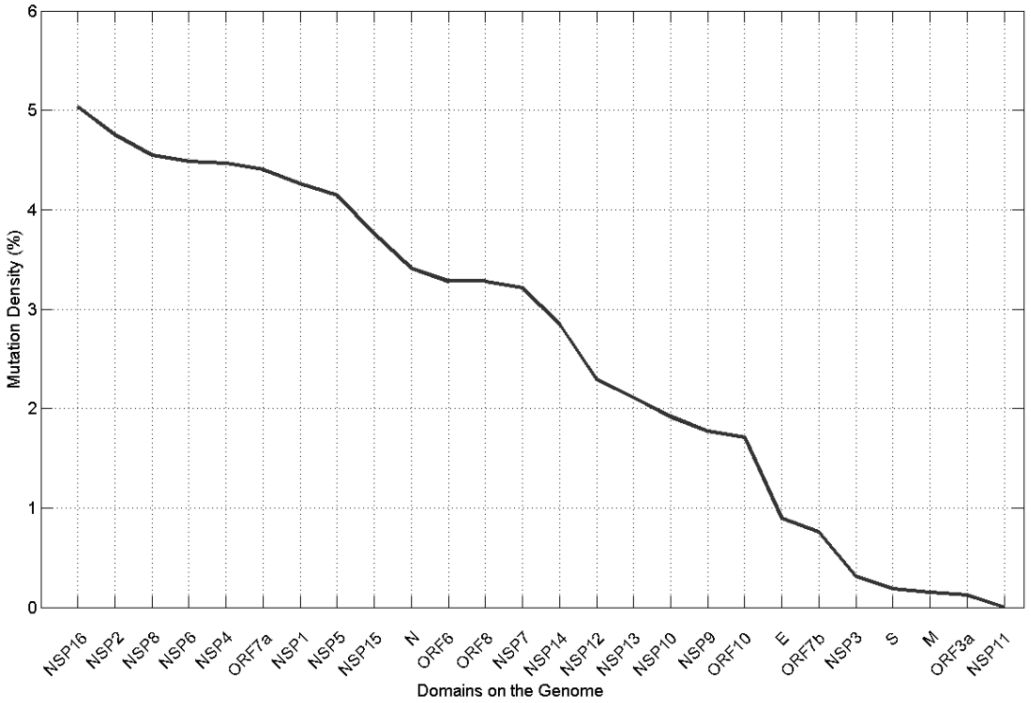

Figure 6. The Mutation Densities in Domains of the SARS-CoV-2 Genome Variants Detected in China

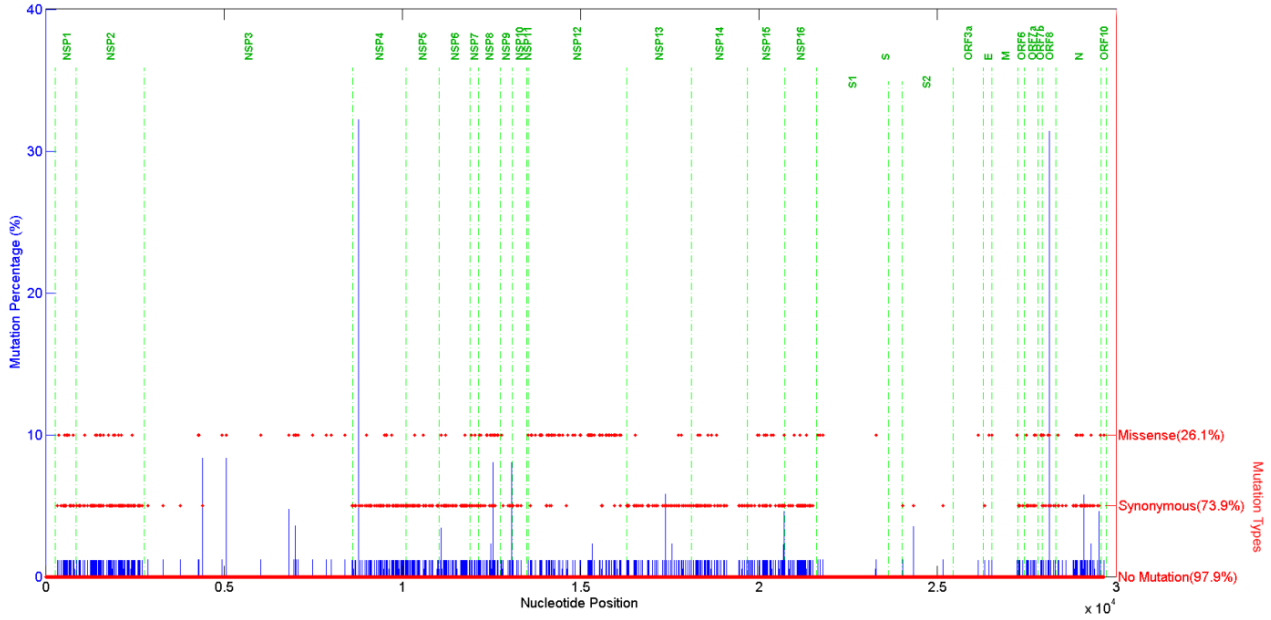

Figure 7. The Most Common Mutation Types in the Coding Region of the SARS-CoV-2 Genome Variants Detected in China

$$Synonymus\ \%_i = \frac{(Total\ number\ of\ synonymous\ mutations\ in\ domain\ i)}{(\#mutations\ in\ the\ entire\ coding\ region\ of\ genome)} \cdot 100 \qquad (3)$$

$$Missense\ \%_i = \frac{(Total\ number\ of\ missense\ mutations\ in\ domain\ i)}{(\#mutations\ in\ the\ entire\ coding\ region\ of\ genome)} \cdot 100 \qquad (4)$$

where "#mutations" means the total number of mutations.

$$SynonymusDens\ \%_i = \frac{(Total\ number\ of\ synonymous\ mutations\ in\ domain\ i)}{(Total\ number\ of\ nucleotides\ in\ domain\ i)} \cdot 100 \qquad (5)$$

$$MissenseDens\ \%_i = \frac{(Total\ number\ of\ missense\ mutations\ in\ domain\ i)}{(Total\ number\ of\ nucleotides\ in\ domain\ i)} \cdot 100 \qquad (6)$$

Percentage values in Figure 8 were calculated using the formula shown in (3) and (4). The ordering of domain names on the x-axis of the graph was made by considering the descending order of the missense mutations. When the graph is examined, it can be seen that the missense mutations are mostly detected in the NSP12 domain.

However, the nucleotide numbers in the domains are different from each other. Therefore, it is possible to talk about the mutation density for each domain using the formulas shown in (5) and (6). The missense and synonymous mutation density percentages for each domain are shown in Figure 9.

As in the previous analysis, the ordering of domain names on the x-axis of the graph was made by considering the descending order of the missense mutations. The missense mutation graph in Figure 9 shows that the domain with the highest missense mutation density is the NSP8. In the next analysis, the amino acid sequence produced according to the reference genome was compared with the amino acid sequences produced for all variants and the most common amino acid changes were detected. The results are shown in Figure 10. In other words, Figure 10 shows the percentage values of amino acid changes resulting from missense mutations.

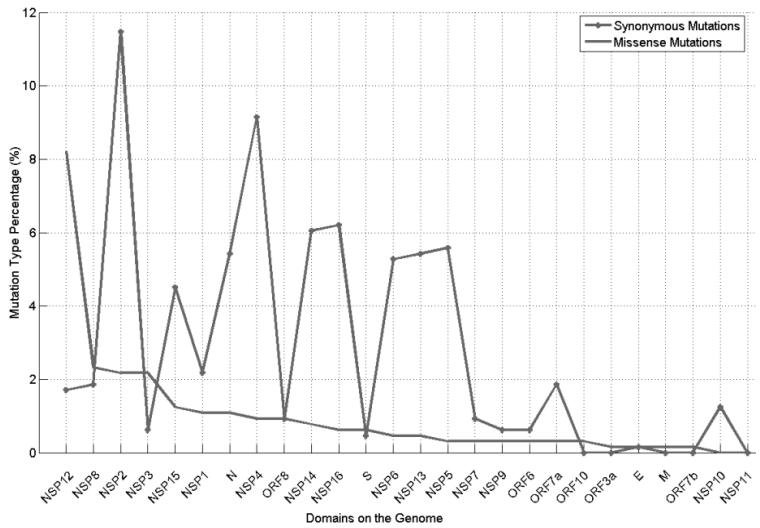

Figure 8. The Percentages of Most Common Synonymous and Missense Mutations in Each Domain of the SARS-CoV-2 Genome Variants Detected in China

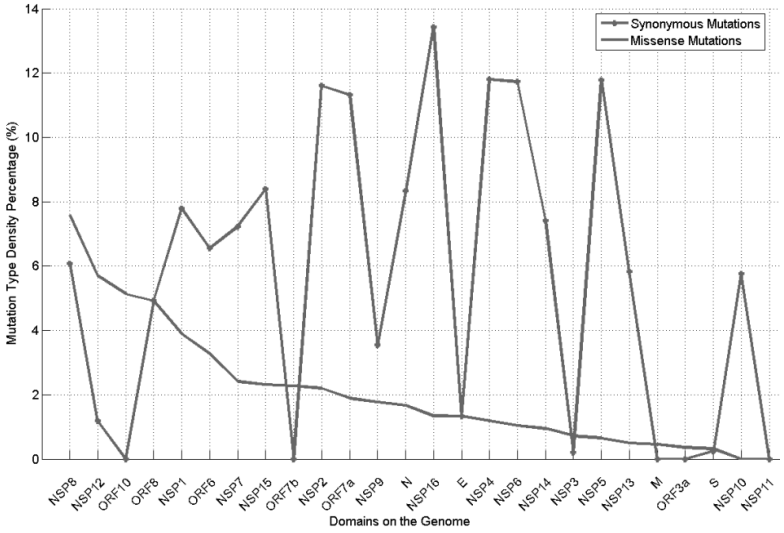

Figure 9. The Density Percentages of Most Common Synonymous and Missense Mutations in Each Domain of the SARS-CoV-2 Genome Variants Detected in China

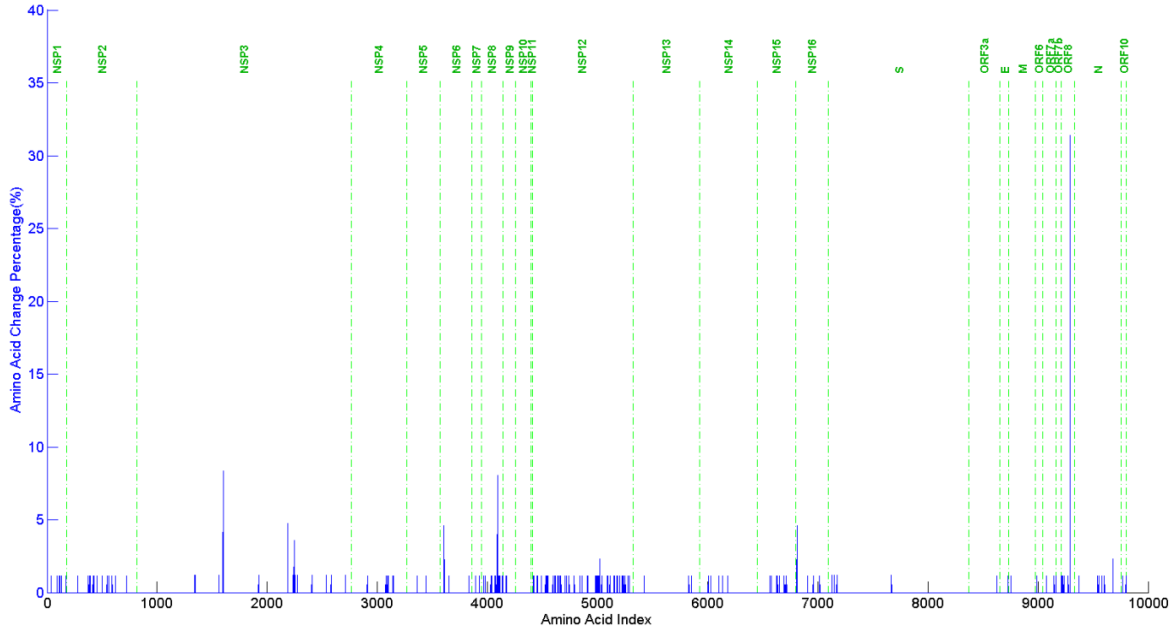

Figure 10. The Most Common Amino Acid Changes and Their Positions in the Coding Region of the SARS-CoV-2 Genome Variants Detected in China

Figure 10 shows that amino acid changes were detected most extensively in ORF8. This change is the change from amino acid "L" to amino acid "S". The percentage values of all other amino acid changes are less than 10%. This means that the most dominant missense mutation is in the ORF8 domain for the China variants. As shown in Table 1, ORF8 may play a role in host-virus interaction. When Figure 4 and Figure 10 are compared, it can be seen that the dominant nucleotide mutation in the NSP4 domain (Figure 4) does not cause an amino acid change (Figure 10), meaning that the nucleotide mutation in the NSP4 domain is a synonymous mutation. On the other hand, the dominant nucleotide mutation in the ORF8 domain (Figure 4) causes an amino acid change (Figure 10).

Amino acids can be categorized into four different classes based on their polarity properties. These categories are expressed as the aminoacids with non-polar R groups (A, V, L, I, P, F, W, M), the amino acids with polar R groups (G, S, T, Y, C, N, Q), the amino acids with negatively charged R groups (D, E) and the amino acids with positively charged R groups (K, R, H). Nucleotide mutations can cause an amino acid of a certain category to change to an amino acid of a different category. The results obtained after analyzing these category changes are shown in Figure 11. It can be easily seen that (non-polar R group) > (non-polar R group) is the most common category change. Such changes are mostly detected in the NSP12 domain. This result supports the analysis result shown in Figure 8.

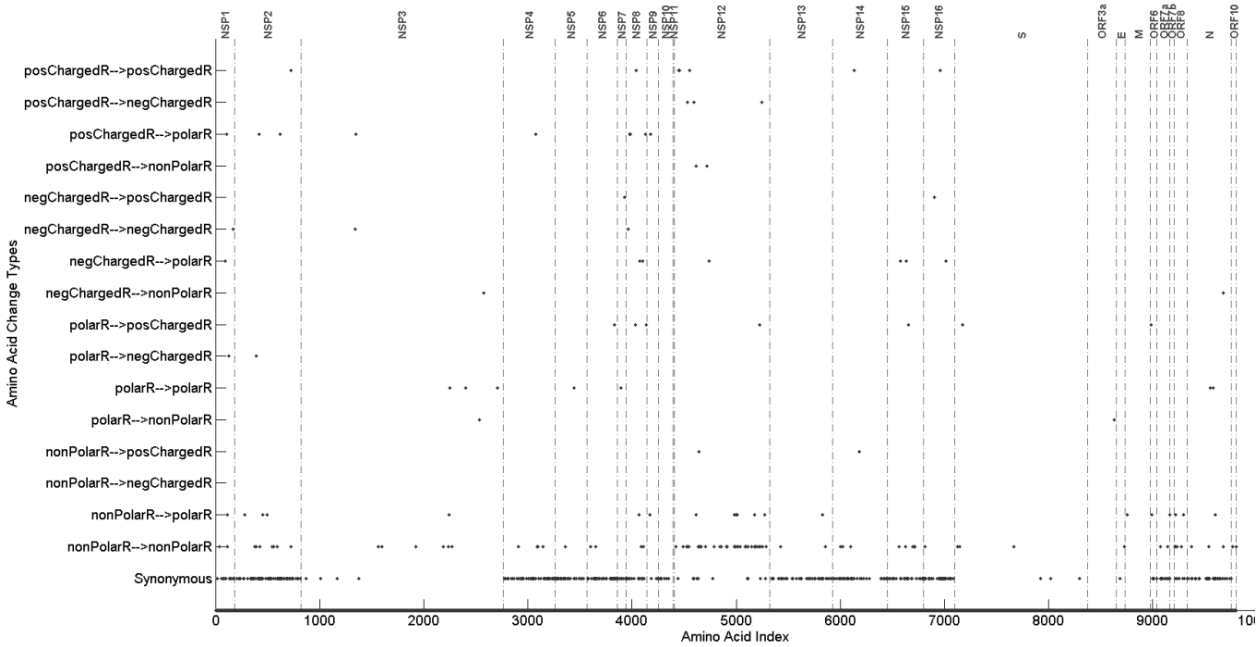

Figure 11. The Most Common Changes of Amino Acid Types in the Coding Region of the SARS-CoV-2 Genome

To show the total number of amino acid changes in each domain for each amino acid change type, we used a color-coded display method shown in Figure 12. The red regions are the regions with the highest number of amino acid changes. For example, the synonymous amino acid change in the NSP2 domain is the most occurring amino acid change among all other amino acid changes. If the color map on the right is checked, it can be seen that the synonymous amino acid change number in the NSP2 domain is 74. The number of amino acid changes for (non-polar R group) > (non-polar R group) change type in the NSP12 domain is statistically significantly higher than all other missense mutation type numbers for all domains. This result is consistent with the result shown in Figure 11.

Analysis of Mutations in SARS-CoV-2 USA Variants

After examining the variants detected in China, we investigated the SARS-CoV-2 variants detected in the USA as a second step. The total number of variant genomes obtained for the USA is 200. The most common types of nucleotide mutations, corresponding mutation positions, and the domain boundaries are shown in Figure 13.

Similar to the analysis results obtained for China variants (Figure 4), high nucleotide mutation percentages were detected at nucleotide positions

8782 (C> T mutation) and 28144 (T> C mutation) for variants detected in the USA. This result is consistent with the previous studies (Koyama et al., 2020; Lu et al., 2020; Rozhgar et al. 2020). However, there are 8 additional mutation positions with mutation percentage values greater than 10%. These nucleotide positions and the corresponding most common nucleotide mutation types are listed in Table 2. Mutations at nucleotide positions 3037, 14408, 23403, and 25563 have higher percentage values than mutations at nucleotide positions 8782 and 28144. The most common type of nucleotide mutation is C>T (40.95%). This result differs from the analysis result obtained for China variants. In China results the most common nucleotide mutation type was T>C (30.39%). When the domain-based mutation percentages and mutation densities are examined for the USA variants, the result graphs given in Figure 14 and Figure 15 are obtained. If the analysis results obtained for the China and USA variants are compared (Figure 5 vs Figure 14 and Figure 6 vs Figure 15) it can be said that the domains of the most common nucleotide mutation positions are not consistent for these two variant groups.

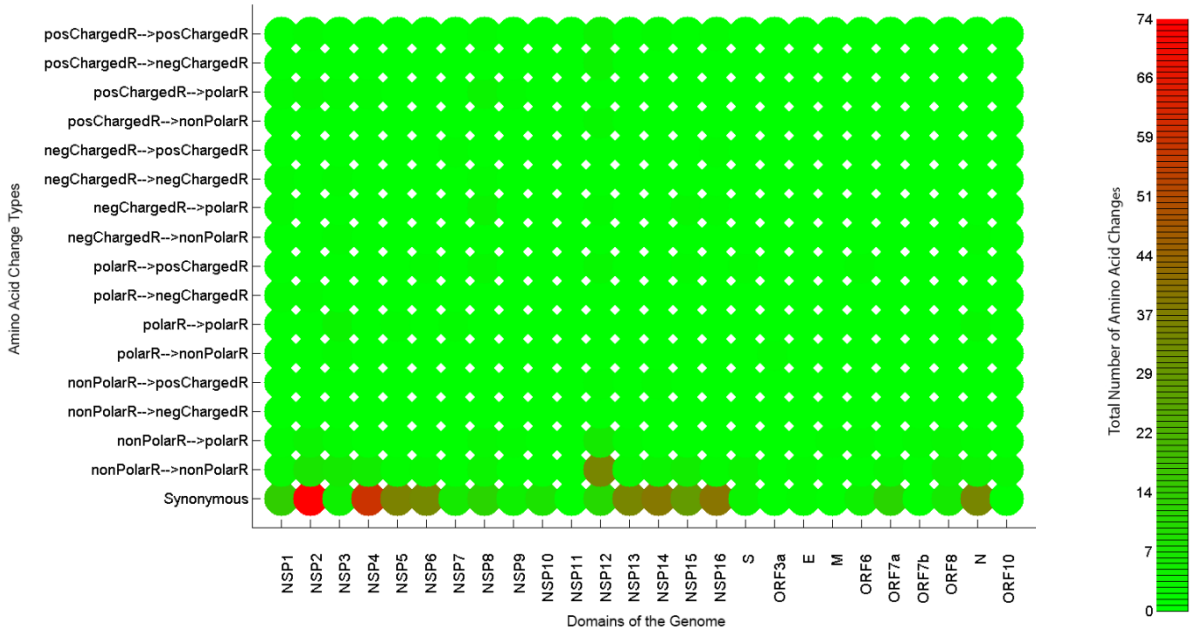

Figure 12. The Number of Most Common Changes of Amino Acid Types in the Coding Region of the SARS-CoV-2 Genome Variants Detected in China

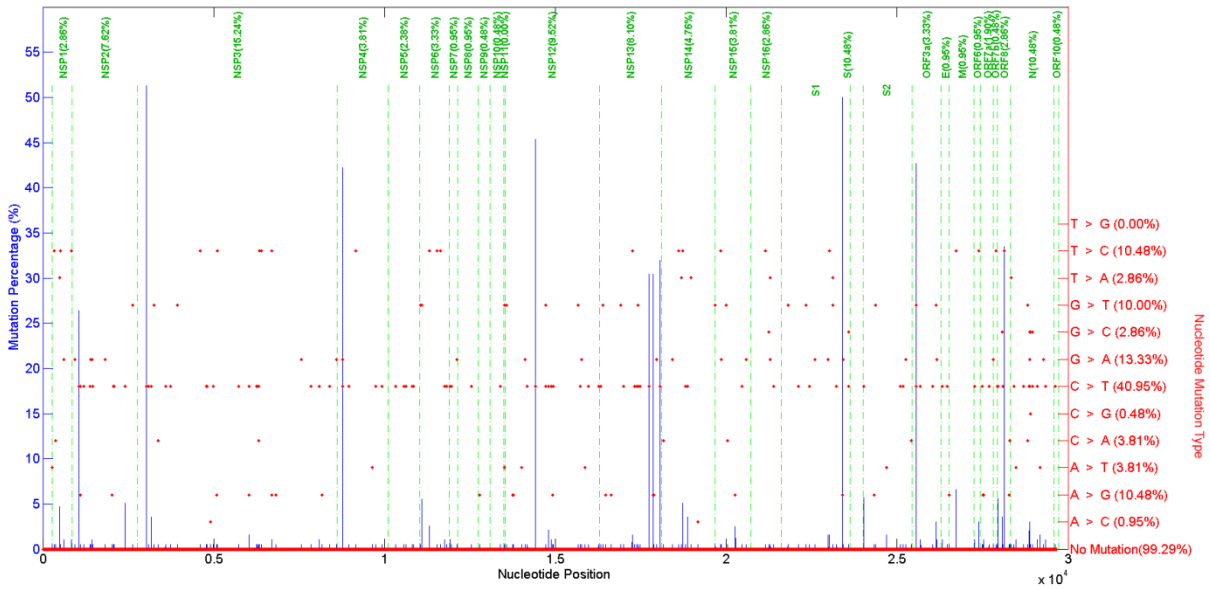

Figure 13. The Most Common Nucleotide Mutations in the Coding Region of the SARS-CoV-2 Genome Variants Detected in the USA

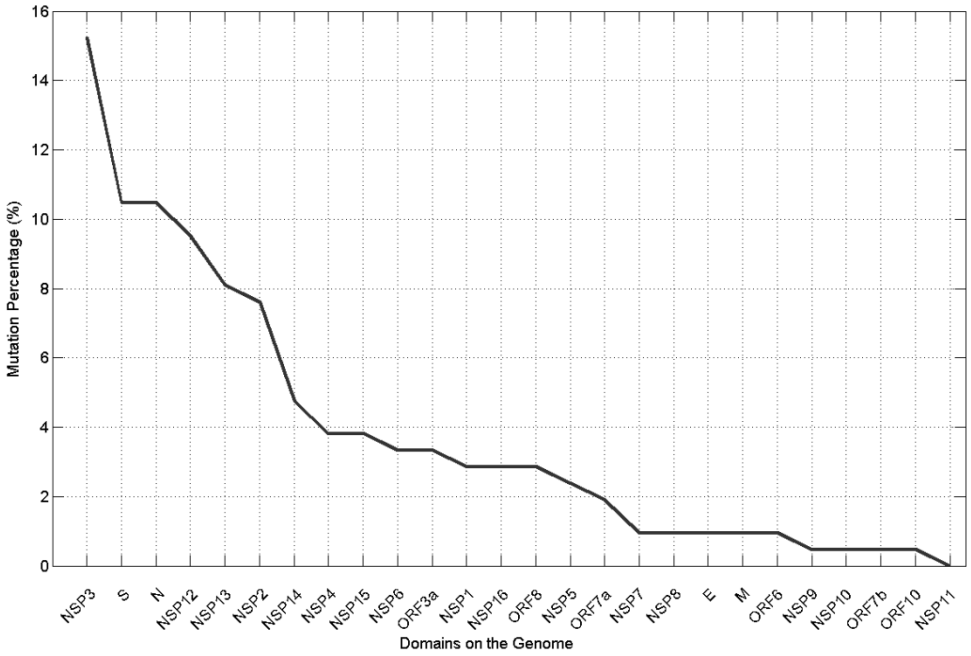

Figure 14. The Mutation Percentages in Domains of the SARS-CoV-2 Genome Variants Detected in the USA

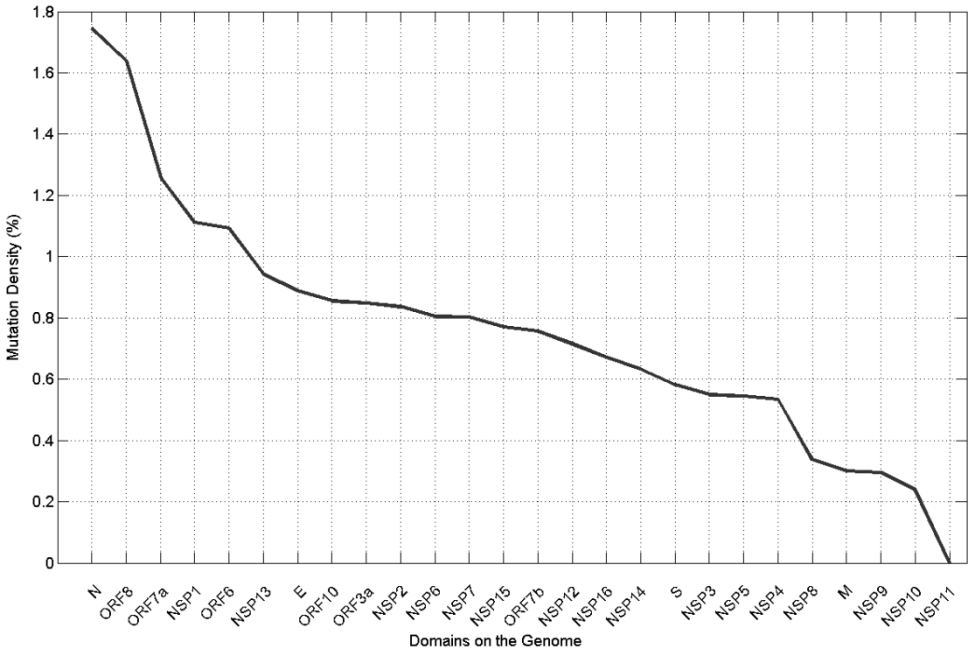

Figure 15. The Mutation Densities in Domains of the SARS-CoV-2 Genome Variants Detected in the USA

Figure 16 shows the most common types of mutations in the coding regions of SARS-CoV-2 genome variants detected in the USA. The percentage of synonymous mutations is 37.4% of all mutations while this percentage value is 62.6% for missense mutations. In the analysis result of China variants (Figure 7) the percentage values for synonymous mutations and missense mutations are 73.9% and 26.1%, respectively, i.e. number of synonymous mutations is greater than the number of missense mutations. However, the analysis result of the USA variants shown in Figure 16 shows that the number of missense mutations is greater than the number of synonymous mutations. This result shows that the percentage of synonym mutations is higher for China variants than the percentage of missense mutations, but the opposite is true for the USA variants.

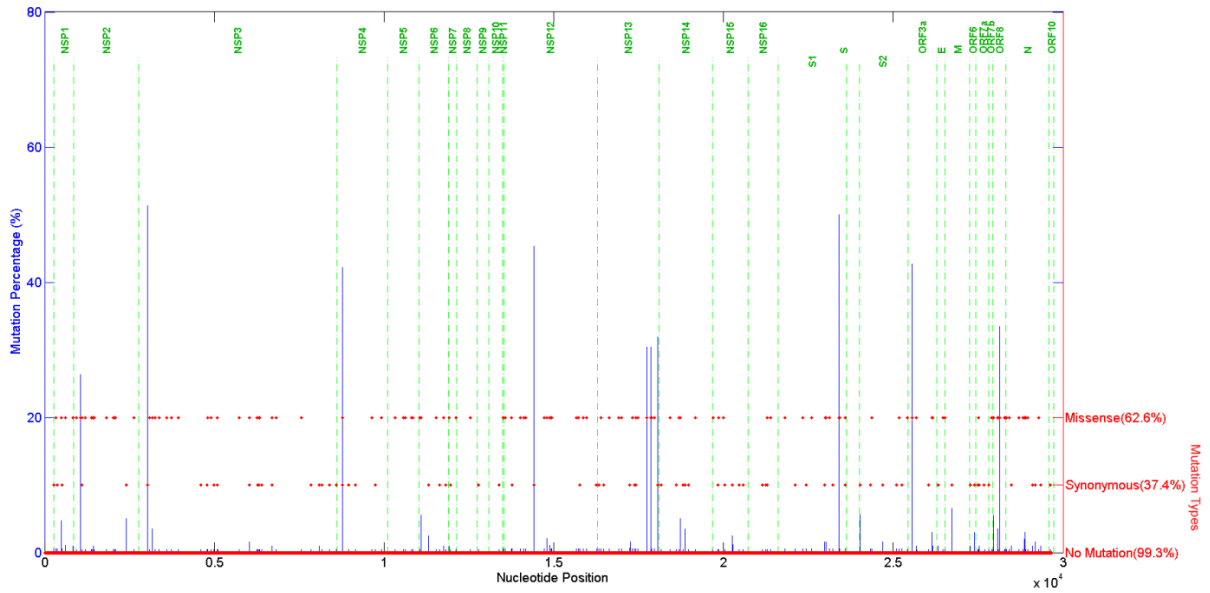

Figure 16. The Most Common Mutation Types in the Coding Region of the SARS-CoV-2 Genome Variants Detected in the USA

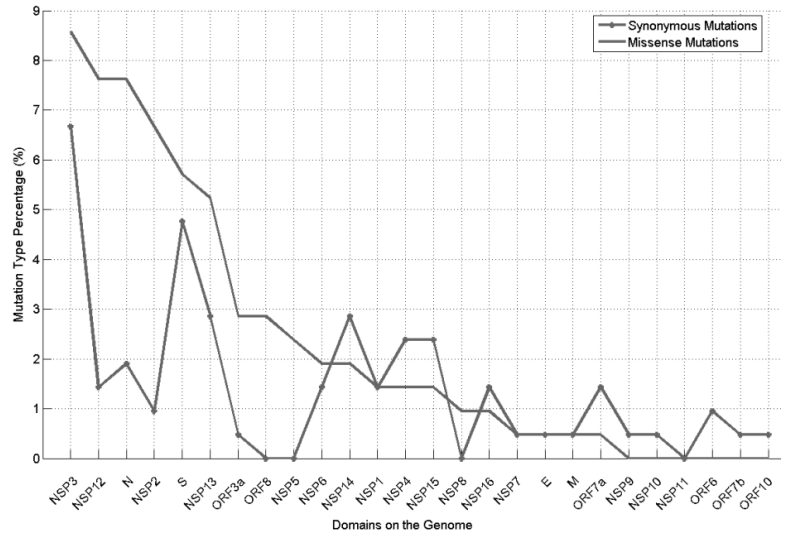

Figure 17. The Percentages of Most Common Synonymous and Missense Mutations in Each Domain of the SARS-CoV-2 Genome Variants Detected in the USA

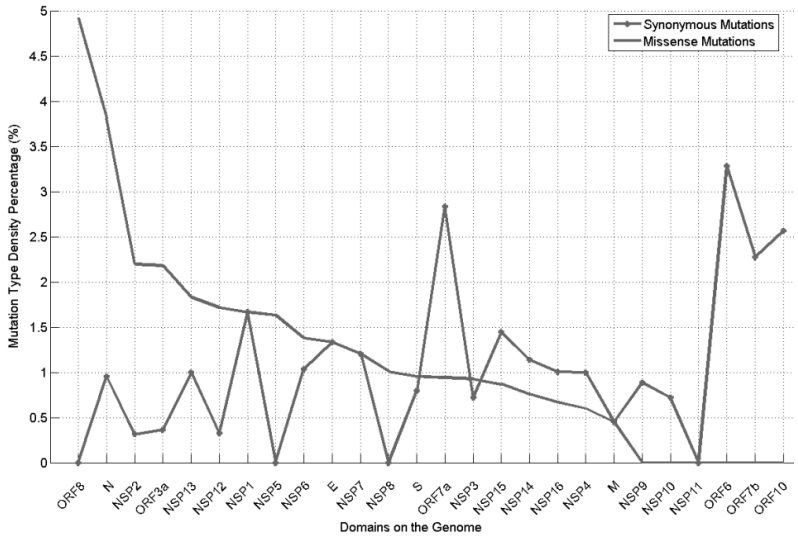

Figure 18. The Density Percentages of Most Common Synonymous and Missense Mutations in Each Domain of the SARS-CoV-2 Genome Variants Detected in the USA

When the percentage values of synonymous and missense mutations in each domain are examined, the graph given in Figure 17 is obtained. The graphs are sketched by considering the descending order of the missense mutations. If Figure 17 is compared with Figure 8, it can be noticed that NSP2, NSP3, and NSP12 are in the top four domains for both variant groups. Finally, mutation density graphs for synonymous and missense mutations are shown in Figure 18. For variants detected in the USA, missense mutations were most densely detected in the ORF8 domain. On the other hand, ORF8 is the fourth dense domain for variants detected in China (Figure 9).

As the next analysis step, the amino acid sequence produced using the reference genome was compared with the amino acid sequence products of all variants detected in the USA, and the most common amino acid changes were determined for the USA variants. The resulting graph is shown in Figure 19. Similar to the analysis result on the China variants (Figure 10), the amino acid change in the ORF8 domain has a percentage of more than 10% for the USA variants as well. This change is again from "L" to "S" amino acid. However, there are 5 additional amino acid changes with an amino acid change percentage greater than 10%. These missense mutation results are listed in Table 3.

Table 2. The Nucleotide Positions of Most Common Nucleotide Mutations in the Variant Genomes Detected in the USA

Nucleotide Position	Mutation	Domain
1059	C>T	NSP2
3037	C>T	NSP3
8782	C>T	NSP4
14408	C>T	NSP12
17747	C>T	NSP13
17858	A>G	NSP13
18060	C>T	NSP13
23403	A>G	S (S1)
25563	G>T	ORF3a
28144	T>C	ORF8

Table 3. The Most Occurred Amino Acid Changes Because of Missense Mutations in USA Variants

Domain	Amino Acid Change
NSP2	M>P
NSP13	M>P
NSP13	M>P
S (S1)	D>G
ORF3a	Q>H
ORF8	L>S

The results obtained for the amino acid category change for the USA variants are shown in Figure 20. The category with the most amino acid changes is (non-polar R group) > (non-polar R group), as in the analysis results for China variants.

The numbers of total amino acid changes in each site for each amino acid change type are shown in Figure 21 by a color-coded display method.

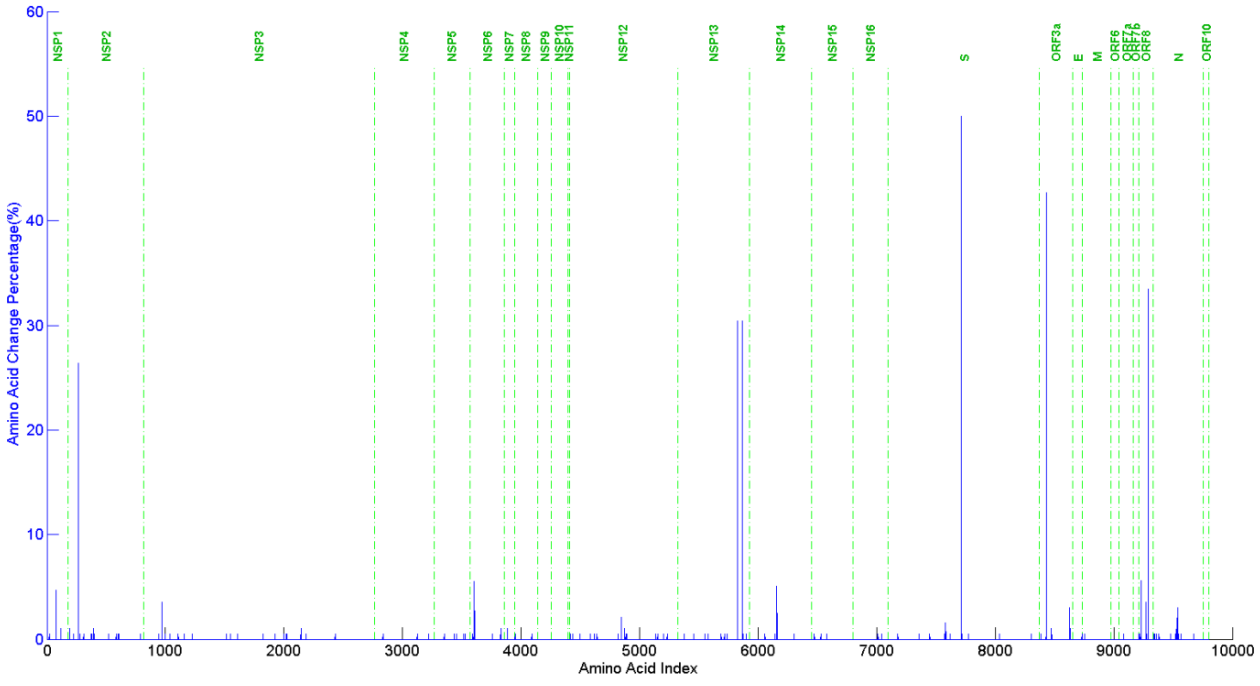

Figure 19. The Most Common Amino Acid Changes and Their Positions in the Coding Region of the SARS-CoV-2 Genome Variants Detected in the USA

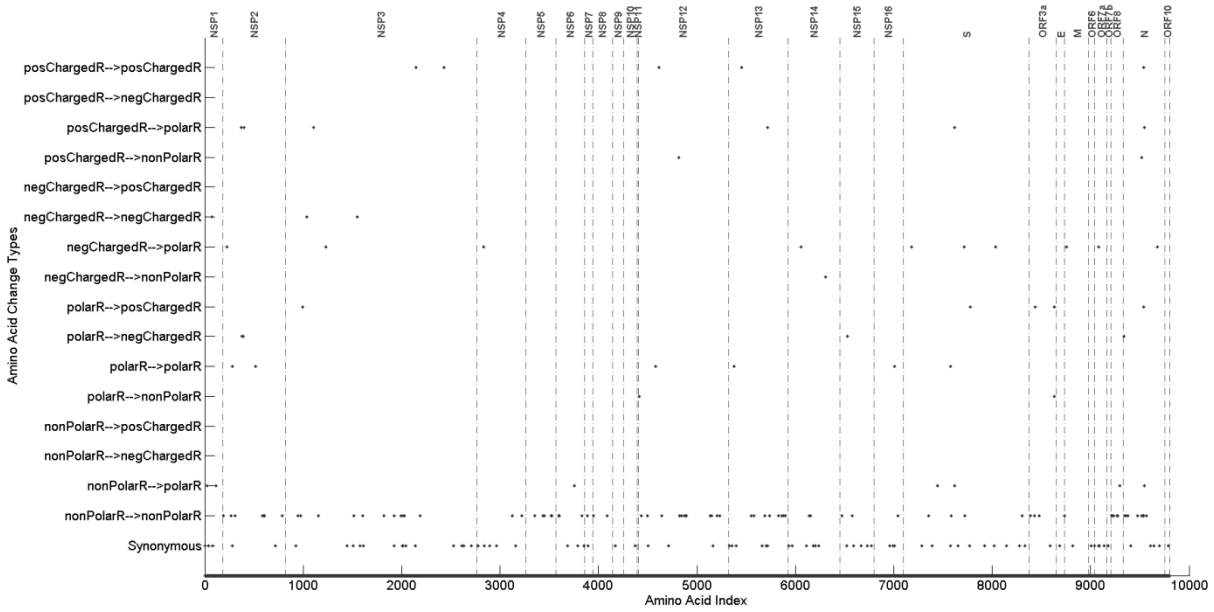

Figure 20. The Most Common Changes of Amino Acid Types in the Coding Region of the SARS-CoV-2 Genome Variants Detected in the USA

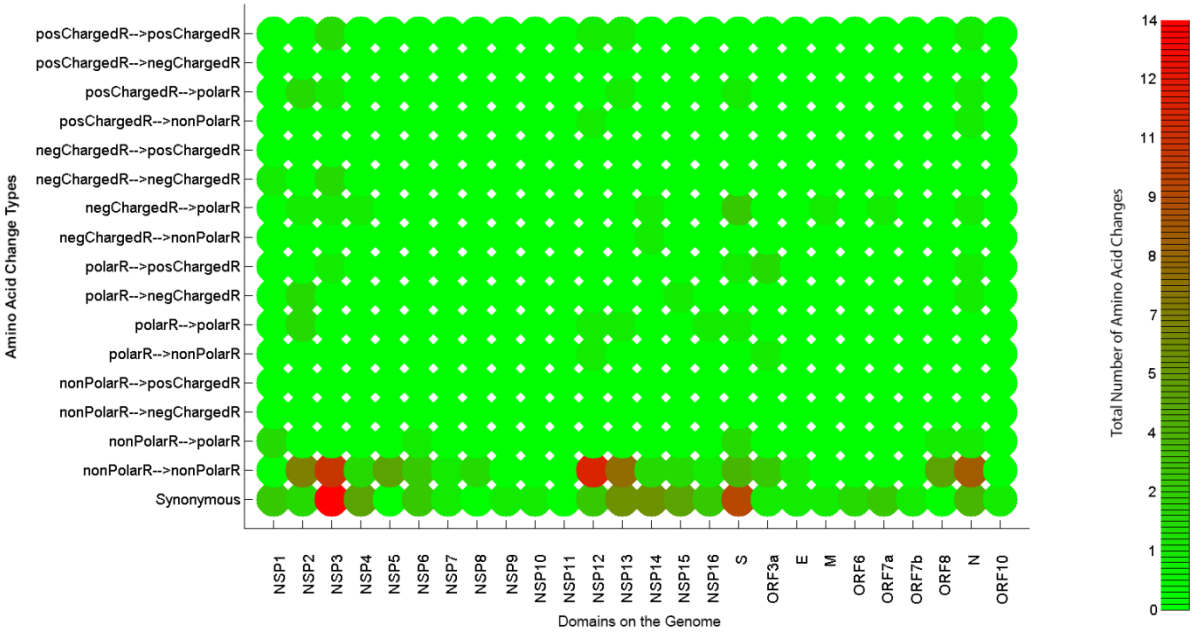

Figure 21. The Number of Most Common Changes of Amino Acid Types in the Coding Region of the SARS-CoV-2 Genome Variants Detected in the USA

The number of amino acid changes for the category (non-polar R group) > (non-polar R group) in the NSP12 domain is higher than any other missense mutation results for all domains. This result is consistent with the results obtained in the analysis for the China variants (Figure 12). However, some other important amino acid change categories have been detected, such as (non-polar R group) > (non-polar R group) changes in NSP3 and N domains.

Conclusion

In this study, we examined some statistical characteristics of two SARS-CoV-2 variant groups which are detected in two different locations of the world, China and the USA. Nucleotide mutations at positions 8782 (C>T mutation) and 28144 (T>C mutation) are common for both variant groups. However, some other mutation positions were

detected on the genome of variants in the USA variant group. The domains where the most common nucleotide mutations were detected are not consistent for these two groups of variants. However, the most common type of amino acid change is (non-polar R group) > (non-polar R group) for both the China and USA variants. The percentage of synonymous mutations in the variant group detected in China is higher than that of missense mutations. On the other hand, the opposite is true for the variant group detected in the USA. This causes an increase in the number of missense mutations in the USA variant group. As the virus continues to spread, it is thought that other mutation sites and missense mutations may occur. Results for the USA variant group support this estimate, as COVID-19 disease caused by the SARS-CoV-2 virus has spread more widely in the USA than China as of May 2020 (WHO COVID-19 Dashboard, 2020).

If we consider the analysis results for all variants (both China and USA variants) and focus on missense mutations, it is clear that the mutations leading to amino acid change are concentrated in the following three domains: ORF8, S, and ORF3a. (Figure 10 and Figure 19). As shown in Table 1, ORF8 may play a role in host-virus interaction, S glycoprotein attaches the virion to the cell membrane by interacting with the host receptor and has a role in initiating the infection and ORF3a forms homotetrameric potassium sensitive ion channels (viroporin) and may modulate virus release. The common feature of these three domains is that they directly modulate the interaction of the SARS-CoV-2 virus with the host. Phylogenetic analysis suggested that the corresponding virus family probably originated in bats and spread to humans (Cotten et al, 2013) and the genome of the virus continues to mutate. One of the results obtained in this study shows us that the areas (ORF8, S, and ORF3a) where the most false mutations are detected in the genome are the regions that affect the interaction of the virus with the host. This indicates that mutations in these domains are probably also effective in cross-host evolution.

As of May 2020, almost all countries of the world are fighting against the SARS-CoV-2 virus and the COVID-19 disease. It seems that this will be a long struggle until the required vaccines or other effective treatment methods are developed. However, we believe that investigating the variants of the virus and sharing information about variants will be effective in combating the virus. In this study, we tried to show the differences between the two variant groups detected in only two different locations of the world, China and the USA. However, to better understand the virus and to prevail in this fight, it is necessary to continue such analysis studies on more variants.

Acknowledgement

Endless thanks to hero healthcare professionals who work devotedly in the process of combating COVID-19.

References

Antonarakis SE, Nomenclature Working Group (1998) Recommendations for a nomenclature system for human gene mutations. Hum. Nut. 11: 1-3.

Cotten M, Watson SJ, Kellam P, Al-Rabeeah AA, Makhdoom HQ, Assiri A, et al. (2013) Transmission and evolution of the middle east respiratory syndrome coronavirus in Saudi Arabia: A descriptive genomic study. Lancet 382:1993-2002.

Den Dunnen JT, Antonarakis SE (2001) Nomenclature for the description of sequence variants. Hum. Genet 109(1):121-124.

Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, et al. (2020) First case of 2019 novel coronavirus in the United States. The New England Journal of Medicine 382:929-936.

Koyama T, Platt D, Parida L (2020) Variant analysis of COVID-19 genomes. Bull World Health Organ. E-pub: 24 February 2020.

Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. (2020) Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. The New England Journal of Medicine 382:1199-1207.

Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. (2020) Genome characterization and epidemiology of 2019 novel coronavirus: implication for virus origins and receptor binding. Lancet 395:565-574,

NGDC (China's National Genomics Data Center), https://bigd.big.ac.cn/ncov, (accessed 25 April 2020).

Rozhgar AK, Safrad M, Ozaslan M (2020) Genomic characterization of a novel SARS-CoV-2. Gene Reports 19:100682.

USCS Genome Browser, https://genome.ucsc.edu, (accessed 25 April 2020).

Vijgen L, Keyaerts E, Moes E, Thoelen I, Wollants E, Lemey P, Vandamme AM, Van Ranst M (2005) Complete genomic sequence of human coronavirus OC43: Molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. Journal of Virology 79(3):1595-1604.

WHO, Naming the Coronavirus Disease (COVID-19) and the virus that causes it, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it, (accessed 19 May 2020).

WHO COVID-19 Dashboard, World Health Organization, https://covid19.who.int, (accessed 11 May 2020).

Worldometer, COVID-19 Coronavirus Pandemic, https://www.worldometers.info/coronavirus/, (accessed 15 May 2020).

Wu F, Zhao S, Yu B, Chen YM, Wang W, Hu Y, et al. (2020) Complete genome characterization of a novel coronavirus associated with severe human respiratory disease in Wuhan, China. bioRxiv 2020: 2020.01.24.919183.