Comparison Of SARS-CoV-2 Virus Variant Genomes Detected In China and USA

Mehmet Cihan Sahingil (mehmet.sahingil@tubitak.gov.tr)
TUBITAK Bilgi Guvenligi Ileri Teknolojileri Arastirma Merkezi https://orcid.org/0000-0002-4900-2256
Yakup OZKAZANC
Hacettepe University

Research article

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

Posted Date: June 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-34079/v1

Abstract

In 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 generates new SARS-CoV-2 variants. In this study we investigated the variant genomes which were detected in China and USA. The publicly available SARS-CoV-2 virus genomes, which were detected in human body are multi-aligned and the obtained results reported. There are 87 genomes for China variants and 200 for USA variants in the used data. The analyses are made for each domain of the genomes. The analysis results show that, the variant genomes in the investigated two groups of SARS-CoV-2 have some similar mutation characteristics as well as some characteristic features that differ from each other. The nucleotide mutations in 8782 (C>T mutation) and 28144 (T>C mutation) are common for both of the variant groups. However, in USA variant group some other mutation positions on the variants' genome were detected. The percentage of missense mutations detected in USA variants is higher than the percentage of synonymous mutations. On the other hand, the percentage of synonymous mutations is higher than the percentage of missense mutations for the variants detected in China. Additionally, the domains where the most mutations are detected on the genome are the regions that affect the interaction of the virus with the host.

1. Background

A new coronavirus, which is also called as SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2), was discovered in December 2019 ([1]). This name was chosen because the virus is genetically related to the coronavirus responsible for the SARS outbreak of 2003 ([2]). While related, the two viruses are differing from each other. The name of the disease which is caused by the virus is COVID-19 (Coronavirus Disease-19). The virus is discovered firstly in Wuhan city of China. Because the property of infectious of the virus is very high, the virus spread firstly in China and then in many countries globally. The first confirmed case of COVID-19 infection in the USA reported on January 20, 2020 ([3]). The virus is affecting 213 countries and territories around the world and two international conveyances with 5,814,886 cases, including 357,984 deaths as of May 28, 2020 ([4]).

Coronaviruses are RNA viruses with a positive stranded genome of approximately 30 kb. The two-thirds of the genome consists of a large open reading frame (ORF) which is called as ORF1ab ([10]). 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) ([11]). 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 in the genome and size info of the domains are listed in Table 1 with short descriptions ([12]).

Table 1
The Details Of The SARS-CoV-2 Genome Domains ([12]).

Domain Name	Position	Size	Short Description
NSP1 (Non-structural protein 1)	266–805	540	Inhibits host translation by interacting with the 40S ribosomal subunit. The NSP1-40S ribosome complex further induces an endonucleolytic cleavage near the 5' UTR of host mRNAs, targeting them for degradation. Viral mRNAs are not susceptible to NSP1-mediated endonucleolytic RNA cleavage thanks to the presence of a 5'-end leader sequence and are therefore protected from degradation. By suppressing host gene expression, NSP1 facilitates efficient viral gene expression in infected cells and evasion from host immune response.
NSP2 (Non-structural protein 2)	806–2719	1914	May play a role in the modulation of host cell survival signaling pathway by interacting with host PHB and PHB2. Indeed, these two proteins play a role in maintaining the functional integrity of the mitochondria and protecting cells from various stresses.
NSP3 (Non-structural protein 3)	2720–8554	5835	Responsible for the cleavages located at the N-terminus of the replicase polyprotein. In addition, PL-PRO possesses a deubiquitinating/delSGylating activity and processes both 'Lys-48'- and 'Lys-63'-linked polyubiquitin chains from cellular substrates. Participates together with NSP4 in the assembly of virally-induced cytoplasmic double-membrane vesicles necessary for viral replication. Antagonizes innate immune induction of type I interferon by blocking the phosphorylation, dimerization and subsequent nuclear translocation of host IRF3. Prevents also host NF-kappa-B signaling.
NSP4 (Non-structural protein 4)	8555–10054	1500	Participates in the assembly of virally-induced cytoplasmic double-membrane vesicles necessary for viral replication.
NSP5(3CL-PRO) (3C-like proteinase)	10055–10972	918	Cleaves the C-terminus of replicase polyprotein at 11 sites. It is able to bind an ADP-ribose-1"-phosphate (ADRP).
NSP6 (Non-structural protein 6)	10973–11842	870	Plays a role in the initial induction of autophagosomes from host reticulum endoplasmic. Later, limits the expansion of these phagosomes that are no longer able to deliver viral components to lysosomes.
NSP7 (Non-structural protein 7)	11843–12091	249	Forms a hexadecamer with NSP8 (8 subunits of each) that may participate in viral replication by acting as a primase. Alternatively, may synthesize substantially longer products than oligonucleotide primers.
NSP8 (Non-structural protein 8)	12092–12685	594	Forms a hexadecamer with NSP7 (8 subunits of each) that may participate in viral replication by acting as a primase. Alternatively, may synthesize substantially longer products than oligonucleotide primers.

Domain Name	Position	Size	Short Description
NSP9 (Non-structural protein 9)	12686–13024	339	May participate in viral replication by acting as a ssRNA-binding protein.
NSP10 (Non-structural protein 10)	13025–13441	417	Plays a pivotal role in viral transcription by stimulating both NSP14 3'-5' exoribonuclease and NSP16 2'-O-methyltransferase activities. Therefore plays an essential role in viral mRNAs cap methylation.
NSP11 (Non-structural protein 11)	13442–13480	39	No additional description
NSP12(Pol) (RNA-directed RNA polymerase)	13442–16236	2795	Responsible for replication and transcription of the viral RNA genome.
NSP13(Hel) (Helicase)	16237–18039	1803	Multi-functional protein with a zinc-binding domain in N-terminus displaying RNA and DNA duplex-unwinding activities with 5' to 3' polarity. Activity of helicase is dependent on magnesium.
NSP14(ExoN) (Proofreading exoribonuclease)	18040–19620	1581	Enzyme possessing two different activities: an exoribonuclease activity acting on both ssRNA and dsRNA in a 3' to 5' direction and a N7-guanine methyltransferase activity. Acts as a proofreading exoribonuclease for RNA replication, thereby lowering The sensitivity of the virus to RNA mutagens.
NSP15 (Uridylate-specific endoribonuclease)	19621–20658	1038	Mn(2+)-dependent, uridylate-specific enzyme, which leaves 2'-3'-cyclic phosphates 5' to the cleaved bond.
NSP16 (2'-O-methyltransferase)	20659–21552	894	Methyltransferase that mediates mRNA cap 2'-O-ribose methylation to the 5'-cap structure of viral mRNAs. N7-methyl guanosine cap is a prerequisite for binding of NSP16. Therefore plays an essential role in viral mRNAs cap methylation which is essential to evade immune system.

Domain Name	Position	Size	Short Description
S glycoprotein (Spike glycoprotein)	21599–25381	3783	21599–23617 region: Spike protein S1: attaches the virion to the cell membrane by interacting with host receptor, initiating the infection. Binding to human ACE2 receptor and internalization of the virus into the endosomes of the host cell induces conformational changes in the Spike glycoprotein. Uses also human TMPRSS2 for priming in human lung cells which is an essential step for viral entry. Proteolysis by cathepsin CTSL may unmask the fusion peptide of S2 and activate membranes fusion within endosomes. 24008–25381 region: Spike protein S2': Acts as a viral fusion peptide which is unmasked following S2 cleavage occurring upon virus endocytosis.
ORF3a (Protein 3a)	25393–26217	825	Forms homotetrameric potassium sensitive ion channels (viroporin) and may modulate virus release. Up-regulates expression of fibrinogen subunits FGA, FGB and FGG in host lung epithelial cells. Induces apoptosis in cell culture. Downregulates the type 1 interferon receptor by inducing serine phosphorylation within the IFN alpha- receptor subunit 1 (IFNAR1) degradation motif and increasing IFNAR1 ubiquitination.
ORF4 (E protein) (Envelope small membrane protein)	26245–26469	225	Envelope small membrane protein (E protein): Plays a central role in virus morphogenesis and assembly. Acts as a viroporin and self-assembles in host membranes forming pentameric protein-lipid pores that allow ion transport. Also plays a role in the induction of apoptosis.
M protein (Membrane protein)	26523–27188	666	Membrane Protein (M protein): Component of the viral envelope that plays a central role in virus morphogenesis and assembly via its interactions with other viral proteins.
ORF6 (ns6) (Non-structural protein 6)	27202–27384	183	Non-structural protein 6 (ns6): Could be a determinant of virus virulence, since, when expressed in an otherwise attenuated JHM strain of murine coronavirus, it can dramatically increase the lethality of the latter. Seems to stimulate cellular DNA synthesis in vitro.
ORF7a (Protein 7a)	27439–27756	318	Non-structural protein which is dispensable for virus replication in cell culture.
ORF7b (ns7b) (Non-structural protein 7b)	27756–27887	129	No additional description
ORF8	27894–28259	366	May play a role in host-virus interaction.

Domain Name	Position	Size	Short Description
N	28274–29533	1260	Nucleoprotein (NC): Packages the positive strand viral genome RNA into a helical ribonucleocapsid (RNP) and plays a fundamental role during virion assembly through its interactions with the viral genome and membrane protein M. Plays an important role in enhancing the efficiency of subgenomic viral RNA transcription as well as viral replication.
ORF10	29558–29674	117	No additional description

As the virus is transmitted from person to person via droplet transmission ([5]) easily, more patients are infected and as a result the virus is expected to accumulate more variants. So, while the virus spreads continuous globally, it is seen that the genome of the virus mutates and this mutation processes generates new SARS-CoV-2 variants ([6]).

Some mutation types are synonymous which means the mutation of the corresponding nucleotide doesn't 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 some other type of mutation types as well ([8], [9]), in this study we mostly concentrated on these two type of mutations in coding domains of the variant genomes.

Some early works consider the SARS-CoV-2 variant analysis for the variants detected in a specific location, especially in China ([6], [10]). In this paper we compared the variation characteristics of two variant genome groups which are gathered from two distinct locations of the world, China and USA. The variant genomes in both of the groups are multi-aligned and some statistical characteristics are obtained for each of the domains of the SARS-CoV-2 genome. The obtained analysis results are given in Sect. 2. The discussion and conclusion notes are given in Sect. 3 and 4, respectively. In Sect. 5 of this paper, the used material and method are described. Finally, the abbreviations are given in Sect. 6.

2. Results

We firstly investigated the non-coding parts, which is also called 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 (Fig. 1) and the starting nucleotide position of NSP-1 is 266. Therefore, the first 265 nucleotides are taken as non-coding region (5' UTR). The most common nucleotide mutation types and the corresponding mutation locations are shown in Fig. 2. To make this analysis we compared the reference sequence with 287 variant genomes (87 for China, 200 for USA) of SARS-CoV-2 virus detected in China and 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 in x-axis of the graph. The mutation percentage values are shown in left side of the figure.

On the other hand, points show the most common nucleotide mutation types in corresponding nucleotide position. The nucleotide mutation types are shown in right side of the figure. There are 17 different cases: 1 “No Mutation” case, 12 mutation between nucleotide types cases and four deletion cases. The percentage value for the “No-Mutation” type shows the percentage of variants which have the same nucleotide type with the reference genome for the investigated genome region. 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 5’ UTR region. In the following parts of the paper this graphical display method will be used for expressing mutation percentages.

The last domain (ORF10) of the SARS-CoV-2 genome finishes in 29674th nucleotide. The rest of the genome sequence from this point is also taken as non-coding region (3’ UTR). The most common nucleotide mutation types and the corresponding mutation locations are shown in Fig. 3. To make this analysis we used the same data of reference and variant genomes which were used in 5’ UTR analysis. The mutations occur mostly in the last 40 nucleotide of this non-coding region. The most occurred mutation types are T > C mutation (35.61%), A>- deletion (32.87%) and C > T mutation (19.17%) like 5’ UTR. The mutation percentages in other nucleotide positions are below 10%.

2.1 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. We considered only the variants gathered from China as the first step. The total number of variants in China group is 87. The most common nucleotide mutation percentages, the mutation types, the corresponding mutation locations and the domain boundaries are shown in Fig. 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 on the top of the graph. 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).

$$DomainMutation\%_i = \frac{(Totalmutationnumberindomaini)}{\sum_{j=1}^{Domaincount}(Totalmutationnumberindomainj)} \bullet 100 \qquad (1)$$

If Fig. 4 is examined, it can be easily seen that the mutations in 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 ([6], [7], [10]). All the other mutations in specific nucleotide positions have the percentage values less than 10%. Most common nucleotide mutations are T > C (30.39%) and C > T (28.68%). If the nucleotide positions of the mutations are considered, it is seen that the total mutation number for some genome domains are very low. The domains with mutation percentages less than 2% are NSP7, NSP9-11, S, ORF3a, E, M, ORF6, ORF7b, ORF8 and ORF10. The domain mutation percentages [illegible] wn in Fig. 5.

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

$$MutationDensity\%_i = \frac{(Totalmutationnumberindomain_i)}{(Totalnucleotidecountindomain_i)} \bullet 100 \quad (2)$$

As known, all the nucleotide mutations don't cause an amino acid change in translation step of gene regulation process. In these synonymous mutation types, the nucleotide mutates but the changed codon codes the same amino acid type with the unchanged codon. On the other hand, some nucleotide mutations causes a change in amino acid type. This type mutations which are called as missense mutation may cause the structure of the result protein to change. Because of this reason, missense mutations are more critical. Figure 7 shows the most common mutation types in coding region of the SARS-CoV-2 genome variants detected in China. The red dots are indicating the most common mutation types in the specific nucleotide positions. The red dots in "No Mutation" line are the nucleotide locations where no mutation was detected among all variants. The percentage value next to the "No Mutation" label shows the percentage of non-mutated nucleotides in the entire genome to the total number of nucleotides in the entire genome. Likewise, the red dots in "Missense" and "Synonymous" lines are showing the nucleotide positions where missense and synonymous mutations are detected, respectively. The percentage values next to the "Missense" and "Synonymous" labels show the percentage of missense and synonymous mutations in the entire genome to the number of mutations in the entire genome, respectively.

$$Synonymus\%_i = \frac{(Totalsynonymousmutationnumberindomain_i)}{(Totalmutationcountinallcodingregionofgenome)} \bullet 100 \quad (3)$$

$$Missense\%_i = \frac{(Totalmissensemutationnumberindomain_i)}{(Totalmutationcountinallcodingregionofgenome)} \bullet 100 \quad (4)$$

However, the nucleotide numbers of domains differ from each other. Therefore, the mutation density for each domain can be also considered by using the formula shown in (5) and (6). The mutations are occurred most densely in NSP8 domain (Fig. 9).

$$SynonymusDens\%_i = \frac{(Totalsynonymousmutationnumberindomain_i)}{(Totalnucleotidecountindomain_i)} \bullet 100 \quad (5)$$

$$MissenseDens\%_i = \frac{(Totalmissensemutationnumberindomain_i)}{(Totalnucleotidecountindomain_i)} \bullet 100 \quad (6)$$

As said before, the mutations in nucleotides may affect the generated amino acid sequence. To analyze this phenomenon, the generated amino acid sequence of the reference genome was compared with all ... mmon amino acid changes were detected. The

result graph is shown in Fig. 10. In other words, Fig. 10 shows the amino acid change percentages which are due to missense mutations.

Amino acids can be categorized into four different classes according to their polarity characteristics. These categories are called as the amino acids 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 negative charged R groups (D, E) and the amino acids with positive charged R groups (K, R, H). The nucleotide mutations may cause the change of an amino acid which is in a specific category to another amino acid which is in a different category. The obtained category change analysis results are shown in Fig. 11. It is easily seen that most amino acid change type is detected in (non-polar R group) > (non-polar R group). This kind of change is mostly detected in NSP12 domain. This fact supports the analysis result shown in Fig. 8.

To show the total amino acid change counts in each of the domains for each amino acid change types, we used a color coded display method shown in Fig. 12 with color map bar. This display method gives information about the relative quantities of the resulting mutation types. For example, synonymous amino acid change in NSP2 domain is the most amino acid change occurred among all other amino acid changes. If the color map on the right side is controlled, it can be seen that the exact synonymous amino acid change count in NSP2 domain is 74. This means that in NSP2 domain 74 synonymous amino acid changes are detected. The amino acid change count for (non-polar R group) > (non-polar R group) change type in NSP12 domain is higher than all the other missense mutation results for all domains, which is consistent with the results shown in Fig. 11.

2.2 Analysis of Mutations in SARS-CoV-2 USA Variants

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

Like in China case (Fig. 4), in nucleotide positions 8782 (C > T mutation) and 28144 (T > C mutation) high nucleotide mutation percentage values are detected for the variants detected in the USA. This result is consistent with the previous studies ([6], [7], [10]). However, there are 8 additional mutation positions which have mutation percentage values greater than 10%. These nucleotide positions and the corresponding most common nucleotide mutation types are listed in Table 2. The mutations in nucleotide position 3037, 14408, 23403 and 25563 have higher percentage values than the mutations in nucleotide positions 8782 and 28144.

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
Most common nucleotide mutation type is C > T (40.95%). This result differs from the result of China variants. In China case the most common nucleotide mutation type was T > C (30.39%). If the domain based mutation percentages and densities are considered for USA variants, the result graphs given in Fig. 14 and Fig. 15 are obtained. If China and USA cases are compared (Fig. 5 vs Fig. 14 and Fig. 6 vs Fig. 15) it can be said that the domains of most common nucleotide mutation positions are not consistent for these two variant groups.		

If the percentage of synonymous and missense mutations in each domain are considered, the graph given in Fig. 17 is obtained. The graphs were sketched by considering the descending order of the missense mutations. If Fig. 17 is compared with Fig. 8, it can be noticed that NSP2, NSP3 and NSP12 are in first four domains for both of the variant groups. Finally, the mutation density graphs for synonymous and missense mutations are shown in Fig. 18. For the variants detected in the USA, the missense mutations are occurred more densely in ORF8 domain. In China case, ORF8 was fourth dense domain (Fig. 9).

The generated amino acid sequence of the reference genome was compared with all the variants' amino acid sequence products and most common amino acid changes were detected for USA variants. The result graph is shown in Fig. 19. Like in China case (Fig. 10), for USA variants there is an amino acid change which has more than 10% percentage in ORF8 domain. This change is again from "L" amino acid to "S" amino acid. However, there are 5 additional amino acid changes which have more than 10%. These missense mutation results are listed in Table 3.

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 amino acid category change result for USA variants is shown in Fig. 20. The most amino acid change type is detected in (non-polar R group) > (non-polar R group) like in China case.

The total amino acid change counts in each of the domains for each amino acid change types are shown in Fig. 21 with a color coded display method. The amino acid change count for (non-polar R group) > (non-polar R group) change type in NSP12 domain is again higher than all the other missense mutation results for all domains like China case (Fig. 12). However, there are some other amino acid change count for USA variants like (non-polar R group) > (non-polar R group) change type in NSP3 domain or (non-polar R group) > (non-polar R group) change type in N domain.

3. Methods

The links for the genomics data of the reference genome and the variants are publicly available in a web page of China's National Genomics Data Center (NGDC) which is a COVID-19 dedicated page ([13]). As the first step we gathered the genomes detected in China and the genomes detected in the USA. The numbers of variant genomes in the China group and USA group are 87 and 200, respectively. Because most of the analyses are percentage based, the difference between the numbers of variant genomes in the groups don't effect the result dramatically.

We multi-aligned the variants by using Bioinformatics toolbox of MATLAB software as the second step. The gap penalty and extension penalty are accepted as 10 and 0.5, respectively. In third step of the study we used NC_045512 ([12]) as the reference genome and compare the multi-aligned variant genomes with NC_045512 genome. The main reason of using this genome as the reference genome is to make the results comparable with the previous studies ([6], [7], [10]) where this genome is also used as the reference genome.

4. Discussion

In this study we showed some statistical characteristics of two SARS-CoV-2 variant groups which are detected in two different locations in the world, China and USA. The nucleotide mutations in 8782 (C > T mutation) and 28144 (T > C mutation) are common for both of the variant groups. However, in USA variant group some other mutation positions on the variants' genome were detected. The domains where the most common nucleotide mutations are detected are not consistent for these two variant groups. However, the most amino acid change type is detected as (non-polar R group) > (non-polar R group) for both China and USA variants. The variant group which is detected in China shows more synonymous mutations than missense mutations as percentage. On the other hand, the opposite situation is true for the variant group which is detected in the USA. This causes the number of missense mutations to increase in USA variant group. It is considered that as long as the virus continues to spread, other mutation sites and missense mutations may also appear. Because the COVID-19 disease which is caused by SARS-CoV-2 virus spread in the USA more widely than China as of May 2020 ([14]), the obtained results for USA variant group supports this prediction.

If we consider the analysis results for all variants (both China and USA variants) and focus on missense mutations, it is seen that there are three domains in which mutations leading to amino acid change are most prominent: ORF8, S and ORF3a (Fig. 10 and Fig. 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 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 domains is that they directly modulate the interaction of SARS-CoV-2 virus with the host. Phylogenetic analysis suggested that the corresponding virus family probably originated in bats and spread to people ([15]) and the genome of the virus continues to mutate. This result shows us that the considered domains where the most mutations are detected on the genome are regions that affect the interaction of the virus with the host and the mutations in these domains are probably effective also in cross-host evolution.

5. Conclusions

As of May 2020, almost all the countries of the world are fighting against the SARS-COV-2 virus and the corresponding disease COVID-19. It seems that this will be a long-fighting process until the required vaccines or other effective treatment methods are developed. However, we believe that investigating the variants of the virus and sharing knowledge on variants will be effective in the fight with the virus. In this study we tried to show the differences of two variant groups which are detected in only two different locations, China and USA. But this type of analysis works must continue for more variants for better understanding the virus and prevail in this fight.

Abbreviations

COVID-19: Coronavirus Disease 2019

SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2

NGDC: National Genomics Data Center

UTR: Untranslated Region

NSP: Non-Structural Protein

ORF: Open Reading Frame

UCSC: University of California Santa Cruz

USA: United States of America

WHO: World Health Organization

Declarations

Ethics approval and consent to participate

All the used data for the corresponding virus is publicly available data.

Consent for publication

Not applicable.

Availability of data and materials

The data sets used during the current study are publicly available in the NGDC ([13]). Additionally, the used data MATLAB codes are given in the following link: https://data.mendeley.com/datasets/mw8ycyxwgm/draft?a=9e25025e-87af-49d4-8f2c-a02673e85589

Competing interests

The authors declare that they have no competing interests

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

MCS analyzed the genome data of variants', prepared the result figures, reported the obtained results and re-ordered the manuscript according to the major comments given by YO. YO checked the manuscript and added major comments on the report. Both authors read and approved the final manuscript.

Acknowledgments

Endless thanks to the heroic healthcare professionals working with great devotion in the process of fighting against COVID-19.

References

1. Wu F, Zhao S, Yu B, Chen YM, Wang W, Hu Y, et al. Complete genome characterization of a novel coronavirus associated with severe human respiratory disease in Wuhan, China. 2020; bioRxiv. 2020:2020.01.24.919183.
2. World Health Organization. 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.
3. Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, et al. First Case of 2019 Novel Coronavirus in the United States. The New England Jounal of Medicine. 2020; 382:929-36; DOI: 10.1056/nejmOA2001191.
4. COVID-19 Coronavirus Pandemic, https://www.worldometers.info/coronavirus/ . Accessed 28 May 2020.
5. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. The New England Jounal of Medicine. 2020; DOI: https://doi.org/10.1056/NEJMoa2001316.
6. Koyama T, Platt D, and Parida L. Variant analysis of COVID-19 genomes. [Submitted]. Bull World Health Organ. E-pub: 24 February 2020. DOI: http://dx.doi.org/10.2471/BLT.20.253591.
7. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genome characterisation and epidemiology of 2019 novel coronavirus: implication for virus origins and receptor binding. 2020; Vol. 395, 565-574, DOI: https://doi.org/10.1016/S0140-6736(20)30251-8.
8. Antonarakis SE, and Nomenclature Working Group. Recommendations for a Nomenclature System for Human Gene Mutations. Hum. Nut. 1998; 11: 1-3.
9. Den Dunnen JT, Antonarakis SE. Nomenclature for the Description of Sequence Variants. Hum. Genet. 2001; 109(1), 121-124.
10. Rozhgar AK, Safrad M, Ozaslan M. Genomic Characterization of a Novel SARS-CoV-2. Gene Reports. 2020; 19(2020)100682; DOI: https://doi.org/10.1016/j.genrep.2020.100682.
11. Vijgen L, Keyaerts E, Moes E, Thoelen I, Wollants E, Lemey P, Vandamme AM, and Van Ranst M. Complete Genomic Sequence of Human Coronavirus OC43: Molecular Clock Analysis Suggests a Relatively Recent Zoonotic Coronavirus Transmission Event. Journal of Virology. 2005; Vol. 79 No.3.
12. USCS Genome Browser. https://genome.ucsc.edu. Accessed 25 April 2020.
13. China's National Genomics Data Center (NGDC). https://bigd.big.ac.cn/ncov. Accessed 25 April 2020.
14. WHO Coronavirus Disease (COVID-19) Dashboard. World Health Organization.

15. Cotten M, Watson SJ, Kellam P, Al-Rabeeah AA, Makhdoom HQ, Assiri A, et al. Transmission and Evolution of the Middle East Respiratory Syndrome Coronavirus in Saudi Arabia: A Discriptive Genomic Study. Lancet. 2013.; Vol. 382, 1993-2002; DOI: http://dx.doi.org/10.1016/S0140-6736(13)61887-5

Figures

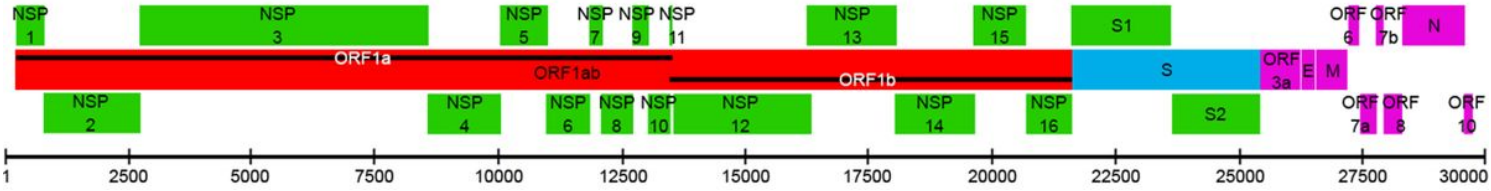


Figure 1

Gene Structure of the SARS-CoV-2 Genome

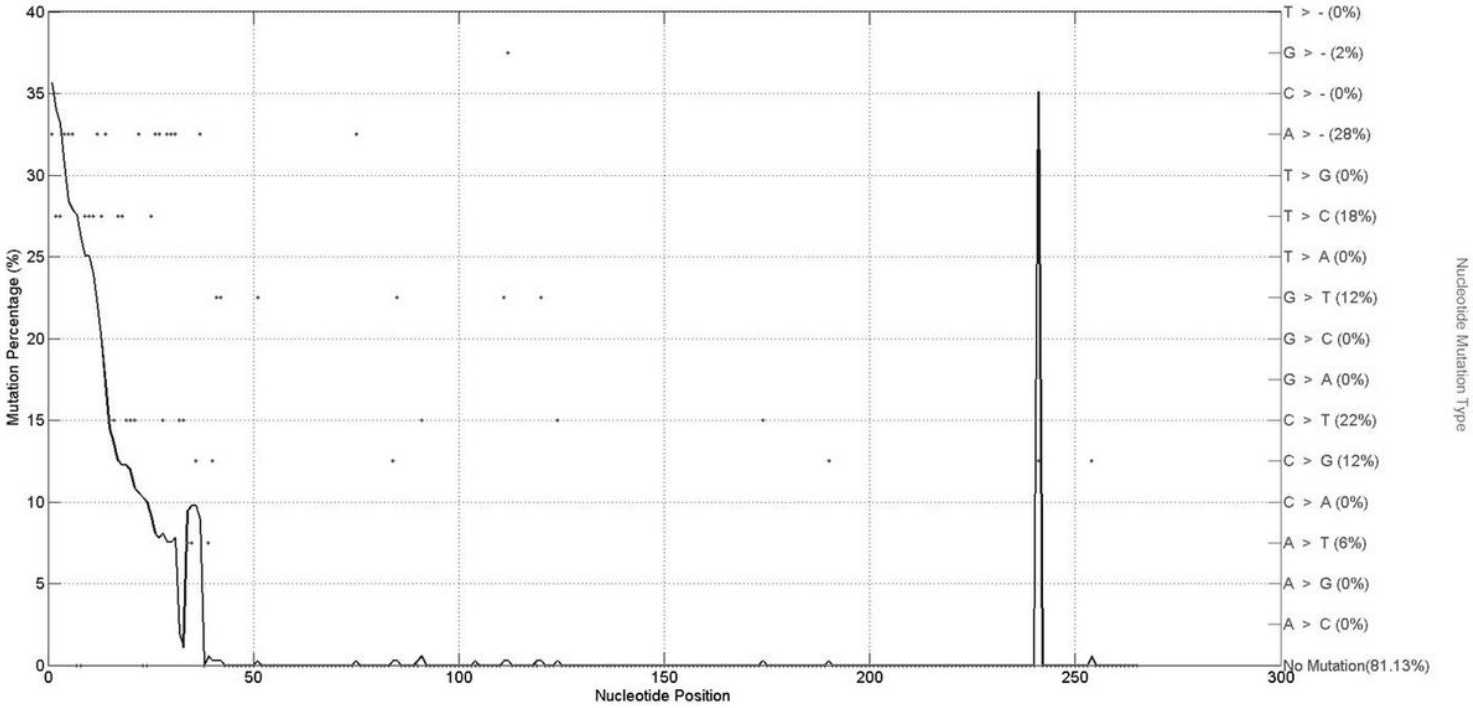


Figure 2

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

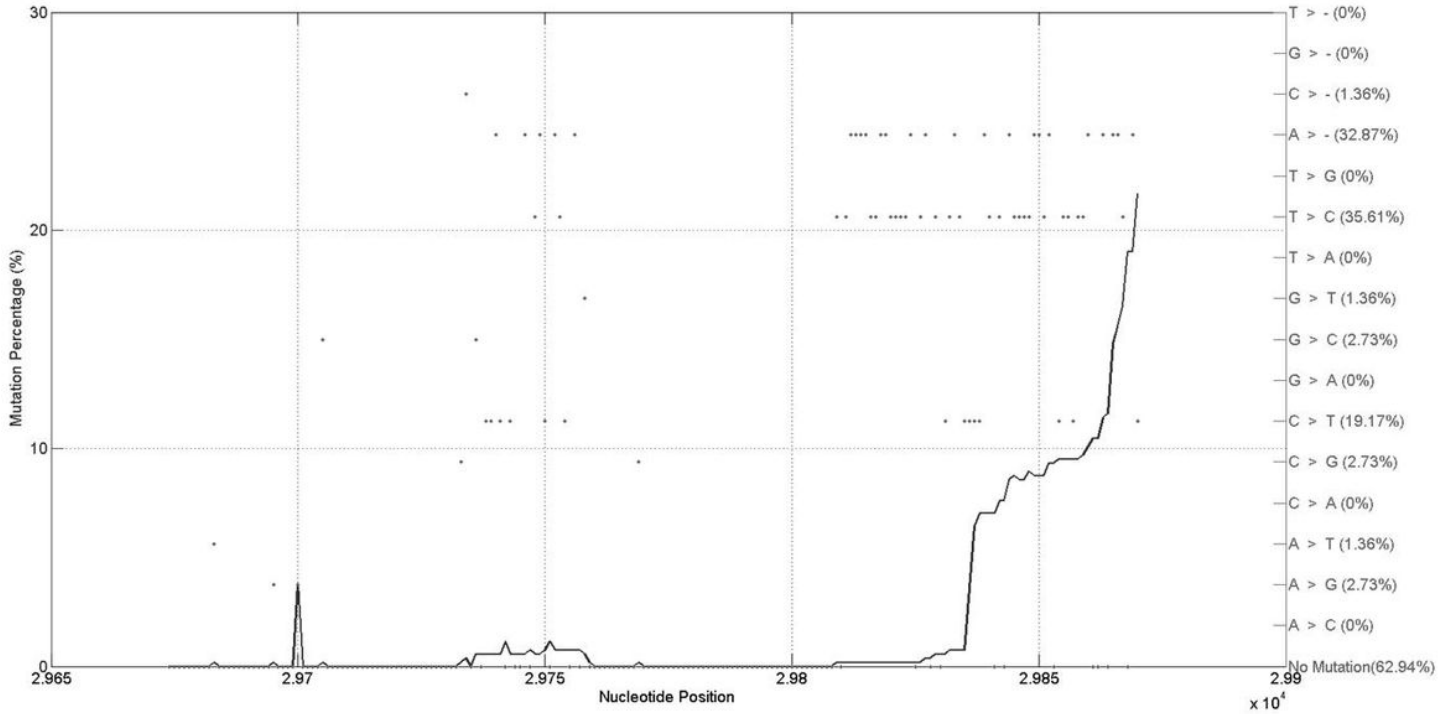


Figure 3

The Most Common Nucleotide Mutations in 3’ UTR of the SARS-CoV-2 Genome

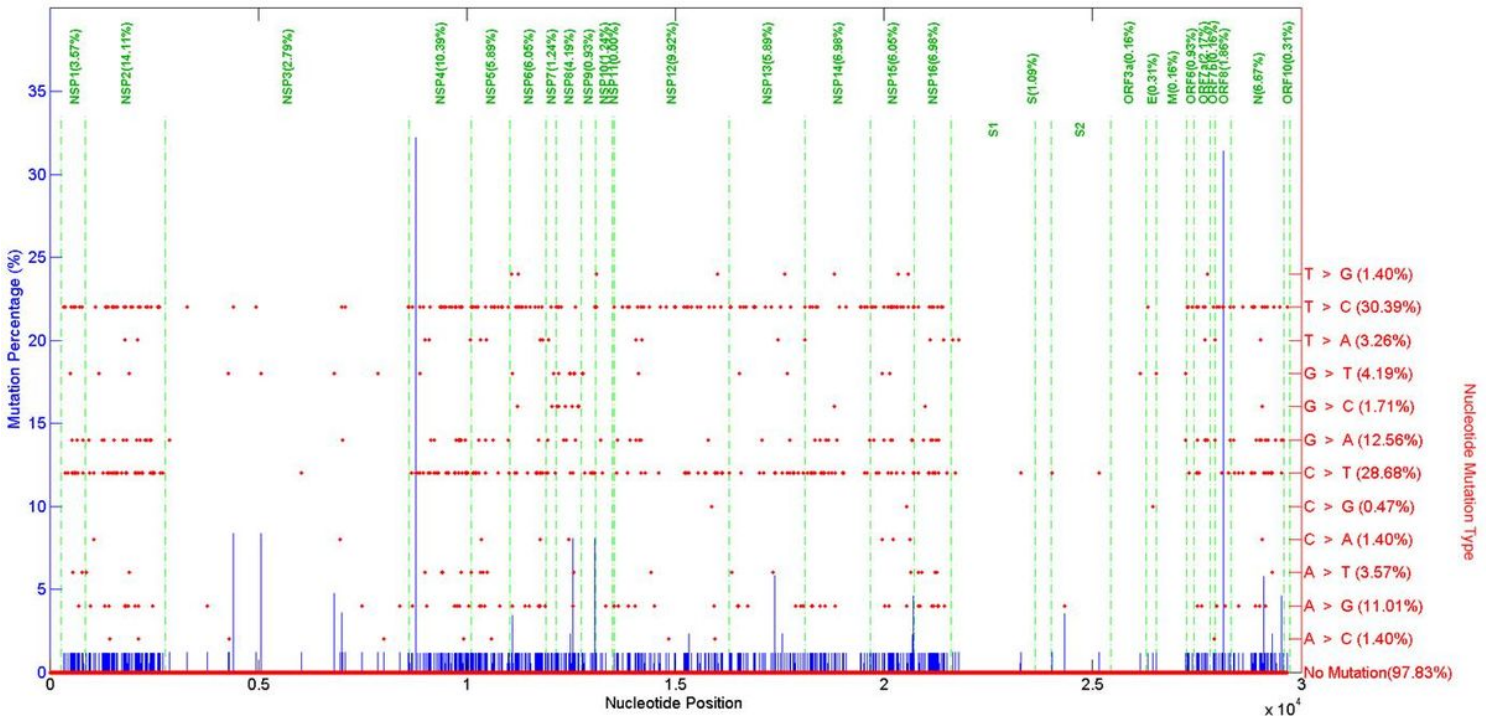


Figure 4

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

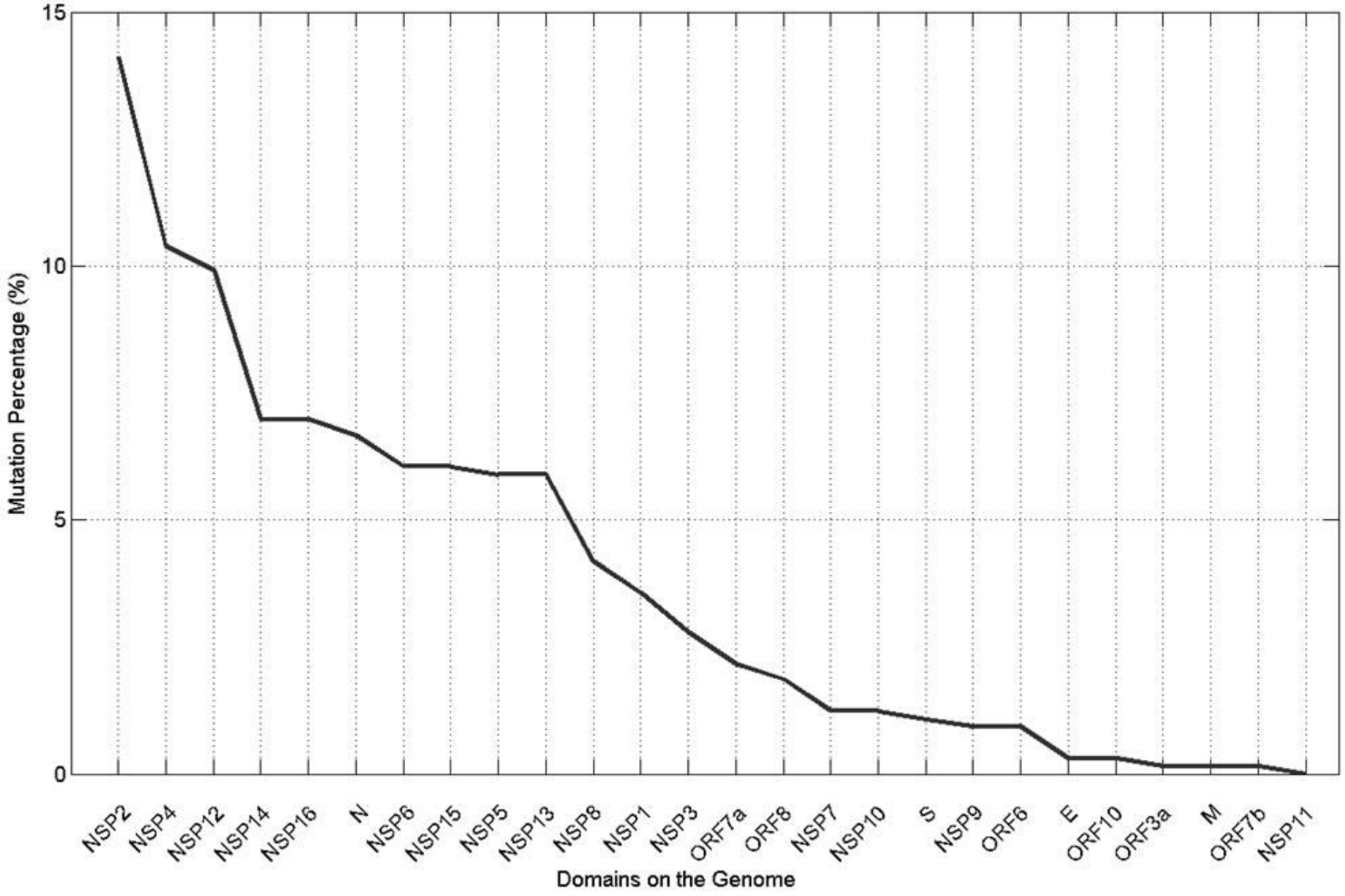


Figure 5

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

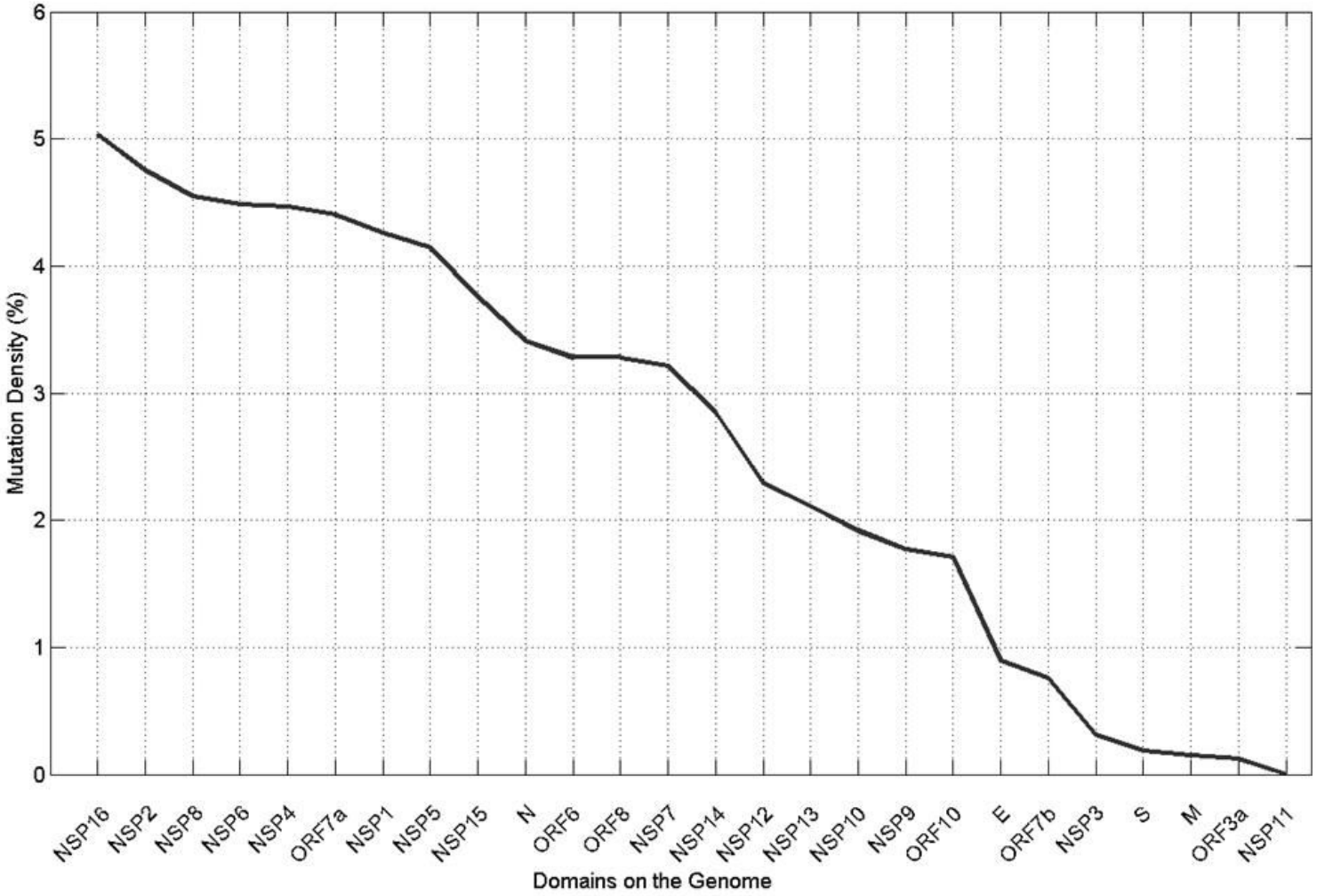


Figure 6

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

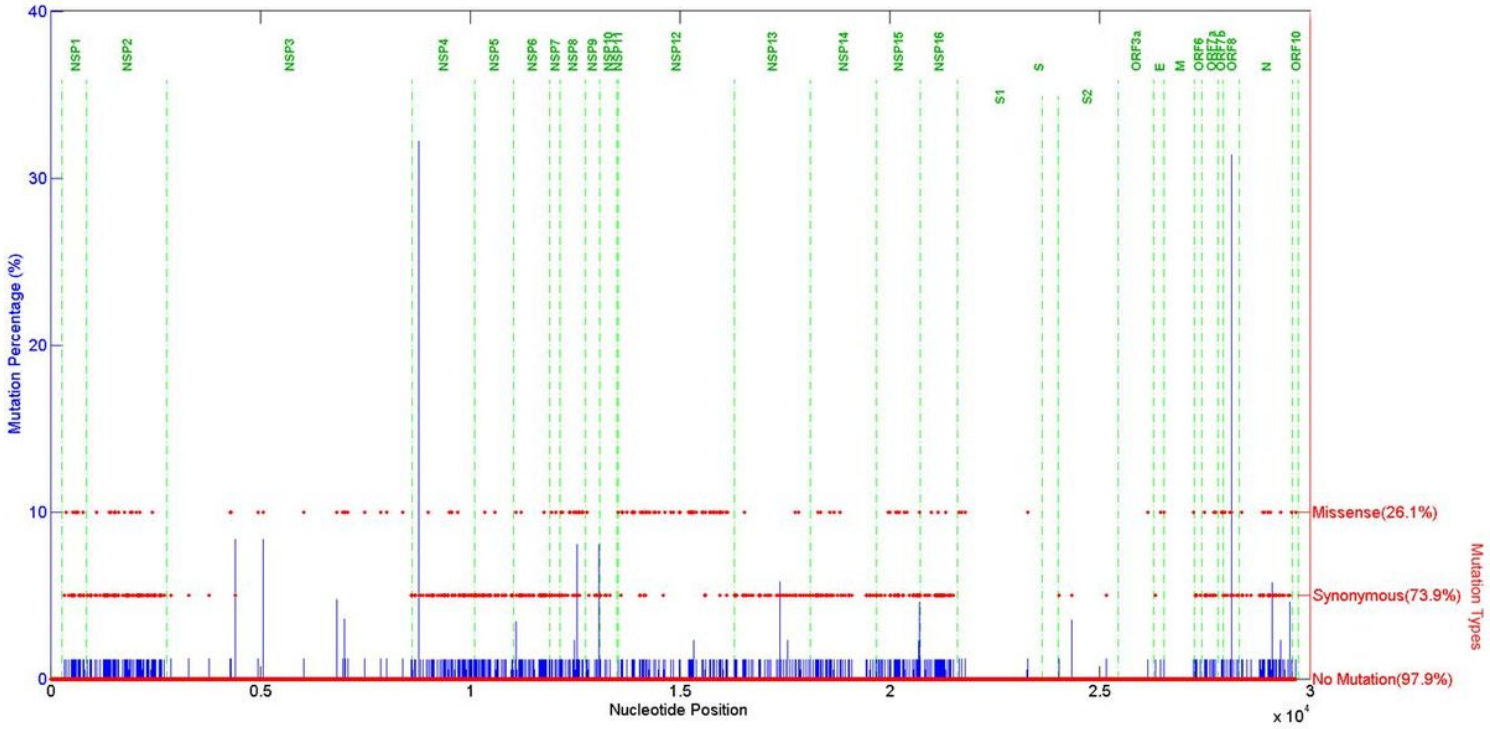


Figure 7

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

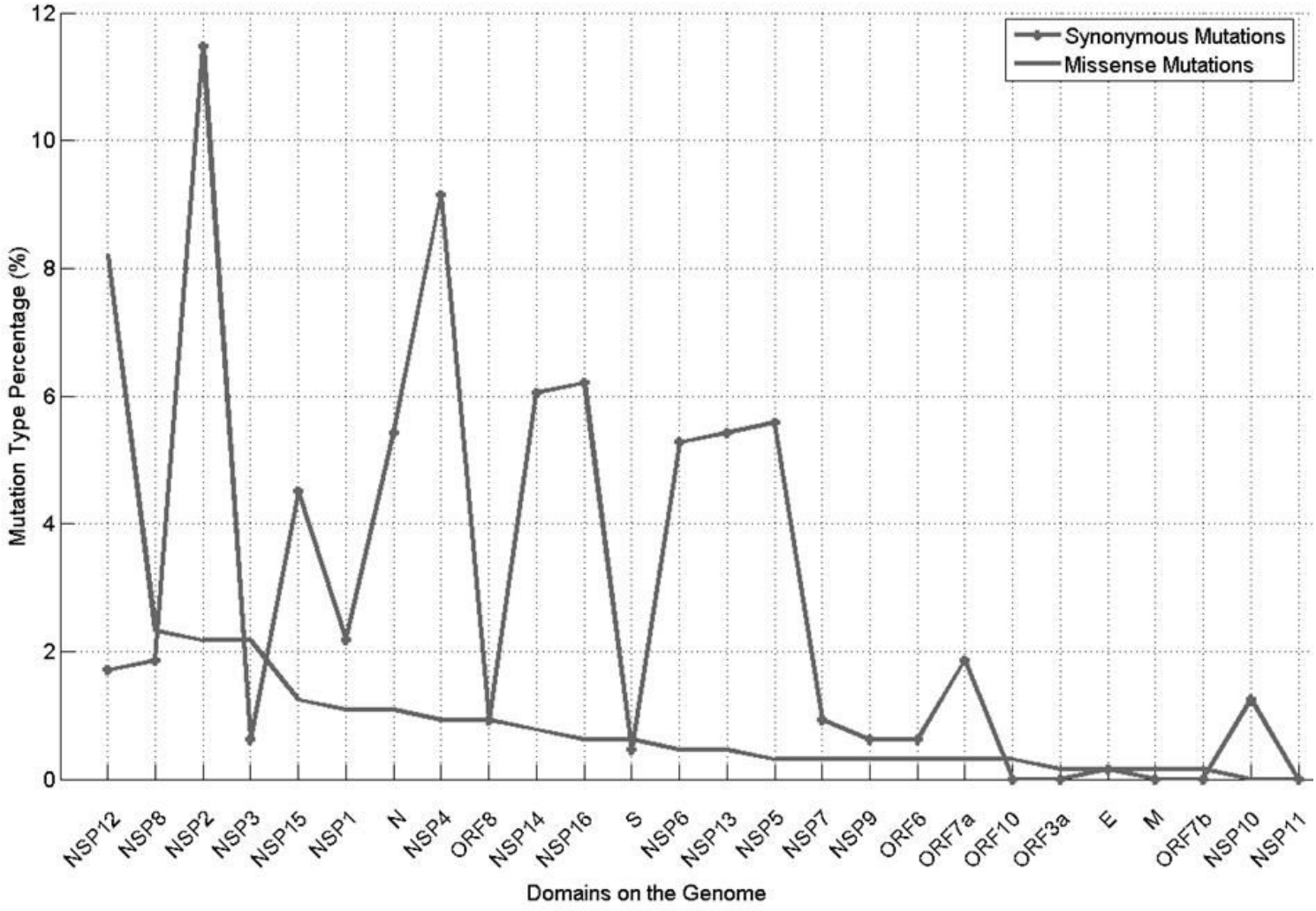


Figure 8

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

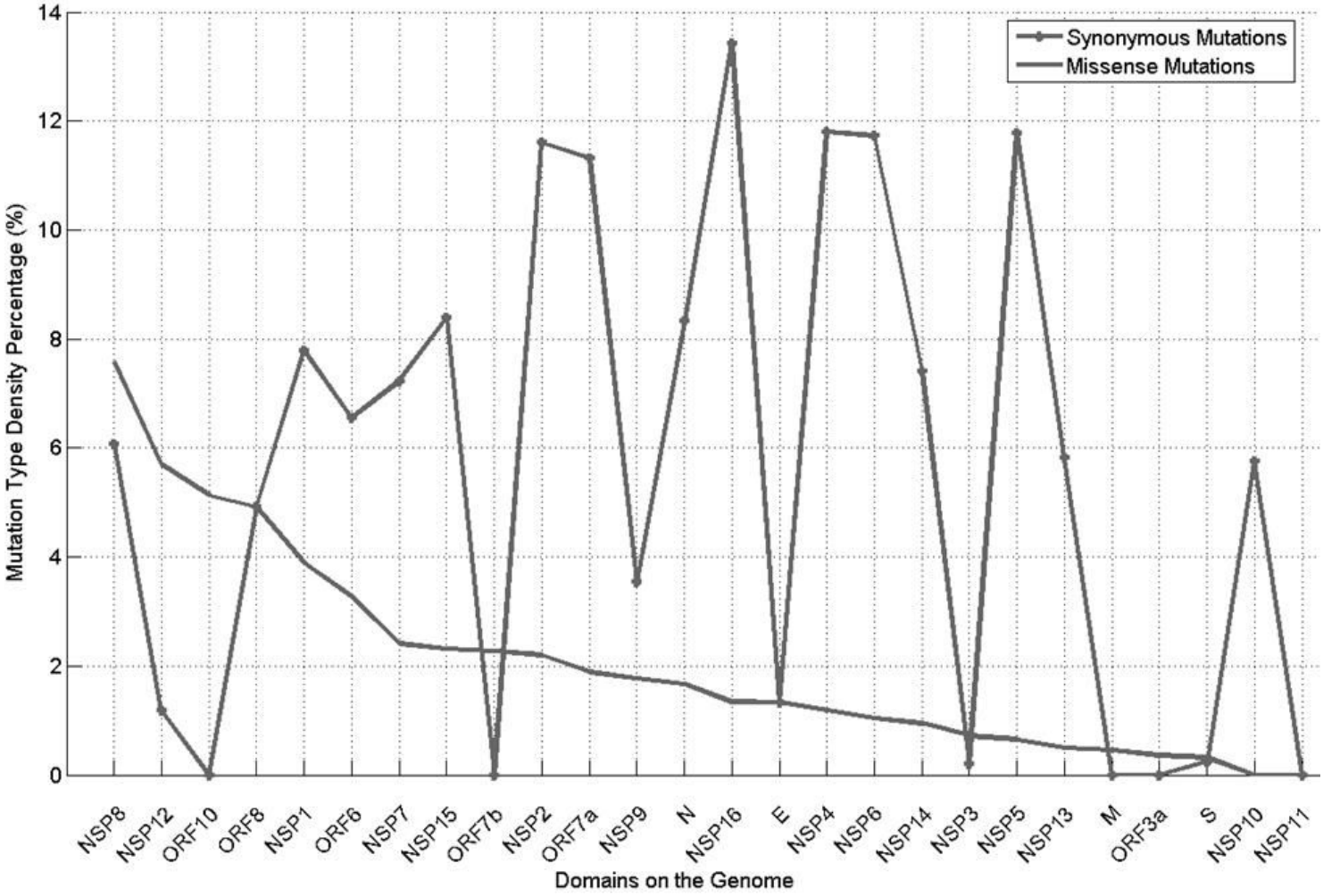


Figure 9

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

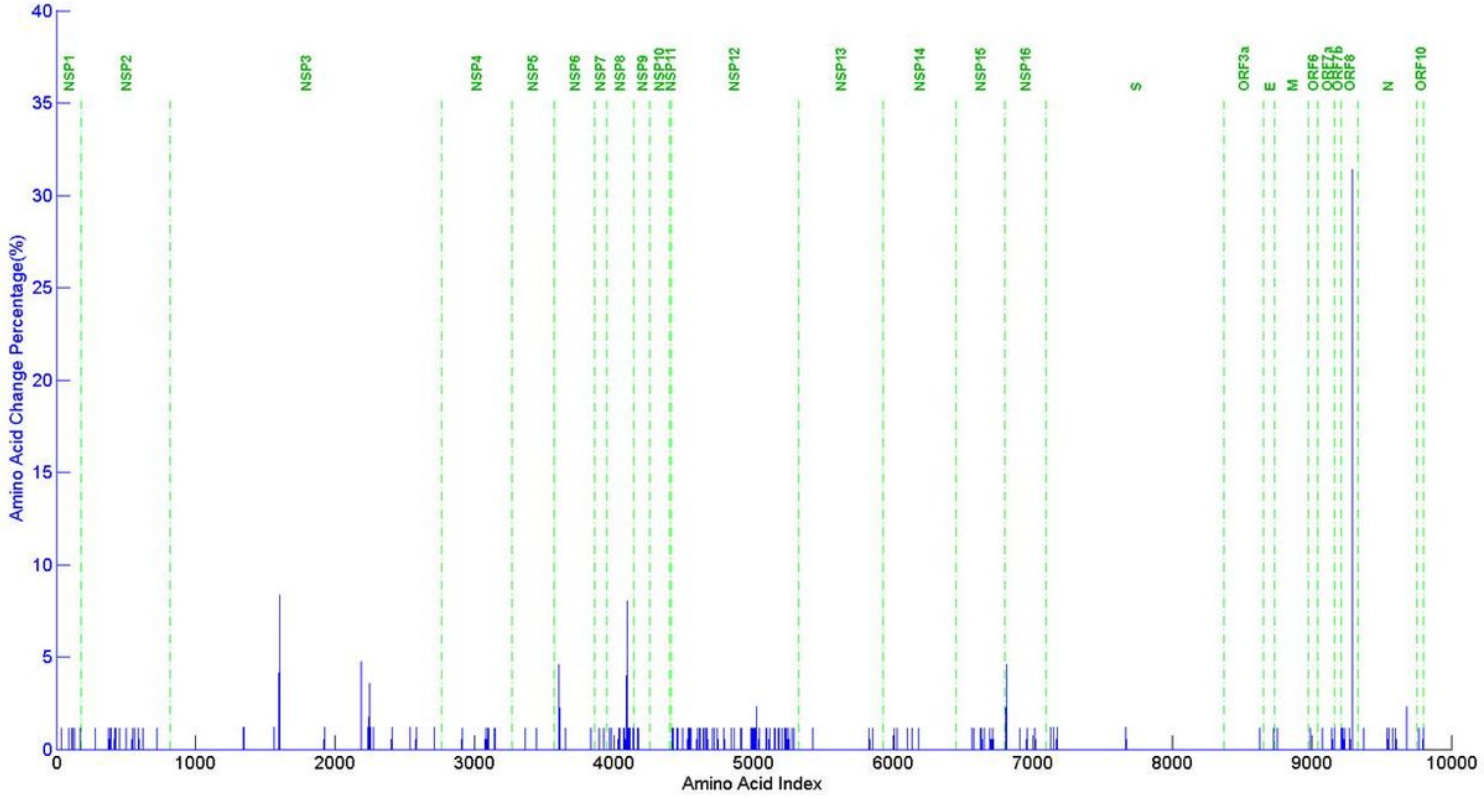


Figure 10

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

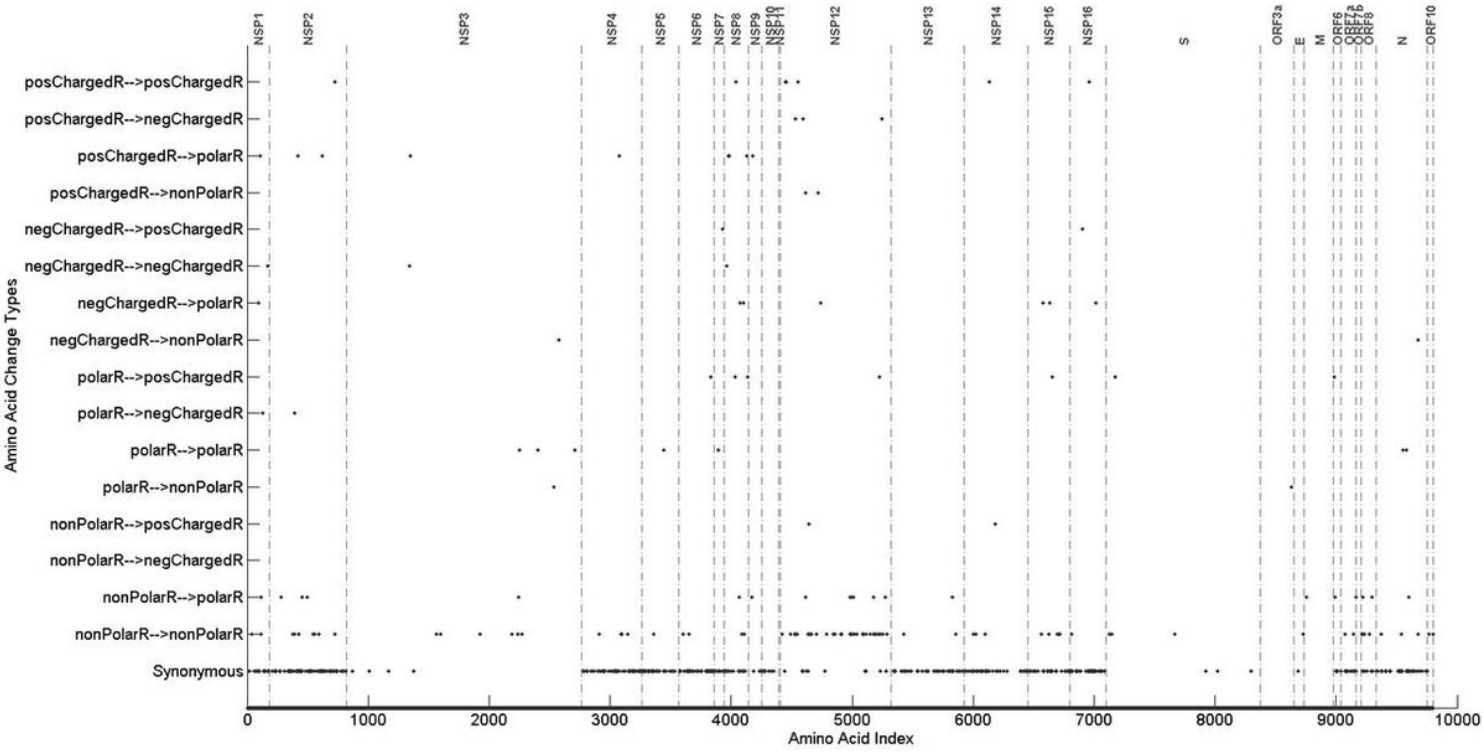


Figure 11

The Most Common Changes of Amino Acid Types in Coding Region of the SARS-COV-2 Genome Variants Detected in China

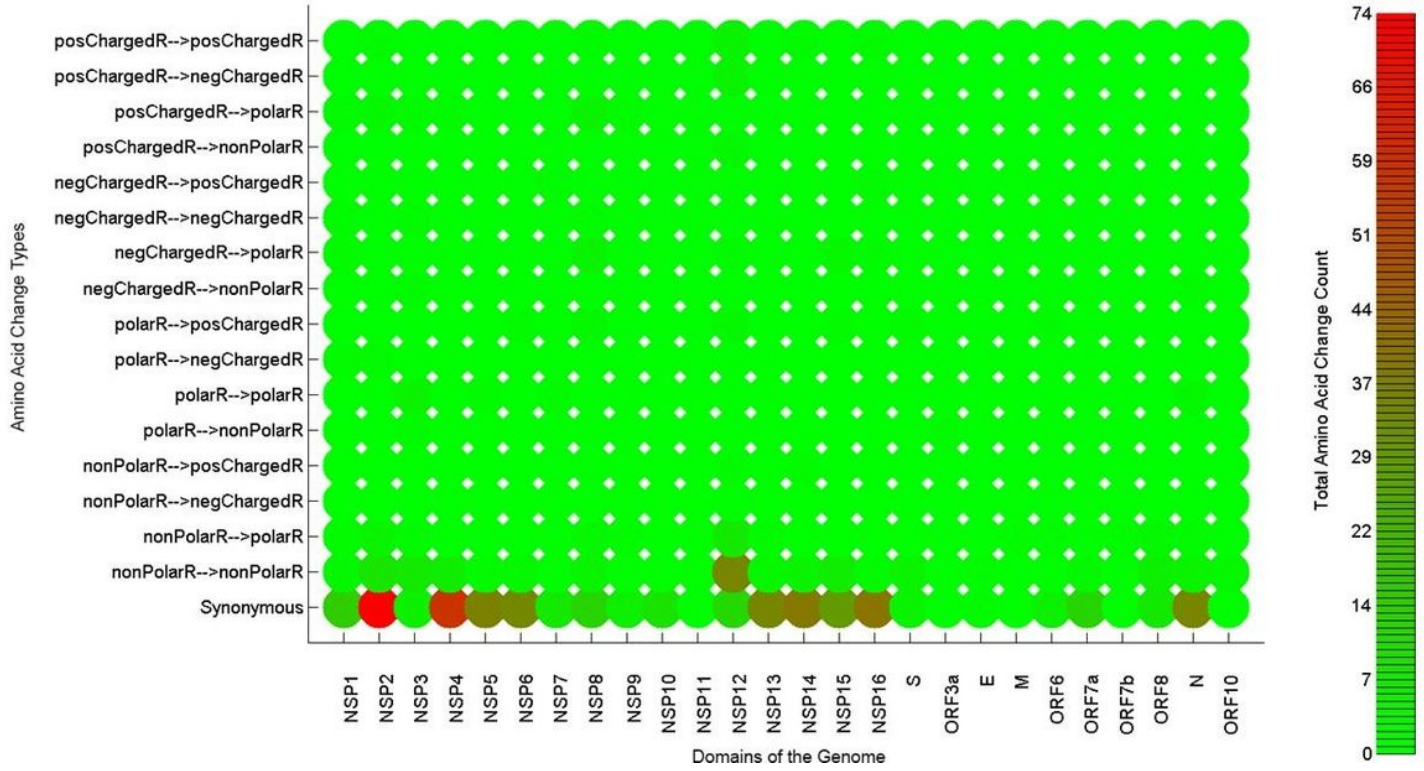


Figure 12

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

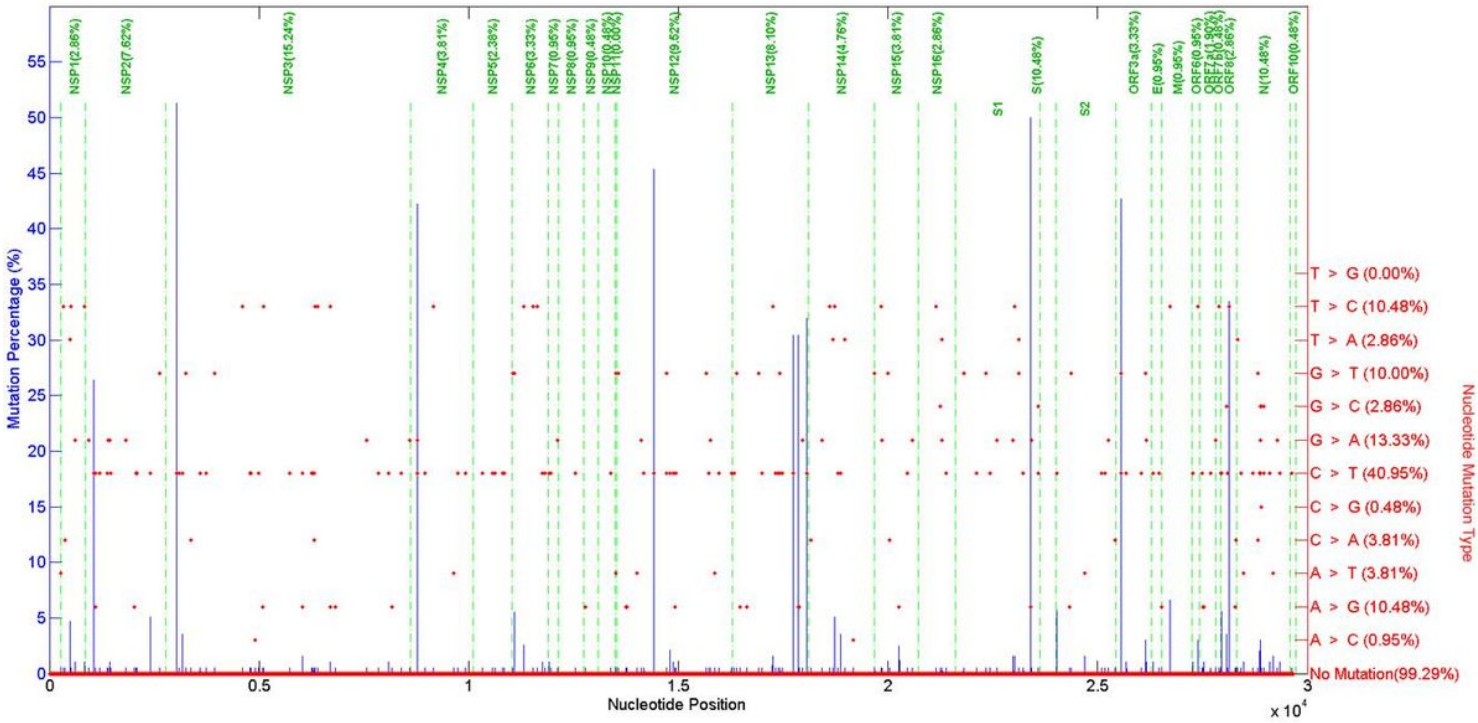


Figure 13

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

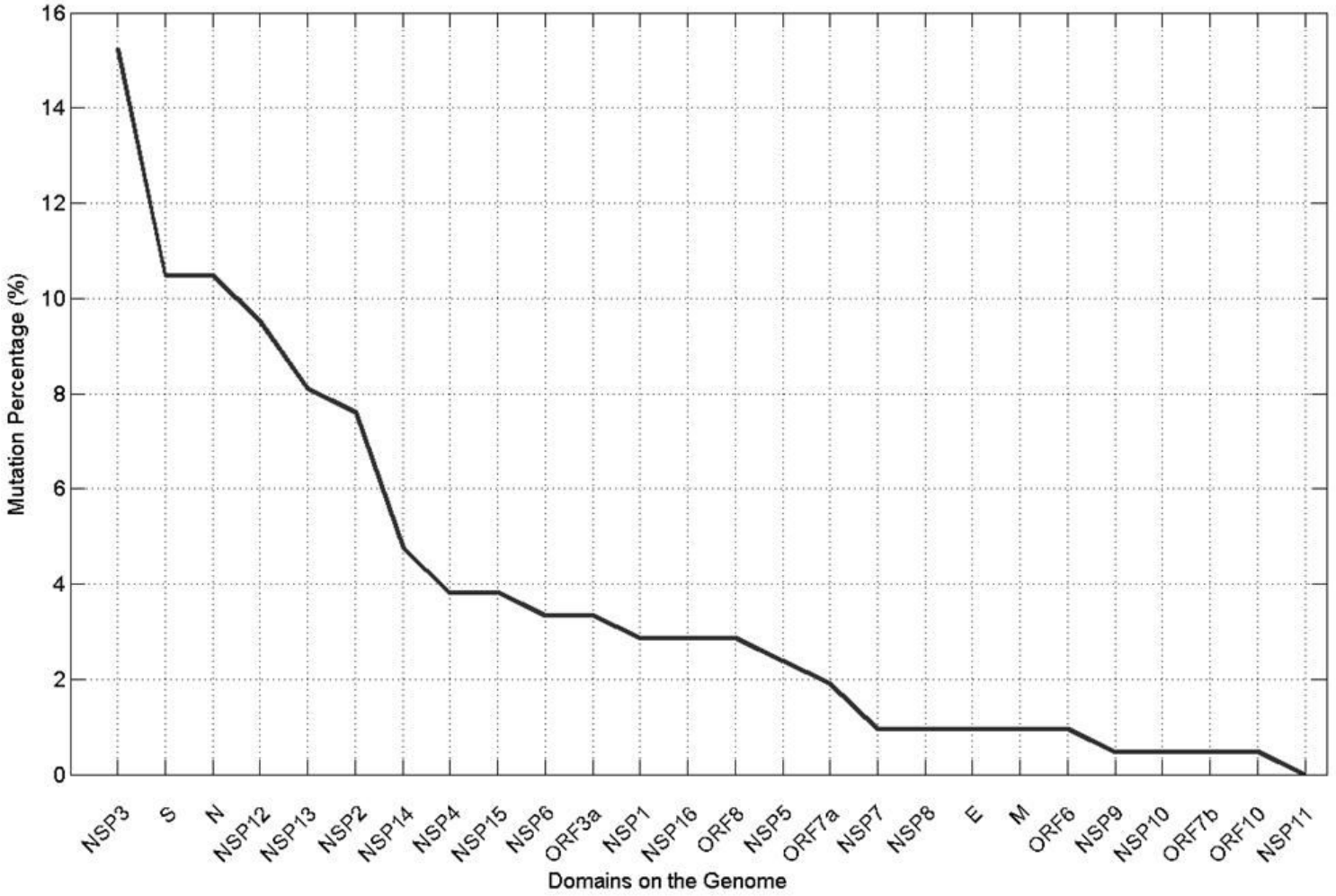


Figure 14

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

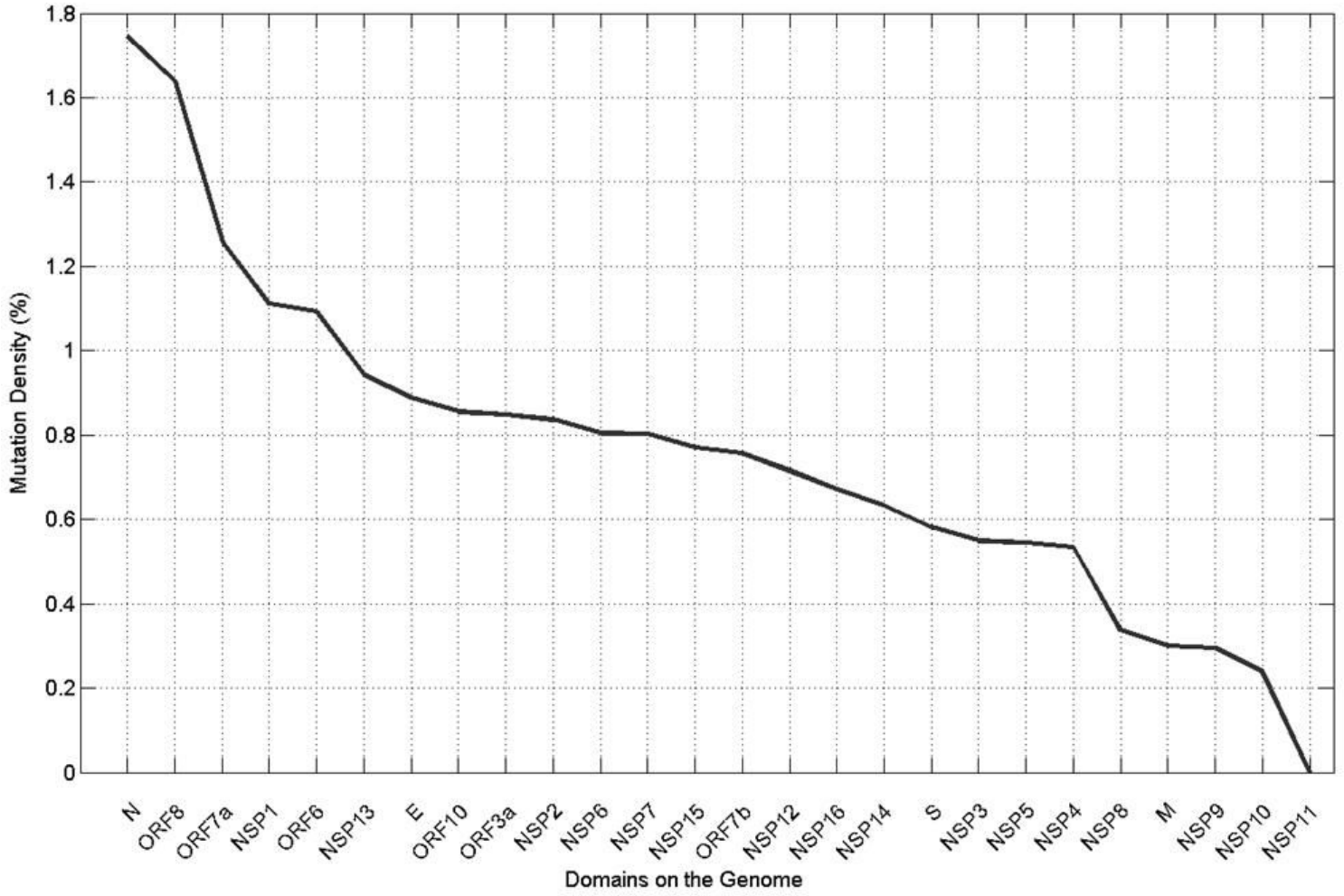
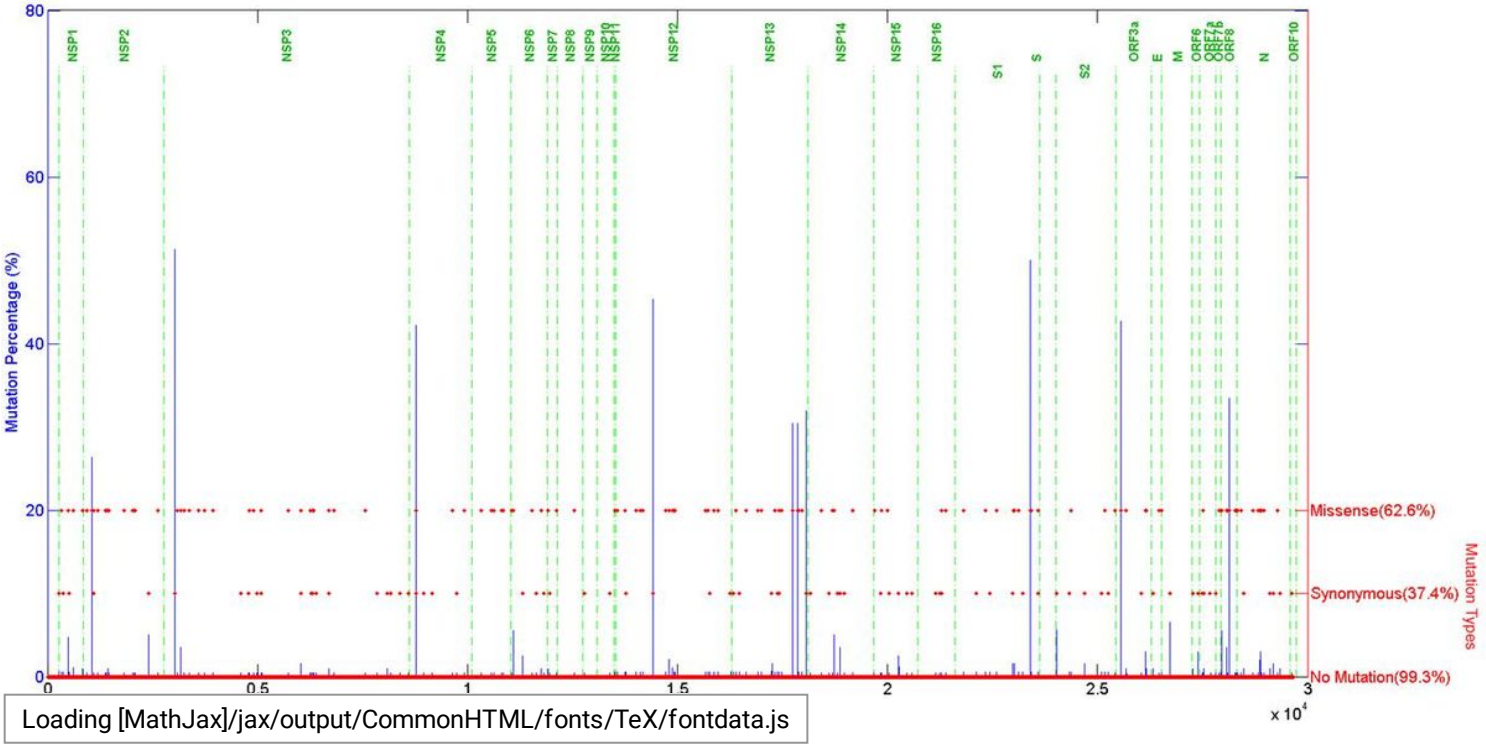


Figure 15

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



Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Figure 16

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

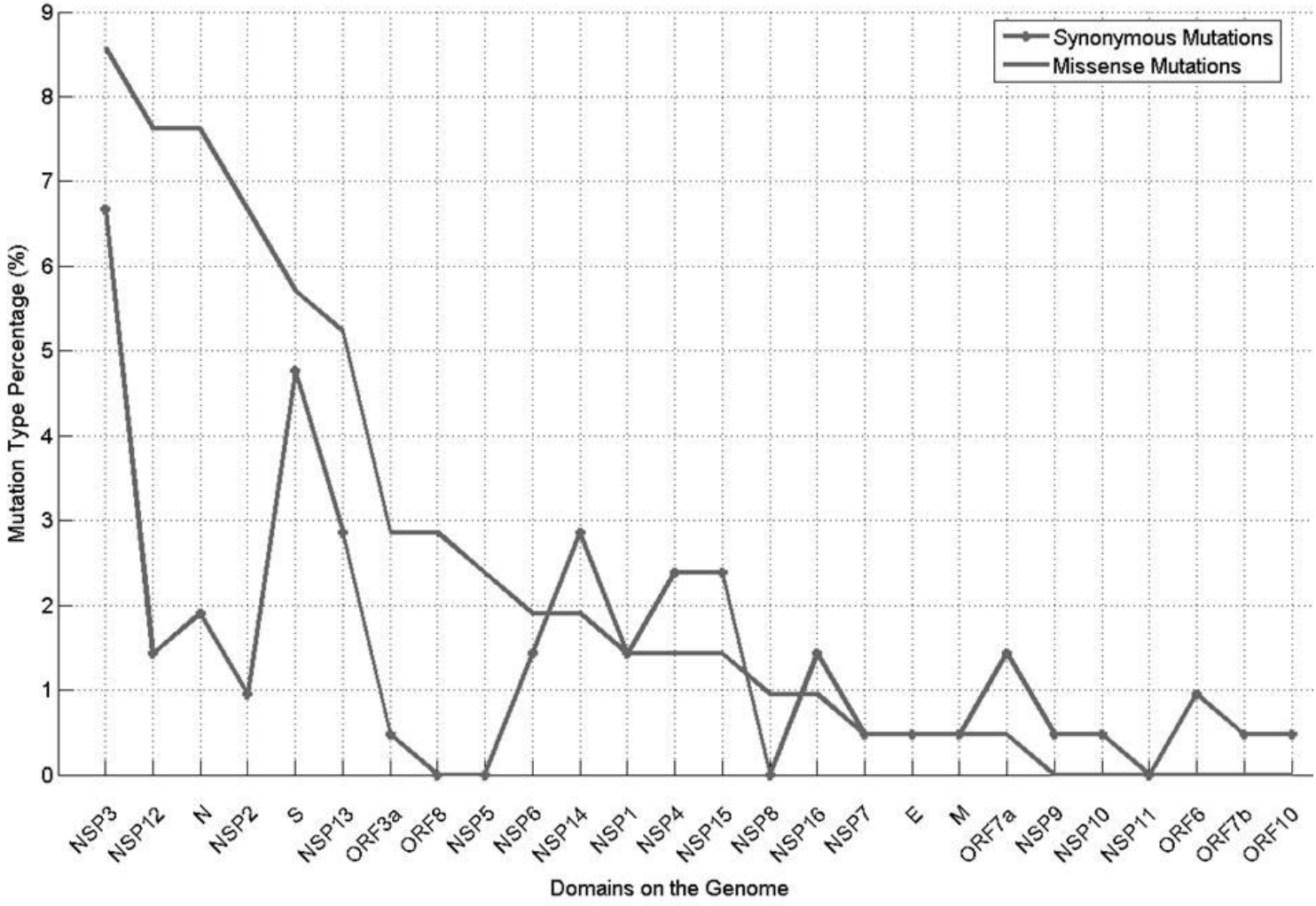


Figure 17

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

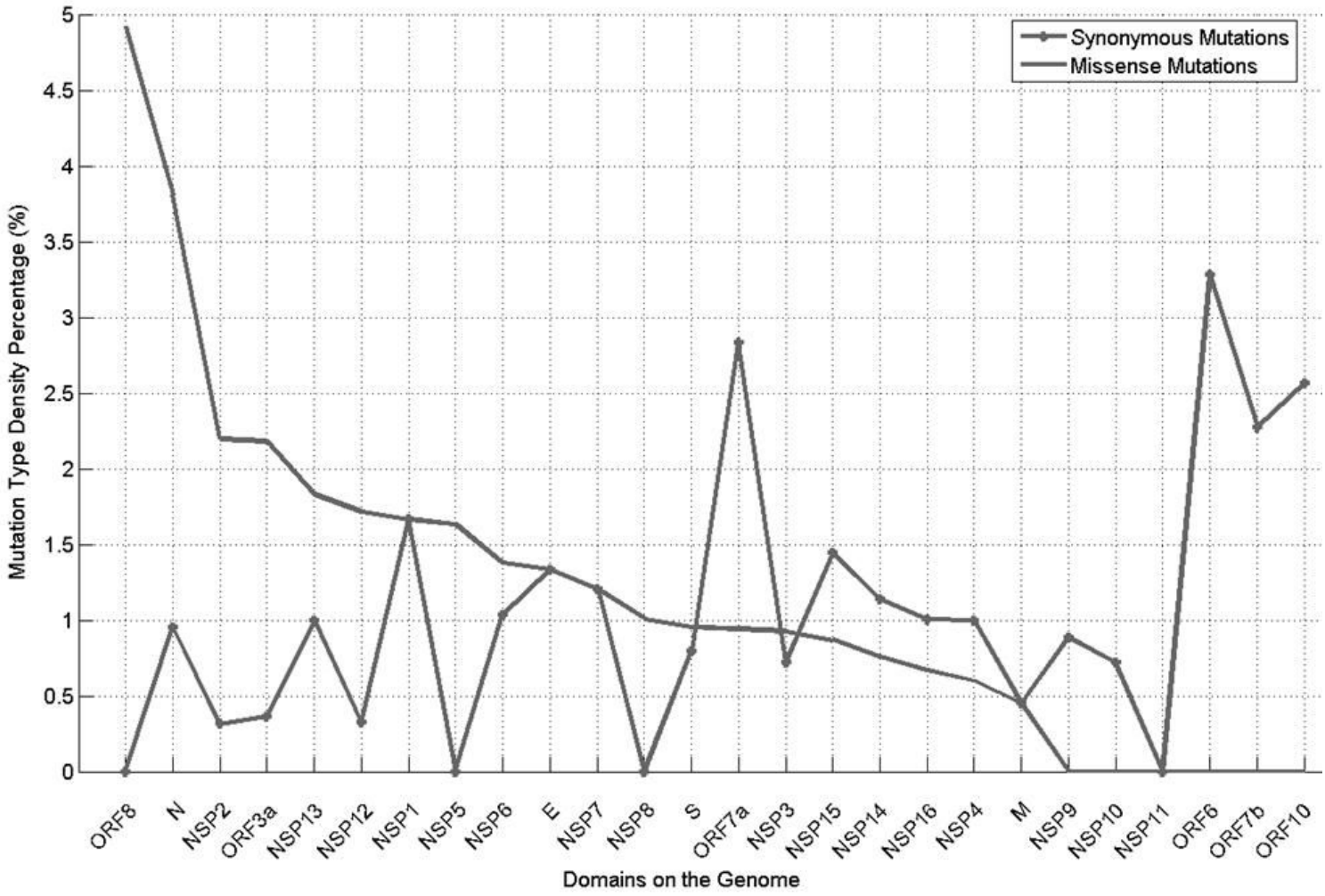


Figure 18

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

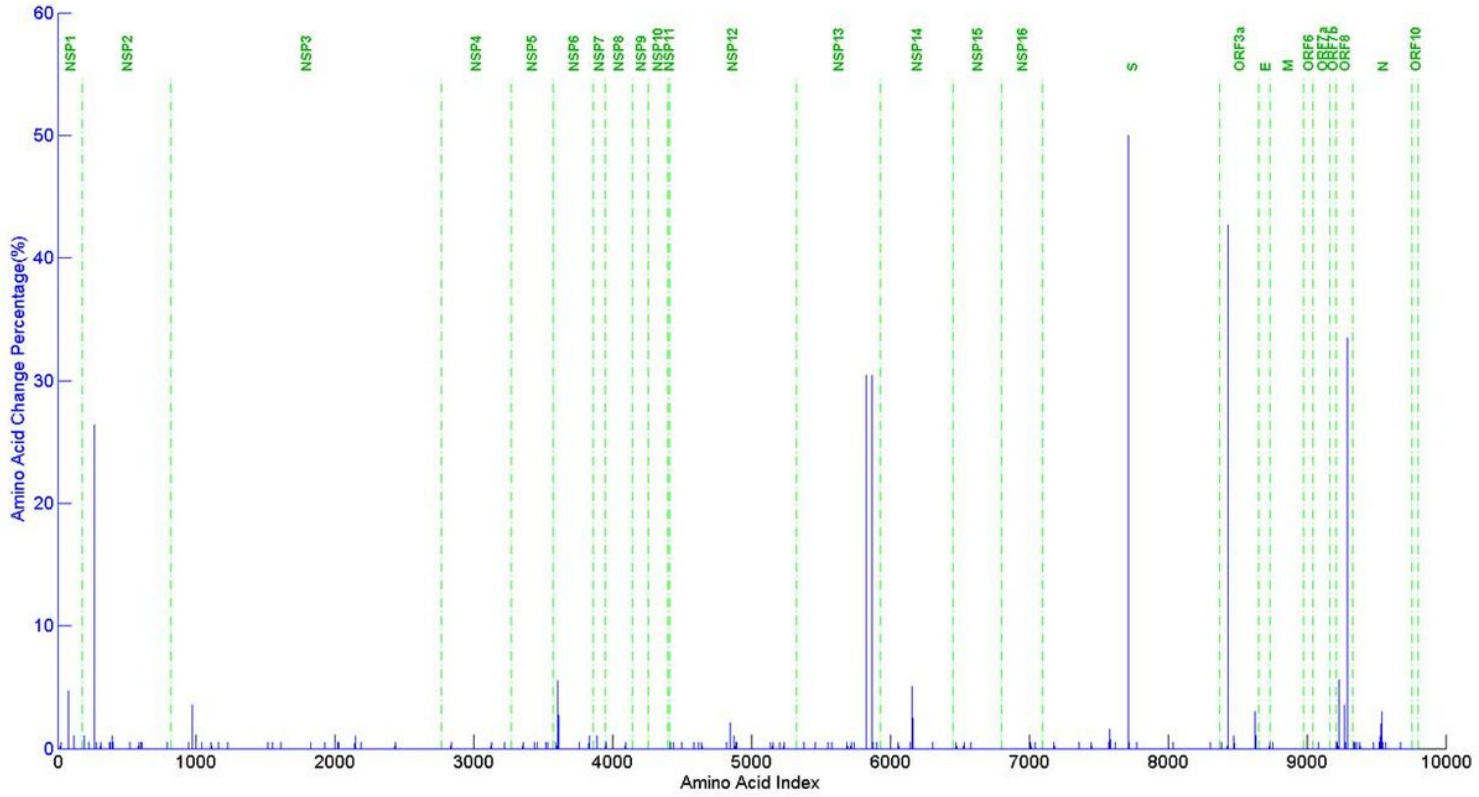


Figure 19

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

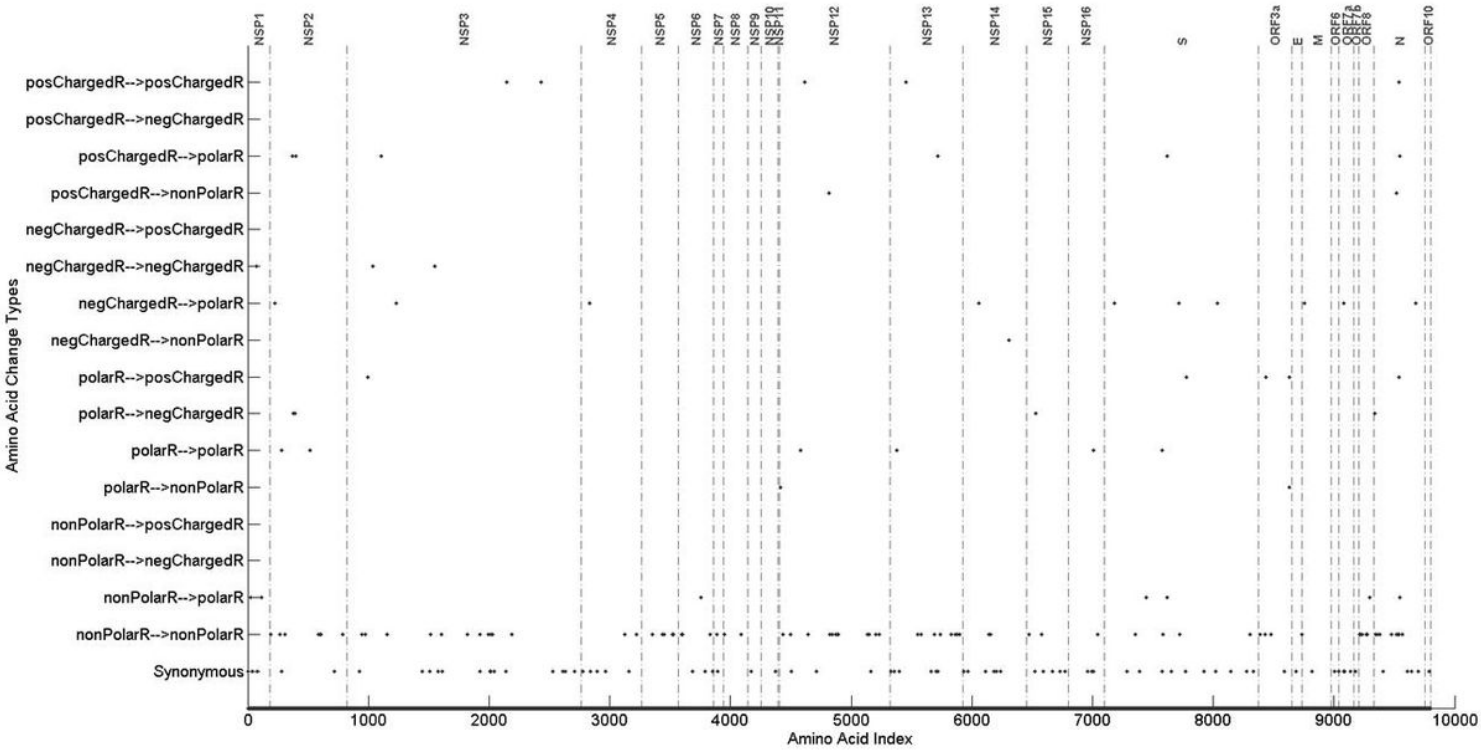


Figure 20

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

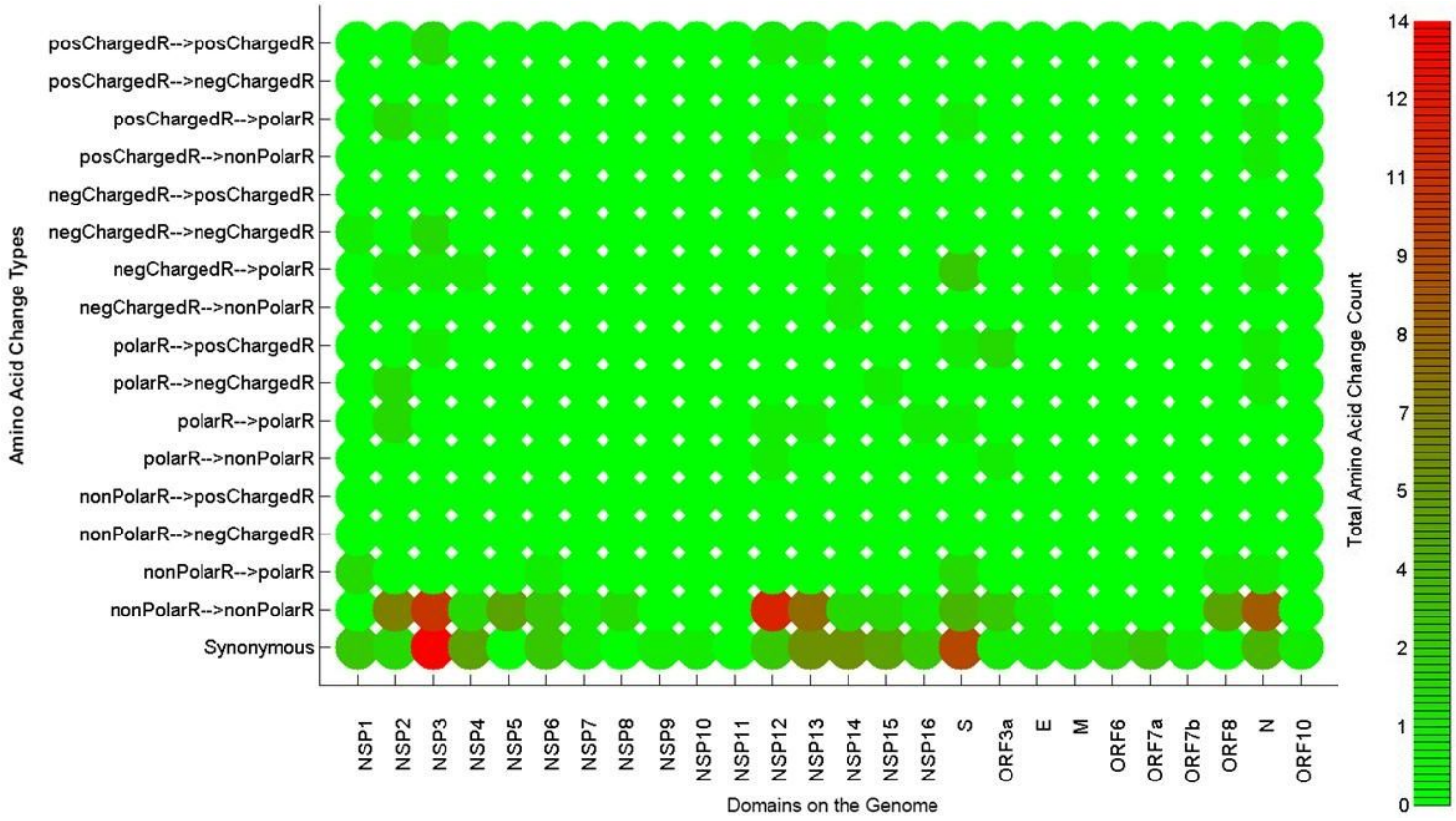


Figure 21

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

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- VirusGeneSequences.rar
- UsedData.rar
- MultiAlignmentResults.rar
- MatlabCodes.rar
- SupplementaryMaterialDiscriptions.txt
- GeneProperties.xlsx