JJBS

Volume 15, Number 4,December 2022
ISSN 1995-6673
Pages 553 – 559
https://doi.org/10.54319/jjbs/150403

Jordan Journal of Biological Sciences

Characterizing New Genomic and Proteomic Variations among SARS-CoV-2 Strains

Nzar A. A. Shwan[1,*] , Sarkar S. Aziz[2], Bahra K. Hamad[1], Omar A. Hussein[3]

[1] *Medical Laboratory Technology Department, Erbil Technical Health and Medical College, Erbil Polytechnic University, Erbil, Iraq*
[2] *Medical Laboratory Technology Department, Soran Technical College, Erbil Polytechnic University, Erbil, Iraq* [3] *Medical Laboratory Technology Department, Shaqlawa Technical College, Erbil Polytechnic University, Erbil, Iraq*

Received: Jan 15, 2022; Revised: April 24, 2022; Accepted May, 11, 2022

Abstract

Since its emergence, COVID-19 has caused severe health problems, and reached more than 220 countries. The viral genome is prone to mutations leading to the appearance of new variants that might be more infectious. Many new genomic sequences of SARS-CoV-2 are uploaded to the public database; monitoring these sequences for possible variations can significantly help in the process of vaccine development and prevention plans. This study aimed to explore whole genome sequences of SARS-CoV-2 that are recently submitted to global databases from different geographical areas for possible new mutations. For this purpose, forty complete genomic sequences of SARS-CoV-2 from 20 countries were downloaded from GISAID (12 Dec 2020 -20 Mar 2021) and converted to their corresponding amino acid sequences using Expasy online software. Both the DNA and amino acid sequences were aligned with the reference genome (Accession number; NC_045512) by the multiple-sequence alignment tool Clustal Omega. The aligned sequences were then examined for any change compared to the reference genome.

The results showed a total of 1264 nucleotide variations; 93.43% were SNPs, 6.49% deletions, and 0.08% insertions. About 59% were non-synonymous mutations and 41% were synonymous mutations. Most of the non-synonymous mutations that lead to amino acid changes were in the Spike (36.63%) and Nucleocapsid (15.60%) genes. Among these changes 24 unique amino acid variations were repeated more than five times, dispersed among the following proteins NSP3, NSP6, NSP12, Spike, ORF3a, and Nucleocapsid.

The analysis in this study revealed an increase in the number of variations accumulated throughout the pandemic, and most of the non-synonymous mutations were in the Spike and Nucleocapsid genes. Sustained molecular surveillance of SARS-CoV-2 is essential to identify new variants and their impact on control measures of the pandemic and also important in the process of vaccine production.

Keywords : COVID-19, SARS-CoV-2 Sequences, Variations, Bioinformatics

1. Introduction

The recent pandemic, Coronavirus Disease-2019 (COVID-19) caused by Severe Acute Respiratory Syndrome-Coronavirus2 (SARS-CoV-2) has made a threatful health problem since its appearance in Wuhan City, China(Wu *et al.*, 2020, Li *et al.*, 2020). This pandemic has reached 220 countries, infected at least 206,987,517 individuals, and caused 4,358,629 deaths. In Iraq, the number of cases has reached 1,761,143 and 19,541 deaths, as of 10th May 2021 (Worldometer, 2021).

Coronaviruses (CoVs) are members of a diverse family of enveloped, positive-sense, single-stranded RNA-viruses, called Coronaviridae(Alluwaimi *et al.*, 2020). The CoVs include four genera, the α- and β-CoV infect mammals, whereas γ- and δ-CoV are related to birds (Worldometer, 2021, Guo *et al.*, 2020). Seven species of CoVs are capable of infecting humans. The α-CoVs HCoV-229E, HCoV-NL63, β-CoVs HCoV-HKU1, and HCoV-OC43 give rise to mild respiratory symptoms similar to the common cold (Guo *et al.*, 2020, Liu *et al., 2021*), while the β-CoVs, SARS-CoV, and Middle East Respiratory Syndrome-Coronavirus (MERS-CoV) lead to severe and possibly fatal respiratory tract infections (Guo *et al.*, 2020). The results of whole genome sequencing analysis have revealed that the SARS-CoV-2 is 96.2% similar to bat CoV (RaTG13), 79.5% to SARS-CoV, and 50% to MERS (Guo *et al.*, 2020, Hu *et al.*, 2021).

The genome size of SARS-CoV-2 is 29903 nucleotides, which contains 12 open reading frames (ORFs) encoding 27 proteins (Rahimi *et al.*, 2021, Gordon *et al.*, 2020, Wu *et al.*, 2020). It starts with 265 nucleotides, 5′ UTR, and ends with 358 nucleotides, 3′ UTR. The first ORF spans over 67% of the viral genome encoding 16 non-structural proteins (NSPs); these NSPs are mainly involved in the transcription and replication processes of the viral genome. Followed by the structural genes; membrane (M), Spike glycoprotein (S), Nucleocapsid (N), and Envelope (E) genes respectively. The other accessory proteins are

[*] Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 ; e-mail: author@institute.xxx .

encoded by the remaining ORFs dispersed between the structural genes (Kumar *et al*., 2020).

Variations in the nucleotides and amino acid sequences are necessary for the viruses to adapt and evolve in the environment or the host. These variations enable viruses to evade the host immune system (Agudelo-Romero *et al*., 2008). Furthermore, some variations might alter the pathogenicity or the rate of infectivity of the virus (Abdullahi *et al*., 2020). For instance, mutations in the furin cleavage site have made SARS-CoV-2 more contagious than SARS-CoV (Huang *et al*., 2020).

Due to the rapid spread of the disease, several studies have been conducted seeking changes in the viral genome. Koyama *et al.* (2020) analyzed 10,022 SARS CoV-2 genomes from 68 countries; most of the genome sequences were isolated in the United States of America, the United Kingdom, and Australia. Overall, their analysis showed 5,775 distinct genome variants (Koyama *et al*., 2020).

Analysis of further genomic sequences in a wider range of countries (28 countries) has shown variants correlated with increased transmissibility, infectivity, and fatality rate (Dumonteil *et al*., 2021, Toyoshima *et al*., 2020).

Since its emergence, many genomic sequences of SARS-CoV2 were uploaded to the global databases from diverse geographic areas and at different times (NCBI, 2021, Gisaid.org, 2021). It is crucial to monitor these genomic sequences for possible variations to understand and trace the evolution and spread of the virus; in turn, this will have a significant contribution to versatile planning in the prevention and development of therapeutic vaccines for the virus, as well as self and community protective measures. This study aimed to analyze genomic and proteomic sequences of SARS-CoV-2, specifically those newly submitted to global databases from different geographical areas and explore the mutation rate and the effect of these mutations on the proteins produced by the virus.

2. Methodology

2.1. *Whole genomic sequences*

Forty genomic sequences of SARS-CoV-2 were downloaded from the GISAID (Gisaid.org, 2021) from 20 countries as FASTA format file (Mexico, Bulgaria, France, Italy, Belgium, Ukraine, Botswana, Brazil, Spain, Indonesia, Russia, Romania, Ghana, USA, Czech Republic, Monaco, Austria, Cameroon, England, and Scotland) (Table 1). The reference sequence of SARS-CoV2 (Accession number; NC_045512, Dec/2019) was obtained from NCBI (NCBI, 2021).

Table 1 The continents where the genomic sequences of SARS-CoV-2 were downloaded from.

Continents	*No. of countries*	*No. of sequences*
Asia	1	2
Europe	13	29
North America	2	5
South America	1	1
Africa	3	3
Sum	20	40

The downloaded sequences in this study were originally uploaded to GISAID (Gisaid.org, 2021) between 12/Dec/2020 to 20/Mar/2021. For the bioinformatic analysis only high-quality and full-length sequences with known collection date were included in this study. Incomplete and low coverage sequences with more than 5% ambiguous bases (Ns) were excluded.

2.2. Sequence processing and alignment

Before we proceed with the analysis, a single FASTA format file was created, including all the 40 genomic sequences along with the reference genome. This file was then used to convert the genomic sequences to their relevant amino acid sequences by the online bioinformatic software Expasy (Duvaud *et al*., 2021). Then, both the nucleotide and amino acid reference sequences were annotated by SnapGene software (V 5.1.5) to determine and highlight the exact position of the genes and proteins, based on the information provided by the reference sequence of SARS-CoV2 on NCBI (NCBI, 2021).

Finally, the FASTA files for both the genomic sequences and the amino acid sequences were used in the multiple-sequence alignment tool (Clustal Omega) to align both nucleotide and amino acid sequences in two separate runs (Sievers *et al*., 2011, Gasteiger *et al*., 2003).

2.3. Result visualization

The results of the alignments were visualized by the UGENE software (V.37). The nucleotide and amino acid variations were observed in the aligned sequences when compared to the reference sequences. The observed variations in the genomes and proteins were recorded in Microsoft Excel (V. Professional plus 2016), which then were analysed accordingly.

3. Results

Forty genomic and proteomic sequences of SARS-CoV2 derived from twenty countries downloaded were compared and analyzed with the reference sequence (Accession number; NC_045512). The total number of nucleotide variations was 1264, of which 1181 (93.43%) variations were SNPs, 82 (6.49%) deletions and 1 (0.08%) insertion. The average nucleotide variation per sequence was 31.6, as shown in (Figure 1).

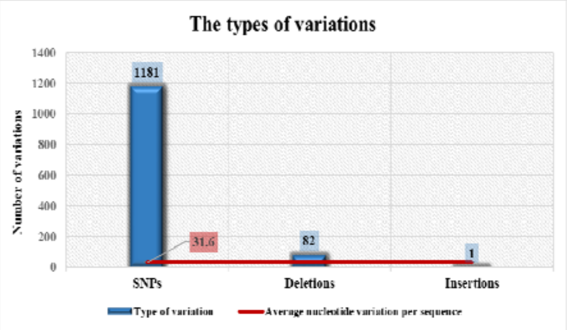

Figure 1. The types of variations. Of the total 1264 variations observed, 1181 (93.43%) are SNPs, 82 (6.49%) deletions, and 1 (0.08%) insertion (Blue columns). The average nucleotide variation per sequence (Redline) is 31.6.

Of these 1264 nucleotide variations, 745 (59%) had changed amino acid sequences (non-synonymous mutations), while the other 519 (41%) were synonymous, which had not altered the codon sequences in a manner to change the corresponding amino acid compared to the reference sequence (Figure 2).

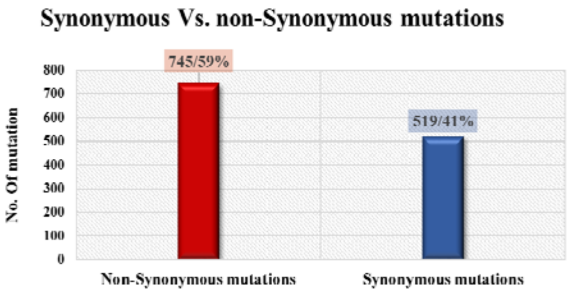

Figure 2: Synonymous Vs. Non-synonymous mutations. 59% of the observed variations changed the amino acid sequence (Non-synonymous mutations shown as red column), while the remained 41% were silent mutations that didn't alter the amino acid sequence (Synonymous mutations shown as blue column).

Based on the type of the nucleotide change, the most frequent nucleotide conversion was Cytosine (C) changed to Thymine (T) (521 /44.1%) (Figure 3), followed by Adenine (A) to Guanine (105/ 8.9%), G to T (90/ 7.63%), T to C (86/ 7.28%), G to A (80/ 6.77%), and A to T (79/ 6.7%). C to A and G to C both 74 times (6.265%), T to A (28/ 2.38%), T to G (27/ 2.28%), A to C (11/ 0.93%) and C to G with the least number (6/ 0.5%). The C substitution to other nucleotides was about 50.8%. The G substitution to other nucleotides was 20.6%, T to other nucleotides was 11.9%, and A to other nucleotides was 16.5%.

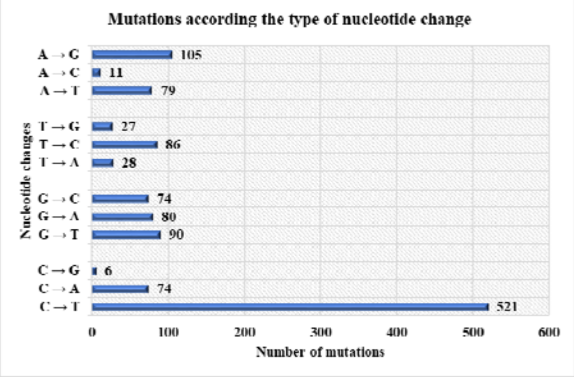

Figure 3: Types of nucleotide changes. The C substitution to other nucleotides is the most frequent nucleotide change with 50.8%. The C → T change shown at the bottom of the graph is the most prominent nucleotide change (44.1%). Other mutations are close to each other in distribution (The C substitutions: 20.6%, T substitutions: 11.9%, and A substitutions: 16.5%).

On the protein level, the most variations were located in the spike protein with 273 (36.63%) of the total 745 amino acid variations (Figure 4). Oppositely, the lowest number of variations was 1(0.12%) in NSP8 and Envelope protein, and 0 in the NSP1, NSP15, ORF6, and ORF7b. Other proteins were NSP7 and ORF7a with 2 (0.29%) variations, NSP16 (3/ 0.40%), NSP14 and ORF10 (4/ 0.53%), Membrane protein and NSP5 (5/ 0.67%), NSP9(8/ 1.10%), NSP4 (10/ 1.34%), NSP13 (18/ 2.40%), NSP2(20/ 2.69%), ORF3a (23/ 3.08%), NSP6 (32/ 4.28%), NSP12 (53/ 7.11%), ORF8 (71/ 9.52%), NSP3 (92/ 12.33%) and Nucleocapsid protein with 116 variations (15.60%). The average amino acid variation was 18.625.

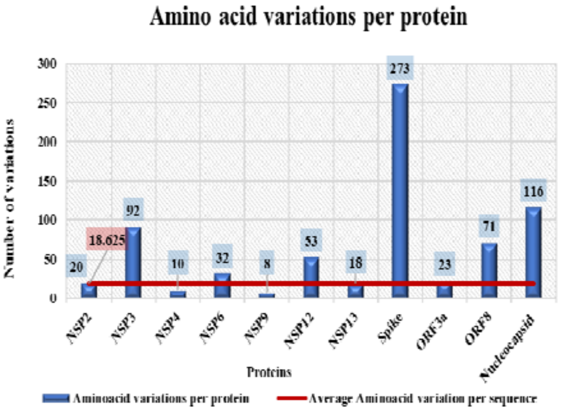

Figure 4: Amino acid variations of proteins for SARS-CoV-2 (Blue columns). The highest number of variations have been recorded in the spike protein (273), followed by Nucleocapsid (116) and to a lesser extent in NSP3 (92). Proteins with a low number of variations (NSP8, Envelope protein, NSP1, NSP15, ORF6, ORF7b, NSP7, ORF7a, NSP16, NSP14, ORF10, Membrane Protein, NSP5) are not shown here. The average amino acid variation per sequence is 18.6 (Red line).

Finally, this study showed the presence of significant amino acid variations revealed by their high frequency. In the analysis, twenty-four amino acid variations were repeated more than five times for 7 different proteins (Figure 5). Of these, three variations were from NSP3, T183I, A890D, and I1412T, all found in 19 different aligned sequences. 106-108 Del from NSP6 has been repeated 21 times. There were 39 times repetitions of P323L at NSP12. In the spike protein, there were nine amino acid variations which are 68-70 Del and A570D with both 20 times, P681H (22 times), D614G (39 times), N501Y (23 times), 144 Del (19 times), and 19 times repetition for T716 I, S982A and D1118H. ORF3a contained one such variation, which is Q57H that is repeated seven times. The Q27, R52 I, and Y73C with the frequency of 19, and K68 with nine presented in ORF8. Five frequent variations were found in the N protein; these are R203K, and G204R shared in 26 sequences, T205I found in 6 sequences, S235F, and D3L repeated in 19 sequences. The remaining amino acid variations with frequencies lower than five were neglected.

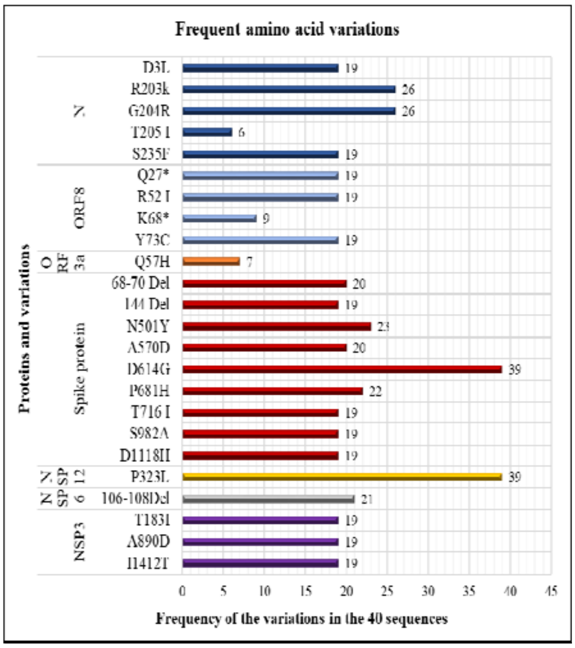

Figure 5: Frequency of highly repeated amino acid variations (more than five times) within the 40 SARS-Cov-2 genome sequences. D614G in spike protein and P323L in NSP12 (39 times), with R203S and G204R in Nucleocapsid protein, are the most repeated changes among the frequent amino acid variation.

4. Discussion

This study analysis showed a high number of nucleotide variations, with an average of 31.6 from 40 genomic sequences when compared to the reference sequence (Accession number; NC_045512). The predominant type of variation was SNPs. This phenomenon is suggested to be associated with the error-prone viral RNA-dependent RNA polymerase or by mechanisms of the host cell RNA editing enzymes as a defense mechanism (Mercatelli and Giorgi, 2020).

The type of nucleotide change (shown in Figure 3) showed C→T change to score the highest value. The average nucleotide variation in this study was higher than in the study by (Mercatelli and Giorgi, 2020). It might be caused by the higher number of mutations accumulated over time, and a lower number of sequences analyzed in this study. The reason for this mutational bias toward C→T is not completely understood, but there are two possible explanations, the codon usage bias, and the RNA editing host enzyme APOBEC (Pollpeter *et al.*, 2018, Ghosh and Chakraborty, 2020). However, it has been observed that the translational preference of a few codons is strongly correlated with the mutational bias imposed by genome compositional constraint and influenced by natural selection, especially in the second and third codon position, which is more biased towards the AT/U content. This is determined by the relative synonymous codon usage (RSCU) value with an average of 64.19% at the second position and 65.24% at the third position, respectively (Ghosh and Chakraborty, 2020).

Of the 1264 nucleotide variations, there were more non-synonymous mutations in comparison to synonymous mutations (shown in Figure 2). Analysis of 3067 SARS-CoV-2 whole genome sequences isolated from 55 countries revealed 782 variant sites, 65.98% non-synonymous, and 28.39% synonymous mutations. The remaining 5.63% was in the intergenic regions of the genome (Laamarti *et al.*, 2020). Despite the higher number of sequences included in the study, the number of variations is lower compared to the current study. These differences might be due to different submission dates of the downloaded sequences. Furthermore, Laamarti *et al.* (2020) analyzed sequences from the first three months of the emergence of the disease, while the sequences analyzed in this study were after twelve months (December 2020); for this reason, a higher number of variations potentially accumulated since the emergence of the virus in Wuhan, China at December 2019 (Laamarti *et al.*, 2020).

Non-synonymous mutation distribution (shown in Figure 4) shows the predominance of variants in S and N proteins. Other proteins contained a lower number of variations, and some of them had no variation at all. The high mutation rate in the spike protein is due to its receptor-binding properties and immunogenicity, and it is supposed to be the major target for antibodies (Singh *et al.*, 2020). The N protein is also one of the critical targets for B cells to be targeted by antibodies, the most abundant protein in coronaviruses, and highly immunogenic (Oliveira *et al.*, 2020). Both proteins are under immune system pressure (Forni *et al.*, 2020).

We only recorded the most repeated mutations, equal to or more than five-time frequency. Of these recorded mutations, the 68-70 Del, A570D, P681H, D614G, N501Y, 144 Del, T716 I, S982A, and D1118H were in the S protein (shown in Figure 5). Similar mutations were found in VOC belonging to B.1.1.7 from the viral sequences in Czech Republic, France, Ukraine, Ghana, Italy, England, Bulgaria, Spain, Belgium and Botswana (Davies *et al.*, 2020, Ramirez *et al.*, 2021, ASSESSMENT, 2020).

Some preliminary data on the effect of 68-70 Del, N501Y, D614G, and P681H has been discovered. The G614 variant, which has emerged as a predominant clade in Europe and is spreading worldwide, predominates over time in locales where it is found, implying that this change enhances viral transmission. This mutation increases the entry to ACE2-expressing cells more efficiently due to the decreased shedding of the S1-domain and higher incorporation of S-protein into the virion (Zhang *et al.*, 2020, Isabel *et al.*, 2020). The N501Y mutation appears to increase the affinity of interaction with murine and human ACE2, as it is one of the key residues in the receptor-binding domain of Spike protein (Ramirez *et al.*, 2021, Gu *et al.*, 2020, Starr *et al.*, 2020). The P681H is placed immediately in the spike furin cleavage site (Peacock *et al.*, 2021, Kemp *et al.*, 2020). However, the functional effect of this mutation is not well understood. The 69-70 deletion-mutation has been determined to increase the viral infectivity in vitro, associated with immune evasion in immunocompromised patients, and has also been shown to be related to the problems in the SARS-CoV-2 RT-PCRs assays targeting the S gene (Ramírez *et al.*, 2021, Kemp *et al.*, 2020b). However, the exact impacts of these mutations on transmissibility, infectivity, and clinical severity are not known up to this time and remain to be fully elucidated.

The N protein contained five frequent variations, R203K and G204R (Russia, Brazil, Cameroon, Mexico, France, Ukraine, Ghana, Italy, England, Bulgaria, Spain, and Belgium), T205I (USA, Indonesia, Monaco, and Botswana), S235F and D3L (France, Ukraine, Ghana, Italy, England, Bulgaria, Spain, and Belgium). It has been shown that R203K and G204R increase positively charged site, and might increase local rigidity with the removal of Gly204. (Garvin *et al.*, 2020). The exact role of mutations in the N beyond RNA and protein interaction interfaces in the pathogenesis of the virus requires further investigations (Singh *et al.*, 2021).

The NSP3 follows the N protein based on the number of mutations per protein with 92 mutations. Three variations were frequent in our results in the NSP3, which are T183I, A890D, and I1412T (Bulgaria, Italy, Belgium, Ukraine, Spain, France, Ghana, and England). Loconsole *et al.* (2021) identified these mutations from a patient traveling back to the Apulia Region in Italy from London, UK (Loconsole *et al.*, 2021). The functional effect remains to be explored.

Four recurrent mutations, Q27*, R52I, Y73C, and K68*, were identified in ORF8 (France, Ukraine, Ghana, Italy, England, Bulgaria, Spain, and Belgium). The functional effect of Q27* and K68* is the truncation in the protein structure(Pereira, 2021). These mutations do not seem to have a harmful effect on the virus, as the morbidity and mortality rates have not decreased despite the presence of these mutations in the circulating variants (Pereira, 2021).

The P323L mutation in NSP12 had a frequency of 39 times (Bulgaria, Italy, Belgium, Ukraine, Italy, Spain, France, Ghana, Monaco, England). The P323L is unlikely to influence polymerase enzymatic activity directly as it is located distal to the NSP12 catalytic core. Instead, this residue is located at the surface of NSP12, near one of the binding sites for NSP8. However, this mutation could modify the NSP8 interaction or interaction with a yet unknown viral or host factor (Peacock *et al.*, 2021). Also, P323L is associated with epitope loss, which could influence the pathogenesis of antibody escape variants (Hasan *et al.*, 2021). However, Hasan *et al.* (2020) describe that this mutation might affect the proofreading activity of RNA-dependent RNA polymerase (RdRp), provoking other changes (Pachetti *et al.*, 2020).

The amino acid deletions 106–108 in NSP6 known as the 'SGF deletion (Bulgaria, Italy, Belgium, Ukraine, Italy, Spain, France, Ghana, Monaco, England), identically found in B.1.1.7 lineage, the P.1 lineage, and several isolates from the B.1.351 lineage. NSP6 is a multi-pass transmembrane protein that is thought to be involved in autophagy and antagonism of innate immune responses, but the influence of this deletion on virus phenotype remains unclear (Peacock *et al.*, 2021). The three successive amino acid deletion 106–108 has caused these variants (P.1, B.1.1.7, and B.1.351) more transmissible than previous circulating variants with possible increased risk of hospitalization, severity, and mortality in the (B.1.1.7). No impact was reported in-hospital mortality (B.1.351), and the effect on (P.1) is under investigation (WHO, 2021).

The Q57H is located in ORF3a (US, Indonesia, Monaco, France, Botswana), which codes for a protein that regulates inflammation, antiviral responses, and apoptosis in the infected cells (Joshi *et al.*, 2021). Q57H mutation might affect inflammasome activation (Hassan *et al.*, 2020). Other proteins had fewer variations and did not attract much of the study's attention, although they did not contain variations repeated five or more times.

5. Conclusions

The analysis in this study showed an increased number of mutations accumulated over time, revealed by the average number of mutations per sequence (31.6) in comparison to the studies mentioned previously, in the course of the pandemic, a year after the onset. A higher number of non-synonymous mutations were recorded compared to the silent mutations. Most of these mutations were located in the Spike and Nucleocapsid proteins due to their immunogenic properties as being major targets of B cells. Our results also showed some major amino acid variations that have been shown to increase viral entry to ACE2 expressing cells.

The amino acid variations have the possibility to either positively or negatively alter proteins, subsequently changing the phenotype of the virus such as infectivity, virulence, and tissue tropism. Moreover, it can also affect the decision on the process of vaccine development.

Having applied restricted criteria for the collected sequences, one limitation of our study could be a low number of whole genome sequences included in the analysis. Including further genomic sequences and restricting the analysis on the structural genes is the ideal starting point for future analysis. Finally, continued molecular observation of the SARS-CoV-2 genome is necessary to identify new emerging variants and their impact on the control measures and prevention plans of the pandemic.

References

Abdullahi, I. N., Emeribe, A. U., Ajayi, O. A., Oderinde, B. S., Amadu, D. O. & Osuji, A. I. 2020. Implications of SARS-CoV-2 genetic diversity and mutations on pathogenicity of the COVID-19 and biomedical interventions. *J Taibah Univ Med Sci,* **15,** 258-264.

Agudelo-Romero, P., Carbonell, P., Perez-Amador, M. A. & Elena, S. F. 2008. Virus Adaptation by Manipulation of Host's Gene Expression. *PLoS One,* **3,** e2397.

Alluwaimi, A. M., Alshubaith, I. H., Al-Ali, A. M. & Abohelaika, S. 2020. The Coronaviruses of Animals and Birds: Their Zoonosis, Vaccines, and Models for SARS-CoV and SARS-CoV2. *Front Vet Sci,* **7,** 655.

Assessment, R. R. 2020. Risk related to spread of new SARS-CoV-2 variants of concern in the EU/EEA.

Davies, N.G., Abbott, S., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J.D., Pearson, C.A., Russell, T.W., Tully, D.C., Washburne, A.D. and Wenseleers, T., 2021. Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1. 7 in England. Science, 372(6538), p.eabg3055.

Dumonteil, E., Fusco, D., Drouin, A. & Herrera, C. 2021. Genomic Signatures of SARS-CoV-2 Associated with Patient Mortality. *Viruses,* **13,** 227.

Duvaud, S., Gabella, C., Lisacek, F., Stockinger, H., Ioannidis, V. and Durinx, C., 2021. Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Res*, **49(W1)**, pp.W216-W227.

Forni, D., Cagliani, R., Pontremoli, C., Mozzi, A., Pozzoli, U., Clerici, M. and Sironi, M., 2021. Antigenic variation of

SARS-CoV-2 in response to immune pressure. *Mol Ecol*, 30(14), pp.3548-3559.

Garvin, M. R., Prates, E. T., Pavicic, M., Jones, P., Amos, B. K., Geiger, A., Shah, M. B., Streich, J., Gazolla, J. G. F. M. & Kainer, D. 2020. Potentially adaptive SARS-CoV-2 mutations discovered with novel spatiotemporal and explainable AI models. *Genome Biol,* **21,** 1-26.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. & Bairoch, A. 2003. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res,* **31,** 3784-8.

Ghosh, S. & Chakraborty, S. 2020. Phylogenomics analysis of SARS-CoV2 genomes reveals distinct selection pressure on different viral strains. *Biomed Res Int,* 2020.

Global Initiative on Sharing Avian Influenza Data.Org. 2021. *GISAID - Initiative* [Online]. Available: https://www.gisaid.org/ [Accessed].

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., O'meara, M. J., Rezelj, V. V., Guo, J. Z., Swaney, D. L., Tummino, T. A., Hüttenhain, R., Kaake, R. M., Richards, A. L., Tutuncuoglu, B., Foussard, H., Batra, J., Haas, K., Modak, M., Kim, M., Haas, P., Polacco, B. J., Braberg, H., Fabius, J. M., Eckhardt, M., Soucheray, M., Bennett, M. J., Cakir, M., Mcgregor, M. J., Li, Q., Meyer, B., Roesch, F., Vallet, T., Mac Kain, A., Miorin, L., Moreno, E., Naing, Z. Z. C., Zhou, Y., Peng, S., Shi, Y., Zhang, Z., Shen, W., Kirby, I. T., Melnyk, J. E., Chorba, J. S., Lou, K., Dai, S. A., Barrio-Hernandez, I., Memon, D., Hernandez-Armenta, C., Lyu, J., Mathy, C. J. P., Perica, T., Pilla, K. B., Ganesan, S. J., Saltzberg, D. J., Rakesh, R., Liu, X., Rosenthal, S. B., Calviello, L., Venkataramanan, S., Liboy-Lugo, J., Lin, Y., Huang, X. P., Liu, Y., Wankowicz, S. A., Bohn, M., Safari, M., Ugur, F. S., Koh, C., Savar, N. S., Tran, Q. D., Shengjuler, D., Fletcher, S. J., O'neal, M. C., Cai, Y., Chang, J. C. J., Broadhurst, D. J., Klippsten, S., Sharp, P. P., Wenzell, N. A., Kuzuoglu-Ozturk, D., Wang, H. Y., Trenker, R., Young, J. M., Cavero, D. A., Hiatt, J., Roth, T. L., Rathore, U., Subramanian, A., Noack, J., Hubert, M., Stroud, R. M., Frankel, A. D., Rosenberg, O. S., Verba, K. A., Agard, D. A., Ott, M., Emerman, M., Jura, N., *et al.* 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature,* **583,** 459-468.

Gu, H., Chen, Q., Yang, G., He, L., Fan, H., Deng, Y.-Q., Wang, Y., Teng, Y., Zhao, Z. & Cui, Y. 2020. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science,* **369,** 1603-1607.

Guo, Y.-R., Cao, Q.-D., Hong, Z.-S., Tan, Y.-Y., Chen, S.-D., Jin, H.-J., Tan, K.-S., Wang, D.-Y. & Yan, Y. 2020. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status. *Mil Med Res,* **7,** 11.

Hasan, M. M., Das, R., Rasheduzzaman, M., Hussain, M. H., Muzahid, N. H., Salauddin, A., Rumi, M. H., Rashid, S. M., Siddiki, A. Z. & Mannan, A. 2021. Global and local mutations in Bangladeshi SARS-CoV-2 genomes. *Virus Res,* **297,** 198390.

Hassan, S. S., Attrish, D., Ghosh, S., Choudhury, P. P. & Roy, B. 2021. Pathogenic perspective of missense mutations of ORF3a protein of SARS-CoV-2. *Virus Res,* **300,** 198441.

Hu, B., Guo, H., Zhou, P. & Shi, Z.-L. 2021. Characteristics of SARS-CoV-2 and COVID-19. *Nat Rev Microbio.,* **19,** 141-154.

Huang, Y., Yang, C., Xu, X. F., Xu, W. & Liu, S. W. 2020. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol Sin*, **41**, 1141-1149.

Isabel, S., Graña-Miraglia, L., Gutierrez, J. M., Bundalovic-Torma, C., Groves, H. E., Isabel, M. R., Eshaghi, A., Patel, S. N., Gubbay, J. B. & Poutanen, T. 2020. Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. *Sci Rep,* **10,** 1-9.

Joshi, M., Puvar, A., Kumar, D., Ansari, A., Pandya, M., Raval, J., Patel, Z., Trivedi, P., Gandhi, M., Pandya, L., Patel, K., Savaliya, N., Bagatharia, S., Kumar, S., & Joshi, C. 2021. Genomic Variations in SARS-CoV-2 Genomes From Gujarat: Underlying Role of Variants in Disease Epidemiology. *Front Genet, 12*, 586569.

Kemp, S.A., Collier, D.A., Datir, R.P., Ferreira, I.A., Gayed, S., Jahun, A., Hosmillo, M., Rees-Spear, C., Mlcochova, P., Lumb, I.U. and Roberts, D.J., 2021. SARS-CoV-2 evolution during treatment of chronic infection. *Nature*, *592*(7853), pp.277-282.

Meng, B., Kemp, S.A., Papa, G., Datir, R., Ferreira, I.A., Marelli, S., Harvey, W.T., Lytras, S., Mohamed, A., Gallo, G. and Thakur, N., 2021. Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B. 1.1. 7. *Cell Rep*, *35*(13), p.109292.

Koyama, T., Platt, D. & Parida, L. 2020. Variant analysis of SARS-CoV-2 genomes. *Bull World Health Organ,* **98,** 495-504.

Kumar, S., Nyodu, R., Maurya, V.K., Saxena, S.K. 2020. Morphology, Genome Organization, Replication, and Pathogenesis of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). In: Saxena, S. (Eds). **Coronavirus Disease 2019 (COVID-19). Medical Virology: From Pathogenesis to Disease Control**. Springer, Singapore, pp. 21-23.

Laamarti, M., Alouane, T., Kartti, S., Chemao-Elfihri, M., Hakmi, M., Essabbar, A., Laamarti, M., Hlali, H., Bendani, H. & Boumajdi, N. 2020. Large scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations. *PloS one,* **15,** e0240345.

Li, H., Liu, S.-M., Yu, X.-H., Tang, S.-L. & Tang, C.-K. 2020. Coronavirus disease 2019 (COVID-19): current status and future perspectives. *Int J Antimicrob Agents,* **55,** 105951.

Liu D.X., Liang J.Q., Fung T.S. 2021. Human Coronavirus-229E, -OC43, -NL63, and -HKU1 (Coronaviridae). In: Bamford, D. H. and Zuckerman, M, (eds). **Encyclopedia of Virology**, Academic Press, United States, pp. 428–440.

Loconsole, D., Sallustio, A., Accogli, M., Centrone, F., Capozzi, L., Del Sambro, L., Parisi, A. & Chironna, M. 2021. Genome sequence of a SARS-CoV-2 VUI 202012/01 strain identified from a patient returning from London, England, to the Apulia Region of Italy. *Microbiol Resour Announc,* **10,** e01487-20.

Mercatelli, D. & Giorgi, F. M. 2020. Geographic and genomic distribution of SARS-CoV-2 mutations. *Front Microbiol,* **11,** 1800.

National Center for Biotechnology Information, N. C. F. B. I. 2021. *NCBI Virus* [Online]. Available: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/. [Accessed].

Oliveira, S. C., De Magalhães, M. T. & Homan, E. J. 2020. Immunoinformatic analysis of SARS-CoV-2 Nucleocapsid protein and identification of COVID-19 vaccine targets. *Front Immunol,* **11,** 2758.

Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., Masciovecchio, C., Angeletti, S., Ciccozzi, M. & Gallo, R. C. 2020. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J Transl Med,* **18,** 1-9.

Peacock, T.P., Goldhill, D.H., Zhou, J., Baillon, L., Frise, R., Swann, O.C., Kugathasan, R., Penn, R., Brown, J.C., Sanchez-David, R.Y. and Braga, L., 2021. The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. *Nat Microbiol*, *6*(7), pp.899-909.

Peacock, T. P., Penrice-Randal, R., Hiscox, J. A. & Barclay, W. S. 2021. SARS-CoV-2 one year on: evidence for ongoing viral adaptation. *J Gen Virol,* **102,** 001584.

Pereira, F. 2021. SARS-CoV-2 variants combining spike mutations and the absence of ORF8 may be more transmissible and require close monitoring. *Biochem Biophys Res Commun,* **550,** 8-14.

Pollpeter, D., Parsons, M., Sobala, A. E., Coxhead, S., Lang, R. D., Bruns, A. M., Papaioannou, S., Mcdonnell, J. M., Apolonia, L. & Chowdhury, J. A. 2018. Deep sequencing of HIV-1 reverse transcripts reveals the multifaceted antiviral functions of APOBEC3G. *Nat Microbiol,* **3,** 220-233.

Rahimi, A., Mirzazadeh, A. & Tavakolpour, S. 2021. Genetics and genomics of SARS-CoV-2: A review of the literature with the special focus on genetic diversity and SARS-CoV-2 genome detection. *Genomics,* **113,** 1221-1232.

Ramírez, J. D., Muñoz, M., Patiño, L. H., Ballesteros, N. & Paniz-Mondolfi, A. 2021. Will the emergent SARS-CoV2 B. 1.1. 7 lineage affect molecular diagnosis of COVID-19? *J Med Virol,* **93,** 2566-2568.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., Mcwilliam, H., Remmert, M., Söding, J., Thompson, J. D. & Higgins, D. G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol,* **7,** 539.

Singh, J., Samal, J., Kumar, V., Sharma, J., Agrawal, U., Ehtesham, N. Z., Sundar, D., Rahman, S. A., Hira, S. & Hasnain, S. E. 2021. Structure-function analyses of new SARS-CoV-2 variants B. 1.1. 7, B. 1.351 and B. 1.1. 28.1: clinical, diagnostic, therapeutic and public health implications. *Viruses,* **13,** 439.

Singh, P. K., Kulsum, U., Rufai, S. B., Mudliar, S. R. & Singh, S. 2020. Mutations in SARS-CoV-2 leading to antigenic variations in spike protein: a challenge in vaccine development. *J Lab Physicians,* **12,** 154-160.

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., King, N. P., Veesler, D. & Bloom, J. D. 2020. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell,* **182,** 1295-1310 e20.

Toyoshima, Y., Nemoto, K., Matsumoto, S., Nakamura, Y. & Kiyotani, K. 2020. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. *J Hum Genet,* **65,** 1075-1082.

World Health Organization. 2021. *COVID-19 Weekly Epidemiological Update* [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20210309_weekly_epi_update_30.pdf [Accessed March 7 2021].

Worldometer. 2021. *COVID-19 Coronavirus Pandemic* [Online]. Available: https://www.worldometers.info/coronavirus/ [Accessed May 10 2021].

Wu, P., Duan, F., Luo, C., Liu, Q., Qu, X., Liang, L. & Wu, K. 2020. Characteristics of Ocular Findings of Patients With Coronavirus Disease 2019 (COVID-19) in Hubei Province, China. *JAMA ophthalmol,* **138,** 575-578.

Zhang, L., Jackson, C. B., Mou, H., Ojha, A., Peng, H., Quinlan, B. D., Rangarajan, E. S., Pan, A., Vanderheiden, A. & Suthar, M. S. 2020. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat Commun,* **11,** 1-9.