Scientific Research Publishing

Journal of Biosciences and Medicines, 2022, 10, 84-97
https://www.scirp.org/journal/jbm
ISSN Online: 2327-509X
ISSN Print: 2327-5081

Bioinformatic Analysis of SARS-CoV-2 Genomes to Develop a Universal Coronavirus Vaccine

Anya Vaish[1*], James McSwiggen[2]

[1]Tesla STEM High School, Redmond, USA
[2]McSwiggen Biotech Consulting LLC, Arvada, USA
Email: *anyavaish@gmail.com, Jim@McSwiggenBiotech.com

How to cite this paper: Vaish, A. and McSwiggen, J. (2022) Bioinformatic Analysis of SARS-CoV-2 Genomes to Develop a Universal Coronavirus Vaccine. *Journal of Biosciences and Medicines*, **10**, 84-97.
https://doi.org/10.4236/jbm.2022.1010006

Received: September 1, 2022
Accepted: October 11, 2022
Published: October 14, 2022

Abstract

COVID-19 is caused by the SARS-CoV-2 virus. Current RNA vaccines Pfizer/BioNTech's BNT162b2 and Moderna's mRNA-1273 are more than 94% successful in preventing infection. The spike protein of the virus is essential for the interaction and internalization of the virus in the host cell and is considered a prime target for vaccine development against the SARS virus. This study aims to identify highly conserved sequences in spike protein or other sections of the viral genome that can potentially be used to develop a universal coronavirus vaccine. Bioinformatic analysis of 258,269 full-length SARS-CoV-2 genomic sequences in the NCBI database was carried out using a custom Perl Script. All sequences were compared to the spike protein and full-length viral genome reference to find 100 nucleotide-long segments that were at least 99% conserved across SARS-CoV-2 sequences. The analysis resulted in a >99.5% conserved 114-nucleotide segment on the spike protein and a 99.49% conserved 104-nucleotide segment on the non-spike protein section of the viral genome. The conserved sequences from this study may be useful in developing an RNA or protein vaccine that may be effective against future SARS-CoV-2 strains or could act as a universal vaccine if these sequences are present in other coronavirus families.

Keywords

Coronavirus, S-Protein, Spike Protein, Conservation, Universal Vaccine, Bioinformatics

1. Introduction

This study aims to identify highly conserved SARS-CoV-2 RNA sequences that can potentially be used to develop a universal coronavirus vaccine. Coronavirus

disease 2019 (COVID-19) is caused by the SARS-CoV-2 virus, which is an RNA virus. As of April 2022, over 500 million people have been infected by SARS-CoV-2 and over 6.1 million have died worldwide [1]. Several vaccines for SARS-CoV-2 were quickly developed in response to the outbreak of COVID-19 [2] [3]. Among these vaccines, two mRNA vaccines from Moderna (mRNA-1273) and Pfizer/BioNTech (BNT162b2), both of which encode the prefusion-stabilized full-length spike glycoprotein of the SARS-CoV-2, are widely used and have shown more than 94% efficacy against symptomatic COVID-19. Other vaccines have shown 60% - 80% efficacy against symptomatic COVID-19 [4]. However, the virus is prone to mutation, which is demonstrated by the emergence of new variants including Delta and Omicron variants. Current vaccinations may not be as effective against new strains [5] [6]. Therefore, the development of a universal vaccine, which may be effective against future strains of coronaviruses, is highly desirable [7] [8]. SARS-CoV-2 is a β-coronavirus that is closely related to SARS-CoV-1 and distantly related to MERS-CoV, both of which have previously caused epidemics and are still serious threats to human health, as well as distantly related to common-cold coronaviruses [9]. A pancoronavirus vaccine that covers all strains of SARS-CoV and MERS viruses will be even more desirable to quickly respond to the emergence of new diseases [10]. A pancoronavirus vaccine could potentially be developed by targeting a highly conserved sequence from SARS-CoV-2, as this vaccine would be able to target the same location in multiple coronavirus variants and therefore be at least somewhat effective against them. Conservation analysis can be performed to identify certain conserved sequences for this purpose.

Previous studies have performed conservation analysis on SARS-CoV-2 as well as on other coronaviruses. A study conducted conservation and phylogenetic analysis to trace the evolutionary history of SARS-CoV-2 to determine that a bat coronavirus, rather than a pangolin coronavirus, was more likely to be the lowest common ancestor of SARS-CoV-2 [11]. Another study identified larger regions of genomes of betacoronaviruses lineage B, a group that includes SARS-CoV-2 and SARS-CoV that were conserved, such as the 3'-UTR and 5'-UTR [12]. However, since the 3'-UTR and the 5'-UTR are untranslated regions, they cannot be used for the development of a vaccine. Neither of these studies identified specific nucleotide sequences that are highly conserved and can be targeted by a vaccine. Another study performed phylogenetic network analysis using the median joining network algorithm to trace the evolution of SARS-CoV-2, which is useful since tracing the movement of SARS-CoV-2 variants can help researchers predict how future variants will behave [13]. This study did not relate their findings to the development of a universal or pancoronavirus vaccine. One study analyzed 3132 viral protein sequences across multiple families of coronavirus using sequence alignment and identified several 9-amino acid epitopes with 89% exact match in the spike protein region that included two epitopes identified from recovered COVID-19 patients [14].

This bioinformatic analysis project sought to analyze 258,269 unique full-length SARS-CoV-2 genomes submitted into the National Center for Biotechnology Information (NCBI) database as of Feb 13, 2022 to determine highly conserved RNA sequences across SARS-CoV-2 genomes. These conserved sequences may prove to be significant for the development of a universal coronavirus vaccine, a vaccine that targets a conserved sequence and may be effective against multiple variants of coronaviruses. The RNA sequences of the S-protein genomes were of particular interest since the spike protein is an essential part of virus transmission [15], which suggests that the spike protein is likely to be more conserved across SARS-CoV-2 genomes relative to other parts of the genome. Additionally, two approved RNA vaccines for SARS-CoV-2 target the S-protein [4]. Therefore, in addition to the bioinformatic analysis of full-length genomic sequence of SARS-CoV-2, the S-protein genomic sequence was specifically analyzed. The analysis is important to find a conserved RNA sequence on the spike protein that will code for a unique protein or peptide epitope and allow the preparation of RNA vaccine. This project also aimed to perform sequence alignment analysis on common SARS-CoV-2 mutant sequences to gain an understanding of how closely related these mutants were to each other, as well as where the mutations were commonly occurring throughout the whole coronavirus genome and in the spike protein sequence. This information may be useful to develop a universal coronavirus vaccine and may be used to predict how future variants of SARS-CoV-2 will mutate, and therefore help predict how effective a vaccine may be against them.

2. Methodology

Existing nucleotide sequences data was obtained from the publicly available, open-source database of SARS-CoV-2 genomes from the National Center for Biotechnology Information (NCBI). **Figure 1** depicts the procedure of bioinformatic analysis.

All full-length SARS-CoV-2 genomic sequences from North America in the NCBI database as of February 13, 2022 were downloaded in a FASTA file for analysis. A total of 258,269 full-length SARS-CoV-2 genomic sequences were

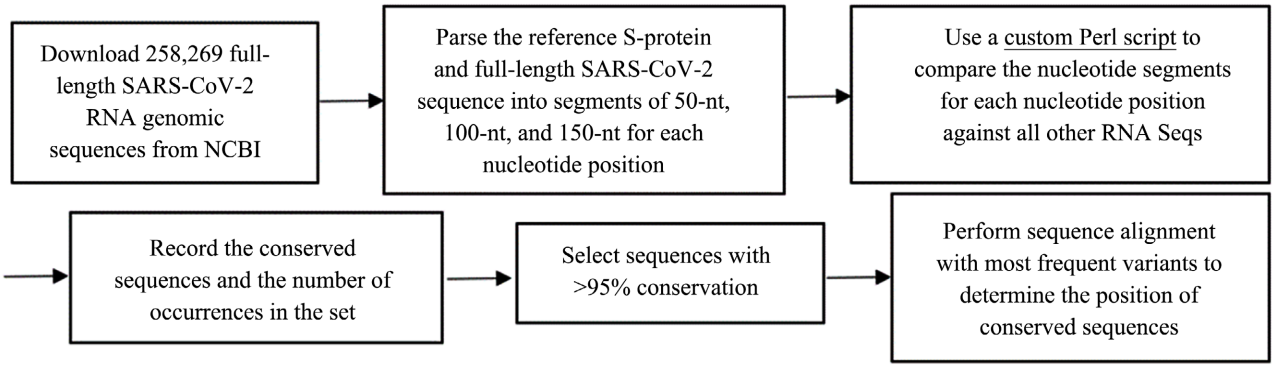

Figure 1. Flow chart of the procedure, including data collection and data analysis.

downloaded in the search database set. The selection criteria for download from NCBI SARS-Cov-2 database was full-length sequence, North America as geographic region, and Homo sapiens as host. The full-length genomic sequences included SARS-CoV-2 variants Alpha B.1.1.7, Beta B.1.351, Gamma P.1, Delta B.1.617.2, Eta B.1.525, Iota B.1.526, Kappa B.1.617.1, and Omicron B.1.1.529. The RefSeq S-protein RNA sequence was isolated from the full-length SARS-CoV-2, reference sequence and the S-protein RefSeq and the full-length RefSeq were used as reference sequences to parse into nucleotide fragments to use for searching similar sequences in the set of SARS-CoV-2 genomes.

A custom Perl script was written to parse the RefSeqs into smaller fragments and search for the sequences in the set of SARS-CoV-2 genomes. The Perl program was run on the reference sequence to first parse it into smaller segments for analysis. The reference sequence was parsed into segments with a length of 50 nucleotides, 100 nucleotides, and 150 nucleotides, with the start of each segment exactly one nucleotide away from the start of the previous segment. The Perl program was tested on a set of 1000 genomic sequences to determine whether the 50-nucleotide, 100-nucleotide, or 150-nucleotide segments should be used for analysis on a larger set of genomic sequences. The Perl program was then used to analyze the 100-nucleotide segments from the full-length RNA and the S-protein RNA reference sequences against each corresponding full-length sequence of the SARS-CoV-2 target database, one genomic sequence at a time. If the fragment segments were a perfect match compared to the corresponding segments in the database genomic sequences, the segment was recorded as a match sequence. This process was then repeated for each of the other S-protein sequence fragments and sequences from the rest of the full-length RefSeq. For each match sequence recorded, the percent conservation was calculated using the number of occurrences of the sequence across all full-length or S-protein sequences. Match sequences that had over 95% conservation (occurred in over 95% of the S-protein sequences) were recorded as highly conserved. If a sequence fragment was not present in at least 95% of sequences in the set, that fragment was removed from analysis. Based on this, sequences with approximately 99.5% conservation were selected as potential candidates for analysis and consideration for a universal vaccine as listed in **Table 1**.

In addition to this, all match sequences were imported into Geneious Prime® 2021.2.2 bioinformatic software and then translated into the amino acid sequence. The lead nucleotide sequences were compared with the sequence of the RefSeq and prominent mutant SARS-CoV-2 genomes to determine the location of the conserved sequences on the RefSeq and which mutations were significant (which mutations resulted in a change in the amino acid sequence).

The HLA (Human Leukocyte antigen) CD4 immunogenicity of the consensus sequences were assessed by immunoinformatic method as described by Dhanda and colleagues [16]. The tool to determine immunogenicity of peptides is freely available at http://tools.iedb.org/CD4episcore/. The method combines the ANN-based immunogenicity prediction with HLA class II binding prediction by

Table 1. 100-nucleotide (nt) long RNA sequence fragments from S-protein (NC_045512.2_SProt) RNA and the RefSeq genome (NC_045512.2) were measured against a set of 258,269 full-length SARS-CoV-2 genomic sequences for perfect match. Frag. Freq. = frequency of segment across all genomes.

RefSeq Position # (NT)	Reference Segment	Frag. Freq.	% Conservation
S-protein			
2945	CACGUCUUGACAAAGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGC	257,123	99.55
2946	ACGUCUUGACAAAGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCU	257,124	99.55
2947	CGUCUUGACAAAGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUG	257,150	99.56
2948	GUCUUGACAAAGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGC	257,116	99.55
2949	UCUUGACAAAGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGCA	257,109	99.55
2950	CUUGACAAAGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGCAG	257,118	99.55
2951	UUGACAAAGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGCAGA	257,116	99.55
2952	UGACAAAGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGCAGAA	257,086	99.54
2953	GACAAAGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGCAGAAA	257,245	99.60
2954	ACAAAGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGCAGAAAU	257,246	99.60
2955	CAAAGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGCAGAAAUC	257,134	99.55
2956	AAAGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGCAGAAAUCA	257,215	99.59
2957	AAGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGCAGAAAUCAG	257,216	99.59

Continued

2958	AGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACU UCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGC AGAAAUCAGA	257,213	99.59
2959	GUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUU CAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGCA GAAAUCAGAG	257,025	99.51
Full-length			
8409	UUAAAGAUUUCAUGUCAUUGUCUGAACAACUACGAAAACAAAUAC GUAGUGCUGCUAAAAAGAAUAACUUACCUUUUAAGUUGACAUGUG CAACUACUAG	256,963	99.49
8410	UAAAGAUUUCAUGUCAUUGUCUGAACAACUACGAAAACAAAUACG UAGUGCUGCUAAAAAGAAUAACUUACCUUUUAAGUUGACAUGUGC AACUACUAGA	256,963	99.49
8411	AAAGAUUUCAUGUCAUUGUCUGAACAACUACGAAAACAAAUACGU AGUGCUGCUAAAAAGAAUAACUUACCUUUUAAGUUGACAUGUGCA ACUACUAGAC	256,964	99.49
8412	AAGAUUUCAUGUCAUUGUCUGAACAACUACGAAAACAAAUACGUA GUGCUGCUAAAAAGAAUAACUUACCUUUUAAGUUGACAUGUGCAA CUACUAGACA	256,966	99.49
8413	AGAUUUCAUGUCAUUGUCUGAACAACUACGAAAACAAAUACGUAG UGCUGCUAAAAAGAAUAACUUACCUUUUAAGUUGACAUGUGCAAC UACUAGACAA	256,956	99.49

7-allele method at the population level. To assess the immunogenicity properties, the conserved region peptide sequences were parsed into 15 amino acid long peptides and the immunogenicity score, 7-allele HLA class II binding score at population level, and combined scores were calculated for each peptide fragment using the immunoinformatic tool.

3. Results

The present study used 258,269 SARS-CoV-2 full-length genomic sequences downloaded from the NCBI database for analysis. The sequences represent all full-length SARS-CoV-2 genomes submitted to the NCBI database corresponding to the North America region as of February 13, 2022. Bioinformatic analysis using a custom Perl script using the whole gene sequence of S-protein and the full-length genomic sequence resulted in a 114-nucleotide segment corresponding to the region 2945 - 3058 on the spike protein (region 24,507 - 24,620 on the full-length genome) and a 104-nucleotide segment corresponding to the region 8409 - 8512 on the non-spike protein section of the viral genome.

The 100-nucleotide sequences corresponding to the above two highly conserved sequence regions are depicted in Table 1. The sequences in the spike protein region 2945 - 3058 were most conserved, showing conservation frequency of more than 99.5%. The next most conserved sequence region 8409 -

8512 was on the non-spike protein region of the genome and had a conservation frequency of 99.49%. Sequences with lower conservation frequencies are not shown.

Figure 2(a) depicts the alignment of three sequences from the region 2945 - 3058 of the spike protein with the spike protein (NC_045512.2_S: 21,563 - 25,384) and full-length RefSeq NC_0455512.2, Delta B.1.617.2, Alpha B.1.1.7, Gamma P.1, Beta B.1.351, Iota B.1.526, ETA B.1.525, and Omicron B.1.1.529 (OM570283). Tick marks indicate the positions of mutations in the variant viral genome compared to the RefSeq. Gaps indicate the missing nucleotide(s). Figure 2(a) depicts that the receptor binding domain (RBD) is highly mutable and that the S2 domain is more stable as indicated by the observation of a highly conserved sequence segment with >99.5% conservation. The next best conserved sequence with 99.49% conservation occurred in the non-spike protein region of the SARS-CoV-2 genome (Table 1). Figure 2(b) depicts the alignment of three sequences with the full-length genomic sequence (NC_045512), spike protein sequence (NC_045512.2Sprot), the RBD (in the S1 domain of the spike protein)

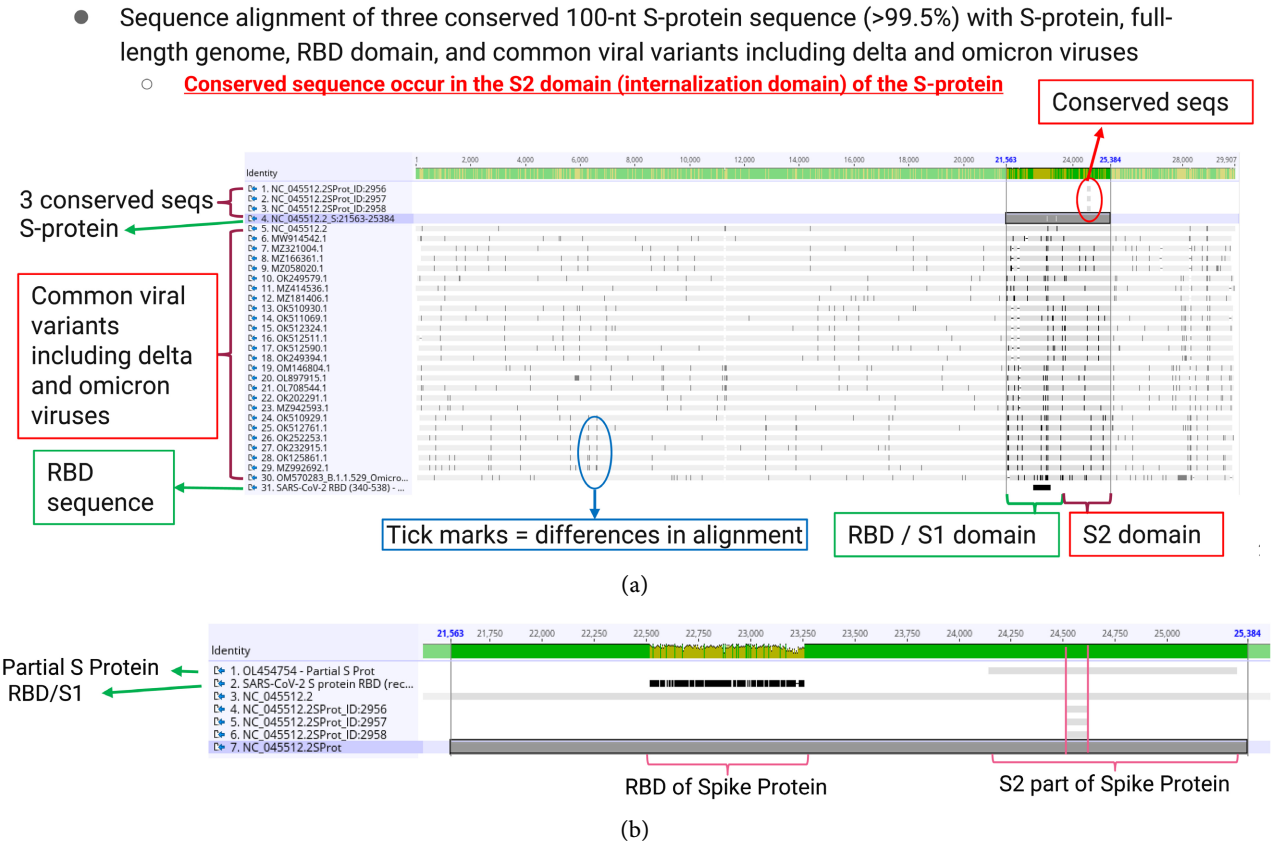

Figure 2. (a) Alignment of three 100-nucleotide sequences from the analysis of 258,269 full-length SARS-CoV-2 genomic sequences against the spike protein and full-length RefSeq NC_0455512.2, Delta B.1.617.2, Alpha B.1.1.7, Gamma P.1, Beta B.1.351, Iota B.1.526, ETA B.1.525, and Omicron B.1.1.529 (OM570283). Tick marks indicate the positions of mutations in the variant viral genome compared to the RefSeq; (b) Three 100-nucleotide sequences are shown aligned with the full-length genomic sequence (NC_045512.2), spike protein sequence (NC_045512.2Sprot), the receptor binding domain (RBD) sequence, and partial S2 part of the spike protein (OL454754). The three highly conserved sequences align with the S2 domain of the spike protein.

sequence, and partial S2 part of the spike protein (OL454754). The three highly conserved sequences on the spike protein align with the S2 domain (internalization domain) of the spike protein.

Figure 3 depicts the alignment of three 100-nucleotide sequences from the conserved S protein region with the RefSeq. The corresponding translated protein sequences are also shown. The sequence ID: 2956 is in frame with the RefSeq and depicts the corresponding protein sequence. This sequence can be used for vaccine design. The sequences ID: 2957 and ID: 2958 are out of frame, so corresponding protein sequences are not shown. The out of frame sequences cannot be used for vaccine design.

Immunogenic Properties of Conserved Sequences

The immunogenicity properties of both the most conserved protein regions on the SARS-CoV-2 genome were evaluated by immunoinformatic analysis according to the method described by Dhanda and colleagues [16]. The in frame 111 nucleotides from 24,509 - 24,619 from the 114-nucleotide segment from 2945 - 3058 region (region 24,507 - 24,620 on the full-length genome) on the spike protein ("CGUCUUGACAAAGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGCAGAAAUCAGA"; the corresponding protein is "RLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIR") and the in frame sequence Seq ID: 8411 (corresponding protein sequence "KDFMSLSEQLRKQIRSAAKKNNLPFKLTCATTR") from the 104-nucleotide segment corresponding to the region 8409 - 8512 on the non-spike protein section of the viral genome were used for parsing into 15 amino acids long sequences and immunoinformatic analysis.

The results of immunoinformatic analysis of the segment of spike protein sequence "RLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIR" are summarized in Table 2. The protein sequence was parsed into 22 peptide segments of 15 amino acids for the immunoinformatic analysis. Eleven peptide sequences out of 22 showed a high immunogenic score of approximately 90.00 or more. A combined immunogenicity and HLA binding score was more than 50 percentiles for the all eleven peptides.

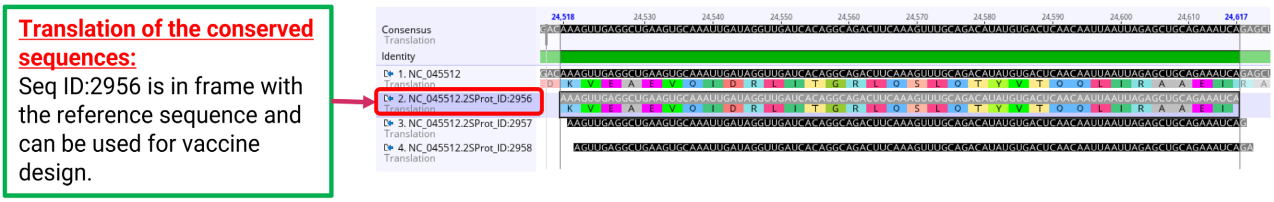

Figure 3. The three 100-nucleotide sequences are shown to align with the RefSeq. The corresponding translated protein sequences are also shown. The sequence ID: 2956 is in frame with the RefSeq and depicts the corresponding protein sequence, and can be used for vaccine design. The sequences ID: 2957 and ID: 2958 are out of frame; no corresponding protein sequences are depicted.

Table 2. Immunoinformatic analysis of the protein sequence "RLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIR" from the spike protein corresponding to the 111 nucleotides from 24,509 - 24,619 on the full-length SAR-CoV-2 genomic sequence. The protein sequence was parsed into 22 peptide segments of 15 amino acids and analyzed by immunoinformatic tools on http://tools.iedb.org/CD4episcore/ at a combined threshold of 60%. Analysis resulted in an immunogenicity score on a scale from 0 - 100, where 100 being the most immunogenic, and a 7-allele HLA binding score in percentile range from 0 - 100. The two scores are represented together into a combined score 0 - 100.

Peptide Number	Peptide	Peptide core	Immunogenicity Score	Median Percentile Rank (7-allele)	Combined Score
4	VEAEVQIDRLITGRL	VQIDRLITG	96.54	25	53.61
5	EAEVQIDRLITGRLQ	VQIDRLITG	97.05	25	53.82
6	AEVQIDRLITGRLQS	IDRLITGRL	97.16	25	53.86
7	EVQIDRLITGRLQSL	IDRLITGRL	97.49	25	54.00
8	VQIDRLITGRLQSLQ	IDRLITGRL	97.74	25	54.09
15	TGRLQSLQTYVTQQL	LQSLQTYVT	95.80	36	59.92
18	LQSLQTYVTQQLIRA	LQSLQTYVT	97.67	33	58.87
19	QSLQTYVTQQLIRAA	YVTQQLIRA	98.67	34	59.87
20	SLQTYVTQQLIRAAE	YVTQQLIRA	97.27	34	59.31
21	LQTYVTQQLIRAAEI	YVTQQLIRA	97.33	29	56.33
22	QTYVTQQLIRAAEIR	QLIRAAEIR	89.74	32	55.10

The results of immunoinformatic analysis of 18 peptide segments of 15 amino acids belonging to the protein sequence "KDFMSLSEQLRKQIRSAAKKNNLPFKLTCATTR" corresponding to the sequence Seq ID: 8411 are summarized in Table 3. Two of the 18 peptide sequences showed an immunogenicity score of more than 90, while four peptide segments showed somewhat respectable immunogenicity scores. A combined immunogenicity and HLA binding score was more than 50 percentiles for five peptides.

Figure 4(a) depicts the design of an RNA vaccine. The 99-nt fragment from the lead sequence ID: 2956 or any in frame sequence from the region 2945 - 3058 on the spike protein can be used for universal vaccine design by repeating and concatenating multiple fragments separated by 3- or 6-nucleotide sequences. The translation of 99 nucleotides will result in a 33-amino acid epitope. A 3'-UTR and 5'-UTR sequence is appended to the designed sequence to facilitate translation. The RNA also contains a 5'G-ppp cap to promote translation. The RNA is encapsulated in a lipid nanoparticle for delivery into cells. Figure 4(b) shows the current Pfizer/BioNTech or Moderna vaccines design that uses prefusion stabilized full-length S-protein mRNA.

4. Discussion

Vaccine efficacy can be compromised by the emergence of viral mutations that affect the binding of neutralizing antibodies raised against the vaccine or previous

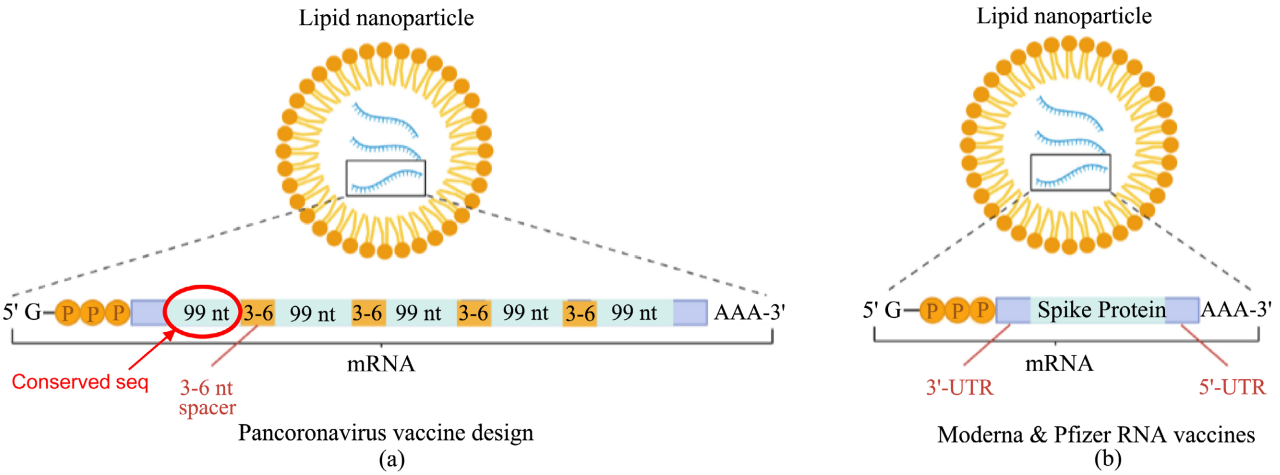

Figure 4. (a) RNA vaccine design is shown. A 99-nt fragment from the lead sequence can be used for universal vaccine design by repeating and concatenating multiple fragments separated by 3- or 6-nucleotide sequences. A 3'-UTR and 5'-UTR sequence is appended to the designed sequence to facilitate translation. The RNA contains a 5'G-ppp cap for translation. The RNA is encapsulated in a lipid nanoparticle for delivery into cells; (b) shows the current Moderna or Pfizer/BioNTech vaccines design that use either the whole S-protein or a large S2 fragment of the S-protein.

Table 3. Immunoinformatic analysis of the protein sequence "KDFMSLSEQLRKQIRSAAKKNNLPFKLTCATTR" corresponding to the sequence Seq ID: 8411 from the region 8409 - 8512 on the non-spike protein section of the viral genome. The protein sequence was parsed into 18 peptide segments of 15 amino acids. An immunogenicity score on a scale from 0 - 100, a 7-allele HLA binding score in percentile range from 0 - 100, and the combined score for the most relevant immunogenic peptide is listed.

Protein Number	Peptide	Peptide core	Immunogenicity Score	Median Percentile Rank (7-allele)	Combined Score
4	SLSEQLRKQIRSAAK	LRKQIRSAA	90.76	34	56.70
5	LSEQLRKQIRSAAKK	KQIRSAAKK	78.51	33	51.21
6	SEQLRKQIRSAAKKN	LRKQIRSAA	74.33	32	48.93
7	EQLRKQIRSAAKKNN	KQIRSAAKK	78.36	32	50.54
8	QLRKQIRSAAKKNNL	KQIRSAAKK	81.05	34	52.82
18	KKNNLPFKLTCATTR	KKNNLPFKL	93.34	35	58.34

viral infection(s). The emergence of Delta, Omicron, and other variants of the SARS-CoV-2 virus compromising the efficacy of vaccination raised the concern of vaccine efficacy and required booster doses of vaccines to neutralize new viral variants [17].

Because the spike protein is integral to virus interaction with the cell followed by internalization into cells, the spike protein is considered a prime target for vaccine design [18]. The RBD within the S1 domain of the spike protein is responsible for binding to the ACE2 receptor on the cell surface. The S2 domain of the spike protein is responsible for virus internalization. As the RBD is poorly conserved between SARS-CoVs and other pathogenic human coronaviruses, the RBD represents a promising antigen for detecting coronavirus-specific antibodies in humans [9]. A structure-function and antigenicity study suggested that the S-protein of the Delta variant has evolved to optimize the fusion step to enter

cells while the overall structure of the RBD is preserved among all SARS-CoV-2 variants and the reoccurring mutations appear to be limited to a number of sites. Therefore, the study suggested that therapeutic antibodies or universal vaccines should not target the N-terminal domain (NTD) because the escape from anti-NTD antibodies appear to be little cost to the virus; instead, the RBD is a better target for therapeutic antibodies [19]. A comparative efficacy study of two mRNA vaccines BNT162b1, which encodes a secreted trimerized SARS-CoV-2 receptor-binding domain, and BNT162b2, which encodes a prefusion stabilized membrane-anchored SARS-CoV-2 full-length spike, revealed that the vaccine containing full-length spike protein mRNA resulted in less systemic reactogenicity compared to the vaccine containing RBD mRNA [20]. On the contrary, the current analysis of 258,269 full-length SARS-CoV-2 genomic sequences revealed that the RBD is highly mutable and the S2 domain is more stable as indicated by the observation of a highly conserved sequence segment (**Figure 2(a)**), suggesting that the S2 domain of the spike protein is a more suitable target for vaccine development across present and future SARS-CoV-2 strains.

RNA vaccines contain an mRNA code that is translated in cells to produce a peptide/protein antigen. Longer peptides are processed into shorter peptides by antigen presenting cells (APC) that are recognized by $CD8^+$ cytotoxic T lymphocytes (CTL) and helper ($CD4^+$) T-cells. Typically, $CD8^+$ CTL recognizes 8 - 11-amino acid linear peptides presented in association with Class I MHCs (Major Histocompatibility Class) on APC, and helper ($CD4^+$) T-cells recognize 11 - 30-amino acid long peptides presented in association with Class II MHC [21] [22]. Studies have suggested that long 30-amino acid peptides encompassing short minimal epitopes may be more effective immunogens [22]. The use of full-length spike protein mRNA, as used in Moderna and Pfizer/BioNTech mRNA vaccines, will result in a heterogeneous population of short peptide antigens that will in turn result in a heterogeneous population of antibodies. Some of these antibodies are overrepresented and stored in memory B-cells to neutralize follow on viral infection, while other antibodies are underrepresented and suppressed [23]. The overrepresented antibody may not be effective against mutant viruses because antibodies may not be most immunodominant or target non-conserved or non-neutralizing epitopes [21]. The current study proposes an mRNA vaccine design with a unique highly conserved 99 - 114 nucleotides from spike protein RNA. The immunoinformatic analysis of protein sequence from the 111-nucleotide segment from the highly conserved 2945 - 3058 nucleotide S-protein sequence suggest a highly immunogenic protein sequence. This study also revealed that the next best highly conserved but lesser immunogenic sequence belonged to non-Spike protein sequence of the SARS-CoV-2 genome. This study supports the notion that spike protein is a highly important target for SARS-CoV-2 vaccine development. To increase the robust production of corresponding peptides upon vaccination, the nucleotide segment can be repeated 5 - 30 or more times, thus producing epitopes for production of unique antibodies for a conserved sequence as shown in **Figure 4**. The other lesser conserved RNA

sequences can be combined with the most conserved sequence to produce a number of antibodies targeting only highly conserved antigen sequences on the virus. Antibodies target their epitopes in either conformation-independent or -dependent manner. Conformation-independent epitopes are generally linear stretched of amino acids that are usually found in protein loops and are strong candidate for peptide vaccine design. However, rigidifying peptide epitope conformation has advantages because they may more closely match the antigen structure. To fix the conformation of the resulting epitopes from the 99 - 114 nucleotides from spike protein RNA, disulfide bonds can be introduced to circularize the resulting peptide by introducing cysteine residues at the C-terminus and N-terminus of the peptide [21].

5. Conclusion

The conserved sequences selected from this study may be useful in developing an RNA or protein vaccine that could be effective against future SARS-CoV-2 strains or could act as a universal coronavirus vaccine if these sequences are present in other coronavirus families. The method used in this study can also be used to find conserved sequences across all coronavirus families, such as MERS and SARS-CoV-1. A conserved sequence among all families of coronaviruses can be used to design a universal coronavirus vaccine targeting all families of coronaviruses. Designing, manufacturing, and testing the vaccine for effectiveness in laboratory settings is under investigation.

Acknowledgements

Authors thank Ms. Kate Allender for advice on the project.

Author Contributions

"Conceptualization, methodology, formal analysis, investigation, writing—original draft preparation, visualization, software, A.V.; software, data curation, writing—review and editing, supervision, J.M. All authors have read and agreed to the published version of the manuscript".

Conflicts of Interest

Authors declare no conflict of interest.

References

[1] World Health Organization (2022) Coronavirus Disease (COVID-19) Pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[2] Krammer, F. (2020) SARS-CoV-2 Vaccines in Development. *Nature*, **586**, 516-527. https://doi.org/10.1038/s41586-020-2798-3

[3] Nagy, A. and Alhatlani, B. (2021) An Overview of Current COVID-19 Vaccine Platforms. *Computational and Structural Biotechnology Journal*, **19**, 2508-2517. https://doi.org/10.1016/j.csbj.2021.04.061

[4] Corbett, K.S. (2021) Immune Correlates of Protection by mRNA-1273 Vaccine

against SARS-CoV-2 in Nonhuman Primates. *Science*, **373**, eabj0299.

[5] Pegu, A., *et al.* (2021) Durability of mRNA-1273 Vaccine-Induced Antibodies against SARS-CoV-2 Variants. *Science*, **373**, 1372-1377.

[6] Cao, Y., *et al.* (2022) Omicron Escapes the Majority of Existing SARS-CoV-2 Neutralizing Antibodies. *Nature*, **602**, 657-663. https://doi.org/10.1038/s41586-021-04385-3

[7] Dai, L., *et al.* (2020) A Universal Design of Betacoronavirus Vaccines against COVID-19, MERS, and SARS. *Cell*, **182**, 722-733. https://doi.org/10.1016/j.cell.2020.06.035

[8] Koff, W.C. and Berkley, S.F. (2021) A Universal Coronavirus Vaccine. *Science*, **371**, 759-759. https://doi.org/10.1126/science.abh0447

[9] Premkumar, L., *et al.* (2020) The Receptor Binding Domain of the Viral Spike Protein Is an Immunodominant and Highly Specific Target of Antibodies in SARS-CoV-2 Patients. *Science Immunology*, **5**, eabc8413.

[10] Cohen, J. (2021) The Dream Vaccine. *Science*, **372**, 227-231. https://doi.org/10.1126/science.372.6539.227

[11] Lei, K.C. and Zhang, X.D. (2020) Conservation Analysis of SARS-COV-2 Spike Suggests Complicated Viral Adaptation History from Bat to Human. *Evolution, Medicine, and Public Health*, **2020**, 290-303. https://doi.org/10.1093/emph/eoaa041

[12] Chan, A.P., Choi, Y. and Schork, N.J. (2020) Conserved Genomic Terminals of SARS-COV-2 as Coevolving Functional Elements and Potential Therapeutic Targets. *MSphere*, **5**, e00754-20. https://doi.org/10.1128/mSphere.00754-20

[13] Forster, P., *et al.* (2020) Phylogenetic Network Analysis of SARS-COV-2 Genomes. *Proceedings of the National Academy of Sciences*, **117**, 9241-9243. https://doi.org/10.1073/pnas.2004999117

[14] Li, M., *et al.* (2021) Rational Design of a Pan-Coronavirus Vaccine Based on Conserved CTL Epitopes. *Viruses*, **13**, 333. https://doi.org/10.3390/v13020333

[15] Xia, X. (2021) Domainss and Functions of Spike Protein in SARS-COV-2 in the Context of Vaccine Design. *Viruses*, **13**, 109. https://doi.org/10.3390/v13010109

[16] Dhanda, S.K., Karosiene, E., *et al.* (2018) Prediction of HLA CD4 Immunogenicity in Human Populations. *Frontiers in Immunology*, **9**, Article No. 1369. https://doi.org/10.3389/fimmu.2018.01369

[17] Planas, D., *et al.* (2021) Considerable Escape of SARS-CoV-2 Omicron to Antibody Neutralization. *Nature*, **602**, 671-675. https://doi.org/10.1038/s41586-021-04389-z

[18] Li, X., *et al.* (2021a) Possible Targets of Pan-Coronavirus Antiviral Strategies for Emerging or Re-Emerging Coronaviruses. *Microorganisms*, **9**, Article No. 1479. https://doi.org/10.3390/microorganisms9071479

[19] Zhang, J., *et al.* (2021) Membrane Fusion and Immune Evasion by the Spike Protein of SARS-CoV-2 Delta Variant. *Science*, **374**, 1353-1360. https://doi.org/10.1126/science.abl9463

[20] Walsh, E.E., *et al.* (2020) Safety and Immunogenicity of Two RNA-Based Covid-19 Vaccine Candidates. *The New England Journal of Medicine*, **383**, 2439-2450. https://doi.org/10.1056/NEJMoa2027906

[21] Malonis, R.J. (2020) Peptide-Based Vaccines: Current Progress and Future Challenges. *Chemical Reviews*, **120**, 3210-3229. https://doi.org/10.1021/acs.chemrev.9b00472

[22] Slingluff Jr., C.L. (2011) The Present and Future of Peptide Vaccines for Cancer:

Single or Multiple, Long or Short, Alone or in Combination? *Cancer Journal*, **17**, 343-50. https://doi.org/10.1097/PPO.0b013e318233e5b2

[23] Cho, A., *et al.* (2021) Anti-SARS-CoV-2 Receptor-Binding Domain Antibody Evolution after mRNA Vaccination. *Nature*, **600**, 517-522. https://doi.org/10.1038/s41586-021-04060-7