doi:10.20944/preprints202004.0024.v1

EMERGENCE OF DRIFT VARIANTS THAT MAY AFFECT COVID-19 VACCINE DEVELOPMENT AND ANTIBODY TREATMENT

Takahiko Koyama[1], Dilhan Weeraratne[2], Jane L. Snowdon[2], Laxmi Parida[1]

[1] IBM TJ Watson Research Center, Yorktown Heights, NY 10598, USA
[2] IBM Watson Health, Cambridge, MA 02142, USA

Address for Correspondence: Takahiko Koyama, PhD, IBM T. J. Watson Research Center, 1101 Kitchawan Rd., Yorktown Heights, NY 10598, USA (email: tkoyama@us.ibm.com).

ABSTRACT

New coronavirus (SARS-CoV-2) treatments and vaccines are under development to combat the COVID-19 disease. Several approaches are being used by scientists for investigation including 1) various small molecule approaches targeting RNA polymerase, 3C-like protease, and RNA endonuclease and 2) exploration of antibodies obtained from convalescent plasma from patients who have recovered from COVID-19. The coronavirus genome is highly prone to mutations that lead to genetic drift and escape from immune recognition; thus, it is imperative that sub-strains with different mutations are also accounted for during vaccine development. As the disease has grown to become a pandemic, new B-cell and T-cell epitopes predicted from SARS coronavirus have been reported. Using the epitope information along with variants of the virus, we have found several variants which might cause drifts. Among such variants, 23403A>G variant (p.D614G) in spike protein B-cell epitope is observed frequently in European countries such as the Netherlands, Switzerland and France.

INTRODUCTION

In late 2019, a new coronavirus, SARS-CoV-2, causing acute respiratory distress syndrome was first reported in Wuhan, China. Despite a lockdown of the city, the number of patients kept increasing exponentially while in parallel the virus spread across the globe. The World Health Organization (WHO) declared a pandemic on March 11, 2020. Currently, no treatments or vaccines are scientifically proven to be effective against the virus. Safe and effective vaccines of SARS-CoV-2 are urgently needed to put an end to this pandemic. To that end, a clinical trial of mRNA-1273 with full spike protein as an antigen started on March 8, 2020(1).

To date, no approved treatments for COVID-19 exist. Pharmaceutical companies are currently investigating repurposed compounds from other infections. For instance, lopinavir and ritonavir are both HIV protease inhibitors; however, the treatment benefit derived was dubious in a lopinavir–ritonavir clinical trial just reported(2). Remdesivir, an RNA polymerase inhibitor originally intended to treat Ebola virus appears to have in-vitro activity against SARS-CoV-2(3). Additionally, convalescent immunoglobulins derived from recovering patients is currently being investigated as a potential treatment for the disease(4). Before vaccines become widely available, these treatments are the best hope for striving to keep the mortality rate low.

In order to make vaccines, antigens from pathogens are used. Typically, surface proteins outside of the viral virion are selected for antigens so that antibodies generated from vaccine trained B-cell can bind to the virus for neutralization. In addition to the B-cell epitope requirement, the antigens must generate antigenic peptides which bind to major histocompatibility complex (MHC) molecules to be presented. By presenting a peptide, a B-cell can get stimulated from a helper T-cell and become a plasma cell to generate antibodies. Some portion of such stimulated B-cell is transferred to the germinal center where B-cells are further enhanced from random somatic mutagenesis induced by AID to make binding stronger to the antigen. Therefore, the resulting antibodies have differences in binding epitope and protein sequences in antibody variable regions. The antigens introduced as vaccines needs to account for current major sub-strains to prevent potential immune escapes.

Genetic drift takes place when the occurrence of alleles, or variant forms of a gene, increase or decrease over time(5). Genetic drift is measured by the changes in allele frequencies and continues until one of two possible events occurs: the involved allele is lost by a population or the involved allele is the only allele present in a population at a particular locus. Genetic drift may cause a new population to be genetically distinct from the original population. This study's objective is to interrogate currently identified sub-strains of SARS-CoV-2 and identify genetic drifts and potential immune recognition escape sites that would be integral for a development of a successful vaccine.

RESULTS

Twelve distinct variants were found within B-cell epitopes of S (spike protein), N (nucleocapsid protein), and M (membrane protein), respectively as listed in Table 1. Also, twenty-two distinct variants were identified in T-cell epitopes.

Among the twelve variants in the B-cell epitopes, 23403A>G variant (p.D614G) in one of the epitopes in spike protein between residue 601 and 640 stands out with 175 samples in 615 total samples. The variant is located in the middle of that epitope and the amino acid change in 23403A>G variant (p.D614G) involves a change of large acidic residue D (aspartic acid) into small hydrophobic residue G (glycine). Such large differences in both size and hydrophobicity in the middle of the epitope would make binding affinity to antibodies trained by vaccines with wild type

Table 1. SARS-CoV2 variants which occur in the predicted epitopes

CELL TYPE	EPITOPE	PROTIEN	RESIDUES	AMINO ACID CHANGE	BASE CHANGE	NUMBER OF SAMPLES
B-CELL	GTNTSNQVAVLYQD**V**NCTEVPVAIHADQLTPTWRVYSTGS	S	601-640	p.V615L	23405G>C	1
B-CELL	GTNTSNQVAVLYQ**D**VNCTEVPVAIHADQLTPTWRVYSTGS	S	601-640	p.D614G	23403A>G	175
B-CELL	FSQILPDPSKPSKRS**F**IE	S	802-819	p.F817L	24011T>C	1
B-CELL	FSQILPDPSK**P**SKRSFIE	S	802-819	p.P812S	23996C>T	1
B-CELL	FGAGAALQIPFAMQ**M**AYRFNGI	S	888-909	p.M902fs	24268del	1
B-CELL	MA**D**SNGTITVEELKKLLEQWNLVI	M	1-24	p.D3G	26530A>G	5
B-CELL	RPQGL**P**NNTASWFTALTQHGK	N	41-61	p.P46S	28409C>T	1
B-CELL	NNN**A**ATVLQLPQGTTLPKGF	N	153-172	p.A156S	28739G>T	2
B-CELL	NKHIDAYKTFPPTEPKKDKKKKTD**E**AQPLPQRQKKQPTVTLLPAADM	N	355-401	p.E378Q	29405G>C	1
B-CELL	NKHIDAYKTFPPTEPKKD**K**KKKTDEAQPLPQRQKKQPTVTLLPAADM	N	355-401	p.K373N	29392G>T	1
B-CELL	NKHIDAYKTFPPTEP**K**KDKKKKTDEAQPLPQRQKKQPTVTLLPAADM	N	355-401	p.K370N	29383G>T	1
B-CELL	NKHIDAYKTF**P**PTEPKKDKKKKTDEAQPLPQRQKKQPTVTLLPAADM	N	355-401	p.P365S	29366C>T	1
T-CELL	QPFLMDLE**G**KQGN	S	173-185	p.G181V	22104G>T	1
T-CELL	TRFQTLLALHRSYLTPGD**S**SSGW	S	236-258	p.S254F	22323C>T	2
T-CELL	TRFQTLLALHR**S**YLTPGDSSSGW	S	236-258	p.S247R	22303T>A/G	3
T-CELL	TRFQT**LLA**LHRSYLTPGDSSSGW	S	236-258	p.L241_A243del	22281_22289del	1
T-CELL	TRF**Q**TLLALHRSYLTPGDSSSGW	S	236-258	p.Q239K	22277C>A	6
T-CELL	NLDSKVGGNYNYLYRL**F**R	S	440-457	p.F456fs	22928del	1
T-CELL	YLYRLFR**K**SNLKPFERDI	S	451-468	p.K458R	22935A>G	1
T-CELL	YLYRL**F**RKSNLKPFERDI	S	451-468	p.F456fs	22928del	1
T-CELL	TECSN**L**LLQYGSFCTQL	S	747-763	p.L752F	23816C>T	1
T-CELL	VKQIYKTPPIKD**F**GGFNF	S	785-802	p.F797C	23952T>G	1
T-CELL	VKQIYK**T**PPIKDFGGFNF	S	785-802	p.T791I	23934C>T	1
T-CELL	DSLSSTA**S**ALGKLQDVV	S	936-952	p.S943T	24390G>C	4
T-CELL	DSLSSTA**S**ALGKLQDVV	S	936-952	p.S943R	24389A>C	3
T-CELL	DSLSS**T**ASALGKLQDVV	S	936-952	p.T941A	24383A>G	1
T-CELL	DSLS**S**TASALGKLQDVV	S	936-952	p.S940F	24381C>T	2
T-CELL	DSL**S**STASALGKLQDVV	S	936-952	p.S939F	24378C>T	2
T-CELL	RLNEV**A**KNL	S	1185-1193	p.A1190G	25131C>G	1
T-CELL	RL**N**EVAKNL	S	1185-1193	p.N1187K	25123T>A	1
T-CELL	RI**F**TIGTVTLKQGEI	ORF3a	6-20	p.F8L	25414T>C	1
T-CELL	GMSRIG**M**EV	N	316-324	p.M322I	29239G>T	1
T-CELL	MEVTP**S**GTWL	N	322-331	p.S327L	29253C>T	1

spike protein compromised. Most of the samples with the variant were collected in Europe, in particular Netherlands (66 out of 112), Switzerland (29 out of 30) and France (21 out of 32) as shown in Table 2. In these countries, the majority of infected patients possess the variant; therefore, vaccine design and convalescent plasma antibody treatment might require further considerations to accommodate the drift.

Table 2. Statistics of 23403A>G variant (p.D614G) in spike protein observed by country.

COUNTRY	VARIANT COUNT	TOTAL COUNT
NETHERLANDS	66	112
SWITZERLAND	29	30
FRANCE	21	32
UNITED KINGDOM	12	30
USA	9	123
BRAZIL	8	13
BELGIUM	7	8
FINLAND	6	7
PORTUGAL	2	2
ITALY	2	6
IRELAND	2	3
GERMANY	2	9
DENMARK	2	2
CHINA	2	151
RUSSIA	1	1
MEXICO	1	1
LUXEMBURG	1	1
GEORGIA	1	3
CHILE	1	7

METHODS

Predicted epitopes for B-cell and T-cell were obtained from results of assays performed for SARS-CoV and sequence alignments between SARS-CoV and SARS-CoV-2(6). The sequence identity and similarity of spike protein between the strains is 76.3% and 87.0% after running Needle pairwise alignment(7). As shown in Figure 1, the spike protein sequences of SARS-CoV and SARS-CoV-2 have high similarity in the regions of interest, which are colored in blue. For instance, in the segment between 601-640, 32/41 (78%) of the residues are identical, 5/41 (12%) of the residues are similar, and 4/41 (10%) of the residues are dissimilar, respectively.

615 variants data files in GFF3 format were downloaded from China's National Genomics Data Center (NGDC) (https://bigd.big.ac.cn/ncov/release_genome?lang=en) on March 20, 2020. They provide the variant information from GISAID (Global Initiative on Sharing All Influenza Data)(8), GenBank, NGDC Genome Warehouse, and National Microbiology Data Center (NMDC). Sample information is provided in Supplementary Table 1. Samples with hyper mutations and large gaps are considered of low quality and discarded from the analysis. GFF3 were processed to extract sample information including geographic location, coordinate information, base changes, genes, amino acid changes, and variant types and organized into a database.

For each epitope, variants located within the predicted epitope segments were tabulated. Country based statistics of the prevalence of 23403A>G variant (p.D614G) were created.

DISCUSSION

Reports of reinfection and relapse of COVID-19 indicate that eliciting an effective host immune response to facilitate viral clearance can be a challenge at least in some patients. The variant 23403A>G (spike protein p.D614G) in the epitope will make even more challenging for antibodies, which are trained by antigens derived by wild type spike protein, to neutralize the virus. There is a possibility that a patient can be infected with both WT and p.D614G sub-strains in a short period of time if antibodies from the first infection mainly bind to the epitope of 601-640 residue of spike protein. Also, as viruses mutate during replication, host antibodies generated in the earlier phase of the infection may not be as effective later on (9).

23403A>G (p.D614G) variant sub-strains are prevalent in not just three countries. However, the number of total samples are rather small for Brazil, Finland, Belgium and other countries; it is difficult to determine which strains in terms of D614 status are dominant for those countries. Within the patient cohort we analyzed, the variant is first observed in EPI_ISL_406862 collected on January 28, 2020 in a sample from Germany. Subsequently the variant was detected in EPI_ISL_412982 collected on February 7, 2020 in a sample from Wuhan, China. Notably, these two samples do not share common variants besides p.D614G. It is not clear whether the variant emerged in China and disseminated to Europe or this variant emerged independently in China and Europe. Intriguingly, p.D614G variant was detected only in 2/151 Chinese patients analyzed and studies are required to understand the immunogenicity conferred by the two different alleles.

There are three other variants in spike protein B-cell epitopes besides p.D614G. However, these other variants are not observed in any other samples. Thus, these sub-strains might have lost fitness or it may be too early to evaluate the prevalence. These variants along with future variants in the epitopes need to be vigilantly monitored for potential drifts.

CONCLUSION

The highly prevalent 23403A>G (p.D614G) variant in the European population may cause antigenic drift resulting in vaccine mismatches offering little protection to that group of patients. Innovative vaccine design methods including using highly conserved internal epitopes, recombinant proteins spanning epitopes or pooling multiple vaccines will be required to combat the inherent antigenic drift. Consideration of drift variants in SARS-CoV-2 will offer cross-

doi:10.20944/preprints202004.0024.v1

protection across different sub-strains and obviate the need for reformulation of the vaccine for each distinct sub-strain. Additionally, consideration of drift variants in convalescent immunoglobulin treatment strategies will also result in better patient outcome. In conclusion, consideration of antigenic drift in the different subs-trains of the virus is imperative to the design of “one fit all” universal vaccine to offer protection against the deadliest outbreak in this century.

ACKNOWLEDGMENTS

The authors are grateful to medical and technical staff treating COVID-19 patients, collecting specimens from patients, sequencing genomes in a timely manner, and sharing the genomic sequences in publicly available repositories as listed in Supplemental Table 1.

REFERENCES

1. Safety and Immunogenicity Study of 2019-nCoV Vaccine (mRNA-1273) to Prevent SARS-CoV-2 Infection. https://ClinicalTrials.gov/show/NCT04283461.

2. Cao B, Wang Y, Wen D, Liu W, Wang J, Fan G, et al. A Trial of Lopinavir–Ritonavir in Adults Hospitalized with Severe Covid-19. New England Journal of Medicine. 2020.

3. Wang M, Cao R, Zhang L, Yang X, Liu J, Xu M, et al. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. Cell Research. 2020;30(3):269-71.

4. Chen L, Xiong J, Bao L, Shi Y. Convalescent plasma as a potential therapy for COVID-19. The Lancet Infectious Diseases.

5. Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, et al. Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet. 2016;17(11):704-14.

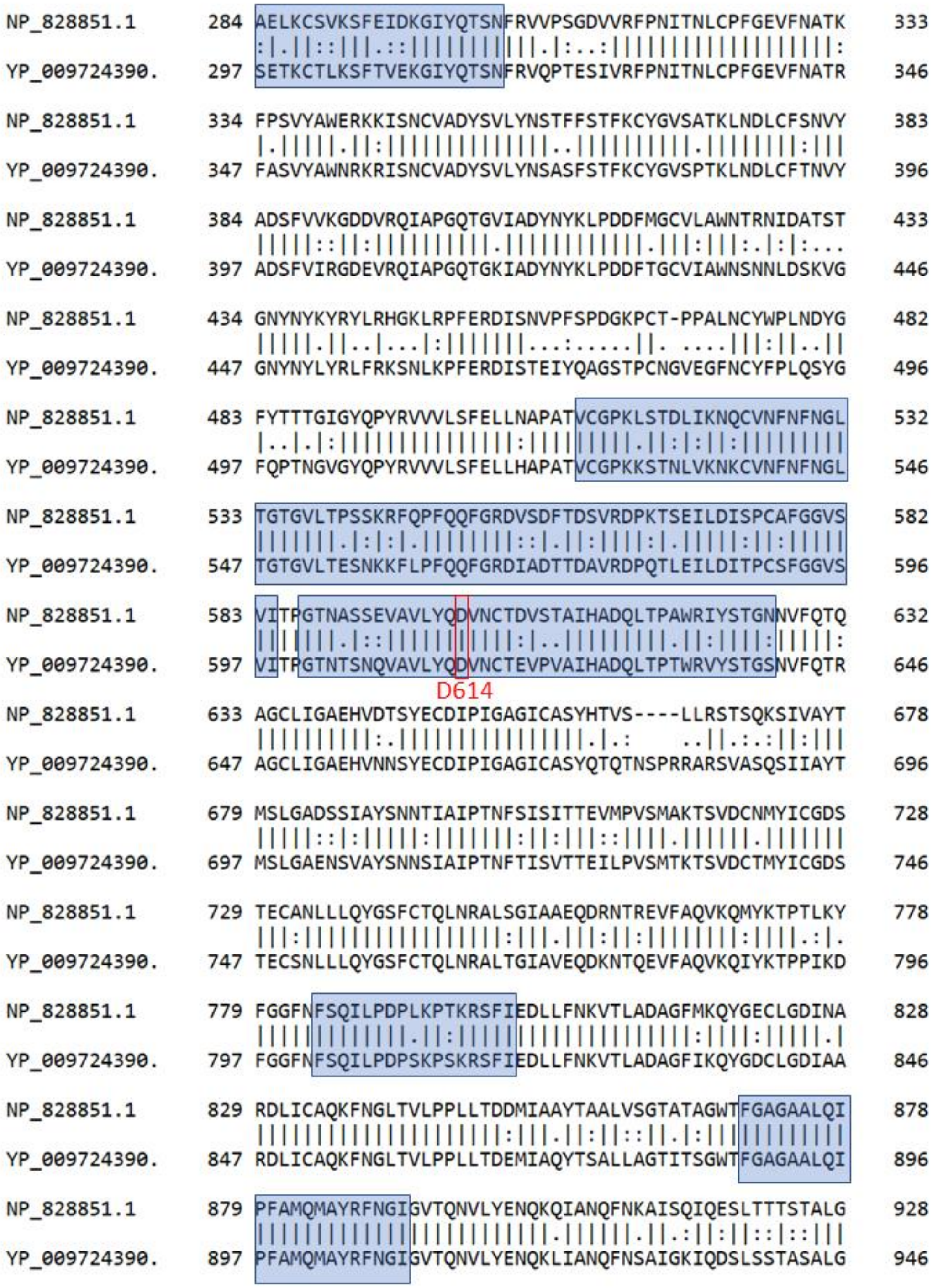
6. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host Microbe. 2020.

7. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology. 1970;48(3):443-53.

8. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 2017;22(13).

9. Shi Y, Wang Y, Shao C, Huang J, Gan J, Huang X, et al. COVID-19 infection: the perspectives on immune responses. Cell Death & Differentiation. 2020.

doi:10.20944/preprints202004.0024.v1



```
NP_828851.1       284 AELKCSVKSFEIDKGIYQTSNFRVVPSGDVVRFPNITNLCPFGEVFNATK    333
                      :|.||::|||.::|||||||||||.|:..:|||||||||||||||||||:
YP_009724390.     297 SETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATR    346

NP_828851.1       334 FPSVYAWERKKISNCVADYSVLYNSTFFSTFKCYGVSATKLNDLCFSNVY    383
                      |.|||||.||:||||||||||||||..||||||||||.||||||||:|||
YP_009724390.     347 FASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVY    396

NP_828851.1       384 ADSFVVKGDDVRQIAPGQTGVIADYNYKLPDDFMGCVLAWNTRNIDATST    433
                      |||||::||:||||||||||.||||||||||||.|||:|||:.|:|:...
YP_009724390.     397 ADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVG    446

NP_828851.1       434 GNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYG    482
                      |||||.||..|...|:|||||||...:.....||. ....|||:||..||
YP_009724390.     447 GNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYG    496

NP_828851.1       483 FYTTTGIGYQPYRVVVLSFELLNAPATVCGPKLSTDLIKNQCVNFNFNGL    532
                      |..|.|:|||||||||||||||:|||||||||.||:|:||:|||||||||
YP_009724390.     497 FQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGL    546

NP_828851.1       533 TGTGVLTPSSKRFQPFQQFGRDVSDFTDSVRDPKTSEILDISPCAFGGVS    582
                      |||||||.|:|:|.||||||||::|.||:||||:|.|||||:||:|||||
YP_009724390.     547 TGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVS    596

NP_828851.1       583 VITPGTNASSEVAVLYQDVNCTDVSTAIHADQLTPAWRIYSTGNNVFQTQ    632
                      |||||||.|::|||||||||||:|..|||||||||.||:||||:|||||:
YP_009724390.     597 VITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTR    646
                                       D614
NP_828851.1       633 AGCLIGAEHVDTSYECDIPIGAGICASYHTVS----LLRSTSQKSIVAYT    678
                      ||||||||||:.||||||||||||||||.|.:    ..||.:.:||:|||
YP_009724390.     647 AGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYT    696

NP_828851.1       679 MSLGADSSIAYSNNTIAIPTNFSISITTEVMPVSMAKTSVDCNMYICGDS    728
                      |||||::|:|||||:|||||||:||:|||::||||.||||||.|||||||
YP_009724390.     697 MSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDS    746

NP_828851.1       729 TECANLLLQYGSFCTQLNRALSGIAAEQDRNTREVFAQVKQMYKTPTLKY    778
                      |||:|||||||||||||||||:|||.|||:||:||||||||:||||.:|.
YP_009724390.     747 TECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKD    796

NP_828851.1       779 FGGFNFSQILPDPLKPTKRSFIEDLLFNKVTLADAGFMKQYGECLGDINA    828
                      |||||||||||||.||:||||||||||||||||||||:||||:|||||.|
YP_009724390.     797 FGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAA    846

NP_828851.1       829 RDLICAQKFNGLTVLPPLLTDDMIAAYTAALVSGTATAGWTFGAGAALQI    878
                      |||||||||||||||||||||:|||.||:||::||.|:||||||||||||
YP_009724390.     847 RDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQI    896

NP_828851.1       879 PFAMQMAYRFNGIGVTQNVLYENQKQIANQFNKAISQIQESLTTTSTALG    928
                      |||||||||||||||||||||||||.||||||.||.:||:||::|::|||
YP_009724390.     897 PFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALG    946
```

Figure 1. Pairwise sequence alignments of spike protein (S) between SARS-CoV (NP_828851.1) and SARS-CoV-2 (YP_009724390). Similarities in the predicted B-cell epitopes in blue are high. D614 residue is marked in a red rectangle.