EVALUATION OF THE EFFECT OF D614G, N501Y AND S477N MUTATION IN SARS-CoV-2 THROUGH COMPUTATIONAL APPROACH

Sarmilah Mathavan, Suresh Kumar*

Faculty of Health and Life Sciences, Management and Science University, Seksyen 13, 40100, Shah Alam, Selangor, Malaysia

***Corresponding author: Suresh Kumar, Tel: +60-14-2734893, Fax: +60-35-5112848, Email address: sureshkumar@msu.edu.my

ABSTRACT

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that causes an outbreak of COVID-19 disease in humans with the aid of spike protein. It contains a receptor-binding domain (RBD) that recognizes and binds to the host receptor angiotensin-converting enzyme 2 (ACE2). The aim of this research was to examine the mutational effect of spike protein on the sequence, the structural level and the interaction study between the mutant Spike protein and the human ACE2 protein. A total of 17,227 spike proteins from Asia, Africa, Europe, Oceania, South America and North America were retrieved and compared to the Wuhan spike protein reference sequence (Wuhan-Hu-1). The structural and stability implications of D614G, N501Y and S477N mutations were evaluated. The binding affinity between mutated RBD and human ACE2 protein was also studied. The D614G mutation may have originated in Germany, Europe on the basis of date of first sample collection report. It is now widely circulated all over the world with most occurrences in North America. The mutations N501Y and S477N may have originated from Oceania on the basis of date of first sample collection report and also have the highest occurrences in Oceania. Based on the computational analysis of mutational effects, the D614G, N501Y and S477N mutations decreased stability and were tolerated. For disease propensity prediction, N501Y was more prone to disease compared to D614G, while S477N was not prone to disease. The mutation of D to G at position 614 and S to N at position 477 for secondary structure prediction shows no changes in secondary structure while remaining in the coil region, whereas the mutation of N to Y at position 501 changes from coil structure to extended strand. N501Y mutation has a higher affinity to human ACE2 protein compared to D614G and S477N based on a docking study. D614G spike mutation identified to exist between the two hosts based on comparison of SARS-CoV-2 derived between the mink and human. Further research is needed on the link between the mink mutation N501T and the mutation N501Y in humans, which has evolved as a separate variant.

Keywords: COVID-19, SARS-CoV-2, D614G, N501Y, S477N, mink, VOC 202012/01, B.1.1.7

doi:10.20944/preprints202012.0710.v1

INTRODUCTION

Coronaviruses are a complex group of viruses that invade several distinct animals and therefore can cause respiratory illnesses in humans that are moderate to severe (Cui, Li, and Shi 2019). In the town of Wuhan, China, a novel coronavirus named SARS-CoV-2 emerged at the end of 2019, triggering an epidemic of rare viral pneumonia. This novel coronavirus disease, also known as coronavirus disease 2019 (COVID-19), has spread quickly all over the world as it is highly transmissible (J. T. Wu, Leung, and Leung 2020). It became an unprecedented vulnerability to global public health (Deng and Peng 2020).

Most of the first 27 recorded hospitalised patients were epidemiologically associated with the Huanan Seafood Wholesale Market, a downtown Wuhan wet market. The first known case dates back to 8 December 2019. Later, multiple patients were found with no background of Huanan Seafood Wholesale Market exposure. All these cases provided clear evidence for human-to-human transmission of the new virus (Jiang, Du, and Shi 2020).

Coronaviruses are enveloped viruses that have a single-stranded RNA genome of positive sense (26-32 kb). So far, four coronavirus genera (α, β, γ, δ) have also been established and α coronavirus (HCoV-229E and NL63) and β coronavirus (MERS-CoV, SARS-CoV, HCoV-OC43 and HCoV-HKU1) genera (Su et al. 2016). The existence of a previously unrecognized β-CoV strain in all of them was discovered by virus genome analysis of five patients with pneumonia hospitalised from December 18 to December 29, 2019 (Lu et al. 2020). The novel β-CoV was then named "SARS-CoV-2" by the International Virus Classification Commission. There are four main structural proteins of SARS-CoV-2 as shown in **Figure 1** (Begum et al. 2020).

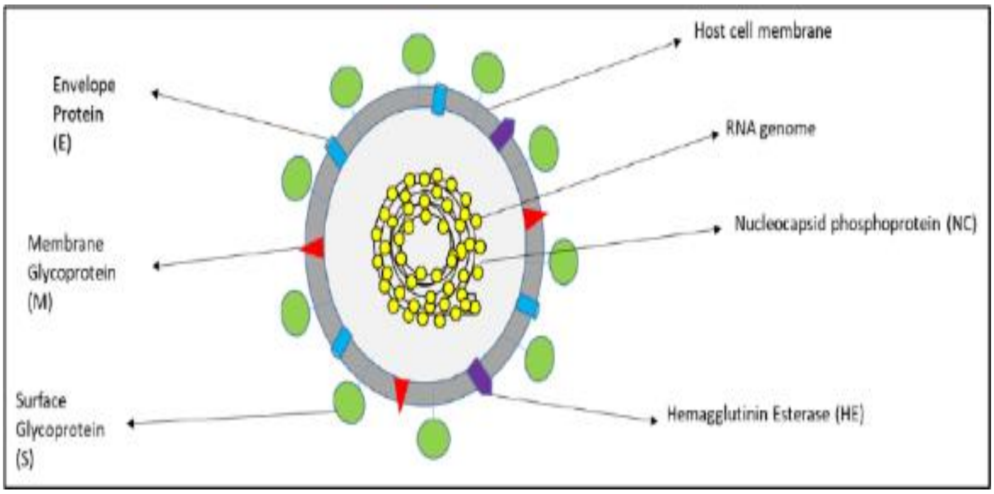

Host cell membrane
Envelope Protein (E)
RNA genome
Nucleocapsid phosphoprotein (NC)
Membrane Glycoprotein (M)
Surface Glycoprotein (S)
Hemagglutinin Esterase (HE)

Figure 1: Schematic Representation of SARS-CoV-2. It consists of functional protein of surface glycoprotein (S), an envelope protein (E), nucleocapsid phosphoprotein (N), and membrane glycoprotein (M).

SARS-CoV-2 contains six main open-label readings, including frames (ORFs) of the coding enzyme replication region (ORF 1a and 1b), the gene E (envelope protein), the gene M (membrane protein), the gene S (spike protein), and the gene N (nucleocapsid protein). Spike

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2020 doi:10.20944/preprints202012.0710.v1

glycoprotein is used to direct coronavirus penetration into host cells. The foundation of this research is Spike glycoprotein. It comprises two subunits that form the receptor-binding domain and the fusion peptide domain, respectively S1 and S2 (Zhou et al. 2020). Spike proteins are assembled into trimers upon this virion structure to produce a unique "corona" or crown-like profile. Spike protein is among the significant structural proteins, 1273 amino acids long, including two significant sub-domains, S1 and S2, comprises an N-terminal S1 subunit and a C-terminal membrane-proximal S2 subunit. The S1 subunit consists of S1A, S1B, S1C, and S1D domains. The S1A domain, referred to as the N-terminal domain (NTD), recognises carbohydrates, such as sialic acid required for attachment of the virus to the host cell surface. The S1B domain referred to as the receptor-binding domain (RBD) of the SARS-CoV-2 spike protein interacts with the human ACE-2 receptor (H. Zhang et al. 2020). RBD length from 319aa until 541aa **(Figure 2)**. The structural elements within the S2 subunit comprise three long α-helices, multiple α-helical segments, extended twisted β-sheets, membrane-spanning α-helix, and an intracellular cysteine-rich segment. This protein is heavily glycosylated because it comprises 21 to 35 N-glycosylation domains. Between the S1 and S2 subunits in SARS-CoV-2 presents furin (Golgi-resident host protease) cleavage site (Ou et al. 2020). In the S2 subunit, a second proteolytic cleavage site S2, upstream of the 2-fusion peptide is present.

The spike protein of coronaviruses is incorporated into the viral envelope and facilitates viral entry into target cells (Kumar. 2020). S1 identifies and attach to host receptors while synthesis between both the viral envelope as well as the host cell membrane is aided by corresponding conformational deviations in S2. Membrane fusion depends on S protein cleavage by host cell proteases at the S1/S2 site, which results in S protein activation. Cleavage of the S protein can occur in the constitutive secretory pathway of infected cells or during viral entry into target cells and is essential for viral infectivity (Böttcher-Friebertshäuser, Garten, and Klenk 2018). The receptor binding and membrane fusion responsibilities of S protein imply that it is S protein-dependent vaccines may stimulate antibody to inhibit virus attachment and fusion or incapacitate viral infections. S protein is an essential antigenic portion of all SARS-CoV-2 structural proteins that have been responsible for triggering the host immune system, neutralizing antibodies, and/or maintaining immunity against viral infections. A study has reported that while studying the surface of S protein, they noticed that certain mutations emerged, creating an especially close "ridge" via X-ray crystallography upon this S protein. This ridge is smaller than the SARS virus, which might be one of the factors why SARS-CoV-2 has become so contagious. S protein mutation arises in a conservative receptor-binding domain (RBD) domain that is specifically engaged in the attachment of the host receptor. The mutation is the foundation of virus evolution whereby changes of amino acids could affect protein structure and protein function. Therefore, the difference arises from the original protein, which eventually causes the virus to become more susceptible to vaccination and difficult to treat. Mutation of S protein encourages the virus to benefit from the immune response and to obtain a strategic transmission gain (Heald-Sargent and Gallagher 2012).

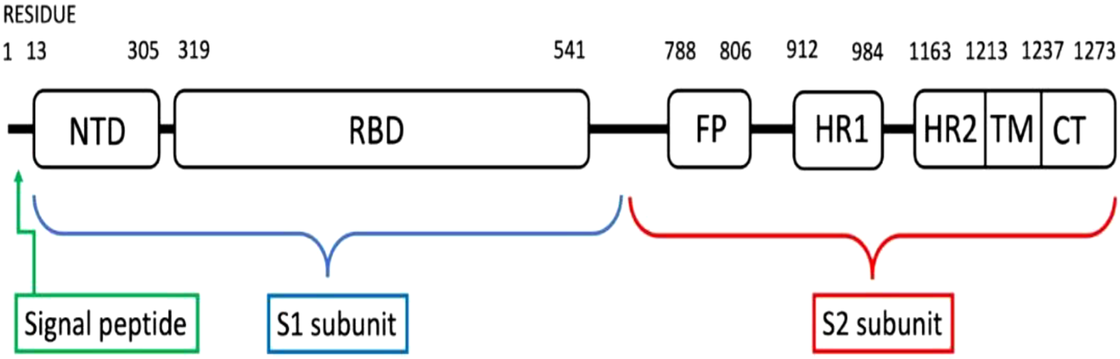


Figure 2.: A schematic diagram of the SARS-CoV-2 spike protein. It is comprised of the S1 subunit and the S2 subunit shown by the residue and the length.

SARS-CoV-2 transmission from human to human occurs mainly through bodily contact or through bodily fluid produced by an infected individual when sneezing and coughing (Rothan and Byrareddy 2020). The COVID-19's signs occur after 5.2 days of incubation. Most typical symptoms of COVID-19 disease include fever, coughing, and exhaustion, although other indications involve sputum, nausea, haemoptysis, diarrhea, shortness of breath, and lymphopenia. COVID-19 is suspected to invade lung alveolar epithelium as an input channel using the angiotensin-converting enzyme II receptor-mediated endocytosis (Velavan and Meyer 2020). Patients hospitalized with COVID-19 reported an increased leukocyte quantity, abnormal breathing patterns, and higher plasma amounts of pro-inflammatory cytokines Kumar. 2020).

Studies have found a variation between countries throughout the scenario of mortality rates, potentially due to diverse demographic makeup and the type of measures being taken to constrain viral spread in various countries (Dowd et al. 2020). Compared to a reference genome obtained on January 5, 2020, SARS-CoV-2 has experienced over 10,000 documented single mutations (Wang et al. 2020). In general, owing to the unavailability of exonuclease proofreading action of the virus-encoded RNA polymerases, RNA viruses are susceptible to random mutations (Ferron et al. 2017). Nidoviruses, along with coronaviruses, have an enzyme to excrete RNA polymerase-inserted improper mutagenic nucleotides and therefore preserve a high recognition accuracy in the transcription and replication of viruses (Chen et al. 2020).

Intervention by the human immune system triggers viral mutations. Global spread and transmission of COVID-19 offer significant opportunities for the virus to pick unique but advantageous mutations naturally (Chen et al. 2020). However, to classify clinical approaches at the local level, it seems feasible to define specific regional distributions of virus variants as more complete sequences become accessible. Therefore, it is likely that functional genomic research connected to epidemiological data from different countries may provide further perspectives into the pathogenicity and infectivity of this virus (Ceraolo and Giorgi 2020).

In the last few months, in SARS-CoV-2 spike protein shifts in amino acids between aspartic acid (D) and glycine (G) have occurred at position 614 as well as serine (S) and asparagine (N) at position 477, in comparison to the reference genome of Wuhan (Wuhan-Hu-1). Some mutations can trigger changes in both the structural and sequence levels. Prior studies have shown that such mutations have been linked to higher disease transmission and viral strains, human cell transduction, risk pathogenesis, and instances of deaths (McAuley et al. 2020).

Viruses develop by repeatedly mutating their gene sequences within and outside the receptor-binding domain (RBD) of spike protein sites that make it possible to enter human host cells, and the frequency of mutations is often due to the lack of re-reading by virus assembly machines. The SARS-CoV-2 virus had proof-reading capabilities. Yet mutations at a slow rate are emerging. When developing drug/vaccine candidates, the link between mutations and viral protein/function appears to be important (Begum, Banerjee, and Ray 2020). There has been high heterogeneity in the genome owing to such mutation, and thus posing a barrier for scientists to find an appropriate drug or vaccination (Kumar et al. 2020). Numerous current researches concentrate on vaccine design, drug repurposing, discovering the pathogenicity of the virus. Spike protein mutation analysis is used to understand the virus' virulence, its modes of antibody escape, including its cellular virus (Ashwaq, Manickavasagam, and Haque 2020).

The mutation has steadily emerged and there is a dynamic interplay between amino acids that can provide the virus immune resistance and the functional landscape of the unique variant in which they occur. Human ACE2 is now established as a receptor for the SARS-CoV-2 spike protein (F. Wu et al. 2020). A previous study reported main mutations found including D614G across the geographic distribution which increases the infectivity and transmissibility into the ACE2 receptor. However, certain studies claimed that 614 positions at the S1 domain of spike protein outside of RBD unlikely to directly affect the ACE2 binding activity since it is being released before S2 proteolytic processing which activates the S2 domain (fusion domain) (Ogawa et al. 2020). Also, the role of mutation of S477N and N501Y lying within the RBD of Spike (aa 319-541) need to studied further. Moreover, to this date, there has been no literature documentation on the physicochemical properties of mutated spike protein sequences specifically in terms of stability, amino acid composition, and point mutation analysis in terms of stability changes upon mutation, effects of missense variant, and disease propensity as well as secondary protein structure changes of the potential mutational effect in the transmission of the disease. Hence, there is an urgent need to develop an effective vaccine against SARS-CoV-2, as well as antibody-based therapeutics through a clear understanding of characterization of mutations as well as its relation with infectivity and transmissibility by studying specifically on RBD and protein-protein interaction. This research aims to analyze the mutational effect of spike protein in SARS-CoV-2 by identifying the occurrence rate of spike protein mutations from different geographical locations, to identify distinct mutations from different geographical locations, to predict how mutations affect the physiochemical, sequence, structural stability of spike proteins. Also, to study how mutations have an impact on human ACE2 protein interaction.

2. METHODOLOGY

2.1 Retrieval of the SARS-CoV-2 Spike Protein sequences

Seventeen thousand two hundred and twenty-eight sequence of the SARS CoV-2 spike protein sequences were obtained in the FASTA format (accessed 2nd October 2020) from the NCBI virus database, available at https://www.ncbi.nlm.nih.gov/labs/virus/vssi/ (Brister et al. 2015). The parameters are set such as only complete protein sequences, surface glycoprotein, a human host. The sequences were obtained from all the geographic locations including Asia (1183), South America (123), North America (11100), Africa (278), Europe (436), and Oceania (4107). SARS-CoV-2 spike protein sequence from Wuhan-Hu-1, China (NCBI accession number: QHD43416.1) (F. Wu et al. 2020) which has been collected in December 2019 was used as a reference sequence to examine the mutations.

The Asia group comprises genome obtained from patients located in Bahrain, Bangladesh, China, Georgia, Hong Kong, India, Iran, Iraq, Israel, Japan, Jordan, Kazakhstan, Lebanon, Malaysia, Nepal, Pakistan, Philippines, Saudi Arabia, South Korea, Sri Lanka, Taiwan, Thailand, Timor-Leste, Vietnam, and West Bank. The Europe group comprises genomes from Poland, Finland, Czech Public, France, Germany, Greece, Italy, Malta, Netherlands, Russia, Serbia, Spain, Sweden, and United Kingdom whereas the Africa group comprises genomes obtained from patients located in Egypt, Ghana, Morocco, Nigeria, South Africa, Tunisia, and Zambia. North America includes every genome obtained from patients located in Belize, Guatemala, Jamaica, Mexico, Puerto Rico, and the USA. The Oceania group comprises genomes from Australia’s patients, Guam and New Zealand’s patient. Finally, the South America group contains the genome from Brazil, Chile, Colombia, Peru, Uruguay, and Venezuela.

2.2. Prediction of mutational effects

Mutational effects from all geographic regions were performed by using the Sequence Editor, Database, and Analysis Platform (SSE) version 1.4 tool. SSE provides an integrated environment where sequences can be aligned, annotated, classified, and directly analyzed by several built-in bioinformatics programs. SSE incorporates a sequence editor for the creation of multiple sequence alignments. The sequences from geographical locations were aligned with spike protein sequence from Wuhan-Hu-1, China (NCBI accession number: QHD43416.1) by using CLUSTAL multiple alignments available at SSE tool. To recognize shared patterns between functionally or structurally associated genes, multiple sequence alignment is important for analyzing highly associated genes or proteins (Simmonds 2012).

2.3. Identification of a distinct mutation

The Venn diagram has been used to evaluate the number of distinct mutational effects of genomes derived from geographical regions worldwide using one of the Bioinformatics tools which are InteractiVenn (http://www.interactivenn.net/) (Heberle et al. 2015). From the previous approach, more than 30 mutation occurrences were chosen to overlap with each other in each geographical location. Lists comprising accession numbers displayed only a single number of accessions on each line. Input lists are filtered and rendered non-redundant and two major data, including the distinct and intersection components, are generated. The distinct mutation site distribution was analyzed and the collection date of each distinct mutation was obtained to determine the likelihood of the first mutation worldwide.

2.4. Analysis of physicochemical parameter

The mutated genome sequences were analyzed according to their presence within different regions of the spike protein sequences and further analyses were carried out using the software suite of programs developed ExPASy ProtParam web server. The parameters computed by ProtParam include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index, and grand average of hydropathicity (GRAVY). Molecular weight and theoretical pI are calculated as in Compute pI/Mw. The amino acid and atomic compositions are self-explanatory (Gasteiger et al. 2003).

2.5 Analysis of sequence and structural stability

Mutations across spike proteins worldwide focusing on more than 20 occurrences of mutations were analysed on their functional alterations in stability, transmission, and adaptability using various Bioinformatics tools.

2.6 Prediction of protein stability changes upon point mutation

I-Mutant 3.0 can predict to which extent a mutation in a protein sequence will or will not affect the stability of the folded protein. This tool dependent on a support vector machine for automatic prediction of protein stability. This tool was chosen for its accuracy of more than 89% computing the stability of free energy changes ($\Delta\Delta G$ value). Positive free energy values denoted increasing stability while negative free energy values denoted decreasing stability. Estimation (RI) reliability index score determined with the DSSP software. This happens when, beginning from the protein sequence, the indication of the stability switch is calculated from the DDG value of the Gibbs free energy. The index makes it possible to filter out more accurate predictions from the subset (Capriotti, Fariselli, and Casadio 2005).

2.7 Effects of amino acid substitution on protein function

The web-server Sorting Intolerant from Tolerant (SIFT) can assess the degree to which a mutation (amino acid substitution) in a query protein can influence protein expression. The SIFT algorithm is capable of promoting human genetic research based on infectious diseases, primarily because the variant causing disease is usually rare and appears to occur in peculiar coding regions, which comprise just 1% of the total genome, which consists of millions of variants of single nucleotides.

SIFT examines the structure of the amino acids and determines the ranking. The SIFT score is the standardized probability of discovering the new amino acid at that location, ranging from 0 to 1. A value between 0 and 0.05 is believed to impact protein function. SIFT scores are highly sensitive, high coverage, and balanced. The basic principle is that highly conserved regions appear to be much less tolerant of mutations, and therefore amino acid substitutions or insertions/deletions in these domains are much more likely to influence function (Sim et al. 2012).

2.8 Effects of a variant of interest in a host

VarSite is a web server map that recognized disease-associated variants from UniProt and ClinVar, along with common variants from gnomAD, to protein three-dimensional structure in the Protein Data Bank. UniProt comprises disease variants that influence human proteins. This research is mainly image-based and offers both a summary of each human protein and a summary of any particular variant of concern. Information is useful in determining whether the variant may be pathogenic or benign. The disease propensity is often a straightforward, standardized ratio of several diseases to natural variants of the form in amino acid changes. The greatest risk is 3.27 contributed by amino acid changes from C (Cysteine) to R (Arginine) said to be observed over 3 times in disease data than in natural data while the minimal disease risk predicted by I (Isoleucine) to V (Valine) is 0.25. A score of more than 1 is denoted that variant of interest is more disease prone (Laskowski et al. 2020).

2.9 Prediction of secondary structural changes in spike protein

The very first existing approaches established by Chou and Fasman, Lim, and Garnier had a reliability of ~60 percent. Structural knowledge could provide transparency into protein function and as a result, a high-accuracy protein structure prediction from its sequence is hugely valuable. In the output of GOR V, a-helix is expressed by letter H, ß-sheet by E, and coil by C. The GOR (Garnier–Osguthorpe–Robson) program includes both information theory and Bayesian statistics to assess the secondary protein structure. The concept behind the integration of multiple sequence alignments through GOR is to improve the knowledge for enhanced differentiation between secondary structures (Garnier, Gibrat, and Robson 1996; Sen et al. 2005).

2.10 Protein-protein docking analysis

The purpose of this analysis is to emphasize mutations that occur precisely at RBD locations. The highest mutations were chosen for the study for their implications on tertiary protein stability as well as for further interaction between mutant spike protein and wild type human ACE2 protein.

2.11 Prediction of Receptor Binding Domain (RBD) mutational

A compilation of all RBD associated mutations was obtained from all geographic regions and assessed for more than 30 occurrences in each geographical region. RBD is situated between 319aa and 541aa. RBD seemed to be the target of drug discovery and vaccine development known for its role in receptor binding. For our study, more than 30 RBD mutation occurrences were reported.

2.12 Prediction of RBD mutation structural impact on spike protein

Only highly contributed RBD mutations were selected for this structural impact on spike protein by using the PremPS tool. This tool is well known for its specialty in predicting the effects of single mutations on protein stability by calculating the changes in unfolding Gibbs free energy. PremPS which requires a novel score function consisting of just ten attributes and operates on a fair comparison of five thousand mutations, half of them destabilizing mutations and the other half are stabilizing mutations whereby decreasing stability, $\Delta\Delta G$ <0 and increasing stability, $\Delta\Delta G$> 0. Two measurements of the Pearson correlation coefficient (R) and root-mean-square error (RMSE) were used to check the agreement among experimental and expected values of unfolding free energy variations. Besides, the location of mutations can be predicted with a range of scores measured by this tool. If the residue submerged in the polypeptide chains, it is said to be that the proportion of a solvent available surface area of this residue in the protein and the stretched tripeptide is less than 0.2, elsewhere positioned on the surface of the protein. The PremPS produces the involvement of each variable in the target function with each mutation and offers a 3d virtual display that shows the non-covalent relationships between the mutated site and its neighboring residues (Yuting et al. 2020).

2.13 Molecular docking of RBD mutational spike protein with ACE2 human protein

Hex is an interactive molecular graphics program for calculating and displaying feasible docking modes of pairs of protein. It is also the only program for dock and superposition using spherical polar polarity. Fourier (SPF) distinctions to speed up measurements, and it is still one of the few molecular dockings to present the status with built-in visuals.

Docking helps the researcher to practically monitor a compound repository and determine the best binders focusing on various scoring functions. It examines ways in which two proteins like ACE-2 enzyme receptor and mutational spike protein-ligand work together and dock well with each other. Their relative stability was assessed using simulations of free energy including molecular dynamics and respective binding affinities. As the variable, Orthogonal Projections to Latent Structures (OPLS) energies were preferred and set to default because of robust statistical modeling techniques and accurate measurement of free energy mixing (Ritchie and Orpailleur 2013).

2.1.4 Comparison of spike mutation between human and mink host

The mink-derived SARS-CoV-2 isolates of Netherland and Denmark which are available publically at NCBI were retreived. There were 13 isolates from Netherland (QJS39496, QJS39507, QJS39519, QJS39531, QJS39543, QJS39555, QJS39567, QJS39579, QJS39591, QJS39603, QJS39615, QJS39627, and QJF11995) and 12 from Denmark (QNJ45106, QNJ45118, QNJ45130, QNJ45142, QNJ45154, QNJ45166,QNJ45178, QNJ45190, QNJ45202, QNJ45214, QNJ45226 and QNJ45238) were compared with spike protein sequence, human-derived SARS-CoV-2, Wuhan-Hu-1(QHD43416.1) through multiple alignment using clustal at default parameters.

3. RESULTS

3.1 SARS-CoV-2 Spike protein mutation from different geographical regions

i) **South America:** Hundred twenty-three SARS-CoV-2 mutational spike protein sequences were aligned and compared to the Wuhan reference sequence (Wuhan-Hu-1). According to the figure below, the highest mutation ranked by D614G in South America country **(Supplementary table S1)**. The graph below was plotted to interpret the raw table produced by mutational effects of sequences from all continents present in South America **(Figure 3)**.

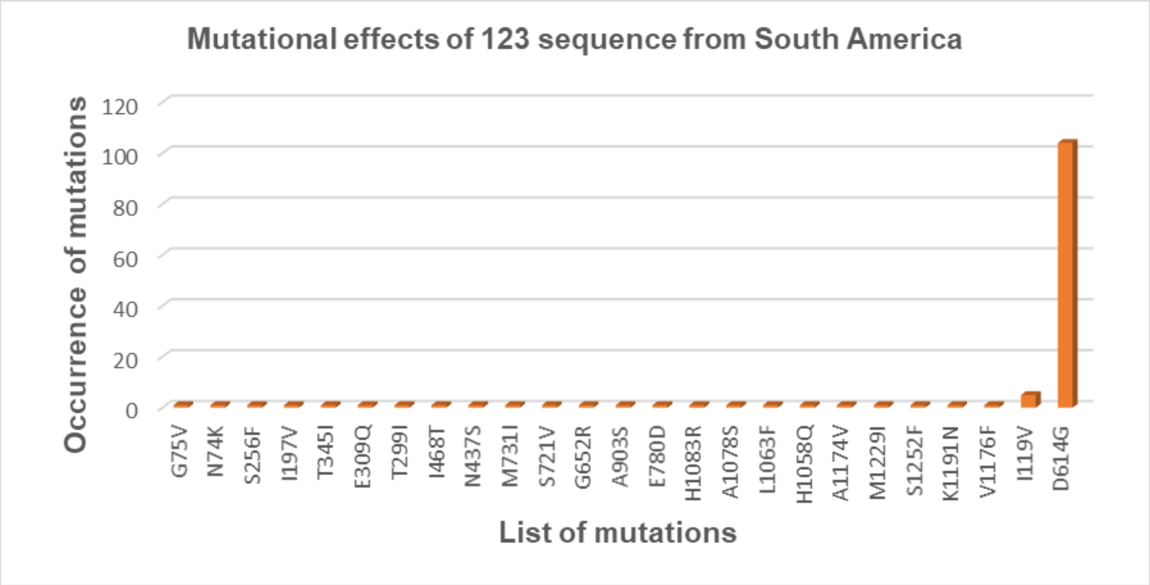


Figure 3: South America mutational effects. D614G mutation showing the highest number of occurrences compared to other mutations in South America. Diagram obtained through mutational effects analysis.

ii) **Africa:** Two hundred seventy-eight sequences of SARS-CoV-2 mutational spike protein were aligned and compared to the Wuhan reference sequence (Wuhan-Hu-1). According to the figure below, the highest mutation ranked by D614G in Africa country **(Supplementary table S2)**. The graph below was plotted to interpret the raw table produced by mutational effects of sequences from all continents present in Africa **(Figure 4)**.

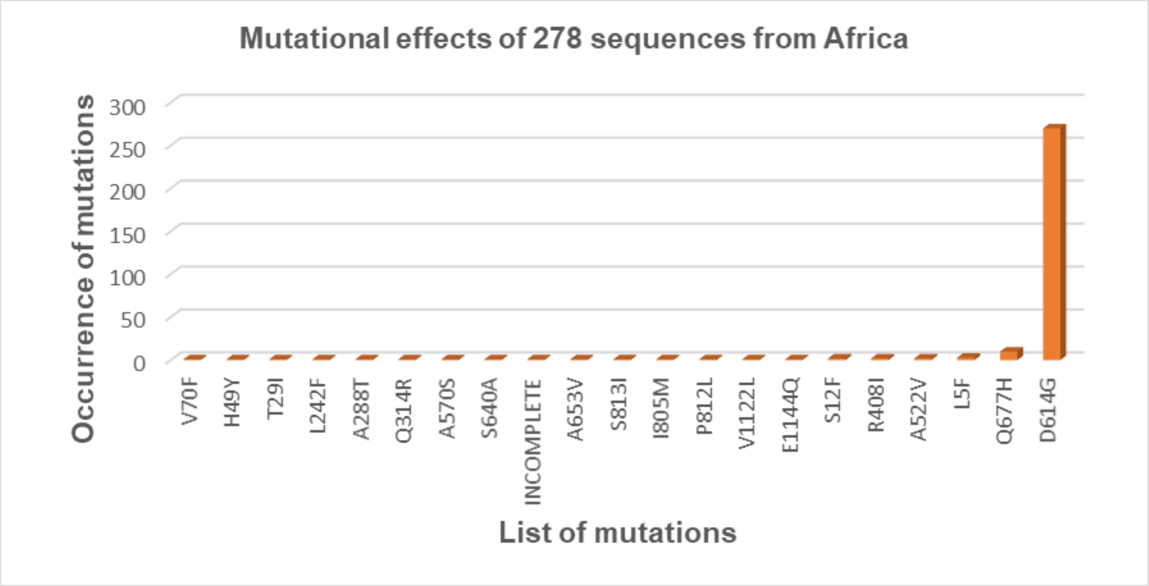


Figure 4: Africa mutational effects. D614G mutation showing the highest number of occurrences compared to other mutations in Africa. Diagram obtained through mutational effects analysis.

iii) **Europe:** Four hundred thirty-six sequences of SARS-CoV-2 mutational spike protein were aligned and compared to the Wuhan reference sequence (Wuhan-Hu-1). According to the figure below, the highest mutation ranked by D614G in Europe country **(Supplementary table S3)**. The graph below was plotted to interpret the raw table produced by mutational effects of sequences from all continents present in Europe **(Figure 5)**.

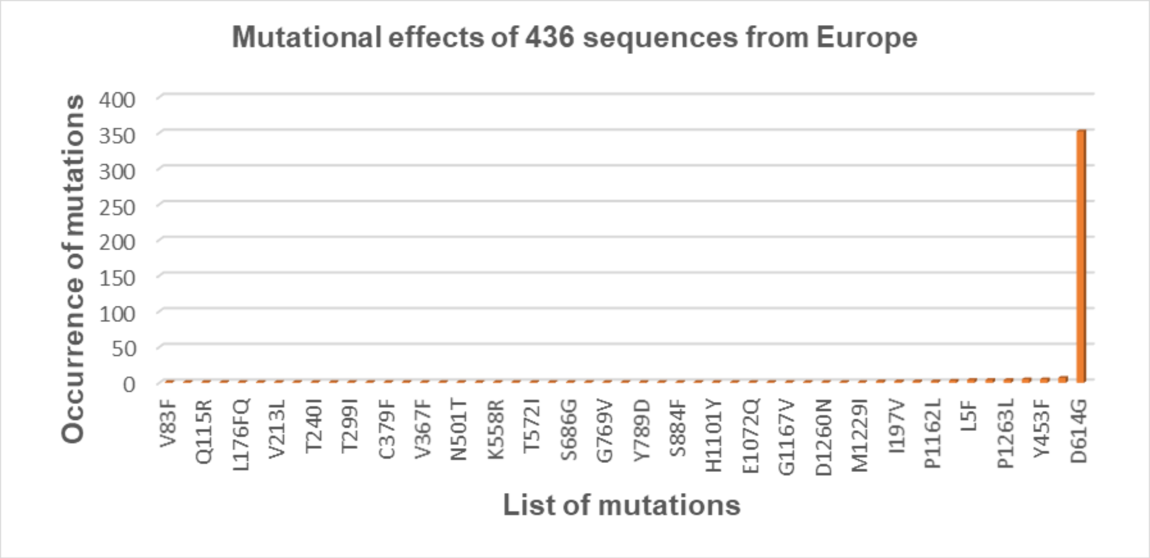


Figure 5: Europe mutational effects. D614G mutation showing the highest number of occurrences compared to other mutations in Europe. Diagram obtained through mutational effects analysis.

iv) **Asia:** Thousand one hundred eighty-three sequences of SARS-CoV-2 mutational spike protein were aligned and compared to the Wuhan reference sequence (Wuhan-Hu-1). According to the figure below, the highest mutation ranked by D614G in Asia countries **(Supplementary table S4)**. The graph below was plotted to interpret the raw table produced by mutational effects of sequences from all continents present in Asia **(Figure 6)**.

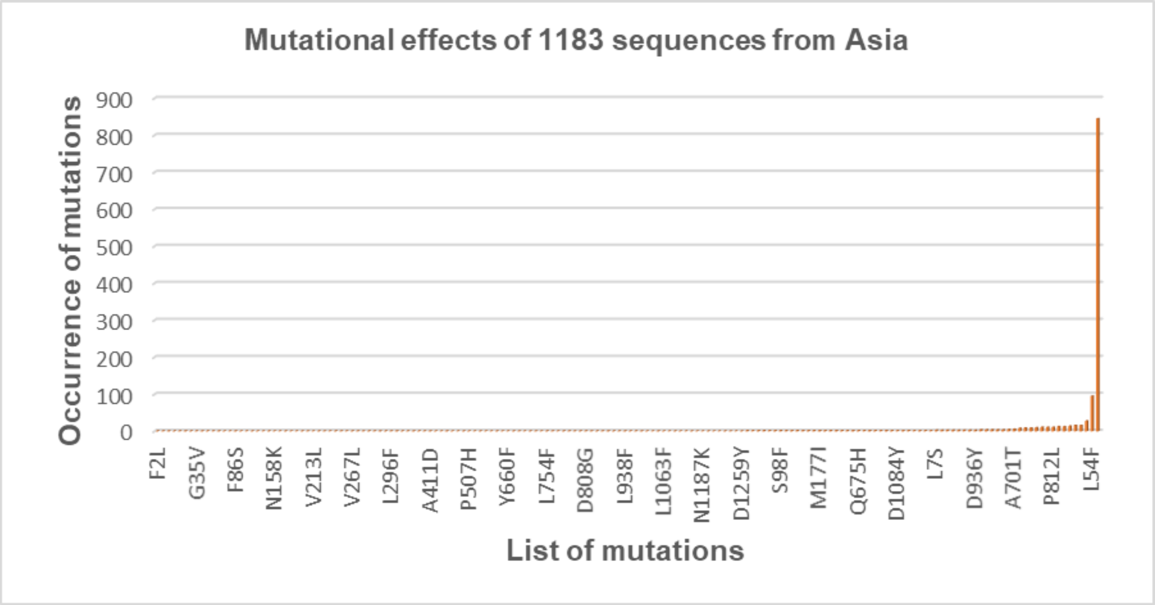


Figure 6: Asia mutational effects. D614G mutation showing the highest number of occurrences compared to other mutations in Asia. Diagram obtained through mutational effects analysis.

v) **Oceania:** Four thousand one hundred seven sequences of SARS-CoV-2 mutational spike protein were aligned and compared to the Wuhan reference sequence (Wuhan-Hu-1). According to the figure below, the highest mutation ranked by D614G and major contribution of S477N in Oceania country **(Supplementary table S5)**. The graph below was plotted to interpret the raw table produced by mutational effects of sequences from all continents present in Oceania **(Figure 7)**.

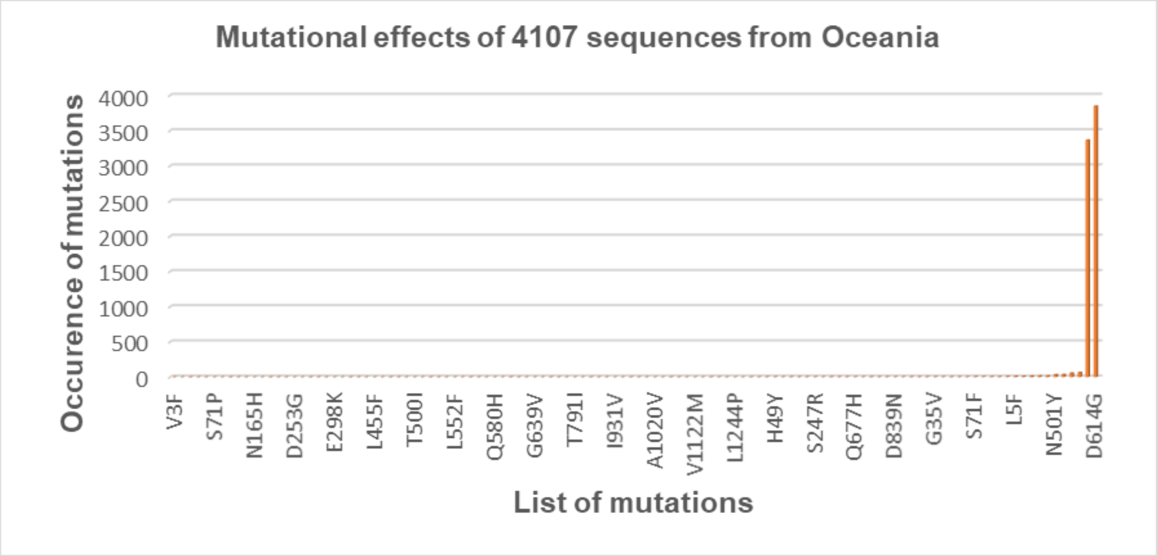


Figure 7: Oceania mutational effects. D614G mutation showing the highest number of occurrences compared to other mutations in Oceania. Diagram obtained through mutational effects analysis.

vi) **North America:** Eleven thousand one hundred seven sequences of SARS-CoV-2 mutational spike protein were aligned and compared to the Wuhan reference sequence (Wuhan-Hu-1). According to the figure below, the highest mutation ranked by D614G in North America country **(Supplementary table S6)**. The graph below was plotted to interpret the raw table produced by mutational effects of sequences from all continents present in North America **(Figure 8)**.

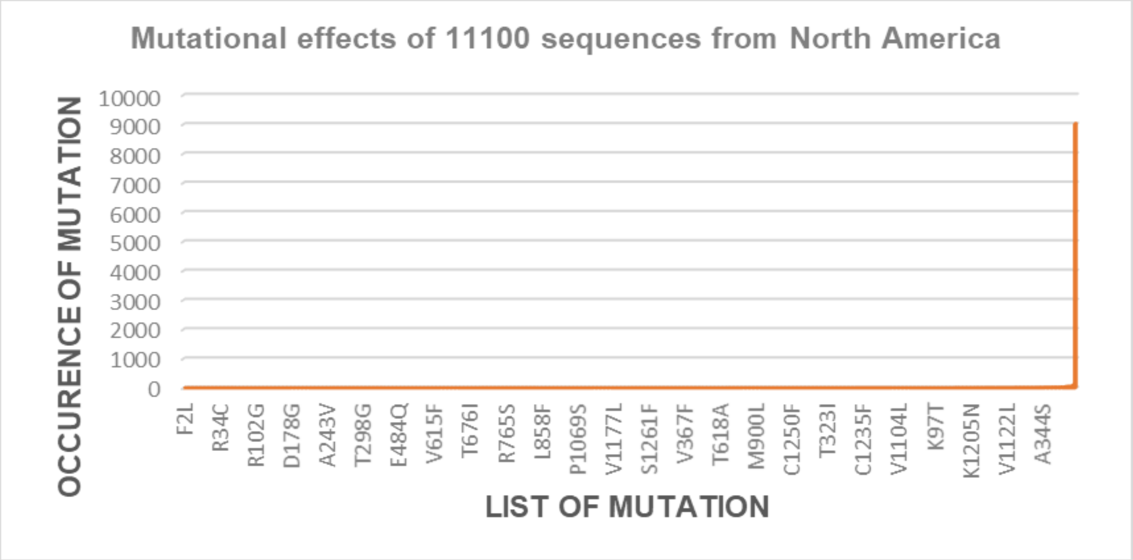


Figure 8: North America mutational effects. D614G mutation showing the highest number of occurrence compared to other mutations in North America. Diagram obtained through mutational effects analysis.

Thirteen mutations occurred substantially higher due to random amino acid changes in the SARS-CoV-2 spike protein, which may result in increased transmission worldwide. **Table 1** indicates that all mutations occurred more than 30 resulting from all geographic locations.

Table 1: The number of mutations occurring in each geographical region more than 30 frequency of occurrence. S477N and D614G have the highest prevalence in Oceania and North America compared to other mutations worldwide with more than 1,000 mutation sequences deposited.

List of mutations	Geographic location	Number of occurrences
N501Y	OCEANIA	34
T778I	OCEANIA	37
P681L	NORTH AMERICA	49
E780Q	NORTH AMERICA	53
A845D	NORTH AMERICA	57
G1124V	OCEANIA	58
P1263L	NORTH AMERICA	59
T632N	OCEANIA	64
L54F	ASIA	95
D614G	SOUTH AMERICA	104
L5F	NORTH AMERICA	115
D614G	AFRICA	270
D614G	EUROPE	352
D614G	ASIA	845
S477N	OCEANIA	3366
D614G	OCEANIA	3847
D614G	NORTH AMERICA	9023

3.2 SARS-CoV-2 spike protein distinct mutation from different geographical regions

The number of distinct mutations is important for identifying the origin of the mutation. A distinct mutation for mutational spike protein sequences that occurred more than 30 times were obtained from the InteractiVenn web server. The distinct mutation consists of only one distinct sequence which is important for this analysis. **Table 2** reveals that most mutations in the world contain a distinct mutation sequence.

Table 2: Distinct mutation. One distinct mutation sequence was chosen from a large number of distinct mutations to represent each mutation in the world, but the observation indicates that certain mutations have more than 1 mutation (shown as -)

List of mutations	Geographic location	Number of occurrences	Number of distinct mutation (all geogrpahical locations)	Selected distinct mutated sequence
N501Y	OCEANIA	34	1	QNO58931.1
T778I	OCEANIA	37	1	QNO81779.1
P681L	NORTH AMERICA	49	-	-

E780Q	NORTH AMERICA	53	-	-
A845D	NORTH AMERICA	57	-	-
G1124V	OCEANIA	58	9	QKV37476.1
P1263L	NORTH AMERICA	59	-	-
T632N	OCEANIA	64	1	QNP04771.1
L54F	ASIA	95	9	QLR07160.1
D614G	SOUTH AMERICA	104	84	QNH88895.1
L5F	NORTH AMERICA	115	34	QJX45019.1
D614G	AFRICA	270	238	QNM81526.1
D614G	EUROPE	352	310	QJW69295.1
D614G	ASIA	845	541	QKK14635.1
S477N	OCEANIA	3366	-	-
D614G	OCEANIA	3847	361	QJR94581.1
D614G	NORTH AMERICA	9023	7666	QMT91816.1

The date of collection for each distinct mutation sequence obtained from the NCBI Virus database compared to that the wild type of Wuhan reference sequence (Wuhan-Hu-1) was collected in December 2019, in addition to the mutation in spike protein sequences, ranges from February 2020 to August 2020. **Table 3** reveals that the first mutation occurred in February 2020.

Table 3: Date of collection of distinct mutation sequence. (For example, Sequence QJW69295.1 represents the D614G mutation from the Europe country collected in February 2020 that reveals the origin of the mutation from the EUROPE: Germany: Bavaria and QNO58931.1 represents N501Y mutation from OCEANIA: Australia: Victoria collected in July 2020 reveals the origin of mutation from OCEANIA)

Accession number	Collection date	Mutation	Geo Location
QHD43416.1	2019-12	Reference	ASIA: China
QJW69295.1	2020-02	D614G	EUROPE: Germany: Bavaria
QKV37476.1	2020-03-23	G1124V	OCEANIA: Australia: Victoria
QJX45019.1	2020-03-24	L5F	NORTH AMERICA: USA: CA
QJR94581.1	2020-04-03	D614G	OCEANIA: Australia: Victoria
QNH88895.1	2020-04-10	D614G	South America : Venezuela
QKK14635.1	2020-05-23	D614G	ASIA: Bangladesh
QMT91816.1	2020-06-01	D614G	NORTH AMERICA:

			USA: Washington, Yakima County
QLR07160.1	2020-06-15	L54F	ASIA: India: Vadodara
QNO58931.1	2020-07-01	N501Y	OCEANIA: Australia: Victoria
QNO81779.1	2020-07-19	T778I	OCEANIA: Australia: Victoria
QNP04771.1	2020-07-22	T632N	OCEANIA: Australia: Victoria
QNM81526.1	2020-08-17	D614G	Africa: Egypt

3.3 Physiochemical characterization of distinct spike protein mutation

The table below was tabulated with the completion of the study of physicochemical parameters obtained from the ExPASy ProtParam tool. Significant changes in molecular weight, decrease in instability index, but all proteins are stable and no changes in the aliphatic index are detected, but increasing GRAVY shows an increase in hydropathicity. **Table 4** shows various physicochemical parameters of the mutational spike protein.

Table 4: Physicochemical parameters of each mutation. Drastic changes due to mutation have been observed in all primary structures of the mutational spike protein.

Parameters	Molecular weight	Theoretical pI	Instability index	Aliphatic index	GRAVY	Total number of negatively charged residues (Asp + Glu)	Total number of positively charged residues (Arg + Lys)	Formula
QHD43416.1 (Wuhan-Hu-1 reference)	141178.47	6.24	33.01 Stable protein	84.67	-0.079	110	103	C6336H9770N1656O1894S54
QNP05659.1 (Oceania-G1124V)	141189.54	6.32	32.79 Stable protein	84.90	-0.075	109	103	C6338H9775N1657O1892S54
QNA40386.1 (North America-A845D)	141164.44	6.24	32.86 Stable protein	84.60	-0.081	110	103	C6335H9768N1656O1894S54
QMI91946.1 (North America-P1263L)	141160.59	6.32	32.49 Stable protein	85.59	-0.064	109	103	C6339H9780N1656O1890S54

QNO83159.1 (Oceania-T632N)	141160.46	6.32	32.76 Stable protein	84.67	-0.081	109	103	C6335H9768N1658O1892S54
QNO81779.1 (Oceania-T778I)	141159.51	6.32	33.20 Stable protein	84.98	-0.075	109	103	C6337H9773N1657O1891S54
QNN26420.1 (Asia-L54F)	141154.45	6.32	32.80 Stable protein	84.37	-0.078	109	103	C6337H9766N1656O1892S54
QNS00745.1 (North America-L5F)	141154.45	6.32	32.86 Stable protein	84.37	-0.078	110	103	C6337H9766N1656O1892S54
QNO58895.1 (Oceania-S477N)	141147.46	6.32	32.73 Stable protein	84.67	-0.079	109	103	C6335H9769N1657O1892S54
QNO58763.1 (Oceania-D614G)	141147.46	6.32	32.73 Stable protein	84.67	-0.079	109	103	C6335H9769N1657O1892S54
QMT93412.1 (North America-P681L)	141136.48	6.32	32.73 Stable protein	84.98	-0.073	109	103	C6335H9772N1656O1892S54
QNV49482.1 (South America-D614G)	141120.43	6.32	32.86 Stable protein	84.67	-0.077	109	103	C6334H9768N1656O1892S54
QNR61256.1 (Africa-D614G)	141120.43	6.32	32.86 Stable protein	84.67	-0.077	109	103	C6334H9768N1656O1892S54
QNT10048.1 (Europe-D614G)	141120.43	6.32	32.86 Stable protein	84.67	-0.077	109	103	C6334H9768N1656O1892S54
QKM77228.1 (Asia-D614G)	141120.43	6.32	32.86 Stable protein	84.67	-0.077	109	103	C6334H9768N1656O1892S54
QOE45083.1 (North America-D614G)	141120.43	6.32	32.86 Stable protein	84.67	-0.077	109	103	C6334H9768N1656O1892S54
QNV69966.1 (North America-E780Q)	141119.45	6.41	32.80 Stable protein	84.67	-0.077	108	103	C6334H9769N1657O1891S54
QNO95711.1 (Oceania-N501Y)	141114.42	6.24	33.03 Stable protein	84.67	-0.072	109	102	C6337H9766N1652O1893S54

The table below is tabulated according to the completion of the sequence and structural stability shifts produced by I-Mutant, SIFT, and Varsite servers. The majority of mutations that have occurred have reduced protein stability, tolerated protein function and several mutations are not disease prone compared to D614G and N501Y. **Table 5** indicates the substantial stability, protein function, and disease propensity of SARS-CoV-2 spike protein mutations.

Table 5: Changes in stability, function, and disease propensity of each mutation. D614G mutation indicates that it reduces stability, tolerates protein function, and induces more disease, but S477N is less disease prone

AA change	I-Mutant	SIFT	Varsite -disease propensity (>1 more disease prone)
S477N	Decrease stability (-1.53)	TOLERATED (0.84)	0.46
E780Q	Decrease stability (-0.18)	TOLERATED (0.29)	0.59
L54F	Increase stability (0.14)	TOLERATED (0.47)	0.64
L5F	Decrease stability (-0.08)	TOLERATED (0.10)	0.64
T778I	Increase stability (0.50)	TOLERATED (0.77)	0.66
T632N	Decrease stability (-0.54)	TOLERATED (0.49)	0.72
P681L	Increase stability (0.17)	TOLERATED (0.37)	0.95
P1263L	Decrease stability (-0.16)	AFFECT PROTEIN FUNCTION (0.04)	0.95
D614G	Decrease stability (-2.41)	TOLERATED (0.62)	1
N501Y	Decrease stability (-0.49)	TOLERATED (0.09)	1.3
A845D	Decrease stability (-2.14)	TOLERATED (0.07)	1.84
G1124V	Decrease stability (-0.70)	TOLERATED (0.10)	1.86

Secondary structure predicted and compared to Wuhan sequence (Wuhan-Hu-1) in the discovery of secondary structural variations. Figure below from the GOR4 server. **Figure 9** indicates that there are no major changes observed in the mutated secondary protein structure which is composed random coil, D614(A), G614(B), S477(C), and N477 (D) and mutation of N501 (E), 501Y(F) changes from random coil to extended strand.

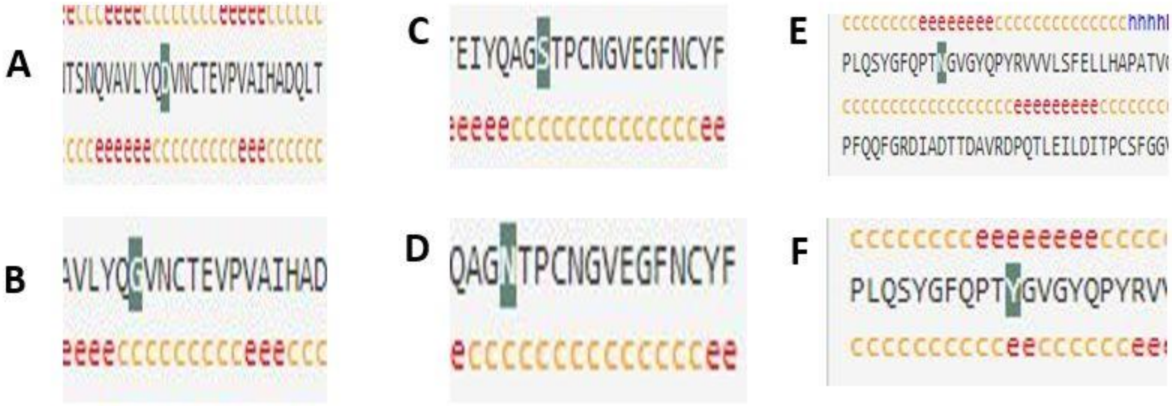

A
ITSNQVAVLYQDVNCTEVPVAIHADQLT
C
EIYQAGSTPCNGVEGFNCYF
E
PLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATV
PFQQFGRDIADTTDAVRDPQTLEILDITPCSFGG
B
AVLYQGVNCTEVPVAIHAD
D
QAGNTPCNGVEGFNCYF
F
PLQSYGFQPTYGVGYQPYRV

Figure 9: Comparison of secondary structure. (**A-B)** D to G at position 614 and **(C-D)** S to N at position 477 shows no changes in secondary structure while remaining in the coil region (C). **(E-F)** Secondary structure N to Y changes from coil structure to extended strand

3.4 Mutational analysis of spike protein of RBD

The receptor binding domain of spike protein is from 319-541 amino acids (aa). Mutations that arise at RBD of spike protein are very less however very significant in drug and vaccine discovery about the role carried by the RBD region in the transmission of this disease. **Table 6** reveals that Oceania ranks the highest RBD mutational followed by North America and Asia. The RBD mutation was analyzed throughout the geographical location. The mutation A522V has been found in all geographical locations, namely in Oceania, North America, Asia, and Africa. H519Q, S477N mutation was found in Oceania, North America, Asia. A475V, G446V, P521S are found in Oceania and North America. A520S mutation was found in North America and Asia.V367F was found in Europe and Asia.

Table 6. The RBD mutation was recorded in all geographical locations with several occurrences till October 2^{nd,} 2020.

Mutation	Present in	No.of occurences
A522V	Oceania, North America, Asia and Africa	32
H519Q	Oceania, North America, Asia	3
S477N	Oceania, North America, Asia	3424
A475V	Oceania and North America	2
G446V	Oceania and North America	2
P521S	Oceania and North America	1
A520S	North America and Asia.	3
V367F	Europe (host:mink) and Asia (Hong Kong) (host:human)	5
P330A	Oceania	3
K444N	Oceania	1

doi:10.20944/preprints202012.0710.v1

L455F	Oceania	1
I469T	Oceania	1
V483F	Oceania	1
G485R	Oceania	10
S494L	Oceania	1
T500I	Oceania	1
N501Y	Oceania	34
R408I	Africa	2
P337R	Asia	1
R346T	Asia	1
L368P	Asia	1
V382L	Asia	1
A411D	Asia	1
E471Q	Asia	1
C488R	Asia	1
P491L	Asia	1
Q506H	Asia	1
P507S	Asia	1
P507H	Asia	1
Y508N	Asia	2
L518I	Asia	3
A372V	Europe	1
C379F	Europe	1
V382E	Europe	1
T393P	Europe	1
Y453F	Europe (host:mink)	5
F486L	Europe (host:mink)	1
Q321L	North America	2
T323I	North America	3
P330S	North America	2
A344S	North America	18
T345S	North America	1
A348T	North America	1
A348S	North America	1
R357K	North America	1
F374L	North America	1
P384L	North America	2
V395I	North America	2
R403K	North America	10
V407I	North America	1
A411S	North America	6
G413R	North America	1
L441I	North America	1
F456Y	North America	1
R457K	North America	1
K458Q	North America	1
G476S	North America	8
S477N	North America	3

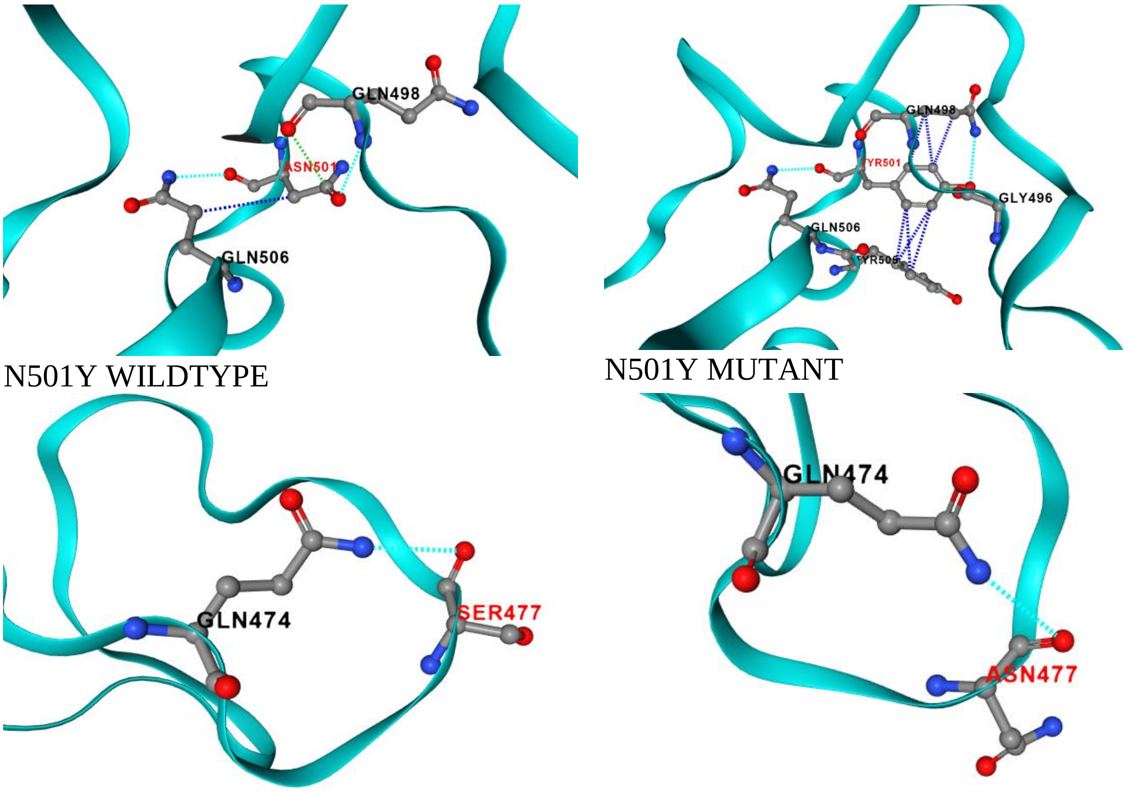
T478K	North America	1
P479L	North America	1
V483A	North America	26
E484Q	North America	1
Q493L	North America	1
S494P	North America	1
G503F	North America	1
Y505H	North America	2
Y508H	North America	1
A522S	North America	12
A522G	North America	3
K529N	North America	1
K529E	North America	1
T345I	South America	1
N437S	South America	1
I468T	South America	1

The table below is tabulated according to the completion of the RBD mutational effect analysis of the protein structure obtained from the PremPS web server. The ΔΔG value for each mutation is calculated and demonstrates a substantial comparison between the two where S477N was found to destabilize the protein structure whereas N501Y was found to stabilize the protein structure and to be found at a similar place. **Table 7** reveals the unfolding energy and location of the mutation in the RBD region.

Table.7: Changes in the structure of RBD mutational spike protein. S477N had a negative destabilization value and N501Y had a positive value showing the stabilization of the protein structure and both were located on the protein surface.

Mutation	ΔΔG (kcal mol^{-1}) Unfolding energy	Location
S477N	-0.14 (destabilizing)	SURFACE
N501Y	0.26 (stabilizing)	SURFACE

Mutations arise at RBD of spike protein associated with several other variants to display the stability of the protein structure. N501Y variant has a high association indicating more stability compared to S477N. **Figure 10** reveals the comparison between variant wild type with a mutant type in three- dimensional view.



N501Y WILDTYPE N501Y MUTANT

S477N WILDTYPE S447N MUTANT

Figure 10: Location and association of mutants. N501Y has more bonding to other amino acids, which enhances the stability of the spike protein, while S477N has less bonding to others and reduces stability.

The table below is tabulated according to the completion of the protein docking analysis obtained from Hex software. Binding energy was calculated by the dock with ACE2 human protein for each mutant spike protein and compared with binding energy obtained by Wuhan wild type (Wuhan-Hu-1) dock with ACE2 human protein. In the interaction analysis, **Table 8** indicated the highest and lowest binding energy which reveals the binding affinity towards the human host.

Table 8: Binding energy of RBD mutational spike protein. N501Y has the highest binding energy and thus has a greater affinity and is strongly bound to the human host for successful transmissibility and infectivity compared to S477N which has a lower binding affinity.

Mutation	Etotal
Wuhan (Wuhan-Hu-1) (wild)	-484.34
S477N (mutant)	-476.30
N501Y (mutant)	-496.97

3.5 Comparision of spike mutation between human and mink host

We also compared spike mutation between humans and mink. Based on the analysis D614G, V367F mutation was found in both mink and human host **(Table 9)**. Regarding N501T mutation happened at Mink and was also found in human N501Y. Other spike mutation in mink like Y453F, G261D, V367F, F486L, ΔN710, ΔA892, and ΔS943 were only identified in mink and not in humans (as of 2nd October 2020).

Table 9: Summary of amino acid mutations in the spike protein of mink-derived SARS-CoV-2 in Denmark and the Netherlands, based on publicly available data

Accession number	Collection date	Mutation	Host (No. of occurences)	Geo Location
QHD43416.1	2019-12	Reference	Human	ASIA: China
QJW69295.1	2020-02	D614G	Human (>9000)	EUROPE: Germany: Bavaria
QJS39496	2020-04-29	D614G	Mink: Mustela lutreola	Europe: Netherlands
QJS39591	2020-04-25	Y453F	Mink: Mustela lutreola (5)	Europe: Netherlands
QNO58931	2020-07-01	N501Y	Human (34)	Oceania:Australia: Victoria
QJS39507	2020-04-29	N501T	Mink: Mustela lutreola (1)	Europe: Netherlands
QJS39591	2020-04-25	G261D	Mink: Mustela lutreola (4)	Europe: Netherlands
QJS39627	2020-05-06	V367F	Mink: Mustela lutreola	Europe: Netherlands
QKF95522	2020-01-22	V367F	Human (4)	Asia: Hong Kong
QJS39567	2020-04-29	F486L	Mink: Mustela lutreola (1)	Europe: Netherlands
QNJ45106	2020-06-14	ΔN710	Mink: Neovison vison (2)	Europe:Denmark
QNJ45178	2020-06-14	ΔA892	Mink: Neovison vison (1)	Europe:Denmark
QNJ45226	2020-06-17	ΔS943	Mink: Neovison vison (1)	Europe:Denmark

4. DISCUSSION

4.1 Distribution of mutations

The result obtained in the present study suggests that sequences of patients suffering from SARS-CoV-2 which have been deposited in NCBI Virus associated with a large amount of mutational effect from all over the geographical region. Based on the result earlier, the total number of occurrences in conjunction with the list of mutations showed that a large number of occurrences were found to be contributed by two variants such as D614G and S477N. D614G observed in 9023 sequences from North America while in Oceania from 3847 sequences co-occurrences with S477N which was found in 3366 sequences. While the spike protein sequence had several mutations distributed at different residues, few mutations led to a significant increase (Begum et al. 2020). The frequency of mutations was higher due to RNA viruses are susceptible to random mutation. Mutation at 614 revealed that amino acid alteration from Aspartic acid to Glycine. Mutation at 477 revealed that amino acid alteration from Serine to Asparagine. This could also contribute to the development of an antibody evasion mutant if location 614 and 477 are representative of the immunogenic epitope. If this location is a component of the epitope, this mutation may allow the virus to escape the immune system and proliferate into a new, more evolved cluster. Currently, D614G and S477N, N501Y mutations have become a topic of debate and the functional importance of this mutation to transmissibility has not been identified. Also, several new mutations have been identified in the spike glycoprotein sequence of isolated regions that have not been detected in other countries. This new mutation has not been identified in any other country as per the sequences examined, thus indicating either a minor independent mutation that may have arisen after the outbreak (Begum et al. 2020).

4.2 Root of mutations

The earliest sequence carrying the first mutation of D614G is QJW69295.1 was found in February 2020 in Europe. Although the virus spread internationally at the beginning of 2020 until borders were sealed and viral strains existing across the world were transmitted, intercontinental travel stayed under control throughout the summer of 2020. The lack of intercontinental travel made it possible for continent-specific variants to evolve. However, travel around Europe continued in the summer of 2020. Here we report on the novel SARS-CoV-2 variant D614G, which appeared in the early summer of 2020, possibly in Germany, and spread widely to multiple countries in Europe. It has risen in pace over the period in many countries at the same time. As can be predicted from a much earlier sample date, the cluster increases its frequency in Germany, initially rising to about 60 percent occurrence within a month of the sequence being identified (Hodcroft et al. 2020). The earliest sequence of the first mutation of S477N is

4.3 Changes on the physicochemical parameter of spike protein upon mutation

The result obtained for the physicochemical parameter in the present study reveals that significant decrease observed in molecular weight and instability index however no changes in the aliphatic index but highly increase in hydropathicity. This prediction helps in understanding the exact conformational behavior of a protein, its physical and chemical properties and make a comparison between wild type and mutant spike protein sequence. The predicted molecular weight of Wuhan HU-1 (141178.47), S477N (141147.46), D614G (141120.43), N501Y (141114.42). D614G showed a molecular weight of 141120.43 whereby wild type showed 141178.47. This is because Glycine (G) is lighter compared to Asparagine (D) in correlation with changes upon mutation in a linear chain. Furthermore, molecular weight-related with instability index however all the mutant proteins are considered a stable proteins in primary structure level since all the score ranges lesser than 40. The aliphatic index remains as it is because of no changes observed in the aliphatic side chain occupied by amino acids including Alanine, Valine, Isoleucine, and Leucine. Hydropathicity is measured by adding the hydropathy value of amino acid residues and divide by the length of the sequence. Assessing the hydrophobic or hydrophilic character and topology of the protein is also critical. The GRAVY score and topology analysis were conducted for this purpose. It lies in the range from -2 to +2 where hydrophobic is a positive value and hydrophilic protein is indicated by a negative value. It is also an indicator of whether a protein would be observed on 2-D gels, as proteins with GRAVY scores > 0.4 are not in the solubility range and are therefore difficult to detect. The sequence of Wuhan-Hu-1 (-0.079), D614G(-0.077), S477N (-0.079), N501Y (-0.072) with a less negative value indicating a low hydrophobic nature and hence good solubility. An increasing positive score indicates a greater hydrophobicity. D614G and N501Y mutation is said to be increasing tremendously compared to wild type due to the presence of more hydrophobic residues which likely to increase pathogenicity of viral protein. Hydrophobic molecules represent non-polar and non-charged hence it does not dissolve in aqueous solution. More of these molecules increase the propensity of viral substance to a stronger adhesion (Basu et al. 2020).

4.4 Impact of D614G, S477N, and N501Y towards stability, function, and disease propensity

The result obtained for sequence and structural stability were obtained and tabulated according to score ranges which are unique and yet to be analysed by others. D614G variant was found to be changed in amino acid such as from Aspartic acid to Glycine. This transition is a potentially significant alteration in a sequence as Aspartic acid is a large negatively charged and acidic amino acid, while Glycine is a small neutral amino acid and thus a transition from D to G could be led to electrostatic changes. Electrostatic interaction plays a major role in stability, flexibility, and function as well as protein folding. This event took place in the patient’s genome since changes of amino acids and the formation of electrostatic interactions changes contribute to improper protein folding which eventually decreases the stability of the spike protein. Non charged amino acid usually will not be the hotspot for binding. Mutation at position 477 from Serine to Asparagine causes a slight alteration in a sequence as Serine is a polar and non-charged amino acid while Asparagine is also a polar and non-charged amino acid. Polar defines that more hydrophilic group so can form a hydrogen bond with a water molecule. Salt bridges formed by positively and negatively charged amino acids are found to be important for the stabilization of protein three-dimensional structure hence if Serine and Asparagine are non-charged hence it leads to an assumption where this variant is weaker compared to D614G (Singh et al. 2020a). Hydrophilic interfaces are also not fit to be a hotspot for binding. Salt bridges are lacking in this transition leads to fluctuating stability (Begum et al. 2020; Kumar and Nussinov 2002).

This residue and aspartate 614 are directed in opposite directions with glycosylated residue at position 616 implying that glycine substitution also has a null effect on this relationship. The glycosylated residue (asparagine) Finally, at positions 854, 859, 860, 861 of the spike protein, neither position 614 nor the inter-atomic contacts lie in a polybasic cleavage area that is of interest to SARS-CoV-2 as it was proposed to enable the membrane fusion protein. Mutation D614G is considered conservative and unlikely to affect protein function which is proven with a negative score (-2.41) measured by the I-Mutant tool (Isabel et al. 2020). S477N is considered to also follow the mechanism of D614G in terms of tolerating protein function regardless of risk mutation since S477N was also found to be not affecting protein function however less negative score (-1.53) was recorded compared to D614G. Similarly, N501Y is also able to tolerate protein function with a decreasing stability score (-0.49).

Compared to S477N (0.46), the disease propensity is higher for N501Y (1.3) and D614G (1). Over the last few months, the spread of D614G has increased concerning the number of cases of infection in all the geographical regions concerned. Currently N501Y mutation is also predicted to be associated with the number of cases of infection and mortality. The virulence of N501Y and D614G is higher as it scores more than 1 and this result has shown that N510Y and D614G virulence have been transmitted in large quantities to infect humans. Virulence of N501Y and D614G is 10-fold more infectious than wild spike protein. A small, independently folded S1 subdomain, described as the Receptor-binding Domain (RBD), binds ACE2 directly when the virus engages the target cell (Korber et al. 2020; L. Zhang et al. 2020).

4.5 Comparison of secondary structure prediction

The result obtained for secondary structure prediction were tabulated which is unique and yet to be discovered by any other researchers. The secondary structure involves alpha helix, beta stranded, and coil region represented by (hh, ee, cc) respectively. D614 and G614 while S477 and N477 remain in random coils and do not impose any structural changes in secondary structure level upon mutation. The secondary structure generated consists of 274 alpha-helix, 281 extended strands, and 718 random coils which clearly shows that mutated spike protein and wild-type spike protein composed of a larger amount of random coils. The coil is highly sensitive to point mutation on the functional level. Mutant stability is better in the coil region. This analysis broadens the view of the correlation between amino acid properties and protein flexibility which reveals that it causes by mutation partially buried. Effects of mutation on irregular structure such as random coils are neutral (Gromiha et al. 1999).

4.6 RBD mutational changes the stability of tertiary protein structure

RBD mutational focuses on S477N and N501Y which are positioned at 319aa until 541aa which represent the RBD region in spike protein structure. S477N mutation was observed to be decreasing the protein stability however N501Y increases the protein stability. This can be seen in **Figure 11** whereby S477N is unlikely to associate with neighboring variants to form any bond which plays a major role in enhancing the protein stability but N501Y are forming many non-covalent interactions with surrounding variants at a different position such as 496 498, 506 which likely to contribute in enhancing the spike protein stability. Non-covalent interaction with adjacent drives the spontaneous protein folding which increases the stability of a protein (Karshikoff 2006). Hence, it can be concluded as S477N RBD mutational are not effective mutation which has any significant raise in protein stability. Both mutations are found on the surface of the protein structure.

4.7 RBD mutational affects the protein-protein interaction

Based on the result obtained on the effect of RBD mutational on protein-protein interaction between mutant spike protein and wild type spike protein, it shows that N501Y produce less binding energy. Less binding energy (more negative energy) contributes to high affinity. The more the negative energy, the better the mutant. Moreover, these mutants bound spontaneously without consuming much energy. Binding energy is released when a protein associates with its target leading to a lowering of the overall energy of the complex (Pantsar and Poso 2018). The amino acid substitutions and the longer capping loops could explain the increase in binding affinities in SARS-CoV-2. Higher affinity values might be related to the dynamic of infection and the rapid spread observed for this virus (Ortega et al. 2020).

The structural analysis of the RBD is therefore a very critical step for designing a potent inhibitor or an antibody to potentially block hACE2 and RBD interactions. It has also been reported that the region, which corresponds to residues 475-483 in RBD, is a site of interest due to a higher frequency of mutations in this region. Although S477N imposes stronger binding energy however compared to wild type and N501Y, the result reveals that S477N is unlikely to be a hotspot to bind with ACE2. N501Y has a higher binding affinity to affect local conformation to attach to ACE2 represent that position in ACE2 protein structure is suitable for inhibitors to bind to inhibit the infectivity and transmissibility of SARS-CoV-2 (Ortega et al. 2020; Singh et al. 2020b).

4.8. Comparison of spike mutation between human and mink host

Netherland and Danish mink-derived SARS-CoV-2 isolates have been reported to contain several mutations in the spike protein. Mink was probably infected by humans, representing naive and susceptible hosts, with a rapid spread of virus between mink farms. Mutations in the human virus spike protein have been transferred to the mink. This is supported by the available data: the D614G mutation that confers higher transmissibility of the virus between humans arose in the human-derived virus in China and perhaps Germany around February 2020 and, within three months, the original D614 became the dominant virus in the world. The global spread of the D614G mutant over the period February-May 2020 coincided temporarily with mink outbreaks in the Netherlands and Denmark (March and August 2020 respectively). This appears to be reflected in the increased frequency of the D614G mutant in the mink-derived virus in Denmark compared to the Netherlands. Also, Mink and ferrets are known to be susceptible to SARS-CoV-2 infection, but less so than human infection. Some adaptive changes in the spike protein and other viral proteins are therefore expected to occur after human transmission to the mink. Further research needs to be done on the link between the mink mutation N510T and the human mutation N501Y. A more in-depth study is needed to test whether this has been transmitted back to humans through adaptive changes and this paved way for the evolution of the new variant recently discovered in the United Kingdom VOC 202012/01(B.1.1.7) and South Africa. (table 9).

5. CONCLUSION

The present study indicated that D614G and N501Y may lead to increases in infectivity and transmissibility based on high binding affinity to human ACE2. Due to the alarming rate of cases in daily terms, scientists are working on current vaccine development to inhibit the transmission of SARS-CoV-2 however became a debate upon mutation has emerged lately. It is observed D614G have high number of occurrences in North America and widely circulated around the world. N501Y mutation recorded in RBD region, through computation study predicted it has an increase in proteins stability and also has a higher affinity to human ACE2 protein, which have higher chance of increase in infectivity and transmission. Although S477N is positioned at RBD, it has a less binding affinity towards the host cell and only found in Oceania. However, more studies need to carried out for an overall understanding of harmful variants that have been distributed all over the world on the advantageous escape from immune response and spread and does not affect antibody formation hence antibodies will be generated and there is no need for a new set of antibodies or vaccines. Therefore, a vaccine developed to be successful in inhibiting the transmission as well as work on all kinds of mutations.

There are possible limitations encountered in this study is that significant ranges could be detected with more representation of many other mutations in different countries as the number of sequences deposited in the database increasing after October 2020. As for the future study, this research can be brought narrower to identify target virus epitope which will serve effectively in developing target vaccine. Observation on highly conserved T and B cell epitopes is essential since they are vaccine candidates in terms of affinity receptor antibodies and defense efficacy. THE potential T cell epitope of S protein represents a significant source towards speedy development and production of the SARS-CoV-2 vaccine.

Funding: No funding for this study

Acknowledgment:

The author acknowledges the support of Management & Science University

ORCID ID:

Suresh Kumar: https://orcid.org/0000-0001-5682-0938,

Sarmilah Mathavan: https://orcid.org/0000-0001-6105-502X

REFERENCES

Ashwaq, Omar, Pratibha Manickavasagam, and Manirul Haque. 2020. "V483a-an Emerging Mutation Hotspot of Sars-Cov-2." (September).

Basu, Souradip, Suparba Mukhopadhyay, Rajdeep Das, Sarmishta Mukhopadhyay, Pankaj Kumar Singh, and Sayak Ganguli. 2020. "Impact of Clade Specific Mutations on Structural Fidelity of SARS-CoV-2 Proteins." *BioRxiv : The Preprint Server for Biology*.

Begum, Feroza, Arup Kumar Banerjee, and Upasana Ray. 2020. "Mutation Hot Spots in Spike Protein of SARS-CoV-2 Virus." *Preprints 2020, 2020040281* 1(1):1–31.

Böttcher-Friebertshäuser, Eva, Wolfgang Garten, and Hans Dieter Klenk. 2018. "Activation of Viruses by Host Proteases." *Activation of Viruses by Host Proteases* 1–335.

Brister, J. Rodney, Danso Ako-Adjei, Yiming Bao, and Olga Blinkova. 2015. "NCBI Viral Genomes Resource." *Nucleic Acids Research* 43(D1):D571–77.

Capriotti, Emidio, Piero Fariselli, and Rita Casadio. 2005. "I-Mutant2.0: Predicting Stability Changes upon Mutation from the Protein Sequence or Structure." *Nucleic Acids Research* 33(SUPPL. 2):306–10.

Ceraolo, Carmine, and Federico M. Giorgi. 2020. "Genomic Variance of the 2019-NCoV Coronavirus." *Journal of Medical Virology* 92(5):522–28.

Chen, Jiahui, Rui Wang, Menglun Wang, and Guo Wei Wei. 2020. "Mutations Strengthened SARS-CoV-2 Infectivity." *Journal of Molecular Biology* 432(19):5212–26.

Cui, Jie, Fang Li, and Zheng Li Shi. 2019. "Origin and Evolution of Pathogenic Coronaviruses." *Nature Reviews Microbiology* 17(3):181–92.

Deng, Sheng-Qun, and Hong-Juan Peng. 2020. "Characteristics of and Public Health Responses to the Coronavirus Disease 2019 Outbreak in China." *Journal of Clinical Medicine* 9(2):575.

Dowd, Jennifer Beam, Liliana Andriano, David M. Brazel, Valentina Rotondi, Per Block, Xuejie Ding, Yan Liu, and Melinda C. Mills. 2020. "Demographic Science Aids in Understanding the Spread and Fatality Rates of COVID-19." *Proceedings of the National Academy of Sciences of the United States of America* 117(18):9696–98.

Ferron, François, Lorenzo Subissi, Ana Theresa Silveira De Morais, Nhung Thi Tuyet Le, Marion Sevajol, Laure Gluais, Etienne Decroly, Clemens Vonrhein, Gérard Bricogne, Bruno Canard, and Isabelle Imbert. 2017. "Structural and Molecular Basis of Mismatch Correction and Ribavirin Excision from Coronavirus RNA." *Proceedings of the National Academy of Sciences of the United States of America* 115(2):E162–71.

Garnier, Jean, Jean François Gibrat, and Barry Robson. 1996. "[32] GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence." *Methods in Enzymology* 266(1995):540–53.

Gasteiger, Elisabeth, Alexandre Gattiker, Christine Hoogland, Ivan Ivanyi, Ron D. Appel, and Amos Bairoch. 2003. "ExPASy: The Proteomics Server for in-Depth Protein Knowledge and Analysis." *Nucleic Acids Research* 31(13):3784–88.

Gromiha, M. Michael, Motohisa Oobatake, Hidetoshi Kono, Hatsuho Uedaira, and Akinori Sarai. 1999. "Role of Structural and Sequence Information in the Prediction of Protein Stability Changes: Comparison between Buried and Partially Buried Mutations." *Protein Engineering* 12(7):549–55.

Heald-Sargent, Taylor, and Tom Gallagher. 2012. "Ready, Set, Fuse! The Coronavirus Spike Protein and Acquisition of Fusion Competence." *Viruses* 4(4):557–80.

Heberle, Henry, Vaz G. Meirelles, Felipe R. da Silva, Guilherme P. Telles, and Rosane Minghim. 2015. "InteractiVenn: A Web-Based Tool for the Analysis of Sets through Venn Diagrams." *BMC Bioinformatics* 16(1):1–7.

Hodcroft, Emma B., Moira Zuber, Sarah Nadeau, Iñaki Comas, Fernando González Candelas, SeqCOVID-SPAIN Consortium, Tanja Stadler, and Richard A. Neher. 2020. "Emergence and Spread of a SARS-CoV-2 Variant through Europe in the Summer of 2020." *MedRxiv* 2020(October):2020.10.25.20219063.

Isabel, Sandra, Lucía Graña-Miraglia, Jahir M. Gutierrez, Cedoljub Bundalovic-Torma, Helen E. Groves, Marc R. Isabel, Ali Reza Eshaghi, Samir N. Patel, Jonathan B. Gubbay, Tomi Poutanen, David S. Guttman, and Susan M. Poutanen. 2020. "Evolutionary and Structural Analyses of SARS-CoV-2 D614G Spike Protein Mutation Now Documented Worldwide." *Scientific Reports* 10(1):1–9.

Jiang, Shibo, Lanying Du, and Zhengli Shi. 2020. "An Emerging Coronavirus Causing Pneumonia Outbreak in Wuhan, China: Calling for Developing Therapeutic and Prophylactic Strategies." *Emerging Microbes and Infections* 9(1):275–77.

Karshikoff, Andrey. 2006. "Non-Covalent Interactions in Proteins." *Non-Covalent Interactions in Proteins* (March):1–334.

Korber, Bette, Will M. Fischer, Sandrasegaram Gnanakaran, Hyejin Yoon, James Theiler, Werner Abfalterer, Nick Hengartner, Elena E. Giorgi, Tanmoy Bhattacharya, Brian Foley, Kathryn M. Hastie, Matthew D. Parker, David G. Partridge, Cariad M. Evans, Timothy M. Freeman, Thushan I. de Silva, Adrienne Angyal, Rebecca L. Brown, Laura Carrilero, Luke R. Green, Danielle C. Groves, Katie J. Johnson, Alexander J. Keeley, Benjamin B. Lindsey, Paul J. Parsons, Mohammad Raza, Sarah Rowland-Jones, Nikki Smith, Rachel M. Tucker, Dennis Wang, Matthew D. Wyles, Charlene McDanal, Lautaro G. Perez, Haili Tang, Alex Moon-Walker, Sean P. Whelan, Celia C. LaBranche, Erica O. Saphire, and David C. Montefiori. 2020. "Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus." *Cell* 182(4):812-827.e19.

Kumar, Sandeep, and Ruth Nussinov. 2002. “Close Range Electrostatic Interactions in Proteins.” 604–17.

Laskowski, Roman A., James D. Stephenson, Ian Sillitoe, Christine A. Orengo, and Janet M. Thornton. 2020. “VarSite: Disease Variants and Protein Structure.” *Protein Science* 29(1):111–19.

Lu, Roujian, Xiang Zhao, Juan Li, Peihua Niu, Bo Yang, Honglong Wu, Wenling Wang, Hao Song, Baoying Huang, Na Zhu, Yuhai Bi, Xuejun Ma, Faxian Zhan, Liang Wang, Tao Hu, Hong Zhou, Zhenhong Hu, Weimin Zhou, Li Zhao, Jing Chen, Yao Meng, Ji Wang, Yang Lin, Jianying Yuan, Zhihao Xie, Jinmin Ma, William J. Liu, Dayan Wang, Wenbo Xu, Edward C. Holmes, George F. Gao, Guizhen Wu, Weijun Chen, Weifeng Shi, and Wenjie Tan. 2020. “Genomic Characterisation and Epidemiology of 2019 Novel Coronavirus: Implications for Virus Origins and Receptor Binding.” *The Lancet* 395(10224):565–74.

McAuley, Alexander J., Michael J. Kuiper, Peter A. Durr, Matthew P. Bruce, Jennifer Barr, Shawn Todd, Gough G. Au, Kim Blasdell, Mary Tachedjian, Sue Lowther, Glenn A. Marsh, Sarah Edwards, Timothy Poole, Rachel Layton, Sarah Jane Riddell, Trevor W. Drew, Julian D. Druce, Trevor R. F. Smith, Kate E. Broderick, and S. S. Vasan. 2020. “Experimental and in Silico Evidence Suggests Vaccines Are Unlikely to Be Affected by D614G Mutation in SARS-CoV-2 Spike Protein.” *Npj Vaccines* 5(1):1–5.

Ogawa, Junko, Wei Zhu, Nina Tonnu, Oded Singer, Tony Hunter, Amy Ryan (Firth), and Gerald Pao. 2020. “The D614G Mutation in the SARS-CoV2 Spike Protein Increases Infectivity in an ACE2 Receptor Dependent Manner.” *BioRxiv : The Preprint Server for Biology* 90033(4).

Ortega, Joseph Thomas, Maria Luisa Serrano, Flor Helene Pujol, and Hector Rafael Rangel. 2020. “Role of Changes in SARS-CoV-2 Spike Protein in the Interaction with the Human ACE2 Receptor: An in Silico Analysis.” *EXCLI Journal* 19:410–17.

Pantsar, Tatu, and Antti Poso. 2018. “Binding Affinity via Docking: Fact and Fiction.” *Molecules (Basel, Switzerland)* 23(8):1–11.

Ritchie, Dave, and Team Orpailleur. 2013. “Hex 8.0.0 User Manual.” *Protein Docking Using Spherical Polar Fourier Correlations*.

Rothan, Hussin A., and Siddappa N. Byrareddy. 2020. “The Epidemiology and Pathogenesis of Coronavirus Disease (COVID-19) Outbreak.” *Journal of Autoimmunity* (February):102433.

Sen, Taner Z., Robert L. Jernigan, Jean Garnier, and Andrzej Kloczkowski. 2005. “GOR V Server for Protein Secondary Structure Prediction.” *Bioinformatics* 21(11):2787–88.

Sim, Ngak Leng, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C. Ng. 2012. “SIFT Web Server: Predicting Effects of Amino Acid Substitutions on Proteins.” *Nucleic Acids Research* 40(W1):452–57.

Simmonds, Peter. 2012. “SSE: A Nucleotide and Amino Acid Sequence Analysis Platform.” *BMC Research Notes* 5(January).

Singh, Amit, Georg Steinkellner, Katharina Köchl, Karl Gruber, and Christian C. Gruber. 2020a. “Serine 477 Plays a Crucial Role in the Interaction of the SARS-CoV-2 Spike Protein with the Human Receptor ACE2 .” *BioRxiv : The Preprint Server for Biology* 1–28.

Singh, Amit, Georg Steinkellner, Katharina Köchl, Karl Gruber, and Christian C. Gruber. 2020b. “Serine 477 Plays a Crucial Role in the Interaction of the SARS-CoV-2 Spike Protein with the Human Receptor ACE2 .” 1–28.

Velavan, Thirumalaisamy P., and Christian G. Meyer. 2020. “The COVID-19 Epidemic.” *Tropical Medicine and International Health* 25(3):278–80.

Wang, Rui, Yuta Hozumi, Changchuan Yin, and Guo-Wei Wei. 2020. “Decoding SARS-CoV-2 Transmission and Evolution and Ramifications for COVID-19 Diagnosis, Vaccine, and Medicine.” *Journal of Chemical Information and Modeling*.

Wu, Fan, Su Zhao, Bin Yu, Yan Mei Chen, Wen Wang, Zhi Gang Song, Yi Hu, Zhao Wu Tao, Jun Hua Tian, Yuan Yuan Pei, Ming Li Yuan, Yu Ling Zhang, Fa Hui Dai, Yi Liu, Qi Min Wang, Jiao Jiao Zheng, Lin Xu, Edward C. Holmes, and Yong Zhen Zhang. 2020. “A New Coronavirus Associated with Human Respiratory Disease in China.” *Nature* 579(7798):265–69.

Wu, Joseph T., Kathy Leung, and Gabriel M. Leung. 2020. “Nowcasting and Forecasting the Potential Domestic and International Spread of the 2019-NCoV Outbreak Originating in Wuhan, China: A Modelling Study.” *The Lancet* 395(10225):689–97.

Yuting, Chen, Haoyu Lu, Ning Zhang, Zefeng Zhu, Shuqin Wang, and Minghui Li. 2020. “PremPS: Predicting the Effects of Single Mutations on Protein-RNA Interactions.” *BioRxiv : The Preprint Server for Biology*.

Zhang, Haibo, Josef M. Penninger, Yimin Li, Nanshan Zhong, and Arthur S. Slutsky. 2020. “Angiotensin-Converting Enzyme 2 (ACE2) as a SARS-CoV-2 Receptor: Molecular Mechanisms and Potential Therapeutic Target.” *Intensive Care Medicine* 46(4):586–90.

Zhang, Lizhou, Cody Jackson, Huihui Mou, Amrita Ojha, Erumbi Rangarajan, Tina Izard, Michael Farzan, and Hyeryun Choe. 2020. “The D614G Mutation in the SARS-CoV-2 Spike Protein Reduces S1 Shedding and Increases Infectivity.” *BioRxiv : The Preprint Server for Biology*.

Zhou, Peng, Xing Lou Yang, Xian Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao Rui Si, Yan Zhu, Bei Li, Chao Lin Huang, Hui Dong Chen, Jing Chen, Yun Luo, Hua Guo, Ren Di Jiang, Mei Qin Liu, Ying Chen, Xu Rui Shen, Xi Wang, Xiao Shuang Zheng, Kai Zhao, Quan Jiao Chen, Fei Deng, Lin Lin Liu, Bing Yan, Fa Xian Zhan, Yan Yi Wang, Geng Fu Xiao, and Zheng Li Shi. 2020. “A Pneumonia Outbreak

Associated with a New Coronavirus of Probable Bat Origin." *Nature* 579(7798):270–73.

Kumar, S., Mathavan, S., Jin, W. J., Azman, N. A. B., Subramanaiam, D., Zainalabidin, N. A. B., ... & Taqiyuddin, J. A. 2020. "COVID-19 Vaccine Candidates by Identification of B and T Cell Multi-Epitopes Against SARS-COV-2". Preprints doi: 10.20944/preprints202008.0092.v1

Kumar, S. 2020. "COVID-19: A drug repurposing and biomarker identification by using comprehensive gene-disease associations through protein-protein interaction network analysis". Preprints doi: 10.20944/preprints202003.0440.v1

Kumar, S. 2020. "Drug and vaccine design against Novel Coronavirus (2019-nCoV) spike protein through Computational approach". Preprints doi: 10.20944/preprints202002.0071.v1