Emergence of mutations and possible antigenic drift in the surface glycoprotein of SARS-CoV-2 (COVID-19)

Saeed Mujahid Hashimi[1]

[1]Griffith University Griffith Health

April 28, 2020

Abstract

Recently (2019), a novel coronavirus (SARS-CoV-2) first reported in Wuhan, China has been declared a pandemic by the World Health Organization and is rapidly spreading throughout the globe which is associated with high morbidity and mortality, especially in the elderly and those with existing chronic conditions. SARS-CoV-2 infects cells through interaction of its surface glycoprotein with the human angiotensin converting enzyme 2 (ACE-2). This study conducted a analysis of mutation frequency in the surface glycoprotein of 796 sequenced SARS-CoV-2 isolates from different geographical locations in the GISAID and GenBank databases. Multiple sequence alignment analysis of the surface glycoprotein identified 64 different mutations at the protein level spanning multiple geographic locations globally. A cluster of mutations was identified in the receptor binding domain (RBD) of the surface glycoprotein. Significantly, the analysis showed that 68.5% of the isolates contain a D614 residue compared to 31.5% which contain a G614 suggesting virus is spreading in two forms. Furthermore, our investigation found that one isolate from Belgium had acquired 5 cumulative mutations in the surface glycoprotein indicating possible antigenic drift. The findings of this study are of critical importance for the design of vaccines and novel drugs against this severe acute respiratory syndrome coronavirus.

Emergence of mutations and possible antigenic drift in the surface glycoprotein of SARS-CoV-2 (COVID-19)

Saeed Mujahid Hashimi[1*]

[1] School of Medical Science and Menzies Health Institute Queensland, Griffith University, Gold Coast, Queensland 4333, Australia

Correspondence: Dr Saeed M Hashimi, School of Medical Science and Menzies Health Institute Queensland, Griffith University, Gold Coast, Queensland 4333, Australia

E-mail: s.hashimi@griffith.edu.au

Running Title: Possible antigenic drift of SARS-CoV-2

Keywords: SARS-CoV-2; surface glycoprotein; ACE2; severe acute respiratory syndrome coronavirus; COVID-19, antigenic drift

Summary

Recently (2019), a novel coronavirus (SARS-CoV-2) first reported in Wuhan, China has been declared a pandemic by the World Health Organization and is rapidly spreading throughout the globe which is associated with high morbidity and mortality, especially in the elderly and those with existing chronic conditions. SARS-CoV-2 infects cells through interaction of its surface glycoprotein with the human angiotensin converting enzyme 2 (ACE-2). This study conducted a analysis of mutation frequency in the surface glycoprotein of 796 sequenced SARS-CoV-2 isolates from different geographical locations in the GISAID and GenBank

databases. Multiple sequence alignment analysis of the surface glycoprotein identified 64 different mutations at the protein level spanning multiple geographic locations globally. A cluster of mutations was identified in the receptor binding domain (RBD) of the surface glycoprotein. Significantly, the analysis showed that 68.5% of the isolates contain a D614 residue compared to 31.5% which contain a G614 suggesting virus is spreading in two forms. Furthermore, our investigation found that one isolate from Belgium had acquired 5 cumulative mutations in the surface glycoprotein indicating possible antigenic drift. The findings of this study are of critical importance for the design of vaccines and novel drugs against this severe acute respiratory syndrome coronavirus.

Introduction

The novel 2019 SARS-CoV-2 coronavirus which causes significant morbidity and mortality has been declared a pandemic by the World Health Organization (Chan et al., 2020). Due to the highly infectious nature of this contagion, urgent and coordinated efforts are required to develop a vaccine against SARS-CoV-2.

SARS-CoV-2 is genetically related to the SARS-CoV coronavirus which infected 8096 in 25 countries around the world (Tang et al., 2020). Coronaviruses are RNA viruses with characteristic spikes on their envelopes resembling a crown-like structure (Siddell, 1995). They are zoonotic and therefore most likely crossed from animals such as bats into humans through intermediaries (Andersen, Rambaut, Lipkin, Holmes, & Garry, 2020; Huynh et al., 2012; Lai, Bergna, Acciarri, Galli, & Zehender, 2020). Clinical features of SARS-CoV-2 range from asymptomatic to severe respiratory distress syndrome (Wu et al., 2020).

The mode entry of SARS-CoV-2 has been recently shown to be through its direct interaction of the surface glycoprotein with human angiotensin-converting enzyme 2 (hACE2) (Hoffmann et al., 2020; Yan et al., 2020). This mode of entry is the same as hCoV-NL63 and SARS-CoV coronaviruses (Hofmann et al., 2005; Kuba et al., 2005). The receptor-binding domain of the surface glycoprotein was shown to be important for the entry of SARS-CoV-2 into the host cells as evidence by the crystal structure of the SARS-CoV-2 RBD domain in complex with hACE2 (Tai et al., 2020). Since the surface glycoprotein is surfaced exposed, it makes an ideal candidate for the development of neutralizing antibodies, therapeutic drug design and vaccine development (Ahmed, Quadeer, & McKay, 2020). However, due to the emergence of high mutation rates in RNA viruses, therapeutic modalities might be hampered (Elena & Sanjuán, 2005). Furthermore, antigen drift as has been a problem in the development of influenza vaccines needs to be considered in SARS-CoV-2 (Hensley et al., 2009).

This study analyzed the surface glycoprotein from emerging isolates of SARS-CoV-2 sequenced globally. Mutations in several isolates were detected from different geographical locations, which might have important implications for vaccine design and therapeutic development.

Methods and Materials

Sequences used in this study

94 protein sequences for the surface (S) glycoprotein from different isolates of SARS-CoV-2 were retrieved (22/03/2020) from the GenBank database through the National Center for Biotechnology Information (NCBI). A further 731 genomes from different isolates of the SARS-CoV-2 virus were retrieved from the GISAID EpiCoV database (22/03/2020). The genomes from GISAID were selected to contain complete sequences with high coverage and exclude low coverage sequences. A list of sequences used in this study can be found in Supplementary Figure 1 and 2.

Sequence trimming

As the DNA sequences from GISAID were not annotated, it was necessary to firstly trim the sequences and isolate the coding sequence for the surface glycoprotein. BioPerl was used to isolate the gene for the surface glycoprotein from a total of 731 SARS-CoV-2 genomes. A simple Perl script was designed to remove sequences upstream of the ATG start codon and downstream from the TAA stop codon (Supplementary

Figure 3). Subsequently, incomplete sequences were removed which resulted in 702 sequences for further analysis.

EMBOSS Transeq was used to translate the DNA sequences of the surface glycoprotein into amino acid sequences.

Multiple alignment of the surface glycoprotein and identification of sequence variants

Surface glycoproteins sequences obtained from the GenBank were aligned using the multiple sequence alignment feature of the BLAST sequence analysis tool from NCBI (Supplementary Figure 1) (Altschul et al., 1997). The first sequence SARS-CoV-2 from isolate Wuhan-Hu-1 (Accession: MN908947.3; Wuhan, China) was used as the consensus sequence to determine subsequent variant sequences from the different isolates (Wu et al., 2020).

Furthermore, the surface glycoprotein sequences obtained from GISAID were analyzed for similarity by using the multiple sequence alignment tool Clustal Omega (Supplementary Figure 2) (Madeira et al., 2019). Again, the sequences were compared to the Wuhan-Hu-1 isolate (Accession: MN908947.3) to determine divergent sequences. Also, multiple sequence alignment for all isolates from Belgium was conducted separately (Supplementary Figure 4). The multiple sequence alignments were analyzed using the software BioEdit to identify the sequence differences between isolates.

Results

The SARS-CoV-2 surface glycoprotein is rapidly mutating

Multiple alignment of all sequences (retrieved: 22/03/2020) that were available in GenBank and GISAID for the surface glycoprotein of the SARS-CoV-2 protein from different isolates showed that there is a rapid emergence of mutations in the amino acid sequences (Figure 1). All the mutations were substitutions, except for one isolate from India which had a deletion of Y145del. The analysis showed that the mutations are scattered in isolates from different regions around the globe, with a significant proportion from European states (Supplementary Table 1). In total 63 different substitution mutations and 1 deletion were observed.

Mutations in SARS-CoV-2 surface glycoprotein are clustered

The mutations were mapped onto the surface glycoprotein sequence to determine clusters around specific domains (Figure 1). Interestingly, the mutations were found to cluster around the N-terminal domain (NTD), the receptor-binding domain (RBD), and the N-terminal of the S2 fragment.

The N-terminal domain (NTD) showed 14 substitutions and a single deletion mutation which were identified from 28 different isolates. H49Y was identified in six isolates originating from China (5) and the United States of America (1) (Figure 2A). Furthermore, Q239K was found in six isolates from the Netherlands (5) and Finland (1). While M153T (2 isolates from China) and S254F (Isolates from Germany and Netherlands) were identified in 2 isolates each.

The receptor-binding domain (RBD) showed 12 substitution mutations that were identified from 19 different isolates. V367F was found in six isolates that originated from France (5) and Hong Kong (1). Also, V483A was identified in three isolates from the United States of America.

Analysis of mutations in the C-terminal of the S1 fragment revealed that the SARS-CoV-2 virus is spreading in two forms around the globe. Specifically, 481 isolates showed an aspartic acid (D) at residue 615 accounting for 68.5% of the total isolates analyzed (Figure 2B). Nearly all isolates from China, including the first reported SARS-CoV-2 genome, contained this amino acid substitution except for two isolates. Furthermore, D614 was found in multiple isolates from America, Europe, Asia, and Australasia. However, G614 was found in 221 isolates accounting for 31.5% of all sequences (Figure 2B). Interestingly, nearly all the isolates G164 were from outside of China (G614 was found in Wuhan_HBCDC-HB-06_2020 and China_Shanghai_SH0025). The G614 variant first emerged in China on the 6^{th} of February 2020 in isolate China_Shanghai_SH0025. In fact, both the D614 and G614 variants emerged in China and subsequently have spread throughout the world.

Emergence of cumulative mutations SARS-CoV-2 S glycoprotein

Analysis of the S glycoprotein sequences of SARS-CoV-2 showed that several isolates have rapidly accumulated multiple mutations. Isolate Beijing_IVDC-BJ-005_2020 from China was found to contain substitutions V860C, L861K, and F970S. While isolate Belgium_BA-02291_2020 from Belgium contained substitutions K458R, H519P, D614G, T941A, and S943P. Further analysis also revealed 51 different isolates from Belgium (Supplementary Figure 4) had the D614G and S943P substitutions, indicating that Belgium_BA-02291_2020 most likely evolved from these isolates. The number of isolates that had accumulated 2 mutations from around the world was found to be 30.

Discussion

SARS-CoV-2 or commonly known as COVID-19 is a novel RNA coronavirus that has caused significant morbidity and mortality around the globe and continues to spread rapidly (*Coronavirus disease 2019 (COVID-19): situation report, 64* , 2020). As of the 24$^{\text{th}}$ of March 2020, according to the WHO situation report 64, the current confirmed cases are 372,757 and 16,231 deaths since the first reports in early January 2020 (*Coronavirus disease 2019 (COVID-19): situation report, 64* , 2020). Therefore, there is an urgent need for new therapeutic regimes such as a vaccine against this novel aggressive contagion. To this end, a promising SARS-CoV-2 antigen candidate for vaccine development is the characteristic surface (spike) glycoprotein of the virus. Several current clinical trials are looking at developing vaccines for SARS-CoV-2. However, the success of these trials depends on the vaccine providing protective immunity against the virus. Recently, it has been shown that SARS-CoV-2 enters the epithelial cells via interaction of its surface glycoprotein with the human angiotensin-converting enzyme 2 (ACE2) found on the surface of the epithelium (Hoffmann et al., 2020; Yan et al., 2020). This study sought to investigate the emergence of mutations in the surface glycoprotein of SARS-CoV-2 and their implications on the development of an effective vaccine to disrupt the entry of the virus into human cells.

Sequences for the surface glycoprotein of SARS-CoV-2 from different isolates in the GISAID and GenBank databases were analyzed for mutations using Multiple clustalW Sequence alignment. In this study, the first sequenced SARS-CoV-2 genome from China was used as a reference to map the mutations to (Tang et al., 2020). The analysis found that SARS-CoV-2 is rapidly evolving and is acquiring mutations that may cause antigenic drift as seen in influenza viruses (Boni, 2008). It was found that the surface glycoprotein had 64 different mutations in the protein sequence from the pool of 796 SARS-CoV-2 isolates analyzed from around the world. This rapid mutation rate combined with the highly infectious nature of the virus has great implications for the development of a therapeutic vaccine. The 1918 pandemic caused by the H1N1 influenza virus, infected over one-third of the global population and killed over 40 million people (Taubenberger, 2005). Current vaccines for influenza need to be updated seasonally to catch up on the rapid mutation rates of the virus, particularly in the haemagglutinin (HA) protein which is used by the virus to enter epithelial cells (Boni, 2008).

SARS-CoV-2 enters the host cell through binding to the surface-bound ACE2 enzyme (Tai et al., 2020). The mode of entry is through the interaction of the receptor-binding domain (RBD) of the surface glycoprotein with the ACE2 host cell membrane enzyme (Yan et al., 2020). Therefore, disrupting this interaction through the production of neutralizing antibodies has important implications for vaccine development. However, continuous mutations in this region need to be considered for the optimal design of vaccine antigens. The analysis from this study found that among the 796 isolates, 12 substitution mutations were identified in the RBD of the surface glycoprotein in 19 isolates originating from China, USA and Europe. Therefore, it is critical that mutations in the RBD region of the surface glycoprotein need to be considered in the design of any vaccine.

Antigenic drift is the process by which viruses accumulate multiple mutations in the sequence of proteins that the immune system recognizes as non-self and mount neutralizing antibodies against (Carrat & Flahault, 2007). If enough mutations accumulate over time in the viral proteins, the immune system can longer recognize the viral antigens and as such this can cause a significant epidemic or pandemic (Both, Sleigh, Cox,

& Kendal, 1983). SARS-CoV-2 is a novel virus to which there are no current vaccines in clinical use, while several accelerated clinical trials are underway to develop a vaccine. This study reports that SARS-CoV-2 is rapidly moving towards antigenic drift as shown by the sequence of one isolate from Belgium (Belgium_BA-02291_2020) which contains substitutions K458R, H519P, D614G, T941A, and S943P. Furthermore, K458R and H519P are located in the RBD region of the surface glycoprotein, which suggests the virus may be evolving to improve its binding capacity the ACE2 host surface enzyme. Also, of significance is the fact that Belgium_BA-02291_2020 has most likely evolved from one of the other 51 isolates in Belgium which had the D614G and S943P substitutions. Another isolate from China (Beijing_IVDC-BJ-005_2020) was found to contain substitutions V860C, L861K, and F970S, indicating possible antigenic drift.

To conclude, this study has established that the SARS-CoV-2 surface glycoprotein is rapidly mutating and evolving. More importantly, the virus is moving rapidly towards antigenic drift and this has important implications for the development and efficacy of a vaccine. The sequence analysis of the surface glycoprotein from SARS-CoV-2 has shed light on the rapid nature of SARS-CoV-2 to mutate which needs to be considered in future studies investigating possible therapeutics against the virus.

ACKNOWLEDGMENT

The author is grateful for the support from the School of Medical Science, Griffith University in providing online resources. Furthermore, we thank GISAID and GenBank for providing the genome sequences used in this analysis.

Disclaimers

The authors declare no competing interests.

Data availability

Data is available from the authors upon request. Sequence data were extracted from NCBI and GISAID.

Ethical approval

Ethical approval is not applicable as the sequence data for the SARS-CoC-2 used in this study was extracted from the GenBank and GISAID databases.

References

Ahmed, S. F., Quadeer, A. A., & McKay, M. R. (2020). Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. *Viruses, 12* (3), 254.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research, 25* (17), 3389-3402.

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nature Medicine* , 1-3.

Boni, M. F. (2008). Vaccination and antigenic drift in influenza. *Vaccine, 26* , C8-C14.

Both, G. W., Sleigh, M., Cox, N., & Kendal, A. (1983). Antigenic drift in influenza virus H3 hemagglutinin from 1968 to 1980: multiple evolutionary pathways and sequential amino acid changes at key antigenic sites. *Journal of virology, 48* (1), 52-60.

Carrat, F., & Flahault, A. (2007). Influenza vaccine: the challenge of antigenic drift. *Vaccine, 25* (39-40), 6852-6862.

Chan, J. F.-W., Yuan, S., Kok, K.-H., To, K. K.-W., Chu, H., Yang, J., . . . Poon, R. W.-S. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet, 395* (10223), 514-523.

Coronavirus disease 2019 (COVID-19): situation report, 64(64). (2020). Retrieved from

Elena, S. F., & Sanjuan, R. (2005). Adaptive value of high mutation rates of RNA viruses: separating causes from consequences. *Journal of virology, 79* (18), 11555-11558.

Hensley, S. E., Das, S. R., Bailey, A. L., Schmidt, L. M., Hickman, H. D., Jayaraman, A., . . . Bennink, J. R. (2009). Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. *Science, 326* (5953), 734-736.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Kruger, N., Herrler, T., Erichsen, S., . . . Nitsche, A. (2020). SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* .

Hofmann, H., Pyrc, K., van der Hoek, L., Geier, M., Berkhout, B., & Pohlmann, S. (2005). Human coronavirus NL63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry.*Proceedings of the National Academy of Sciences, 102* (22), 7988-7993.

Huynh, J., Li, S., Yount, B., Smith, A., Sturges, L., Olsen, J. C., . . . Gates, J. E. (2012). Evidence supporting a zoonotic origin of human coronavirus strain NL63. *Journal of virology, 86* (23), 12816-12825.

Kuba, K., Imai, Y., Rao, S., Gao, H., Guo, F., Guan, B., . . . Deng, W. (2005). A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus–induced lung injury. *Nature Medicine, 11* (8), 875-879.

Lai, A., Bergna, A., Acciarri, C., Galli, M., & Zehender, G. (2020). Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. *Journal of medical virology* .

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., . . . Finn, R. D. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research, 47* (W1), W636-W641.

Siddell, S. G. (1995). The coronaviridae. In *The coronaviridae*(pp. 1-10): Springer.

Tai, W., He, L., Zhang, X., Pu, J., Voronin, D., Jiang, S., . . . Du, L. (2020). Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cellular & Molecular Immunology* , 1-8.

Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., . . . Qian, Z. (2020). On the origin and continuing evolution of SARS-CoV-2.*National Science Review* .

Taubenberger, J. (2005). The virulence of the 1918 pandemic influenza virus: unraveling the enigma. In *Infectious Diseases from Nature: Mechanisms of Viral Emergence and Persistence* (pp. 101-115): Springer.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., . . . Pei, Y.-Y. (2020). A new coronavirus associated with human respiratory disease in China. *Nature, 579* (7798), 265-269.

Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., & Zhou, Q. (2020). Structural basis for the recognition of the SARS-CoV-2 by full-length human ACE2. *Science* , eabb2762. doi:10.1126/science.abb2762

Figure Legends

Figure 1. Mutation map of the SARS-CoV-2 surface glycoprotein. Multiple sequence alignment of the surface glycoprotein of SARS-CoV-2 was performed on 796 different isolates sequenced from around the world. The sequences were compared to the first sequenced genome of SARS-CoV-2 from China (Tang et al., 2020). Substitution and deletion mutations are mapped to the surface glycoprotein. SP, signal peptide; NTD, N-terminal domain; RBD, receptor-binding domain; S1, cleavage fragment 1; S2 cleavage fragment 2; S cleavage site, potential furin protease site.

Figure 2. Enumeration of mutations SARS-CoV-2 surface glycoprotein occurring frequently . **A.** Frequency of the mutations in the surface glycoprotein. H49Y, Q239K, and V367F were found in six different isolates from around the world. **B.** Frequency of D614 compared to G614 in the analyzed isolates.

68.5% of isolates contained the D614 amino acid residue while 31.5% of isolates contained G614. Both variants of the surface glycoprotein originated from China.

Supplementary Tables

Supplementary Table 1. List of mutations detected in all SARS-CoV-2 isolates analyzed.

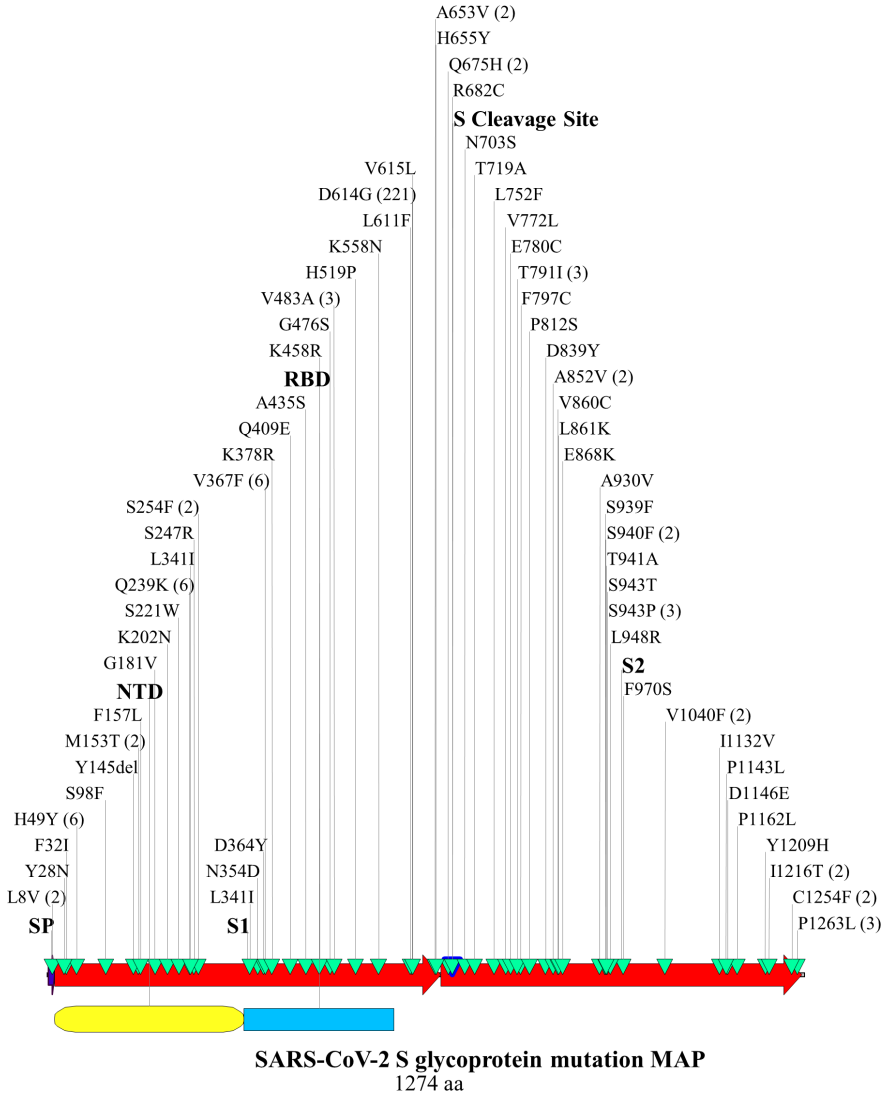
Supplementary Figure Legend

Supplementary Figure 1. Multiple Alignment of SARS-CoV-2 surface glycoprotein sequence from isolates in GenBank.

Supplementary Figure 2. SARS-CoV-2 surface glycoprotein multiple sequence alignment of isolates from GISAID.

Supplementary Figure 3. Perl script used to trim the SARS-CoV-2 surface glycoprotein sequences from different isolates in GISAID.

Supplementary Figure 4. SARS-CoV-2 surface glycoprotein multiple sequence alignment of isolates from Belgium in GISAID.



SARS-CoV-2 S glycoprotein mutation MAP
1274 aa