Comprehensive annotations of the mutational spectra of SARS-CoV-2 spike protein: a fast and accurate pipeline

M. Shaminur Rahman[1*], M. Rafiul Islam[1*], M. Nazmul Hoque[1*,2], A. S. M. Rubayet Ul Alam[1,3], Masuda Akther[1], J. Akter Puspo[1], Salma Akter[1,4], Azraf Anwar[5], Munawar Sultana[1], M. Anwar Hossain[1,6**]

[1]Department of Microbiology, University of Dhaka, Dhaka 1000, Bangladesh

[2]Department of Gynecology, Obstetrics and Reproductive Health, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur-1706, Bangladesh

[3]Department of Microbiology, Jashore University of Science and Technology, Jashore 7408, Bangladesh

[4]Department of Microbiology, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh

[5]Independent Researcher, 47-07 41st Street, New York, USA, Email: aa3641@columbia.edu

[6]Present address: Vice-Chancellor, Jashore University of Science and Technology, Jashore 7408, Bangladesh

*Equal contribution

**Corresponding to:

M. Anwar Hossain, PhD
Professor
Department of Microbiology
University of Dhaka, Dhaka 1000, Bangladesh
E-mail: hossaina@du.ac.bd

Abstract

In order to explore nonsynonymous mutations and deletions in the spike (S) protein of SARS-CoV-2, we comprehensively analyzed 35,750 complete S protein gene sequences from across six continents and five climate zones around the world, as documented in the GISAID database as of June 24th, 2020. Through a custom Python-based pipeline for analyzing mutations, we identified 27,801 (77.77 % of spike sequences) mutated strains compared to Wuhan-Hu-1 strain. 84.40% of these strains had only single amino-acid (aa) substitution mutations, but an outlier strain from Bosnia and Herzegovina (EPI_ISL_463893) was found to possess six aa substitutions. The D614G variant of the major G clade was found to be predominant across circulating strains in all climates. We also identified 988 unique aa substitution mutations distributed across 660 positions within the spike protein, with eleven sites showing high variability – these sites had four types of aa variations at each position. Besides, 17 in-frame deletions at four major regions (three in N-terminal domain and one just downstream of the RBD) may have possible impact on attenuation. Moreover, the mutational frequency differed significantly (p= 0.003, Kruskal–Wallis test) among the SARS-CoV-2 strains worldwide. This study presents a fast and accurate pipeline for identifying nonsynonymous mutations and deletions from large dataset for any particular protein coding sequence and presents this S protein data as representative analysis. By using separate multi-sequence alignment with MAFFT, removing ambiguous sequences and in-frame stop codons, and utilizing pairwise alignment, this method can derive nonsynonymus mutations (Reference:Position:Strain). We believe this will aid in the surveillance of any proteins encoded by SARS-CoV-2, and will prove to be crucial in tracking the ever-increasing variation of many other divergent RNA viruses in the future.

Key Words: SARS-CoV-2, Spike (S) Protein, Mutations, Geography, Climate

1. Introduction

Mutations in the viral genomes serve as the building blocks of viral evolution, and remain the main reason for the novelty in evolution (Baer, 2008; Duffy, 2018). In most cases, mutations are not beneficial for the organisms developing them, and lead to them having fewer descendants over time. Thus, a large portion of mutations, either at nucleotides (nt) and/or change in amino-acids (aa) levels, are harmful (Loewe and Hill, 2010). RNA viruses like SARS-CoV-2 generally have higher mutation rates; in them, however, these mutations are correlated with differential virulence, evolving ability, and traits considered beneficial for viruses (Duffy, 2018; Islam et al., 2020). SARS-CoV-2's inherently high mutation rate has already produced many descendants from the original Wuhan strain; this complicates its genotyping. The ability of the structural proteins (spike protein especially) in different strains of the SARS-CoV-2 to undergo rapid changes have enabled their genomes to emerge in novel hosts, escape vaccine-induced immunity, and evolve in diverse geo-climatic conditions (Duffy, 2018; Islam et al., 2020; Loewe and Hill, 2010). Moreover, spontaneous mutation is a key parameter in modelling the genetic structure, and evolution of populations (Drake and Holland, 1999). Therefore, investigation of the increased rate of synonymous mutations in the SARS-CoV-2 genomes could be an important tool in assessing the genetic health of populations.

SARS-CoV-2 comprises of four major structural proteins– specifically Spike (S) glycoproteins, envelope (E) proteins, membrane (M) proteins, and nucleocapsid (N) proteins (Ahmed et al., 2020; Rahman et al., 2020; Wu et al., 2020). The entry of SARS-CoV-2 into the host cells is mediated by the transmembrane S protein which consists of two functional subunits responsible for binding to the host cell receptor (S1 subunit), and for fusing the viral and cellular membranes (S2 subunit) (Walls et al., 2020). The higher antigenic and surface exposure properties of the S protein facilitate the attachment and entry of viral particles into the host cells through the host angiotensin-converting enzyme 2 (ACE2) receptor (Grant et al., 2020; Shang et al., 2020; Zhou et al., 2019). Therefore, the spike contains highest variations and determines, to some extent, the viral host range (Coutard et al., 2020; Wu et al., 2020). Furthermore, the S protein is the main target of neutralizing antibodies (Abs) upon infection, and is thus one of the most important structures for therapeutics and vaccine design (Rahman et al., 2020; Walls et al., 2020).

The continuing rapid transmission, and global spread of COVID-19 have raised intriguing questions regarding the evolution and adaptation of SARS-CoV-2 in diverse geographic and climatic conditions driven by synonymous mutations, deletions and/or replacements (Bal et al., 2020; Islam et al., 2020; Pachetti et al., 2020). The capability of the different strains of SARS-CoV-2 strains for swiftly adapting to diverse environments could be linked with their geographic distributions. Though not yet well-studied, evidence suggests that the transmission of SARS-CoV-2 infections and per day mortality rate from this infection is positively associated with weather conditions, and the diurnal temperature range (DTR) (Brassey et al., 2020; Su et al., 2016). However, the exact role of geo-climatic conditions on SARS-CoV-2 is unknown, but it would be worth keeping in mind that this novel disease originated from wildlife before spreading to humans (Harvey, 2020). Therefore, genomic mutation analysis of SARS-CoV-2 strains, integrated with geographic and climatic data, would provide a fuller understanding of the origin, dispersal and dynamics of the evolving SARS-CoV-2 virus. Although several reports predicted possible adaptations at the nucleotide and aa-level, along with structural heterogeneity in viral proteins, especially in the S protein (Armijos□Jaramillo et al., 2020; Islam et al., 2020; Phan, 2020; Sardar et al., 2020), most of these studies were carried out few complete representative genomes from a limited geographic area. As the genome number is increasing day by day, regular in-house monitoring of the crucial components such as the S protein is urgently necessary to understand the genomic basis and evolution of the diagnostic RT-PCR primer. There are a few pipelines (Yin, 2020) and websites (https://mendel.bii.astar.edu.sg/METHODS/corona/beta/MUTATIONS/hCoV19_Human_2019_WuhanWIV04/hCoV-19_Spike_new_mutations_table.html) in GSAID where aa change or substitution can be observed. In order to provide an alternative tool with a wider range of functions, we present an easy, rapid pipeline that will assist in the alignment of large volumes of viral genomes, remove low quality sequences and in-frame stop codons and provide in-house non-synonymous mutation analysis of large volumes of sequences while requiring minimal knowledge of the command line. This tool can perform this analysis for any other proteins as required. This study aimed to investigate the mutational spectra of aa utilizing this novel methodology in the S proteins in 35,750 complete genome sequences of the SARS-CoV-2 belonging to 135 countries and/regions, and five climatic zones around the world, retrieved from the global initiative on sharing all influenza data (GISAID) (https://www.gisaid.org/) up to June 24, 2020 (Supplementary Data 1).

2. Methodology

2.1 Genomic data collection, and processing

To decipher the genetic variations of the S glycoprotein, we retrieved 53,981 complete (or near-complete) genome sequences of SARS-CoV-2, available at the global initiative on sharing all influenza data (GISAID) (https://www.gisaid.org/) up to June 24, 2020. These sequences belonged to infected patients from 135 countries and/or regions from across six continents (Supplementary Data 1). Using pyfasta (https://github.com/brentp/pyfasta), we split the total genome into 6 separate files having around 8,900 sequences in each. We aligned each file through the MAFFT (maximum limit 10,000 sequences) online server (https://mafft.cbrc.jp/alignment/server/add_fragments.html?frommanual) using default parameters (Katoh et al., 2002). The complete genome sequence of SARS-CoV-2 Wuhan-Hu-1 strain (Accession NC_045512, Version NC_045512.2) was used as a reference genome.

2.2 Mutational frequency analysis

MEGA 7 was used to differentiate the spike protein of SARS-CoV-2 from multiple sequence alignment (Kumar et al., 2016). Sequence cleaner (https://github.com/metageni/Sequence-Cleaner) with set parameters of minimum length (m=3822), percentage N (mn=0), keep_all_duplicates, and remove_ambiguous was employed to remove all ambiguous, and low-quality sequences. We utilized SeqKit toolkit (seqkit grep -s -p "-" in.fa > out.fa) to apprehend gap containing strains for deletion analysis (Shen et al., 2016). Internal stop codon containing sequences were removed by using SEquence DAtaset builder (SEDA; https://www.sing-group.org/seda/). Amino-acid mutation analysis was done with bio-python program using pairwise alignment (https://github.com/SShaminur/Mutation-Analysis). The custom Venn diagrams (http://bioinformatics.psb.ugent.be/webtools/Venn/) server was used to make the Venn diagrams, and visualize the data. Swiss-Model, a structure homology-modelling server (https://swissmodel.expasy.org/) was used to predict the 3D structure (template, PDB ID:6VSB) of the S protein of the reference genome and the structure was visualized in PyMOL (DeLano, 2002; Rahman et al., 2020; Waterhouse et al., 2018). Furthermore, we divided the S glycoprotein mutation of SARS-CoV-2 data according to their geographic origins from six continents - Europe, Asia, North America, South America, Africa, and Australia, and five related climatic zones - temperate, tropical, diverse, dry and continental (Kissler et al.). To estimate the case fatality (mortality) rates of SARS-CoV-2 infections, we collected information on total infected cases, and total reported deaths in these countries from the World Health Organization (WHO) COVID-19 Reports up to June 12, 2020 (WHO Reports, 2020). Microsoft Excel 2016 was used for all the statistical analyses (David, 2017). Detailed step by step methods are described in Mutation_analysis.pdf (https://github.com/SShaminur/Mutation-Analysis).

3. Results and discussions

3.1 Genomic data collection and processing

Trimming low quality, ambiguous and non-human host RNA sequences resulted in 35,750 (66.23 %) cleaned and full length S protein sequences (Supplementary Data 1). These sequences belonged to 135 countries and/or regions of from six continents (Europe, Asia, North America, South America, Africa, and Australia), and five major climatic zones (temperate, tropical, diverse, dry and continental) around the world (Supplementary Data 1). European countries and/or regions had the highest percentage (58.90%) of S protein sequences, followed by North American (25.78%), Asian (9.34%), Australian (3.61%), South American (1.21%), and African (1.18%) countries or regions. On the other hand, the temperate climatic zone covered the majority of these S protein sequences (60.18%), followed by diverse (33.08%), continental (3.25%), tropical (2.81%), and dry (0.69%) climatic conditions (Supplementary Data 1). We selected the complete genome sequence SARS-CoV-2 Wuhan-Hu-1 strain (Accession NC_045512, Version NC_045512.2) as a reference genome. Through synonymous mutations analysis, we found 27,801 (77.77 %) mutated strains of the SARS-CoV-2 in the cleaned sequences (n= 35,750). Furthermore, country or region-specific aa change patterns revealed the highest number of mutated SARS-CoV-2 strains in England (7,067) followed by USA (6,501), Wales (3,002), Scotland (1,463), Netherlands (1,194), Australia (681), Belgium (596), and Denmark (582) (Supplementary Data 1).

3.2 Screening for mutational evolution throughout S protein

Our mutational analyses revealed a total of 988 unique amino acid (aa) change(s)/substitution(s) distributed across 660 unique positions in the S glycoprotein (Supplementary Data 2). The primary structure of the S-protein is 1274 aa, of them 51.81% aa positions (n=660) undergo aa-level evolution worldwide. We found eleven highly variable sites (Position: 32, 142, 146, 215, 261, 477, 529, 570, 622, 778, 791, 1146, 1162) showing four types of aa variations in a single position (Table 1). We also found that positions 52, 185 and 410 in the S glycoprotein had aa variation numbers of 3, 2 and 1, respectively (Fig. 1c, Table 1, Supplementary Data 2). Notably, position 614 showed two variants, substitution D614G (Aspartic acid □ Glycine) found in □74.82 % (n=26,749) of the cleaned sequences (□96.22% of the mutated sequences), and another variant D614N (Aspartic acid □ Asparagine) observed only in four strains from England and Wales (EPI_ISL_439400, EPI_ISL_443658 and EPI_ISL_445498, EPI_ISL_472913). The variant D614G in the S protein has overcome the wild-type variant from China since its first appearance in Germany on January 28, 2020 (Comandatore et al., 2020; Eaaswarkhanth et al., 2020; Kim et al., 2020; Trucchi et al., 2020).

A strain from Bosnia_and_Herzegovina (EPI_ISL_463893) had the highest number of aa changes/substitutions (n=6) at six positions (R246I, L276I, T430A, D614G, S750N, L922V) of S protein. Also, we found that 84.8 % (n=23,576) of the mutated sequences carried just a single aa mutation throughout the S proteins. The remaining 13.44 %, 1.63 %, 0.11 % and 0.01 % of the mutated sequences contained 2, 3, 4 and 5 aa changes, respectively (Fig. 1b, Supplementary Data 2). Moreover, no synonymous mutation was found in the full length S protein of 18 countries and/or regions including Anhui, Brunei, Cambodia, Changzhou, Chongqing, Foshan, Ganzhou, Guam, Hefei, Jiangxi, Jingzhou, Jiujiang, Lishui, Nepal, Philippines, Qatar, Yingtan, Yunnan. This indicates S protein homogeneity of these countries/regions with the reference sequence from Wuhan, China (Supplementary Data1).

The RBD region (Wrapp et., al 2020) (aa position: 338-530) showed nonsynonymous mutations at 82 different positions in 516 strains, whereas in the S1 site and S2 site, there were 362 and 297 positional mutations, respectively. Moreover, in the furin cleavage site (R685 and S686), we also observed a nonsynonymous mutation (S686G) in a single strain (Russia/Krasnodar-63401/2020|EPI_ISL_428867|2020-03-11) (Fig. 1a). We also found aa substitutions at six positions within the RBD region that are directly involved in binding with ACE-2 receptor (Wang et al., 2020; Yuan et al., 2020) including N439K (Scotland, Romania), L455F (England), A475V (USA, Australia), and F456L, Q493L and N501Y (USA) (Supplementary Data 2). All these mutations were found between March and April at a lower frequency (N439K with maximum frequency in 41 Scottish strains and one Romanian strain), except Q493L found in two USA strains reported in May. Q493R position showed variation in an English strain (EPI_ISL_470150) found in April. Furthermore, 18 substitutions at fourteen positions, previously reported to interact with anti-SARS-CoV-2 antibody (Yuan et al., 2020), were found in the strains from Bangladesh, England, Portugal, Wales, Shanghai, France, USA, Scotland, Russia, Latvia, Netherlands, South Africa, Bosnia and Herzegovina, Belgium, Bosnia and Australia (Supplementary Data 2) during the time frame March to May. Discontinuation of the mutants globally may be linked to reduction of virus pathogenicity and virulence fitness affecting transmission dynamics. However, the unavailability of these variants may result due to rejection of the variants with a lower ratio when generating the final consensus sequences as well as insufficient sequences reporting from unusual asymptomatic patients. Moreover, eight glycosylated sites of S protein underwent aa conversions including three substitutions in the NTD region (N17K, N74K, N149H), including a total five substitutions at four sites in the S1 region (N17K, N74K, N149H, N603S, N603K) and four mutations in the S2 region (N717T, N1074D, N1158S, N1194S) (Watanabe et al., 2020). Furthermore, a total of 50 aa substitutions within the S protein observed that incorporated asparagine (N) in S-protein of SARS-CoV-2 including seven within the RBD region (S359N, K378N, K417N, K458N, S477N, T523N and K529N) (Supplementary Data 2). These substitutions alter glycosylation sites and it nature, though it needs further investigations. Overall, the aa substitutions related to asparagine in the RBD (ACE binding domain) and/or in S1/2 domains nearer to the glycosylated sites may affect the glycosylation shield, folding of S protein, host-pathogen interactions, viral entry and finally immune modulation, thus affecting antibody recognition and viral pathogenicity (Ou et al., 2020; Watanabe et al., 2020).

3.3 Deletion analysis of SARS-CoV-2 S glycoprotein

Besides site-specific mutations, our analysis revealed 17 in-frame deletions of ranged nucleotides across the SARS-CoV-2 S protein sequences originating from different countries worldwide (Table 2, Supplementary Data 2). Notably, we considered the deletions that occurred in at least two strains at a certain position as deletions. All of the identified deletions distributed throughout the nucleotide sequence 200-2035 fall into four major regions of S protein i.e. nt-positon ranges 179-226 (61-76 aa: NVTWFHAIHVSGTNGT), 413-433 (138-144 aa:

DPFFLGVY), 724-732 (241-244: LLAL) and 2021-2035 (675-679 aa: QTQTN). Amino acid deletions at positions 61-76, 138-144, and 241-244 are near the RBD region. Among them, deletions of positions 61-76 and 141-144 are surface exposed, but 241-244 are situated at the inner surface of the predicted S protein (Fig. 2). Also, deleted aa at positions 675-679 are located in the C-terminal transmembrane domain of S protein. Surface exposed deletions near the RBD region may have significant impact on host-pathogen interaction and immune modulation.

Among the deletions, nucleotide deletion positioned at 418-433 (aa position 140-144) faced frequent overlapped deletions among strains of multiple countries (Table 2). Notably, a single aa in-frame deletion of nucleotides positioned 429-431 (aa position 145) with the highest frequency in 48 strains from multiple countries and/or regions including Australia, England, Canada, Slovenia, Jordan, Netherlands, Saudi_Arabia, Scotland, USA, Spain, Wales and India. A strain from Taiwan (EPI_ISL_444275) showed two coevolving deletions at nt positions 200-226 (68-76 aa:IHVSGTNGT) and nt positions 2021-2035 (675-679 aa:QTQTN). Moreover, two deletions at nt positions 418-420 (140 aa:F) and 727-732 (243-244 aa:AL) were coevolved in a Sichuan strain (EPI_ISL_451369). No other strain had such coevolving deletions, thereby indirectly indicating the negative impact of the deletions on virus fitness and human to human transmissibility. Noteworthy, a 5-aa deletion (675-679 aa: QTQTN) at the upstream of the polybasic cleavage site of S1-S2, and a 21-nt deletion 23596–23617 (aa- NSPRRAR) including the polybasic cleavage site in clinical samples and cell-isolated virus strain likely benefit the SARS-CoV-2 replication or infection in vitro and under strong purification selection in vivo (Liu et al., 2020). Moreover, attenuated SARS-CoV-2 variants with 15-30-bp deletions (Del-mut) at the S1/S2 junction were reported to show less virulence in an animal model (Lau et al., 2020).

These deletions may affect viral adaptations to human, virus-host interactions for infections, attenuation, pathogenicity, and immune-modulations by potentially influencing the tertiary structures and functions of the associated proteins (Phan, 2020). However, further studies are required for the mechanistic clarification and functional implication of these deletions in the SARS- CoV-2 S glycoprotein. The deletion mutations identified in this study should be also considered for current vaccine development.

3.4 Geo-climatic scenario of amino-acid changes in the spike protein of SARS-CoV-2, and associated disease severity

Considering geo-climatic impacts on aa changes in the S protein of the SARS-CoV-2, we sought to determine the possible residue positions, and total number of mutations in the S protein gene sequences from 135 countries and/or territories and five climatic zones worldwide. Eight hundred and eighty-eight (988) unique aa replacements across 660 positions along the S protein were identified which differed significantly (p= 0.003, Kruskal–Wallis test) among the genomes of SARS-CoV-2. We found that the frequency of aa changes in the S protein remained substantially higher in the SARS-CoV-2 genome sequences of Europe (62.02%), followed by North America (25.50%), Asia (6.83%), Australia (2.89%), South America (1.41%), and Africa (1.35%) (Supplementary Data 1). Among these replacements, only one aa residue at position 5 (L5F) and 614 (D614G) were found to be the common in Asia, Europe, North America, South America, Africa, and Australia (Fig. 2a). Moreover, 408, 127, 139, 17, 10, and 8 unique aa replacements, and 244, 146, 194, 61, 19, and 23 accessory aa replacements (mutations shared with at least two continents) were found in the SARS-CoV-2 genomes sequenced from Europe, Asia, North America, Australia, South America, and Africa, respectively (Fig. 3a, Supplementary Data 3). Higher unique mutations in European, Asian and American sequences point out the geographical clustering predisposition of the virus. However, further phylogenic study targeting those unique and accessory mutations may lead to a better understanding of global phylodynamics, and thereby guiding the regional control strategy for the COVID-19 pandemic.

This study also explores the non-synonymous mutations in the S protein of the SARS-CoV-2 genomes across five different climatic conditions worldwide. This revealed significant (p= 0.017, Kruskal–Wallis test) variations in mutation patterns. Our analysis showed that only two core aa substitutions at positions 614 (D614G) and 936 (D936Y) were shared across all the climatic zones (Fig. 3b). Similarly, 426, 231, 29, 29, and 1 unique aa replacement were found in the S protein sequences of the temperate, diverse, tropical, continental and dry climatic conditions, respectively. Moreover, 252, 239, 47, 76, and 14 residue positions in the S protein sequences were identified where nonsynonymous mutations occurred in at least two climatic zones (Fig. 3b, Supplementary Data 3). RNA viruses like SARS-CoV-2 might have remarkable capabilities to adapt to new environments, and confront different selective pressures they encounter (Watanabe et al., 2020).

The genomic variability of SARS-CoV-2 strains manifested by mutations in the spike protein scattered across the globe underly geographically specific etiological effects. One important effect of mapping mutations is the development of antiviral therapies targeting specific regions, for example the spike region of the SARS-CoV-2 genomes (Callaway, 2020). Our current findings corroborate the study completed by Deshwal (2020), who reported the highest SARS-CoV-2 infections and case fatality rates in European countries. In another study, Pachetti et al. (2020) reported two non-synonymous mutations (R203K and L3606F) that were shared across ORFs of the SARS-CoV-2 genomes of six continents, and co-occurrence mutations were also common in different countries along with unique mutations. Nevertheless, mutations in the structural proteins of the SARS-CoV-2, especially in the spike proteins, are driven by the geographic locations that diverged differently, possibly due to the environment, demography, and the low fidelity of reverse transcriptase (Brassey et al., 2020; Pachetti et al., 2020; Su et al., 2016).

Investigating the continental and/or regional impacts of aa substitutions in the SARS-CoV-2 genomes, we found higher case fatality rates in temperate European countries such as United Kingdom (14.16%), Italy (11.72%), France (10.05%), Spain (9.31%), Belgium (3.30%), Germany (3.00%), Russia (2.30%), Netherlands (2.07%), Sweden (1.65%) and Turkey (1.63%) (Supplementary Data 3). Among the tropical Asian countries, higher mortality rates from SARS-CoV-2 infections were estimated in Iran (4.76%), India (4.72%), China (2.56%), Pakistan (1.38%), and Indonesia (1.11%), and rest of the countries had less than 1.0% case fatality rates. Moreover, in the diverse climatic conditions of the American countries or territories (both North and South Americans), United States of America (5.67%), and Brazil (5.14%) had relatively higher mortality rates from SARS-CoV-2 pandemics, and rest of the countries in these continents had substantially lower disease severity rates (< 1.0%). Case fatality or mortality rates from SARS-CoV-2 infections in rest of the two continents (Africa and Australia) remained much lower, and only 2.19%, 1.40%, and 1.26% death rates were found in South Africa, Australia, and Algeria, respectively. The rest of the countries and/or territories of these two continents had less than 1.0% mortality rates (Supplementary Data 3).

The predominantly higher mortality rates in European temperate countries might be correlated with higher unique mutations found in the S proteins reported from this climate Thus, our present study revealed that the predicted rates of unique aa changes in the European sequences could be associated with higher pathogenicity of the virus. However, it is worth noting that reported disease severity (may not represent the actual severity) might be affected by several other factors like health care facilities, average age group and genetic context of the population and control strategies adopted by the countries. Irrespective of the significance of geography for emerging infectious disease epidemiology, the effects of global mobility upon the genetic diversity and molecular evolution of SARS-CoV-2 are under-appreciated and only beginning to be understood. The recent monograph on the spatial epidemiology of COVID-19 makes no reference to the genetic disparity of SARS-CoV-2 (Brassey et al., 2020; Harvey, 2020; Pachetti et al., 2020; Su et al., 2020).

3.5 Pipeline validations

The SARS-CoV-2 genomes are increasing very rapidly in the Global initiative on sharing all influenza data (GISAID), but not all genomes are of high quality or complete. So, nonsynonymous mutation analysis with particular crucial part of the virus like S or other structural protein gives statistically more significant insights rather considering the complete genome of the SARS-CoV-2 virus. In this study, we found 33.77 % (18,231/53,981) sequences are in low quality or having ambiguous sequences. Sequence cleaner has removed those sequences and give us cleaned sequences. Among them, we found ten in frame stop codon containing sequences and we have removed it using SEDA. SeqKit toolkit were used to arrest gap containing sequences, and we found around 453 sequences from there, we have to carefully checked the in-frame deletion and 103 strains contains in frame deletions. SNP-sites is a very efficient tools for nucleotide variation detection in different format like multi-fasta alignment, variant call format (VCF), and relaxed phylip format (Page et al., 2016; Seemann, 2015) but this tool is highly dedicated for nucleotide. Snippy (Seemann, 2015) is another tool where nucleotide and protein variation can also be detected, but for large data set with ambiguous sequences will require a separate processing to entrust more accurate results. Here we will get the nonsynonymous mutation that alter aa results in a file format (Sequence_ID Reference_amino_acid:Mutation_Position:Strain_amino_acid) that will assist in the downstream analysis like unique mutation, unique position mutation, mutational frequency, strains having number of mutation. For deletion analysis, this pipeline helps in decreasing the size of sequences (just 453 sequences from 53,981 sequences) for deletion analysis.

4. Conclusions

Analyses of genome sequences of 30,493 SARS-CoV-2 strains from across 135 countries and/or regions, and five climatic conditions worldwide revealed the presence of synonymous and non-synonymous mutations, deletions and/or replacements at different positions of the S protein gene, which was reflected in the S-protein primary sequence. These findings of previously unreported mutations in the spike protein of SARS-CoV-2 genomes suggest that the virus is evolving, and European, North American and Asian strains might coexist, each of them characterized by a different mutation patterns, and associated case fatality rates. Moreover, the geo-climatic distribution of the mutations in the spike deciphered higher mutations rates as well as disease severity in the European temperate countries. Furthermore, the structural validations of the mutations in the reference genomes of Wuhan-Hu-1 strain further validated the results of our current study. However, there is no experimental evidence to suggest a difference in aggressiveness of such mutations amongst the studied genome sequences. Moreover, the geo-climate effects of the observed mutations in the spike protein of SARS-CoV-2 on the properties of the diverse strain variants are yet to be evaluated in clinical or experimental studies.

Therefore, these results need to be interpreted cautiously given the existing uncertainty about SARS-CoV-2 genomic data to develop potential prophylaxis and mitigation for tackling the pandemic COVID-19 crisis. So, the pipeline developed will help in the easy and accurate way investigate the nonsynonymous mutation, frequency, deletion analysis from large number of data with a shortest possible time without having knowledge of much bioinformatics.

Figures

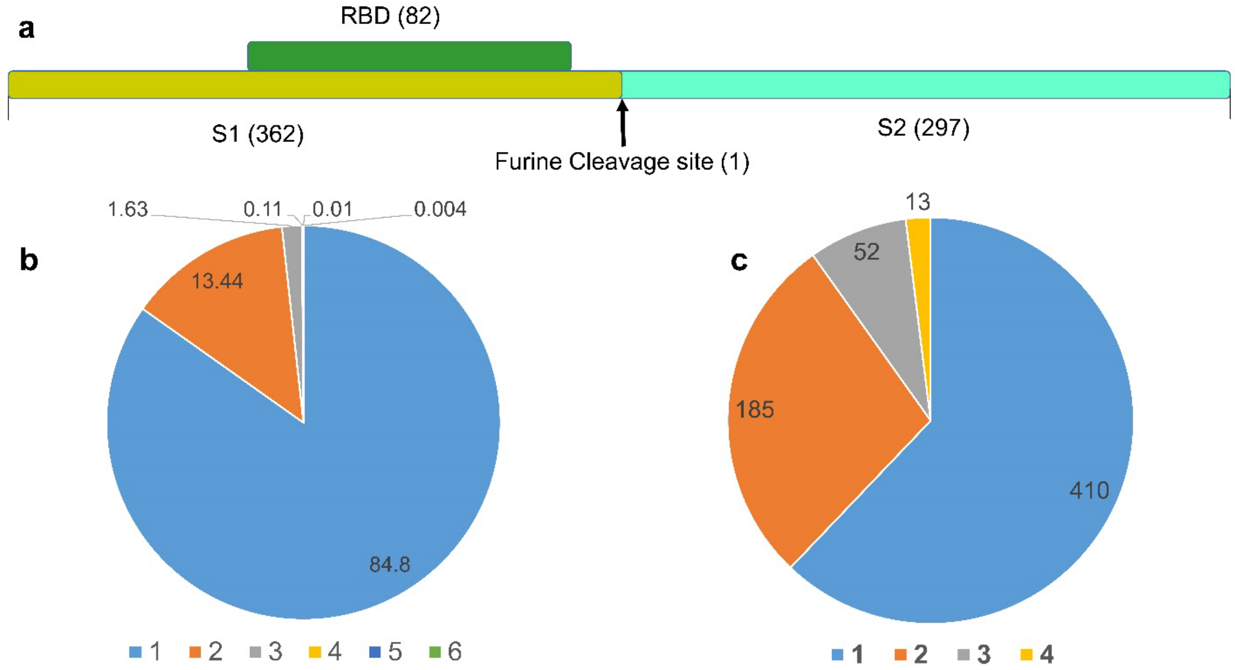


Fig. 1: Mutational frequency and distribution of S glycoprotein of SARS-CoV-2. (a) Represents different structural regions in spike protein where aa mutations occurred worldwide. The receptor binding domain (RBD) had 82 positions where aa mutations were found whereas the S1 and S2 subunits have 362 and 297 positions for aa mutation, respectively. The furin cleavage site (R685, S686) also possessed one mutation (S686G) in of the Russian SARS-CoV-2 strains (EPI_ISL_428867). **(b)** Denotes the number of mutations in different strains of SARS-CoV-2 where 1, 2, 3, 4, 5 and 6 codes for one, two, three, four, five and six different types of aa mutations across the studied strains. In this study, most of the strains (84.40%) had single aa variation while 13.44%, 1.63%, 0.11% and 0.01% sequences harbored 2, 3, 4 and 5 aa mutations, respectively. **(c)** Positional aa variations in S protein of SARS-CoV-2 where 1, 2, 3 and 4

represent the aa variation in one, two, three and four different positions. 13 positions in the protein were found to having 4 types of aa variations, and 52, 185 and 410 positions in the spike undergone to three, two and one type of aa variations, respectively.

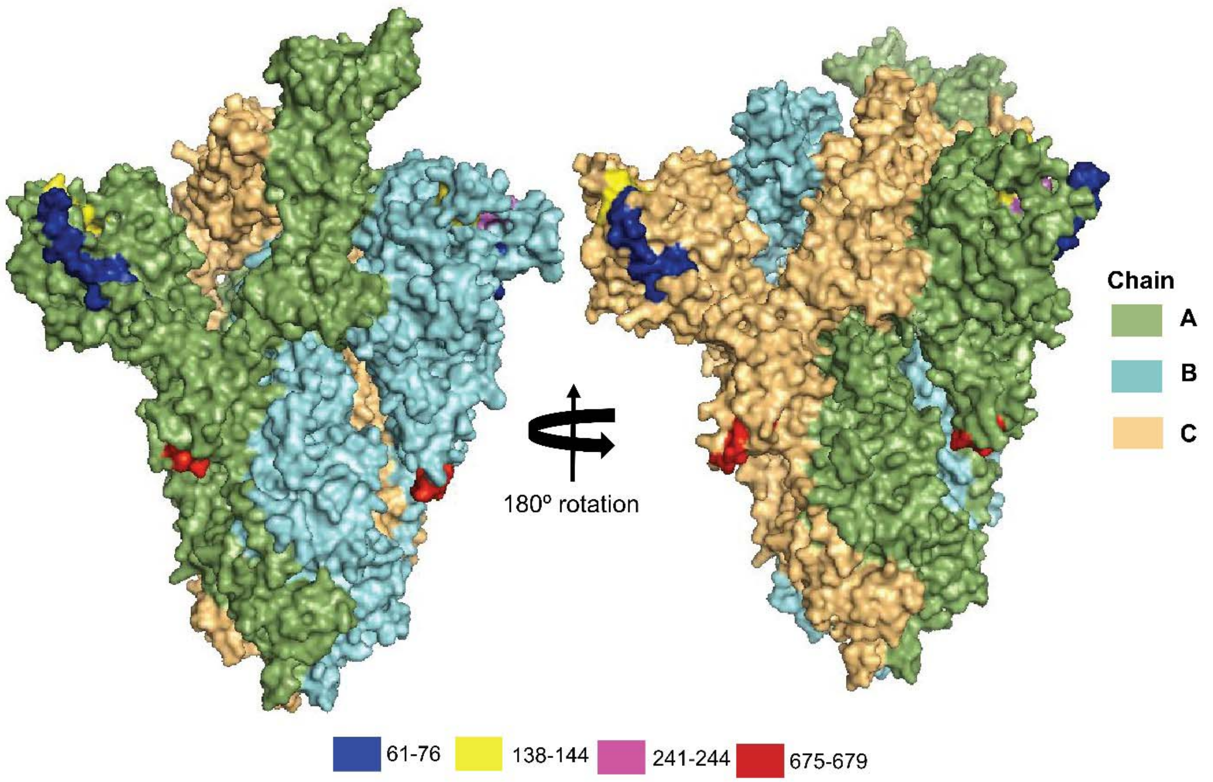


Fig. 2: The four amino acids deleted positions (61-76, 138-144, 241-244, and 675-679) in the spike (S) protein of the reference genome, SARS-CoV-2 Wuhan-Hu-1 strain (Accession NC_045512, Version NC_045512.2). The positions are visualized in the tertiary (3D) structure of S protein in PyMOl.

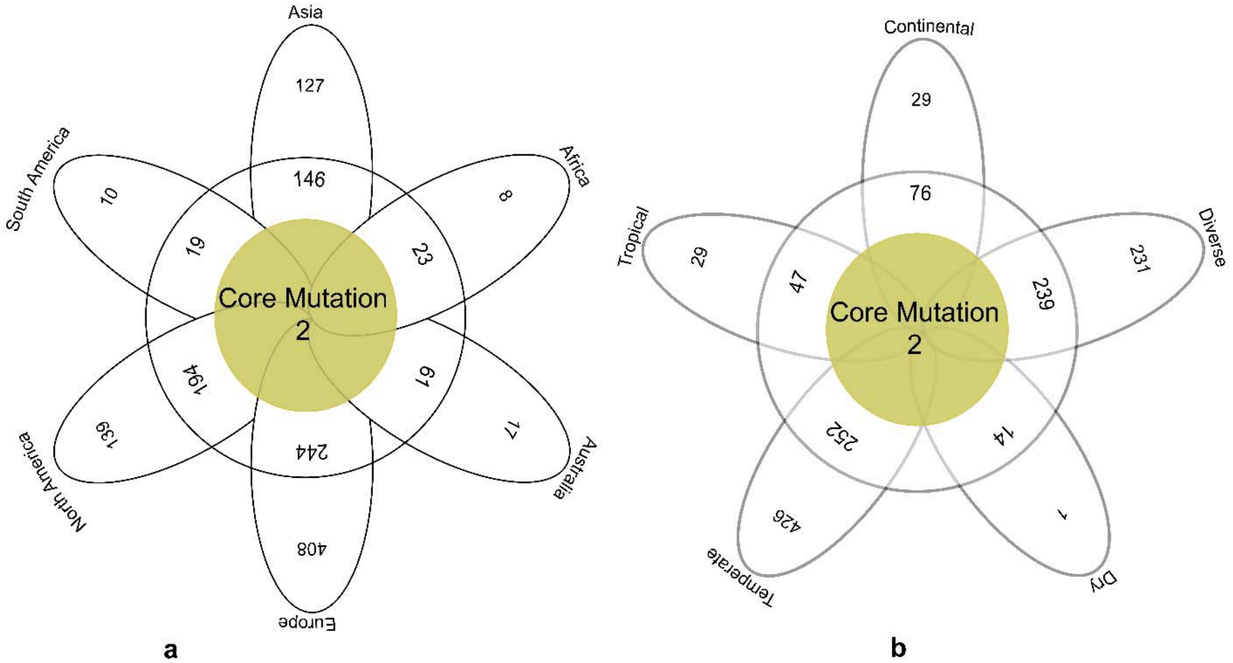


Fig. 3: The frequency spectra of amino-acid mutations in the spike protein of SARS-CoV-2. Amino-acid (aa) mutations are represented according to **(a)** geographic areas and **(b)** different climate zones. We found two core shared aa mutation at residue position 5 (L5F) and 614 (D614G) in Asia, Europe, Africa, Australia, North America, and South America, and two core shared mutations at residue positions of 614 (D614G), and 936 (D936Y) in continental, diverse, dry, tropical and temperate climatic conditions. In both cases (a and b), the middle brown circles represent frequency of aa substitutions shared by all variables, and the frequency of aa substitutions shared by at least two continents/climate zones are shown in white circle. The white colored outer ribbons represent unique aa mutations in each individual region and climate zone.

Table 1 Amino acid variations of S glycoprotein according to their positions. Here, the position where variation more than two aa variations found are represented.

Positon in S	No. of variations	Name of Amino Acid	Position in S	No. of variations	Name of Amino Acid
32	4	F32L, F32Y, F32I, F32V	273	3	R273M, R273K, R273S
142	4	G142D, G142A, G142V, G142S	354	3	N354D, N354K, N354S
146	4	H146Q, H146N, H146Y, H146R	414	3	Q414R, Q414K, Q414P
215	4	D215Y, D215H, D215G, D215N	468	3	I468F, I468T, I468V
261	4	G261V, G261S, G261D, G261R	483	3	V483F, V483I, V483A
477	4	S477I, S477N, S477R, S477G	558	3	K558N, K558Q, K558R
529	4	K529M, K529N, K529R, K529E	615	3	V615I, V615F, V615L
570	4	A570S, A570V, A570D, A570T	654	3	E654D, E654Q, E654K
622	4	V622F, V622L, V622I, V622A, A623V	675	3	Q675H, Q675R, Q675K
778	4	T778S, T778A, T778N, T778I	677	3	Q677H, Q677R, Q677Y
791	4	T791I, T791A, T791K, T791P	681	3	P681H, P681L, P681S
1146	4	D1146Y, D1146H, D1146E, D1146N	684	3	A684V, A684T, A684S
1162	4	P1162L, P1162T, P1162A, P1162S	747	3	T747A, T747I, T747N
19	3	T19P, T19I, T19S	750	3	S750N, S750R, S750I
21	3	R21I, R21T, R21K	752	3	L752I, L752R, L752F
22	3	T22N, T22I, T22A	765	3	R765L, R765H, R765C
26	3	P26L, P26S, P26R	772	3	V772L, V772I, V772A
27	3	A27V, A27T, A27S	780	3	E780D, E780Q, E780V
72	3	G72E, G72W, G72R	812	3	P812S, P812T, P812L
75	3	G75D, G75V, G75R	831	3	A831S, A831V, A831T
80	3	D80N, D80Y, D80A	836	3	Q836H, Q836P, Q836L

97	3	K97E, K97N, K97R	838	3	G838S, G838V, G838D
102	3	R102S, R102I, R102G	839	3	D839Y, D839E, D839N
148	3	N148Y, N148K, N148S	845	3	A845S, A845V, A845D
153	3	M153T, M153I, M153V	847	3	R847T, R847I, R847K
183	3	Q183H, Q183R, Q183L	870	3	I870S, I870T, I870V
218	3	Q218R, Q218E, Q218L	879	3	A879S, A879V, A879T
222	3	A222V, A222S, A222P	930	3	A930S, A930V, A930T
239	3	Q239K, Q239R, Q239H	1085	3	G1085R, G1085E, G1085L
246	3	R246I, R246K, R246S	1129	3	V1129L, V1129A, V1129I
247	3	S247R, S247I, S247N	1153	3	D1153A, D1153H, D1153Y
251	3	P251S, P251H, P251L	1170	3	S1170T, S1170Y, S1170P
263	3	A263T, A263S, A263V			

Table 2 Deletion-sites observed across the S glycoprotein. Countries represent the origin of strains where the deletions found. We considered the deletions that occurred in at least two strains in a certain position.

Nucleotide positions	Amino acid positions	Deleted amino acid	Countries	No. of strains
179-217	61-73	NVTWFHAIHVSGT	England	1
200-226	68-76	IHVSGTNGT	Taiwan, Malaysia	2
201-224	68-75	IHVSGTNG	Thailand	1
203-208	69-70	HV	Sweden, England, Australia	3
413-421	138-140	DPF	Sweden	1
418-420	140	F	England, Sichuan	3
420-431	141-144	LGVY	England, Iceland, USA, Scotland, Kenya	16
420-422	141	L	England	1
422-430	141-143	LGV	Portugal, England, Iceland, Scotland	4
423-431	142-144	GVY	England, Netherlands	3
428-430	143	V	USA, Belgium	4
428-433	143-144	VY	England	2
429-431	145	Y	England, Canada, Slovenia, Jordan, Netherlands, Saudi_Arabia, Scotland, USA, Spain, Wales, India, Australia	48
724-732	241-243	LLA	China, England, Belgium, Scotland, Netherlands	6
724-726	241	L	USA	2
727-732	243-244	AL	England, Wales, Spain, Sichuan	6
2021-2035	675-679	QTQTN	Taiwan, Malaysia	2

Conflicts of Interest Statement

The authors of this manuscript declare that they have no conflict of interest.

Acknowledgements

The authors sincerely appreciate the researchers worldwide who had deposited and shared the complete genomes data of SARS-CoV-2 and other coronaviruses to GISAID (https://www.gisaid.org/). This research utilized these precious data. The authors would also like to extend thanks to Geni Gueiros who was kind to modify his tools (Sequence cleaner) upon request from Md. Shaminur Rahman.

Data availability

This study utilized the SARS-CoV-2 genome sequences retrieving from the publicly available open database, GISAID. Detailed step by step methods are described in Mutation_analysis.pdf (https://github.com/SShaminur/Mutation-Analysis).

Author contributions

MSR, MRI, MNH, ASMRUA, MA, JA, and SA conducted the overall study. MSR, MRI, and MNH drafted the manuscript. MNH finally compiled the manuscript. AA, MS, and MAH contributed intellectually to the interpretation and presentation of the results.

Supplementary Information

Supplementary information supporting the findings of this study are available in this article as Supplementary Files, or from the corresponding author on request.

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, 254.

Armijos□Jaramillo, V., Yeager, J., Muslin, C., Perez□Castillo, Y., 2020. SARS□CoV□2, an evolutionary perspective of interaction with human ACE2 reveals undiscovered amino acids necessary for complex stability. Evolutionary Applications, DOI: 10.1101/2020.03.21.001933.

Baer, C.F., 2008. Does mutation rate depend on itself. PLoS Biology, 6, e52.

Bal, A., Destras, G., Gaymard, A., Bouscambert-Duchamp, M., Valette, M., Escuret, V., Frobert, E., Billaud, G., Trouillet-Assant, S., Cheynet, V., 2020. Molecular characterization of SARS-CoV-2 in the first COVID-19 cluster in France reveals an amino acid deletion in nsp2 (Asp268del). Clinical Microbiology and Infection, 26(7), 960–962.

Brassey, J., Heneghan, C., Mahtani, K. R.& Aronson, J. K. 2020. Do weather conditions influence the transmission of the coronavirus (SARS-CoV-2)? Centre for Evidence-Based Medicine, Nuffield Department of Primary Care Health Sciences, University of Oxford, March 22, 2020.

Callaway, E., 2020. Coronavirus vaccines: five key questions as trials begin. Nature 579, 481.

Comandatore, F., Chiodi, A., Gabrieli, P., Biffignandi, G.B., Perini, M., Ramazzotti, M., Ricagno, S., Rimoldi, S.G., Gismondo, M., Micheli, V., 2020. Identification of variable sites in Sars-CoV-2 and their abundance profiles in time. bioRxiv.

Coutard, B., Valle, C., de Lamballerie, X., Canard, B., Seidah, N., Decroly, E., 2020. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Research, 176, 104742.

David, M., 2017. Statistics for managers, using Microsoft excel. Pearson Education India.

DeLano, W.L., 2002. The PyMOL molecular graphics system. http://www. pymol. org.

Drake, J.W., Holland, J.J., 1999. Mutation rates among RNA viruses. Proceedings of the National Academy of Sciences 96, 13910-13913.

Duffy, S., 2018. Why are RNA virus mutation rates so damn high? PLoS biology 16, e3000003.

Eaaswarkhanth, M., Al Madhoun, A., Al-Mulla, F., 2020. Could the D614 G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality? International Journal of Infectious Diseases, 96, 459-460.

Grant, O.C., Montgomery, D., Ito, K., Woods, R.J., 2020. 3D Models of glycosylated SARS-CoV-2 spike protein suggest challenges and opportunities for vaccine development. bioRxiv. doi: https://doi.org/10.1101/2020.04.07.030445.

Harvey, C. What Could Warming Mean for Pathogens like Coronavirus? E&E News, March 9, (2020).

Islam, M.R., Hoque, M.N., Rahman, M.S., Puspo, J.A., Akhter, M., Akter, S., Rubayet-Ul-Alam, A., Sultana, M., Crandall, K.A., Hossain, M.A., 2020. Genome Wide Analysis of Severe Acute Respiratory Syndrome Coronavirus-2 Implicates World-Wide Circulatory Virus Strains Heterogeneity. Preprints 2020040137. doi: 10.20944/preprints202004.0137.v1.

Katoh, K., Misawa, K., Kuma, K.i., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research 30, 3059-3066.

Kim, J.-S., Jang, J.-H., Kim, J.-M., Chung, Y.-S., Yoo, C.-K., Han, M.-G., 2020. Genome-Wide Identification and Characterization of Point Mutations in the SARS-CoV-2 Genome. Osong Public Health and Research Perspectives 11, 101.

Kissler, S.M., Tedijanto, C., Goldstein, E., Yonatan, H., Grad, and Marc Lipsitch. 2020.‘Projecting the Transmission Dynamics of SARS-CoV-2 through the Postpandemic Period’. Science, 368 (6493), 860-868.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular Biology and Evolution 33, 1870-1874.

Lau, S.-Y., Wang, P., Mok, B.W.-Y., Zhang, A.J., Chu, H., Lee, A.C.-Y., Deng, S., Chen, P., Chan, K.-H., Song, W., 2020. Attenuated SARS-CoV-2 variants with deletions at the S1/S2 junction. Emerging Microbes & Infections 9, 837-842.

Liu, Z., Zheng, H., Yuan, R., Li, M., Lin, H., Peng, J., Xiong, Q., Sun, J., Li, B., Wu, J., 2020. Identification of a common deletion in the spike protein of SARS-CoV-2. bioRxiv.

Loewe, L., Hill, W.G., 2010. The population genetics of mutations: good, bad and indifferent. The Royal Society, 365(1544), 1153–1167.

Ou, X., Liu, Y., Lei, X., Li, P., Mi, D., Ren, L., Guo, L., Guo, R., Chen, T., Hu, J., 2020. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nature Communications, 11, 1-12.

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

Page, A.J., Taylor, B., Delaney, A.J., Soares, J., Seemann, T., Keane, J.A., Harris, S.R., 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial Genomics 2, 2(4), e000056.

Phan, T., 2020. Genetic diversity and evolution of SARS-CoV-2. Infection, Genetics and Evolution 81, 104260.

Rahman, M.S., Hoque, M.N., Islam, M.R., Akter, S., Rubayet-Ul-Alam, A., Siddique, M.A., Saha, O., Rahaman, M.M., Sultana, M., Hossain, M.A., 2020. Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2 etiologic agent of global pandemic COVID-19: an in silico approach. bioRxiv. doi: https://doi.org/10.1101/2020.03.30.015164.

Sardar, R., Satish, D., Birla, S., Gupta, D., 2020. Comparative analyses of SAR-CoV2 genomes from different geographical locations and other coronavirus family genomes reveals unique features potentially consequential to host-virus interaction and pathogenesis. bioRxiv. Seemann, T., 2015. Snippy: rapid haploid variant calling and core SNP phylogeny. Available.Shang, W., Yang, Y., Rao, Y., Rao, X., 2020. The outbreak of SARS-CoV-2 pneumonia calls for viral vaccines. npj Vaccines 5, 1-3.

Shen, W., Le, S., Li, Y., Hu, F., 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PloS One 11, e0163962.Su, S., Wong, G., Shi, W., Liu, J., Lai, A.C., Zhou, J., Liu, W., Bi, Y., Gao, G.F., 2016. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends in Microbiology 24, 490-502.

Trucchi, E., Gratton, P., Mafessoni, F., Motta, S., Cicconardi, F., Bertorelle, G., D'Annessa, I., Di Marino, D., 2020. Unveiling diffusion pattern and structural impact of the most invasive SARS-CoV-2 spike mutation. bioRxiv.

Walls, AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. 2020. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell, 181, 281-292.e6.

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., Lu, G., Qiao, C., Hu, Y., Yuen, K.-Y., 2020. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell, 181(4), 894-904.e9.

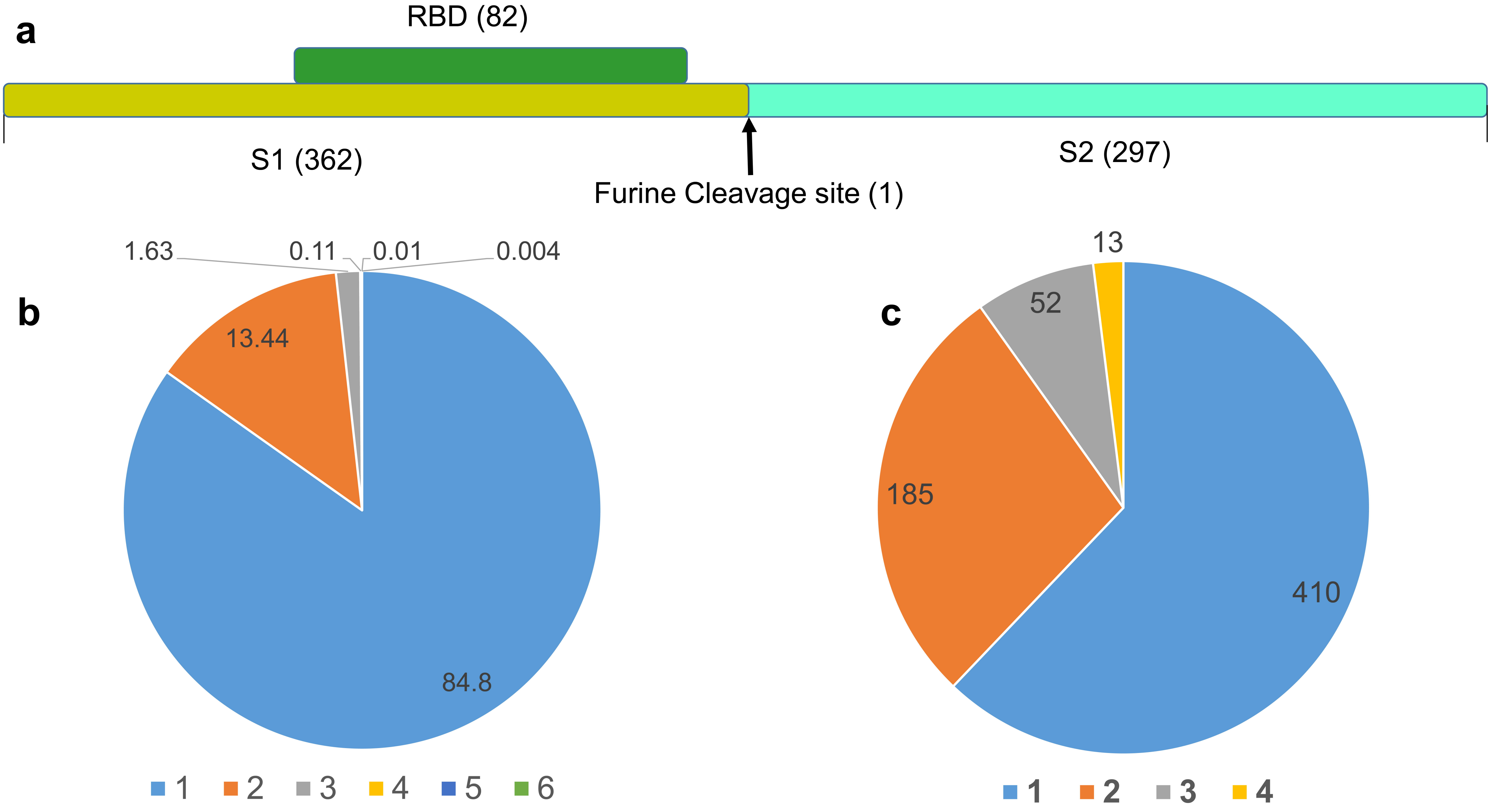
Watanabe, Y., Allen, J.D., Wrapp, D., McLellan, J.S., Crispin, M., 2020. Site-specific glycan analysis of the SARS-CoV-2 spike. Science, eabb9983.

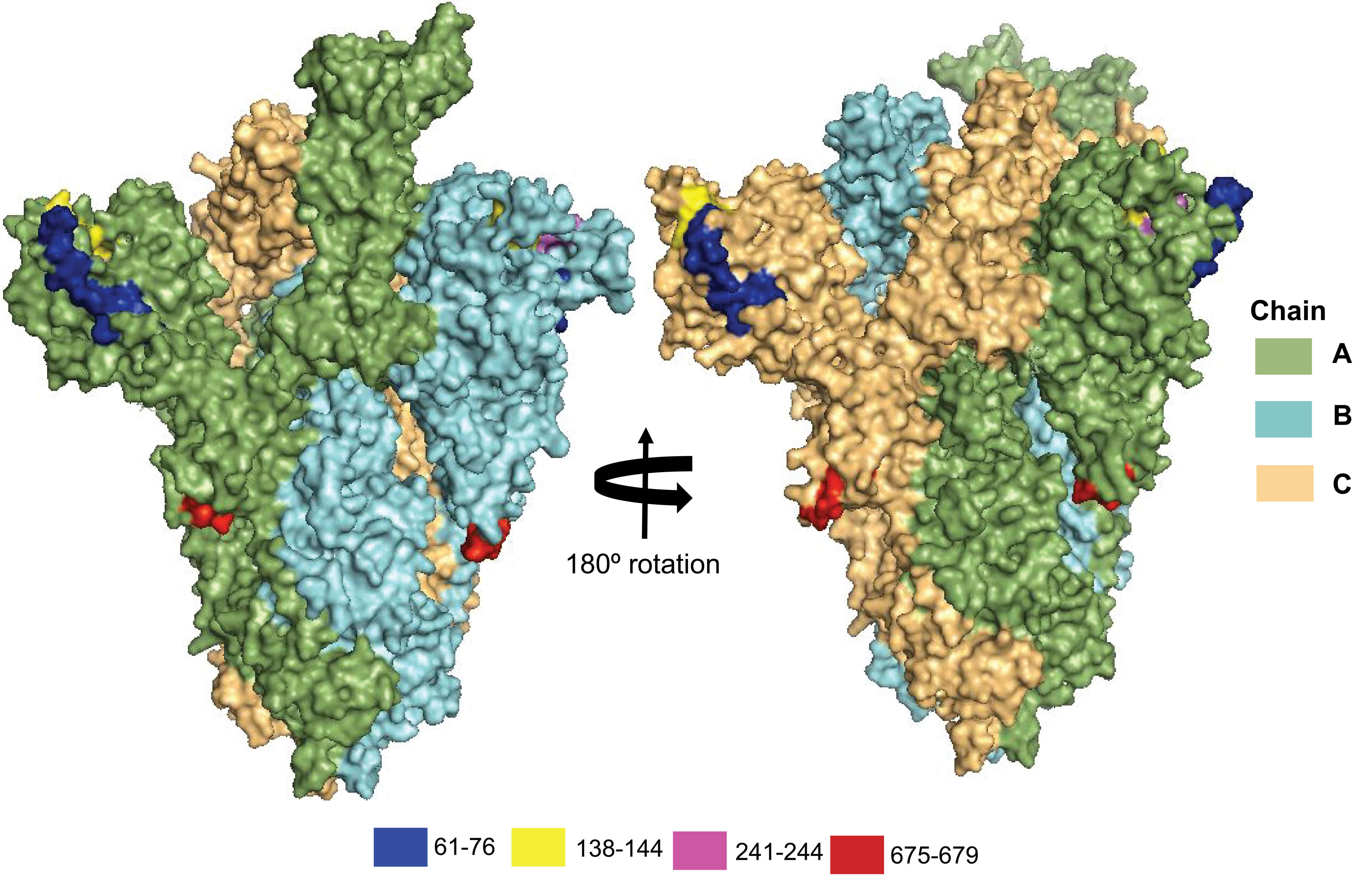
Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., 2018. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Research, 46, W296-W303.

Wu, C., Liu, Y., Yang, Y., Zhang, P., Zhong, W., Wang, Y., Wang, Q., Xu, Y., Li, M., Li, X., 2020. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. Acta Pharmaceutica Sinica B, 10(5), 766-788.Yin, C., 2020. Genotyping coronavirus SARS-CoV-2: methods and implications. Genomics, https://doi.org/10.1016/j.ygeno.2020.04.016.

Yuan, M., Wu, N.C., Zhu, X., Lee, C.-C.D., So, R.T., Lv, H., Mok, C.K., Wilson, I.A., 2020. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. Science, 368, 630-633.

Zhou, H., Chen, Y., Zhang, S., Niu, P., Qin, K., Jia, W., Huang, B., Zhang, S., Lan, J., Zhang, L., Tan, W. 2019. Structural definition of a neutralization epitope on the N-terminal domain of MERS-CoV spike glycoprotein. Nature Communications, 10, 1-13.


a
RBD (82)
S1 (362)
S2 (297)
Furine Cleavage site (1)
b
1.63
0.11
0.01
0.004
13.44
84.8
■ 1 ■ 2 ■ 3 ■ 4 ■ 5 ■ 6
c
13
52
185
410
■ 1 ■ 2 ■ 3 ■ 4


Chain
A
B
C
180º rotation
61-76
138-144
241-244
675-679

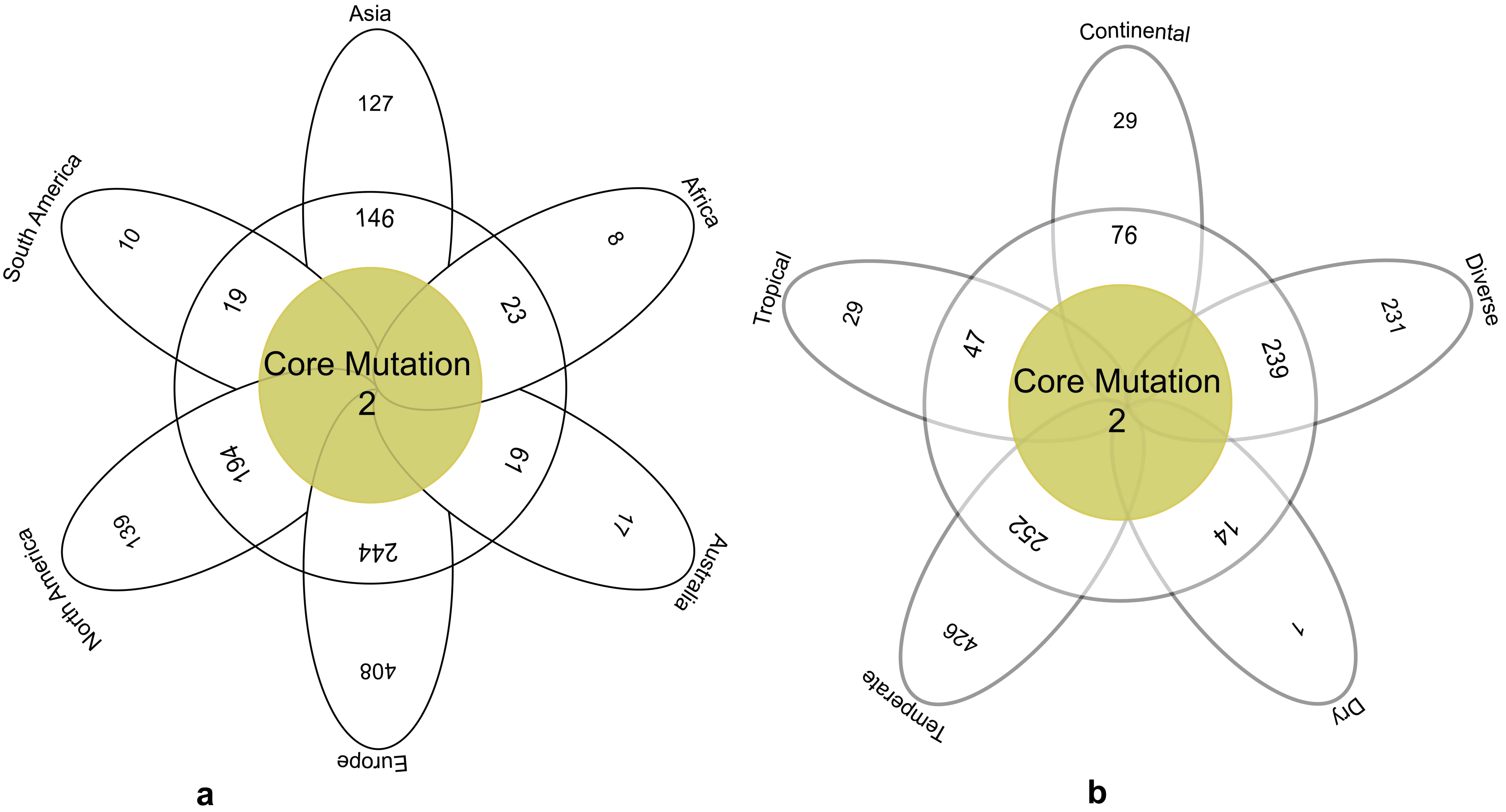

Asia
127
146
South America
10
19
Africa
8
23
Core Mutation
2
North America
139
194
Australia
17
61
Europe
408
244
a
Continental
29
76
Tropical
29
47
Diverse
231
239
Core Mutation
2
Temperate
426
252
Dry
1
14
b

Table 1: Amino Acid variation of S glycoprotein according to their position. Here, the position where variation more than 2 are represented.

Positon In S	Number of Variation	Name of Amino Acid	Positon In S	Number of Variation	Name of Amino Acid
32	4	F32L, F32Y, F32I, F32V	273	3	R273M, R273K, R273S
142	4	G142D, G142A, G142V, G142S	354	3	N354D, N354K, N354S
146	4	H146Q, H146N, H146Y, H146R	414	3	Q414R, Q414K, Q414P
215	4	D215Y, D215H, D215G, D215N	468	3	I468F, I468T, I468V
261	4	G261V, G261S, G261D, G261R	483	3	V483F, V483I, V483A
477	4	S477I, S477N, S477R, S477G	558	3	K558N, K558Q, K558R
529	4	K529M, K529N, K529R, K529E	615	3	V615I, V615F, V615L
570	4	A570S, A570V, A570D, A570T	654	3	E654D, E654Q, E654K
622	4	V622F, V622L, V622I, V622A, A623V	675	3	Q675H, Q675R, Q675K
778	4	T778S, T778A, T778N, T778I	677	3	Q677H, Q677R, Q677Y
791	4	T791I, T791A, T791K, T791P	681	3	P681H, P681L, P681S
1146	4	D1146Y, D1146H, D1146E, D1146N	684	3	A684V, A684T, A684S
1162	4	P1162L, P1162T, P1162A, P1162S	747	3	T747A, T747I, T747N
19	3	T19P, T19I, T19S	750	3	S750N, S750R, S750I
21	3	R21I, R21T, R21K	752	3	L752I, L752R, L752F
22	3	T22N, T22I, T22A	765	3	R765L, R765H, R765C
26	3	P26L, P26S, P26R	772	3	V772L, V772I, V772A
27	3	A27V, A27T, A27S	780	3	E780D, E780Q, E780V
72	3	G72E, G72W, G72R	812	3	P812S, P812T, P812L
75	3	G75D, G75V, G75R	831	3	A831S, A831V, A831T
80	3	D80N, D80Y, D80A	836	3	Q836H, Q836P, Q836L
97	3	K97E, K97N, K97R	838	3	G838S, G838V, G838D
102	3	R102S, R102I, R102G	839	3	D839Y, D839E, D839N
148	3	N148Y, N148K, N148S	845	3	A845S, A845V, A845D
153	3	M153T, M153I, M153V	847	3	R847T, R847I, R847K
183	3	Q183H, Q183R, Q183L	870	3	I870S, I870T, I870V
218	3	Q218R, Q218E, Q218L	879	3	A879S, A879V, A879T
222	3	A222V, A222S, A222P	930	3	A930S, A930V, A930T
239	3	Q239K, Q239R, Q239H	1085	3	G1085R, G1085E, G1085L
246	3	R246I, R246K, R246S	1129	3	V1129L, V1129A, V1129I
247	3	S247R, S247I, S247N	1153	3	D1153A, D1153H, D1153Y
251	3	P251S, P251H, P251L	1170	3	S1170T, S1170Y, S1170P
263	3	A263T, A263S, A263V			

Table: 2 Deletion-sites observed across the S glycoprotein. Countries represent the origin of strains where the deletions found. We considered the deletions that occurred in at least two strains in a certain position.

Nucleotide Position	Amino acid position	Deleted amino acid	Countries	Number of Strains
179-217	61-73	NVTWFHAIHVSGT	England	1
200-226	68-76	IHVSGTNGT	Taiwan, Malaysia	2
201-224	68-75	IHVSGTNG	Thailand	1
203-208	69-70	HV	Sweden, England, Australia	3
413-421	138-140	DPF	Sweden	1
418-420	140	F	England, Sichuan	3
420-431	141-144	LGVY	England, Iceland, USA, Scotland, Kenya	16
420-422	141	L	England	1
422-430	141-143	LGV	Portugal, England, Iceland, Scotland	4
423-431	142-144	GVY	England, Netherlands	3
428-430	143	V	USA, Belgium	4
428-433	143-144	VY	England	2
429-431	145	Y	England, Canada, Slovenia, Jordan, Netherlands, Saudi_Arabia, Scotland, USA, Spain, Wales, India, Australia	48
724-732	241-243	LLA	China, England, Belgium, Scotland, Netherlands	6
724-726	241	L	USA	2
727-732	243-244	AL	England, Wales, Spain, Sichuan	6
2021-2035	675-679	QTQTN	Taiwan, Malaysia	2