Mutations in human SARS-CoV-2 spike proteins, potential drug binding and epitope sites for COVID-19 therapeutics development

Kunchur Guruprasad

1 Founder, ABREAST™, Plot Nos. 14/A & 15, Sitaramnagar, R.K. Puram – P.O., Hyderabad, 500056, India

ARTICLE INFO

Handling editor: Chandra Shekhar Verma

Keywords:
Human SARS-CoV-2 mutations
Mutation propensity
Invariant sites
Epitope sites
Drug design sites

ABSTRACT

The comparison of 303,250 human SARS-CoV-2 spike protein sequences with the reference protein sequence Wuhan-Hu-1, showed ~96.5% of the spike protein sequence has undergone the mutations till date, since outbreak of the COVID-19 pandemic disease that was first reported in December 2019. A total of 1,269,629 mutations were detected corresponding to 1,229 distinct mutation sites in the spike proteins comprising 1,273 amino acid residues. Thereby, ~3.5% of the human SARS-CoV-2 spike protein sequence has remained invariant in the past two years. Considering different mutations occur at the same mutation site, a total of 4,729 distinct mutations were observed and are catalogued in the present work. The WHO/CDC, U.S.A., classification and definitions for the current variants being monitored (VBM) and variant of concern (VOC) are assigned to the SARS-CoV-2 spike protein mutations identified in the present work along with a list of other amino acid substitutions observed for the variants. All 195 amino acid residues in receptor binding domain (Thr333-Pro527) were associated with mutations in SARS-CoV-2 spike protein sequence including Lys417, Tyr449, Tyr453, Ala475, Asn487, Thr500, Asn501 and Gly502 that make interactions with the ACE-2 receptor <3.2 Å distance as observed in the crystal structure complex available in the Protein Data Bank (PDB code:6LZG). However, not all these residues were mutated in the same spike protein. Especially, Gly502 mutated only in two spike protein sequences and Tyr449 mutated only in seven spike protein sequences among the spike protein sequences analysed constitute potential sites for the design of suitable inhibitors/drugs. Further, forty-four invariant residues were observed that correspond to ten domains/regions in the SARS-CoV-2 spike protein and some of the residues exposed to the protein surface amongst these may serve as epitope targets to develop monoclonal antibodies.

1. Introduction

The outbreak of the ongoing COVID-19 pandemic disease caused due to the human SARS-CoV-2 infection was first reported from the city of Wuhan, Hubei-1 province, China, during December 2019 (Wu et al., 2020). The disease has since, spread rapidly all across the world causing serious infections to millions of people and leading to the loss of several human lives (https://www.worldometers.info/coronavirus/). The SARS-CoV-2 that belongs to the Coronavirinae and SARS-CoV genera (https://www.ncbi.nlm.nih.gov/taxonomy/) is a 30 kb positive-stranded RNA viral genome comprising genes translated into structural and non-structural proteins. One of the proteins, the spike glycoprotein (S-protein), which is a homotrimer presents itself on the surface of the virion as a ‘crown’ and is involved in the recognition of human host cell surface ACE-2 receptor, an essential requirement for viral-host cellular membranes fusion and transfer of the viral nucleocapsid into host cells (Zhang et al., 2020). The SARS-CoV-2 is known to have its origins in bats (Zhou et al., 2020) and transmitted to humans via pangolins intermediate host species (Han, 2020; Lam et al., 2020; Guruprasad, 2020a,d). The disease is currently known to spread mainly via human-to-human contact through respiratory droplets released in air while coughing or sneezing by infected persons or via contact with virus contaminated surfaces. The spike protein comprises an N-terminal S1 subunit and a C-terminal membrane proximal S2 subunit. The S1 subunit contains four domains; S1A, S1B, S1C and S1D. The S1A or N-terminal domain (NTD), recognises sialic acid carbohydrate required for attachment of the virus to the host cell surface and the S1B or the receptor-binding domain (RBD) interacts with the human ACE-2 receptor (Zhang et al., 2020; Wang et al., 2020a). The S2 subunit comprises three long α-helices, multiple α-helical segments, extended twisted β-sheets, membrane spanning α-helix and an intracellular cysteine rich segment (Guruprasad, 2021). A furin-cleavage

E-mail addresses: kunchur.guruprasad@gmail.com, abreastkgp@gmail.com.

1 Website: https://www.abreast.in.
Currently, there are no approved drugs to specifically treat COVID-19 patients. However, certain known drugs to treat other diseases have been approved under emergency use authorization (EUA) by the U.S. Food and Drugs Administration (F.D.A) to treat COVID-19 under strict medical supervision. The antiviral drugs; Remdesivir (Veklury), favipiravir (Avigan), rheumatoid arthritis drug; baricitinib (Olumiant), monoclonal antibodies; combinations of bamlanivimab and etesevimab by Eli Lilly (U.S.A. and Germany). Antibodies; combinations of bamlanivimab and etesevimab by Eli Lilly, U.S.A., and casirivimab and imdevimab by Regeneron, U.S.A., are some of the drugs in use and there are several different therapies being researched (https://www.mayoclinic.org/, https://www.goodrx.com/).

The vaccines approved by the W.H.O. (U.S.A.) are: Moderna COVID-19 (mRNA-1273) (U.S.A.), Oxford/AstraZeneca COVID-19 (U.K. and Sweden), Johnson & Johnson COVID-19 (U.S.A.), Pfizer BioNTech COVID-19 (U.S.A. and Germany). The other vaccines approved for use in one or more countries include; Oxford/AstraZeneca vaccine - COVSHIELD (manufactured by Serum Institute of India), COVAXIN developed by Bharat Biotech (India) in collaboration with ICMR, SPUTNIK V (Russia), Sinopharm COVID-19 (China), CUREVAC (Germany). A draft landscape and tracker of COVID-19 candidate vaccines currently under different stages of clinical trials and awaiting approvals is available at (https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines).

Viruses are known to constantly evolve through mutations. The genetic differences between viruses associated with one or more mutations leads to genetic variants of the virus. Ever since the outbreak of the COVID-19 pandemic, the human SARS-CoV-2 has been undergoing several mutations. Sequences with similar variants are grouped into lineages and multiple lineages can have the same amino acid substitution. Also, different amino acid substitutions can be observed at the same mutation site. The Centers for Disease Prevention and Control (CDC), U.S.A., in collaboration with the SARS-CoV-2 Interagency Group (SIG), established under the US Department of Health and Human Sciences (HHS) designed a classification scheme that was recently revised to include a fourth class of variant classification named Variant Being Monitored (VBM) along with the previously defined three classes of SARS-CoV-2 variants, i.e., Variants Of Interest (VOI), Variants Of Concern (VOC) and Variant Of High Consequence (VOHC). The SARS-CoV-2 variant classification include definitions and attributes of the variants along with the resulting public health action. The attributes associated with the VOIs are: changes to receptor binding, reduced neutralization by antibodies generated against previous infection or vaccination, reduced efficacy of treatments, potential diagnostic impact or predicted increase in transmissibility or disease severity. The VOIs are defined where there is evidence of an increase in transmissibility, severe disease leading to hospitalization or deaths, significant reduction in neutralization by antibodies generated during previous infection or vaccination, reduced effectiveness of treatments or vaccines, or diagnostic failures. A variant of high consequence is defined as one which has clear evidence that prevention measures or medical countermeasures have significantly reduced effectiveness relative to previously circulating variants. The VBM class includes variants with substitutions of concern along with the previously designated VOIs and VOHCs (https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html). The Phylogeographic Assignment of Named Global Outbreak Lineages, a web-based application developed by the Centre for Genomic Pathogen Surveillance (CGPS) in South Cambridgeshire (https://www.sanger.ac.uk/collaboration/centre-global-pathogen-surveillance-cgps/) assists genomic epidemiology by implementing a dynamic nomenclature for SARS-CoV-2 lineages or (PANGO nomenclature) (Rambaut et al., 2020) available at (https://cov-lineages.org/lineage_list.html). The WHO proposed the use of labels corresponding to the different PANGO lineages comprising Greek alphabets assigned to the variants useful to track the transmission and spread of the SARS-CoV-2 including the VOC (https://www.who.int/en/activities/tracking-sars-cov-2-variants/). At the time of communicating this manuscript, there are nine VBMs: alpha, beta, gamma, epsilon, eta, iota, kappa, zeta, mu and one VBM not yet classified with PANGO lineage (B.1.617.3) and 1 VOC, delta and none from the VOI or VOHC categories.

A number of studies have been reported on the mutations in human SARS-CoV-2 spike proteins based on either infectivity and reactivity to a panel of neutralizing antibodies and sera from convalescent patients (Li et al., 2020), or computational analyses of the protein sequences (Guruprasad, 2020a, 2020b, 2021; Kaushal et al., 2020; Korber et al., 2020a; Emary et al., 2021; Phelan et al., 2020; Mercatelli and Giorgi, 2020; Yadav et al., 2020; Saha et al., 2020). The D614G mutation that established itself as the dominant form was one of the high consequence defined under the variant classification. After circulating in Europe (Emary et al., 2021). It has been shown based on the neutralization studies with the spike protein receptor-binding domain monoclonal antibodies and convalescent sera from people with and without the mutation, that viruses with the D614G mutation spread more quickly than viruses without this mutation (Korber et al., 2020b) and that the mutation was not a hindrance to vaccine development (Wu et al., 2021; Shen et al., 2021). The variants characterisation through constant genomic surveillance, especially, in regions where there are

### Table 1

| Domains/Regions | Number of amino acid residues in domains/regions | Total number of distinct mutation sites in domains/regions | Total number of observed mutations |
|-----------------|-----------------------------------------------|--------------------------------------------------------|----------------------------------|
| S1α (NTD) (1-302) | 302                                           | 302                                                    | 422247                           |
| S1β–S2 linker (303-332) | 30                                           | 30                                                     | 2194                             |
| S1β (RBD) (333-527) | 195                                          | 195                                                    | 246622                           |
| S1β – S1 linker (528-533) | 6                                            | 6                                                      | 190                              |
| S1β domain (534-589) | 56                                           | 56                                                     | 5732                             |
| S1β – S1 linker (590-593) | 4                                            | 4                                                      | 85                               |
| S1β domain (594-674) | 81                                           | 81                                                     | 322034                           |
| Protease cleavage site (675-692) | 18                                           | 18                                                     | 87140                            |
| S1-S2 subunits linker (693-710) | 18                                           | 18                                                     | 21631                            |
| Central β-strand (711-737) | 27                                           | 26                                                     | 19294                            |
| Downward helix (738-782) | 45                                           | 41                                                     | 5676                            |
| S2’ cleavage site (783-815) | 33                                           | 32                                                     | 3125                             |
| Fusion peptide (816-828) | 13                                           | 13                                                     | 357                              |
| Connecting region (829-911) | 83                                           | 79                                                     | 6745                             |
| Heptad repeat region (912-983) | 72                                           | 62                                                     | 53655                            |
| Central helix (984-1034) | 51                                           | 40                                                     | 21907                            |
| β-hairpin (1035-1068) | 34                                           | 30                                                     | 1225                             |
| β-sheet domain (1069-1133) | 65                                           | 63                                                     | 11775                            |
| Heptad repeat region (1134-1213) | 80                                          | 75                                                     | 26561                            |
| Transmembrane region (1214-1236) | 23                                           | 23                                                     | 3971                             |
| Cytoplasmic region (1237-1273) | 37                                           | 35                                                     | 7463                             |
viral surges, is useful to inform local outbreak investigations and understand national trends. Such studies aid in defining policies to contain spread of the virus, as well as, monitor the potential impact of the variants on the diagnostics, vaccines and therapeutics, so that variants with reduced susceptibility to treatments are detected quickly.

In the present study, a catalogue of all the known mutations identified among 303,250 human SARS-CoV-2 spike protein sequences with respect to the reference spike protein sequence Wuhan-Hu-1 is presented. The mutations were classified according to the WHO/CDC variant classification and definitions. The other amino acid substitutions observed at the variant site are also listed. The amino acid residues in the viral human SARS-CoV-2 spike protein RBD that makes interactions with human host ACE-2 receptor were identified from the crystal structure complex available in the Protein Data Bank and the spike protein mutations analysed to suggest potential drug design sites. Further, the invariant amino acid residues associated with the different domains/regions in human SARS-CoV-2 spike proteins were analysed to suggest potential epitope sites.

The analyses provide information on the distinct mutation sites observed in human SARS-CoV-2 spike protein with respect to the reference sequence, distribution of the total number of mutations observed at the distinct mutation sites and according to the different domains/regions in the spike protein sequence, mutation percentages, propensities for mutated residues and non-mutated sites corresponding to domains/regions in the spike protein, classification of the variants according to WHO labels along with other amino acid substitutions observed in the present work and identification of potential inhibitor/epitope sites for drug/antibody design.

2. Method

The human SARS-CoV-2 spike protein (surface glycoprotein) sequences available in the NCBI virus database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/) as on November 2, 2021 were obtained in the FASTA format. The spike protein sequences comprising 1,273 amino acid residues as in the reference protein sequence (NCBI Accession code: YP_009724390.1) were considered for the analyses. The mutations were identified by comparing individual spike protein sequences to the reference spike protein sequence in the human SARS-CoV-2 genome isolate collected from an infected individual during December 2019 in

![Fig. 1. A. Distribution of the total number of mutations at the mutation sites in human SARS-CoV-2 spike protein sequence. B. Top 26 mutation sites in human SARS-CoV-2 proteins arranged in decreasing order of the total number of observed mutations.](image)
The WHO classification of variants along with the other amino acid substitutions observed [mentioned within square brackets] represented among the 303,250 human SARS-CoV-2 spike protein sequences.

Table 2

| Mutation | Description |
|----------|-------------|
| D614G   | Most frequent mutation reported previously (Guruprasad et al., 2020). |
| D614N   | Associated with all sub-lineages or a new lineage. |
| A501T   | Associated with the current VBMs and the VOC. |
| L452Q   | Associated with the current SARS-CoV-2 lineages. |

3. Results and discussions

A total of 303,250 human SARS-CoV-2 spike protein sequences comprising 1,273 amino acid residues available in the NCBI Virus database (as on November 2, 2021) were analysed. These spike protein sequences contained a total of 1,269,629 mutations that represented 1,229 distinct mutation sites, suggesting ~96.54% of the human SARS-CoV-2 spike protein sequence has undergone the mutations since outbreak of the COVID-19 pandemic during December 2019. The mutation sites represented a total of 4,729 distinct mutations. The total number of distinct mutation sites observed along with their distribution according to the different domains/regions in the human SARS-CoV-2 spike protein is listed in Table 1. A catalog of all the mutations identified is attached in the Supplementary Table 1.

The distribution of the total number of mutations observed at the different mutation sites along the spike protein sequence is shown in Fig. 1A. The most frequent D614G mutation reported previously (Guruprasad et al., 2020a, 2021; Li et al., 2020) is predominant in the present dataset too and accounts for ~23.27% of the total number of observed mutations, whereas, the mutation percentages at all the remaining sites in spike protein were <5%. The top 26 mutation sites in the dataset arranged in the decreasing order of mutation percentages is shown in Fig. 1B. These represent the mutations at: D614L, D614S, D614N, D614A, D614V, D614S.

The classification of human SARS-CoV-2 mutations identified in the present work according to different PANGO lineages (https://cov-lineages.org/lineage_list.html) using the WHO Greek key classification labels is listed in Table 2. According to the variant classification and definitions by the CDC, U.S.A., (https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html), these are grouped by the WHO into four categories; Variants Being Monitored (VBM), Variants of Interest (VOI), Variants of Concern (VOC), Variants of High Consequence (VHO). Currently there are ten VBMs and 1 VOC with none in the VOI or VHO categories. The table contains 46 variants (indicated in bold) along with other amino acid substitutions observed that may constitute a sub-lineage or a new lineage. The D614G mutation is associated with all the current VBMs and the VOC. The other amino acid substitutions observed at the most common mutated position 614 in human SARS-CoV-2 spike protein sequences were; D614N, D614A, D614S.

Mutations were associated with 1,229 amino acid residues positions in the human SARS-CoV-2 spike protein distributed in the following domains/regions; S1 (NTD) domain (1-302), S1A domain (1-302), S1B domain (303-527), S1B-S1C linker (303-332), S1D domain (593-710), S1D-S2 subunits linker (693-710), fusion peptide region (711-727), and S2 subunits linker (693-710), S2 subunits linker (693-710).
(RBD domain). The S1\textsuperscript{A} (NTD) domain is known to contain two regions; M153ESEFR158 and S247YLTPG252 specific to human SARS-CoV-2 compared to SARS-CoV, both of which via their glycosylated spike proteins recognize the human angiotensin converting enzyme-2 (ACE-2) receptor (Guruprasad, 2020b). Although, all residues at the positions M153 to R158 and S247 to G252 are known to be mutated (Supplementary Table 1), it is interesting to note that the residue W152 immediately preceding the first region and the residue D253 following the second region were associated with relatively high mutation percentages and are among the top 26 mutations in the spike protein (Fig. 1B). However, residues in the above two regions mentioned were associated with far fewer mutations, suggesting perhaps the importance of these regions in the recognition of ACE-2 receptor by the human SARS-CoV-2 spike protein.

The mutation propensity distribution according to the domains/regions shown in Fig. 3, shows the protease cleavage site comprising only 18 amino acid residues is associated with a maximum value (4.85), followed by the S1D domain comprising 81 amino acid residues with mutation propensity value (3.98). The mutations at P681 (with 66,197 mutations) and at D614 (with 295,479 mutations) out of a total of 1,269,629 observed mainly contribute to the high propensity values for the protease cleavage site and S1D domain, respectively. The variants associated with the protease cleavage site (residues 675-692) according to WHO classification were; Q677H (Eta), P681H (Alpha, Mu), P681R (Kappa, Delta) and variants associated with the S1D domain (residues 594-674) were; D614G (Alpha, Beta, Gamma, Epsilon, Eta, Iota, Kappa, Zeta, Mu, Delta) and H655Y (Gamma). The mutations observed at the furin-cleavage site that plays a role in virus-host cell entry represented by the ‘PRRA’ sequence motif were; P681R, P681H, P681L, P681Y, P681S, R682Q, R682L, R682W, R682G, R682P, R683W, R683L, R683Q, R683P, A684V, A684S, A684T, A684E, A684P.

Currently, two anti-SARS-CoV-2 monoclonal antibody treatments, in the U.S.A., for healthcare providers with FDA Emergency Use Authorization (EUA) for the treatment of COVID-19 are: bamlanivimab plus etesevimab (https://www.fda.gov/media/145802/download) and casirivimab plus imdevimab (https://www.fda.gov/media/143892/download) and in laboratory studies, SARS-CoV-2 variants containing L452R or E484K substitution in the spike protein were reported to cause a marked reduction in susceptibility to bamlanivimab and possibly lower sensitivity to etesevimab and casirivimab and have therefore been defined as SARS-CoV-2 substitutions of therapeutic concern according to the CDC, U.S.A. (https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html). These mutations were observed in the present analyses (Supplementary Table 1) and the additional substitutions observed were: L452M, L452Q, L452W, L452P and E484Q, E484Z, E484G, E484A, E484D, E484F, E484R, E484V and...
E484S. According to the WHO classification, the L452R variant is associated with Epsilon, Iota, Kappa, Delta (Table 2), E484K is associated with Alpha, Beta, Gamma, Eta, Iota, Zeta, Mu and E484Q variant is associated with Kappa.

In order to verify correlation between in-silico prediction and experimental evidence, the literature was examined to infer mutations that affect neutralizing antibodies. A comprehensive map of the spike protein RBD escape mutants against the monoclonal antibodies; LY-CoV016 (etesevimab) (Starr et al., 2021a) and LY-CoV555 (bamlanivimab) along with a combination cocktail of the two antibodies have been reported (Starr et al., 2021b). Accordingly, these antibody binding are known to be affected due to mutations at the following twenty-three sites; D405, E406, K417, D420, Y421, L452, L455, F456, N460, I472, Y473, A475, G476, V483, E484, G485, F486, N487, Y489, F490, Q493, S494, G504. Mutations at all these sites were observed in the present work and the different mutations are listed in Supplementary Table 1. Among the mutations that escape either monoclonal antibody or their combination, the E484K for the LY-CoV555 antibody and K417N for the LY-CoV016 antibody show >1000 fold change in the IC50 values according to the neutralization assays (Wang et al., 2021). The mutations at E484 and K417 are among the six spike protein RBD mutations/C203.2 Å interacting distance of the ACE-2 receptor as observed in the crystal structure complex (PDB code:6LZG). The L452R mutation, which is the predominant spike protein RBD mutation at position 452 is also within interacting distance of the ACE-2 receptor and is known to escape the LY-CoV555 antibody (Starr et al., 2021b). The SARS-CoV-2 spike protein mutations reported here taken together with experimental studies on the escape mutants for the antibodies and their cocktail combinations.

Fig. 4. Invariant amino acid residue positions and ACE-2 interacting sites in human SARS-CoV-2 spike protein.

Note: Forty-four invariant amino acid residues (highlighted in green) were not observed to be mutated among 303,250 human SARS-CoV-2 spike protein sequences in the NCBI database with respect to the reference sequence (NCBI accession code: YP_009724390.1). The RBD region comprising residues Thr333-Pro527 (underlined) and eight residues in RBD that make interactions <3.2 Å with ACE-2 receptor (in PDB code: 6LZG) are highlighted (cyan).
suggest that more specific drugs/antibodies need to be developed for COVID-19. The amino acid residue sites that have not undergone any mutation during the past two years since outbreak of the COVID-19 pandemic disease in SARS-CoV-2 spike protein or those in the RBD that interact with ACE-2 receptor and known to have undergone relatively least mutations may serve as useful targets for the design.

The SARS-CoV-2 spike protein residues that have remained invariant since the outbreak of COVID-19 pandemic ~2 years ago, are listed in Table 3 along with their associated domains/regions. These residue sites highlighted (in green) in Fig. 4, constitute ~3.4% of the human SARS-CoV-2 spike glycoprotein sequence. The propensity values evaluated for the non-mutated sites in the spike protein according to the domain/region is shown in Fig. 5. The invariant residues were associated with ten domains/regions; the central helix with maximum propensity value (6.24) followed by heptad repeat region (residues 912-1213) (4.01), β-hairpin region (3.4), downward helix (2.57), heptad repeat region (residues 1134-1213) (1.8), cytoplasmic region (1.56), connecting region (1.39), central β-strand (1.07), β-sheet domain (0.89) and the S2 cleavage site (0.87). The three-dimensional electron microscopy structure of the SARS-CoV-2 spike glycoprotein (closed state) comprising the three subunits (PDB code: 6VXX_A-chain, B-chain, C-chain) (Walls et al., 2020) was used to map the position of invariant residues. Based on the molecular visualization in graphics using PyMol, some of the invariant residues representing different domains/regions that were exposed on the protein surface are shown in Fig. 6(A-D). These correspond to some of the residues in the heptad repeat region (residues 912-983) listed in Table 3. The residues; Gln920, Gln935, Asn953, Asn955 are shown highlighted (red) in the three-dimensional structure of the protein (PDB code: 6VXX_A-chain) in Fig. 6A and the equivalent residues were highlighted in the B-chain (green) and in C-chain (blue). The hydrophobic residues; Leu959 in the heptad repeat region and Ile1013 in central helix that are in close proximity in three-dimensional structure of the spike protein and the glycosylated residue; Asn801 in S2’ cleavage site is also exposed on the protein surface. The heptad repeat region (residues 912-983) and the central helix (residues 984-1034) are associated with relatively high propensity values for the invariant residue sites in the human SARS-CoV-2 spike protein as shown in Fig. 5. Also, the invariant hydrophobic residue; Leu1145 (in heptad repeat region 1134-1213) from all the three protein subunits form a hydrophobic cluster at the base of the spike protein as shown in Fig. 6B. It is noted that residues beyond this region are not defined in the three-dimensional structure for the individual chains. Further, the invariant residue; Cys1126 in the β-sheet domain (1069-1133) close to the N-acetyl-D-glucosamine (NAG) sites (NAG1319, NAG1320) in the A-chain are also exposed on the exterior of the protein surface as shown in Fig. 6C. The invariant residues; Gln988 and Gln992 in the central helix were exposed within the interior on the protein surface as shown in Fig. 6D. These invariant sites exposed on the protein surface may serve as potential epitope targets for antibody/inhibitor design. The epitopes may be constituted by the invariant residues contributed from more than one chain among the three spike protein subunits; A-chain (red), B-chain (green) and C-chain (blue) as shown in Fig. 6B.

The human SARS-CoV-2 spike protein receptor binding domain (RBD) comprising amino acid residues Thr333-Pro527 is involved in the interactions with human host cell ACE-2 receptor. All 195 amino acid residues in RBD have undergone the mutations and the top six mutation sites in SARS-CoV-2 spike protein along with their mutation percentages were; L452 (5.5%), T478 (3.77%), E484 (3.32%), N501 (1.95%), S477 (1.8%), K417 (1.54%). The rest of the mutation sites in the RBD domain were associated with ≤0.37%. The L452 mutation site also ranked second among the top 26 mutation sites observed in the overall SARS-CoV-2 spike protein (Fig. 1B) suggesting high susceptibility to the mutations at these sites relative to the other residues in spike protein RBD. In the crystal structure of the human SARS-CoV-2 RBD protein complexed with ACE-2 receptor (PDB code: 6LZG) (Wang et al., 2020c), the amino acid residue atomic interactions between the spike protein RBD and ACE-2 receptor defined by interactions ≤3.2 Å were; Lys417(NZ)-Asp30(OD1), Tyr449(OD)-Asp38(OD1), Tyr449(OH)-Glu42(NE2), Tyr453(OH)-His34(ND1), Ala475(O)-Ser19(OG), Asn487(OD1)-Glu24(NE2), Asn487(ND2)-Tyr83(OD), Thr500(OG1)-Tyr41(OG), Asn501(ND2)-Tyr41(OG) and Gly502(N)-Lys353(O). The interacting amino acid residues in the spike protein are highlighted (in cyan) in Fig. 4. Except Lys417, none of the spike protein RBD residues involved in the interactions with ACE2 were among the top six mutations mentioned above. Therefore, most of these residue interactions may play an important role in host cell infection and virus transmission and their mutations are likely to affect the human SARS-CoV-2 spike protein and ACE-2 interactions. Only two Gly502 mutations were observed out of the total 1,269,629 mutations among 303,250 human SARS-CoV-2 spike proteins. These were; G502E in the spike protein with NCBI Accession code (UAT83124.1) and G502V in (QTY96446.1). Likewise, only five Tyr449 mutations were observed that correspond to 7 proteins. These were; Y449S (QWN56156.1, QZJ78555.1), Y449D.
An examination of the other mutations in spike proteins (NCBI Accession codes) mentioned above comprising the Gly502 and Tyr449 mutations revealed that none of the individual spike proteins contained mutations involving all residues that make interactions with the ACE-2 receptor as shown in Table 4. This observation suggests that all the SARS-CoV-2 spike protein RBD – ACE-2 receptor interactions outlined earlier in this work may not be necessary for virus-host cell transmission. The Tyr449 makes side-chain to side-chain hydrogen bond interactions with Asp38 and Gln42 as shown in Fig. 7A. The Asp38 and Gln42 residues are associated with helix conformation in the ACE-2 receptor (as in PDB code: 6LZG). This tyrosine residue represented in the ‘VGGNY’ sequence is one of the two sequences in human SARS-CoV-2 spike protein RBD previously identified among structural determinants for recognition of the human ACE-2 receptor (Guruprasad, 2020b). The amide nitrogen of Gly502 is involved in main-chain to main-chain hydrogen bond with carbonyl oxygen of Lys353 as shown in Fig. 7B. The Asn487 makes side-chain interactions with the side-chains of Gln24 and Tyr83 as shown in Fig. 7C. Asn487 was observed to be mutated among twenty-four proteins in the dataset analysed. While all the sites of interacting residues in human SARS-CoV-2 spike protein RBD with human ACE-2 receptor offer attractive targets for inhibitor/drug design, Gly502 and Tyr449 sites may be more promising considering their relatively low mutation rates during the past two years. Therefore, inhibiting interactions made by the human SARS-CoV-2 spike protein RBD with the ACE-2 receptor are suggested as potential sites for COVID-19 drug design.
proteins corresponding to the variants classified according to WHO suggest, variants that may be examined to delineate whether they are likely to affect virus transmission, cause severity of infection, escape host immune recognition or efficacy of existing vaccines and monoclonal antibodies. The epitope targets and drug design sites proposed in this work may be examined for further development of drugs and monoclonal antibodies.

4. Conclusions

The comparison of 303,250 human SARS-CoV-2 spike protein sequences of length comprising 1,273 amino acid residues with the first reported human SARS-CoV-2 spike protein reference sequence Wuhan-Hu-1, showed 1,229 distinct mutation sites that represented 4,729 distinct mutations. All these mutations are catalogued in the present work along with the classification of the variants according to the WHO definitions of Variants Being Monitored (VBM) and Variant Of Concern (VOC). The other amino acid substitutions observed for the variants are also presented. The D614G mutation in spike protein S1 domain is common to Alpha, Beta, Gamma, Epsilon, Eta, Iota, Kappa, Zeta, Mu, Delta variants that represent the current VBM and VOC. The protease cleavage site (amino acid residues; 675-692) was associated with the maximum mutation propensity (4.85). Forty-four sites or nearly 3.45% of the human SARS-CoV-2 spike protein sequence has not undergone any mutations since outbreak of the COVID-19 pandemic. The invariant residue sites were associated with ten domains/regions of the SARS-CoV-2 spike protein with a maximum propensity value (6.24) for the central helix (984-1034), followed by the heptad repeat region (912-983) with propensity value (4.01) and the beta-hairpin (1035-1068) with propensity value (3.4). Some of the invariant site residues identified that are exposed to the protein surface may serve as potential epitope targets for monoclonal antibodies or inhibitor design. Eight residues; Lys417,
