Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of the Indian Chemical Society 98 (2021) 100156

Contents lists available at ScienceDirect

Journal of the Indian Chemical Society

journal homepage: www.editorialmanager.com/JINCS/default.aspx

Genomic variation and point mutations analysis of Indian COVID-19 patient samples submitted in GISAID database

Shikha Mudgal [a], Rohitash Yadav [b,**], Hoineiting Rebecca Haokip [c], Ananya Pandit [d], Y. Sheena Mary [e,*]

[a] Division of Molecular Biology Proteomics, And Metabolomics, All India Institute of Medical Sciences, Rishikesh, India
[b] Department of Pharmacology, All India Institute of Medical Sciences, Rishikesh, India
[c] Medical Allied ICU, All India Institute of Medical Sciences, Rishikesh, India
[d] Uttaranchal Institute of Pharmaceutical Sciences, Dehradun, Uttarakhand, India
[e] Thushara, Neethinagar-64, Pattathanam, Kollam, Kerala, India

ARTICLE INFO

Keywords:
COVID-19
SARS CoV-2
Point mutation
Non-structural protein

ABSTRACT

Corona virus disease 2019 (COVID-19) endemic has havoc on the world; the causative virus of the pandemic is SARS CoV-2. Pharmaceutical companies and academic institutes are in continuous efforts to identify anti-viral therapy or vaccines, but the most significant challenge faced is the highly evolving genome of SARS CoV-2, which is imparting evolutionary selective benefits to the virus. To understand the viral mutations, we have retrieved nine hundred and thirty-four samples from different states of India via the GISAID database and analyzed the frequency of all types of point mutation in all structural, non-structural proteins, and accessory factors of SARS CoV-2. Spike glycol protein, nsp3, nsp6, nsp12, N and NS3 were the most evolving proteins. High frequency point mutations were Q496P (nsp2), A380V (nsp4), A994D (nsp3), L37F (nsp6), P323L & A97V (nsp12), Q57H (ns3), D614G (S), P13L (N), R203K (N), G204R (N) and S194L (N).

1. Introduction

Wuhan, China, was the site for the COVID-19 outbreak in December 2019. Around the world, the endemic had spread to 206 countries and territories, with 19,73,15,905 confirmed cases as of July 31, 2021, and 42,07,830 deaths. In India 31,612,794 confirmed cases and 423,842 deaths as of July 31, 2021 were reported [1]. The World Health Organization (WHO) has declared the novel Corona virus (COVID-19) out break a global pandemic on March 11, 2020. Corona viridae (CoVs) are the largest known positive-sense, single-stranded RNA viruses with 30 kbp genome. This novel beta-CoVs corona virus (2019-nCoV) originating from Wuhan, China, has been linked to severe respiratory infections in humans. The genome of SARS CoV-2 comprises of 5' and 3' un-translated regions constituting 265 and 358 nucleotides, respectively. The 5' 20 kb region of the genome encodes for two ORFs (ORF1a/1 ab) which contain 16 non-structural proteins (nsp's) from nsp1-16 required to form the virus replication and transcription complex. The 3' proximal region encodes additional factors and, four structural proteins spike glycoprotein, envelop, membrane and nucleocapsid (Fig. 1) [2].

Research institutes and pharmaceutical industries worldwide continuously conduct clinical trials to identify the suitable drug/vaccine against SARS CoV-2. The most potential drug target receptor identified as ACE2 receptor [3,4], Trans membrane protease serine 2 (TMPRSS2) [4], 3CLpro (3C like protease) [5], and RdRp (RNA dependent RNA polymerases) [6]. Even after vigorous trials and studies, a potential breakthrough is not achieved; the most likely reason attributable is rigorous rate in which mutations occur in the SARS CoV-2 genome. Yadav et al. reported the DFT and MD simulations of drug molecules and inhibition of SARS-CoV-2 proteins recently [7–11].

GISAID has developed a unique vocabulary of hCoV-19 virus centred

Abbreviations: COVID-19, Corona virus disease 2019; SARS-CoV-2, Severe acute respiratory syndrome corona virus 2; NSP, Non-structural protein; RdRp, RNA dependent RNA polymerase; ACE2, Angiotensin-Converting Enzyme 2; RBD, Receptor binding domain; RCT, Replication and Transcription Complex; S protein, Spike glycol protein; N protein, Nucleocapsid protein; M protein, Membrane protein; E protein, Envelope protein.

* Corresponding author.
** Corresponding author.
E-mail addresses: sh.mudgal@gmail.com (S. Mudgal), rohitashyadav1@gmail.com (R. Yadav), rebeccahaokip9620@gmail.com (H.R. Haokip), ananyapanditdgp@gmail.com (A. Pandit), marysheena2018@rediffmail.com (Y.S. Mary).

https://doi.org/10.1016/j.jics.2021.100156
Received 31 March 2021; Received in revised form 2 August 2021; Accepted 11 September 2021

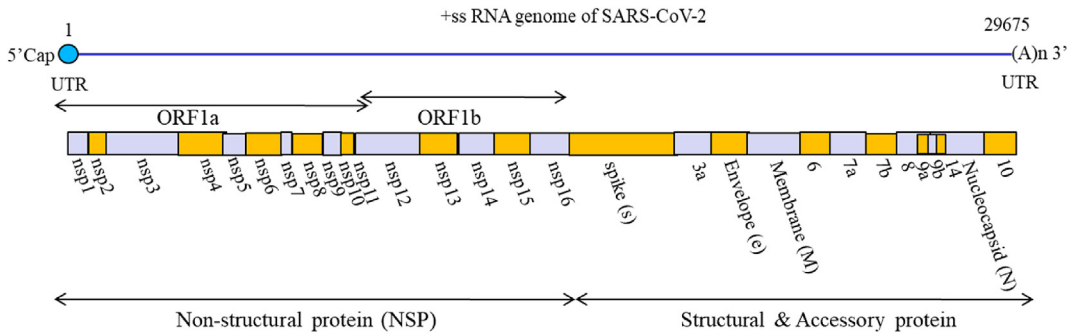


Fig. 1. Genomic representation of structural, non-structural, and accessory proteins of SARS CoV-2.

on variation markers dividing SARS CoV-2 in eight clads, G, GH, GR, GV, GRY, S, L and others [12]. GISAID clads are formed by the statistical calculation of genome distance in phylogenetic clusters and by further merging small linages into major clads based on variation markers [13]. Clad GRY consist linage B.1.1.7, GR (B.1.1.1), GV (B.1.177), GH (B.1.*), G (B.1), L(B), V(B.2), S(A) [14].

It is essential to know the frequency of mutations/variations in all the significant proteins of the SARS CoV-2. It will help to understand the conserved (protein with fewer mutations) and rapidly evolving (protein with high number of conversions) proteins. This information might help researcher to know the effect of mutations on resistance to antiviral therapy. Hence, we planned the study to account for the mutation rate per protein by collecting Indian samples from GISAID database.

2. Material and methods

2.1. Data collection

The genomic sequences of SARS-CoV-2 COVID-19 patients were retrieved from the GISAID database between September to October 2020 [15]. Samples included in the study should be SARS CoV-2 sequencing samples collected from the host human, complete (above 29 KDa), high coverage whole genome sequence. All the COVID-19 isolates that occurred at the start of every Indian location will be collected.

We took a total of nine hundred and thirty four (934) complete and high coverage whole genome sequences of COVID-19 isolates from nineteen states of India (number of samples) namely Gujarat (100), Odisha (100), Delhi (100), Karnataka (100), Maharashtra (100), Telangana (100), West Bengal (100), Uttarakhand (55), Haryana (54), Madhya Pradesh (41), Tamil Nadu (28), Uttar Pradesh (27), Rajasthan (7), Punjab (7), Ladakh (6), Bihar (4), Andhra Pradesh (2), Assam (2), Nepal (1), Jammu and Kashmir (1) (Fig. 2a). In Indian data, identified clades were seven and, the sample's per clad were G (321), GH (104), GR (308), S (19), V (2), L (7), and others (173) (Fig. 2b). Common variations among clads were detected using bioinformatics web tool Venn of Bioinformatics and Evolutionary genomics [16].

India form GISAID database; (b) Clad wise distribution of samples; (c) Comprehensive view of dispersal of samples in the seven clades in Gujrat, Odisha, Delhi, Karnataka, Maharashtra, Telangana and West Bengal.

2.2. Data extraction from GISAID

We used the Epicov server to search Indian COVID-19 isolated statewise; the filters used were complete and had high coverage. Manually Indian COVID-19 sample data mutation details were collected and recorded location wise 10 per time using the CoVserver tool of the GISAID database for all the 25 structural and non-structural proteins of the SARS CoV-2. The sample sequences compared with the reference

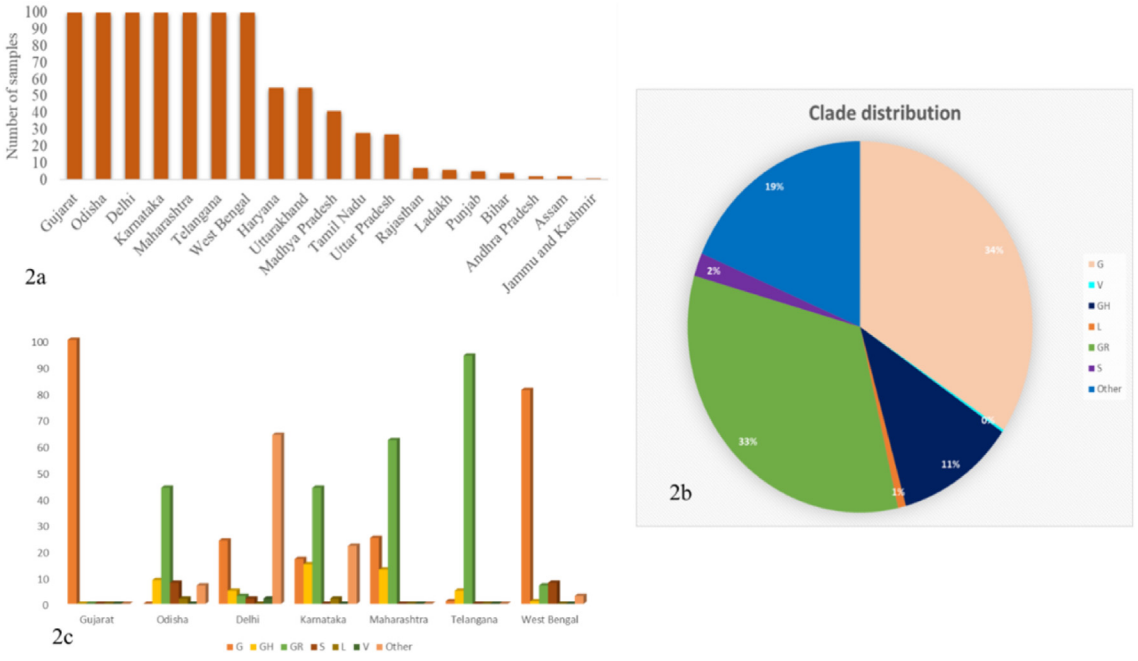


Fig. 2. Sample distribution (a) Number of samples collected from different states of.

genome "hCoV19/Wuhan/WIV04/2019" (it is the sequence of the initially identified SARS CoV-2 isolate from Wuhan, China) simultaneously via CoVserver.

2.3. Mutation frequency calculation

We imported the list of mutations of all the nine hundred and thirty four samples in the R3.6 environment. R base package "table" function calculated the frequency of mutations for all the 25 structural and non-structural proteins. Its function "sort" sorted it in decreasing order to get the highest frequency per protein. Formula used for frequency calculation was:

$$Frequency = \sum COVID - 19 \; cases \; of \; "point \; mutation'"$$

2.4. Mutated protein selection criteria

Two phases of the selection criteria set: In the first phase, all the proteins with mutations in more than 200 COVID isolates were selected (Fig. 3), and in the second phase, the selected proteins were further analyzed and segregated based on the frequency of mutations per protein (cut off set to 100 point mutations per protein). We grouped the SARS CoV-2 proteins into conserved and evolving proteins based on mutation-selection criteria defined earlier. Evolving proteins are proteins in which more than 200 samples recorded modifications and the rate of frequency of point mutations was crossing over 100. Conserved proteins definition is vice versa of evolving proteins.

3. Results

3.1. CoVserver result

The CoVserver compared all the COVID-19 isolates selected, with the reference genome "hCoV19/Wuhan/WIV04/2019" (it is the sequence of the initially identified SARS CoV-2 isolate from Wuhan, China), and we got the list of changes in amino acid for all the genome searched along with the clad information.

3.2. Data description

We collected nineteen (19) Indian states data from the GISAID consortium. Till November 4, 2020 total of 2458 samples of sequencing data was submitted for all the 19 states in the database, out of which we analyzed 934 sample data. Clads identified in India were G, GH, GR, S, V, L and others. Samples collected from clads were G (321), GH (104), GR (308), S (19), L (7) and others (173). Dominant clad of SARS CoV-2 found were G and GR (Fig. 2c).

Gujarat had 610 submitted samples till November 4, 2020 and we analyzed 100 COVID-19 samples. All the 100 SARS CoV-2 samples were of clad G.

Odisha had 610 submitted samples till November 4, 2020, and we analyzed 100 COVID-19 samples. The distribution of SARS CoV-2 samples along clads G (30/100), GH (9/100), GR (44/100), S (8/100), L (2/100) and other (7/100).

Delhi had 147 sequence data and we analyzed 100 COVID19 samples data. Dominant clad in Delhi was other (64/100 samples) followed by G (24/100), GH (5/100), GR (3/100), S (2/100) and V (2/100).

Karnataka data in the GISAID consortium was 115 COVID 19 samples and for the mutation analysis, the team collected 100 sample data. The highest SARS CoV-2 samples were of GR clad (44/100), other (22/100), G (17/100), GH (15/100), L (2/100).

Maharashtra had a high number of COVID 19 samples, followed by Gujarat and Telangana, 457, and we scrutinized 100. The state had the GR (62/100) SARS CoV-2 dominant variant, followed by G (25/100) and GH (13/100).

Telangana submitted 539 COVID-19 sample data to the GISAID and analyzed among them were 100. Maximum SARS CoV-2 samples were from GR clad (94/100), GH (5/100), G (1/100).

West Bengal had 187 COVID-19 samples, and analyzed among them were 100, G clad had 81/100, S (8/100), GR (7/100), other (3/100), GH (1/100).

3.3. First phase selection criteria

SARS CoV-2 proteins mutations noted from 934 COVID-19 isolates were: nsp1 (56 samples), nsp2 (360 samples), nsp3 (827 samples), nsp4 (204 samples), nsp5 (65 samples), nsp6 (231 samples), nsp7 in 14, nsp8

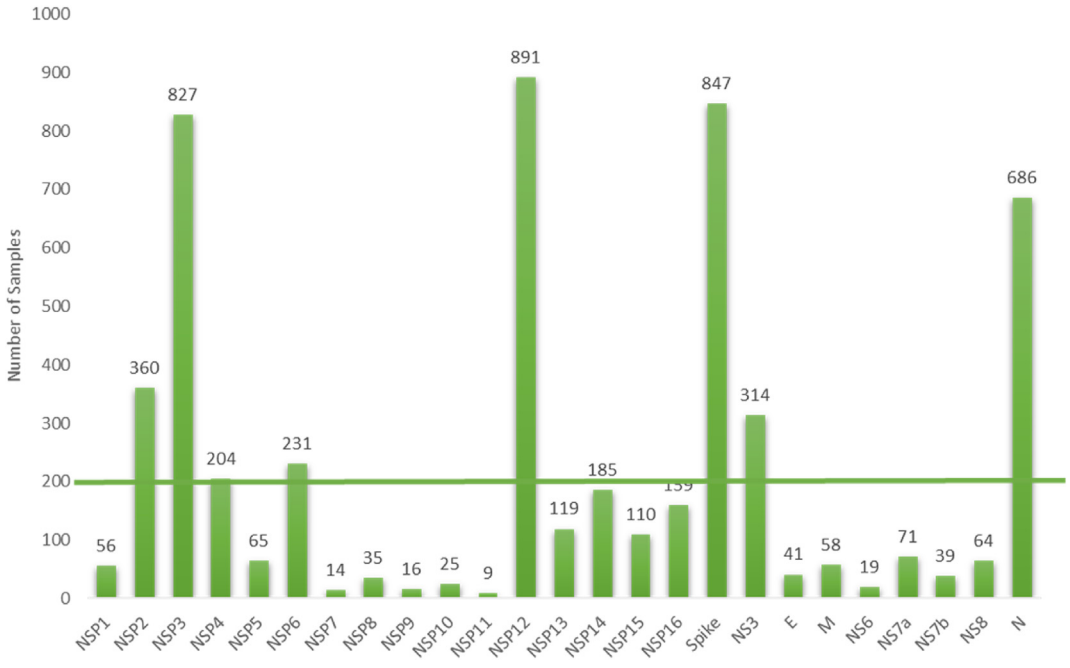


Fig. 3. The number of samples with mutations, distributed per protein.

in 35, nsp9 in 16, nsp10 in 25, nsp11 in 9, nsp12 in 891, nsp13 in 119, nsp14 in 185, nsp15 in 110, nsp16 in 159, Spike glycoprotein (s) in 847, ns3 in 314, Envelope (E) protein in 41, Membrane (M) protein in 58, ns6 in 19, ns7a in 71, ns7b in 39, ns8 in 64 and Nucleocapsid (N) in 686 samples (Fig. 3). The first phase selection criteria with mutations in more than 200 COVID 19 isolates were nsp2, nsp3, nsp4, nsp6, nsp12, ns3, Spike glycoprotein and Nucleocapsid (N) protein.

3.4. Second phase selection

Second phase selection – criteria are selecting protein-based on point mutation frequency calculated for all the SARS CoV-2 proteins. In nsp2, 213 types of point mutations were present; the highest frequency was for Q496P (39 samples). The nsp4 had 165 varieties of conversions, and the highest frequency was for A380V (55). The nsp3 had 540 types of mutations, and the highest frequency was for A994D (132 samples). The

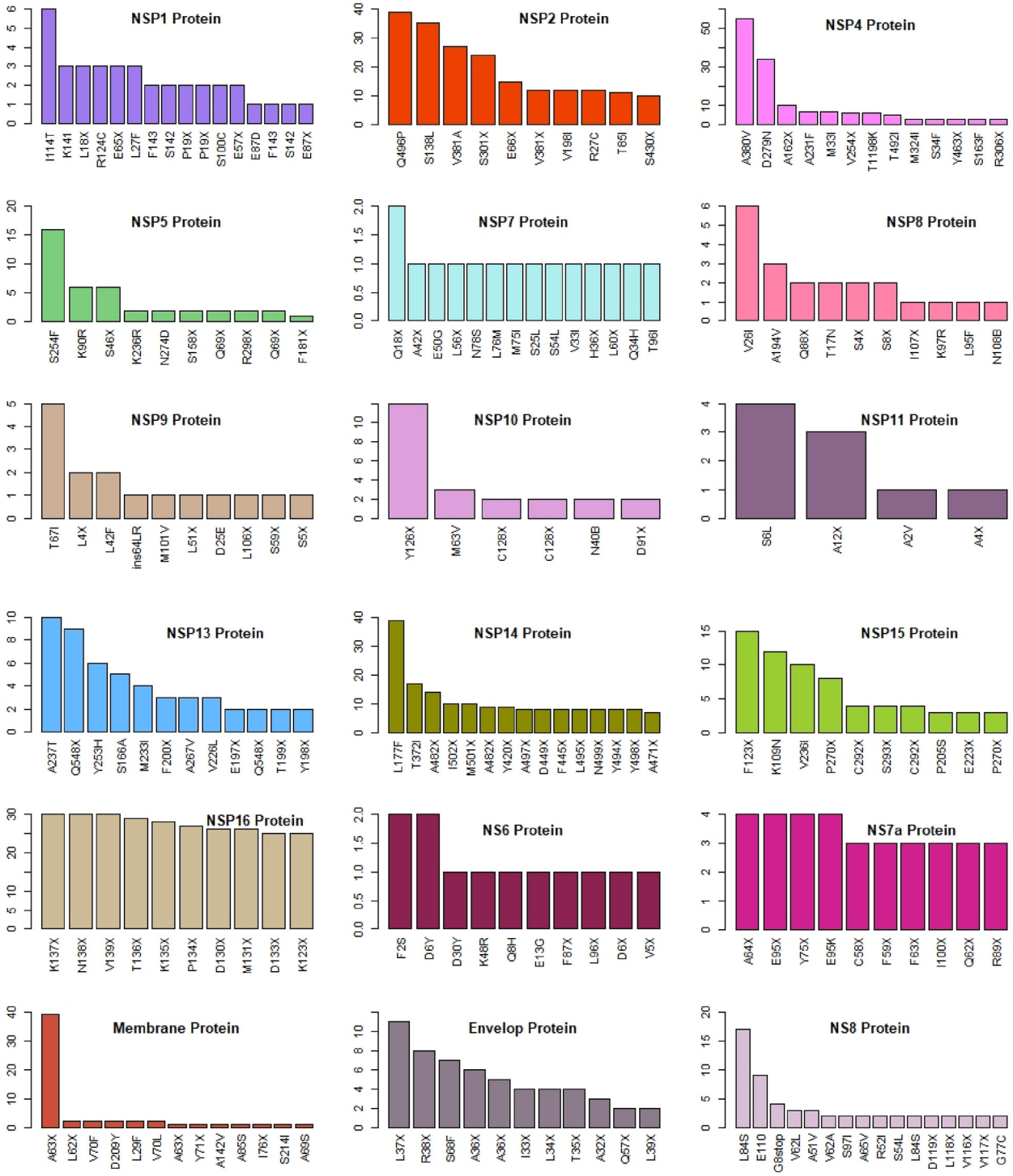


Fig. 4. Frequency of mutations in various proteins of SARS CoV-2: nsp1, nsp2, nsp4, nsp5, nsp7, nsp8, nsp9, nsp10, nsp11, nsp13, nsp14, nsp15, nsp16, ns8, ns6, ns7a, ns7b (no mutation identified), M and E proteins have less frequency of mutations hence considered conserved protein.

nsp6 had 60 types of mutations; L37F had the highest frequency (140 samples). The nsp12 had 272 types of mutations, and the highest frequency mutations were P323L (414) and A97V (132). The ns3 had 112 types of mutations; the highest frequency mutation calculated was Q57H (159). Spike glycoprotein (S) had 478 types of mutations; D614G (669) had the highest frequency. Lastly, the Nucleocapsid (N) protein had 171 mutations, and the highest mutational frequency was for P13L (123), R203K (291), G204R (312), S194L (137). The nsp2 and nsp4 did not meet the second phase selection criteria hence were excluded. After the complete analysis, we found nsp3, nsp6, nsp12, ns3, Spike and N proteins to be evolving proteins (Fig. 4) and nsp1, nsp2, nsp4, nsp5, nsp7, nsp8, nsp9, nsp11, nsp13, nsp14, nsp15, nsp16, ns6, ns8, ns7a, ns7b, M, E were conserved proteins (Fig. 5).

3.5. Variations common among clads

GISAID consortium has it system of nomenclature for the hCoV-19 virus, segregating viral strain in eight clads. Our study included five central clads G, GH, GR, S and Others (O), including all the B.1 linages and B.2(S) and mother of all the diversity. We studied the common variations present in clads for Spike protein (considered a major virulent factor). G GH GR shared five common variations tabulated in Fig. 6, G GH (4), G GR (13), G O S (5), O S (4), G O (13), GH O S (2), GH O (4), G O (13), GH O S (2), GH O (4), O S (2), GR O S (2), GR O (9), O S (2) and common between all five clad G GR GH O S are two mutations namely D614G and S943X.

4. Discussion

SARS CoV-2 is a rapidly evolving genome. The study showed that the highest number of point mutations was in Spike glycoprotein followed by nsp12, N, ns3, nsp6 and nsp3 in the order mentioned. These proteins play an essential role in the life cycle of viruses and are also potential therapeutic targets. Elevated amounts of point mutations in the proteins might be the reason behind the failure of the drugs explored as a possible therapy.

Spike protein is a tri-meric protein containing 21 to 35 *N*-glycosylation sites giving the virus the characteristic crown-like structure. The outer region of spike protein has an S1 domain at *N*-terminal comprising receptor-binding domain (RBD) and S2 domain at C-terminal contain fusion peptides. Spike RBD binds to the host cell's ACE2 receptor, resulting in proteolytic cleavage of S1 and S2 domain, exposing fusion peptides which then inserts in the host plasma membrane and thus facilitating the entry of the virus into the host cell [17]. Spike protein is the main virulence factor triggering an antigenic immune response; therefore, it is a potential antiviral therapy target. The fast-evolving genome of SARS CoV-2 consistently changes the spike protein, rendering it stable and resistant to antiviral treatment.

Two-thirds of the SARS CoV-2 genome comprises two reading frames ORF1a and ORF1b encoding two poly protein pp1a and pp1*b* or one poly protein pp1ab. The poly proteins process into 16 non-structural proteins and most of them are part of the replication and transcription complex (RTC) of the virus. RTC is a compartment where replication and translation of virus occur. It's clearly formation is essential to protect the viral genome from host immune response, which results in increased replication proficiency [18].

Non-structural protein (nsp3) is the most significant multi-domain protein, approximately 200 kDa of RTC of SARS CoV-2. The nsp3 has many roles to play; (i) It has a papain-like protease domain which helps it to release from poly peptide pp1ab [19]; (ii) It also lowers the effect of host immune response by interfering in cytokine expression [20]; (iii) nsp3, 4 and 6 all are trans membrane proteins involved in the convolution of endoplasmic reticulum membrane forming double-membrane vesicle (DMV) causing the formation of replication sites [21].

RNA dependent RNA polymerase (RdRp or nsp12) is a catalytic unit complex with cofactors nsp7 and nsp8 [22]. As its name signifies, its primary function is the replication of the viral genome. It might be the conserved domain among most RNA viruses, but the SARS CoV-2 genome shows an elevated frequency of mutations in the protein rendering more survival advantages to the virus [23].

The nsp6 is a *trans*-membrane protein that functions to induce auto phagosomes from the endoplasmic reticulum and reduces their size. This help virus escapes lysosomal degradation [24]. The ns3 is an accessory protein encoded by ORF8 which is a hyper variable protein evolving rapidly. Its function in masking virus from class I MHC (major histocompatibility factors) thereby overwhelming the type I interferon response [25].

Nucleocapsid (N) protein plays various vital roles in the virus life cycle. Still, the most crucial part is the packaging of the viral genome in a protected, flexible shell called ribonucloprotein (RNP) complex, which further ensures viral genome timely replication [26]. It is present in replication and transcription complex (RTC), where it is associated with nsp3 and facilitates viral RNA replication [27]. Amongst all the 25 proteins of SARS CoV-2, the conserved protein could prove to be a better therapeutic target, which might impart success to combat the COVID-19 challenge.

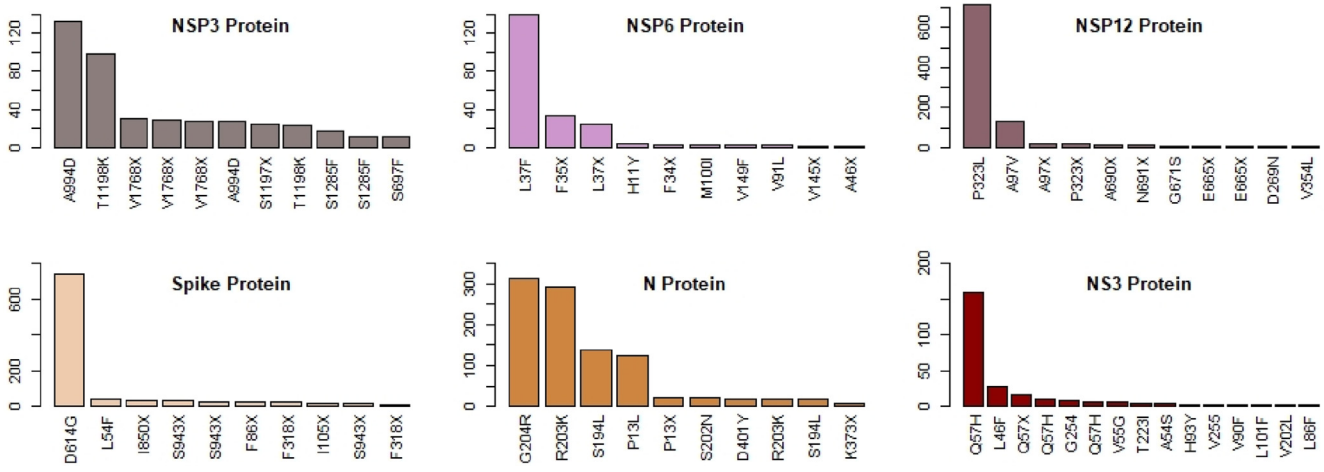
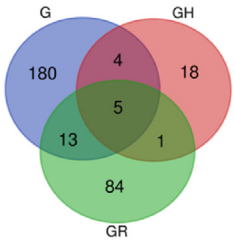


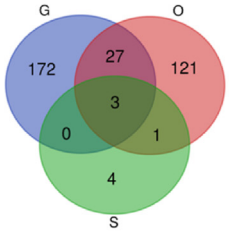
Fig. 5. Frequency of mutations in various proteins of SARS CoV-2 proteins: nsp3, nsp6, nsp12, Spike, N, and ns3 proteins have a high frequency of mutations and hence considered mutated proteins.



6.1.a

Name	Total	Variations
G GH GR	5	F318X, N801X, Q677H, S943X, D614G
G GH	4	K1181X, A262S, T299I, E583D
G GR	13	G75A, K278X, L387X, I850X, A263X, F86X, D88X, G832X, G769X, Q675H, N764X, Y266X, I105X

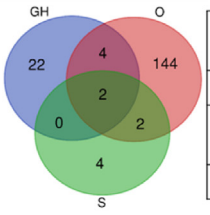
6.1.b



6.2.a

Name	Total	Variations
G O S	5	H519X, S943X, D614G
O S	4	A262X
G O	13	N122X, D294X, N99X, S1175X, A263X, V120X, E96X, F318X, P1263L, K558N, P384X, G832X, N801X, T393X, D614X, D80X, T114X, M1237X, Y266X, K97X, N1173X, A771X, K77X, A123X, A397X, L54F, L110X

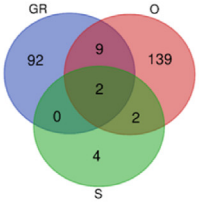
6.2.b



6.3.a

Name	Total	Variations
GH O S	2	S943X, D614G
GH O	4	Q23R, N801X, F318X ,D138Y
O S	2	A262X, H519X

6.3.b



6.4.a

Name	Total	Variations
GR O S	2	S943X, D614G
GR O	9	N801X, G1251V, A243S, A263X, Y144X, F318X, Y266X, G832X, D138Y
O S	2	A262X, H519X

6.4.b

Name	Total	Variations
G GR GH O S	2	S943X, D614G

6.5

Fig. 6. Venn diagram of Spike protein for all the clads, 6.1.a)Venn diagram of interaction in G GH GHR clad; 6.2.a) Interactions between G O S; 6.3.a) Interactions between GH O S; 6.4.a) Venn of GR S O: Part b of figures, 6.1, 6.2, 6.3, 6.4: Tables of common variations in the area of interest; 6.5) Table depicting two common variations shared by all the five clads analyzed, G, GH, GR, O, S.O: others.

5. Conclusion

Analysis of the SARS CoV-2 virus genome revealed and proved the changing character of the virus, which is a well-known fact. Viruses evolve very rapidly to the changing environment that favours their survival. The history states the disasters caused by the Polio virus, H1N1, SARS CoV, HIV etc., and all the so far known viruses use different strategies to combat the human immune system. GISAID data base platform had allowed studying the SARS CoV-2 protein mutations. SARS CoV-2 proteins involved in combating host immune response or developing a protective shell around the virus are rapidly changing, a probable reason for the survival advantage for SARS CoV-2 and failures of the potential drugs tested. We found that most of the mutations are substituted, followed by few deletions. These mutations and their patterns might help in devising strategies to overcome the problem caused. The low efficacy of vaccines tested so far show the high evolution rate of the virus.

Funding sources

Self-Funded.

Ethical statement

Not applicable.

Declaration of competing interest

No potential conflict of interest was reported by the authors.

Acknowledgments

Thanks to the Department of Pharmacology, All India Institute of Medical Sciences, Rishikesh, and Molecules of Life research lab, Society of Young Biomedical Scientists, India for providing the facility.

References

[1] https://www.worlddometers.info/coronavirus/.
[2] Saxena SK. Corona Virus Disease 2019 (COVID-19): Epidemiology, Pathogenesis. Diagnosis and Therapeutics. Springer nature; 2020. https://doi.org/10.1007/978-981-15-4814-7.
[3] Zoufaly A, Poglitsch M, Aberle JH, Hoepler W, Seitz T, Traugott M, Grieb A, Pawlka E, Laferl H, Wenisch C, Nuehold S, Haider D, Stiasny K, Bergthaler A, Puchhammer-Stoeckl E, Mirazimi A, Montserrat N, Zhang H, Slutsky AS,

Penninger JM. Human recombinant soluble ACE21 in severe COVID-19. Lancet Respir. Med. 2020;8:1154–8. https://doi.org/10.1016/S2213-2600(20)30418-5.
[4] McKee DL, Sternberg A, Stange U, Laufer S, Naujokat C. Candidate drugs against SARS-CoV-2 and COVID-19. Pharmacol. Res. 2020;157:104859. https://doi.org/10.1016/j.phrs.2020.104859.
[5] Rathanayake AD, Zheng J, Kim Y, Perera KD, Mackin S, Meyerholz DK, Kashipathy MM, Battaile KP, Lovell S, Perlman S, Groutas WC, Chang KO. 3C-like protease inhibitors block corona virus replication in vitro and improve survival in MERS-CoV infected mice. Sci. Transl. Med. 2020;12. https://doi.org/10.1126/scitranslmed.abc5332. eabc5332.
[6] Zhu W, Chen CZ, Gorshkov K, Xu M, Lo DC, Zheng W. RNA-dependent RNA polymerase as a target for COVID-19 drug discovery. SLAS Discov 2020;25:1141–51. https://doi.org/10.1177/247255520942123.
[7] Mary YS, Mary YS, Yadav R, Celik I, Rad AS, Sarala SMD. DFT investigations and inhibition of the novel SARS-CoV-2 mainprotease in three cocrystals of hydrochloro-thiazide. Anal. Chem. Lett. 2021;11:450–68. https://doi.org/10.1080/22297928.2021.1934538.
[8] Yadav R, Hasan S, Mahato S, Celik I, Mary YS, Kumar A, Dhamija P, Sharma A, Choudhary N, Chaudhary PK, Kushwah AS, Chaudhary JK. Molecular docking, DFT analysis and dynamics simulation of natural bioactive compounds targeting ACE2 and TMPRSS2 dual binding sites of spike protein of SARS CoV-2. J. Mol. Liq. 2021:116942. https://doi.org/10.1016/j.molliq.2021.116942.
[9] Yadav R, Chaudhary JK, Jain N, Chaudhary PK, Khanra S, Dhamija P, Sharma A, Kumar A, Handu S. Role of structural and non-structural proteins and therapeutic targets of SARS-CoV-2 for COVID-19. Cells 2021;10:821. https://doi.org/10.3390/cells10040821.
[10] Yadav R, Imran M, Dhamija P, Chaurasia DK, Handu S. Virtual screening, ADMET prediction and dynamics simulation of potential compounds targeting the main protease of SARS-CoV-2. J. Biomol. Struct. Dyn. 2020;25:1–16. https://doi.org/10.1080/07391102.2020.1796812.
[11] Yadav R, Imran M, Dhamija P, Suchal K, Handu S. Virtual screening and dynamics of potential inhibitors targeting RNA binding domain of nucleocapsid phosphoprotein from SARS-CoV-2. J. Biomol. Struct. Dyn. 2021;39:4433–48. https://doi.org/10.1080/07391102.2020.1778536.
[12] Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-Stroh S, Melidou A, Neher RA, O'Toole A, Pereyaslov D. Geographical and temporal distribution of SARS-CoV-2 clads in the WHO European region, january to june 2020. Euro Surveill. 2020;25:2001410. https://doi.org/10.2807/1560-7917.ES.2020.25.32.2001410.
[13] Han AX, Parker E, Scholer F, Maurer-Stroh S, Russell CA. Phylogenetic clustering by linear integer programming (PhyCLIP). Mol. Biol. Evol. 2019;36:1580–95. https://doi.org/10.1093/molbev/msz053.
[14] https://www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses.
[15] https://www.gisaid.org/.
[16] https://bioinformatics.psb.ugent.be/webtools/Venn/.
[17] Huang Y, Yang C, Xu XF, Xu W, Liu SW. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol. Sin. 2020;41:1141–9. https://doi.org/10.1038/s41401-020-0485-4.
[18] Prajapat M, Sarma P, Shekhar N, Avti P, Sinha S, Kaur H, Kumar S, Bhattacharyya A, Kumar H, Bansal S, Medhi B. Drug targets for corona virus: a systematic review. Indian J. Pharmacol. 2020;52:56–65. https://doi.org/10.4103/ijp.IJP_115_20.
[19] Santerre M, Arjona SP, Allen CN, Shcheribik N, Sawaya BE. Why do SARS-CoV-2 NSPs rush to the ER? J. Neurol. 2021;268:2013–22. https://doi.org/10.1007/s00415-020-10197-8.
[20] Yoshimoto FK. The proteins of severe acute respiratory syndrome corona virus-2 (SARS-CoV-2 or n-COV19), the cause of COVID-19. Protein J. 2020;39:198–216. https://doi.org/10.1007/s10930-020-09901-4.
[21] Serrano P, Johnson MA, Chatterjee A, Neuman BW, Joseph JS, Buchmeirer MJ, Kuhn P, Wuthrich K. Nuclear magnetic resonance structure of the nucleic acid-binding domain of severe respiratory syndrome corona virus nonstructural protein 3. J. Virol. 2009;83:12998–3008. https://doi.org/10.1128/JVI.01253-09.
[22] Stertz S, Reichelt M, Spiegel M, Kuri T, Maretinez-Sobrido L, Garcia-Sastre A, Weber F, Kochs G. The intracellular sites of early replication and budding of SARS-corona virus. Virology 2007;361:304–15. https://doi.org/10.1016/j.virol.2006.11.027.
[23] Peng Q, Peng R, Yuan B, Zhao J, Wang M, Wang X, Wang Q, Sun Y, Fan Z, Qi J, Gao GF, Shi Y. Structural and biochemical characterization of the nsp12-nsp7-nsp8 core polymerase complex from SARS-CoV-2. Cell Rep. 2020;31:107774. https://doi.org/10.1016/j.celrep.2020.107774.
[24] Aftab SO, Ghouri MZ, Masood MU, Haider Z, Khan Z, Ahmad A, Munawar N. Analysis of SARS-CoV-2 RNA-dependent RNA polymerase as a potential therapeutic drug target using a computational approach. J. Transl. Med. 2020;18:275. https://doi.org/10.1186/s12967-020-02439-0.
[25] Benvenuto D, Angeletti S, Giovanetti M, Bianchi M, Pascarella S, Cauda R, Ciccozzi M, Cassone A. Evolutionary analysis of SARS-CoV-2: how mutation on Non-structural protein 6 (NSP6) could affect viral autophagy. J. Infect. 2020;81:e24–7. https://doi.org/10.1016/j.jinf.2020.03.058.
[26] Zinzula L. Lost in deletion: the enigmatic ORF8 protein of SARS-CoV-2. Biochem. Biophys. Res. Commun. 2021;538:116–24. https://doi.org/10.1016/j.bbrc.2020.10.045.
[27] McBride R, van Zyl M, Fielding BC. The corona virus nucleocapsid is a multifunctional protein. Viruses 2014;6:2991–3018. https://doi.org/10.3390/v6082991.