Geographical Distribution of Amino Acid Mutations in Human SARS-CoV-2 Orf1ab Poly-Proteins Compared to the Equivalent Reference Proteins from China

Dr. Kunchur Guruprasad, Ph.D

ABREAST™, Plot Nos.14/A & 15, Sitaramnagar, Safilguda, Hyderabad-500056, India

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

Website: https://www.abreast.in

ABSTRACT

The amino acid mutations among 28,345 poly-protein sequences corresponding to human SARS-CoV-2 orf1AB gene representing the six geographical locations; Africa, Asia, Europe, North America, Oceania and South America were identified by comparing with the equivalent reference poly-protein sequences derived from the first human SARS-CoV-2 genome sequence, reported from Wuhan-Hu-1, China. The mutations were analysed according to the following three datasets; i) 27,956 poly-proteins comprising 7,096 amino acid residues, ii) 373 poly-proteins comprising between 7,051-7,095 amino acid residues and iii) 16 poly-proteins comprising between 7,097-7,099 amino acid residues. In all, 3,204 distinct mutation sites were observed among the poly-proteins comprising 7,096 amino acid residues contributing to ~45% of the poly-protein sequence in SARS-CoV-2 orf1AB gene that have undergone mutations since the outbreak of COVID-19 pandemic disease in December 2019. Fifteen proteins of the poly-protein sequence were associated with mutations and the mutation propensities for the "leader protein", nsp2, nsp3, nsp6, nsp7, nsp8, endoRNAse proteins was higher (> 1) compared to nsp4, nsp9, nsp10, 3C-like proteinase, RdRp, helicase, 3'-to-5' exonuclease and 2'-O-ribose methyltransferase proteins. Relatively higher mutation percentages were observed for the RdRp (35.32%), nsp2 (26.42%), nsp3 (11.73%) and helicase (7.88%) proteins, whereas, mutation percentages for the remaining proteins ranged between 0.16% for nsp10 protein to 4.11% for the 3' -to-5' exonuclease proteins. Five mutations; T265I in nsp2 protein, T1246I in nsp3, G3278S in 3C-like proteinase, L3606F in nsp6 and P4715L in RdRp were common across all six geographical locations. The P4715L RdRp mutation was predominant in all geographical locations, except Africa, where G5215S mutation was predominant. The maximum number of distinct mutation sites were observed for the nsp3 protein. In 373 orf1AB poly-protein sequences comprising between 7,051-7,095 amino acid residues, deletion mutations were observed that were associated with "leader protein" between positions; 82-86

(GHVMV) and positions 141-143 (KSF). Among 16 orf1AB poly-proteins comprising between 7,097-7,099 amino acid residues, certain insertion mutations were observed that were associated with the nsp2 (517K), nsp3 (938E, 1901Y), 2' -O-ribose methyltransferase (7046F) and nsp6 (3610F, 3611L) proteins. In this work, all mutations observed among the 28,345 orf1AB poly-proteins of human SARS CoV-2 relative to the reference sequences are presented.

Keywords: human SARS-CoV-2; orf1AB poly-proteins; mutations; geographical locations; leader protein; nsp2; nsp3; nsp4; nsp6; nsp7; nsp8; nsp9; nsp10; 3C-like proteinase; RNA dependent RNA polymerase; helicase; 3'-to-5' exonuclease; endoRNAse; 2'-O-ribose methyltransferase.

INTRODUCTION

It is 18 months since outbreak of the COVID-19 pandemic disease was first reported from the city of Wuhan, Hubei-1 province, China in December 2019 (Wu et al., 2020). The severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) responsible for causing the COVID-19 disease has resulted in 191,229,700 coronavirus cases and 4,105,820 deaths worldwide (https://www.worldometers.info/coronavirus/). The SARS-CoV-2 is a positive sense single stranded RNA genome comprising 29,903 base pairs (NCBI Accession code: NC_045512.2) and the orf1AB gene expresses a poly-protein comprising 7096 amino acids (Yoshimoto, F.K. et al., 2020). The gene codes for non-structural proteins (NSPs) represented by "leader protein" (180 amino acid residues), nsp2 (638 aa), nsp3 (1945 aa), nsp4 (500 aa), nsp5 or 3C-like proteinase (306 aa), nsp6 (290 aa), nsp7 (83 aa), nsp8 (198 aa), nsp9 (113 aa), nsp10 (139 aa), nsp11 (13 aa), nsp12 or RNA dependent RNA polymerase (RdRp) (932 aa), nsp13 or helicase (601 aa), nsp14 or 3'-to-5' exonuclease (527 aa), nsp15 or endoRNAse (346 aa) and nsp16 or 2'-O-ribose methyltransferase (298 aa). The poly-proteins of the SARS-CoV-2 orf1AB gene are suggested to play a role in virus pathogenesis distinct from or in addition to functions directly involved in viral replication (Graham, R.L. et al., 2008). Viruses are known to undergo mutations facilitating their survival in host by evading host immune recognition. Early on during the outbreak of the COVID-19 pandemic, I reported mutations observed in the protein sequences obtained in the genome sequences available in the NCBI databank (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) for 22 infected individuals with the novel coronavirus in India between January 2020 to mid-April 2020 (Guruprasad, 2020a). The observed mutations were also mapped on to the protein three-dimensional structures for the RdRp, helicase, endoRNAse and the spike proteins and the secondary structure of the proteins corresponding to the mutations, proximity of the mutations to the functionally important residues in the different proteins and to the remdesivir and tipiracil drug binding sites in RdRp and endoRNAse targets, respectively, were analysed (Guruprasad, 2020b). The SARS-CoV-2 has been undergoing mutations world-wide since the outbreak of COVID-19 pandemic disease. The mutations present in 9 samples of the human SARS-CoV-2 from Eastern India were reported (Maitra et al., 2020). In another study, 220 complete SARS-CoV-2 genome sequences derived from patients infected by the SARS-CoV-2 worldwide between December 2019 to mid-March 2020 were randomly collected from the GISAID database and 8 novel recurrent nucleotide mutations were identified (Pachetti et al., 2020). In yet another study, the human SARS-CoV-2 orf1AB poly-proteins corresponding to 1,134 complete sequences derived from

affected patients in USA between December 2019 to April 2020 reported 4 significant mutations (Banerjee, S., 2021). Mutations corresponding to 1,516 variations at the nucleotide level at different positions throughout the genome were reported by analysing 2,492 genomic sequences of SARS-CoV-2 strains (Islam, M.R. et al., 2020). Khailany, R.A. et al., 2020 reported 116 mutations among 95 complete SAR-CoV-2 genome sequences. The mutations in 908 SARS-CoV-2 orf1AB poly-proteins from North America were carried out to study the effects of certain mutations in RdRp and helicase proteins on the structure, dynamics and function (Begum et al., 2020). Further, the SARS CoV and SARS-CoV-2 sequences from 13 different countries have been compared in order to identify and study the effects of mutations on major target proteins (Khan, M.I. et al., 2020). These studies indicate that the knowledge of the known mutations in proteins of human SARS-CoV-2 are useful from the perspective of understanding structure, function and drug-resistance.

In the present work, I have revisited my previous study (Guruprasad, 2020c) and analysed the mutations present in poly-proteins of the human SARS-CoV-2 orf1AB gene in an enlarged dataset comprising 28,345 genomes. The distinct mutation sites and mutation types observed among the different proteins are identified with respect to the reference sequence and a catalogue of the mutations is presented that serves to recognize protein-specific and geographical location-specific mutations and mutation ‘hot-spots’. The knowledge of mutations observed in poly-proteins of the human SARS-CoV-2 orf1AB gene is useful to analyse their likely impact on the structure and function of the proteins. Also, the known mutations in a protein target would be useful to consider while designing drugs and is therefore of pharmaceutical relevance in studies to treat COVID-19 disease.

MATERIALS AND METHODS

The protein sequences corresponding to the human SARS-CoV-2 orf1AB gene were obtained from the NCBI databank (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/). Three datasets were prepared: i) comprising human SARS-CoV-2 orf1AB poly-protein sequences of length 7,096 amino acid residues, ii) poly-protein sequences of length comprising between 7,051-7,095 amino acid residues and iii) poly-protein sequences of length comprising between 7,097-7,099 amino acid residues. The multiple sequence alignments were generated separately for the different datasets using the NGphylogeny.fr suite (Lemoine et al., 2019) available at (https://ngphylogeny.fr) and using the human SARS-CoV-2 orf1AB poly-protein sequence (NCBI code: YP_009724389.1) from the Wuhan-Hu-1, China isolate as the reference sequence. The identification and analyses of mutations in the individual datasets were carried out using the in-house computer software developed at ABREAST™ (https://www.abreast.in).

RESULTS AND DISCUSSION

The human SARS-CoV-2 orf1AB poly-protein sequences comprised between 7,051-7,099 amino acid residues with the predominant occurrence of sequences comprising 7,096 amino acid residues as in the reference sequence from Wuhan, China (NCBI code: YP_009724389.1).

Mutations in poly-proteins of the human SARS CoV-2 orf1AB gene (comprising 7,096 amino acid residues)

The poly-protein sequence data corresponding to the human SARS-CoV-2 orf1AB gene analysed in the present work was ~2.55 times the data in previous analyses (Guruprasad, 2020c) and is represented in **Figure 1**. The data comprised a total of 27,956 human SARS-CoV-2 poly-proteins corresponding to the six geographical locations as follows; Africa (290), Asia (1523), Europe (581), North America (16039), Oceania (9396) and South America (127). These sequences represented a total of 79,040 mutations distributed according to the geographical locations as: Africa (472), Asia (3774), Europe (1229), North America (48442), Oceania (24787) and South America (336). The mutations represented a total of 4,841 distinct mutation positions along the SARS-CoV-2 orf1AB poly-protein sequence and distributed as: Africa (114), Asia (873), Europe (205), North America (2620), Oceania (931) and South America (98). Excluding redundant mutations observed across geographical locations, the total number distinct mutation positions reduced to 3,204 mutations. Thereby, ~45% human SARS-CoV-2 poly-protein orf1AB gene has undergone mutations since outbreak of the COVID-19 pandemic in December 2019. Fifteen proteins of the poly-protein sequence were associated with mutations and their protein-wise distribution is shown in **Figure 2.** The mutation percentages are presented in **Table 1** and the RdRp protein was observed to be associated with the maximum number of mutations (35.32%), followed by nsp2 protein (26.42%), nsp3 protein (11.73%) and the helicase (7.88%) proteins. The mutation percentages for remaining proteins ranged between (0.16%) for nsp10 to (4.11%) for 3'-to-5' exonuclease proteins. The relatively large number of mutations observed for the RdRp, nsp2, nsp3 and helicase proteins was consistent with previous analysis based on 10,929 human SARS-CoV-2 orf1AB poly-protein sequences (Guruprasad, 2020c). The protein-wise distribution of the total number of mutations

according to the geographical locations is shown in **Figure 3.** The RdRp, helicase, nsp2, nsp3 proteins represented in the human SARS-CoV-2 orf1AB gene from North America and the RdRp and nsp2 proteins represented from Oceania showed relatively large number of mutations as compared to other proteins and to the mutations in proteins representing other geographical locations. The total number of distinct mutation sites observed in poly-proteins of human SARS-CoV-2 orf1AB gene representing the different geographical locations is listed in **Table 2**. Accordingly, the nsp3 protein is associated with the maximum number of distinct mutation sites along the poly-protein sequence across all the six geographical locations. As more than one mutation type could be observed at some of the mutation sites, the total number of mutations increased to 5,631. These mutations were distributed among the different geographical locations as: Africa (114), Asia (935), Europe (208), North America (3278), Oceania (998) and South America (98). The mutation sites along with the mutation types are catalogued in the **Supplementary Table 1.** This table can be used to consult the presence or absence of specific mutations in proteins corresponding to the different geographical locations and to identify the mutation 'hot-spots'. The RdRp protein among all other proteins in the poly-protein sequence was associated with the maximum mutation percentages across all six geographical locations as shown in **Table 3**, with the maximum mutation percentage observed in Africa (59.74%). However, among the remaining proteins there appeared to be a preference according to the geographical location as shown in **Figure 4**. For instance, the next most mutated protein after RdRp was the nsp3 proteins from Africa (12.9%), Asia (19.2%), Europe (18.3%) and South America (23.5%), whereas, it was the nsp2 protein in North America (22.4%) and Oceania (37.2%) and the 3C-like proteinase in South America (16.96%) (refer Table 3). According to the mutation propensities calculated as the ratio of mutation frequency to the protein sequence length frequency and represented in **Figure 5**, the proteins associated with relatively higher mutation propensity (> 1) were: "leader protein", nsp2, nsp3, nsp6, nsp7, nsp8, endoRNAse as compared to the nsp4, nsp9, nsp10, 3C-like proteinase, RdRp, helicase, 3'-to-5' exonuclease and 2'-O-ribose methyltransferase proteins. The "leader protein" was associated with the maximum mutation propensity and nsp10 protein with least mutation propensity.

Certain mutation sites were associated with more than one mutation type. The total number of such mutation sites according to the geographical locations observed were: Asia (59), North America (575), Europe (3) and Oceania (66). A list of these mutation sites and the corresponding mutations is attached in **Supplementary Table 2.** It was noticed that the N-

terminal 117 amino acid residues corresponding to the "leader protein" in human SARS-CoV-2 orf1AB gene from Europe was not associated with mutations.

The distribution of the total number of mutations observed in the poly-proteins of human SARS-CoV-2 orf1AB gene from Africa, Asia, Europe, North America, Oceania, South America are shown in **Figures 6-11**, respectively. The mutations occurring ≥ 25% were: G5215S (30.5%), P4715L (25%) from Africa, P4715L/F (30%) from Asia, P4715L (38.7%) from Europe, P4715L/F/H (27.6%) from North America, P4715L (36.8%), I300F (35.3%) from Oceania and P4715L (33.3%) from South America. Five mutations were common to all six geographical locations. These were; T265I (nsp2), T1246I (nsp3), G3278S (3C-like proteinase), L3606F (nsp6) and P4715L (RdRp). The T1246I, G3278S common mutations were new observations in the present dataset. Also, it was observed in the present dataset, that the P4715L mutation represented the predominant mutation across all geographical locations, except Africa, where the G5215S mutation was predominant and in Oceania, the I300F represented the maximum mutation percentage. The P4715L mutation is associated with the 'interface domain' and the G5215S mutation with the Thumbs sub-domain in the RdRp protein structure according to the mutation sites mapped along the human SARS-CoV-2 RdRp protein sequences (Guruprasad, 2021). According to the regions along the human SARS-CoV-2 orf1AB poly-protein sequences that are associated with relatively higher mutations as shown in **Figures 6-11** for the different geographical locations, it appears there is a certain similarity in the patterns shared for Asia and Europe. Likewise, there is certain similarity shared in the patterns observed for North America and Oceania. These patterns are different as observed for either Africa or South America, suggesting that the SARS-CoV-2 poly-proteins may have a mutational preference that is geographically dependent.

Mutations in the poly-proteins of human SARS-CoV-2 orf1AB gene (comprising 7051-7095 amino acids)

In 373 human SARS-CoV-2 orf1AB poly-protein sequences, the length of the poly-protein sequence varied between 7,051-7,095 amino acid residues. These were geographically represented as follows: Africa (2), Asia (12), Europe (15), North America (234), Oceania (108) and South America (2). A total of 146 deletion mutation positions were observed in these poly-proteins besides other mutations. The deletion mutations were observed in: ["leader protein"]; V54, GHVMV (82-86), Y136, KSF (141-143), QENW (158-161), [nsp2]; N389, GLNDNL

(445-450), V649, T770, C784, [nsp3]; V868, GQQDGSE (989-995), IEVN (1023-1026), F1028, G1095, EIPKEEVKPF (1205-1214), QRK (1224-1226), S1539, NF (1927-1928), LKSEDAQGMDNLACED (2028-2043), NNSLKIT (2081-2087), Y2141, INI (2227-2229), [nsp4]; FLF (2780-2782), FDTW (2834-2837), G2879, FYWFF (3153-3157), L3198, [nsp6]; MVD (3669-3671), SGF (3675-3677), F3760, [nsp7]; T3904, [RdRp]; YHFRELGVV (4738-4746), [3' -to-5' exonuclease]; YV (6345-6346), FFY (6369-6371), LYLDAYN (6418-6424), LWV (6433-6435), [endoRNAse]; L6679, F6692, I6721, and [2'-O-ribose methyltransferase]; RENNRVVISSDVLVNN (7081-7096). The "leader protein" was associated with maximum number of deletion mutations and between (141-143) KSF and (82-86) GHVMV. However, deletion mutations were not observed in 3C-like proteinase, nsp8, nsp9, nsp10 and helicase proteins. The genome accession codes according to the NCBI database corresponding to the deletion mutations along with the other mutations observed is provided in **Supplementary Table 3**.

Mutations in poly-proteins of human SARS CoV-2 orf1AB gene (comprising 7097-7099 amino acids)

In 16 human SARS-CoV-2 orf1AB poly-protein sequences, the length of the sequence varied between 7,097-7,099 amino acid residues. The geographical distribution of these sequences were as follows: Asia (1), North America (11) and Oceania (4). In these poly-protein sequences certain insertion mutations, besides other mutations were observed with respect to the reference protein sequences from Wuhan-Hu-1, China. According to sequential numbering for these poly-protein sequences, the insertion mutations were observed at 517K (nsp2), at 938E, 1901Y (nsp3), at 3610F, 3611L (nsp6), and at 7046F (2'-O-ribose methyltransferase). The genome accession codes according to the NCBI database corresponding to the insertion mutations along with the other mutations observed is provided in **Supplementary Table 4**.

CONCLUSIONS

The comparison of 28,343 human SARS-CoV-2 orf1AB poly-protein sequences representing six different geographical locations with the equivalent reference protein sequences from Wuhan-Hu-1, China, revealed the presence of 3,204 distinct mutation sites. The mutations were observed among fifteen proteins of the human SARS-CoV-2 orf1AB poly-protein gene. The RdRp, nsp2, nsp3 and helicase proteins were associated with relatively large number of mutations compared to the other proteins. Five mutations; T265I in nsp2, T1246I in nsp3, G3278S in 3C-like proteinase, L3606F in nsp6 and P4715L in RdRp were common to all six geographical locations; Africa, Asia, Europe, North America, Oceania and South America. The maximum number of distinct mutation sites was observed for the nsp3 protein. The P4715L mutation in RdRp protein was maximally observed across all the geographical locations, except Africa, where the G5215S mutation was predominant. All the mutation sites and mutation types observed in the present work are catalogued. Nearly, 45% human SARS-CoV-2 orf1AB poly-protein sequence has undergone mutations since outbreak of the COVID-19 pandemic disease first reported from the city of Wuhan, China during December 2019.

REFERENCES

Banerjee, S., Seal, S., Dey, R., Mondal, K. K., & Bhattacharjee, P. (2021). Mutational spectra of SARS-CoV-2 orf1ab polyprotein and signature mutations in the United States of America. *Journal of Medical Virology*, *93*(3), 1428-1435.

Begum, F., Mukherjee, D., Das, S., Thagriki, D., Tripathi, P. P., Banerjee, A. K., & Ray, U. (2020). Specific mutations in SARS-CoV2 RNA dependent RNA polymerase and helicase alter protein structure, dynamics and thus function: Effect on viral RNA replication. *bioRxiv*.

Graham, R. L., Sparks, J. S., Eckerle, L. D., Sims, A. C., & Denison, M. R. (2008). SARS coronavirus replicase proteins in pathogenesis. *Virus research*, *133*(1), 88-100.

Guruprasad, Kunchur (2020a): Amino Acid Mutations in the Protein Sequences of Human SARS CoV-2 Indian Isolates Compared to Wuhan-Hu-1 Reference Isolate from China. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.12300860.v1

Guruprasad K (2020b). Mapping Mutations in Proteins of SARS CoV-2 Indian Isolates on to the Three-Dimensional Structures. ChemRxiv. Cambridge: Cambridge Open Engage; 2020; This content is a preprint and has not been peer-reviewed. https://doi.org/10.26434/chemrxiv.12683771.v1

Guruprasad, Kunchur (2020c): Geographical Distribution of Amino Acid Mutations in Human SARS-CoV-2 Orf1ab Poly-Proteins Compared to the Equivalent Reference Proteins from China. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.12951617.v1

Guruprasad, Kunchur (2021): Mutation Sites Are Distant from Remdesivir Binding Site in Human SARS-CoV-2 RNA-Dependent RNA Polymerase. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.13515116.v1

Islam, M. R., Hoque, M. N., Rahman, M. S., Alam, A. R. U., Akther, M., Puspo, J. A., ... & Hossain, M. A. (2020). Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. *Scientific reports*, *10*(1), 1-9.

Khailany, R. A., Safdar, M., & Ozaslan, M. (2020). Genomic characterization of a novel SARS-CoV-2. *Gene reports*, *19*, 100682.

Khan, M. I., Khan, Z. A., Baig, M. H., Ahmad, I., Farouk, A. E., Song, Y. G., & Dong, J. J. (2020). Comparative genome analysis of novel coronavirus (SARS-CoV-2) from different geographical locations and the effect of mutations on major target proteins: An in silico insight. *PLoS One*, *15*(9), e0238344.

Lemoine, F., Correia, D., Lefort, V., Doppelt-Azeroual, O., Mareuil, F., Cohen-Boulakia, S., & Gascuel, O. (2019). NGPhylogeny. fr: new generation phylogenetic services for non-specialists. *Nucleic acids research*, *47*(W1), W260-W265.

Maitra, A., Sarkar, M. C., Raheja, H., Biswas, N. K., Chakraborti, S., Singh, A. K., ... & Ghosh, T. (2020). Mutations in SARS-CoV-2 viral RNA identified in Eastern India: Possible implications for the ongoing outbreak in India and impact on viral structure and host susceptibility. *Journal of Biosciences*, *45*(1).

Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., ... & Zella, D. (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of Translational Medicine*, *18*, 1-9.

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

Yoshimoto, F. K. (2020). The proteins of severe acute respiratory syndrome coronavirus-2 (SARS CoV-2 or n-COV19), the cause of COVID-19. *The protein journal*, *39*, 198-216.

ACKNOWLEDGEMENTS

The author acknowledges several researchers and sequencing centres for making publicly available the complete genomes of human SARS-CoV-2.

CONFLICT OF INTEREST

The author declares no conflict of interest

FUNDING

None

TABLE 1. Mutation percentages in poly-proteins of the human SARS-CoV-2 orf1AB sequence

Proteins	No	Total number of Mutations	Mutation percentage
leader protein	1	1282	1.621964
nsp2	2	20890	26.42966
nsp3	3	9277	11.7371
nsp4	4	1321	1.671306
3C-like proteinase	5	2157	2.728998
nsp6	6	1985	2.511387
nsp7	7	776	0.981781
nsp8	8	720	0.910931
nsp9	9	239	0.302379
nsp10	10	127	0.160678
RdRp	11	27919	35.32262
helicase	12	6233	7.885881
3’-to-5’ exonuclease	13	3249	4.110577
endoRNAse	14	1417	1.792763
2’-O-ribose methyltransferase	15	1448	1.831984

TABLE 2. Geographical distribution of the total number of distinct mutation sites in poly-proteins of the human SARS-CoV-2 orf1AB gene comprising 7,096 amino acid residues.

Human SARS-CoV-2 orf1AB poly-proteins	AFRICA	ASIA	EUROPE	NORTH AMERICA	OCEANIA	SOUTH AMERICA
leader protein	2	24	1	89	33	4
nsp2	11	103	20	329	95	13
nsp3	32	257	76	744	261	30
nsp4	8	55	9	163	68	5
3C-like proteinase	3	38	11	106	35	4
nsp6	3	39	2	102	41	4
nsp7	3	7	5	35	13	2
nsp8	2	34	4	63	23	0
nsp9	0	11	4	42	13	2
nsp10	1	11	3	36	7	3
RdRp	16	86	25	280	96	10
helicase	13	65	13	204	90	7
3’ -to-5’ exonuclease	9	57	11	189	75	6
endoRNAse	5	51	14	143	48	5
2’ -O-ribose methyltransferase	6	35	7	95	33	3
TOTAL	114	873	205	2620	931	98

TABLE 3. Protein-wise mutation percentages according to the different geographical locations

Proteins	AFRICA	ASIA	EUROPE	N_AMERICA	OCEANIA	S_AMERICA
leader protein	0.423728	1.059883	0.081366	2.190248	0.701980	1.190476
nsp2	5.932203	16.42819	11.79820	22.41443	37.20498	5.059523
nsp3	12.92372	19.21038	18.30756	12.80293	8.008230	23.51190
nsp4	3.177966	2.517223	1.952807	1.709260	1.407996	2.976190
3C-like proteinase	3.389830	2.252252	5.044751	3.711655	0.560777	16.96428
nsp6	1.059322	5.034446	3.905614	2.437967	2.231008	2.380952
nsp7	1.059322	0.370959	0.650935	1.414062	0.246096	0.892857
nsp8	2.330508	1.298357	1.464605	0.986747	0.661637	0
nsp9	0	0.423953	0.488201	0.386028	0.112962	0.595238
nsp10	0.211864	0.370959	0.244100	0.179596	0.076653	0.892857
RdRp	59.74576	36.22151	46.05370	30.89674	42.80873	37.5
helicase	4.449152	3.285638	2.115541	11.54783	1.855811	2.380952
3’-to-5’ exonuclease	2.118644	5.988341	1.220504	4.541513	3.183120	2.678571
endoRNAse	1.906779	3.709591	5.858421	2.171669	0.560777	1.488095
2’-O-ribose methyltransferase	1.271186	1.828298	0.813669	2.609305	0.379231	1.488095

FIGURE 1

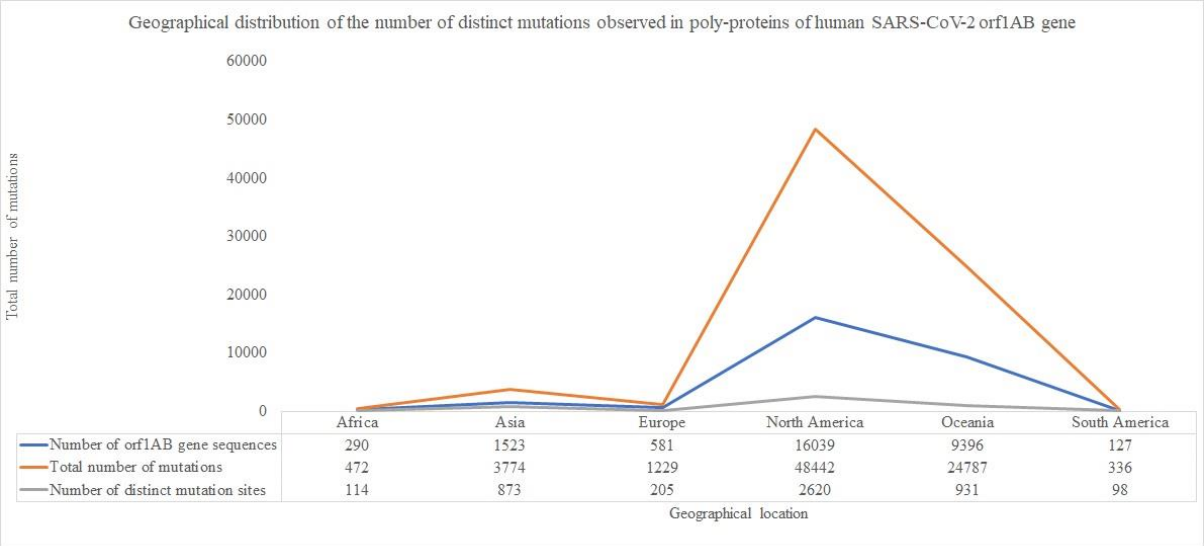

Geographical distribution of the number of distinct mutations observed in poly-proteins of human SARS-CoV-2 orf1AB gene
Total number of mutations
60000
50000
40000
30000
20000
10000
0
Africa Asia Europe North America Oceania South America
Number of orf1AB gene sequences 290 1523 581 16039 9396 127
Total number of mutations 472 3774 1229 48442 24787 336
Number of distinct mutation sites 114 873 205 2620 931 98
Geographical location

FIGURE 2

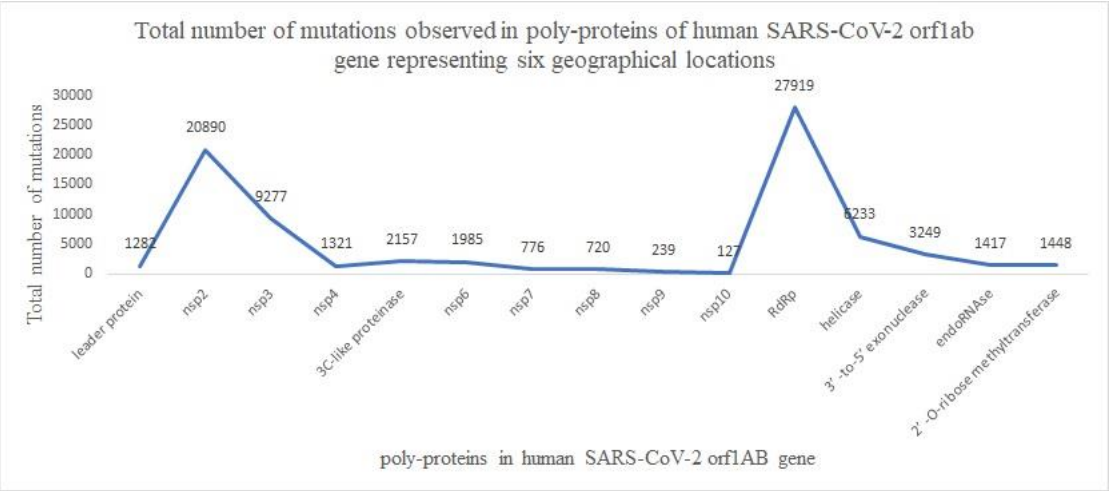
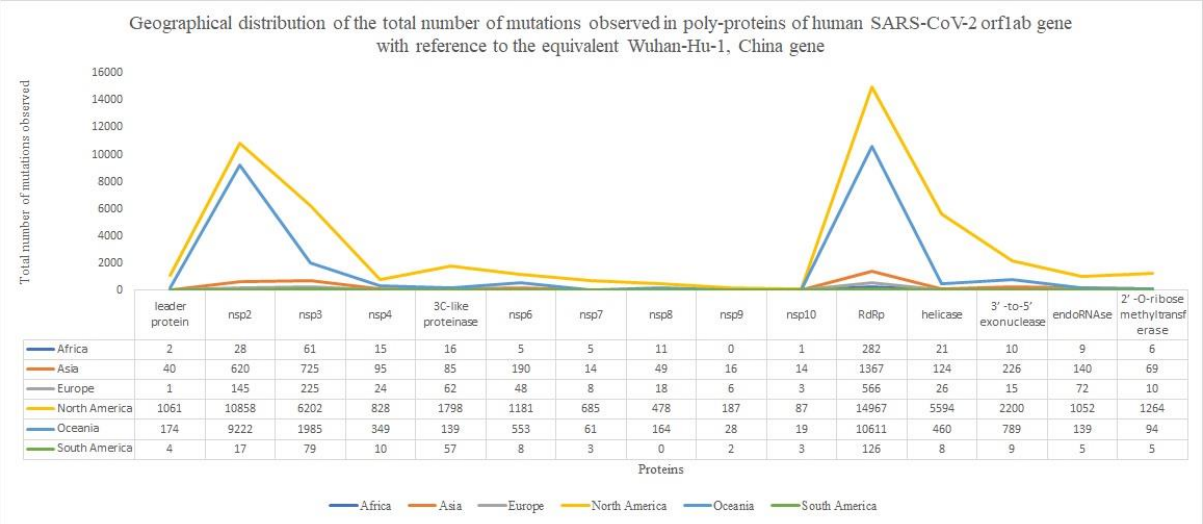

Total number of mutations observed in poly-proteins of human SARS-CoV-2 orf1ab gene representing six geographical locations
Total number of mutations
30000
25000
20000
15000
10000
5000
0
1282
20890
9277
1321
2157
1985
776
720
239
127
27919
5233
3249
1417
1448
leader protein
nsp2
nsp3
nsp4
3C-like proteinase
nsp6
nsp7
nsp8
nsp9
nsp10
RdRp
helicase
3'-to-5' exonuclease
endoRNAse
2'-O-ribose methyltransferase
poly-proteins in human SARS-CoV-2 orf1AB gene

FIGURE 3



	leader protein	nsp2	nsp3	nsp4	3C-like proteinase	nsp6	nsp7	nsp8	nsp9	nsp10	RdRp	helicase	3′ -to-5′ exonuclease	endoRNAse	2′ -O-ribose methyltransferase
Africa	2	28	61	15	16	5	5	11	0	1	282	21	10	9	6
Asia	40	620	725	95	85	190	14	49	16	14	1367	124	226	140	69
Europe	1	145	225	24	62	48	8	18	6	3	566	26	15	72	10
North America	1061	10858	6202	828	1798	1181	685	478	187	87	14967	5594	2200	1052	1264
Oceania	174	9222	1985	349	139	553	61	164	28	19	10611	460	789	139	94
South America	4	17	79	10	57	8	3	0	2	3	126	8	9	5	5

FIGURE 4

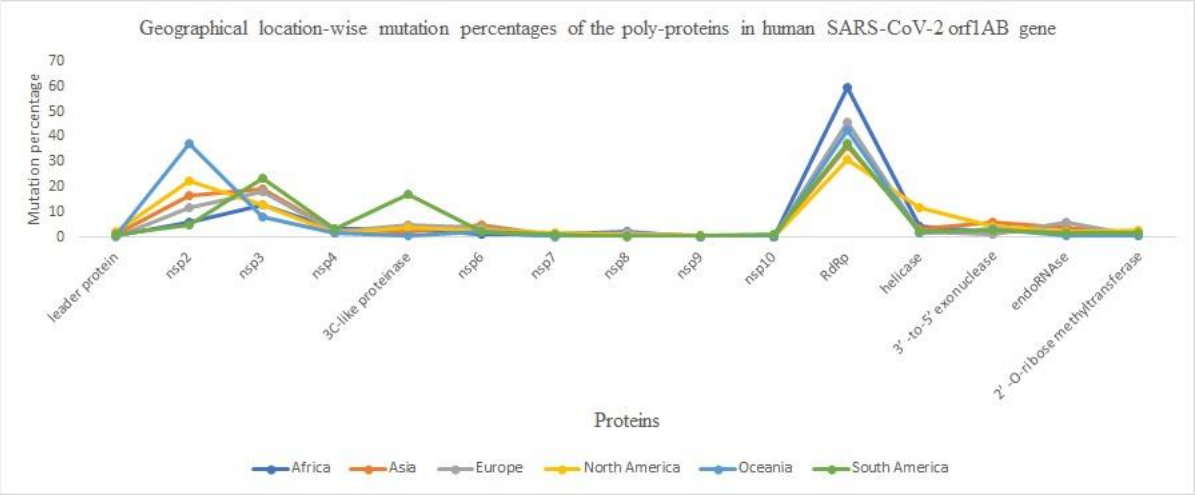

Geographical location-wise mutation percentages of the poly-proteins in human SARS-CoV-2 orf1AB gene
Mutation percentage
70
60
50
40
30
20
10
0
leader protein
nsp2
nsp3
nsp4
3C-like proteinase
nsp6
nsp7
nsp8
nsp9
nsp10
RdRp
helicase
3' -to-5' exonuclease
endoRNAse
2' -O-ribose methyltransferase
Proteins
Africa
Asia
Europe
North America
Oceania
South America

FIGURE 5

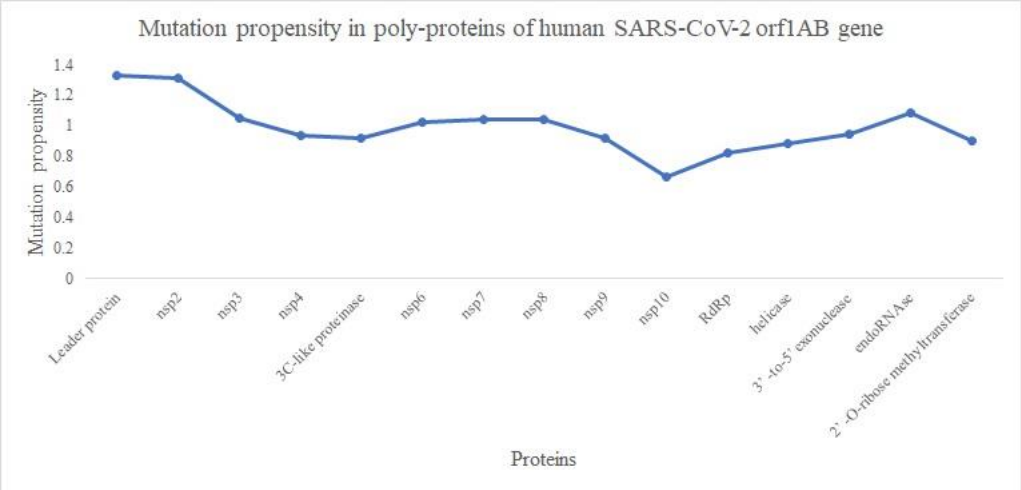

Mutation propensity in poly-proteins of human SARS-CoV-2 orf1AB gene
Mutation propensity
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Leader protein
nsp2
nsp3
nsp4
3C-like proteinase
nsp6
nsp7
nsp8
nsp9
nsp10
RdRp
helicase
3' -to-5' exonuclease
endoRNAse
2' -O-ribose methyltransferase
Proteins

FIGURE 6

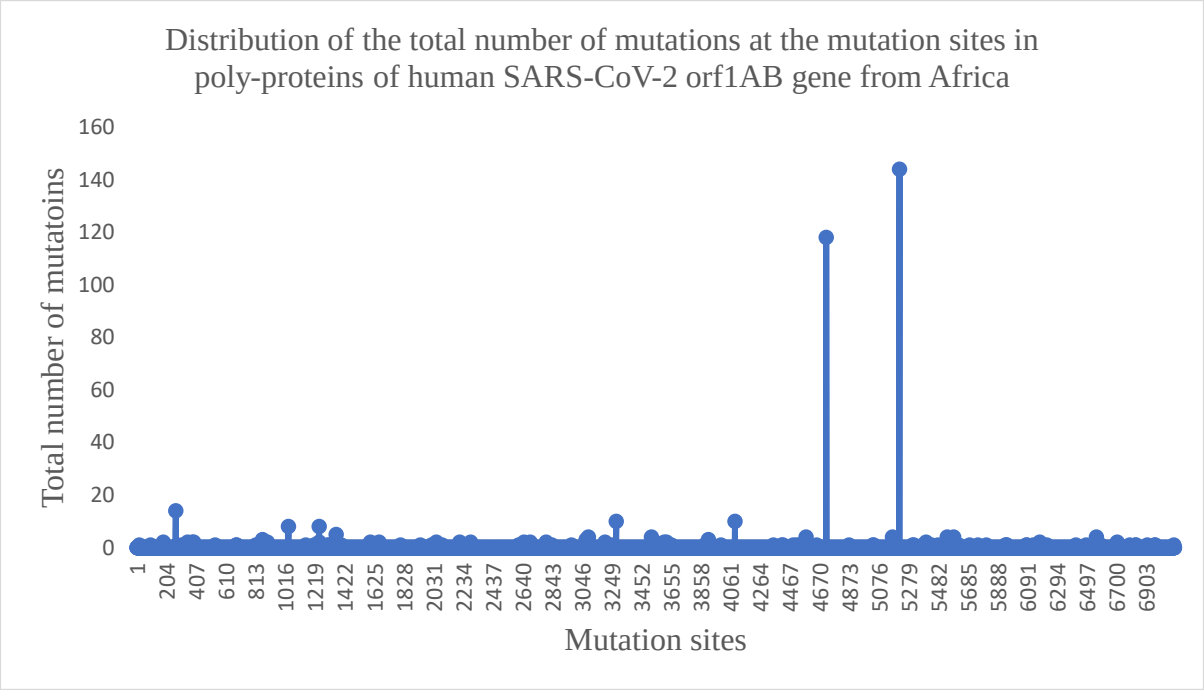

Distribution of the total number of mutations at the mutation sites in poly-proteins of human SARS-CoV-2 orf1AB gene from Africa
Total number of mutatoins
160
140
120
100
80
60
40
20
0
1
204
407
610
813
1016
1219
1422
1625
1828
2031
2234
2437
2640
2843
3046
3249
3452
3655
3858
4061
4264
4467
4670
4873
5076
5279
5482
5685
5888
6091
6294
6497
6700
6903
Mutation sites

FIGURE 7

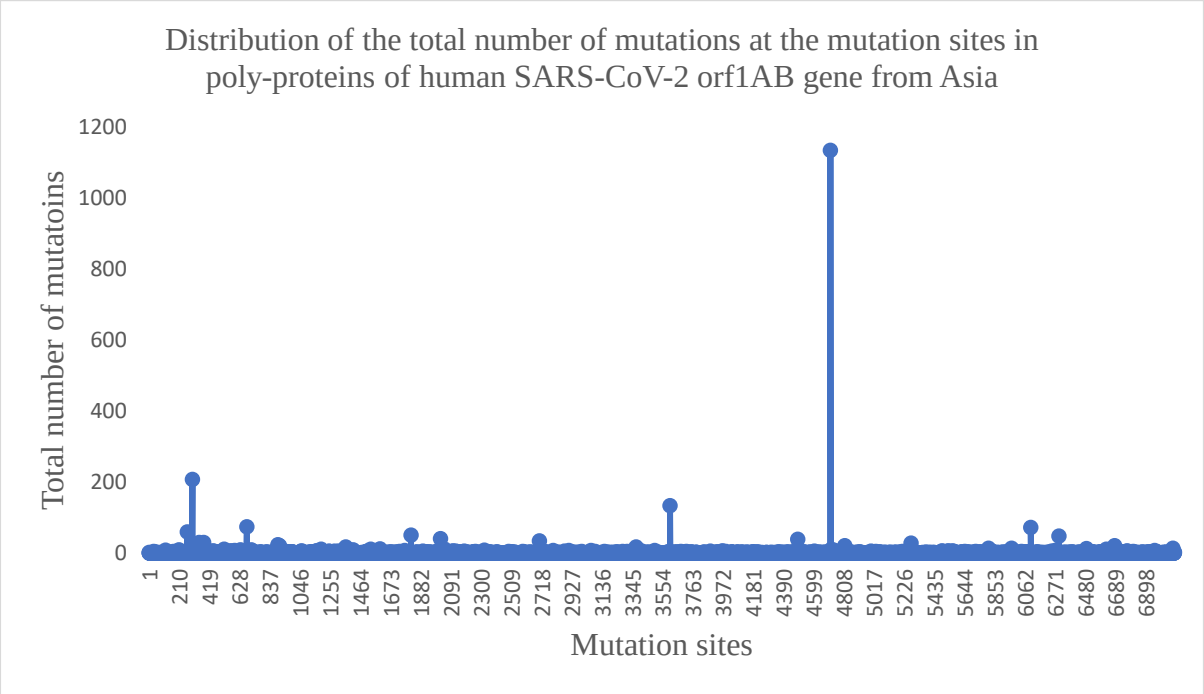

Distribution of the total number of mutations at the mutation sites in poly-proteins of human SARS-CoV-2 orf1AB gene from Asia
Total number of mutatoins
1200
1000
800
600
400
200
0
1
210
419
628
837
1046
1255
1464
1673
1882
2091
2300
2509
2718
2927
3136
3345
3554
3763
3972
4181
4390
4599
4808
5017
5226
5435
5644
5853
6062
6271
6480
6689
6898
Mutation sites

FIGURE 8

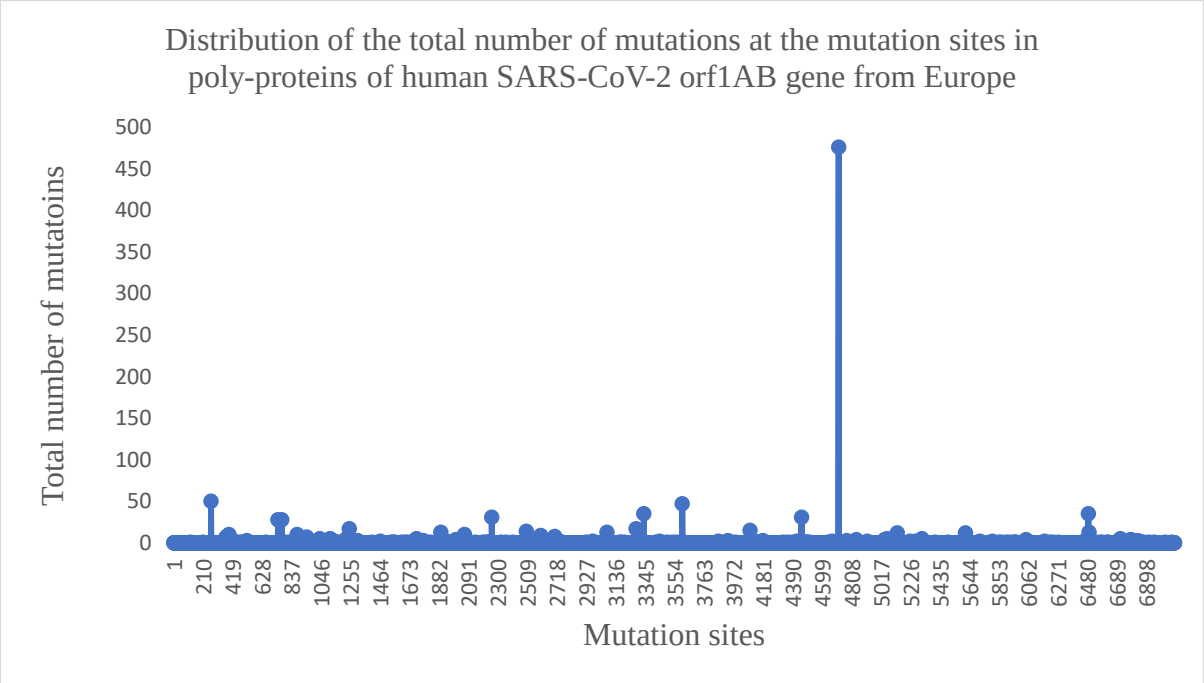

Distribution of the total number of mutations at the mutation sites in poly-proteins of human SARS-CoV-2 orf1AB gene from Europe
Total number of mutatoins
500
450
400
350
300
250
200
150
100
50
0
1
210
419
628
837
1046
1255
1464
1673
1882
2091
2300
2509
2718
2927
3136
3345
3554
3763
3972
4181
4390
4599
4808
5017
5226
5435
5644
5853
6062
6271
6480
6689
6898
Mutation sites

FIGURE 9

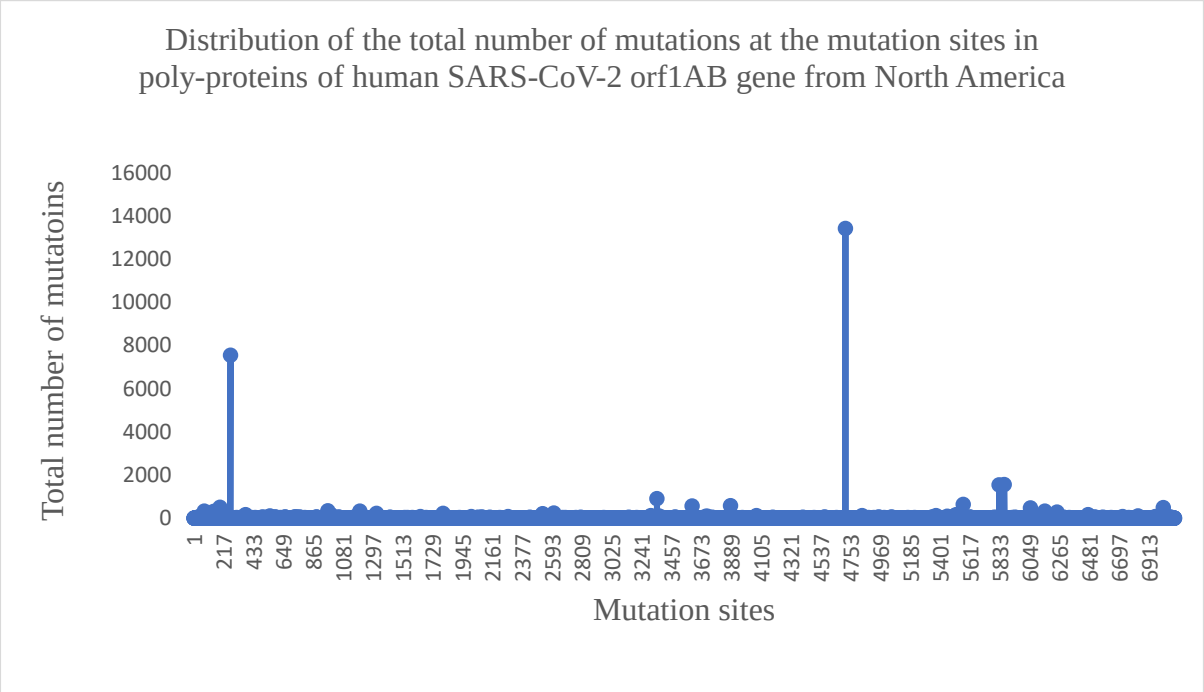

Distribution of the total number of mutations at the mutation sites in poly-proteins of human SARS-CoV-2 orf1AB gene from North America
Total number of mutatoins
16000
14000
12000
10000
8000
6000
4000
2000
0
1
217
433
649
865
1081
1297
1513
1729
1945
2161
2377
2593
2809
3025
3241
3457
3673
3889
4105
4321
4537
4753
4969
5185
5401
5617
5833
6049
6265
6481
6697
6913
Mutation sites

FIGURE 10

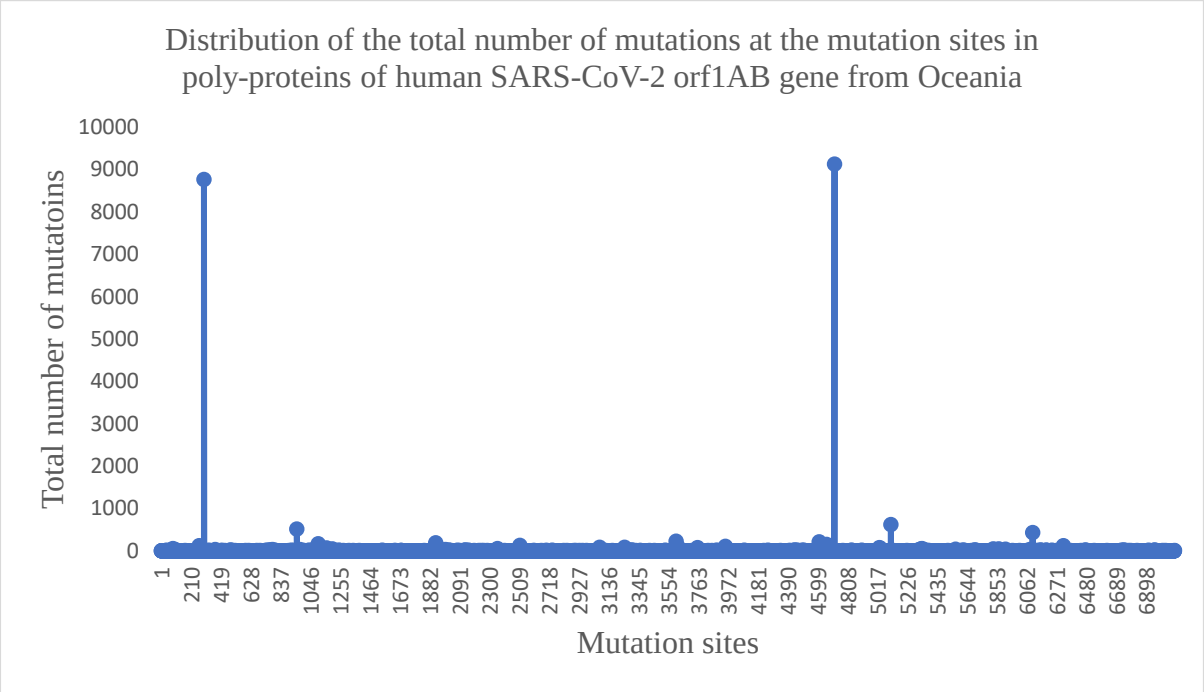
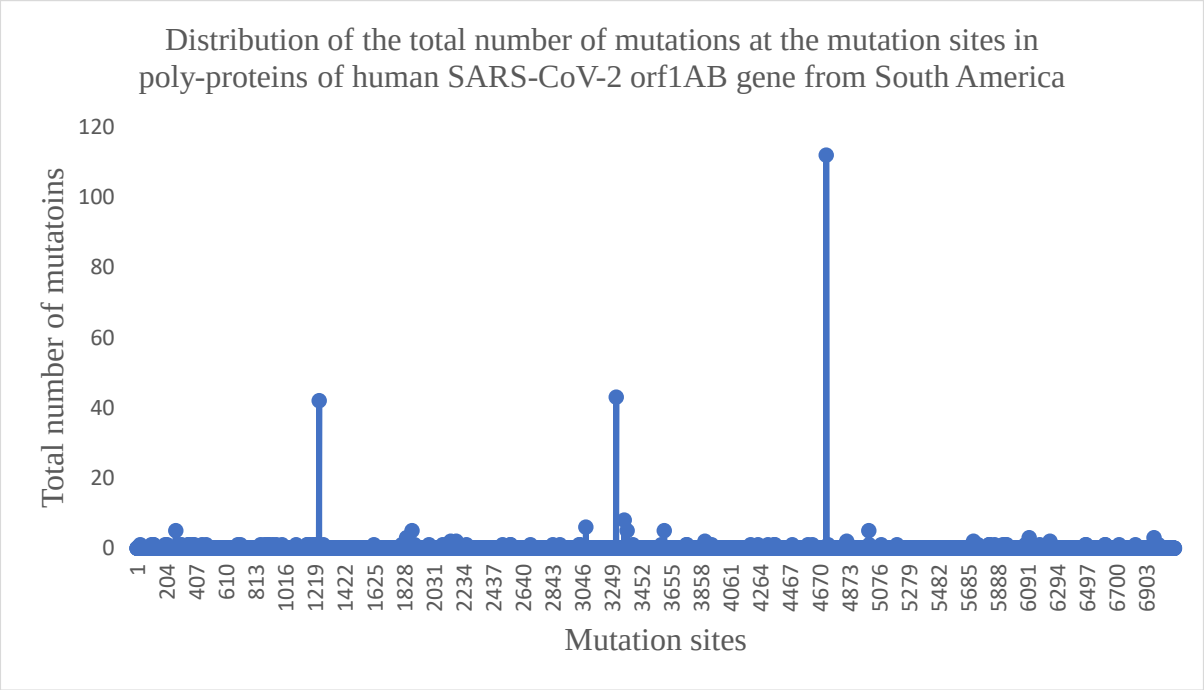

Distribution of the total number of mutations at the mutation sites in poly-proteins of human SARS-CoV-2 orf1AB gene from Oceania
Total number of mutatoins
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1
210
419
628
837
1046
1255
1464
1673
1882
2091
2300
2509
2718
2927
3136
3345
3554
3763
3972
4181
4390
4599
4808
5017
5226
5435
5644
5853
6062
6271
6480
6689
6898
Mutation sites

FIGURE 11


Distribution of the total number of mutations at the mutation sites in poly-proteins of human SARS-CoV-2 orf1AB gene from South America
Total number of mutatoins
120
100
80
60
40
20
0
1
204
407
610
813
1016
1219
1422
1625
1828
2031
2234
2437
2640
2843
3046
3249
3452
3655
3858
4061
4264
4467
4670
4873
5076
5279
5482
5685
5888
6091
6294
6497
6700
6903
Mutation sites