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

Mobile: +91 7337554324

Website: https://www.abreast.in

Abstract:

Mutations in orf1ab poly-protein sequences from human SARS-CoV-2 isolates representing six geographical locations were identified by comparing with the equivalent reference sequences from the Wuhan-Hu-1, China isolate, epicentre of the current COVID-19 pandemic disease. The orf1ab poly-proteins of sequence length 7096 amino acid residues representing 10,929 genomes from six geographical locations comprised a total of 27,895 mutations that corresponded to 2,095 distinct mutation sites. The percentage of mutations was significantly high for RdRp (33.47%), nsp2 (20.04%), helicase (15.95%) and nsp3 (12.61%) proteins, compared to rest of the proteins which ranged between (0.14%) for nsp10 to (2.79%) for nsp6 proteins. A total of 2715 mutations were observed for the unique mutation sites identified for each of the six geographical locations. The distribution of the mutations was; Africa (87), Asia (605), Europe (134), North America (1677), Oceania (200) and South America (12). The RdRp protein contained significantly high mutation percentage (>31%) that varied among the different geographical locations. The nsp2 proteins from Asia, North America, Oceania and South America, the nsp3 proteins from Africa and Europe and the helicase proteins from North America showed high mutation percentage next to the RdRp proteins. The P4715L mutation in RdRp, T265I in nsp2 and L3606F in nsp6 were observed in all the geographical locations with the RdRp P4715L mutation being predominant among the orf1ab poly-proteins. In another dataset comprising 158 genomes in which the orf1ab poly-proteins comprised sequences of variable length between 7084-7095 amino acid residues, 88 additional distinct mutations were observed for the six geographical locations that included deletion mutations. The proteins containing deletion mutations were; leader protein, nsp2, nsp3, nsp4, nsp6, RdRp, 3’ -to-5’ exonuclease and endoRNAse.

In this work, all the mutations observed in 11,087 orf1ab poly-proteins of human SARS CoV-2 comprising between 7084-7096 amino acid residues with reference to the human SARS-CoV-2 orf1ab poly-protein sequences from Wuhan-Hu-1, China and representing the six geographical locations; Africa, Asia, Europe, North America, Oceania and South America 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:

The novel severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) responsible for the current COVID-19 pandemic disease has resulted in 29,206,638 coronavirus cases and 928,830 deaths worldwide as on 14th September 2020 (https://www.worldometers.info/coronavirus/). The SARS CoV-2 is a positive sense single stranded RNA genome and the first complete genome sequence for the novel coronavirus was reported from infected individual during December 2019 in Wuhan-Hu-1, China (Wu et al., 2020). The complete genome contains 29,903 base pairs single stranded RNA (NCBI Accession code: NC_045512.2). The poly-proteins encoded by the orf1ab gene in the genome comprises between 7084-7096 amino acid residues which represents fifteen proteins comprising the leader protein, (non-structural proteins) nsp2, nsp3, nsp4, nsp6, nsp7, nsp8, nsp9, nsp10 and the 3C-like proteinase, RNA dependent RNA polymerase (RdRp), helicase, 3' -to-5' exonuclease, endoRNAse and 2' -O-ribose methyltransferase.

In earlier report (Guruprasad, 2020a), we analysed mutations present in all the protein sequences of 22 human SARS-CoV-2 Indian isolates collected between January 2020 to mid-April 2020 from the novel coronavirus infected individuals from the states of Kerala, Karnataka, Telangana, Gujarat and available in the publicly accessible NCBI database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/). Around the same time, mutations present in 9 samples of the human SARS-CoV-2 from Eastern India was also reported (Maitra et al., 2020). We further mapped the mutations observed among the 22 human SARS-CoV-2 Indian isolates on to the protein three-dimensional structures of the RdRp, helicase, endoRNAse and the spike proteins and specifically analysed the secondary structure 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 (Guruprasad, 2020b). In another study, 220 complete SARS-CoV-2 genome sequences from the GISAID database that were derived from patients infected by the SARS-CoV-2 worldwide between December 2019 to mid-March 2020 were randomly collected and 8 novel recurrent nucleotide mutations in the genome sequences were identified that included a novel RdRp variant (Pachetti et al., 2020). (Begum et al., 2020) analysed 908 SARS-CoV-2 orf1ab poly-proteins from North America and carried out normal mode analysis to understand the effects of certain mutations in RdRp and helicase proteins on the structure, dynamics and function.

In this work, we identify and analyse all the mutations present among the human SARS-CoV-2 orf1ab poly-protein sequences. The sequences represent six geographical locations; Africa, Asia, Europe, North America, Oceania and South America and the mutations were identified by comparing the individual orf1ab protein sequences with corresponding reference protein sequences from Wuhan-Hu-1, China (Wu et al., 2020), epicentre of the current pandemic COVID-19 disease.

Materials and Methods:

The human SARS CoV-2 orf1ab poly-protein sequences were obtained from the NCBI databank (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/). The multiple sequence alignments of the proteins were generated using the NGphylogeny.fr suite (Lemoine et al., 2019) available at (https://ngphylogeny.fr). The mutations in the orf1ab poly-proteins were identified and analysed using computer programs developed for the purpose. The human SARS-CoV-2 orf1ab poly-protein sequence is of variable length ranging between 7084-7096 amino acid residues. Protein sequences of length 7096 amino acid residues were predominant. We therefore, split the data into two datasets, one comprising orf1ab poly-proteins with length 7096 amino acid residues (referred as 7096 dataset) and the other comprising all the remaining orf1ab poly-proteins comprising length between 7084-7095 amino acid residues (referred as 7084-7095 dataset). The analysis of the mutations in the two datasets with reference to the poly-proteins of the Wuhan-Hu-1, China sequence (NCBI code: YP_009724389.1) were carried out separately.

Results and Discussion:

Mutations in human SARS CoV-2 orf1ab poly-protein sequences (7096 dataset)

The 7096 dataset comprised 10,929 orf1ab poly-protein sequences of the human SARS CoV-2 isolates representing the six geographical locations; Africa, Asia, Europe, North America, Oceania and South America. A total of 27,895 mutations were observed that represented 2,095 distinct mutation sites corresponding to the 15 poly-proteins of orf1ab.

The geographical distribution of the total number of orf1ab poly-protein sequences analysed, the number of unique mutation sites identified and the total number of mutations observed for the 15 proteins of the orf1ab poly-protein sequences is shown in **Figure 1.** Accordingly, the total number of orf1ab poly-protein sequences representing the different geographical locations were; Africa (137), Asia (999), Europe (378), North America (8860), Oceania (526) and South America (29). The number of distinct mutation sites observed were; Africa (87), Asia (605), Europe (134), North America (1677), Oceania (200) and South America (12). The total number of mutations observed for the unique mutation sites were; Africa (253), Asia (2336), Europe (656), North America (23,366), Oceania (1244) and South America (40). The large number of sequences available from North America provided the maximum number of mutations.

The distribution of the total number of mutations observed for the orf1ab poly-proteins is shown in **Figure 2**. The RdRp protein is associated with the maximum number of mutations (33.47%), followed by nsp2 protein (20.04%), helicase (15.95%) and nsp3 protein (12.61%). The mutations for remaining proteins ranged between (0.14%) for nsp10 to (2.79%) for nsp6 proteins.

The distribution of the total number of mutations observed for the individual proteins of the orf1ab poly-proteins corresponding to the different geographical locations is shown in **Table 1.** The RdRp protein was associated with the maximum number of mutations among the fifteen poly-proteins of the human SARS-CoV-2 orf1ab gene and distributed in the geographical locations as follows; Africa (47.43%), Asia (37.67%), Europe (48.62%), North America (32.56%), Oceania (31.02%), South America (57.5%). The mutations for nsp2 protein from Asia (18.36%), North America (20.35%), Oceania (22.02%), South America (15%), and the mutations for nsp3 protein from Africa (19.36%) and Europe (17.68%) were next highest

compared to the percentage of mutations in the RdRp protein. The helicase protein from North America also showed significant mutations (18.2%).

The geographical distribution of the total number of mutations observed for the 15 orf1ab poly-proteins is shown in **Figure 3**. The RdRp, nsp2, nsp3 and the helicase are associated with relatively high number of mutations compared to the other proteins and this is distinctly observed for the proteins represented in North America.

The distribution of the total number of mutations observed at the different mutation sites for the human SARS-CoV-2 orf1ab poly-proteins from Africa, Asia, Europe, North America, Oceania, South America are shown in **Figures 4-9**, respectively. The RdRp mutation at position 4715 was significantly high compared to the mutations observed at other positions and this observation was consistent across all the geographical locations. The mutations observed for the human SARS-CoV-2 orf1ab poly-proteins in the different geographical locations arranged according to the decreasing order of total numbers observed were: P4715L, G5215S, T265I, G3278S, T4090I, D1036E, T1246I, E1363D, A3518V, P6563S, E3073A (Africa); P4715L, I300F, L3606F, Q676P, L6102F, T6297I, T2016K, T265I, A4489V, V5272I (Asia); P4715L, T265I, L3606F, I739V, P765S, G392D, A876T, V2061I, N2603D, F3071Y (Europe); P4715L, T265I, Y5865C, P5828L, E5568D, L3606F, S3884L, K1202N, D144A, F190L (North America); P4715L, T265I, L3606F, I300F, Y5865C, P5828L, V378I, D75E, P271S, P971L, F6158L, V6474L (Oceania); and P4715L, T265I, L3606F, S3884L, A6245V (South America).

The list of all the mutations observed in human SARS-CoV-2 orf1ab poly-proteins distributed according to the geographical locations is attached in the **Supplementary Table-1**. The mutations; T265I in nsp2 protein, L3606F in nsp6 and P4715L mutation in RdRp were the only mutations among the human SARS-CoV-2 orf1ab poly-proteins that were common across all the geographical locations.

Mutations in human SARS CoV-2 orf1ab poly-protein sequences (7084-7095 dataset)

The total number of orf1ab poly-protein sequences in the 7084-7095 dataset comprised 158 sequences corresponding to the human SARS CoV-2 isolates from the six geographical

locations; Africa (2), Asia (9), Europe (7), North America (113), Oceania (26) and South America (1).

A total of 270 mutation sites were observed for the 7084-7095 orf1ab poly-proteins dataset. These mutation sites represented in the different geographical locations were; Africa (17), Asia (48), Europe (21), North America (141), Oceania (34) and South America (9). The total number of mutations corresponding to the mutation sites were; Africa (23), Asia (58), Europe (32), North America (587), Oceania (140) and South America (9). By excluding the redundant mutation sites present between the geographical locations for the orf1ab poly-proteins in the 7084-7095 dataset and by further excluding the redundant mutations by comparing with the mutation sites in the 7096 dataset, we identified 88 additional mutation sites for the orf1ab proteins corresponding to the 7084-7095 dataset. The mutations in the orf1ab poly-proteins in the 7084-7095 dataset are listed in the **Supplementary Table 2**. A number of deletion mutants were observed for the human SARS-CoV-2 orf1ab poly-proteins, besides certain additional mutations. The leader protein was associated with deletion mutants represented in all the six geographical locations analysed. The orf1ab poly-proteins associated with the deletion mutations were; leader protein, RdRp (Africa), leader protein, nsp2, nsp3, nsp4, 3' -to-5' exonuclease (Asia), leader protein, nsp2 (Europe), leader protein, nsp2, nsp3, nsp4, nsp6, 3' -to-5' exonuclease, endoRNAse (North America), leader protein, nsp2, nsp3, nsp7 (Oceania) and leader protein (South America). The deletion mutations observed for the orf1ab poly-proteins were the following: (leader protein) V54del, G82-V86, Y136del, K141-F143; (RdRp) Y4738-V4746; (nsp2) G445-L450; (nsp3) V868del, I1023-N1026, G1095del, E1205-F1214, Q1224-K1226, S1539del, N2082-S2083, K2085-I2086, Y2141del; (nsp4) F3153-F3157, F2834-W2837, G2879del, L3198del; (nsp6) D3668-V3670, F3760del; (nsp7) T3904del; (3' -to-5' exonuclease) L6418-N6424; (endoRNAse) L6679del, F6692del, I6721del.

A number of instances were observed where a different mutation type occurred at the same mutation site. For instance, in nsp3 protein, the mutation R848K (North America) and R848S (Africa); in endoRNAse, S6713P (North America) and S6713L (Asia); in 2' -O-ribose methyltransferase, D6906N (North America) and D6906H (Asia); in RdRp, A4977S and A4977T (North America); in helicase, V5545A and V5545F (North America); in 3' -to-5' exonuclease, P6252S and P6252L (North America).

Conclusions:

Mutations were observed in all of the human SARS-CoV-2 orf1ab poly-proteins compared to the reference protein sequences from Wuhan-Hu-1, China. The 11,087 orf1ab poly-protein sequences of human SARS-CoV-2 representing six geographical locations varying in length between 7084-7096 amino acid residues were associated with 2,183 distinct mutation sites. All the mutations observed are documented in the present work. In some instances, more than one mutation type was associated with a mutation site. The mutations in RdRp, nsp2, helicase and nsp3 proteins were significantly high compared to mutations in the other human SARS-CoV-2 orf1ab poly-proteins. The P4715L mutation in RdRp protein, T265I mutation in nsp2 protein and the L3606F mutation in nsp6 protein were observed in the human SARS-CoV-2 isolates from all six geographical locations; Africa, Asia, Europe, Oceania, North America and South America. Deletion mutations were observed for the leader protein, RdRp, nsp2, nsp3, nsp4, nsp6, nsp7, 3' -to-5' exonuclease and the endoRNAse proteins with mutations in the leader protein observed in orf1ab poly-proteins from all the six geographical locations. The human SARS-CoV-2 has undergone mutations in ~30% of the amino acids in orf1ab poly-proteins within a short span of 9 months, since outbreak of the COVID-19 pandemic in December 2019, with its epicentre in the city of Wuhan, Hubei-1 province, China.

References:

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*.

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, Kunchur (2020b): Mapping Mutations in Proteins of SARS CoV-2 Indian Isolates on to the Three-Dimensional Structures. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.12683771.v1

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.

Acknowledgements:

The author sincerely acknowledges several researchers and genome sequencing centres for making available the complete genomes in the NCBI repository for public access.

Conflict of interest:

The author declares no conflict of interest

Funding:

None

TABLE 1. Geographical distribution of the total number of mutations in human SARS-CoV-2 orf1ab poly-proteins

Human SARS-CoV-2 orf1ab poly-proteins	AFRICA	ASIA	EUROPE	NORTH AMERICA	OCEANIA	SOUTH AMERICA
leader protein	1	23	1	473	33	0
nsp2	18	429	110	4755	274	6
nsp3	49	368	116	2814	171	0
nsp4	9	65	16	328	48	2
3C-like proteinase	12	48	15	519	28	1
nsp6	2	128	33	516	97	3
nsp7	3	5	1	262	3	3
nsp8	9	34	2	215	2	0
nsp9	0	9	6	68	4	0
nsp10	1	10	2	23	4	0
RdRp	120	880	319	7610	386	23
helicase	10	70	7	4254	110	0
3’ -to-5’ exonuclease	7	172	5	809	40	2
endoRNAse	8	64	18	464	38	0
2’ -O-ribose methyltransferase	4	31	5	256	6	0
TOTAL	253	2336	656	23366	1244	40

FIGURE 1.

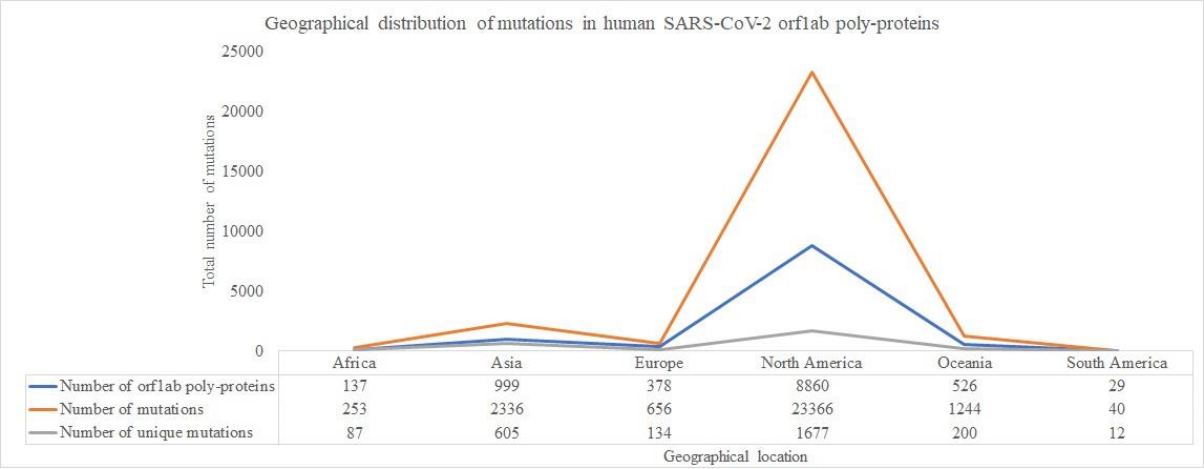

Geographical distribution of mutations in human SARS-CoV-2 orf1ab poly-proteins
Total number of mutations
25000
20000
15000
10000
5000
0
Africa
Asia
Europe
North America
Oceania
South America
Number of orf1ab poly-proteins 137 999 378 8860 526 29
Number of mutations 253 2336 656 23366 1244 40
Number of unique mutations 87 605 134 1677 200 12
Geographical location

FIGURE 2.

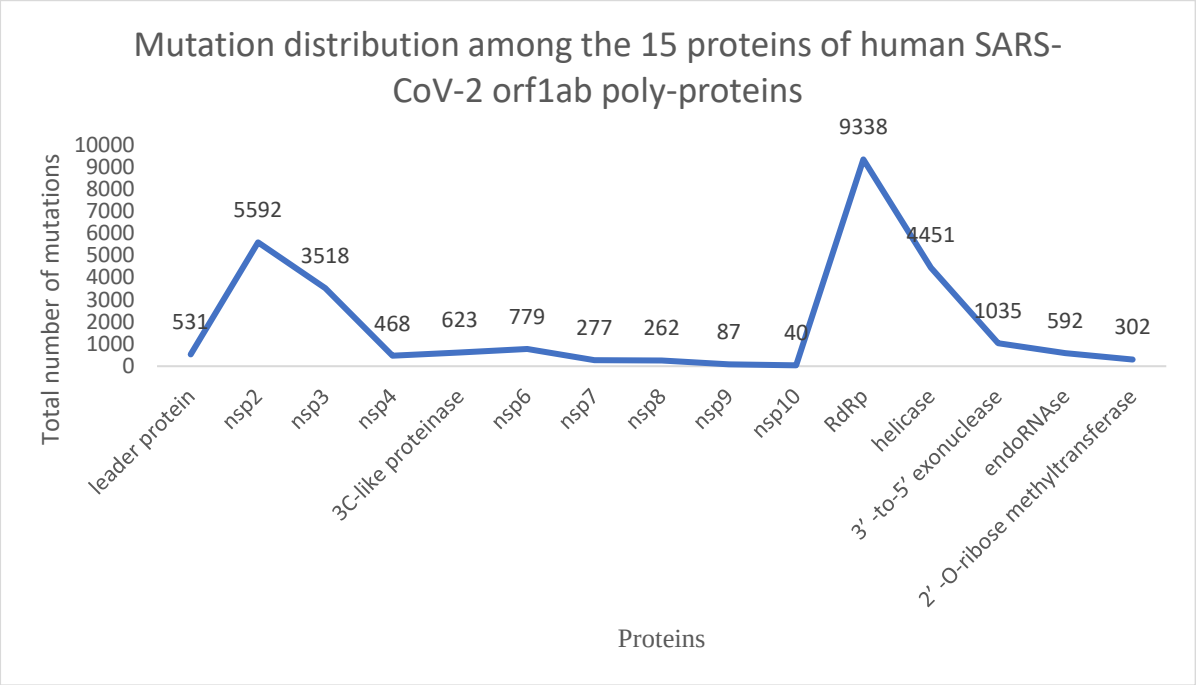

Mutation distribution among the 15 proteins of human SARS-CoV-2 orf1ab poly-proteins
Total number of mutations
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
531
5592
3518
468
623
779
277
262
87
40
9338
4451
1035
592
302
leader protein
nsp2
nsp3
nsp4
3C-like proteinase
nsp6
nsp7
nsp8
nsp9
nsp10
RdRp
helicase
3' -to-5' exonuclease
endoRNAse
2' -O-ribose methyltransferase
Proteins

FIGURE 3.

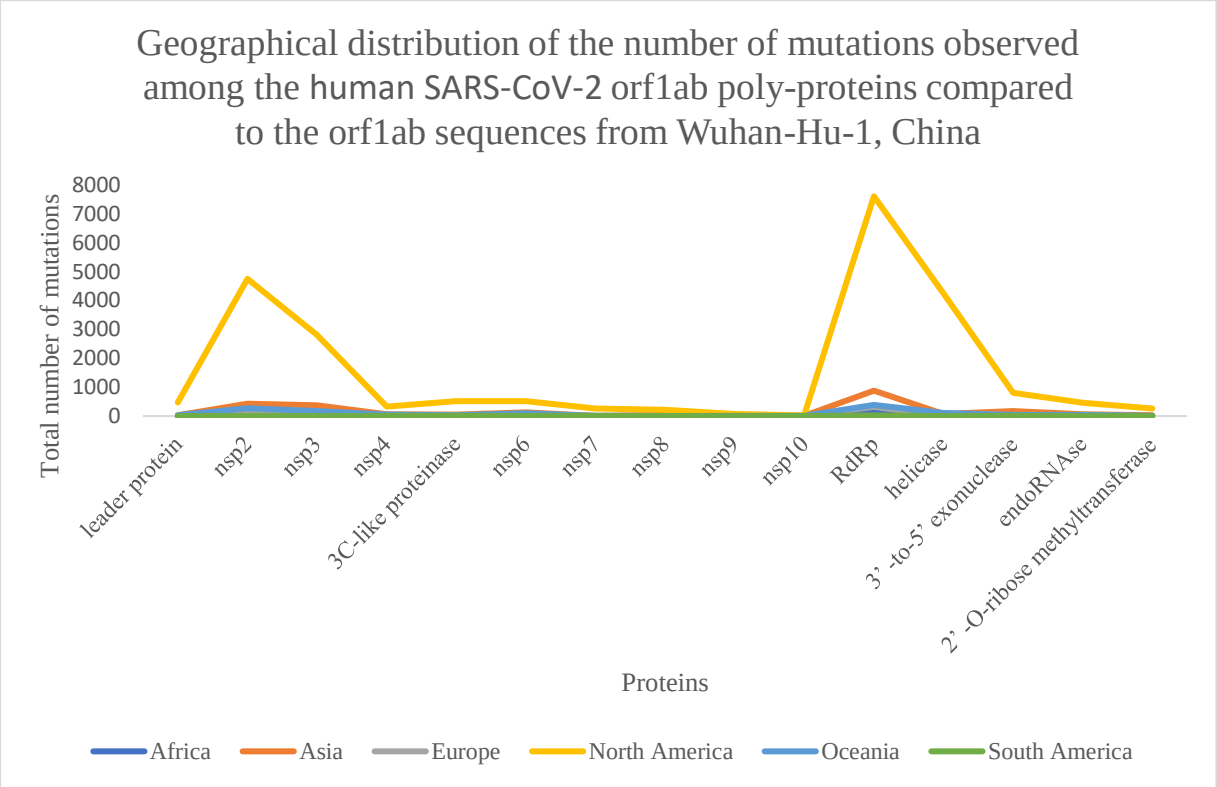

Geographical distribution of the number of mutations observed among the human SARS-CoV-2 orf1ab poly-proteins compared to the orf1ab sequences from Wuhan-Hu-1, China
Total number of mutations
8000
7000
6000
5000
4000
3000
2000
1000
0
leader protein
nsp2
nsp3
nsp4
3C-like proteinase
nsp6
nsp7
nsp8
nsp9
nsp10
RdRp
helicase
3' -to-5' exonuclease
endoRNAse
2' -O-ribose methyltransferase
Proteins
Africa
Asia
Europe
North America
Oceania
South America

FIGURE 4.

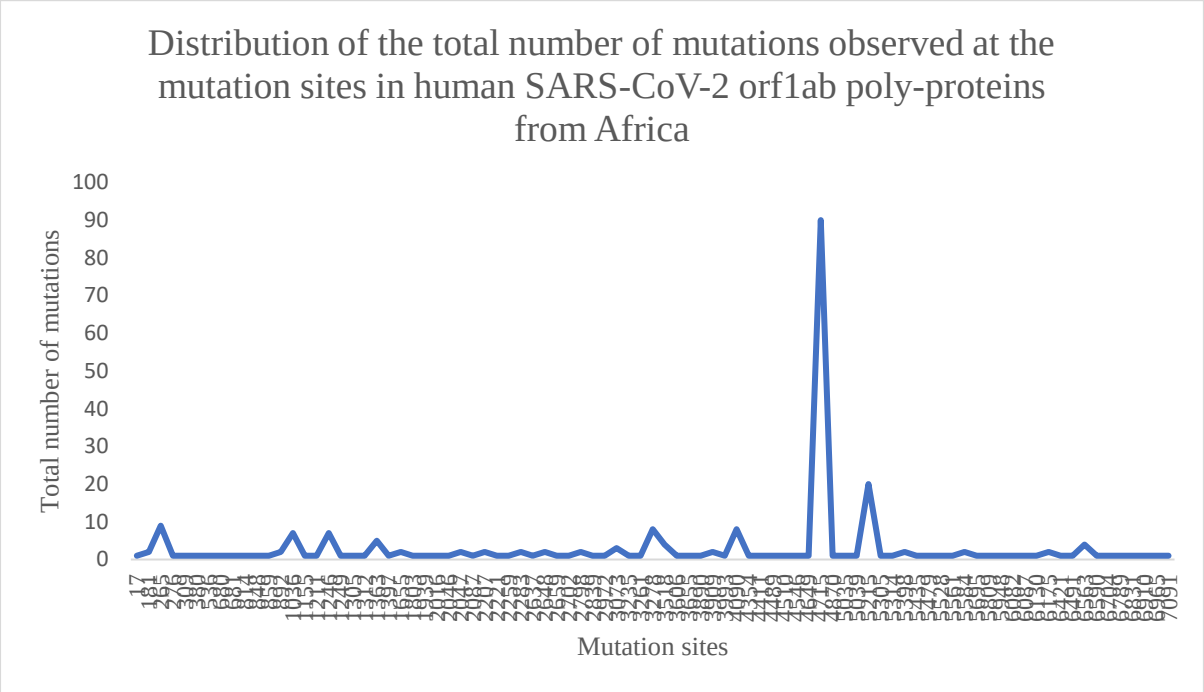

Distribution of the total number of mutations observed at the mutation sites in human SARS-CoV-2 orf1ab poly-proteins from Africa
Total number of mutations
100
90
80
70
60
50
40
30
20
10
0
Mutation sites

FIGURE 5.

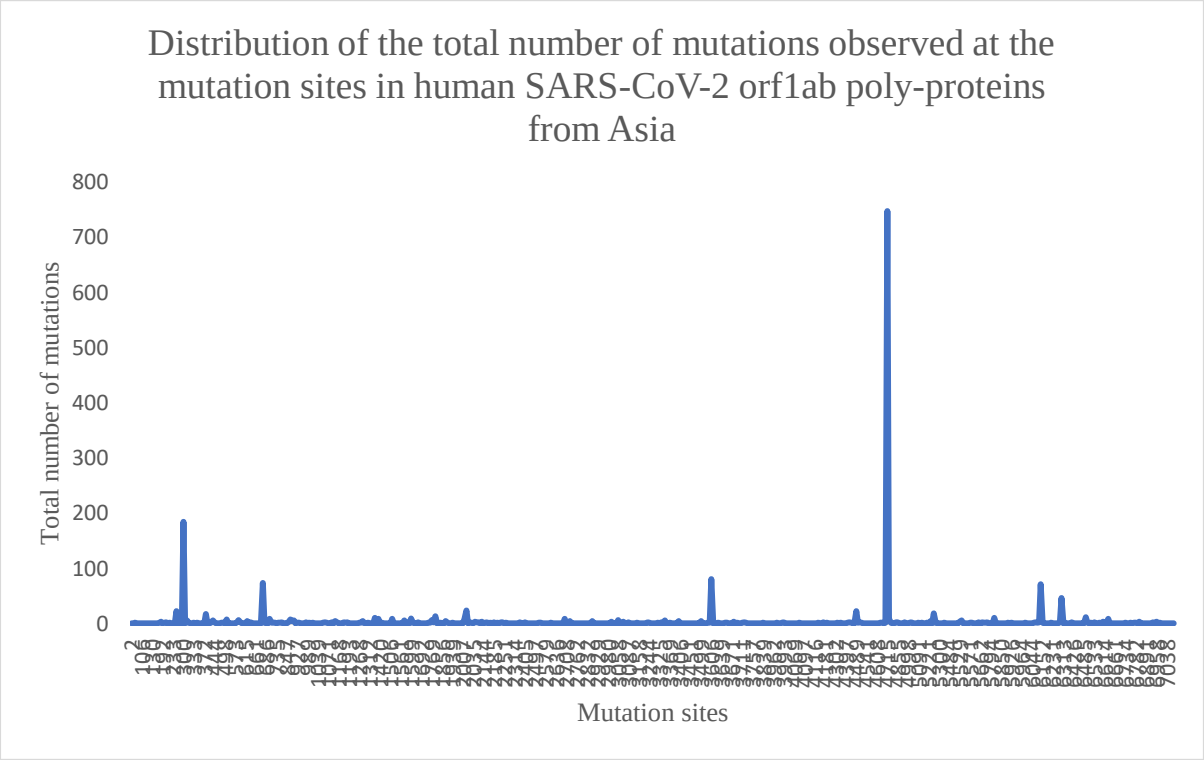

Distribution of the total number of mutations observed at the mutation sites in human SARS-CoV-2 orf1ab poly-proteins from Asia
Total number of mutations
800
700
600
500
400
300
200
100
0
Mutation sites

FIGURE 6.

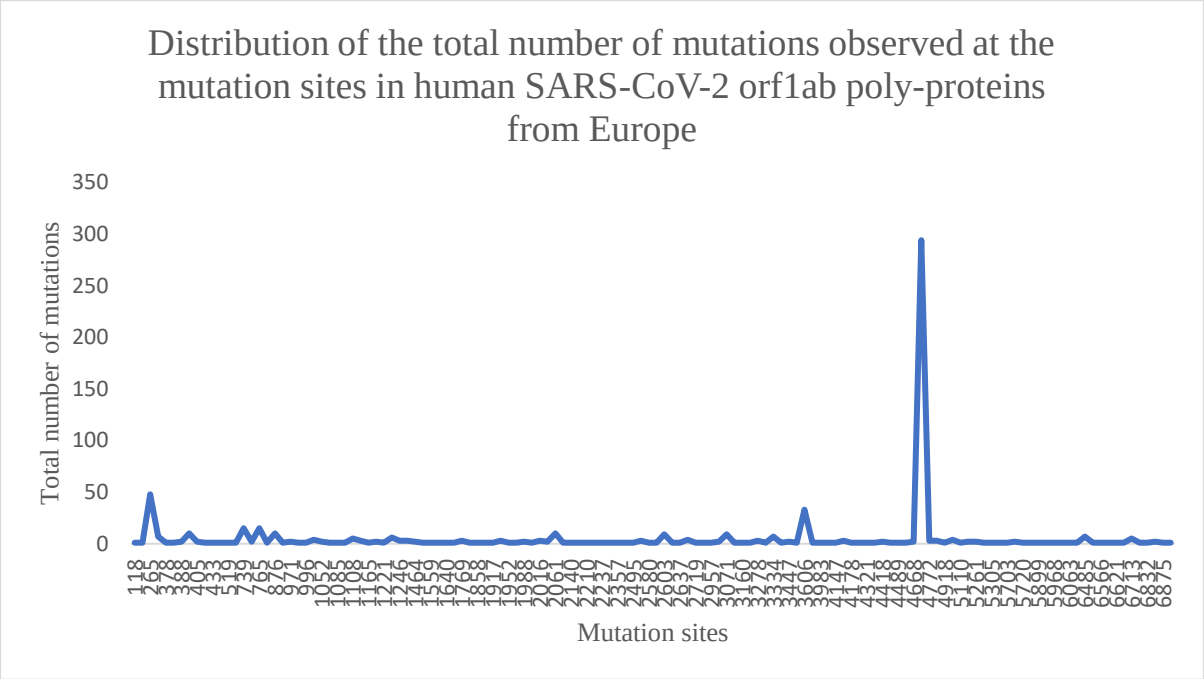

Distribution of the total number of mutations observed at the mutation sites in human SARS-CoV-2 orf1ab poly-proteins from Europe
Total number of mutations
350
300
250
200
150
100
50
0
118
265
378
388
405
433
519
739
765
876
971
996
1052
1085
1108
1165
1221
1246
1464
1559
1640
1769
1889
1917
1952
1986
2016
2061
2140
2210
2237
2357
2495
2580
2603
2637
2719
2957
3071
3160
3208
3224
3334
3347
3606
3983
4147
4178
4321
4418
4488
4698
4772
4918
5110
5161
5305
5703
5720
5869
5968
6063
6485
6566
6621
6713
6832
6875
Mutation sites

FIGURE 7.

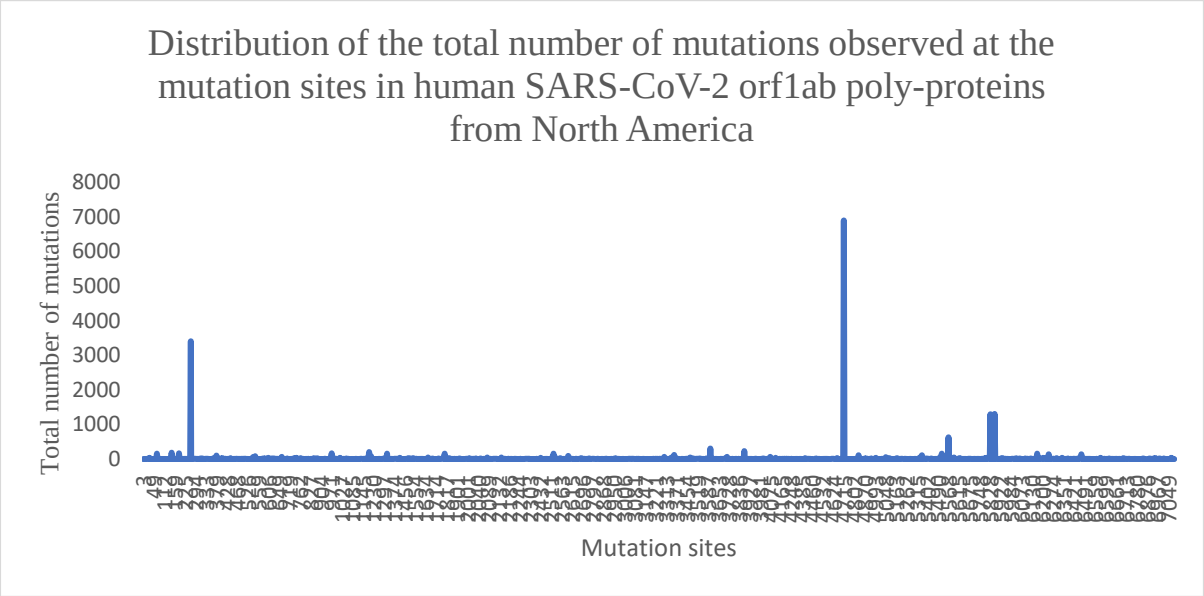

Distribution of the total number of mutations observed at the mutation sites in human SARS-CoV-2 orf1ab poly-proteins from North America
Total number of mutations
8000
7000
6000
5000
4000
3000
2000
1000
0
Mutation sites

FIGURE 8.

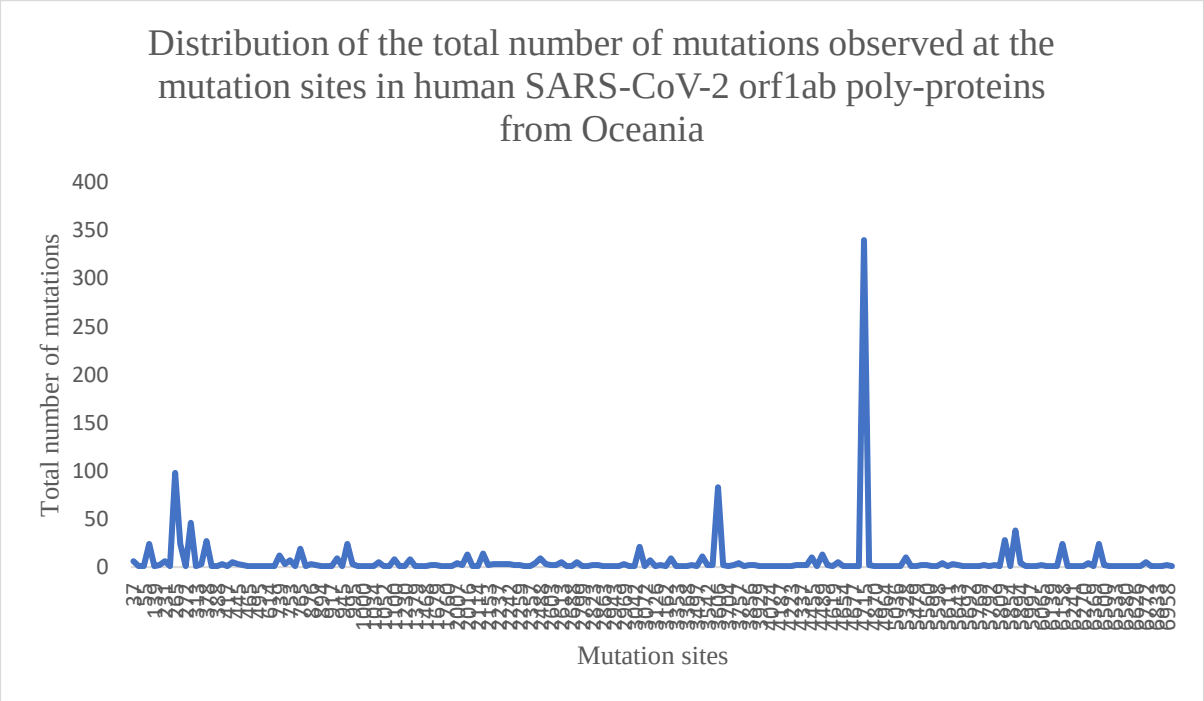
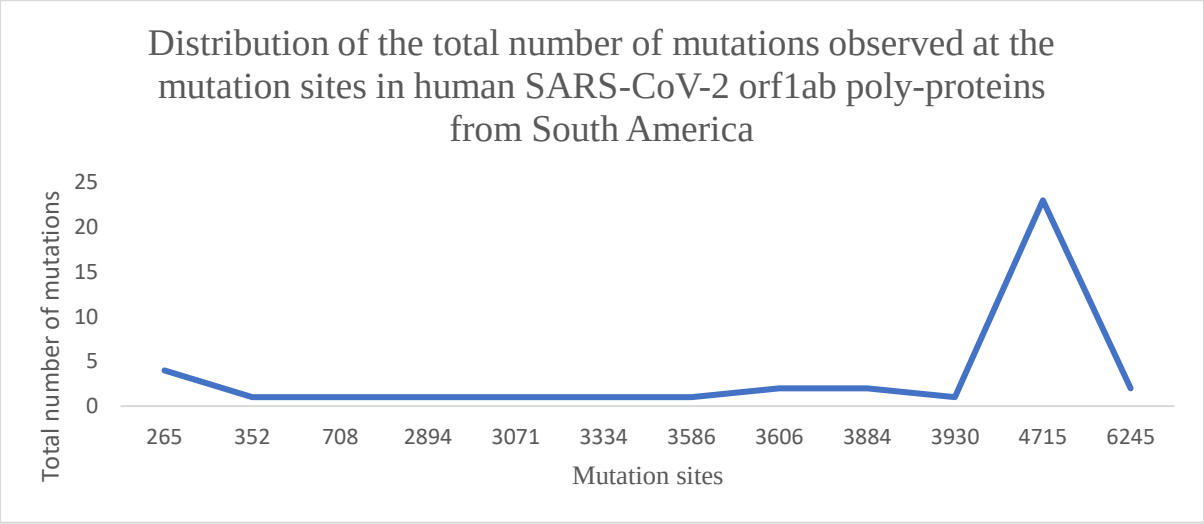

Distribution of the total number of mutations observed at the mutation sites in human SARS-CoV-2 orf1ab poly-proteins from Oceania
Total number of mutations
400
350
300
250
200
150
100
50
0
Mutation sites

FIGURE 9.


Distribution of the total number of mutations observed at the mutation sites in human SARS-CoV-2 orf1ab poly-proteins from South America
Total number of mutations
25
20
15
10
5
0
265
352
708
2894
3071
3334
3586
3606
3884
3930
4715
6245
Mutation sites