Non-Uniform Aspects of SARS-CoV-2 Intraspecies Evolution Reopen Questions on Its Origin

Sk. Sarif Hassan[a,*], Vaishnavi Kodakandla[b], Elrashdy M. Redwan[c], Kenneth Lundstrom[d], Pabitra Pal Choudhury[e], Ángel Serrano-Aroca[f], Gajendra Kumar Azad[g], Alaa A. A. Aljabali[h], Giorgio Palu[i], Tarek Mohamed Abd El-Aziz[j], Debmalya Barh[k], Bruce D. Uhal[l], Parise Adadi[m], Kazuo Takayama[n], Nicolas G Bazan[o], Murtaza Tambuwala[p], Samendra P Sherchan[q], Amos Lal[r], Gaurav Chauhan[s], Wagner Baetas-da-Cruz[t], Vladimir N. Uversky[u,*]

[a] *Department of Mathematics, Pingla Thana Mahavidyalaya, Maligram, Paschim Medinipur, 721140, West Bengal, India*
[b] *Department of Life sciences, Sophia College For Women, University of Mumbai, Bhulabhai Desai Road, Mumbai 400026, India*
[c] *Biological Science Department, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia, Therapeutic and Protective Proteins Laboratory, Protein Research Department, Genetic Engineering and Biotechnology Research Institute, City of Scientific Research and Technological Applications, New Borg EL-Arab, 21934, Alexandria, Egypt*
[d] *PanTherapeutics, Rte de Lavaux 49, CH1095 Lutry, Switzerland*
[e] *Indian Statistical Institute, Applied Statistics Unit, 203 B T Road, Kolkata 700108, India*
[f] *Biomaterials and Bioengineering Lab, Centro de Investigación Traslacional San Alberto Magno, Universidad Católica de Valencia San Vicente Mártir, c/Guillem de Castro, 94, 46001 Valencia, Valencia, Spain*
[g] *Department of Zoology, Patna University, Patna, Bihar, India*
[h] *Department of Pharmaceutics and Pharmaceutical Technology, Yarmouk University, Faculty of Pharmacy, Irbid 566, Jordon*
[i] *Department of Molecular Medicine, University of Padova, Via Gabelli 63, 35121, Padova, Italy*
[j] *Department of Cellular and Integrative Physiology, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Dr, San Antonio, TX 78229-3900, USA, & Zoology Department, Faculty of Science, Minia University, El-Minia 61519, Egypt*
[k] *Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, WB, India, & Departamento de Genética, Ecologia e Evolucao, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil*
[l] *Department of Physiology, Michigan State University, East Lansing, MI 48824, USA*
[m] *Department of Food Science, University of Otago, Dunedin 9054, New Zealand*
[n] *Center for iPS Cell Research and Application, Kyoto University, Kyoto 6068507, Japan*
[o] *Neuroscience Center of Excellence, School of Medicine, LSU Health New Orleans, New Orleans, LA 70112, USA*
[p] *School of Pharmacy and Pharmaceutical Science, Ulster University, Coleraine BT52 1SA, Northern Ireland, UK*
[q] *Department of Environmental Health Sciences, Tulane University, New Orleans, LA, 70112, USA*
[r] *Department of Medicine, Division of Pulmonary and Critical Care Medicine, Mayo Clinic, Rochester, Minnesota, USA*
[s] *School of Engineering and Sciences, Tecnologico de Monterrey, Av. Eugenio Garza Sada 2501 Sur, 64849 Monterrey, Nuevo León, Mexico*
[t] *Translational Laboratory in Molecular Physiology, Centre for Experimental Surgery, College of Medicine, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil*
[u] *Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA*

Abstract

Several hypotheses have been presented on the origin of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) from its identification as the agent causing the current coronavirus disease 19 (COVID-19) pandemic. So far, no hypothesis has managed to identify the origin, and the issue has resurfaced. Here we have unfolded a pattern of distribution of several mutations in the SARS-CoV-2 proteins across different continents comprising 24 geo-locations. The results showed an evenly uneven distribution of unique protein variants, distinct mutations, unique frequency of common conserved residues, and mutational residues across the 24 geo-locations. Furthermore, ample mutations were identified in the evolutionarily conserved invariant regions in the SARS-CoV-2 proteins across almost all geo-locations we have considered. This pattern of mutations potentially breaches the law of evolutionary conserved functional units of the beta-coronavirus genus. These mutations may lead to several novel SARS-CoV-2 variants with a high degree of transmissibility and virulence. A thorough investigation on the origin and characteristics of SARS-CoV-2 needs to be conducted in the interest of science and to be prepared to meet the challenges of potential future pandemics.

Keywords: SARS-CoV-2, Mutations, Furin Cleavage Site (FCS), Evenly-uneven, Invariant regions.

1. Introduction

SARS-CoV-2 is the etiological agent causing the COVID-19 pandemic. Since its very onset, the understanding of the origin of the SARS-CoV-2 has been of utmost importance in the fight against this virus, and the potential emergence of

*Corresponding author
Email addresses: sksarifhassan@pinglacollege.ac.in (Sk. Sarif Hassan), vaishnavikodakandla13@gmail.com (Vaishnavi Kodakandla), lradwan@kau.edu.sa (Elrashdy M. Redwan), lundstromkenneth@gmail.com (Kenneth Lundstrom), pabitrapalchoudhury@gmail.com (Pabitra Pal Choudhury), angel.serrano@ucv.es (Ángel Serrano-Aroca), gkazad@patnauniversity.ac.in (Gajendra Kumar Azad), alaaj@yu.edu.jo (Alaa A. A. Aljabali), giorgio.palu@unipd.it (Giorgio Palu), mohamedt1@uthscsa.edu (Tarek Mohamed Abd El-Aziz), dr.barh@gmail.com (Debmalya Barh), bduhal@gmail.com (Bruce D. Uhal), pariseadadi@gmail.com (Parise Adadi), kazuo.takayama@cira.kyoto-u.ac.jp (Kazuo Takayama), nbazan@lsuhsc.edu (Nicolas G Bazan), m.tambuwala@ulster.ac.uk (Murtaza Tambuwala), sshercha@tulane.edu (Samendra P Sherchan), manavamos@gmail.com (Amos Lal), gchauhan@tec.mx (Gaurav Chauhan), wagner.baetas@gmail.com (Wagner Baetas-da-Cruz), vuversky@usf.edu (Vladimir N. Uversky)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0472.v1

new pathogens and exposure risks [1, 2, 3, 4]. A great source of unfolding the COVID-19 pandemic is the access to SARS-CoV-2 hub at the National Center for Biotechnology Information (NCBI) [5]. In this context, a careful time-based dynamic surveillance of mutations and associated functional changes are most productive due to the potential link to changes in properties such as immune-escape, pathogenesis, and virulence, among others [6]. The surveillance should focus on the analysis of the viral genome and identification of mutations [7, 8, 9]. At the beginning of the pandemic, the largely accepted consensus was that the SARS-CoV-2 would not mutate too fast because, unlike most RNA viruses, it has a "proofreading" protein (ExoN-nsp14), whose function is to prevent too many changes to the viral genome [10, 11]. On the contrary, SARS-CoV-2 sequences from COVID-19 patients showed that the receptor-binding domain (RBD) of the spike protein possessed eight mutations, which assist in initiating infection of the host cells [12, 13, 14]. Many of the mutations in SARS-CoV-2 are inessential, and some are disadvantageous to the virus itself. Some mutations may allow it to propagate more easily from host to host, and these mutations make SARS-CoV-2 variants more transmissible [15]. The majority of mutations in SARS-CoV-2 do not appear to make people sicker, but just make the virus more contagious [16]. The mutation rate is defined as the probability that a change in genetic information is passed to the next generation [17, 18]. For viruses, a generation is simply defined as a cell infection cycle, which includes initiating attachment to the cell surface, entry, replication, encapsidation, and release of infectious particles [19]. It was previously reported that an inverse correlation exists between mutation rates and genome size in RNA-viruses [20]. Coronaviruses have the largest genomes among RNA viruses (30–33 kb) and have acquired proofreading capacity in contrast to all other known RNA viruses [21, 22]. Though most mutations in the SARS-CoV-2 are expected to be either deleterious and swiftly purged or relatively neutral, a small proportion will affect functional properties and increase/decrease infectivity and disease severity or interact with host immunity [23, 24]. In SARS-CoV-2, the average mutation rate remains low and steady, being much lower than for other RNA viruses such as influenza viruses, HIV, and HCD [25].

Such atypical characteristics have contributed to the resurfacing of the question of the origin of the SARS-CoV-2. So far, no clear animal progenitor or intermediary host has been confirmed. Therefore, the hypothesis that SARS-CoV-2 originated as a leak from the Wuhan lab or its unusual origin is now being taken seriously. Primarily, a zoonotic source was thought to have spilled over to humans through the 'wet market' in Wuhan, China, where the virus was first detected in December 2019 [26, 27, 28, 29, 30]. But later, several other orthogonal hypotheses reverted to the old question about SARS-CoV-2 origin [31, 32, 33, 34, 35]. In this study, the apparent uneven distribution of the identified mutations in several proteins of SARS-CoV-2 across the 24 geo-locations questions the natural origin of the SARS-CoV-2, based on prior knowledge from other beta-coronaviruses. Several other observations, such as mutations in invariant regions of the SARS-CoV-2 proteins, which are conserved across four other beta-coronaviruses, strengthen the case of the pseudo-natural origin of SARS-CoV-2.

2. Data acquisition and Methods

2.1. Data and Informatics

The SARS-CoV-2 spike (S), envelope (E), membrane (M), nucleocapsid (N), ORF3a, ORF6, ORF7a, ORF7b, ORF8, and ORF10 sequences (complete) from 24 geo-locations were exported in Fasta format (as of May 29, 2021) from the NCBI database (*http : //www.ncbi.nlm.nih.gov/*). The 24 geo-locations were chosen from all six continents having a relatively higher frequency of SARS-CoV-2 proteins. The Asian group comprises proteins obtained from patients in India, Hong Kong, Bahrain, Bangladesh, and Pakistan. The Oceanian group comprises Australian patients only, whereas the European one includes patients from Austria, France, Greece, Poland, Serbia, and Spain. The South American group contains proteins from Peru and Chile. The African group contains patients from Egypt, Ghana, and Tunisia. Finally, the North American group contains proteins obtained from patients in California, Florida, Texas, Massachusetts, Minnesota, Michigan, and Pennsylvania. Furthermore, Fasta files were processed in *Matlab-2021a* for extracting unique protein sequences from each geo-location. The frequency of total and unique protein sequences is presented in Tables 1 and 2.

Table 1: Frequencies and percentages of S, E, M, N, ORF3a, and ORF6 protein sequences of SARS-CoV-2 from 24 different geo-locations

Geo-locations	Total # of S	Unique # of S	% of Unique	Total # of E	Unique # of E	% of Unique E	Total # of M	Unique # of M	% of Unique (M)
Australia	9919	1121	11.302	9919	19	0.192	9919	38	0.3831
Austria	97	26	26.804	97	1	1.031	97	2	2.0619
Bahrain	167	56	33.533	167	2	1.198	167	4	2.3952
Bangladesh	402	98	24.378	402	6	1.493	402	11	2.7363
California	15616	3321	21.267	15616	61	0.391	15744	192	1.2195
Chile	290	25	8.621	290	1	0.345	290	2	0.6897
Egypt	700	183	26.143	700	16	2.286	700	22	3.1429
Florida	17180	2527	14.709	17180	49	0.285	17324	131	0.7562
France	90	19	21.111	90	2	2.222	90	4	4.4444
Ghana	167	65	38.922	167	6	3.593	167	7	4.1916
Greece	97	11	11.340	97	2	2.062	97	3	3.0928
Hong Kong	228	48	21.053	228	2	0.877	230	5	2.1739
India	813	178	21.894	813	11	1.353	830	20	2.4096
Massachusetts	8856	1281	14.465	8856	37	0.418	9045	92	1.0171
Michigan	9930	1297	13.061	9930	35	0.352	9998	78	0.7802
Minnesota	13046	2658	20.374	13046	36	0.276	13621	77	0.5653
Pakistan	214	49	22.897	214	4	1.869	214	7	3.2710
Pennsylvania	8779	1343	15.298	8779	33	0.376	8913	105	1.1781
Peru	116	44	37.931	116	3	2.586	116	8	6.8966
Poland	153	26	16.993	153	3	1.961	153	2	1.3072
Serbia	146	23	15.753	146	1	0.685	146	3	2.0548
Spain	134	36	26.866	134	2	1.493	134	4	2.9851
Texas	9251	1546	16.712	9251	33	0.357	9431	101	1.0709
Tunisia	58	30	51.724	58	2	3.448	58	3	5.1724
Geo-locations	**Total # of N**	**Unique # of N**	**% of Unique (N)**	**Total # of ORF3a**	**Unique # of ORF3a**	**% of Unique (ORF3a)**	**Total # of ORF6**	**Unique # of ORF6**	**% of Unique (ORF6)**
Australia	9919	213	2.147	9919	132	1.331	9919	19	0.192
Austria	97	22	22.680	97	14	14.433	97	3	3.093
Bahrain	167	33	19.760	167	27	16.168	167	7	4.192
Bangladesh	402	53	13.184	402	59	14.677	402	9	2.239
California	15616	1345	8.613	15608	1073	6.875	15615	104	0.666
Chile	290	16	5.517	290	16	5.517	290	3	1.034
Egypt	700	116	16.571	700	81	11.571	700	10	1.429
Florida	17180	973	5.664	17173	808	4.705	17178	65	0.378
France	90	6	6.667	90	10	11.111	90	3	3.333
Ghana	167	41	24.551	167	23	13.772	167	10	5.988
Greece	97	9	9.278	97	13	13.402	97	2	2.062
Hong Kong	228	28	12.281	228	17	7.456	228	3	1.316
India	813	86	10.578	808	73	9.035	813	7	0.861
Massachusetts	8856	625	7.057	8856	468	5.285	8856	47	0.531
Michigan	9930	418	4.209	9930	389	3.917	9930	38	0.383
Minnesota	13046	481	3.687	13020	456	3.502	13044	45	0.345
Pakistan	214	33	15.421	212	32	15.094	214	5	2.336
Pennsylvania	8779	643	7.324	8779	561	6.390	8779	52	0.592
Peru	116	19	16.379	116	16	13.793	116	2	1.724
Poland	153	22	14.379	153	21	13.725	153	1	0.654
Serbia	146	22	15.068	146	17	11.644	145	1	0.690
Spain	134	21	15.672	134	13	9.701	134	3	2.239
Texas	9251	644	6.961	9251	532	5.751	9251	61	0.659
Tunisia	58	22	37.931	57	10	17.544	57	1	1.754

doi:10.20944/preprints202106.0472.v1

Table 2: Frequencies and percentages of ORF7a, ORF7b, ORF8, and ORF10 protein sequences of SARS-CoV-2 from 24 different geo-locations

Geo-locations	**Total # of ORF7a**	**Unique # of ORF7a**	**% of Unique (ORF7a)**	**Total # of ORF7b**	**Unique # of ORF7b**	**% of Unique (ORF7b)**
Australia	9919	58	0.585	9919	14	0.141
Austria	97	5	5.155	95	2	2.105
Bahrain	167	18	10.778	167	4	2.395
Bangladesh	402	15	3.731	400	6	1.500
California	15612	330	2.114	15724	89	0.566
Chile	290	5	1.724	290	2	0.690
Egypt	700	20	2.857	700	11	1.571
Florida	17161	314	1.830	17305	63	0.364
France	90	1	1.111	90	1	1.111
Ghana	167	10	5.988	167	7	4.192
Greece	96	2	2.083	97	1	1.031
Hong Kong	230	5	2.174	230	2	0.870
India	828	23	2.778	828	7	0.845
Massachusetts	8853	184	2.078	9044	46	0.509
Michigan	9927	199	2.005	9998	45	0.450
Minnesota	13029	758	5.818	13600	59	0.434
Pakistan	212	6	2.830	206	2	0.971
Pennsylvania	8779	202	2.301	8913	38	0.426
Peru	116	9	7.759	116	1	0.862
Poland	152	8	5.263	153	2	1.307
Serbia	146	3	2.055	146	1	0.685
Spain	134	2	1.493	130	2	1.538
Texas	9251	190	2.054	9430	43	0.456
Tunisia	58	7	12.069	58	2	3.448
Geo-locations	**Total # of ORF8**	**Unique # of ORF8**	**% of Unique (ORF8)**	**Total # of ORF10**	**Unique # of ORF10**	**% of Unique (ORF10)**
Australia	9919	54	0.544	9919	16	0.161
Austria	26	3	11.538	97	2	2.062
Bahrain	145	17	11.724	167	3	1.796
Bangladesh	397	19	4.786	402	11	2.736
California	12945	359	2.773	15739	61	0.388
Chile	290	5	1.724	290	1	0.345
Egypt	697	34	4.878	700	8	1.143
Florida	7948	231	2.906	17322	47	0.271
France	90	3	3.333	90	1	1.111
Ghana	69	12	17.391	167	3	1.796
Greece	97	4	4.124	97	1	1.031
Hong Kong	212	10	4.717	230	3	1.304
India	798	27	3.383	830	3	0.361
Massachusetts	5264	137	2.603	9044	29	0.321
Michigan	3061	77	2.516	9998	23	0.230
Minnesota	4619	118	2.555	13608	29	0.213
Pakistan	208	10	4.808	212	3	1.415
Pennsylvania	4564	135	2.958	8913	29	0.325
Peru	115	8	6.957	116	5	4.310
Poland	149	6	4.027	153	2	1.307
Serbia	146	6	4.110	146	2	1.370
Spain	62	3	4.839	134	3	2.239
Texas	4626	154	3.329	9430	39	0.414
Tunisia	56	7	12.500	57	4	7.018

doi:10.20944/preprints202106.0472.v1

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0472.v1

The percentages of each SARS-CoV-2 protein across the 24 geo-locations are presented in Figure 1.

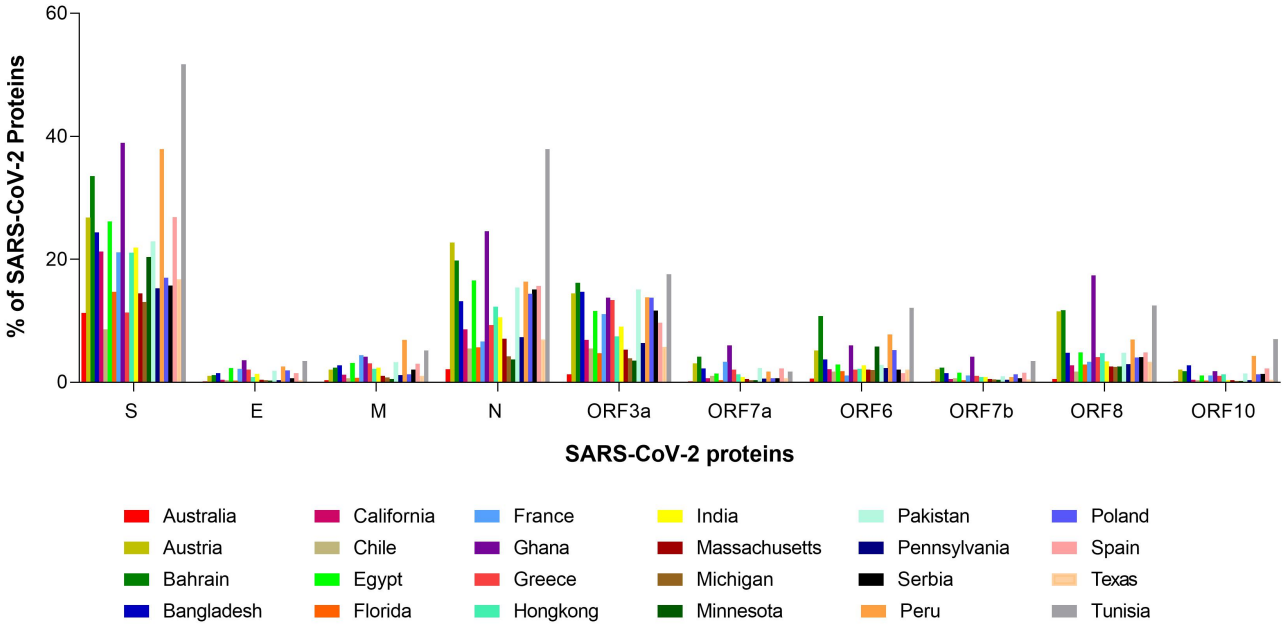

Figure 1: Percentage of each SARS-CoV-2 proteins across 24 geo-locations

Figure 1 indicates that the highest amounts of unique variations across the 24 geo-locations were observed for the S protein. Relatively less unique variations were distributed over the E and ORF3a proteins. Other proteins have a minimal number of unique variations. On the other hand, it was observed that most SARS-CoV-2 proteins possessed the highest unique variations in the viral isolates collected from Tunisia, Ghana, and Greece.

Furthermore, S, E, M, N, ORF3a, ORF6, ORF7a, ORF7b, and ORF8 protein sequences of four other coronaviruses Recombinant SARSr-CoV (taxid-698398), Bat SARS-CoV (taxid-442736), SARS-CoV ExoN1 (taxid-627440), and Bat SARS-like-CoV (taxid-1508227) were downloaded from the NCBI database. In this study, all mutations in SARS-CoV-2 proteins were detected with reference to the SARS-CoV-2 reference sequence, which was deposited in January 2020 by Wu and co-workers formerly called "Wuhan seafood market pneumonia virus" (WSM, NC_045512) [36]. The frequency of total and unique protein sequences is presented in Table 3.

Table 3: Frequencies and percentages of S, E, M, N, ORF3a, ORF6, ORF7a, ORF7b, and ORF8 from four different type of CoVs

Protein	Total	Unique	Percentage	Protein	Total	Unique	Percentage
E-698398	80	6	7.5	**Spike-698398**	36	2	5.56
E-442736	2	1	50	**Spike-627440**	18	2	11.11
E-627440	15	5	33.3	**Spike-442736**	13	7	53.85
E-1508227	2	1	50	**Spike-1508227**	13	13	100
M-698398	116	4	3.45	**ORF3a-442736**	2	1	50
M-442736	2	1	50	**ORF3a-1508227**	11	10	90.91
M-627440	33	3	9.09	**ORF6-1508227**	11	6	54.55
M-1508227	2	1	50	**ORF6-442736**	2	1	50
N-698398	80	4	5	**ORF7a-442736**	2	1	50
N-442736	2	1	50	**ORF7a-1508227**	11	5	45.45
N-627440	15	4	26.67	**ORF7b-1508227**	11	2	18.18
N-1508227	13	12	92.31	**ORF7b-442736**	2	1	50
				ORF8-1508227	10	7	70
				ORF8-442736	2	1	50

The least unique variations of M proteins of four types of beta-coronaviruses were observed. Other proteins of four CoVs had several unique variations, unlike in the case of non-uniformity in unique variations in SARS-CoV-2 proteins.

2.2. Methods

Clustal Omega (*https* : *//www.ebi.ac.uk/Tools/msa/clustalo/*) and MUSCLE (*https* : *//www.ebi.ac.uk/Tools/msa/muscle/*) tools were used to conduct multiple sequence alignment and for mutation detection with reference to the reference sequence

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0472.v1

NC_045512 the web-server ViPR (*https : //www.viprbrc.org/brc/home.spg?decorator = corona*) [37, 38, 39]. At each position of a given protein, the consensus residue is the allele with frequency greater than 50%, regardless of coverage was considered. If no allele exceeds 50%, Xaa (for an amino acid) indicates ambiguity [39]. The effect of mutation was predicted using a webserver, PredictSNP (*https : //loschmidt.chemi.muni.cz/predictsnp1/predictsnp.html*) [40]. The statistical and mathematical computations were performed by Matlab software [].

3. Results

3.1. Unique proteins variants and their mutations

Across the 24 geo-locations, the common amino residues which did not possess any mutations were named as invariant residues. These invariant residues of all unique protein variants from all 24 geo-locations in SARS-CoV-2, were extracted (Table 4) (**Supplementary file-I**). On the other hand, mutated residues common in all 24 geo-locations were also detected (Table 5) (**Supplementary file-I**).

Table 4: Invariant-residues which were common in all unique variants from all 24 geo-locations in SARS-CoV-2 proteins

Invariant residues in SARS-CoV-2 proteins across 24 geo-locations					
S (0.39)	**E (4%)**	**M (9.46%)**		**N (1.91%)**	
1-Met	1-Met	1-Met	180-Lys	1-Met	
953-Asn	2-Tyr	9-Thr	181-Leu	42-Pro	
1051-Ser	3-Ser	65-Phe	190-Asp	49-Thr	
1054-Gln		119-Leu	192-Gly	51-Ser	
1269-Lys		121-Asn	193-Phe	52-Trp	
		156-Leu	195-Ala	57-Thr	
		174-Arg	202-Gly	58-Gln	
		176-Leu	203-Asn	143-Lys	
		177-Ser	218-Ala		
			219-Leu		
			220-Leu		
			222-Gln		
ORF3a (0.73%)	**ORF6 (1.64%)**	**ORF7a (0.83%)**	**ORF7b**	**ORF8 (0.83%)**	**ORF10 (0%)**
1-Met	1-Met	1-Met	1-Met	1-Met	NONE
8-Phe					

From Table 4, it was observed that methionine(M) at the residue position 1 did not change in any of the SARS-CoV-2 proteins listed above, except in ORF10. In ORF10, all amino acid residues from position 1 to 38 were mutated. Even methionine at position 1 was changed to glycine in the only ORF10 sequence QKG88643 from Massachusetts, USA (collected on 18-03-2020). This mutation M1G was found to be a 'neutral' mutation as predicted through the webserver, PredictSNP. Note that there was no homologous sequence to QKG88643 with 100% homology and 100% query coverage (NCBI Blast).

Table 5: Mutation residues that were common in all 24 geo-locations.

Mutation residues in SARS-CoV-2 proteins across 24 geo-locations									
S	**E**	**M**	**N**	**ORF3a**	**ORF6**	**ORF7a**	**ORF7b**	**ORF8**	**ORF10**
D614G/C/N/A			**R203E/K/M/S/T**						
	NONE	NONE	**G204L/P/Q/R/T/V**	**Q57H/E/L/N/R/Y**	NONE	NONE	NONE	NONE	NONE

On the other side, the number of common mutations in the SARS-CoV-2 proteins across 24 geo-locations was surprisingly low (Table 5). D614 was the only mutation possessed by each unique S protein variant from all 24 geo-locations. Similarly, each unique N protein variant from all 24 geo-locations possessed R203 and G204 with changes mutations to multiple amino acids (Table 5). The unique ORF3a variants from all 24 geo-locations had the only common mutation at position 57 with changes to multiple amino acids H/E/L/N/R, and Y. It was noticed that not a single common mutation across 24 geo-locations was found in E, M, ORF6, ORF7a, ORF7b, ORF8, and ORF10.

3.1.1. Spike protein variants and mutations

The total frequency of unique mutations possessed by the S protein of SARS-CoV-2 across the 24 geo-locations is presented in Table 6.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0472.v1

Table 6: Number of unique S protein mutations possessed in each geo-location.

Continent	*Oceania*	*Europe*	*Asia*	*Asia*	*N-America*	*S-America*
Geo-location	**Australia**	**Austria**	**Bahrain**	**Bangladesh**	**California**	**Chile**
# of mutations in S (M_S)	542	98	110	233	1107	63
# of unique seqs. (U_S)	1121	26	56	98	3321	25
Avg. # of mutations per unit unique seqs. ($\frac{M_S}{U_S}$)	0.48	**3.77**	**1.96**	**2.38**	0.33	**2.52**
Continent	*Africa*	*N-America*	*Europe*	*Africa*	*Europe*	*Asia*
Geo-location	**Egypt**	**Florida**	**France**	**Ghana**	**Greece**	**Hong Kong**
# of mutations in S (M_S)	213	995	28	179	11	115
# of unique seqs. (U_S)	183	2527	19	65	11	48
Avg. # of mutations per unit unique seqs. ($\frac{M_S}{U_S}$)	**1.16**	0.39	**1.47**	**2.75**	1.00	**2.40**
Continent	*Asia*	*N-America*	*N-America*	*N-America*	*Asia*	*N-America*
Geo-location	**India**	**Massachusetts**	**Michigan**	**Minnesota**	**Pakistan**	**Pennsylvania**
# of mutations in S (M_S)	219	911	815	970	83	829
# of unique seqs. (U_S)	178	1281	1297	2658	49	1343
Avg. # of mutations per unit unique seqs. ($\frac{M_S}{U_S}$)	**1.23**	0.71	0.63	0.36	**1.69**	0.62
Continent	*S-America*	*Europe*	*Europe*	*Europe*	*N-America*	*Africa*
Geo-location	**Peru**	**Poland**	**Serbia**	**Spain**	**Texas**	**Tunisia**
# of mutations in S (M_S)	218	39	21	88	1122	55
# of unique seqs. (U_S)	44	26	23	36	1546	30
Avg. # of mutations per unit unique seqs. ($\frac{M_S}{U_S}$)	**4.95**	**1.50**	0.91	**2.44**	0.73	**1.83**

We observed that the highest number (495%) of unique mutations possessed by unique S protein variants was from Peru, where 44 unique S sequences had 218 unique mutations. On the other side, the second-highest number of unique S protein variants from California possessed the lowest amount (33%) of unique mutations. Figure 2 shows the average numbers of mutations per unit unique S protein variants.

Figure 2: Geo-location-wise (A): total number of unique mutations and (B): average number of mutation(s) per unique S sequences

From Figure 2(B) shows that the probability of having triple mutants in any randomly chosen unique S protein variant from Austria is nearly 1, since the ratio ($\frac{M_S}{U_S}$) is $3.77 > 3$. Similarly, the probability of having more than quadruple mutants in any

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0472.v1

randomly chosen unique S protein variant from Peru is nearly 1, since the ratio ($\frac{M_S}{U_S}$) is 4.95 > 4. Spectacularly, none of the unique S protein variants from the geo-locations in North America possessed more than one mutation, since the ratio in each case was less than 1, although the total number of unique S variants and mutations were relatively higher than those of others.

The total 23 'variants of concern (VoC)' and 25 'variants of interest (VoI)' mutations in the S protein were reported [41, 42, 43, 44]. Continent-wise, the frequency of common mutations were determined, as well as VoC, VoI among those common S protein mutations possessed by each continental geo-locations (Table 7). It was interesting to note, since Australia was the only geo-location in Oceania considered in this study, common mutations were not observed.

Table 7: Continent-wise common mutations in the S protein and list of Variants of concern (VoC), Variants of Interest(VoI) mutations in S protein.

Continent	**Total # of common mutations in S**	**List of VoC on the continent**	**List of VoI on the continent**
Asia	4	614, 681	5, 142, 614, 681
Europe	1	614	614
Africa	22	80, 452, 484, 614, 681, 701	18, 26, 80, 484, 501, 570 614, 681, 716, 982, 1118
North America	487	13, 18, 20, 26, 80, 138, 152, 190, 215, 417, 452, 484, 501, 570, 614, 655, 655, 681, 701, 716, 982, 1027, 1118, 1191	5, 19, 67, 80, 95, 142, 154, 157, 158, 253, 452, 477, 478, 484, 614, 677, 681, 701, 950, 1071, 1176
South America	45	614	614

It was found that 487 common mutations in the S proteins were from patients from the seven geo-locations in North America, though the only common mutation across 24 geo-locations was D614G. Furthermore, it was noticed that all 23 VoC were presented in each geo-location from North America. On the other hand, the unique S proteins from the European geo-locations possessed only the D614 common mutation. In all African geo-locations, a moderate number of VoC and VoI were found, although the number of common mutations over the geo-locations was not relatively high compared to that of others (Table 7). Also, randomly chosen S protein variants from Ghana has a very high probability of acquiring double VoC/VoI mutants as the ratio ($\frac{M_S}{U_S}$) is 2.75.

Earlier, it was reported that 'RRAR' (amino acid positions: 682-685), a unique furin-like cleavage site (FCS) in the S protein, which was absent in other lineages beta-coronaviruses, such as SARS-CoV, caused high infectivity and transmissibility [45, 46, 47]. Even in this FCS, a single mutation at position 684 was noticed in some unique S protein variants from California, Massachusetts, and Michigan. Details of the protein accessions with associated information are presented in Table 8.

Table 8: Mutations in the unique furin-like cleavage site (FCS) of the S proteins

Accession	**Lineage**	**Length**	**Geo_Location**	**Collection_Date**	**FCS (RRAR)**
QVU70282	B.1.1.7	1270	USA: Massachusetts	06-05-2021	**RRVR**
QVU09331	B.1.1.7	1270	USA: California	16-04-2021	**RRVR**
QVI42615	B.1.1.291	1273	USA: California	24-03-2021	**RRVR**
QVI49490	B.1.427	1273	USA: California	09-02-2021	**RRVR**
QUD47347	B.1.1.7	1270	USA: Michigan	05-04-2021	**RRSR**
QUB14687	B.1.2	1273	USA: Michigan	24-03-2021	**RRSR**
QTU74764	B.1.427	1273	USA: California	09-02-2021	**RRVR**
QTS38722	B.1.429	1271	USA: Michigan	15-03-2021	**RRSR**
QTP22615	B.1.243	1273	USA: Massachusetts	09-09-2020	**RRVR**
QSS81313	B.1.427	1273	USA: California	21-02-2021	**RRVR**
QSL71584	B.1.427	1273	USA: California	10-02-2021	**RRVR**
QSL80009	B.1.2	1273	USA: Michigan	11-02-2021	**RRSR**
QRG20397	B.1.243	1273	USA: CA, Alameda County	12-09-2020	**RRVR**
QQX02259	B.1.561	1273	USA: California	02-01-2021	**RRVR**
QQN04304	B.1.517	1273	USA: Massachusetts	27-11-2020	**RRVR**

The first such mutation, A684V was reported in Massachusetts on September 9, 2020 (Accs. ID: QTP22615). Three days later, the same mutation was identified in California (QRG20397). The mutation A684V/S was 'neutral' (predicted using PredictSNP web-server), and hence it was expected that the ability to infect and transmit remains unchanged [40].

3.1.2. Envelope protein variants and mutations

The total frequency of unique mutations possessed by the E protein of SARS-CoV-2 across the 24 geo-locations is presented in Table 9.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0472.v1

Table 9: Number of unique E protein mutations possessed in each geo-location.

Continent	**Oceania**	**Europe**	**Asia**	**Asia**	**N-America**	**S-America**
Geo-location	**Australia**	**Austria**	**Bahrain**	**Bangladesh**	**California**	**Chile**
# of mutations in E (M_E)	58	0	1	11	61	0
# of unique seqs. E (U_E)	19	1	2	6	61	1
Avg. # of mutations per unit unique seqs. ($\frac{M_E}{U_E}$)	3.05	0.00	0.50	1.83	1.00	0.00
Continent	**Africa**	**N-America**	**Europe**	**Africa**	**Europe**	**Asia**
Geo-location	**Egypt**	**Florida**	**France**	**Ghana**	**Greece**	**Hong Kong**
# of mutations in E (M_E)	12	43	1	5	1	1
# of unique seqs. E (U_E)	16	49	2	6	2	2
Avg. # of mutations per unit unique seqs. ($\frac{M_E}{U_E}$)	0.75	0.88	0.50	0.83	0.50	0.50
Continent	**Asia**	**N-America**	**N-America**	**N-America**	**Asia**	**N-America**
Geo-location	**India**	**Massachusetts**	**Michigan**	**Minnesota**	**Pakistan**	**Pennsylvania**
# of mutations in E (M_E)	32	39	45	54	2	24
# of unique seqs. E (U_E)	11	37	35	36	4	33
Avg. # of mutations per unit unique seqs. ($\frac{M_E}{U_E}$)	2.91	1.05	1.29	1.50	0.50	0.73
Continent	**S-America**	**Europe**	**Europe**	**Europe**	**N-America**	**Africa**
Geo-location	**Peru**	**Poland**	**Serbia**	**Spain**	**Texas**	**Tunisia**
# of mutations in E (M_E)	11	5	0	1	31	1
# of unique seqs. E (U_E)	3	3	1	2	33	2
Avg. # of mutations per unit unique seqs. ($\frac{M_E}{U_E}$)	3.67	1.67	0.00	0.50	0.94	0.50

Almost every unique E protein variant from Australia possessed triple mutations as the ratio $\frac{M_E}{U_E}$ was $3.05 > 3$. Likewise, in India, any E protein contains at least a double mutation ($\frac{M_E}{U_E} = 2.91 > 1$). Compared to this, a much higher number of unique mutations in the unique E proteins from Peru was observed, and any randomly chosen E protein from Puru contains quadruple mutations ($\frac{M_E}{U_E} = 3.67 > 1$). Based on the ratio $\frac{M_E}{U_E} = 0$ that each COVID-19 positive case in Austria, Chile, and Serbia was infected by the SARS-CoV-2 with the wild type E sequence (YP_009724392).

The 12 common mutations at positions 9, 21, 24, 41, 49, 55, 58, 62, 68, 71, 72, and 73 were detected in the unique E protein variants from geo-locations in North America. Among these 12 mutations, 8 mutations (at positions 49, 55, 58, 62, 68, 71, 72, and 73) were shared by the unique E variants from India. Among the 12 mutations, two mutations at positions 21 and 41 were shared with E variants from Bangladesh. No other common mutation was found in geo-locations in Asia, except for the single mutation at position 37 found in India and Bangladesh. E protein variants from the three African geo-locations shared only a single common mutation at position 71.

3.1.3. Membrane protein variants and mutations

The frequency of unique mutations possessed by the M protein of SARS-CoV-2 across the 24 geo-locations is presented in Table 10.

Table 10: Number of unique M protein mutations possessed in each geo-location.

Continent	**Oceania**	**Europe**	**Asia**	**Asia**	**N-America**	**S-America**
Geo-location	**Australia**	**Austria**	**Bahrain**	**Bangladesh**	**California**	**Chile**
# of mutations in M (M_M)	34	1	3	16	139	1
# of unique seqs. M (U_M)	38	2	4	11	192	2
Avg. # of mutations per unit unique seqs. ($\frac{M_M}{U_M}$)	0.89	0.50	0.75	1.45	0.72	0.50
Continent	**Africa**	**N-America**	**Europe**	**Africa**	**Europe**	**Asia**
Geo-location	**Egypt**	**Florida**	**France**	**Ghana**	**Greece**	**Hong Kong**
# of mutations in M (M_M)	19	92	3	6	17	4
# of unique seqs. M (U_M)	22	131	4	7	3	5
Avg. # of mutations per unit unique seqs. ($\frac{M_M}{U_M}$)	0.86	0.70	0.75	0.86	5.67	0.80
Continent	**Asia**	**N-America**	**N-America**	**N-America**	**Asia**	**N-America**
Geo-location	**India**	**Massachusetts**	**Michigan**	**Minnesota**	**Pakistan**	**Pennsylvania**
# of mutations in M (M_M)	16	93	96	64	6	87
# of unique seqs. M (U_M)	20	92	78	77	7	105
Avg. # of mutations per unit unique seqs. ($\frac{M_M}{U_M}$)	0.80	1.01	1.23	0.83	0.86	0.83
Continent	**S-America**	**Europe**	**Europe**	**Europe**	**N-America**	**Africa**
Geo-location	**Peru**	**Poland**	**Serbia**	**Spain**	**Texas**	**Tunisia**
# of mutations in M (M_M)	12	1	2	3	94	2
# of unique seqs. M (U_M)	8	2	3	4	101	3
Avg. # of mutations per unit unique seqs. ($\frac{M_M}{U_M}$)	1.50	0.50	0.67	0.75	0.93	0.67

A relatively large number of mutations were found in the M proteins from Greece. The ratio $\frac{M_M}{U_M} = 5.67 > 5$ for Greece implied that any randomly chosen M protein variants possessed five mutations (Table10). In California, the highest number of unique M proteins possessed relatively very few mutations. Almost surely, no M protein from California contains more than one mutation ($\frac{M_M}{U_M} = 0.72 < 1$), whereas each M protein from Michigan and Massachusetts contains a single mutation ($\frac{M_M}{U_M} > 1$). Most of the unique M protein variants from Peru were likely to contain double mutations ($\frac{M_M}{U_M} = 1.5 > 1$)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0472.v1

(Table10).

All North American geo-locations shared a sum of 24 mutations in the M protein variants at positions 2, 7, 17, 23, 28, 33, 34, 60, 69, 70, 81, 82, 85, 89, 98, 104, 109, 125, 142, 155, 173, 175, 208, and 209 (**Supplementary file-I**). On the other hand, not a single common mutation in the M proteins was noticed in geo-locations from Asia and the same was observed in Africa and Europe. Each M protein from India shared 9 mutations with those of each North American geo-location, at positions 2, 17, 69, 70, 82, 104, 125, 142, and 209. Among the 24 common mutations from geo-locations in North America, only two mutations at positions 17 and 23 were shared with M proteins from Greece.

3.1.4. Nucleocapsid protein variants and mutations

The frequency of unique N protein mutations across the 24 geo-locations is presented in Table 11.

Table 11: Number of unique N protein mutations possessed in each geo-location.

Continent	**Oceania**	**Europe**	**Asia**	**Asia**	**N-America**	**S-America**
Geo-location	**Australia**	**Austria**	**Bahrain**	**Bangladesh**	**California**	**Chile**
# of mutations in N (M_N)	200	21	38	86	362	16
# of unique seqs. N (U_N)	213	22	33	53	1345	16
Avg. # of mutations per unit unique seqs. ($\frac{M_N}{U_N}$)	0.94	0.95	1.15	1.62	0.27	1.00
Continent	**Africa**	**N-America**	**Europe**	**Africa**	**Europe**	**Asia**
Geo-location	**Egypt**	**Florida**	**France**	**Ghana**	**Greece**	**Hong Kong**
# of mutations in N (M_N)	83	356	7	34	9	32
# of unique seqs. N (U_N)	116	973	6	41	9	28
Avg. # of mutations per unit unique seqs. ($\frac{M_N}{U_N}$)	0.72	0.37	1.17	0.83	1.00	1.14
Continent	**Asia**	**N-America**	**N-America**	**N-America**	**Asia**	**N-America**
Geo-location	**India**	**Massachusetts**	**Michigan**	**Minnesota**	**Pakistan**	**Pennsylvania**
# of mutations in N (M_N)	84	363	238	322	31	280
# of unique seqs. N (U_N)	86	625	418	481	33	643
Avg. # of mutations per unit unique seqs. ($\frac{M_N}{U_N}$)	0.98	0.58	0.57	0.67	0.94	0.44
Continent	**S-America**	**Europe**	**Europe**	**Europe**	**N-America**	**Africa**
Geo-location	**Peru**	**Poland**	**Serbia**	**Spain**	**Texas**	**Tunisia**
# of mutations in N (M_N)	20	20	24	17	286	24
# of unique seqs. N (U_N)	19	22	22	21	644	22
Avg. # of mutations per unit unique seqs. ($\frac{M_N}{U_N}$)	1.05	0.91	1.09	0.81	0.44	1.09

It was observed that the least number of mutations was possessed by the unique N proteins from California ($\frac{M_N}{U_N} = 0.27 < 1$), whereas 53 unique N protein variants from Bangladesh had 86 mutations ($\frac{M_N}{U_N} = 1.62 > 1$) (Table 11). Every unique N protein-variant contain at least a single mutation which is followed by the ratio ($\frac{M_N}{U_N} = 1.62 > 1$). Likewise, each unique N variant from Bahrain, Peru, Chile, France, Greece, Hong Kong, India, Serbia, and Tunisia contain at least one mutation (for each geo-location ($\frac{M_N}{U_N} = 1.62 \geq 1$).

Furthermore, it was noticed that 153 mutations were shared among all unique N proteins from each geo-location in North America. Only 6 mutations at positions 3, 194, 202, 203, 204 and 377 were common across Asian geo-locations, whereas only two mutations at positions 203 and 204 were found in the N variants from the European geo-locations. There were 9 mutations at positions 9, 194, 202, 203, 204, 205, 220, 235, and 238 in the N proteins detected in the African geo-locations.

3.1.5. ORF3a protein variants and mutations

The frequency of unique ORF3a protein mutations across the 24 geo-locations is presented in Table 12.

Table 12: Number of unique ORF3a protein mutations possessed in each geo-location.

Continent	**Oceania**	**Europe**	**Asia**	**Asia**	**N-America**	**S-America**
Geo-location	**Australia**	**Austria**	**Bahrain**	**Bangladesh**	**California**	**Chile**
# of mutations in ORF3a (M_{3a})	151	16	28	51	264	15
# of unique seqs. ORF3a (U_{3a})	132	14	27	59	1073	16
Avg. # of mutations per unit unique seqs. ($\frac{M_{3a}}{U_{3a}}$)	1.14	1.14	1.04	0.86	0.25	0.94
Continent	**Africa**	**N-America**	**Europe**	**Africa**	**Europe**	**Asia**
Geo-location	**Egypt**	**Florida**	**France**	**Ghana**	**Greece**	**Hong Kong**
# of mutations in ORF3a (M_{3a})	56	264	9	27	25	13
# of unique seqs. ORF3a (U_{3a})	81	808	10	23	13	17
Avg. # of mutations per unit unique seqs. ($\frac{M_{3a}}{U_{3a}}$)	0.69	0.33	0.90	1.17	1.92	0.76
Continent	**Asia**	**N-America**	**N-America**	**N-America**	**Asia**	**N-America**
Geo-location	**India**	**Massachusetts**	**Michigan**	**Minnesota**	**Pakistan**	**Pennsylvania**
# of mutations in ORF3a (M_{3a})	62	232	235	242	47	225
# of unique seqs. ORF3a (U_{3a})	73	468	389	456	32	561
Avg. # of mutations per unit unique seqs. ($\frac{M_{3a}}{U_{3a}}$)	0.85	0.50	0.60	0.53	1.47	0.40
Continent	**S-America**	**Europe**	**Europe**	**Europe**	**N-America**	**Africa**
Geo-location	**Peru**	**Poland**	**Serbia**	**Spain**	**Texas**	**Tunisia**
# of mutations in ORF3a (M_{3a})	16	23	19	14	247	12
# of unique seqs. ORF3a (U_{3a})	16	21	17	13	532	10
Avg. # of mutations per unit unique seqs. ($\frac{M_{3a}}{U_{3a}}$)	1.00	1.10	1.12	1.08	0.46	1.20

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0472.v1

From Table 12, it was observed that the least number of mutations was possessed by ORF3a variants from California, where the highest number of unique ORF3a variants available though $(\frac{M_{3a}}{U_{3a}}) = 0.25 << 1$. On the other hand, 13 ORF3a variants from Greece had 25 mutations altogether. Therefore, almost every ORF3a variant was likely to contain double mutations $(\frac{M_{3a}}{U_{3a}}) = 1.92 \cong 2$. Furthermore, each ORF3a variant from Australia, Austria, Bahrain, Chile, France, Ghana, Pakistan, Peru, Poland, Serbia, Spain, and Tunisia contains at least one mutation, that is Q57, but not more than two mutations since the ratio, $\frac{M_{3a}}{U_{3a}}$ lies in between 1 and 2.

A total of 167 common mutations in ORF3a variants across the North American geo-locations were detected, whereas the only common mutation, Q57 was detected in the European geo-locations. It was noted that unique ORF3a variants from Texas, Pennsylvania, Florida, Michigan, and Minnesota had common mutations at positions 243, 224, 255, 229, and 238, respectively, from California. ORF3a variants from African geo-locations share five common mutations at positions 57, 100, 155, 171, and 224. Also, three mutations at positions 57, 175, and 223 were possessed by the ORF3a variants from each Asian geo-location. It was noted that unique ORF3a variants shared 225 mutations among 264 in total in both California and Massachusetts.

3.1.6. ORF6 protein variants and mutations

The frequency of unique ORF6 protein mutations across the 24 geo-locations is presented in Table 13.

Table 13: Number of unique ORF6 protein mutations possessed in each geo-location.

Continent	**Oceania**	**Europe**	**Asia**	**Asia**	**N-America**	**S-America**
Geo-location	**Australia**	**Austria**	**Bahrain**	**Bangladesh**	**California**	**Chile**
# of mutations in ORF6 (M_6)	44	2	30	8	59	2
# of unique seqs. ORF6 (U_6)	19	3	7	9	104	3
Avg. # of mutations per unit unique seqs. ($\frac{M_6}{U_6}$)	2.32	0.67	4.29	0.89	0.57	0.67
Continent	**Africa**	**N-America**	**Europe**	**Africa**	**Europe**	**Asia**
Geo-location	**Egypt**	**Florida**	**France**	**Ghana**	**Greece**	**Hong Kong**
# of mutations in ORF6 (M_6)	6	46	2	15	1	10
# of unique seqs. ORF6 (U_6)	10	65	3	10	2	3
Avg. # of mutations per unit unique seqs. ($\frac{M_6}{U_6}$)	0.60	0.71	0.67	1.50	0.50	3.33
Continent	**Asia**	**N-America**	**N-America**	**N-America**	**Asia**	**N-America**
Geo-location	**India**	**Massachusetts**	**Michigan**	**Minnesota**	**Pakistan**	**Pennsylvania**
# of mutations in ORF6 (M_6)	5	45	45	57	4	38
# of unique seqs. ORF6 (U_6)	7	47	38	45	5	52
Avg. # of mutations per unit unique seqs. ($\frac{M_6}{U_6}$)	0.71	0.96	1.18	1.27	0.80	0.73
Continent	**S-America**	**Europe**	**Europe**	**Europe**	**N-America**	**Africa**
Geo-location	**Peru**	**Poland**	**Serbia**	**Spain**	**Texas**	**Tunisia**
# of mutations in ORF6 (M_6)	1	0	0	2	55	0
# of unique seqs. ORF6 (U_6)	2	1	1	3	61	1
Avg. # of mutations per unit unique seqs. ($\frac{M_6}{U_6}$)	0.50	0.00	0.00	0.67	0.90	0.00

The probability of having quadruple mutations in a chosen unique ORF6 variant from Bahrain was nearly 1 as the ratio $\frac{M_6}{U_6} = 4.29 > 4$ (Table 13. Almost surely, each ORF6 variant from Hong Kong ($\frac{M_6}{U_6} = 3.33 > 3$) and Australia ($\frac{M_6}{U_6} = 2.32 > 2$) contains triple and double mutations, respectively. Also, it was noticed that no new ORF6 variant was detected in Poland, Serbia, and Tunisia.

There were 25 common mutations in ORF6 variants in each geo-location of North America, whereas no common mutation in ORF6 was found in the European geo-locations. Likewise, in Asian and African geo-locations, no common mutation was detected for ORF6 variants.

3.1.7. ORF7a protein variants and mutations

The frequency of unique ORF7a protein mutations across the 24 geo-locations is presented in Table 14.

Table 14: Number of unique ORF7a protein mutations possessed in each geo-location.

Continent	**Oceania**	**Europe**	**Asia**	**Asia**	**N-America**	**S-America**
Geo-location	**Australia**	**Austria**	**Bahrain**	**Bangladesh**	**California**	**Chile**
# of mutations in ORF7a (M_{7a})	59	5	15	21	120	5
# of unique seqs. ORF7a (U_{7a})	58	5	18	15	330	5
Avg. # of mutations per unit unique seqs. ($\frac{M_{7a}}{U_{7a}}$)	1.02	1.00	0.83	1.40	0.36	1.00
Continent	**Africa**	**N-America**	**Europe**	**Africa**	**Europe**	**Asia**
Geo-location	**Egypt**	**Florida**	**France**	**Ghana**	**Greece**	**Hong Kong**
# of mutations in ORF7a (M_{7a})	18	108	0	13	7	5
# of unique seqs. ORF7a (U_{7a})	20	314	1	10	2	5
Avg. # of mutations per unit unique seqs. ($\frac{M_{7a}}{U_{7a}}$)	0.90	0.34	0.00	1.30	3.50	1.00
Continent	**Asia**	**N-America**	**N-America**	**N-America**	**Asia**	**N-America**
Geo-location	**India**	**Massachusetts**	**Michigan**	**Minnesota**	**Pakistan**	**Pennsylvania**
# of mutations in ORF7a (M_{7a})	25	114	110	103	5	105
# of unique seqs. ORF7a (U_{7a})	23	184	199	758	6	202
Avg. # of mutations per unit unique seqs. ($\frac{M_{7a}}{U_{7a}}$)	1.09	0.62	0.55	0.14	0.83	0.52
Continent	**S-America**	**Europe**	**Europe**	**Europe**	**N-America**	**Africa**
Geo-location	**Peru**	**Poland**	**Serbia**	**Spain**	**Texas**	**Tunisia**
# of mutations in ORF7a (M_{7a})	29	6	3	1	109	5
# of unique seqs. ORF7a (U_{7a})	9	8	3	2	190	7
Avg. # of mutations per unit unique seqs. ($\frac{M_{7a}}{U_{7a}}$)	3.22	0.75	1.00	0.50	0.57	0.71

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0472.v1

The ratio $\frac{M_{7a}}{U_{7a}} > 3$ in Greece and Peru implied that most unique variants must have at least three mutations (Table 14). Unique ORF7a variants from Australia, Austria, Bangladesh, Chile, Egypt, Ghana, Hong Kong, India, Pakistan, and Serbia must contain at least a single mutation as in each case, the ratio was found greater than or equal/near to 1. Furthermore, it was observed that no new ORF7a sequence was found among 90 infected patients in France, so far.

Ninety-two common mutations were detected in the unique ORF7a variants in the North American geo-locations, whereas no common mutation was observed in the European geo-locations. Only one common mutation at position 28 in Asian geo-locations, and another single common mutation at position 14 in ORF7a was found in African countries. ORF7a protein sequences from Austria had four mutations at positions 79, 99, 102, and 103, commonly found in each geo-location in North America. Likewise, all unique mutations in ORF7a variants detected in Greece, Poland, and Serbia were present in each North American geo-location.

3.1.8. ORF7b protein variants and mutations

The frequency of unique ORF7b protein mutations across the 24 geo-locations is presented in Table 15.

Table 15: Number of unique ORF7b protein mutations possessed in each geo-location.

Continent	**Oceania**	**Europe**	**Asia**	**Asia**	**N-America**	**S-America**
Geo-location	**Australia**	**Austria**	**Bahrain**	**Bangladesh**	**California**	**Chile**
# of mutations in ORF7b (M_{7b})	19	1	3	5	40	1
# of unique seqs. ORF7b (U_{7b})	14	2	4	6	89	2
Avg. # of mutations per unit unique seqs. ($\frac{M_{7b}}{U_{7b}}$)	1.36	0.50	0.75	0.83	0.45	0.50
Continent	**Africa**	**N-America**	**Europe**	**Africa**	**Europe**	**Asia**
Geo-location	**Egypt**	**Florida**	**France**	**Ghana**	**Greece**	**Hong Kong**
# of mutations in ORF7b (M_{7b})	8	36	0	15	0	1
# of unique seqs. ORF7b (U_{7b})	11	63	1	7	1	2
Avg. # of mutations per unit unique seqs. ($\frac{M_{7b}}{U_{7b}}$)	0.73	0.57	0.00	2.14	0.00	0.50
Continent	**Asia**	**N-America**	**N-America**	**N-America**	**Asia**	**N-America**
Geo-location	**India**	**Massachusetts**	**Michigan**	**Minnesota**	**Pakistan**	**Pennsylvania**
# of mutations in ORF7b (M_{7b})	10	35	34	30	1	26
# of unique seqs. ORF7b (U_{7b})	7	46	45	59	2	38
Avg. # of mutations per unit unique seqs. ($\frac{M_{7b}}{U_{7b}}$)	1.43	0.76	0.76	0.51	0.50	0.68
Continent	**S-America**	**Europe**	**Europe**	**Europe**	**N-America**	**Africa**
Geo-location	**Peru**	**Poland**	**Serbia**	**Spain**	**Texas**	**Tunisia**
# of mutations in ORF7b (M_{7b})	0	1	0	1	30	1
# of unique seqs. ORF7b (U_{7b})	1	2	1	2	43	2
Avg. # of mutations per unit unique seqs. ($\frac{M_{7b}}{U_{7b}}$)	0.00	0.50	0.00	0.50	0.70	0.50

Compared to the wildtype ORF7b (YP_009725318), no new ORF7b variant was found in France, Greece, Peru, and Serbia, whereas only one variant other than the wild ORF7b was found in Austria, Chile, Hong Kong, Pakistan, Poland, Spain, and Tunisia. Each ORF7b variant from Australia and India contained at least a single mutant.

There were 17 common mutations at positions 2, 3, 4, 5, 6, 8, 10, 13, 14, 15, 18, 31, 32, 34, 40, 42, and 43 in all North American geo-locations. No ORF7b variants from North America possessed double mutations based on the ratio $\frac{M_{7b}}{U_{7b}} < 1$ for each North American geo-location (Table 15).

3.1.9. ORF8 protein variants and mutations

The frequency of unique ORF8 protein mutations across the 24 geo-locations is presented in Table 16.

Table 16: Number of unique ORF7b protein mutations possessed in each geo-location.

Continent	**Oceania**	**Europe**	**Asia**	**Asia**	**N-America**	**S-America**
Geo-location	**Australia**	**Austria**	**Bahrain**	**Bangladesh**	**California**	**Chile**
# of mutations in ORF8 (M_8)	33	2	14	23	117	4
# of unique seqs. ORF8 (U_8)	54	3	17	19	359	5
Avg. # of mutations per unit unique seqs. ($\frac{M_8}{U_8}$)	0.61	0.67	0.82	1.21	0.33	0.80
Continent	**Africa**	**N-America**	**Europe**	**Africa**	**Europe**	**Asia**
Geo-location	**Egypt**	**Florida**	**France**	**Ghana**	**Greece**	**Hong Kong**
# of mutations in ORF8 (M_8)	26	114	2	43	3	9
# of unique seqs. ORF8 (U_8)	34	231	3	12	4	10
Avg. # of mutations per unit unique seqs. ($\frac{M_8}{U_8}$)	0.76	0.49	0.67	3.58	0.75	0.90
Continent	**Asia**	**N-America**	**N-America**	**N-America**	**Asia**	**N-America**
Geo-location	**India**	**Massachusetts**	**Michigan**	**Minnesota**	**Pakistan**	**Pennsylvania**
# of mutations in ORF8 (M_8)	30	89	69	65	9	69
# of unique seqs. ORF8 (U_8)	27	137	77	118	10	135
Avg. # of mutations per unit unique seqs. ($\frac{M_8}{U_8}$)	1.11	0.65	0.90	0.55	0.90	0.51
Continent	**S-America**	**Europe**	**Europe**	**Europe**	**N-America**	**Africa**
Geo-location	**Peru**	**Poland**	**Serbia**	**Spain**	**Texas**	**Tunisia**
# of mutations in ORF8 (M_8)	7	5	5	3	78	6
# of unique seqs. ORF8 (U_8)	8	6	6	3	154	7
Avg. # of mutations per unit unique seqs. ($\frac{M_8}{U_8}$)	0.88	0.83	0.83	1.00	0.51	0.86

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0472.v1

In each geo-location, wildtype ORF8 protein mutated several times and emerged as a set of unique ORF8 variants in each geo-location. Every unique ORF8 variant from India and Bangladesh contains at least one mutation as the ratio in each case was greater than 1 (Table 16). A total of 32 shared mutations were identified across geo-locations in North America. It was noticed that L84 was the only common mutation found in Asian and African geo-locations.

3.1.10. ORF10 protein variants and mutations

The frequency of unique ORF10 protein mutations across the 24 geo-locations is presented in Table 17.

Table 17: Number of unique ORF7b protein mutations possessed in each geo-location.

Continent	**Oceania**	**Europe**	**Asia**	**Asia**	**N-America**	**S-America**
Geo-location	**Australia**	**Austria**	**Bahrain**	**Bangladesh**	**California**	**Chile**
# of mutations in ORF10 (M_{10})	13	1	2	9	29	0
# of unique seqs. ORF10 (U_{10})	16	2	3	11	61	1
Avg. # of mutations per unit unique seqs. ($\frac{M_{10}}{U_{10}}$)	0.81	0.50	0.67	0.82	0.48	0.00
Continent	**Africa**	**N-America**	**Europe**	**Africa**	**Europe**	**Asia**
Geo-location	**Egypt**	**Florida**	**France**	**Ghana**	**Greece**	**Hong Kong**
# of mutations in ORF10 (M_{10})	6	29	0	2	0	2
# of unique seqs. ORF10 (U_{10})	8	47	1	3	1	3
Avg. # of mutations per unit unique seqs. ($\frac{M_{10}}{U_{10}}$)	0.75	0.62	0.00	0.67	0.00	0.67
Continent	**Asia**	**N-America**	**N-America**	**N-America**	**Asia**	**N-America**
Geo-location	**India**	**Massachusetts**	**Michigan**	**Minnesota**	**Pakistan**	**Pennsylvania**
# of mutations in ORF10 (M_{10})	2	23	16	20	2	22
# of unique seqs. ORF10 (U_{10})	3	29	23	29	3	29
Avg. # of mutations per unit unique seqs. ($\frac{M_{10}}{U_{10}}$)	0.67	0.79	0.70	0.69	0.67	0.76
Continent	**S-America**	**Europe**	**Europe**	**Europe**	**N-America**	**Africa**
Geo-location	**Peru**	**Poland**	**Serbia**	**Spain**	**Texas**	**Tunisia**
# of mutations in ORF10 (M_{10})	8	1	1	2	21	2
# of unique seqs. ORF10 (U_{10})	5	2	2	3	39	4
Avg. # of mutations per unit unique seqs. ($\frac{M_{10}}{U_{10}}$)	1.60	0.50	0.50	0.67	0.54	0.50

The ratio $\frac{M_{10}}{U_{10}} = 0$ implied that other than wildtype ORF10 (YP_009725255), no new ORF10 protein emerged in Chile, France, and Greece, although every amino acid contained mutations at each position starting from 1 to 38. In all 24 geo-locations, every unique ORF10 variant possessed only a single mutation (as in each case $0 < \frac{M_{10}}{U_{10}} < 2$) (Table 17).

In North American geo-locations, a set of common mutations in ORF10 variants at positions 4, 8, 10, 23, 24, 27, 28, 30, and 37 were identified. No other continental geo-locations have common mutations in ORF10. It was noted that an ORF10 variant (QKG88643.1) possessed the M1G mutation.

3.2. Mutations in the invariant residue regions of various proteins of SARS-CoV-2

The ORF10 protein was the unique protein present in SARS-CoV-2, which is not present in any other beta-coronavirus. So except for ORF10, other unique protein variants of four types of beta-coronaviruses were obtained from the NCBI database (Table 3). Further, sequence-based homology using the Clustal-Omega webserver of each unique protein variant of four types with reference protein sequence (NC_045512-China) was obtained (**Supplementary file-II**). Based on the alignment, invariant residue regions of length greater than three amino acids were detected (Table 18). From amino acid homology alignment, it was observed that the SARS-CoV-2 reference protein sequences of *NC_045512* with a set of invariant residues were shared by those proteins of four other different types of beta-coronaviruses. There are several invariant regions identified in all proteins as indicated in Table 18. Each of the S, E, M, N, ORF3a, ORF6, ORF7a, ORF7b, and ORF8 proteins of five different coronaviruses shared 29, 4, 9, 11, 6, 1, 3, 2, and 2 invariant residue regions. Further, it is worth noting that the largest invariant region with a length of 101 was identified in the S protein. These invariant regions possibly served as sets of functional units in the respective proteins, indicating why these were conserved in the beta-coronavirus family.

Table 18: Invariant regions and domain specifications in proteins of four type of CoVs

Protein	Invariant residues	Total # of residues	Protein	Invariant residues	Total # of residues	Protein	Invariant residues	Total # of residues
S	34–38	5	E	3–24	22	ORF3a	31–36	4
S	102–104	3	E	26–36	11	ORF3a	53–58	4
S	165–167	3	E	43–54	12	ORF3a	135–142	8
S	189–191	3	E	57–67	11	ORF3a	154–162	9
S	281–284	4				ORF3a	244–255	12
S	310–320	11	**Protein**	**Invariant residues**		ORF3a	262–275	14
S	374–383	10	M	5–11	7			
S	418–429	12	M	16–26	11	**Protein**	**Invariant residues**	
S	509–518	10	M	41–51	11	ORF6	1–15	15
S	520–528	9	M	53–75	23			
S	538–546	9	M	98–124	27			
S	591–603	13	M	135–144	10			
S	608–618	11	M	156–167	12			
S	659–674	16	M	170–187	18			
S	751–767	17	M	198–210	13	**Protein**	**Invariant residues**	
S	797–809	13				ORF7a	15–31	17
S	814–833	18				ORF7a	37–58	22
S	846–867	22	**Protein**	**Invariant residues**		ORF7a	75–93	19
S	885–921	37	N	38–62	25			
S	944–1044	101	N	66–78	13			
S	1074–1083	10	N	81–93	13			
S	1090–1096	7	N	104–119	16			
S	1115–1122	8	N	132–151	20	**Protein**	**Invariant residues**	
S	1134–1163	30	N	158–181	24	ORF7b	6–25	19
S	1165–1190	26	N	217–231	15	ORF7b	27–33	4
S	1192–1207	16	N	243–266	24			
S	1209–1229	21	N	270–289	20	**Protein**	**Invariant residues**	
S	1234–1246	13	N	297–325	28	ORF8	35–38	3
S	1262–1273	12	N	350–375	26	ORF8	88–91	3

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0472.v1

Over time and due to intraspecies evolution, SARS-CoV-2 proteins have acquired several mutations even in the invariant regions. The total frequency and respective percentage of mutations detected in each invariant residue window of all proteins are presented in Table 19.

Table 19: Frequency and respective percentage of mutations detected in each invariant residue window of S proteins

S proteins invariant residues			Number of mutations										
Invariant residues	**Total # of residues**	**Domain**	**Tunisia**	**Texas**	**Spain**	**Serbia**	**Poland**	**Peru**	**Pennsylvania**	**Pakistan**	**Minnesota**	**Michigan**	**Massachusetts**
34–38	5	S1	0	5	0	0	0	0	4	0	1	1	5
102–104	3	S1	0	3	0	0	0	3	3	1	3	3	3
165–167	3	S1	0	3	0	0	0	0	3	0	3	3	3
189–191	3	S1	0	3	0	0	1	1	3	0	3	1	3
281–284	4	S1	0	4	0	0	0	4	4	0	4	4	4
310–320	11	S1	1	11	0	0	0	0	11	0	11	11	11
374–383	10	S1	0	7	0	0	0	10	4	0	3	3	10
418–429	12	S1	0	12	0	0	0	3	1	0	3	1	12
509–518	10	S1	0	10	6	0	0	0	10	0	10	9	10
520–528	9	S1	1	9	4	0	0	0	2	0	9	2	9
538–546	9	S1	0	9	0	1	0	0	1	0	2	0	9
591–603	13	S1	0	11	0	0	0	0	0	0	3	0	4
608–618	11	S1	1	11	1	1	1	1	6	1	6	4	4
659–674	16	S1	0	16	0	0	0	0	14	1	15	7	4
751–767	17	S2	0	8	1	0	0	0	14	0	11	7	14
797–809	13	S2	0	6	2	0	0	0	5	0	11	1	13
814–833	18	S2 and S2’	0	11	0	0	0	0	9	14	18	8	19
846–867	22	S2’	0	12	0	0	0	2	8	3	6	5	5
885–921	37	S2’	2	16	0	0	0	3	5	1	36	31	8
944–1044	101	S2’	2	88	1	1	1	3	14	1	72	64	28
1074–1083	10	S2’	0	4	0	1	0	2	4	1	5	3	2
1090–1096	7	S2’	0	1	0	0	0	0	2	0	5	7	2
1115–1122	8	S2’	1	4	1	0	1	1	5	1	5	3	5
1134–1163	30	S2’	0	24	1	0	1	0	8	3	19	8	9
1165–1190	26	S2’	0	25	0	0	0	3	12	1	24	7	12
1192–1207	16	S2’	0	10	1	0	1	0	7	1	16	5	5
1209–1229	21	S2’ (1214-1229-TMD)	0	19	0	1	1	1	8	2	21	5	14
1234–1246	13	S2’ (1234-TMD)	2	13	0	0	0	0	9	0	8	4	7
1262–1273	12	S2’	0	4	0	0	1	0	3	1	4	4	5

S proteins invariant residues			Number of mutations											
Invariant residues	**Total # of residues**	**Domain**	**India**	**Hong Kong**	**Greece**	**Ghana**	**France**	**Florida**	**Egypt**	**Chile**	**California**	**Bangladesh**	**Bahrain**	**Austria**	**Australia**
34–38	5	S1	2	0	0	0	0	2	0	0	5	2	0	0	2
102–104	3	S1	0	0	0	0	0	3	1	0	3	1	1	0	3
165–167	3	S1	0	0	0	0	0	3	0	0	3	0	0	0	3
189–191	3	S1	1	0	0	1	0	3	0	0	3	1	0	0	2
281–284	4	S1	0	0	0	4	0	4	0	0	4	0	0	4	4
310–320	11	S1	0	1	0	11	0	11	1	0	11	0	0	11	11
374–383	10	S1	4	0	0	0	2	2	3	10	9	10	0	0	0
418–429	12	S1	0	0	0	0	0	2	2	0	3	0	0	0	0
509–518	10	S1	0	10	0	1	0	10	0	0	10	1	1	0	10
520–528	9	S1	0	9	0	0	0	3	2	0	8	1	0	0	9
538–546	9	S1	0	0	0	0	0	9	0	0	1	0	0	0	0
591–603	13	S1	0	0	0	0	0	4	0	0	5	1	0	0	1
608–618	11	S1	1	1	1	2	2	11	2	1	6	1	1	1	2
659–674	16	S1	0	0	0	1	0	16	0	0	11	1	0	0	6
751–767	17	S2	0	0	0	0	0	7	1	1	14	0	0	1	9
797–809	13	S2	5	0	0	2	0	5	3	0	10	13	0	0	1
814–833	18	S2 and S2’	6	1	0	1	0	7	3	0	10	4	1	0	2
846–867	22	S2’	0	0	0	0	0	15	1	0	10	0	0	0	2
885–921	37	S2’	1	0	1	1	0	6	1	0	32	0	1	1	3
944–1044	101	S2’	1	2	0	2	1	48	5	0	93	4	3	3	12
1074–1083	10	S2’	2	0	0	1	0	10	8	1	4	1	1	0	3
1090–1096	7	S2’	1	1	0	1	0	4	7	0	5	2	1	0	1
1115–1122	8	S2’	3	1	1	2	0	8	1	0	6	2	1	2	2
1134–1163	30	S2’	3	0	0	3	1	14	3	0	25	0	3	0	6
1165–1190	26	S2’	5	2	0	0	0	16	2	0	24	3	0	0	2
1192–1207	16	S2’	1	0	0	0	0	14	0	0	16	0	0	2	0
1209–1229	21	S2’ (1214-1229-TMD)	0	0	0	3	0	14	1	1	21	0	0	0	4
1234–1246	13	S2’ (1234-TMD)	1	0	0	1	0	7	4	1	13	2	1	0	7
1262–1273	12	S2’	2	0	1	0	1	4	7	0	9	4	1	1	4

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0472.v1

In all invariant regions of the S protein, unique variants from California, Florida, Texas, Minnesota, and Massachusetts possessed several mutations (Table 19). Notably, unique S protein variants from California, Texas, and Minnesota had possessed 93, 88, and 72 distinct mutations, respectively, in the invariant region of 101 amino acid residues. Among 29 invariant regions, only seven of the S proteins from Tunisia had a minimal number of mutations, with a maximum of two in each region. Likewise, S protein variants from Spain, Poland, Serbia, Greece, and France got a minimal number of mutations in nine, eight, five, four, and seven invariant regions, respectively. S protein variants from other geo-locations possessed a relatively (with regard to the North American geo-locations) smaller number of mutations in the invariant regions. In more than 50% of the 29 invariant regions, S protein variants from India, Bangladesh, Austria, Egypt, and Pakistan possessed a small number of mutations (Table 19). It was noteworthy that in India, Bangladesh, Austria, Egypt, and Pakistan, only a maximum of five mutations were found in the largest invariant region of the S2 domain of the S proteins.

Several mutations were identified in the S1, S2, S2' domains of the S protein (Table 19). The S1 domain of the S protein attaches the virion to the cell M by interacting with the host ACE2 receptor, initiating the infection. Also, the S2 domain contributes to the fusion of the virion and cellular membranes by acting as a class-I viral fusion protein, and the S2' domain acts as a viral fusion peptide which is unmasked following the S2 cleavage occurring after virus endocytosis [48]. These functions might be modified due to several mutations occurring in the invariant regions (postulated as important functional sites for the virus). Whether these mutations in the invariant regions in the S1, S2 and S2' domains would increase the infectivity of the virus is not clear but definitely remains a matter of concern.

Invariant regions in the E, M, and N proteins of five CoVs which include SARS-CoV-2 too, are presented in Table 20. There were 4, 9, and 11 invariant regions identified in the E, M, and N proteins, respectively.

Table 20: Frequency and respective percentage of mutations detected in each invariant residue window of the E, M, and N proteins

		Number of mutations											
Protein	**Invariant residues**	**Tunisia**	**Texas**	**Spain**	**Serbia**	**Poland**	**Peru**	**Pennsylvania**	**Pakistan**	**Minnesota**	**Michigan**	**Massachusetts**	**India**
E	3–24	0	11	0	0	0	7	10	1	7	12	8	2
E	26–36	0	3	0	0	0	2	1	1	11	11	1	0
E	43–54	0	1	0	0	0	0	3	0	9	4	9	9
E	57–67	0	5	1	0	0	0	4	0	11	5	11	11
Protein	**Invariant residues**	**Hong Kong**	**Greece**	**Ghana**	**France**	**Florida**	**Egypt**	**Chile**	**California**	**Bangladesh**	**Bahrain**	**Austria**	**Australia**
E	3–24	0	0	2	0	11	5	0	15	3	0	0	7
E	26–36	0	0	0	0	5	2	0	11	2	0	0	11
E	43–54	0	0	0	0	6	1	0	10	0	0	0	12
E	57–67	0	0	1	0	8	0	0	9	0	0	0	11
Protein	**Invariant residues**	**Tunisia**	**Texas**	**Spain**	**Serbia**	**Poland**	**Peru**	**Pennsylvania**	**Pakistan**	**Minnesota**	**Michigan**	**Massachusetts**	**India**
M	5–11	1	3	0	0	0	0	3	0	3	2	3	0
M	16–26	0	4	0	0	0	1	4	0	3	11	3	1
M	41–51	0	2	0	0	0	8	4	1	1	7	11	0
M	53–75	0	9	1	1	0	1	10	0	8	7	18	4
M	98–124	0	15	0	0	0	0	6	1	6	16	7	2
M	135–144	0	3	0	0	0	0	3	0	2	3	6	1
M	156–167	0	10	0	0	0	0	8	0	2	0	4	0
M	170–187	0	10	0	0	0	0	5	1	4	2	2	0
M	198–210	0	3	0	0	0	0	4	0	2	4	2	1
Protein	**Invariant residues**	**Hong Kong**	**Greece**	**Ghana**	**France**	**Florida**	**Egypt**	**Chile**	**California**	**Bangladesh**	**Bahrain**	**Austria**	**Australia**
M	5–11	0	0	0	0	4	0	0	3	1	1	0	3
M	16–26	0	8	0	0	4	0	0	10	0	0	0	1
M	41–51	0	0	0	0	4	0	0	9	0	0	0	1
M	53–75	0	0	1	1	9	4	0	20	0	0	0	4
M	98–124	1	0	0	0	16	3	0	15	8	1	0	3
M	135–144	1	0	1	0	3	0	0	10	0	0	0	3
M	156–167	0	0	0	0	4	0	0	4	0	0	0	0
M	170–187	0	0	1	1	2	1	0	5	0	0	0	4
M	198–210	1	4	1	0	5	0	0	5	1	1	0	2
Protein	**Invariant residues**	**Tunisia**	**Texas**	**Spain**	**Serbia**	**Poland**	**Peru**	**Pennsylvania**	**Pakistan**	**Minnesota**	**Michigan**	**Massachusetts**	**India**
N	38–62	0	8	0	0	0	0	8	0	6	8	9	1
N	66–78	0	3	0	1	0	0	3	3	13	3	13	1
N	81–93	1	4	0	1	0	0	5	0	13	5	12	2
N	104–119	0	1	0	0	0	0	3	0	7	16	16	1
N	132–151	0	15	3	1	0	1	12	1	16	16	18	3
N	158–181	0	12	0	2	0	0	13	1	17	7	21	2
N	217–231	1	15	1	1	0	0	12	0	15	5	15	2
N	243–266	0	18	0	0	2	1	15	1	11	14	21	2
N	270–289	0	17	0	0	1	0	18	1	18	6	20	1
N	297–325	2	21	1	0	0	1	17	0	29	8	13	3
N	350–375	1	19	1	0	1	2	17	2	14	16	26	6
Protein	**Invariant residues**	**Hong Kong**	**Greece**	**Ghana**	**France**	**Florida**	**Egypt**	**Chile**	**California**	**Bangladesh**	**Bahrain**	**Austria**	**Australia**
N	38–62	0	0	1	0	14	0	0	11	1	1	0	4
N	66–78	0	0	2	0	13	2	0	6	6	0	0	4
N	81–93	0	0	1	0	13	1	0	7	3	1	0	4
N	104–119	0	0	0	0	11	1	0	15	0	0	0	1
N	132–151	3	0	1	0	15	6	0	16	2	2	0	7
N	158–181	0	0	1	0	19	4	1	22	3	0	0	5
N	217–231	3	0	1	0	12	2	0	15	7	2	1	4
N	243–266	1	0	0	1	16	2	1	21	1	1	1	16
N	270–289	0	0	0	0	20	2	0	17	3	0	0	20
N	297–325	2	0	1	0	29	3	2	19	4	5	0	29
N	350–375	0	1	0	0	19	7	1	20	1	2	1	8

No mutation was identified in the E protein variants from Tunisia, Serbia, Poland, Hong Kong, Greece, France (Table 20). On the other hand, the E protein variants from Chile, Bahrain, Austria, Australia, Texas, Pennsylvania, Minnesota, Michigan, Massachusetts, Florida, and California had a significant number of mutations in each invariant region. Very few mutations were identified in the E protein variants from India, Bangladesh, Spain, Peru, Egypt, Ghana, and Pakistan.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0472.v1

M protein variants in the North American and Oceanian geo-locations contained various mutations in each identified invariant region. In contrast, few mutations in the M proteins in the rest of the geo-locations, were detected in some invariant regions (Table 20).

N proteins from California, Texas, Minnesota, Michigan, Massachusetts, Pennsylvania, Florida, India, Bangladesh, Egypt, and Australia had many mutations in each invariant region. In some of the invariant regions, few mutations were detected in the N proteins from the rest of the geo-locations.

Mutations in the invariant regions of the SARS-CoV-2 ORF proteins are listed in Table 21. There were 6, 1, 3, 2, and 2 invariant regions found in ORF3a, ORF6, ORF7a, ORF7b, and ORF8 variants, respectively.

Table 21: Frequency and respective percentage of mutations detected in each invariant residue window of ORF3a, ORF6, ORF7a, ORF7b, and ORF8 proteins

		Number of mutations											
Protein	**Invariant residues**	**Tunisia**	**Texas**	**Spain**	**Serbia**	**Poland**	**Peru**	**Pennsylvania**	**Pakistan**	**Minnesota**	**Michigan**	**Massachusetts**	**India**
ORF3a	31–36	0	5	0	1	0	0	6	0	6	5	5	1
ORF3a	53–58	1	5	2	2	2	1	6	1	4	5	6	4
ORF3a	135–142	0	4	1	0	0	0	4	1	5	2	8	0
ORF3a	154–162	1	9	0	1	1	0	4	0	9	9	9	1
ORF3a	244–255	0	12	1	0	1	3	7	2	12	12	6	3
ORF3a	262–275	0	13	0	0	0	0	14	0	12	13	12	1
Protein	**Invariant residues**	**Hong Kong**	**Greece**	**Ghana**	**France**	**Florida**	**Egypt**	**Chile**	**California**	**Bangladesh**	**Bahrain**	**Austria**	**Australia**
ORF3a	31–36	0	1	0	1	6	3	1	6	0	1	1	2
ORF3a	53–58	1	1	2	1	5	3	1	6	3	1	1	5
ORF3a	135–142	0	0	0	0	8	0	1	8	1	1	0	7
ORF3a	154–162	0	1	1	1	9	2	0	9	1	0	0	9
ORF3a	244–255	1	6	0	1	12	3	1	11	3	0	0	4
ORF3a	262–275	1	0	0	0	13	1	0	14	2	1	0	5
Protein	**Invariant residues**	**Tunisia**	**Texas**	**Spain**	**Serbia**	**Poland**	**Peru**	**Pennsylvania**	**Pakistan**	**Minnesota**	**Michigan**	**Massachusetts**	**India**
ORF6	1–15	0	13	0	0	0	0	7	1	13	11	12	3
Protein	**Invariant residues**	**Hong Kong**	**Greece**	**Ghana**	**France**	**Florida**	**Egypt**	**Chile**	**California**	**Bangladesh**	**Bahrain**	**Austria**	**Australia**
ORF6	1–15	0	0	2	1	11	0	1	14	4	11	1	12
Protein	**Invariant residues**	**Tunisia**	**Texas**	**Spain**	**Serbia**	**Poland**	**Peru**	**Pennsylvania**	**Pakistan**	**Minnesota**	**Michigan**	**Massachusetts**	**India**
ORF7a	15–31	0	13	0	0	0	2	9	1	8	11	15	1
ORF7a	37–58	0	22	1	1	3	8	19	1	20	22	21	2
ORF7a	75–93	0	19	0	1	0	0	19	1	19	19	19	3
Protein	**Invariant residues**	**Hong Kong**	**Greece**	**Ghana**	**France**	**Florida**	**Egypt**	**Chile**	**California**	**Bangladesh**	**Bahrain**	**Austria**	**Australia**
ORF7a	15–31	0	0	1	0	9	2	1	17	4	1	0	5
ORF7a	37–58	1	0	1	0	22	1	1	22	3	2	1	9
ORF7a	75–93	0	0	1	0	19	4	0	19	3	3	1	8
Protein	**Invariant residues**	**Tunisia**	**Texas**	**Spain**	**Serbia**	**Poland**	**Peru**	**Pennsylvania**	**Pakistan**	**Minnesota**	**Michigan**	**Massachusetts**	**India**
ORF7b	6–25	0	14	0	0	0	0	12	1	13	18	17	5
ORF7b	27–33	0	5	0	0	0	0	4	0	3	4	5	1
Protein	**Invariant residues**	**Hong Kong**	**Greece**	**Ghana**	**France**	**Florida**	**Egypt**	**Chile**	**California**	**Bangladesh**	**Bahrain**	**Austria**	**Australia**
ORF7b	6–25	0	0	8	0	17	3	0	19	1	1	0	12
ORF7b	27–33	0	0	4	0	5	2	0	6	2	0	0	2
Protein	**Invariant residues**	**Tunisia**	**Texas**	**Spain**	**Serbia**	**Poland**	**Peru**	**Pennsylvania**	**Pakistan**	**Minnesota**	**Michigan**	**Massachusetts**	**India**
ORF8	35–38	0	2	0	0	1	0	2	0	4	4	4	0
ORF8	88–91	0	0	0	0	0	0	0	1	1	2	2	0
Protein	**Invariant residues**	**Hong Kong**	**Greece**	**Ghana**	**France**	**Florida**	**Egypt**	**Chile**	**California**	**Bangladesh**	**Bahrain**	**Austria**	**Australia**
ORF8	35–38	0	0	4	1	4	1	0	4	1	1	0	1
ORF8	88–91	0	0	0	0	4	0	0	4	0	0	0	0

ORF3a variants in the North American and Oceanian geo-locations had several mutations in each invariant region, whereas very few mutations were detected in some invariant regions (not in all) of ORF3a in India, Bangladesh, Egypt, Chile (Table 21).

No mutations at the invariant region in ORF6 variants were found in Tunisia, Spain, Serbia, Poland, Peru, Hong Kong, Greece, and Egypt. On the other hand, a handful of mutations in the invariant region were detected in the rest of the geo-locations. In the North American geo-locations, the number of mutations in ORF3a proteins was relatively big. In the North American geo-locations, in the invariant regions, a significant number of mutations in ORF3a proteins were found. A small number of mutations were found in the invariant regions of the ORF7a variant in the rest of the geo-locations with the exception of Tunisia, Hong Kong, Greece, and France (Table 21).

No mutations were found in the ORF7b invariant regions for the ORF7b proteins from Tunisia, Spain, Serbia, Poland, Peru, Hong Kong, Greece, France, Chile, and Austria. On the contrary, a significant number of mutations were detected in the two invariant regions of ORF7b from the rest of the geo-locations.

In two invariant regions, ORF8 variants from California possessed four mutations in each region, and in other North American geo-locations several mutations were also detected in the two invariant regions. However, in most of geo-locations, such as India, Tunisia, Spain, France, Greece, and so on, no mutations were found in the two invariant regions (Table 21).

4. Discussions and Remarks

Variants of S, E, M, N, ORF3a, ORF6, ORF7a, ORF7b, ORF8, and ORF10 proteins of SARS-CoV-2 from six continents comprising 24 geo-locations were analyzed. In each geo-location, a non-uniform frequency distribution of unique variants of

all ten proteins was noticed despite the identical number of total proteins. Clearly, various mutations in a given protein gave rise to several unique variants. Therefore, it turned out that during the intraspecies evolution of a given SARS-CoV-2 RNA genome, this later expressed variable amounts/rates of mutations in different genomic segments, which yielded irregularity in the frequency of protein variants (Table 20). So clearly, each SARS-CoV-2 genome from each geo-location accomplished the non-uniform frequency of unique protein variants. Notably, it was not the case for the other beta-coronaviruses. Furthermore, it was noticed that the total number of common invariant residues and common mutations possessed by each unique set of protein variants from all 24 geo-locations were significantly small. In most of the proteins, neither common invariant residues nor mutational residues were found. Therefore, a significantly large percentage of mutations in each protein variant of SARS-CoV-2 were unevenly or non-uniformly distributed over each of the 24 geo-locations. Thus, an equally evenly uneven pattern of distribution of unique variants of ten SARS-CoV-2 proteins over the 24 geo-locations was observed. It was expected that if common invariant residues are markedly small, common mutations must be significantly large. But this natural flow was not observed.

In spite of the factors behind that, the S glycoprotein remains is the main target for mutations reported so far, as it presents is the main structure for the SARS-CoV-2 attachment to host cells. Recent articles have reported a mouse-adapted WBP-1 SARS-CoV-2 strains (via several *in vivo* passages of Wuhan-Hu-1 (NC_045512)) characterized with two (Q493K and Q498H) mutations in its RBD [49, 50, 51, 52, 53, 54, 55, 56]. Both mutations seem responsible for converting resistant mice susceptible to SARS-CoV-2 because of the compatibility of host Ace2 receptor with the mutated RBD in the SARS-CoV-2 S protein. So, the avoiding mutations dynamics adoption is seeming impossible. Just two naturally emerging mutations in the S proteins generate a vigorous natural change in the WBP-1 strain enhanced the affinity to the mouse Ace2 receptor. The severe lung infections in mice closely resemble lung pathologies and symptoms caused by COVID-19 in humans. Is this result sending a new and/or support existing message? Such as the tight interactive human-animal would expect as a source of infection and/or man-made the virulence virus [57, 58, 59].

The frequency of distinct mutations possessed by the SARS-CoV-2 proteins in North American geo-locations, especially in California, was minimal in relative-percentage. In particular, no unique S protein variant contains more than one mutation in each sequence. It was also noticed that a significantly large number of common mutations (in the S protein, 487, 45, 22, 4, and 1 common mutations were found in North America, South America, Africa, Asia, and Africa) in all of the SARS-CoV-2 protein were found in North American geo-locations, unlike in other continental geo-locations. Therefore, it may be possible the uneven mutation across the geo-locations may be due to ethnicity of the population of these locations. Thus, such a non-uniform frequency of shared mutations on different continents led to the single mutation at position 614. A question arises in this regard: why do mutational factors vary in different geo-locations? Are they dependent on viral or host factors or both? The uneven distribution certainly demands a thorough investigation of demographic correlation with several factors of mutations of SARS-CoV-2.

Furthermore, it was observed that the reference proteins (of the SARS-CoV-2, NC_045512) contained several invariant domains across the other four different beta-coronaviruses (Table 18). Mostly in all North American geo-locations, many mutations were detected in each invariant (assumed to be evolutionary conserved) region of the S protein with regard to the reference SARS-CoV-2 S protein. Likewise, on other proteins also, several mutations were noted in all seven geo-locations from North America. In the rest of the 24 geo-locations, a few mutations in some of the invariant regions of the respective proteins were detected. Thus, in a short span of one year, the NC_045512 SARS-CoV-2 changed itself in such a manner that even the evolutionarily conserved domains (invariant regions) were altered, which might lead to emerging new SARS-CoV-2 variants with a different degree of virulence, infectivity, and transmissibility. These observations reopen the possibility to interrogate the SARS-CoV-2 origin. Correctly identifying the characteristics of SARS-CoV-2 would enable scientists to take appropriate measures to contain future pandemics. It could also help scientists to develop better diagnostics, vaccines and therapeutic tools.

Competing interests

The authors declare that there is no conflict of interest in this work.

Acknowledgements

The authors thank the researchers who generated and shared the sequencing data from NCBI SARS-CoV-2 Data Hub on which this research is based.

Authors' contributions

SSH conceptualized the study. SSH, VK, EMR contributed to the implementation of the research, to the analysis of the results. SSH and EMR wrote the initial draft of the manuscript. SSH, KL, PPC, ASA, GKA, AAAA, AL, GP, TMAEA, PA, GC, DB, MT, and SPS reviewed and edited. VNU, KT, BDU, WBC, and NGB provided constructive reviews and suggestions. All authors read final version and approve.

References

[1] B. Hu, H. Guo, P. Zhou, Z.-L. Shi, Characteristics of sars-cov-2 and covid-19, Nature Reviews Microbiology (2020) 1–14.

[2] K.-S. Yuen, Z.-W. Ye, S.-Y. Fung, C.-P. Chan, D.-Y. Jin, Sars-cov-2 and covid-19: The most important research questions, Cell & bioscience 10 (1) (2020) 1–5.

[3] N. J. Matheson, P. J. Lehner, How does sars-cov-2 cause covid-19?, Science 369 (6503) (2020) 510–511.

[4] D. Wu, T. Wu, Q. Liu, Z. Yang, The sars-cov-2 outbreak: what we know, International Journal of Infectious Diseases 94 (2020) 44–48.

[5] J. Zheng, Sars-cov-2: an emerging coronavirus that causes a global threat, International journal of biological sciences 16 (10) (2020) 1678.

[6] M. Lucas, U. Karrer, A. Lucas, P. Klenerman, Viral escape mechanisms–escapology taught by viruses, International journal of experimental pathology 82 (5) (2001) 269–286.

[7] M. R. Islam, M. N. Hoque, M. S. Rahman, A. R. U. Alam, M. Akther, J. A. Puspo, S. Akter, M. Sultana, K. A. Crandall, M. A. Hossain, Genome-wide analysis of sars-cov-2 virus strains circulating worldwide implicates heterogeneity, Scientific reports 10 (1) (2020) 1–9.

[8] S. Srivastava, S. Banu, P. Singh, D. T. Sowpati, R. K. Mishra, Sars-cov-2 genomics: An indian perspective on sequencing viral variants, Journal of Biosciences 46 (1) (2021) 1–14.

[9] S. S. Hassan, P. P. Choudhury, B. Roy, S. S. Jana, Missense mutations in sars-cov2 genomes from indian patients, Genomics 112 (6) (2020) 4622–4627.

[10] M. Pachetti, B. Marini, F. Benedetti, F. Giudici, E. Mauro, P. Storici, C. Masciovecchio, S. Angeletti, M. Ciccozzi, R. C. Gallo, et al., Emerging sars-cov-2 mutation hot spots include a novel rna-dependent-rna polymerase variant, Journal of translational medicine 18 (2020) 1–9.

[11] F. Robson, K. S. Khan, T. K. Le, C. Paris, S. Demirbag, P. Barfuss, P. Rocchi, W.-L. Ng, Coronavirus rna proofreading: molecular basis and therapeutic targeting, Molecular cell (2020).

[12] Y. Huang, C. Yang, X.-f. Xu, W. Xu, S.-w. Liu, Structural and functional properties of sars-cov-2 spike protein: potential antivirus drug development for covid-19, Acta Pharmacologica Sinica 41 (9) (2020) 1141–1149.

[13] W. T. Harvey, A. M. Carabelli, B. Jackson, R. K. Gupta, E. C. Thomson, E. M. Harrison, C. Ludden, R. Reeve, A. Rambaut, S. J. Peacock, et al., Sars-cov-2 variants, spike mutations and immune escape, Nature Reviews Microbiology (2021) 1–16.

[14] S. Belouzard, J. K. Millet, B. N. Licitra, G. R. Whittaker, Mechanisms of coronavirus cell entry mediated by the viral spike protein, Viruses 4 (6) (2012) 1011–1033.

[15] A. S. Lauring, E. B. Hodcroft, Genetic variants of sars-cov-2—what do they mean?, JAMA (2021).

[16] B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, et al., Tracking changes in sars-cov-2 spike: evidence that d614g increases infectivity of the covid-19 virus, Cell 182 (4) (2020) 812–827.

[17] M. Seyran, K. Takayama, V. N. Uversky, K. Lundstrom, G. Palù, S. P. Sherchan, D. Attrish, N. Rezaei, A. A. Aljabali, S. Ghosh, et al., The structural basis of accelerated host cell entry by sars-cov-2, The FEBS journal (2020).

[18] R. Sanjuán, P. Domingo-Calap, Mechanisms of viral mutation, Cellular and molecular life sciences 73 (23) (2016) 4433–4448.

[19] H. Lodish, S. L. Zipursky, Molecular cell biology, Biochem Mol Biol Educ 29 (2001) 126–133.

[20] E. C. Holmes, The comparative genomics of viral emergence, Proceedings of the National Academy of Sciences 107 (suppl 1) (2010) 1742–1746.

[21] S. S. Hassan, A. A. Aljabali, P. K. Panda, S. Ghosh, D. Attrish, P. P. Choudhury, M. Seyran, D. Pizzol, P. Adadi, T. M. Abd El-Aziz, et al., A unique view of sars-cov-2 through the lens of orf8 protein, Computers in biology and medicine 133 (2021) 104380.

[22] S. S. Hassan, K. Lundstrom, P. P. Choudhury, G. Palu, B. Uhal, R. Kandimalla, M. Seyran, A. Lal, S. P. Sherchan, G. K. Azad, et al., Implications derived from s-protein variants of sars-cov-2 from six continents, bioRxiv (2021).

[23] V. Bajaj, N. Gadi, A. P. Spihlman, S. C. Wu, C. H. Choi, V. R. Moulton, Aging, immunity, and covid-19: how age influences the host immune response to coronavirus infections?, Frontiers in Physiology 11 (2021) 1793.

[24] B. T. Rouse, S. Sehrawat, Immunity and immunopathology to viruses: what decides the outcome?, Nature Reviews Immunology 10 (7) (2010) 514–526.

[25] K. Kupferschmidt, The pandemic virus is slowly mutating. but does it matter? (2020).

[26] T. Leitner, S. Kumar, Where did sars-cov-2 come from?, Molecular biology and evolution 37 (9) (2020) 2463–2464.

[27] W. K. Jo, E. F. de Oliveira-Filho, A. Rasche, A. D. Greenwood, K. Osterrieder, J. F. Drexler, Potential zoonotic sources of sars-cov-2 infections, Transboundary and emerging diseases (2020).

[28] K. Lundstrom, M. Seyran, D. Pizzol, P. Adadi, T. Mohamed Abd El-Aziz, S. Hassan, A. Soares, R. Kandimalla, M. M. Tambuwala, A. A. Aljabali, et al., Origin of sars-cov-2 (2020).

[29] V. Kumar, B. Pruthvishree, T. Pande, D. Sinha, B. Singh, K. Dhama, Y. S. Malik, et al., Sars-cov-2 (covid-19): zoonotic origin and susceptibility of domestic and wild animals, J Pure Appl Microbiol 14 (suppl 1) (2020) 741–747.

[30] A. Banerjee, A. C. Doxey, K. Mossman, A. T. Irving, Unravelling the zoonotic origin and transmission of sars-cov-2, Trends in ecology & evolution (2020).

[31] E. Sallard, J. Halloy, D. Casane, E. Decroly, J. van Helden, Tracing the origins of sars-cov-2 in coronavirus phylogenies: a review, Environmental Chemistry Letters (2021) 1–17.

[32] R. Segreto, Y. Deigin, The genetic structure of sars-cov-2 does not rule out a laboratory origin: Sars-cov-2 chimeric structure and furin cleavage site might be the result of genetic manipulation, BioEssays (2020) 2000240.

[33] K. Sirotkin, D. Sirotkin, Might sars-cov-2 have arisen via serial passage through an animal host or cell culture? a potential explanation for much of the novel coronavirus' distinctive genome, BioEssays 42 (10) (2020) 2000091.

[34] M. Seyran, D. Pizzol, P. Adadi, T. M. A. El-Aziz, S. S. Hassan, A. Soares, R. Kandimalla, K. Lundstrom, M. Tambuwala, A. A. Aljabali, et al., Questions concerning the proximal origin of sars-cov-2, Journal of Medical Virology (2020).

[35] A. Maxmen, S. Mallapaty, The covid lab-leak hypothesis: what scientists do and don't know., Nature (2021).

[36] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, et al., A new coronavirus associated with human respiratory disease in china, Nature 579 (7798) (2020) 265–269.

[37] F. Sievers, D. G. Higgins, Clustal omega for making accurate alignments of many protein sequences, Protein Science 27 (1) (2018) 135–145.

[38] R. C. Edgar, Muscle: multiple sequence alignment with high accuracy and high throughput, Nucleic acids research 32 (5) (2004) 1792–1797.

[39] B. E. Pickett, E. L. Sadat, Y. Zhang, J. M. Noronha, R. B. Squires, V. Hunt, M. Liu, S. Kumar, S. Zaremba, Z. Gu, et al., Vipr: an open bioinformatics database and analysis resource for virology research, Nucleic acids research 40 (D1) (2012) D593–D598.

[40] J. Bendl, J. Stourac, O. Salanda, A. Pavelka, E. D. Wieben, J. Zendulka, J. Brezovsky, J. Damborsky, Predictsnp: robust and accurate consensus classifier for prediction of disease-related mutations, PLoS Comput Biol 10 (1) (2014) e1003440.

[41] J. Singh, S. A. Rahman, N. Z. Ehtesham, S. Hira, S. E. Hasnain, Sars-cov-2 variants of concern are emerging in india, Nature medicine (2021) 1–3.

[42] R. P. Walensky, H. T. Walke, A. S. Fauci, Sars-cov-2 variants of concern in the united states—challenges and opportunities, Jama 325 (11) (2021) 1037–1038.

[43] J. R. Mascola, B. S. Graham, A. S. Fauci, Sars-cov-2 viral variants—tackling a moving target, Jama 325 (13) (2021) 1261–1262.

[44] C. B. Vogels, M. I. Breban, I. M. Ott, T. Alpert, M. E. Petrone, A. E. Watkins, C. C. Kalinich, R. Earnest, J. E. Rothman, J. Goes de Jesus, et al., Multiplex qpcr discriminates variants of concern to enhance global surveillance of sars-cov-2, PLoS biology 19 (5) (2021) e3001236.

[45] B. A. Johnson, X. Xie, A. L. Bailey, B. Kalveram, K. G. Lokugamage, A. Muruato, J. Zou, X. Zhang, T. Juelich, J. K. Smith, et al., Loss of furin cleavage site attenuates sars-cov-2 pathogenesis, Nature 591 (7849) (2021) 293–299.

[46] T. P. Peacock, D. H. Goldhill, J. Zhou, L. Baillon, R. Frise, O. C. Swann, R. Kugathasan, R. Penn, J. C. Brown, R. Y. Sanchez-David, et al., The furin cleavage site in the sars-cov-2 spike protein is required for transmission in ferrets, Nature Microbiology (2021) 1–11.

[47] S. Xia, Q. Lan, S. Su, X. Wang, W. Xu, Z. Liu, Y. Zhu, Q. Wang, L. Lu, S. Jiang, The role of furin cleavage site in sars-cov-2 spike protein-mediated membrane fusion in the presence or absence of trypsin, Signal transduction and targeted therapy 5 (1) (2020) 1–3.

[48] L. Yurkovetskiy, X. Wang, K. E. Pascal, C. Tomkins-Tinch, T. P. Nyalile, Y. Wang, A. Baum, W. E. Diehl, A. Dauphin, C. Carbone, et al., Structural and functional analysis of the d614g sars-cov-2 spike protein variant, Cell 183 (3) (2020) 739–751.

[49] K. Huang, Y. Zhang, X. Hui, Y. Zhao, W. Gong, T. Wang, S. Zhang, Y. Yang, F. Deng, Q. Zhang, et al., Q493k and q498h substitutions in spike promote adaptation of sars-cov-2 in mice, EBioMedicine 67 (2021) 103381.

[50] R. Gao, W. Zu, Y. Liu, J. Li, Z. Li, Y. Wen, H. Wang, J. Yuan, L. Cheng, S. Zhang, et al., Quasispecies of sars-cov-2 revealed by single nucleotide polymorphisms (snps) analysis, Virulence 12 (1) (2021) 1209–1226.

[51] M. Maurin, F. Fenollar, O. Mediannikov, B. Davoust, C. Devaux, D. Raoult, Current status of putative animal sources of sars-cov-2 infection in humans: Wildlife, domestic animals and pets, Microorganisms 9 (4) (2021) 868.

[52] R. Frutos, J. Serra-Cobo, L. Pinault, M. Lopez Roig, C. A. Devaux, Emergence of bat-related betacoronaviruses: hazard and risks, Frontiers in microbiology 12 (2021) 437.

[53] A. Graudenzi, D. Maspero, F. Angaroni, R. Piazza, D. Ramazzotti, Mutational signatures and heterogeneous host response revealed via large-scale characterization of sars-cov-2 genomic diversity, Iscience 24 (2) (2021) 102116.

[54] R. Frutos, L. Gavotte, C. A. Devaux, Understanding the origin of covid-19 requires to change the paradigm on zoonotic emergence from the spillover model to the viral circulation model, Infection, Genetics and Evolution (2021) 104812.

[55] D. Ramazzotti, F. Angaroni, D. Maspero, C. Gambacorti-Passerini, M. Antoniotti, A. Graudenzi, R. Piazza, Verso: a comprehensive framework for the inference of robust phylogenies and the quantification of intra-host genomic diversity of viral samples, Patterns 2 (3) (2021) 100212.

[56] H. A. Al Khatib, F. M. Benslimane, I. E. Elbashir, P. V. Coyle, M. A. Al Maslamani, A. Al-Khal, A. A. Al Thani, H. M. Yassine, Within-host diversity of sars-cov-2 in covid-19 patients with variable disease severities, Frontiers in cellular and infection microbiology 10 (2020).

[57] J. Pekar, M. Worobey, N. Moshiri, K. Scheffler, J. O. Wertheim, Timing the sars-cov-2 index case in hubei province, Science 372 (6540) (2021) 412–417.

[58] E. Decroly, J.-M. Claverie, B. Canard, Le rapport de la mission oms peine à retracer les origines de l'épidémie de sars-cov-2, Virologie 1 (1) (2017).

[59] A. Maxmen, Who report into covid pandemic origins zeroes in on animal markets, not labs, Nature 592 (7853) (2021) 173–174.