Received: 24 August 2020 | Revised: 19 October 2020 | Accepted: 13 November 2020

DOI: 10.1111/tbed.13931

ORIGINAL ARTICLE

Transboundary and Emerging Diseases WILEY

Global SNP analysis of 11,183 SARS-CoV-2 strains reveals high genetic diversity

Fangfeng Yuan[1] | Liping Wang[2] | Ying Fang[1] | Leyi Wang[3]

[1]Department of Pathobiology, College of Veterinary Medicine, University of Illinois at Urbana Champaign, Urbana, Illinois, USA

[2]Department of Diagnostic Medicine and Pathobiology, College of Veterinary Medicine, Kansas State University, Manhattan, Kansas, USA

[3]Veterinary Diagnostic Laboratory and Department of Veterinary Clinical Medicine, College of Veterinary Medicine, University of Illinois, Urbana, Illinois, USA

Correspondence
Leyi Wang, Veterinary Diagnostic Laboratory and Department of Veterinary Clinical Medicine, College of Veterinary Medicine, University of Illinois, Urbana, IL, USA.
Email: leyiwang@illinois.edu

Abstract

Since first identified in December of 2019, COVID-19 has been quickly spreading to the world in few months and COVID-19 cases are still undergoing rapid surge in most countries worldwide. The causative agent, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), adapts and evolves rapidly in nature. With the availability of 16,092 SARS-CoV-2 full genomes in GISAID as of 13 May, we removed the poor-quality genomes and performed mutational profiling analysis for the remaining 11,183 viral genomes. Global analysis of all sequences identified all single nucleotide polymorphisms (SNPs) across the whole genome and critical SNPs with high mutation frequency that contributes to five-clade classification of global strains. A total of 119 SNPs were found with 74 non-synonymous mutations, 43 synonymous mutations and two mutations in intergenic regions. Analysis of geographic pattern of mutational profiling for the whole genome reveals differences between each continent. A transition mutation from C to T represents the most mutation types across the genome, suggesting rapid evolution and adaptation of the virus in host. Amino acid (AA) deletions and insertions found across the genome results in changes in viral protein length and potential function alteration. Mutational profiling for each gene was analysed, and results show that nucleocapsid gene demonstrates the highest mutational frequency, followed by Nsp2, Nsp3 and Spike gene. We further focused on non-synonymous mutational distributions on four key viral proteins, spike with 75 mutations, RNA-dependent-RNA-polymerase with 41 mutations, 3C-like protease with 22 mutations and Papain-like protease with 10 mutations. Results show that non-synonymous mutations on critical sites of these four proteins pose great challenge for development of anti-viral drugs and other countering measures. Overall, this study provides more understanding of genetic diversity/variability of SARS-CoV-2 and insights for development of anti-viral therapeutics.

KEYWORDS

3CL^{pro}, complete genome sequence, PL^{pro}, RdRp, S protein, SARS-CoV-2, SNP analysis

1 | INTRODUCTION

Coronavirus (CoV) is a single-stranded positive-sense RNA virus in the order *Nidovirales*, family *Coronaviridae*. Based on genetic characterization, CoVs are classified into four genera, α, β, γ and δ (Fehr & Perlman, 2015). There are seven known human coronaviruses with two (229E and NL63) in α genus and five (OC43, HKU1, severe acute respiratory syndrome coronavirus (SARS-CoV), Middle

East Respiratory Syndrome CoV and SARS-CoV-2) in β genus (Liu et al., 2020; Ye et al., 2020). SARS-CoV-2 was reported to be the causative agent for the novel respiratory disease, COVID-19 (Zhu et al., 2020). The disease was declared to be a pandemic by WHO early this year and has led to more than 32 million infected and 981,000 dead. SARS-CoV-2 RNA genome encodes 16 non-structural proteins (Nsp) and at least 10 structural proteins including spike (S), ORF3a, envelop (E), membrane (M), open reading frame 6 (ORF6), ORF7a, ORF7b, ORF8, nucleocapsid (N) and ORF10 (Cagliani et al., 2020; Kim et al., 2020). S protein contains receptor-binding domain (RBD) that directly binds to human receptor angiotensin-converting enzyme 2 (ACE2) and induces neutralizing antibody response against SARS-CoV-2 (Cao et al., 2020; Lan et al., 2020). Previous studies showed that antibody response against SARS-CoV-2 is mainly against S and N proteins (Erasmus et al., 2020; To et al., 2020). RNA viruses possess a high mutation rate of genome and readily adapt to changing environmental conditions (Elena & Sanjuán, 2005). Thus, a swarm of variants exist in RNA virus populations. A systemic tracking of SARS-CoV-2 mutations allows monitoring of circulating strains around the world (Guan et al., 2020) and provides guidance for development of countering measures.

Since the first report of SARS-CoV-2, whole-genome sequences of the virus have been uploaded to the public available website, GISAID. Nextstrain employed nomenclature through designation of SARS-CoV-2 clades to label well-defined clades that reached geographic spread with significant frequency. Major clades were named by the year that emerged and a letter. Current clades on Nextstrain tree include 19A, 19B, 20A, 20B and 20C (Hadfield et al., 2018). Another clade definition in GISAID used genetic markers and defined six clades including S, L, V, G, GH and GR. L was split into G and V in March (Tang et al., 2020). In order to characterize the mutational patterns and distributions across the whole genome, we performed a mega data analysis of 11,183 high-quality sequences from GISAID as of 13 May. Geographical distribution of mutations was analysed, and we further focused on four key viral proteins including S, RNA-dependent-RNA-polymerase (RdRp), 3C-like protease ($3CL^{pro}$) and Papain-like protease (PL^{pro}). Potential functional impacts of mutations were evaluated. This study provides more evidence of SARS-CoV-2 genetic diversity, and mutations on key viral proteins may affect development of anti-viral therapeutics.

2 | METHODS AND METHODS

2.1 | Sequence source and analysis

As of 13 May, there are 16,092 high coverage full genomes available in GISAID (Shu & McCauley, 2017). All were downloaded and of which 4,909 were removed due to their poor assembly quality resulting in 11,183 complete genomes used for subsequent analysis. MAFFT was employed for sequence alignment referenced to Wuhan-hu-1 strain (MN908947.3). Alignment results were further processed and analysed through CLC Genomics Workbench 11 (QIAGEN) and UGene (http://ugene.net). Statistical data analysis was performed on Excel (Microsoft) and GraphPad Prism software (GraphPad Software, Inc.). To determine the viral diversity and credibility of mutations across the genome, the entropy of nucleotide sequences was calculated using BioEdit software version 7.0.9.0 (Hall, 1999). [Correction added on 27 May 2021, after first online publication: In this paragraph, the reference "Shu & McCauley, 2017" has been included at the end of the first sentence in this current version.]

2.2 | Protein structural analysis

Protein structures for RdRp, S and $3CL^{pro}$ were obtained from the Protein Data Bank (PDB). For SARS-CoV-2 PL^{pro} structure, homology modelling was carried out by using I-TASSER (Yang et al., 2015) based on SARS-CoV PL^{pro} structure. Structural homology with highest *C* scores was selected for analysis. Visualization of protein structures was performed through PyMOL (PyMOL Molecular Graphics System, version 1.7; Schrödinger, LLC).

3 | RESULTS

3.1 | Global SNPs across the genome and their geographical distribution

A total of 16,092 complete genomes with high coverage as of 13 May were downloaded from GISAID. After removal of 4,909 problematic sequences using stringent inclusion criteria (any N in the genome), 11,183 sequences were included for analysis. Since a large number of sequences do not have authentic or high-quality sequences for both 5' and 3' un-translational region (Singh et al., 2020), terminal sequences for both ends were removed and only regions (266–29674nt) from polyprotein to the last open reading frame (Bal et al., 2020) sequences were included. Alignment against the reference strain, Wuhan-hu-1 (MN908947.3), was performed using MAFFT (Katoh et al., 2017; Rozewicki et al., 2019). For global sequences analysed, an initial threshold setting of 1% (>111) was made to identify classified clades around the globe (Table 1). A low threshold of 0.3% (>33) was also set to identify a site of interest (Table S1). A 0.3% threshold was also applied to countries/regions with more than 333 sequences, and for those countries/regions with less than 333 sequences, single nucleotide polymorphisms (SNPs) with at least two sequences were recorded.

Globally, with a threshold above 0.3%, we observed a total of 119 SNPs across the genome with 74 non-synonymous mutations, 43 synonymous mutations and two mutations in intergenic region (Table 1 and Table S1). A new major clade can be proposed if it reaches 20% frequency globally. Five major clades (19A, 19B, 20A, 20B and 20C) are classified based on nomenclature data provided by Nextstrain (Figure S1). As shown in Table 1 and File S1, top SNPs with most counts include A23403G in S gene (Clade 19A, Count:

TABLE 1 Global nucleotide and amino acid mutations across the genome for threshold above 1%

Name (Clade)	Position	Count	Gene	Nucleotide change	Amino acid change	Entropy	Name (Clade)	Position	Count	Gene	Nucleotide change	Amino acid change	Entropy
	313	127	NSP1	C/T	Synonymous	0.06307		18877	380	NSP14	C/T	Synonymous	0.14832
	490	132	NSP1	T/A	D/E	0.06413		18998	133	NSP14	C/T	A/V	0.06453
	514	112	NSP1	T/C	Synonymous	0.06019		19839	112	NSP15	T/C	Synonymous	0.05607
A (20C)	1059	2679	NSP2	C/T	T/I	0.55524		20268	576	NSP15	A/G	Synonymous	0.20476
	1397	140	NSP2	G/A	V/I	0.06728	J (19A)	23403	7590	Spike	A/G	D/G	0.63444
	1440	164	NSP2	G/A	G/D	0.07648		23731	142	Spike	C/T	synonymous	0.07145
	2416	188	NSP2	C/T	Synonymous	0.08535		23929	125	Spike	C/T	Synonymous	0.06319
	2480	258	NSP2	A/G	I/V	0.11068		24034	150	Spike	C/T	Synonymous	0.07208
	2558	282	NSP2	C/T	P/S	0.11862		25429	164	ORF3a	G/T	V/L	0.07820
	2891	152	NSP3	G/A	A/T	0.07873	K (20C)	25563	3276	ORF3a	G/T	Q/H	0.61047
B (19A)	3037	7552	NSP3	C/T	Synonymous	0.63571	L (19A)	26144	772	ORF3a	G/T	G/V	0.25430
	3177	134	NSP3	C/T	P/L	0.06664		26530	120	M	A/G	D/G	0.06549
	6312	128	NSP3	C/A	T/K	0.06766		26729	138	M	T/C	Synonymous	0.06650
C (19B)	8782	1480	NSP4	C/T	Synonymous	0.39487		26735	149	M	C/T	Synonymous	0.07169
	10097	138	3CLpro	G/A	G/S	0.06897		27046	260	M	C/T	T/M	0.11227
D (19A)	11083	1161	NSP6	G/T	L/F	0.35912		27964	251	ORF8	C/T	S/L	0.10741
	11916	179	NSP7	C/T	S/L	0.08298		28077	144	ORF8	G/	V/L	0.08078
	13730	147	RdRp	C/T	A/ L	0.07000	*M* (19A)	28144	1476	ORF8	T/C	L/ S	0.39349
E (20A)	14408	7564	RdRp	C/ T	P/ L	0.63513		28311	149	N	C/T	P/ L	0.07730
F	14805	816	RdRp	C/ T	Synonymous	0.26613		28688	136	N	T/C	Synonymous	0.06756
	15324	292	RdRp	C/ T	Synonymous	0.12187		28854	243	N	C/T	S/ L	0.03369
	17247	317	NSP13	T/ C	Synonymous	0.13213	O (20B)	28881	2046	N	G/A	R/ K	0.48311
G	17747	928	NSP13	C/ T	P/ L	0.28916	P (20B)	28882	2041	N	G/A	synonymous	0.48170
H	17858	946	NSP13	A/ G	Y/ C	0.28985	Q (20B)	28883	2040	N	G/C	G/ R	0.47915
I	18060	956	NSP14	C/ T	Synonymous	0.29368		29540	133	Unknown	G/T	NA	0.06910
	18736	128	NSP14	T/ C	F/ L	0.06254		2553	211	Unknown	G/A	NA	0.09607

7,590, entropy: 0.63444), C14408T in RdRp (Clade 20A, Count: 7,564, entropy: 0.63513), C3037T in NSP3 (Clade 19A, Count: 7,552, entropy: 0.63571), G25563T in ORF3a (Clade 20C, Count: 3,276, entropy: 0.61047), C1059T in NSP2 (Clade 20C, Count: 2,679, entropy: 0.55524), G28881A (Clade 20B, Count: 2046, entropy: 0.48311), G28883C (Clade 20B, Count: 2040, entropy: 0.47915) and G28882A in N gene (Clade 20B, Count 2041, entropy: 0.4817). Clade 19B contains C8782T (Count: 1,480, entropy: 0.39487). Higher entropy value represents the mutational change in more sequences (Saha et al., 2020), and the pattern of entropy was found to be consistent with that of the SNP count (Figure S3, File S2). Another important SNP, C241T, was not included in this analysis. Different clades based on marker variants can also be defined according to GISAID. Clade 20A contains G clade (C241T, C3037T and A23403G), clade 20B contains GR clade (C241T, C3037T, A23403G and G28882A), clade 20C contains GH clade (C241T, C3037T, A23403G and G25563T), clade 19B contains S clade (C8782T and T28144C), and clade 19A contains V clade (G11083T and G26144T) (Hadfield et al., 2018; Rambaut et al., 2020). All clade classification criteria can be informed by statistical distribution of genome distances in phylogenetic clusters (Han et al., 2019). Mutations with high frequency found here contribute to the clade classification.

Among all 119 SNPs across the genome, there are 60 positions with nucleotide substitutions from C to T, accounting for half of SNPs (Figure 1). It has been reported that transition mutations are much more common than transversion mutations in viruses (Caudill et al., 2019). With most positions possessing C to T mutation, CpG sites decreased. The zinc-finger anti-viral protein binds specifically to CpG for degradation of viral RNA genomes. Researchers found

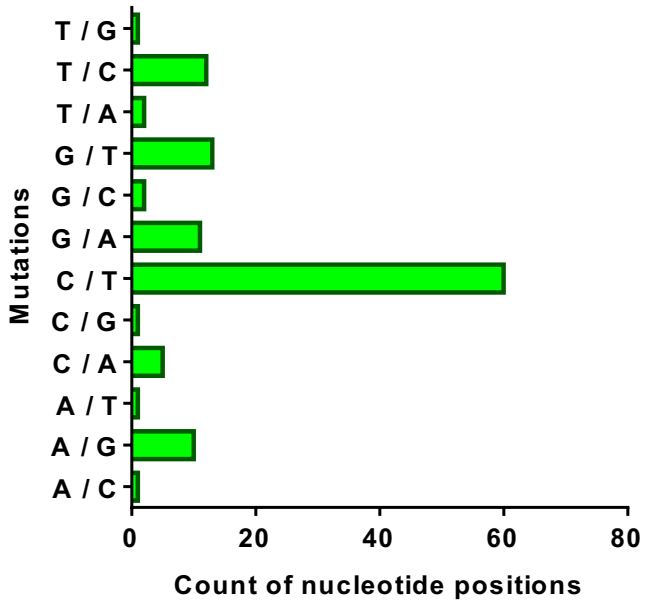

FIGURE 1 Global mutation types across the genome. A total of 119 nucleotide substitutions were analysed by its mutation type. *Y*-axis denotes the type of substitution while the *x*-axis represents the count of each mutation type. C to T (U) mutation represents the majority of nucleotide substitution type

that SARS-CoV-2 has the most extreme CpG deficiency in all known betacoronavirus genomes, indicating viral rapid evolution in the host (di Gioacchino et al., 2020; Xia, 2020). High-frequent C to T mutation found in this study further demonstrates CpG deficiency and SARS-CoV-2 has adapted to new host with high zinc-finger anti-viral protein expression and evolved new ways for immune evasion. More than a third of SNPs across the genome are synonymous mutations (43), and among all non-synonymous mutation sites, 9 were mutated from T to I, 6 from A to V and 6 also from S to L (Data not shown). Although synonymous mutation does not result in change in amino acid sequence, accumulation of these mutations has the capability to erase the characteristic compactness imprint of the single-stranded viral RNA genomes (Tubiana et al., 2015). We also summarized SNPs in each gene of the viral genome. As shown in Figure 2, N gene has 15 nucleotide positions mutated, then nsp2 (13), nsp3 (13), S gene (10), nsp14 (8), nsp12(7), ORF3a (7), nsp13 (6) and nsp5 (5).

To illustrate SNPs landscapes in each country/region, we further did analysis on countries/regions with the number of sequences above 40. Among 11,183 sequences around the globe, 2 North American countries include USA (3,599) and Canada (120); 16 European countries including UK (3,077), Iceland (405), Netherland (401), Denmark (350), Belgium (334), France (274), Austria (224), Spain (181), Russia (139), Germany (109), Sweden (104), Luxembourg (96), Portugal (95), Greece (64), Switzerland (55) and Italy (44); 7 Asia countries/regions including China (294), India (141), Saudi Arabia (127), Singapore (124), Japan (105), Taiwan (80) and Thailand (53); 1 South American country (Brazil, 40); and 1 Oceania country (Australia, 493). The remaining 55 sequences represent the rest of the world. Figure 3a (File S1) demonstrates a landscape comparison between globe and Asia countries/regions. With an exception of China, all other Asia countries/regions displayed a relatively higher mutation frequency across the viral genome, representing potential viral adaptation to hosts. Compared to globe and all other Asia countries/regions, variants from China exhibit much lower SNPs frequencies in terms of B (C3037T), E (C14408T), J (A23403G) and K (G25563T). Instead, SNPs frequencies in China regarding positions in C (C8782T) and M (T28144C) are obviously much higher, which is different from the rest of world that have SNP pattern featuring A23403G (aa: D614G) mutation. It reveals that D614G, which barely exist in China strains, gained more replicative advantages when the virus spread outside of China to the world. Reports from WHO have shown that the new COVID-19 outbreak in Beijing, China exhibits sequence identities more closely to European strains with D614G mutation. For the three major dominant SNPs (Chen et al., 2020; Hillen et al., 2020; Lan et al., 2020; Mercurio et al., 2020; Walls et al., 2020), B (C3037T) and J (A23403G) contribute to 19A clade, and E (C14408T) contributes to 20A clade (Table 1). SNPs G (C17747T), H (A17858G) and I (C18060T) were found predominantly in variants from USA, Canada and Australia (Figure 3b). Interestingly, sequences from Brazil and all European countries displayed an apparently low SNPs frequencies except the three major markers, B (C3037T), J (A23403G) and E (C14408T) (Figure 3b,c, File S1). In other words, SARS-CoV-2 is relatively more stable in these countries. Thus, mutational patterns

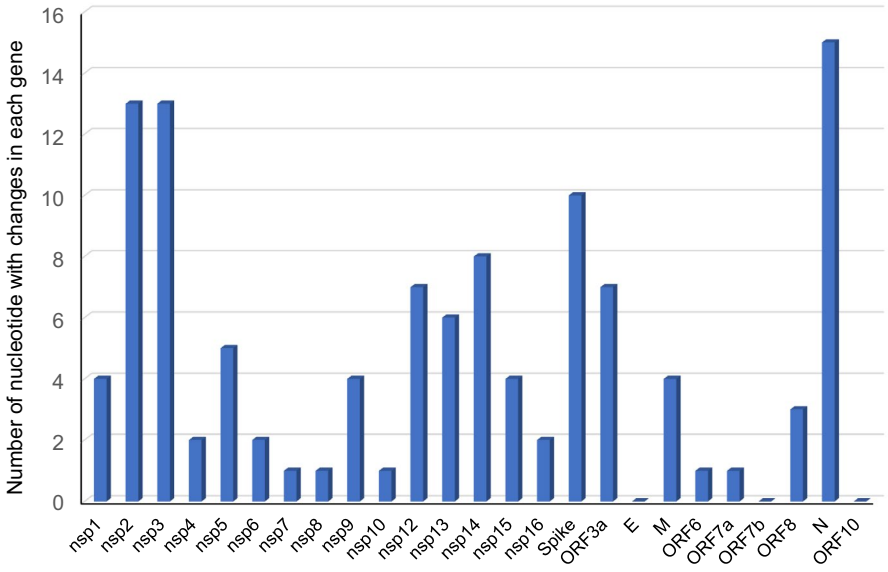

FIGURE 2 Count of nucleotide positions with mutation in each gene. Total global mutations were grouped for each coding gene including sixteen non-structural protein genes and ten structural protein genes. The *x*-axis shows the name of each gene, and *y*-axis indicates the number of nucleotide positions that have substitutions

of SARS-CoV-2 in different regions differ from each other. In order to check the number of SNPs across the genome for different countries, we analysed those countries with more than 333 sequences. We chose threshold of 333 or 0.3% because countries below this threshold theoretically have only one count if SNP was observed. To reduce the inaccuracy, at least two counts should be recorded for a deemed mutation. As is shown in Figure 4, there are 119 SNPs across the genome globally, and Australia sequences contain the most SNPs of 201, while Denmark sequences only have 78 SNPs. Other countries have SNPs of 116 for USA, 132 for UK, 122 for Iceland, 115 for Netherland and 114 for Belgium. In addition, case-fatality rate of selected countries/regions on 13 May has varied a lot from below 1% in Russia, Saudi Arabia, Iceland, Singapore to 19.25% in France, 16.29% in Belgium, 14.44% in UK, etc. (Figure S2). We are trying to find a genetic determinant causing different case-fatality rates among different countries, but we did not find one. According to CDC report, clinical outcomes of COVID-19 patients relate to a variety of factors, such as age, gender, poverty, medical conditions and even blood types (Ellinghaus et al., 2020; Li et al., 2020).

3.2 | Analysis of mutations affecting protein synthesis

Genetic variation/SNPs contribute to alterations of protein translation. We observed multiple deletions and insertions across the genome in different countries/regions (Table 2). Three nucleotide deletion in 1605–1607nt region result in amino acid N deletion in position 267 of nsp2. Twenty-nine counts of ninenucleotides deletion in 686–694nt lead to three amino acids deletions in nsp1 region. Another 9nt deletion (515–520nt) also occurs in nsp1 region, resulting in two amino acids (72V, 73M) missing. Deletion was also found in S gene with thre nucleotides deletion in 21991–21993nt. Accordingly, the single amino acid (Y) was missed in position 144 of S protein. In addition, insertion was found in nsp6. Three consecutive T insertion result in an extra amino acid (F) synthesized. All these deletions/insertions show a globally distributed pattern.

Non-synonymous mutations sometimes result in immediate stop of translation and thus protein truncation. As is shown in Table 3, SNP A12050T in two Denmark strains leads to amino acid change from K to stop codon and a 14aa truncation of nsp7. Forty-nine Belgium strains and 2 Denmark strains have T13402G mutation resulting in 14aa truncation of nsp10. Another SNP T13408A in nsp10 truncated 12aa. A much shorter length of nsp13 (217aa versus 601aa) was generated due to a A16888T mutation in three Denmark strains. A19513T in two Denmark strains results in 36aa truncated in nsp14 C terminal. For structural proteins, two Iceland strains have SNP C27661T resulting in 32aa shorter compared to the original one. Finally, three strains from China have G28041T mutation in ORF8 and also end up with a 72aa deletion in its C terminal. Instead of a change in stop codon, two Germany strains have start codon changed with G25395T and four amino acids are missed in ORF3a. In addition, SNPs in transcriptional regulatory sequence (TRS) may lead to impairment of 3' end structural protein synthesis. Change of protein length could potentially damage its key function in viral replication/assembly/immune system antagonism. However, these may represent quasispecies of SARS-CoV-2 and with those critical mutations, the virus may not get replication advantages. Therefore, further studies are needed for exploring the role of those mutations in the virus replication.

3.3 | Mutations on key viral proteins

S: SARS-CoV-2 S protein is a major target of neutralizing antibodies and contributes to ACE2 binding and entry into host cells. SNPs on S gene potentially impact protein antigenicity and cellular tropism. In this study, there are total 75 non-synonymous mutations found on Spike protein (Table 4), spanning from signal peptide (SP) to cytoplasmic domain (CP). C21575T (L5F) mutation with 70 counts of

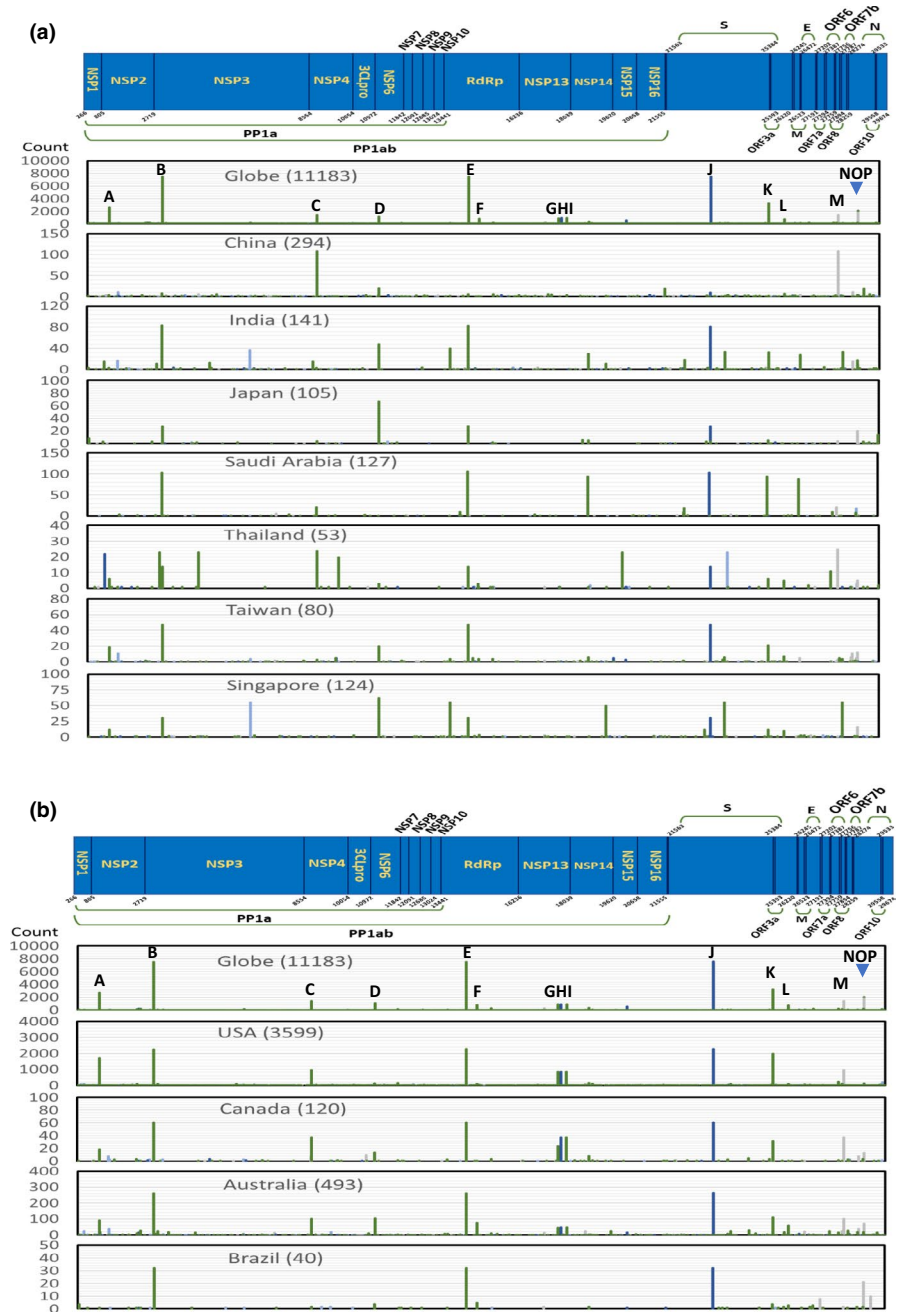

FIGURE 3 Landscape of mutations across the genome for (a) Asian countries, (b) North America, South America, and Oceania countries, and (c) European countries. Sequences up to 13 May were aligned and analysed by UGENE software. Referred to Wuhan-Hu-1 parental strain, each SNP across the genome was recorded. Mutational profiling of whole genome was analysed for both global strains and strains of a specific country/region. Country/region name was indicated with a total number of viral genomes in brackets. A schematic diagram is shown on the top of each figure. Alphabetical letters from A to P indicate corresponding mutations described in Table 1. The y-axis represents counts of strains on each mutation; x-axis denotes the whole-genome landscape of SARS-CoV-2

multiple countries lies in signal peptide region. This SNP was also recorded in Table S1 using a threshold above 0.3% globally. Signal peptides function to translocate spike protein to the membrane. It remains to be determined whether L5F mutation affects S protein translocation or not. A series of mutations with few counts in multiple countries was found in N terminal domain (NTD) of S protein. There are five SNPs found in receptor-binding domain (RBD), among which V483A with 21 counts in USA only, N439K with 31 counts in UK only locate in receptor-binding motif (RBM) and the rest of 3 SNPs (A344S with two counts in Saudi Arabia, N354D with two counts in China, V367F with eight counts in France and Netherland) locate in RBD. The well-known D614G mutation lies in C terminal domain (CTD) of S1 and is close to S2. It has 7,544 counts with a geographic distribution of 27 countries. An increasing trend

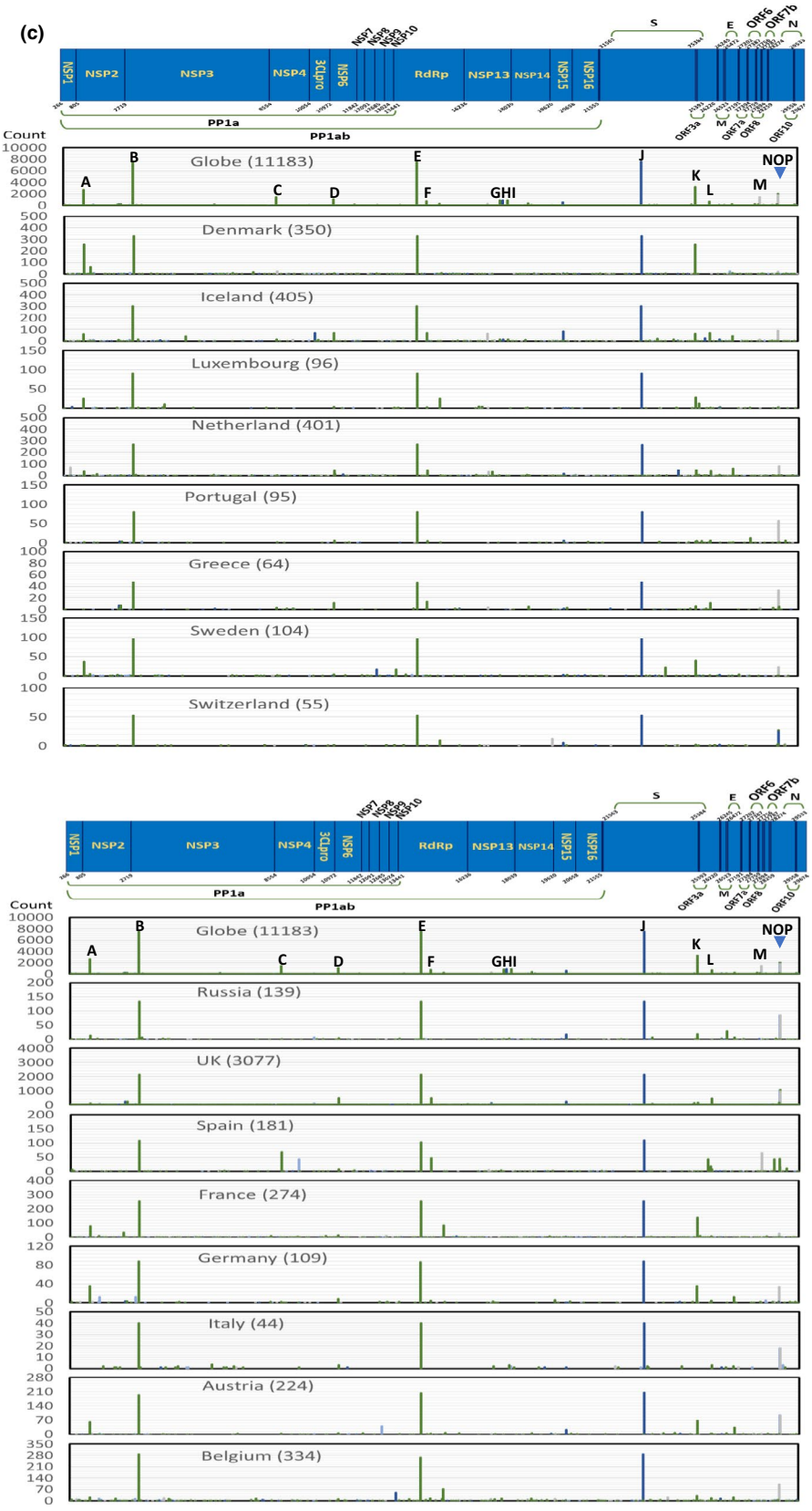

FIGURE 3 (Continued)

of D614G was observed globally, and it was reported that strains with this mutation lead to reduced S1 shedding and increased viral infectivity (Zhang et al., 2020). G614 became the global dominant variant and provided a boost of transmission ability of the virus since outbreaks out of China. However, its impact on therapeutic and vaccine design is limited (Korber et al., 2020). Instead of presence

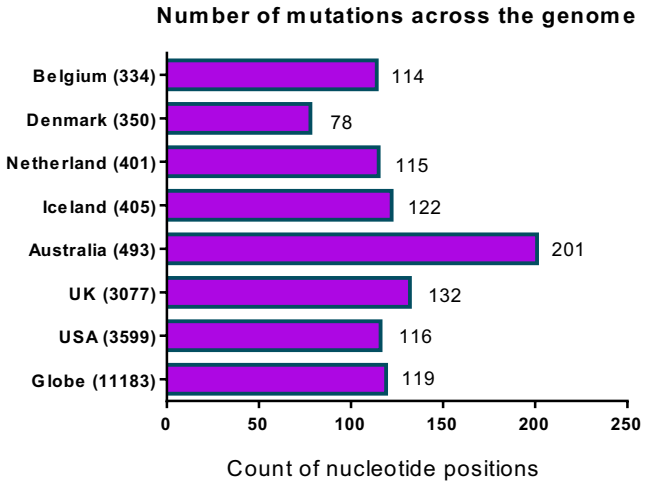

FIGURE 4 Count of the number of nucleotide positions across the genome for countries with more than 300 sequences. The total SNPs across the whole genome were analysed for each country. Y-axis shows countries with more than 300 sequences including Belgium, Denmark, Netherland, Iceland, Australia, UK, USA and the whole world as comparison; x-axis shows the number of nucleotide substitutions across the genome

in the receptor-binding domain (RBD), D614G is located in the interface between the spike protomers and was proposed to cause loss of hydrogen bonds between protomers, thus altering virus infectivity. Antibodies from D614 variant infected patients could cross-neutralize G614 variant, indicating changes in this position have no impacts on antibody-mediated B cell immunity (Grubaugh et al., 2020; Hu et al., 2020; Ozono et al., 2020). Beyond D614G, there are 8 SNPs with few counts in multiple countries located in CTD, followed by 5 SNPs before fusion peptide (FP) in S2 subunit. Taiwan region has six counts of sequences with T791I mutation in FP region. Twenty-three counts of A829T mutation were found only in Thailand. Also, A831V was found only in Iceland samples with 24 counts. D839Y was found in three countries with 11 counts of sequences. Heptad repeat 1 (HR1) and heptad repeat 2 (HR2) interact with each other to form six-helical bundle and facilitate cellular and viral membrane fusion. Six SNPs (D936V, D936Y, S940F, T941A, S943R and S943T) in HR1 and 2 in HR2 (D1163G and V1176F) were found. Interestingly, D936Y was found in 4 countries with total 73 counts of sequences, and S943R (22 counts) and S943T (23 counts) were found only in Belgium samples. Because of the special function in membrane fusion, researchers have been developing potent

TABLE 2 Deletions and insertions found across the whole genome

Type	Nucleotide position in whole genome	Amino acid	Gene	Average Entropy	Total counts	Geographic distribution
Deletion	1605–1607	267 N	NSP2	0.12121	282	UK (153), Netherland (80), Australia (9), Belgium (10), Denmark (4), Iceland (7), USA (4), Portugal (3), France (2), Spain (2), Canada (1), Finland (1), New Zealand (1), Russia (1), Sweden (1), Taiwan (1), Latvia (1)
Deletion	686–694	129 KSF 131	NSP1	0.01895	29	USA (16), UK (8), Swedan (1), Iceland (1), Saudi Arabia (1), France (1), Canada (1)
Deletion	515–520	72 VM 73	NSP1	0.01402	22	USA (13), Australia (3), UK (2), Denmark (1), France (1), Greece (1), Netherland (1)
Deletion	21991–21993	144 Y	spike	0.00779	11	USA (3), Slovenia (2), Saudi Arabia (2), Netherland (2), India (1), Belgium (1)
Insertion	TTT inserted between 11,074 and 11075nt	35 F inserted	nsp6	0.00654	10	Australia (5), England (4), Switzerland (1)

TABLE 3 Key mutations relating to protein expression change

Country	Position	Count	Gene/ region	Entropy	Nucleotide change	Amino acid change	Length Wuhan-hu–1	Length after mutation
Denmark	12050	2	NSP7	0.00320	A/T	K/Stop codon	249nt/83aa	210nt/69aa
Belgium	13402	49	NSP10	0.03333	T/G	Y/Stop codon	417nt/139aa	378nt/125aa
Denmark	13402	2	NSP10	0.03333	T/G	Y/Stop codon	417nt/139aa	378nt/125aa
Belgium	13408	2	NSP10	0.01049	T/A	C/Stop codon	417nt/139aa	384nt/127aa
Denmark	16888	3	NSP13	0.00247	A/T	K/Stop codon	1803nt/601aa	654nt/217aa
Denmark	19513	2	NSP14	0.00092	A/T	R/stop codon	1581nt/527aa	1476nt/491aa
Germany	25395	2	ORF3a	0.00172	G/T	Start codon changed	828nt/275aa	816nt/271aa
Iceland	27661	2	ORF7a	0.00172	C/T	Q/stop codon	366nt/121aa	270nt/89aa
Austria	27393	2	TRS	0.00172	C/T	NA	acgaac	acgaat
France	27893	2	TRS	0.00172	C/T	NA	acgaac	acgaat
China	28041	3	ORF8	0.00247	G/T	G/ Stop codon	366nt/ 121aa	150nt/49aa

TABLE 4 Summarized mutations within spike gene

Nucleotide		Amino acid				Geographic distribution	Region	Nucleotide		Amino acid				Geographic distribution	Region
Position	Change	Position	Change	Count	Entropy			Position	Change	Position	Change	Count	Entropy		
21575	C/ T	5	L/ F	70	0.04689	6 countries	SP	23587	G/T	675	Q/ H	2	0.00952	Iceland	CTD
21614	C/ T	18	L/ F	16	0.01346	UK	NTD	23673	C/T	704	S/ L	2	0.00320	Australia	S2
21648	C/ T	29	T/ I	4	0.00962	Australia, Netherland	NTD	23679	C/T	706	A/ V	5	0.00590	Belgium	S2
21707	C/ T	49	H/ Y	19	0.01780	China	NTD	23732	A/T	724	T/ S	3	0.00000	Denmark	S2
21711	C/ T	50	S/ L	4	0.00458	Australia	NTD	23856	G/A	765	R/ H	4	0.00939	Iceland	S2
21724	G/ T	54	L/ F	4	0.00872	France, India	NTD	23856	G/T	765	R/ L	3	0.00939	Taiwan	S2
21743	A/ T	61	N/ Y	4	0.00000	Denmark	NTD	23934	C/T	791	T/ I	6	0.00590	Taiwan	FP
21846	C/T	95	T/I	2	0.00617	Austria	NTD	24047	G/A	829	A/ T	23	0.01478	Thailand	S2
21855	C/T	98	S/F	2	0.00524	Iceland	NTD	24054	C/T	831	A/ V	24	0.01696	Iceland	S2
21920	G/A	120	V/I	4	0.00412	France	NTD	24077	G/T	839	D/ Y	11	0.01288	3 countries	S2
21974	G/C	138	D/H	3	0.00797	Australia	NTD	24095	G/T	845	A/ S	3	0.00617	Portugal	S2
22020	T/C	153	M/T	2	0.00247	China	NTD	24099	C/T	846	A/ V	6	0.00779	Belgium	S2
22032	T/C	157	F/S	2	0.00172	Canada	NTD	24102	G/C	847	R/ T	2	0.00265	Austria	S2
22103	G/C	181	G/A	3	0.00247	Denmark	NTD	24117	C/T	852	A/ V	3	0.00247	Netherland	S2
22151	A/G	197	I/V	4	0.00390	Greece, Spain	NTD	24197	G/T	879	A/ S	6	0.01459	Austria	S2
22205	G/C	215	D/H	2	0.00567	Saudi Arabia	NTD	24369	A/T	936	D/ V	4	0.00000	Denmark	HR1
22224	C/T	221	S/L	4	0.00412	Australia	NTD	24368	G/T	936	D/ Y	73	0.04120	4 countries	HR1
22277	C/A	239	Q/K	5	0.00590	Netherland	NTD	24381	C/T	940	S/ F	2	0.00247	France	HR1
22289	G/T	243	A/S	2	0.00412	India	NTD	24383	A/G	941	T/ A	2	0.00172	Belgium	HR1
22303	T/G	247	S/R	2	0.00172	China	NTD	24389	A/C	943	S/ R	22	0.03317	Belgium	HR1
22323	C/T	254	S/F	2	0.00630	Belgium	NTD	24390	G/C	943	S/ T	23	0.03528	Belgium	HR1
22344	G/T	261	G/V	3	0.00482	Netherland	NTD	24621	C/T	1020	A/ V	4	0.00524	Austria	S2
22346	G/A	262	A/T	4	0.00412	Australia	NTD	24642	C/T	1027	T/ I	8	0.00654	Austria	S2
22374	A/G	271	Q/R	2	0.00344	India	NTD	24680	G/T	1040	V/ F	3	0.00247	China	S2
22404	A/T	281	E/V	2	0.00587	Taiwan	NTD	24794	G/T	1078	A/ S	6	0.00809	Belgium, Luxembourg	S2
22592	G/T	344	A/S	2	0.00172	Saudi Arabia	RBD	24812	G/T	1084	D/ Y	3	0.00247	China	S2
22622	A/G	354	N/D	2	0.00265	China	RBD	24863	C/G	1101	H/ D	3	0.00524	Denmark	S2
22661	G/T	367	V/F	8	0.00902	France, Netherland	RBD	24933	G/T	1124	G/ V	15	0.01638	Australia, India	S2
22879	C/A	439	N/K	31	0.00524	UK	RBD (RBM)	25050	A/G	1163	D/ G	2	0.00172	Australia	HR2

(Continues)

TABLE 4 (Continued)

Nucleotide		Amino acid				Geographic distribution	Region	Nucleotide		Amino acid				Geographic distribution	Region
Position	Change	Position	Change	Count	Entropy			Position	Change	Position	Change	Count	Entropy		
23010	T/C	483	V/A	21	0.01366	USA	RBD (RBM)	25088	G/T	1176	V/ F	4	0.00458	China, Denmark	HR2
23271	C/T	570	A/V	3	0.00504	China	CTD	25218	G/T	1219	G/ V	3	0.00524	France	TM
23277	C/T	572	T/I	3	0.00524	India	CTD	25249	G/T	1229	M/ I	4	0.00320	Belgium, Iceland	TM
23311	G/T	583	E/D	3	0.00590	India	CTD	25269	G/T	1236	C/ F	2	0.00247	Austria	TM
23393	C/T	611	L/F	4	0.00320	Belgium	CTD	25273	G/T	1237	M/ I	2	0.00492	Portugal	TM
23401	G/T	613	Q/M	3	0.00320	Japan	CTD	25290	G/T	1243	C/ F	2	0.00247	India	CP
23403	A/G	614	D/G	7,544	0.63444	27 countries	CTD	25340	G/A	1260	D/ N	3	0.00320	Australia	CP
23576	G/T	672	A/S	2	0.00000	Denmark	CTD	25350	C/T	1263	P/ L	56	0.03799	4 countries	CP
23588	A/C	676	T/P	2	0.00092	Denmark	CTD								

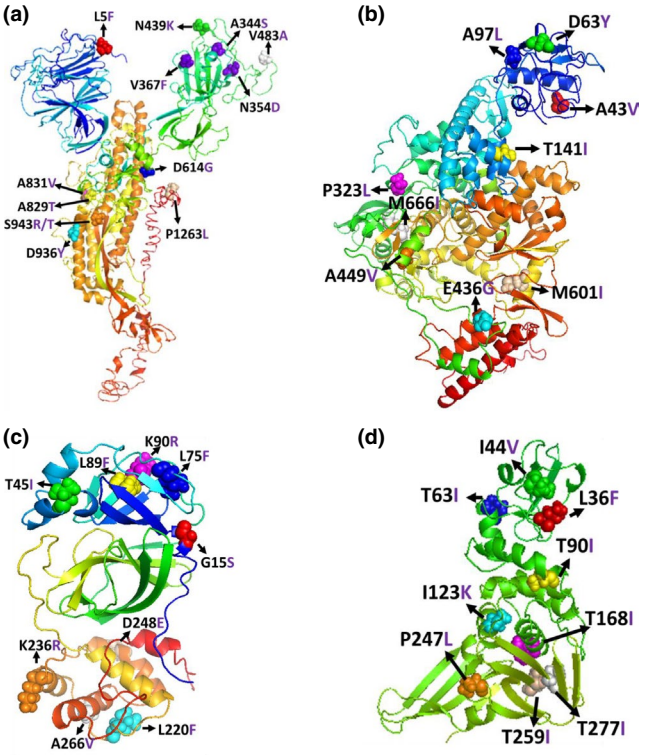

FIGURE 5 Non-synonymous mutations on (a) spike, (b) RdRp, (c) $3CL^{pro}$ and (d) PL^{pro}. Protein structures for RdRp, S and $3CL^{pro}$ were obtained from the Protein Data Bank (PDB) accession 6M71, 6vyb and 6M2Q, respectively. Homology modelling of SARS-CoV-2 PL^{pro} structure was carried out by using I-TASSER (Yang et al., 2015) based on SARS-CoV PL^{pro} structure. PyMOL was used for visualization of protein structure. Sphere with different colours including red, green, blue, yellow, magentas, cyans, oranges, tints and greys indicates corresponding non-sysnonymous mutations on each protein. Amino acid mutations were also coloured in blue after the position number

fusion inhibitors targeting HR1/HR2 of SARS-CoV and MERS-CoV (Xia et al., 2020). Mutations found in these two regions may potentially affect efficacy of fusion inhibitors. Followed by HR2, four and three non-synonymous SNPs were found in transmembrane domain (TM) and cytoplasmic domain (CP), respectively. Notably, P1263L mutation in CP region has 56 counts of sequences from multiple countries. Critical mutations on RBD and mutations with top counts were also denoted through structural analysis (Figure 5a). No mutations were found on the N-linked glycosylation sites, key amino acids for ACE2 binding and SPRRAR↓SV cleavage sites in S protein. Highly genetic variation and diversity observed in S protein poses potential challenge to anti-viral vaccine and therapeutics development. Further studies are needed to determine the functional impacts of key S mutations found in this study.

RdRp: The core component of replication-transcription complex is the catalytic subunit, RdRp (nsp12). In this study, multiple non-synonymous SNPs were found in all regions of RdRp, such as beta-hairpin (2 SNPs), nidovirus RdRp-associated nucleotidyltransferase domain (NiRAN) (4 SNPs), interface domain (8 SNPs), fingers (10

TABLE 5 Summarized non-synonymous mutations within RdRp gene

Nucleotide		Amino acid				Geographic distribution	Region	Nucleotide		Amino acid				Geographic distribution	Region
Position	change	Position	Change	Count	Entropy			position	CHANGE	Position	Change	Count	Entropy		
13517	C/T	26	T/I	6	0.01288	France, Luxembourg		14653	G/T	405	V/F	2	0.00495	Austria	Fingers
13537	A/T	33	R/W	3	0.00000	Denmark	Beta-hairpin	14741	C/T	434	S/F	11	0.00779	UK	Fingers
13568	C/T	43	A/V	17	0.01422	Sweden	Beta-hairpin	14747	A/G	436	E/G	20	0.01310	USA	Fingers
13617	G/T	59	K/N	2	0.00172	Spain		14786	C/T	449	A/V	58	0.03246	6 countries	Fingers
13627	G/T	63	D/Y	21	0.01366	Australia, UK		14912	A/G	491	N/S	12	0.00841	USA	Fingers
13730	C/T	97	A/L	124	0.07000	7 countries		14980	C/T	514	L/F	2	0.00247	Iceland	Fingers
13862	C/T	141	T/I	63	0.03664	3 countries	NiRAN	15243	G/T	601	M/I	18	0.01366	USA	Palm
14109	A/G	223	I/M	2	0.00172	Saudi Arabia	NiRAN	15277	C/T	613	H/Y	12	0.01196	Belgium	Palm
14122	G/T	228	G/C	11	0.00902	Saudi Arabia	NiRAN	15327	G/T	629	M/I	2	0.00550	Spain	Fingers
14183	C/T	248	T/I	4	0.00550	Australia	NiRAN	15380	G/T	647	S/I	10	0.01873	Austria, Germany	Fingers
14195	C/A	252	T/N	3	0.00320	Sweden	Interface	15438	G/T	666	M/I	19	0.01253	UK	Fingers
14198	C/T	253	A/V	2	0.00247	China	Interface	15535	G/T	699	A/S	2	0.00172	Netherland	Palm
14267	C/T	276	T/M	2	0.00247	Greece	Interface	15647	A/G	736	D/G	4	0.00320	China	Palm
14270	A/C	277	E/A	2	0.00172	Belgium	Interface	15760	G/A	774	G/S	16	0.01196	USA	Palm
14274	G/C	278	E/D	2	0.00437	India	Interface	15850	G/T	804	D/Y	2	0.00172	Iceland	Palm
14408	C/T	323	P/L	7,517	0.63513	27 countries	Interface	15860	A/G	807	K/R	9	0.00654	France	Palm
14425	C/A	329	L/I	3	0.00247	India	Interface	15919	G/T	827	V/L	2	0.00172	Greece	Thumb
14511	G/T	357	Q/H	2	0.00172	Russia	Interface	15982	G/T	848	V/L	2	0.00247	Australia	Thumb
14585	C/T	382	A/V	5	0.00458	Taiwan		16054	C/T	872	H/Y	7	0.00524	Iceland	Thumb
14,593	G/A	385	G/S	2	0.00265	Russia		16078	G/A	880	V/I	4	0.00320	India	Thumb
14,636	C/T	399	A/V	2	0.00458	Singapore	Fingers								

TABLE 6 Summarized non-synonymous mutations within 3CLpro gene

Nucleotide		Amino acid				
Position	Change	Position	Change	Count	Entropy	Geographic distribution
10097	G/A	15	G/S	138	0.06897	Austria, Denmark, Iceland, Netherland, Russia, UK
10188	C/T	45	T/I	17	0.01138	USA
10208	C/T	52	P/S	2	0.00172	Russia
10265	G/A	71	G/S	2	0.00962	Denmark
10277	C/T	75	L/F	15	0.01080	USA
10319	C/T	89	L/F	32	0.02106	USA
10323	A/G	90	K/R	76	0.04422	China, Iceland
10376	C/T	108	P/S	12	0.00779	Iceland, UK
10377	C/T	108	P/L	2	0.00482	France
10449	C/T	132	P/L	2	0.00340	Russia
10478	A/C	142	N/H	2	0.00172	India
10479	A/T	142	N/I	2	0.00265	India
10508	A/G	152	I/V	2	0.00172	China
10604	C/T	184	P/S	4	0.00458	China
10631	G/A	193	A/T	3	0.00247	Australia
10641	C/T	196	T/M	2	0.00172	Iceland
10712	C/T	220	L/F	22	0.01422	USA
10761	A/G	236	K/R	13	0.00902	USA
10798	C/A	248	D/E	43	0.02522	UK
10851	C/T	266	A/V	35	0.02169	Australia, USA
10874	A/G	274	N/D	13	0.01113	UK
10889	C/T	279	R/C	3	0.00390	Australia

TABLE 7 Summarized non-synonymous mutations within PLP region

Nucleotide		Amino acid				
Position	Change	Position	Change	Count	Entropy	Geographic distribution
5062	G/T	36	L/F	7	0.01366	China
5084	A/G	44	I/V	4	0.00524	Canada
5142	C/T	63	T/I	40	0.02372	Iceland
5223	C/T	90	T/I	2	0.00172	Australia
5322	T/A	123	I/K	2	0.00172	Belgium
5457	C/T	168	T/I	2	0.00172	Luxembourg
5694	C/T	247	P/L	2	0.00265	Belgium
5730	C/T	259	T/I	2	0.00962	Iceland
5784	C/T	277	T/I	15	0.01138	USA
5845	A/T	297	K/N	2	0.00247	Japan

SNPs), palm (7 SNPs) and thumb (4 SNPs) (Table 5). Notably, SNP C14408T (P323L) with 7,517 counts locates in interface domain and was distributed in 27 countries globally. Interface domain is still poorly studied and presumably interacts with other proteins regulating catalytic activity of RdRp. In most cases, spike D614G was accompanied by RdRp P323L. Structural analysis shows that P323L mutation results in considerable changes in secondary structure at this site and the substitution from proline to leucine could cause damage of structural integrity conferred by proline (Figure 5). Similarly, substitution of valine with a larger side chain at position 97 changes secondary structure of RdRp. It has been reported that A97V and P323L result in alteration of protein stability and intramolecular interactions, thus affecting RdRp functions (Chand et al., 2020). Studies have put more efforts on spike D614G impacts,

whereas RdRp P323L may also play a role in viral genome replication and transcription. Other more frequent mutations include A97L with 124 counts in 7 countries, T141I in NiRAN domain with 63 sequence counts in three countries, A449V in fingers with 58 counts in 6 countries around the world. Some mutations only exist in specific countries, such as A43V in beta-hairpin domain with 17 counts in Sweden only, E436G (fingers, 20 counts) and M601I (Palm, 18 counts) and G774S (Palm, 16 counts) in USA only, G228C (NiRAN, 11 counts) in Saudi Arabia only, S434F (Fingers, 11 counts) and M666I (Fingers, 19 counts) in UK only. Mutations with top counts were also shown on RdRp structure (Figure 5b). RdRp has been proposed to be the target of many anti-viral drugs with nucleotide analogs. So many SNPs in RdRp, especially the high-frequency mutation P323L could potentially reduce effectiveness of anti-viral treatments. Due to the participation of RdRp in viral genome transcription, mutations such as P323L may potentially affect viral replicative ability and transmission. In addition, more knowledge is urgently needed to understand the impacts of RdRp P323L and A97L on polymerase activity and thus viral replication.

3CL^{pro}: 3CL^{pro} serve as a potential target by anti-viral inhibitors due to its crucial cleavage activity and functions in viral replication. As shown in Table 6 and Figure 5c, most frequent mutations found in 3CL^{pro} are G15S (138 counts, globe), T48I (17 counts, USA only), L75F (15 counts, USA only), L89F (32 counts, USA only), K90R (76 counts, China and Iceland), P108S (12 counts, Iceland and UK), L220F (22 counts, USA only), K236R (13 counts, USA only), D248E (43 counts, UK only), A266V (35 counts, Australia and USA) and N274D (13 counts, UK only). Key residues of 3CL^{pro} responsible for SARS-CoV catalytic activity, substrate binding and dimerization were checked, and none get changed in SARS-CoV-2. Anti-viral drugs targeting 3CL^{pro} typically dock within the Cys-His catalytic dyad (Cys145 and His41) which contains active catalytic binding site (Chitranshi et al., 2020). Mutations were not found in these two sites, suggesting that pharmacological inhibitors of 3CL^{pro} may still serve as therapeutics for SARS-CoV-2. However, with multiple high-frequent mutations found in 3CL^{pro} especially G15S, K90R and D248E, more studies about their impacts on cleavage activity and 3CL^{pro} drug efficacies are needed.

PL^{pro}: Same as 3CL^{pro}, proteolytic processing of polyprotein is also mediated by PL^{pro}. In this study, a total of 10 non-synonymous SNPs were found in PL^{pro} region (Table 7, Figure 5d). All of them are specific to a country. Seven sequence counts with L36F mutation were only distributed in China. I44V with 4 counts were only found in Canada. T63I with 40 sequence counts was found in Iceland only. In addition, T277I mutation was only distributed in USA with 15 sequences found. Spacious pockets for binding sites include residues Asp164, Val165, Arg166, Glu167, Met 208, Ala246, Pro247, Pro248, Tyr 264, Gly266, Asn267, Tyr 268, Gln269, Cys217, Gly271, Tyr273, Thr301 and Asp302 (Arya et al., 2020), among which only Proline in positive 247 was substituted to Leucine in two strains from Belgium. Essential properties like deISGylation and deubiquitination of PL^{pro} affect viral replication. Coronavirus PL^{pro} also serves as host innate immune antagonism. All these functions make PL^{pro} to be a potential target for anti-viral therapeutics. However, high-frequent mutations in PL^{pro} such as T63I may have negative effects on anti-viral drug efficacies.

4 | DISCUSSION

By analysis of 11,183 whole genomes of SARS-CoV-2, we demonstrated a high genetic variability between different regions and detailed mutational profiling across the genome and for key viral proteins (S, RdRp, 3CL^{pro} and PL^{pro}). In the present study, 60 out of 119 SNPs are nucleotide substitutions from C to T, representing the most abundant transition. Consistent with previous studies, this observation increases the frequency of codons for hydrophobic amino acids and provides evidence of potential anti-viral editing mechanisms driven by host (Matyasek & Kovarik, 2020; Mercatelli & Giorgi, 2020; Simmonds, 2020). On the other hand, more C to T transitions indicates less CpG abundancy, which is resulted from cytosine methylation and deamination into T. This mutational pattern was also observed in Bat RaTG13 and other coronaviruses, indicating rapid adaptation and evolution of the virus in the host (Matyasek & Kovarik, 2020; Simmonds, 2020). Among all known betacoronaviruses, SARS-CoV-2 represents the most extreme CpG deficiency, which contributes to evasion of host anti-viral defence mechanisms (Xia, 2020).

SARS-CoV-2 mutational pattern in each region varies from each other with North American and European countries more stability and Asian countries more variability (Figure 3). In addition, we did not observe a consistent mutational pattern contributing to the degree of case mortality/morbidity rate although some countries such as France, Belgium and UK do have a much higher fatality rate while countries such as Singapore and Iceland have a much lower fatality rate (Figure S2). Multiple factors were reported to impact the course of COVID-19 pandemic. Stringent measures such as quarantine, social distancing and isolation of infected patients have been implemented in China and result in successful containment of the epidemic (Anderson et al., 2020). Different social and economic factors among different countries also influence spread and outcomes of the disease (Qiu et al., 2020). In addition, according to WHO, the mortality is higher in people older than 65 years and those with underlying comorbidities, such as serious heart conditions, chronic lung disease, high blood pressure, obesity and diabetes (Lai et al., 2020; Ruan, 2020; Weiss & Murdoch, 2020).

SARS-CoV-2 strains from China demonstrate a high nucleotide substitution rate for C (C8782T) and *M* (T28144C) while the global strains feature substitutions on B (C3037T), E (C14408T) and J (A23403G), indicating rapid viral adaptation and evolution in other countries. The rapid spread to the world was reported to be a result from A23403G (D614G) mutation, which is responsible for increased viral infectivity, decreased neutralization sensitivity to individual convalescent serum and enhanced disease transmission thereafter (Daniloski et al., 2020; Hu et al., 2020; Korber et al., 2020; Ogawa et al., 2020; Yurkovetskiy et al., 2020; Zhang

et al., 2020). Virus strains with D614G mutations represents the dominant strains globally (). Also, the recent outbreaks in China during June were due to transmission of viral strains with D614G from Europe (Hu et al., 2020). However, whether or not other critical mutations with highest counts affects viral replicative ability needs to be defined. Landscape of genome-wide mutations globally and in different countries demonstrates high genetic diversity of SARS-CoV-2. Recombination events were reported in some studies (Gallaher, 2020; Korber et al., 2020; Paraskevis et al., 2020; Sashittal et al., 2020).

We also observed that N gene has 15 nucleotide positions mutated, then nsp2 and nsp13 (13), S gene (10), nsp14 (8), nsp7 and ORF3a (7), nsp13 (6) and nsp5 (5). This pattern is consistent with previous results claiming that ORF1a, ORF1b, S and N gene were detected at high frequency (Kim et al., 2020). N represents the most abundant protein expressed by viral genome and is able to induce high level of antibody response which ease serological diagnosis (Azkur et al., 2020; To et al., 2020). Non-synonymous mutations on N gene (C28311T, C28854T, G28881A and G28883C), especially G28881A and G28883C with vast majority of counts that contribute to clade classification, may have impacts on antigenicity of N protein. Further studies are needed to determine the impacts. We also observed here that nsp2 and nsp3 possess high mutation frequency (Figure 2). SARS coronavirus nsp1 and nsp2 are the most variable protein (Graham et al., 2005). However, previous research found that nsp2 are dispensable for SARS viral replication, but attenuates viral growth and genome synthesis (Graham et al., 2005). Nsp3 possesses PL^{pro} domain with protease-cleavage activities and serves as a target for anti-viral development (Rut et al., 2020). With high variability and high-frequency mutations including G2891A, C3037T, C3177T and C6312A, cautions and considerations should be taken for anti-viral therapeutic development. Multiple single nucleotide mutations lead to protein codon change to start/stop codons, which results in protein length change (Table 3). Mutations on TRS sites also may affect viral RNA transcription, thus affecting protein expression. Amino acids deletions and insertions were also observed (Table 2), and protein functions may get changed.

A detailed mutational profiling was performed for multiple key viral proteins including S, RdRp, $3CL^{pro}$ and PL^{pro} (Tables 4–7 and Figure 5). S protein mediates virus binding and entry to host cells, and is able to elicit high level of neutralizing antibody response (BalcioGlu et al., 2020; P. Liu, Cai, et al., 2020; Schmidt et al., 2020). Utilizing monoclonal antibodies (mAbs) to target RBD region as therapeutics have gained promising results and are currently under clinical trials for COVID-19 patients (Alsoussi et al., 2020; Chi et al., 2020; Shi et al., 2020). RdRp, $3CL^{pro}$ and PL^{pro} are conserved among all strains and play critical roles in viral genome replication and polyprotein cleavage to form functional viral proteins (Aftab et al., 2020; Chand et al., 2020; Chitranshi et al., 2020; Gao et al., 2020; Rut et al., 2020; Ul Qamar et al., 2020; Yin et al., 2020). Due to their critical feature of polymerase and protease, structures for RdRp, $3CL^{pro}$ have been decoded (Gao et al., 2020; Ul Qamar et al., 2020; Yin et al., 2020). Anti-viral drugs targeting these proteins are currently under development. Here, we described a detailed mutational profile of these four proteins. Critical mutations potentially impacting protein functions were observed and shown on their structures (Figure 5). Although counts for some of the mutations are not high, it provides insights that SARS-CoV-2 may adapt to environmental changes and gain replicative advantages/fitness to escape anti-viral treatment and being drug-resistant. Thus, further studies are needed to determine whether mutations on key sites affect viral replication and infectivity or not.

In summary, a detailed mutational profiling was described in this study. Landscape of genome-wide mutations across the countries provides insights for SARS-CoV-2 transmission and adaptation as different regions have different mutational patterns. Mutations with high frequency contribute to clade classification of SARS-CoV-2 strains. This study provides more evidence for SARS-CoV-2 genomic diversity around the globe and rapid evolution/adaptation of the virus. Given the detailed mutational profiles of key viral proteins including S, RdRp, $3CL^{pro}$ and PL^{pro}, it also gives some guidance for better design of anti-viral therapeutic to tackle the disease.

ACKNOWLEDGEMENTS

The authors sincerely appreciate the researchers worldwide who sequenced and shared the complete genome data of SARS-CoV-2 from GISAID (https://www.gisaid.org/). Please see the Supplemental PDF file for the acknowledgement table regarding the Authors from the Originating laboratories responsible for obtaining the specimens, as well as the Submitting laboratories where the genome data were generated and shared Via GISAID, on which this research is based.

[Correction added on 27 May 2021, after first online publication: An Acknowledgments section has been included in this current version.]

CONFLICT OF INTEREST

The authors declare that there is no competing interests.

ETHICAL APPROVAL

Ethical statement is not applicable since no human/animal sample handling and gathering were involved in this study.

DATA AVAILABILITY STATEMENT

The data used to support the findings of the manuscript are included within the article.

ORCID

Fangfeng Yuan https://orcid.org/0000-0001-9310-0382
Leyi Wang https://orcid.org/0000-0001-5813-9505

REFERENCES

Aftab, S. O., Ghouri, M. Z., Masood, M. U., Haider, Z., Khan, Z., Ahmad, A., & Munawar, N. (2020). Analysis of SARS-CoV-2 RNA-dependent RNA polymerase as a potential therapeutic drug target using a computational approach. *Journal of Translational Medicine*, *18*(1), 275. https://doi.org/10.1186/s12967-020-02439-0

Alsoussi, W. B., Turner, J. S., Case, J. B., Zhao, H., Schmitz, A. J., Zhou, J. Q., Chen, R. E., Lei, T., Rizk, A. A., McIntire, K. M., Winkler, E. S., Fox, J. M., Kafai, N. M., Thackray, L. B., Hassan, A. O., Amanat, F., Krammer, F., Watson, C. T., Kleinstein, S. H., … Ellebedy, A. H. (2020). A Potently Neutralizing Antibody Protects Mice against SARS-CoV-2 Infection. *The Journal of Immunology*, *205*(4), 915–922. https://doi.org/10.4049/jimmunol.2000583

Anderson, R. M., Heesterbeek, H., Klinkenberg, D., & Hollingsworth, T. D. (2020). How will country-based mitigation measures influence the course of the COVID-19 epidemic? *Lancet*, *395*(10228), 931–934. https://doi.org/10.1016/S0140-6736(20)30567-5

Arya, R., Das, A., Prashar, V., & Kumar, M. (2020). Potential inhibitors against papain-like protease of novel coronavirus (SARS-CoV-2) from FDA approved drugs. *ChemRxiv*, https://doi.org/10.26434/chemrxiv.11860011.v2

Azkur, A. K., Akdis, M., Azkur, D., Sokolowska, M., Veen, W., Brüggen, M.-C., O'Mahony, L., Gao, Y., Nadeau, K., & Akdis, C. A. (2020). Immune response to SARS-CoV-2 and mechanisms of immunopathological changes in COVID-19. *Allergy*, *75*(7), 1564–1581. https://doi.org/10.1111/all.14364

Bal, A., Destras, G., Gaymard, A., Bouscambert-Duchamp, M., Valette, M., Escuret, V., Frobert, E., Billaud, G., Trouillet-Assant, S., Cheynet, V., Brengel-Pesce, K., Morfin, F., Lina, B., & Josset, L. (2020). Molecular characterization of SARS-CoV-2 in the first COVID-19 cluster in France reveals an amino acid deletion in nsp2 (Asp268del). *Clinical Microbiology & Infection*, *26*(7), 960–962. https://doi.org/10.1016/j.cmi.2020.03.020

Balcioğlu, B. K., Denizci öncü, M., Öztürk, H. Ü., Yücel, F., Kaya, F., Serhatli, M., Ülbeği polat, H., Tekin, Ş., & Özdemir bahadir, A. (2020). SARS-CoV-2 neutralizing antibody development strategies. *Turkish Journal of Biology*, *44*(3), 203–214. https://doi.org/10.3906/biy-2005-91

Cagliani, R., Forni, D., Clerici, M., & Sironi, M. (2020). Coding potential and sequence conservation of SARS-CoV-2 and related animal viruses. *Infection, Genetics and Evolution*, *83*, 104353. https://doi.org/10.1016/j.meegid.2020.104353

Cao, Y., Su, B., Guo, X., Sun, W., Deng, Y., Bao, L., Zhu, Q., Zhang, X. U., Zheng, Y., Geng, C., Chai, X., He, R., Li, X., Lv, Q. I., Zhu, H., Deng, W., Xu, Y., Wang, Y., Qiao, L., … Xie, X. S. (2020). Potent neutralizing antibodies against SARS-CoV-2 identified by high-throughput single-cell sequencing of convalescent patients' B cells. *Cell*, *182*(1), 73–84 e16. https://doi.org/10.1016/j.cell.2020.05.025

Caudill, V. R., Qin, S., Winstead, R., Kaur, J., Tisthammer, K., Pineda, E. G., & Pennings, P. S. (2019). CpG-creating mutations are costly in many human viruses. *bioRxiv*, *34*(3), 339–359. https://doi.org/10.1101/702175.

Chand, G. B., Banerjee, A., & Azad, G. K. (2020). Identification of novel mutations in RNA-dependent RNA polymerases of SARS-CoV-2 and their implications on its protein structure. *PeerJ*, *8*, e9492. https://doi.org/10.7717/peerj.9492

Chen, Y. W., Yiu, C. B., & Wong, K. Y. (2020). Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CLpro) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Research*, *9*, 129–https://doi.org/10.12688/f1000research.22457.2

Chi, X., Yan, R., Zhang, J., Zhang, G., Zhang, Y., Hao, M., Zhang, Z., Fan, P., Dong, Y., Yang, Y., Chen, Z., Guo, Y., Zhang, J., Li, Y., Song, X., Chen, Y. I., Xia, L. U., Fu, L., Hou, L., … Chen, W. (2020). A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science*, *369*(6504), 650–655. https://doi.org/10.1126/science.abc6952

Chitranshi, N., Gupta, V. K., Rajput, R., Godinez, A., Pushpitha, K., Shen, T., Mirzaei, M., You, Y., Basavarajappa, D., Gupta, V., & Graham, S. L. (2020). Evolving geographic diversity in SARS-CoV2 and in silico analysis of replicating enzyme 3CL(pro) targeting repurposed drug candidates. *Journal of Translational Medicine*, *18*(1), 278. https://doi.org/10.1186/s12967-020-02448-z

Chitranshi, N., Gupta, V. K., Rajput, R., Godinez, A., Pushpitha, K., Shen, T., Mirzaei, M., You, Y., Basavarajappa, D., Gupta, V., & Graham, S. L. (2020). Evolving geographic diversity in SARS-CoV2 and in silico analysis of replicating enzyme 3CLpro targeting repurposed drug candidates. *Journal of Translational Medicine*, *18*(1), 278. https://doi.org/10.1186/s12967-020-02448-z

Daniloski, Z., Guo, X., & Sanjana, N. E. (2020). The D614G mutation in SARS-CoV-2 Spike increases transduction of multiple human cell types. *bioRxiv*. https://doi.org/10.1101/2020.06.14.151357

di Gioacchino, A., Sulc, P., Komarova, A. V., Greenbaum, B. D., Monasson, R., & Cocco, S. (2020). The heterogeneous landscape and early evolution of pathogen-associated CpG and UpA dinucleotides in SARS-CoV-2. *bioRxiv*. https://doi.org/10.1101/2020.05.06.074039

Elena, S. F., & Sanjuán, R. (2005). Adaptive value of high mutation rates of RNA viruses: Separating causes from consequences. *Journal of Virology*, *79*(18), 11555–11558. https://doi.org/10.1128/JVI.79.18.11555-11558.2005

Ellinghaus, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P., & Karlsen, T. H. (2020). Genomewide association study of severe Covid-19 with respiratory failure. *New England Journal of Medicine*, *383*, 1522–1534.

Erasmus, J. H., Khandhar, A. P., O'Connor, M. A., Walls, A. C., Hemann, E. A., Murapa, P., & Fuller, D. H. (2020). An Alphavirus-derived replicon RNA vaccine induces SARS-CoV-2 neutralizing antibody and T cell responses in mice and nonhuman primates. *Science Translational Medicine*, *12*(555), eabc9396. https://doi.org/10.1126/scitranslmed.abc9396

Fehr, A. R., & Perlman, S. (2015). Coronaviruses: An overview of their replication and pathogenesis. *Methods in Molecular Biology (Clifton. N.J.)*, *1282*, 1–23. https://doi.org/10.1007/978-1-4939-2438-7_1

Gallaher, W. R. (2020). A palindromic RNA sequence as a common breakpoint contributor to copy-choice recombination in SARS-COV-2. *Archives of Virology*, *165*(10), 2341–2348. https://doi.org/10.1007/s00705-020-04750-z

Gao, Y., Yan, L., Huang, Y., Liu, F., Zhao, Y., Cao, L., Wang, T., Sun, Q., Ming, Z., Zhang, L., Ge, J. I., Zheng, L., Zhang, Y., Wang, H., Zhu, Y., Zhu, C., Hu, T., Hua, T., Zhang, B., … Rao, Z. (2020). Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science*, *368*(6492), 779. https://doi.org/10.1126/science.abb7498

Graham, R. L., Sims, A. C., Brockway, S. M., Baric, R. S., & Denison, M. R. (2005). The nsp2 replicase proteins of murine hepatitis virus and severe acute respiratory syndrome coronavirus are dispensable for viral replication. *Journal of Virology*, *79*(21), 13399. https://doi.org/10.1128/JVI.79.21.13399-13411.2005

Grubaugh, N. D., Hanage, W. P., & Rasmussen, A. L. (2020). Making sense of mutation: What D614G means for the COVID-19 pandemic remains unclear. *Cell*, *182*(4), 794–795. https://doi.org/10.1016/j.cell.2020.06.040

Guan, Q., Sadykov, M., Nugmanova, R., Carr, M. J., Arold, S. T., & Pain, A.. (2020). The genomic variation landscape of globally-circulating clades of SARS-CoV-2 defines a genetic barcoding scheme. *bioRxiv*, 2020.2004.2021.054221. https://doi.org/10.1101/2020.04.21.054221

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, *34*(23), 4121–4123. https://doi.org/10.1093/bioinformatics/bty407

Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95–98.

Han, A. X., Parker, E., Scholer, F., Maurer-Stroh, S., & Russell, C. A. (2019). Phylogenetic Clustering by Linear Integer Programming (PhyCLIP).

Molecular Biology and Evolution, *36*(7), 1580–1595. https://doi.org/10.1093/molbev/msz053

Hillen, H. S., Kokic, G., Farnung, L., Dienemann, C., Tegunov, D., & Cramer, P. (2020). Structure of replicating SARS-CoV-2 polymerase. *Nature*, *584*(7819), 154–156. https://doi.org/10.1038/s41586-020-2368-8

Hu, J., He, C.-L., Gao, Q.-Z., Zhang, G.-J., Cao, X.-X., Long, Q.-X., & Huang, A.-L. (2020). The D614G mutation of SARS-CoV-2 spike protein enhances viral infectivity and decreases neutralization sensitivity to individual convalescent sera. *bioRxiv*, 2020.2006.2020.161323. https://doi.org/10.1101/2020.06.20.161323

Katoh, K., Rozewicki, J., & Yamada, K. D. (2017). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *20*(4), 1160–1166. https://doi.org/10.1093/bib/bbx108

Kim, J. S., Jang, J. H., Kim, J. M., Chung, Y. S., Yoo, C. K., & Han, M. G. (2020). Genome-Wide Identification and characterization of point mutations in the SARS-CoV-2 genome. *Osong Public Health and Research Perspectives*, *11*(3), 101–111. https://doi.org/10.24171/j.phrp.2020.11.3.05

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E. E., Bhattacharya, T., Foley, B., Hastie, K. M., Parker, M. D., Partridge, D. G., Evans, C. M., Freeman, T. M., de Silva, T. I., McDanal, C., Perez, L. G., Tang, H., … Wyles, M. D. (2020). Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell*, *182*(4), 812–827.e19. https://doi.org/10.1016/j.cell.2020.06.043

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Montefiori, D. C. (2020). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv*, 2020.2004.2029.069054. https://doi.org/10.1101/2020.04.29.069054

Lai, C. C., Shih, T. P., Ko, W. C., Tang, H. J., & Hsueh, P. R. (2020). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. *International Journal of Antimicrobial Agents*, *55*(3), 105924. https://doi.org/10.1016/j.ijantimicag.2020.105924

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q. I., Shi, X., Wang, Q., Zhang, L., & Wang, X. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, *581*(7807), 215–220. https://doi.org/10.1038/s41586-020-2180-5

Li, J., Wang, X., Chen, J., Cai, Y. I., Deng, A., & Yang, M. (2020). Association between ABO blood groups and risk of SARS-CoV-2 pneumonia. *British Journal of Haematology*, *190*(1), 24–27. https://doi.org/10.1111/bjh.16797

Liu, D. X., Liang, J. Q., & Fung, T. S. (2020). Human Coronavirus-229E, -OC43, -NL63, and -HKU1. *Reference Module in Life Sciences*, B978–970-912-809633-809638.821501-X. https://doi.org/10.1016/B978-0-12-809633-8.21501-X

Liu, P., Cai, J., Jia, R., Xia, S., Wang, X., Cao, L., Zeng, M., & Xu, J. (2020). Dynamic surveillance of SARS-CoV-2 shedding and neutralizing antibody in children with COVID-19. *Emerg Microbes Infect*, *9*(1), 1254–1258. https://doi.org/10.1080/22221751.2020.1772677

Matyasek, R., & Kovarik, A. (2020). Mutation patterns of human SARS-CoV-2 and bat RaTG13 coronavirus genomes are strongly biased towards C>U transitions, indicating rapid evolution in their hosts. *Genes (Basel)*, *11*(7), https://doi.org/10.3390/genes11070761

Mercatelli, D., & Giorgi, F. M. (2020). Geographic and genomic distribution of SARS-CoV-2 mutations. *Frontiers in Microbiology*, *11*, 1800. https://doi.org/10.3389/fmicb.2020.01800

Mercurio, I., Tragni, V., Busto, F., De Grassi, A., & Pierri, C. L. (2020). Protein structure analysis of the interactions between SARS-CoV-2 spike protein and the human ACE2 receptor: From conformational changes to novel neutralizing antibodies. *Cellular and Molecular Life Sciences*, https://doi.org/10.1007/s00018-020-03580-1

Ogawa, J., Zhu, W., Tonnu, N., Singer, O., Hunter, T., Ryan, A. L., & Pao, G. M. (2020). The D614G mutation in the SARS-CoV2 Spike protein increases infectivity in an ACE2 receptor dependent manner. *bioRxiv*. https://doi.org/10.1101/2020.07.21.214932

Ozono, S., Zhang, Y., Ode, H., Seng, T. T., Imai, K., Miyoshi, K., Tokunaga, K.. (2020). Naturally mutated spike proteins of SARS-CoV-2 variants show differential levels of cell entry. *bioRxiv*, 2020.2006.2015.151779. https://doi.org/10.1101/2020.06.15.151779

Paraskevis, D., Kostaki, E. G., Magiorkinis, G., Panayiotakopoulos, G., Sourvinos, G., & Tsiodras, S. (2020). Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, *79*, 104212. https://doi.org/10.1016/j.meegid.2020.104212

Qiu, Y., Chen, X., & Shi, W. (2020). Impacts of social and economic factors on the transmission of coronavirus disease 2019 (COVID-19) in China. *Population Economics*, *33*(4), 1127–1172. https://doi.org/10.1007/s00148-020-00778-2

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., & Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, *5*(11), 1403–1407. https://doi.org/10.1038/s41564-020-0770-5

Rozewicki, J., Li, S., Amada, K. M., Standley, D. M., & Katoh, K. (2019). MAFFT-DASH: Integrated protein sequence and structural alignment. *Nucleic Acids Research*, *47*(W1), W5–W10. https://doi.org/10.1093/nar/gkz342

Ruan, S. (2020). Likelihood of survival of coronavirus disease 2019. *The Lancet Infectious Diseases*, *20*(6), 630–631. https://doi.org/10.1016/S1473-3099(20)30257-7

Rut, W., Lv, Z., Zmudzinski, M., Patchett, S., Nayak, D., Snipas, S. J., Olsen, S. K. (2020). Activity profiling and structures of inhibitor-bound SARS-CoV-2-PLpro protease provides a framework for anti-COVID-19 drug design. *bioRxiv*. https://doi.org/10.1101/2020.04.29.068890

Saha, I., Ghosh, N., Maity, D., Sharma, N., Sarkar, J. P., & Mitra, K. (2020). Genome-wide analysis of Indian SARS-CoV-2 genomes for the identification of genetic mutation and SNP. *Infection, Genetics and Evolution*, *85*, 104457. https://doi.org/10.1016/j.meegid.2020.104457

Sashittal, P., Luo, Y., Peng, J., & El-Kebir, M. (2020). Characterization of SARS-CoV-2 viral diversity within and across hosts. *bioRxiv*, 2020.2005.2007.083410. https://doi.org/10.1101/2020.05.07.083410

Schmidt, F., Weisblum, Y., Muecksch, F., Hoffmann, H. H., Michailidis, E., Lorenzi, J. C. C., Bieniasz, P. D. (2020). Measuring SARS-CoV-2 neutralizing antibody activity using pseudotyped and chimeric viruses. *bioRxiv*. https://doi.org/10.1101/2020.06.08.140871

Shi, R., Shan, C., Duan, X., Chen, Z., Liu, P., Song, J., Song, T., Bi, X., Han, C., Wu, L., Gao, G. E., Hu, X., Zhang, Y., Tong, Z., Huang, W., Liu, W. J., Wu, G., Zhang, B. O., Wang, L., … Yan, J. (2020). A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature*, *584*(7819), 120–124. https://doi.org/10.1038/s41586-020-2381-y

Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. *EuroSurveillance*, *22*(13). https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

Simmonds, P. (2020). Rampant C–>U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: Causes and consequences for their short- and long-term evolutionary trajectories. *mSphere*, *5*(3). https://doi.org/10.1128/mSphere.00408-20

Singh, N., Decroly, E., Khatib, A. M., & Villoutreix, B. O. (2020). Structure-based drug repositioning over the human TMPRSS2 protease domain: Search for chemical probes able to repress SARS-CoV-2 Spike protein cleavages. *European Journal of Pharmaceutical Sciences*, *153*, 105495. https://doi.org/10.1016/j.ejps.2020.105495

Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., Duan, Y., Zhang, H., Wang, Y., Qian, Z., Cui, J., & Lu, J. (2020). On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, *7*(6), 1012–1023. https://doi.org/10.1093/nsr/nwaa036

To, K.-W., Tsang, O.-Y., Leung, W.-S., Tam, A. R., Wu, T.-C., Lung, D. C., Yip, C.-Y., Cai, J.-P., Chan, J.-C., Chik, T.-H., Lau, D.-L., Choi, C.-C., Chen, L.-L., Chan, W.-M., Chan, K.-H., Ip, J. D., Ng, A.-K., Poon, R.-S., Luo, C.-T., … Yuen, K.-Y. (2020). Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: An observational cohort study. *The Lancet Infectious Diseases*, *20*(5), 565–574. https://doi.org/10.1016/S1473-3099(20)30196-1

Tubiana, L., Bozic, A. L., Micheletti, C., & Podgornik, R. (2015). Synonymous mutations reduce genome compactness in icosahedral ssRNA viruses. *Biophysical Journal*, *108*(1), 194–202. https://doi.org/10.1016/j.bpj.2014.10.070

Ul Qamar, M. T., Alqahtani, S. M., Alamri, M. A., & Chen, L. L. (2020). Structural basis of SARS-CoV-2 3CL(pro) and anti-COVID-19 drug discovery from medicinal plants. *Journal of Pharmaceutical Analysis*, *10*(4), 313–319. https://doi.org/10.1016/j.jpha.2020.03.009

Walls, A. C., Park, Y. J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*, *181*(2), 281–292 e286. https://doi.org/10.1016/j.cell.2020.02.058

Weiss, P., & Murdoch, D. R. (2020). Clinical course and mortality risk of severe COVID-19. *Lancet*, *395*(10229), 1014–1015. https://doi.org/10.1016/S0140-6736(20)30633-4

Xia, S., Liu, M., Wang, C., Xu, W., Lan, Q., Feng, S., Qi, F., Bao, L., Du, L., Liu, S., Qin, C., Sun, F., Shi, Z., Zhu, Y., Jiang, S., & Lu, L. U. (2020). Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. *Cell Research*, *30*(4), 343–355. https://doi.org/10.1038/s41422-020-0305-x

Xia, X. (2020). Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. *Molecular Biology and Evolution*, *37*(9), 2699–2705. https://doi.org/10.1093/molbev/msaa094

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: Protein structure and function prediction. *Nature Methods*, *12*(1), 7–8. https://doi.org/10.1038/nmeth.3213

Ye, Z.-W., Yuan, S., Yuen, K.-S., Fung, S.-Y., Chan, C.-P., & Jin, D.-Y. (2020). Zoonotic origins of human coronaviruses. *International Journal of Biological Sciences*, *16*(10), 1686–1697. https://doi.org/10.7150/ijbs.45472

Yin, W., Mao, C., Luan, X., Shen, D.-D., Shen, Q., Su, H., Wang, X., Zhou, F., Zhao, W., Gao, M., Chang, S., Xie, Y.-C., Tian, G., Jiang, H.-W., Tao, S.-C., Shen, J., Jiang, Y. I., Jiang, H., Xu, Y., … Xu, H. E. (2020). Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science*, *368*(6498), 1499. https://doi.org/10.1126/science.abc1560

Yurkovetskiy, L., Pascal, K. E., Tompkins-Tinch, C., Nyalile, T., Wang, Y., Baum, A., Luban, J. (2020). SARS-CoV-2 Spike protein variant D614G increases infectivity and retains sensitivity to antibodies that target the receptor binding domain. *bioRxiv*. https://doi.org/10.1101/2020.07.04.187757

Zhang, L., Jackson, C. B., Mou, H., Ojha, A., Rangarajan, E. S., Izard, T., Choe, H. (2020). The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv*. https://doi.org/10.1101/2020.06.12.148726

Zhu, N. A., Zhang, D., Wang, W., Li, X., Yang, B. O., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., & Tan, W. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine*, *382*(8), 727–733. https://doi.org/10.1056/NEJMoa2001017

[Correction added on 27 May 2021, after first online publication: The full reference for "Shu & McCauley, 2017" has been included in the References list in this current version.]

SUPPORTING INFORMATION

[Correction added on 27 May 2021, after first online publication: A new supplementary table has been added to the online Supporting Information section.]

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Yuan F, Wang L, Fang Y, Wang L. Global SNP analysis of 11,183 SARS-CoV-2 strains reveals high genetic diversity. *Transbound Emerg Dis*. 2021;68:3288–3304. https://doi.org/10.1111/tbed.13931