Journal of
Translational Medicine

RESEARCH

Open Access

Global landscape of SARS-CoV-2 mutations and conserved regions

Mohammad Hadi Abbasian[1†], Mohammadamin Mahmanzar[2†], Karim Rahimian[3†], Bahar Mahdavi[4], Samaneh Tokhanbigli[5], Bahman Moradi[6], Mahsa Mollapour Sisakht[7] and Youping Deng[2*]

Abstract

Background At the end of December 2019, a novel strain of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) disease (COVID-19) has been identified in Wuhan, a central city in China, and then spread to every corner of the globe. As of October 8, 2022, the total number of COVID-19 cases had reached over 621 million worldwide, with more than 6.56 million confirmed deaths. Since SARS-CoV-2 genome sequences change due to mutation and recombination, it is pivotal to surveil emerging variants and monitor changes for improving pandemic management.

Methods 10,287,271 SARS-CoV-2 genome sequence samples were downloaded in FASTA format from the GISAID databases from February 24, 2020, to April 2022. Python programming language (version 3.8.0) software was utilized to process FASTA files to identify variants and sequence conservation. The NCBI RefSeq SARS-CoV-2 genome (accession no. NC_045512.2) was considered as the reference sequence.

Results Six mutations had more than 50% frequency in global SARS-CoV-2. These mutations include the P323L (99.3%) in NSP12, D614G (97.6) in S, the T492I (70.4) in NSP4, R203M (62.8%) in N, T60A (61.4%) in Orf9b, and P1228L (50.0%) in NSP3. In the SARS-CoV-2 genome, no mutation was observed in more than 90% of nsp11, nsp7, nsp10, nsp9, nsp8, and nsp16 regions. On the other hand, N, nsp3, S, nsp4, nsp12, and M had the maximum rate of mutations. In the S protein, the highest mutation frequency was observed in aa 508–635(0.77%) and aa 381–508 (0.43%). The highest frequency of mutation was observed in aa 66–88 (2.19%), aa 7–14, and aa 164–246 (2.92%) in M, E, and N proteins, respectively.

Conclusion Therefore, monitoring SARS-CoV-2 proteomic changes and detecting hot spots mutations and conserved regions could be applied to improve the SARS-CoV-2 diagnostic efficiency and design safe and effective vaccines against emerging variants.

Keywords SARS-CoV-2, COVID-19, Emerging variants, Genome, Amino Acid, Vaccines

†Mohammad Hadi Abbasian, Mohammadamin Mahmanzar and Karim Rahimian have contributed equally to this work

*Correspondence:
Youping Deng
dengy@hawaii.edu

[1] Department of Medical Genetics, National Institute for Genetic Engineering and Biotechnology, Tehran, Iran
[2] Department of Quantitative Health Sciences, John A. Burns School of Medicine, University of Hawaii at Manoa, Honolulu, HI 96813, USA
[3] Department of Bioinformatics, Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran
[4] Department of Computer Science, Tarbiat Modares University, Tehran, Iran
[5] Discipline of Pharmacy, Graduate School of Health, University of Technology Sydney, Sydney, Australia
[6] Department of Biology, Faculty of Sciences, Shahid Bahonar University of Kerman, Kerman, Iran
[7] Department of Biochemistry, Erasmus University Medical Center, 2040, 3000 CA Rotterdam, The Netherlands

Introduction

Over the last two decades, we have seen three lethal coronavirus outbreaks, severe acute respiratory syndrome (SARS, 2002–03) (1), Middle East respiratory syndrome (MERS, since 2012) (2), and now coronavirus disease 2019 (COVID-19, since late 2019) (3). The ecological realities assume that coronaviruses continue to pose a potentially existential threat in the future (4, 5) and already have a significant health, social, and economic impact on millions of people globally. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes COVID-19, appeared in early December 2019 in Wuhan, a city of 11 million populations in China's Hubei province [90, 92]. According to the World Health Organization (WHO), more than 621 million individuals worldwide have been infected with SARS-CoV-2 during the COVID-19 pandemic as of September 2022 [1].

The SARS-CoV-2 is a lipid-enveloped, single-stranded, and positive-sense RNA (+ssRNA) virus with a large genome length of 29,903 nucleotides that contains a 5′-cap structure, a 3′ poly(A) tail, 2 flanking untranslated regions (UTRs), and multiple open reading frames (ORFs) which encoding viral structural proteins and regulatory elements [14, 15] (Fig. 1). SARS-CoV-2's gene content, function, and interactions with host factors have not yet been fully elucidated, notwithstanding its critical medical concern. Therefore, it is essential to have an insight into the basic virology of SARS-CoV-2 to develop better and more efficient therapeutics.

Approximately two-thirds of the entire genome of SARS-CoV-2 contains the ORF1a/b coding region, which is considered the largest ORF at the 5′ terminus. The -1 ribosomal frameshift between ORF1a and ORF1b leads to the formation of two co-terminal polypeptide domains called pp1a and pp1ab. The pp1a protein encoded by ORF1a is proteolytically cleaved into 11 mature non-structural proteins (NSP1-11). In contrast, the pp1ab protein expressed by ORF1ab is processed into 15 NSPs (NSP1-10 and NSP12-16) [90, 92]. The remaining third of the downstream genome near the 3′-terminus comprises ORFs encoding structural proteins (SPs) and accessory proteins [31]. Accessory proteins include ORFs 3a, 3b, 3c, 3d, 6, 7a, 7b, 8a, 8b, 9b, 9c, and ORF10 [32], which are distributed among the four major structural protein genes, namely spike surface glycoprotein (S), an envelope protein (E), membrane glycoprotein (M), and nucleocapsid phosphoprotein (N) [90, 92]. In addition to genomic RNA, several canonical subgenomic (sg) mRNAs are also produced [25, 34] (Fig. 1).

Numerous functional and structural domains of NSPs are well defined, including NSP3 as a papain-like protease (PL2pro) [40] and NSP5 as a 3C-like protease (3CLpro or Mpro) [84, 99, 100] that cleaves pp1a and pp1ab polypeptides in 15 NSPs [36]. NSP12, as a multi-subunit RNA-dependent RNA polymerase (RdRp) [95] in complex with NSP7 and NSP8 as co-factors, forms a replicase complex for replication and transcription of viral genomic RNA [38]. The nidovirus RdRp-associated nucleotidyltransferase (NiRAN) domain featured in NSP12 possesses a β-hairpin domain at its N-terminus [23]. NSP13 as a helicase (Hel) [74] and NSP14 as a proofreading exoribonuclease (ExoN) [44] are critical enzymes that facilitate viral RNA replication and transcription. Other NSPs are almost considered in the host cell and immune suppression. Structural proteins of the 3′-end are involved in viral interaction with the host cell angiotensin-converting enzyme 2 (ACE2) receptor, membrane fusion and entry of the virus into the host cells [63, 87, 89], viral assembly, morphogenesis, and release of virion particles from the host cell [43]. The function and expression of accessory proteins are still largely unknown.

Considerable genomes from all parts of the world have been sequenced and are available at the Global Initiative on Sharing All Influenza Data (GISAID; https://platform.gisaid.org/) [21, 35, 75] and NCBI (https://www.ncb.nlm.nih.gov/) from the onset of the pandemic in January 2020.

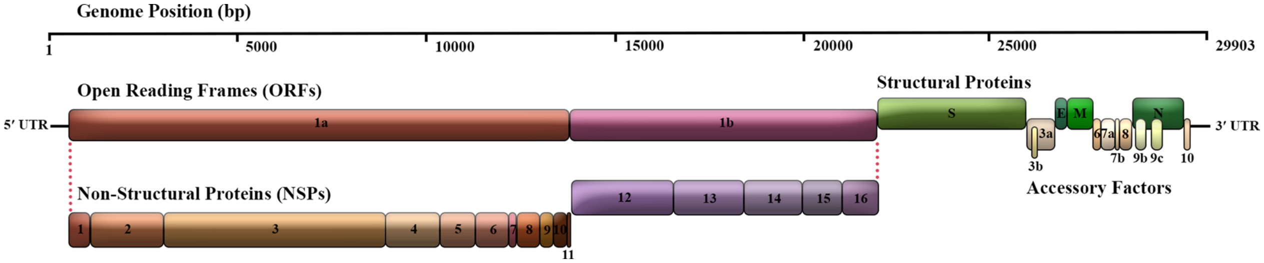

Fig. 1 Schematic view of the SARS-CoV-2 genome arrangement. SARS-CoV-2 is an enveloped single-stranded positive-sense RNA beta coronavirus with a polycistronic genome ~ 30 kb in length. SARS-CoV-2 genome encodes several non-structural proteins (ORF1a and ORF1b, that are processed into NSP1-16) at the 5′-end, in addition to structural proteins (S, E, M, and N), and multiple other accessory proteins (ORF3a, 6, 7a, 7b, 8, 9b, 9c and 10) at the 3′-end

In this study, we used bioinformatics tools to process these massive datasets efficiently and evaluate approximately 10,300,000 SARS-CoV-2 genome sequences worldwide until April 28, 2022. In our high-throughput experiments, we tracked a systematic gene-by-gene comparison analysis with a reference genome (the first sequence data of a patient from Wuhan in the National Center for Biotechnology Information (NCBI) annotation NC_045512.2) to evaluate conserved genomic regions of SARS-CoV-2. Our study results provide a systematic resource to identify novel sequence features or functional elements worth consideration as vaccine candidates and therapeutic development.

Methods

Sequence retrieval

Figure 2 illustrates the research methodology workflow of our study. We obtained credentials to access data in the GISAID database (26–28) with Erasmus Medical Center authorization. The NCBI RefSeq SARS-CoV-2 genome (accession no. NC_045512.2) was considered as the reference sequence. The whole available data of SARS-CoV-2 full-length genome sequences (10,287,271 samples) and their amino acid sequences, including their geographical locations and submission dates of sample annotations, were downloaded in FASTA format from the GISAID databases by April 28, 2022.

Sequence alignment and trimming

Python programming language (version 3.8.0) software was utilized for Pre- and post-processing FASTA files. The entire collected 10,287,271 sample sequences were aligned to the reference sequence using the EV couplings Python package. Excluding the 3′ and 5′ terminus regions was carried out after alignment to eliminate large numbers of missing and ambiguous reads and to achieve better alignment accuracy. Then excessively divergent, short, or long sequences, gaps (including dash and space characters), ambiguous nucleotides (such as N, B, and W), or non-specified amino acids (indicated by X) were removed. Eventually, genomes were filtered from non-human host species (such as mammals and birds) since no significant numbers existed to participate in the study Additional file 1.

Sequence analyses and processing

After the trimming and filtering step, each of the remaining high-quality and high-coverage SARS-CoV-2

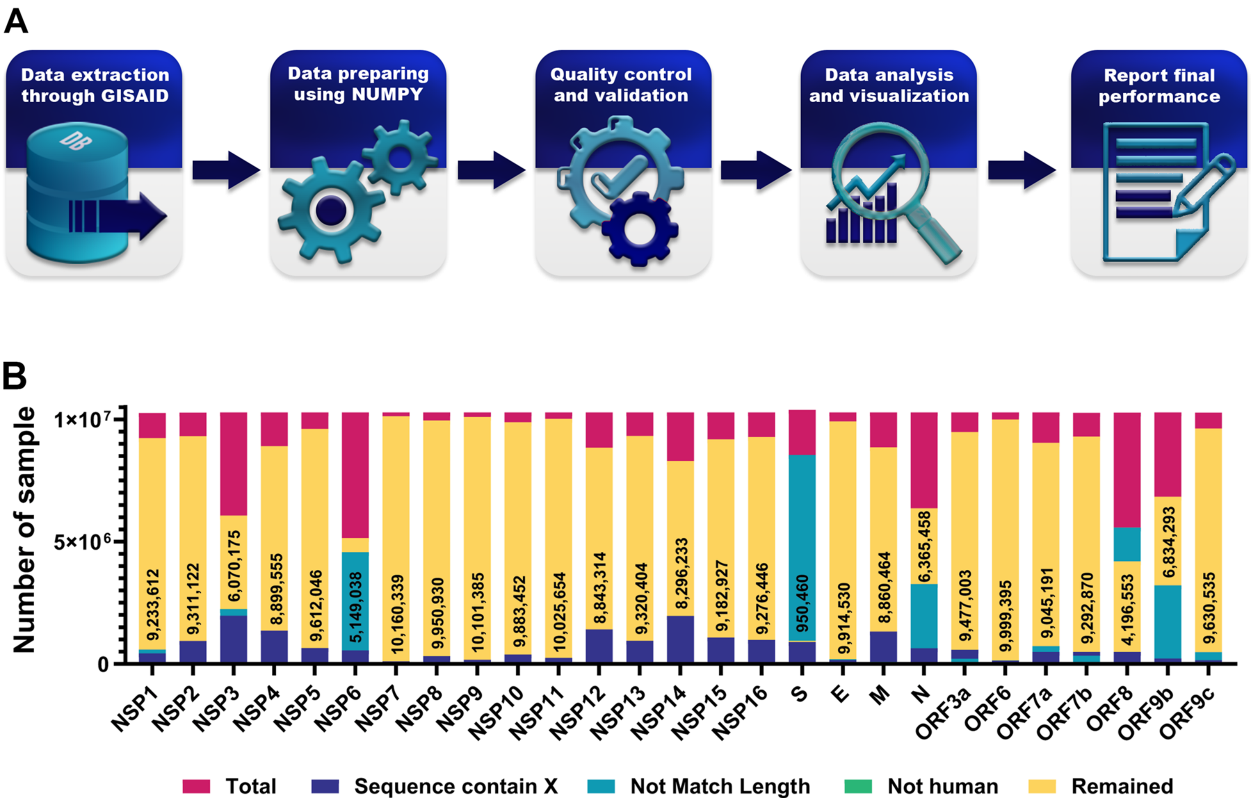

Fig. 2 Overview of the study design. **A** Schematic describing the workflow of the study. **B** Illustration showing the number of SARS-CoV-2 samples. The number of analyzed SARS-CoV-2 samples are mentioned in bars

complete genome sequences was aligned against the reference sequence to identify variants, sequence conservation, and annotate them through the X functions of the Y package. Compared with the reference sequence, each amino acid-changing replacement in samples was defined as a variant. We recognized that as a conserved if there were no intergenic amino acid-changing replacements in the alignment. In such cases, their location of them was reported. The entire high-level processing was optimized using NumPy (version 1.16.2) and Pandas (version 0.22.0) python libraries Additional file 2.

After extracting conserved and hot spot regions, each sample's continent name and geographic coordinates were detailed through pycountry-convert and titlecase Python libraries and displayed in global maps using Matlab programming language (version R2021a) software; Geobubble package. A flowchart outlining the whole procedure in this study is shown in Fig. 2.

In silico analysis of mutations for prediction of protein stability

We have performed four different structure-based bioinformatics tools to examine the effect of several important mutations identified in this study on the stability of SARS-CoV-2 proteins. Different types of computational methods were developed to predict stability changes upon mutation. These methods estimate Gibbs free energy values (ΔΔG) to classify the effect of each mutation as stabilizing or destabilizing on the protein structure. DynaMut [68], DynaMut2 [69], MAESTROweb [39], and SDM [60] are web servers that are able to predict the effects of missense mutations on protein stability. The crystal structures of SARS-CoV-2 wild-type proteins were retrieved from Protein Data Bank (PDB) [8].

Statistical analysis

Exploration, normalization, and visualization of the data were conducted using GraphPad Prism (version 8.0.2) and Microsoft Power BI software as data analytics tools. Clusters were identified through hierarchical clustering analysis using the stats library in the R (version 4.1.3) programming language.

Results and discussion

Top mutations in SARS-CoV-2 in different geographic areas

The mutation is the engine of evolution that generates genetic diversity. It has been demonstrated that mutations in the SARS-CoV-2 genome are responsible for a drastic change in the protein structure and lead to an increase in the infectivity, fitness, and virulence of SARS-CoV-2 [19]. 10,287,271 SARS-CoV-2 samples from February 24, 2020, to April 2022 were downloaded from the GISAID database. High-quality SARS-CoV-2 complete genome sequences for each gene were analyzed (Fig. 2). Mutational Analysis of the SARS-CoV-2 revealed a high rate of recurrent mutations across all regions (Table 1). Six mutations had more than 50% frequency in global SARS-CoV-2. These mutations include the P323L (99.3%) in NSP12, D614G (97.6) in S, the T492I (70.4) in NSP4, R203M (62.8%) in N, T60A (61.4%) in Orf9b, and P1228L (50.0%) in NSP3. Interestingly, the frequency of these top mutations was higher in North America and Europe than in South Africa and Oceania (Table 1).

Mutation rate in the SARS-CoV-2 genome

The mutation rate is the value at which the base pair or a larger region of DNA changes are made during the genome replication. Understanding the rate of mutation is of fundamental importance because the rate of mutations has explained the emergence of new SARS-CoV-2 variants and their establishment in natural populations.

The mutational profile of SARS-CoV-2 genome sequences has been changed during pandemics. Therefore, genomic surveillance of SARS-CoV-2 is critical to monitoring SARSCoV-2 genetic variability, which can improve diagnostic tools, vaccines, and immunotherapeutic interventions against COVID-19 [16]. RNA viruses vary in mutation rates, such as 1.35 × 10−5 in Influenza H3N2 and −1 4 × 10−5 mutations per target in HIV [27, 51]. The SARS coronavirus mutation rate is calculated at 9.0 × 10−7 mutations per nucleotide per replication cycle (m/n/rc) which is lower than most RNA viruses [20].The average mutation rate of 4×10^{-4} nucleotide substitutions/site/year [71]. RNA proofreading capability of Coronavirus preserves its genome since previous studies reported that nsp14 acts as a 3′-5′ exoribonuclease that explains the coronaviruses' extraordinary length single-stranded linear genome [53, 76]. The mutation rate of the SARS-CoV-2 genome has thus been estimated at 1 × 10−3 substitutions per base (30 nucleotides/genome) per year under neutral genetic drift conditions [85] or 1×10^{-5}–1×10^{-4} substitutions per base in each transmission events [86]. Li et al. analyzed the evolution rate of SARS-CoV-2 32 genomes of virus strains between December 24, 2019, and January 23, 2020. The mean evolutionary rate for SARS-CoV-2 32 genomes ranged from 1.7926×10^{-3} to 1.8266×10^{-3} substitutions per site per year [41, 42]. Four months after the pandemic, the mutation rate for a complete SARS-CoV-2 genome with 29,903 nucleotides was 3.95 × 10−4 per nucleotide per year. The SARS-CoV-2 without its non-structural proteins 13 to 16 (Nsp13-Nsp16) exhibits an unusually high mutation rate [73]. This rate is lower than other RNA viruses, such as influenza A/H3N2 (10.9×10^{-6} nucleotide substitutions per site per day) [57]. Motayo et al. reported that the

Table 1 Most common mutations in different SARS-CoV-2 genes

SARS-CoV-2 region	Mutation	Worldwide (%)	Africa(%)	Asia(%)	Europe(%)	North America(%)	South America(%)	Oceania(%)
NSP1	S135R	8.01	3.49	8.10	12.5	2.24	0.41	12.2
NSP2	T85I	3.94	15.6	1.79	0.64	9.27	1.05	1.48
NSP3	P1228L	50.0	32.2	45.1	49.0	53.6	44.1	50.1
NSP4	T492I	70.4	52.6	60.5	73.3	69.4	59.8	77.9
NSP5	P1328	30.8	23.7	25.3	34.6	27.2	39.2	21.9
NSP6	T77A	71.1	50.0	57.2	78.2	66.8	70.3	65.2
NSP7	L71F	0.22	0.14	0.25	0.13	0.20	2.8	0.03
NSP8	Q24R	0.70	0.18	0.06	0.44	1.29	0.12	0.03
NSP9	T35I	0.50	0.33	0.13	0.37	0.80	0.18	0.03
NSP10	T102I	0.16	0.05	0.07	0.17	0.19	0.10	0.05
NSP11	S6L	0.16	0.27	0.08	0.16	0.19	0.10	0.04
NSP12	P323L	99.3	95.8	98.7	99.5	99.3	99.5	97.7
NSP13	P77L	42.6	35.1	44.6	40.0	46.7	32.2	39.2
NSP14	A394V	35.1	22.5	31.7	34.9	37.0	24.9	36.5
NSP15	T112I	8.97	5.36	10.1	13.6	2.56	1.44	12.4
NSP16	R216C	1.26	1.97	0.08	0.09	3.36	0.07	0.12
Spike	D614G	97.6	91.2	95.3	97.5	98.3	99.5	90.1
Orf3a	S26L	43.1	34.4	46.0	40.5	47.2	32.8	37.0
Envelope	T9I	30.6	21.9	26.5	35.0	25.7	21.3	38.8
Membrane	I82T	47.1	42.4	47.8	45.0	52.1	34.1	40.6
Orf6	D61L	9.49	4.82	10.1	14.6	2.70	0.58	13.0
Orf7a	T120I	41.4	33.3	43.9	39.4	45.3	31.2	10.5
Orf7b	T40I	39.5	30.6	33.2	39.2	41.7	33.8	40.6
Orf8a	S24L	3.53	0.14	0.09	0.09	9.42	0.05	0.44
N	D377Y	63.2	48.8	60.4	63.2	65.6	41.7	–
Orf9b	T60A	61.4	45.1	59.1	61.7	63.4	41.0	59.3
Orf9c	G50N	48.8	37.9	47.4	52.8	42.4	–	51.02

evolutionary rate of the Afr-SARS-CoV-2 from February 24 to April 24 was 4.133×10^{-4} substitutions/site/year [54].

The number of studied different sequences of the SARS-CoV-2 proteins is summarized in the Fig. 2. The Analysis of the worldwide SARS-CoV-2 genome revealed that no mutation was observed in more than 90% of nsp11, nsp7, nsp10, nsp9, nsp8, and nsp16 regions. (Fig. 2). Therefore, these regions of SARS-CoV-2 could be considered potential targets for diagnostics, treatment, or vaccine development.

According to our results, 99.72% of Nsp11 protein worldwide (from 99.61% in Africa to 99.93% in Oceania) did not illustrate any mutation. The independent function of NSP11 has not been characterized yet; however, NSP11 contributes to the interaction between the SARS-CoV-2 and host cell membrane [98]. Kaushal et al. analyzed the rate of mutation accumulation between January 19 to April 15, 2020, in the USA SARS-CoV-2 genome. They also found that NSP11 did not accumulate any mutation [33]. Saha et al. analyzed 198 Bangladesh-originated SARS-CoV-2 genomic sequences over 13 weeks. They found that nsp11 did not accumulate any mutation [70].

In the Nsp7 protein, 98.41% of aa sequences (from 95.65% in South America to 99.38% in Oceania) did not accumulate any mutation. Nsp7 forms a supercomplex with nsp8 and nsp12 and participates in coronavirus RdRP machinery that mediates SARS-CoV-2 replication [10]. Previous studies showed that the binding site for the nsp7-nsp8 heterodimer is well conserved, and the high conservation of nsp7 and nsp8 in coronaviruses proposed that nsp7-nsp8 heterodimer is a general component for all coronaviruses [97]. These results were also found in the USA SARS-CoV-2 genome [33]. mutant nsp7 proteins are significantly associated with mutant RdRp and could change the fidelity of genome replication [65].

In our study, 97.96% of Nsp10 protein (from 97.67% in South America to 98.59% in Africa) and 95.87% of nsp9 protein (from 97.67% in South America to 98.59%

in Africa) in the world did not accumulate any mutation. NSP10 is encoded by ORF1a/1b, which comprises the RNA-synthesizing machinery of SARS-CoV2. Previous studies proved that nsp10 interacts with nsp14 and forms the NSP10–NSP14 complex, and this complex is critical for the viral replication process. Anand et al. studied nsp10 had the highest conservation thresholds, and molecular dynamics simulations revealed that the drugs Darifenacin, Nebivolol, Bictegravir, Alvimopan, and Irbesartan is targeted in this nsp. Nsp9 is a highly conserved region in beta coronaviruses and mediates viral replication, overall virulence, and viral genomic RNA reproduction. Saha et al. analyzed 198 Bangladesh-originated SARS-CoV-2 genomic sequences over 13 weeks. They found that nsp9 did not accumulate any mutation [70]. In the study of Kaushal et al. in the USA SARS-CoV-2 genome, no mutations were found [33].

In the nsp8, 95.52% of regions had no mutation (from 85.65 in South America to 98.12 in Oceania). In addition, 90.74% of Nsp16 protein in the world samples (from 61.97% in Oceania to 93.17% in Europe) did not accumulate any mutation. NSP8 is another component of the replication-transcription complex (RTC) in SARS-coronavirus [80]. In the cryoEM structure consisting of nsp7, nsp8, nsp12, and nsp13, the Interface domain is packed against nsp8 [14]. A previous study showed that SARS-CoV nsp8 was a mandatory de novo initiating RNA polymerase [80]. The viral RNA capping machinery protects mRNA from degradation by 5′ exoribonucleases, ensures efficient mRNA translation, and prevents recognition of viral RNA via innate immunity mechanisms. 2′-o-methyltransferase (2′-o-MTase) capping machinery was first discovered in the feline Coronavirus (FCoV) nsp16 [18].

In the worldwide SARS-CoV-2 samples, 97.39% of orf9c, 82.88% of Nsp6, and nsp14 76.01% had at least one, two, or three mutations. The nsp16 protein is an RNA cap modifying enzyme only active in the presence of its activating partner nsp10. In the SARS-CoV-2, Nsp16 is the leading enzymatic partner of the Nsp10-Nsp16 complex and protects from the host's innate immune recognition [45]. Identifying the 2′-O-MTase signature sequence in the SARS-CoV genome added nsp16 to the list of putative targets for antiviral drugs [77].

In our Analysis, different SARS-CoV-2 proteins harbor the maximum rate of mutations, including N, nsp3, S, nsp4, nsp12, and M. The rate of four mutations and above is depicted in Fig. 3. We found that the 10.05% of nsp4 protein in SARS-CoV-2 genomes was the rate of 4 mutations and higher (From 1.28% in South America to 15.40% in Europe). Interestingly, our result showed that the average rate of higher than four mutations in SARS-CoV-2 genomes was higher in South America (6.71%) and Europe (6.71%) in comparison to Africa (3.48%) and Oceania (3.54%). Obermeyer et al. revealed that the highest concentrations of fitness-associated mutations were found in the S, N, M, and ORF1 polyprotein genes [56]. NSP4 is the most significant membrane protein of the NSPs, with nsp6 having roles in the assembly of replication compartments. The highest concentration of fitness-associated mutations is found in nsp4, nsp6, and nsp12–14, suggesting new functions at those sites. [56]. Generally, ORF1ab is a conserved region in the SARS-CoV-2 genome [67]. Comparative sequence analysis showed highly conserved sequences within ORF1ab, particularly in nsp12–16 [79]. Robins et al. aligned and compared 149 proteins in β-coronaviruses and found that nsp12–14 are among the most highly conserved in aa identity [66].

In our study, just 2.05% of the world samples did not have any mutation in the N protein. On the other hand, 77.35% of the N protein of SARS-CoV-2 harbor four mutations and above (From 50.76% in Africa to 83.96% in Europe). Our results showed that 43.6% of the nsp3 protein of SARS-CoV-2 harbor four mutations and above (From 17.00% in Oceania to 47.06% in Europe). In the nsp3, 6.61% of worldwide samples had no mutations. NSP3, also known as papain-like protease, the largest NSP, plays a critical role in viral replication and function as a protease. Papain-like protease, polyprotein processing. Type I IFN inhibition is implicated in membrane structure formation induced upon CoV infection, with which the RTC is thought to be associated. SARS-CoV nsp3 is a 215-kDa, transmembrane, glycosylated, multi-domain protein that is a scaffolding protein for these processes [6]. Troyano-Hernáez evaluated the SARS-CoV-2 proteome in Spain and realized that the nsp3 protein has the highest number of deletions and stop codons. However, The PLpro central catalytic residues were highly conserved [82]. Based on fooladinezhad, results, in North America NSP3 data, 41.47% of aa showed more than four mutations in their sequence [22]. The region corresponding to the C-terminal domain of SARS-CoV NSP3 was found to be significantly less mutated likely due to its vital role in inducing the formation of [2] double-membrane vesicles [6]. The most significant number of mutations was located within the gene encoding for the Nsp3 protein (20.7%), followed by the gene encoding for the spike protein (14.6%) [46].

In the S protein, we found that 65.37% of S protein had at least one, two, or three mutations, and 29.79% of S protein in SARS-CoV-2 genomes have a rate of four mutations and higher (From 3.33% in Oceana to 80.52% in South America).

Conservation in the SARS-CoV-2 genome

We divided each SARS-CoV-2 gene into ten parts, and then we explored the frequency of mutations in each

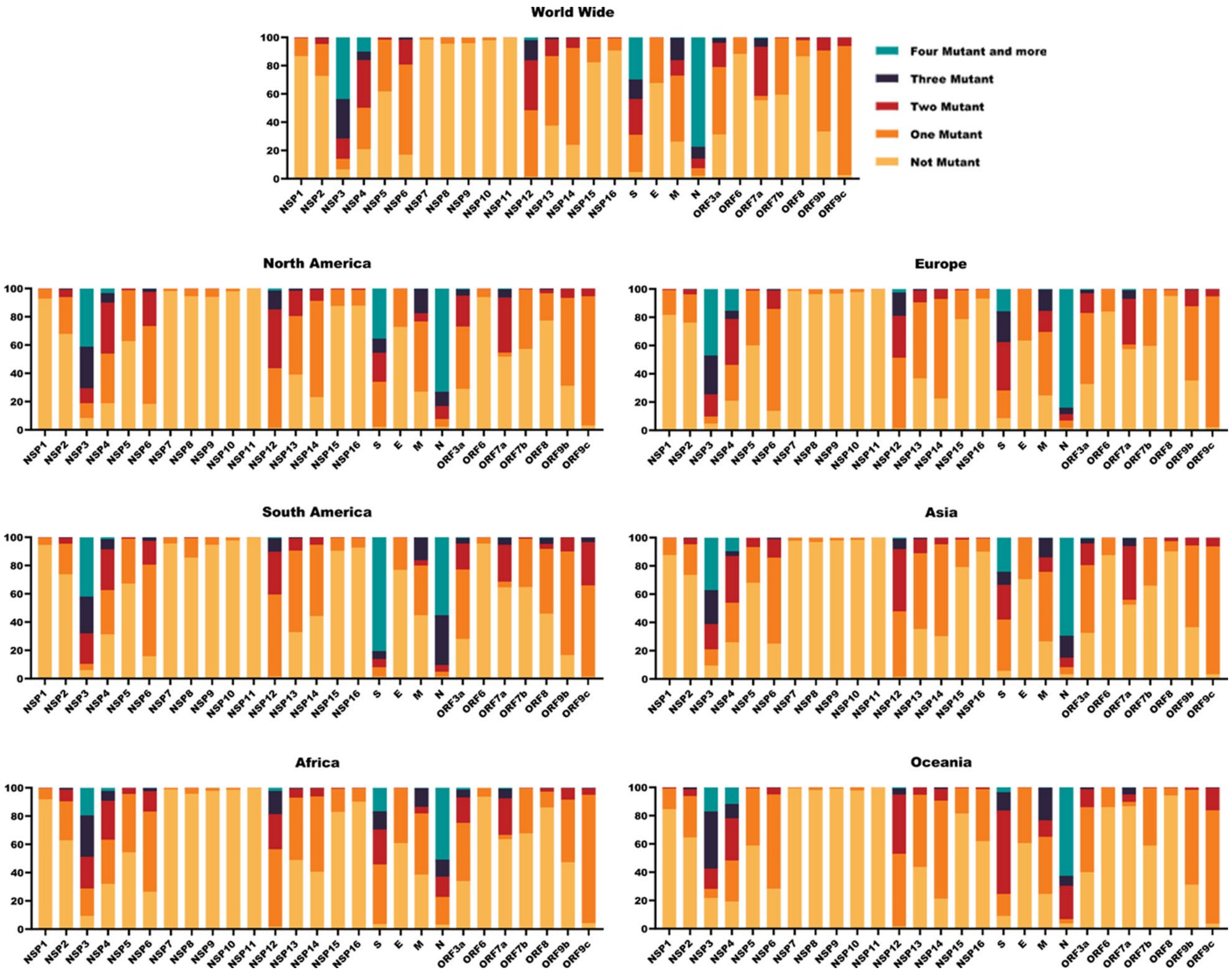

Fig. 3 Mutation rate of SARS-CoV-2 genome in different geographic areas. The graph reports SARS-CoV-2 frequency of one, two, three, and four more mutations in the SARS-CoV-2 genome by April 28, 2022, in seven geographic areas

part. In our study, comparative genome analysis between the SARS-CoV-2 across Asia, North and South America, Europe, Africa, and Oceania revealed that, on average, the frequency of mutations in NSP10, NSP7, NSP8, NSP11, NSP16, ORF6, NSP9 is lower than 0.040% (Fig. 4). Conversely, the average frequency of mutations was high in ORF9c (1.48%), ORF9b (0.84%), ORF7a (0.74%), and ORF7b (0.59%) (Fig. 4).

To date, different types of tests for SARS-CoV-2 detection-based nucleic acid testing (NAT) have developed, including biosensor chips [3], nanopore targeted sequencing (NTS) [99, 100], multiplex real-time reverse transcription–polymerase chain reaction (rRT-PCR) [30] and ATR-FTIR Spectral Analyses [7]. S protein is considered an essential target for a diagnostic test of SARS-CoV-2, which was evaluated in SARS-CoV and MERS-CoV. In the first months of the pandemic majority of primer/probe sets were designed based on the Wuhan-Hu-1, NC_045512.2 sequences. The United States Food and Drug Administration (USFDA) has authorized 277 SARS-CoV-2 molecular diagnostic tests. Reverse transcriptase polymerase chain reaction (RT-PCR) is the main molecular tests used for SARS-CoV-2. Because of the diverse mutations in the SARS-CoV-2 genome, genetic alternation in the primer binding sites and the probe recognition sites may affect the efficiency of RT-PCR-based detection of COVID-19. Different regions of the SARS-CoV-2 genome, such as the RdRp, S, N,or E genes, are common targets for SARS-CoV-2 by PCR assays [17, 83].

The average mutation frequency in the structural proteins was 0.027%, 0.045%, 0.53% and 0.088% in S, E, M, and N, respectively (Fig. 5). Furthermore, we have tested whether the frequency of mutations throughout the

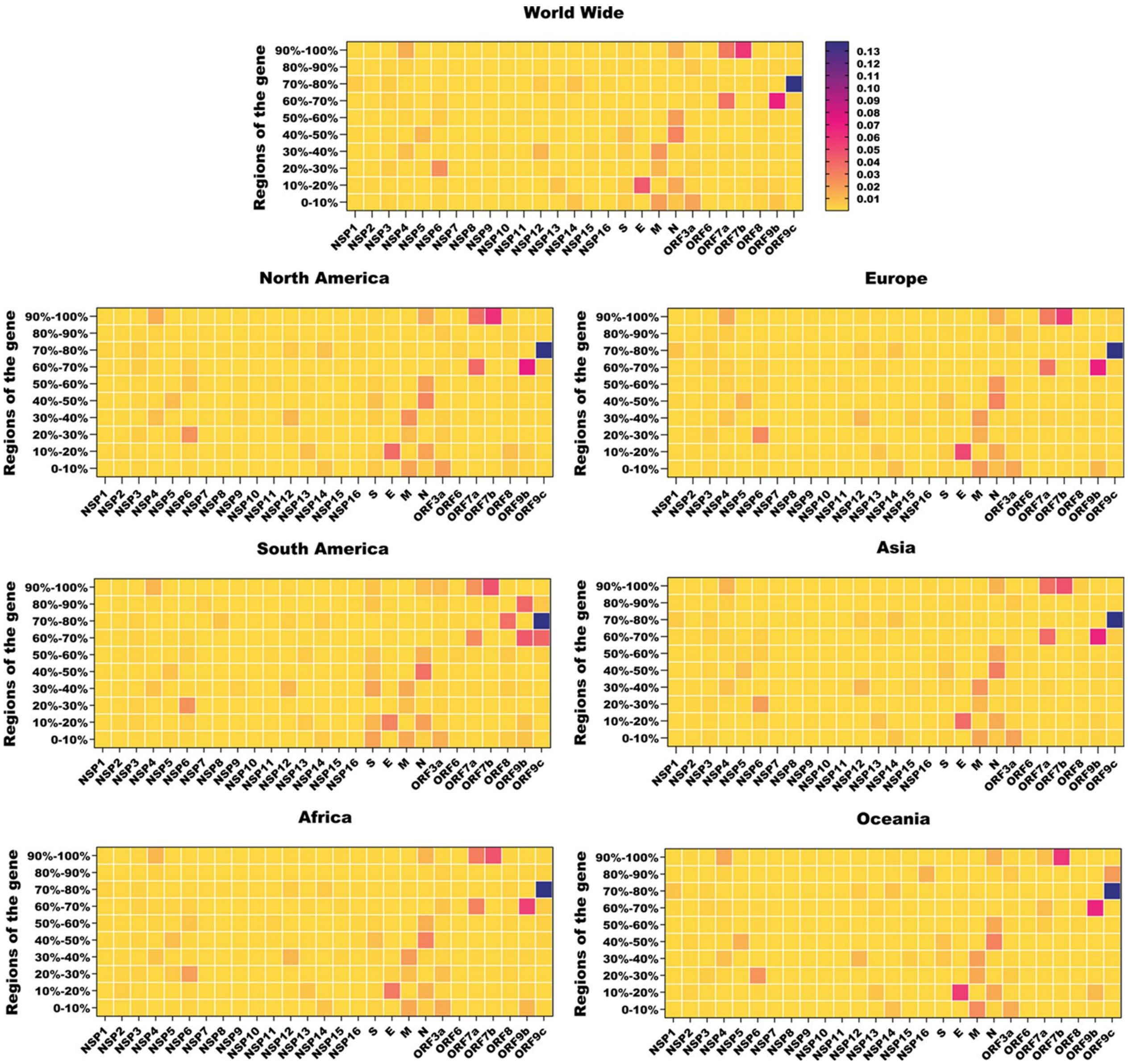

Fig. 4 Heat maps of conserved genomic regions of SARS-CoV-2. SARS-CoV-2 genomes are divided into ten regions, and the frequency of mutations is in different regions worldwide. *Nsp* non-structural protein, *S* Spike protein, *E* Envelope protein, *M* Membrane protein, *N* Nucleocapsid protein

different regions of S, E, M, N and nsp12 can cluster in different areas including Asia, North and South America, Europe, Africa, and Oceania (Fig. 5).

Conserved regions of the S protein are shown in Fig. 5A which aa 254–381 (0.039%) and aa 726–1016 (0.042%) are conserved. No common mutations were observed in these regions. The regions with the highest mutation frequency in the S were aa 508–635(0.77%) and aa 381–508 (0.43%) Hierarchical clustering results classified the spike protein into the two major clusters. The cluster two sub-classified into two other clusters that one of them contained aa 508–635 that enriched with D614G mutation (Fig. 5A). S protein of SARS-CoV-2 is required to attach and fuse into the host cells to initiate infection. It is the primary target of neutralizing antibodies. The S protein of SARS-CoV-2 is the design basis of different generations of COVID-19 vaccines (Pfizer/BioNTech and Moderna's). The S protein of SARS-CoV-2 and SARS-CoV share a 76% similarity in the amino acid sequence [14, 15]. The RBD region targets neutralizing antibodies (nAbs), and mutations in the RBD are present in SARS-CoV-2 variants of concern [78]. Previous

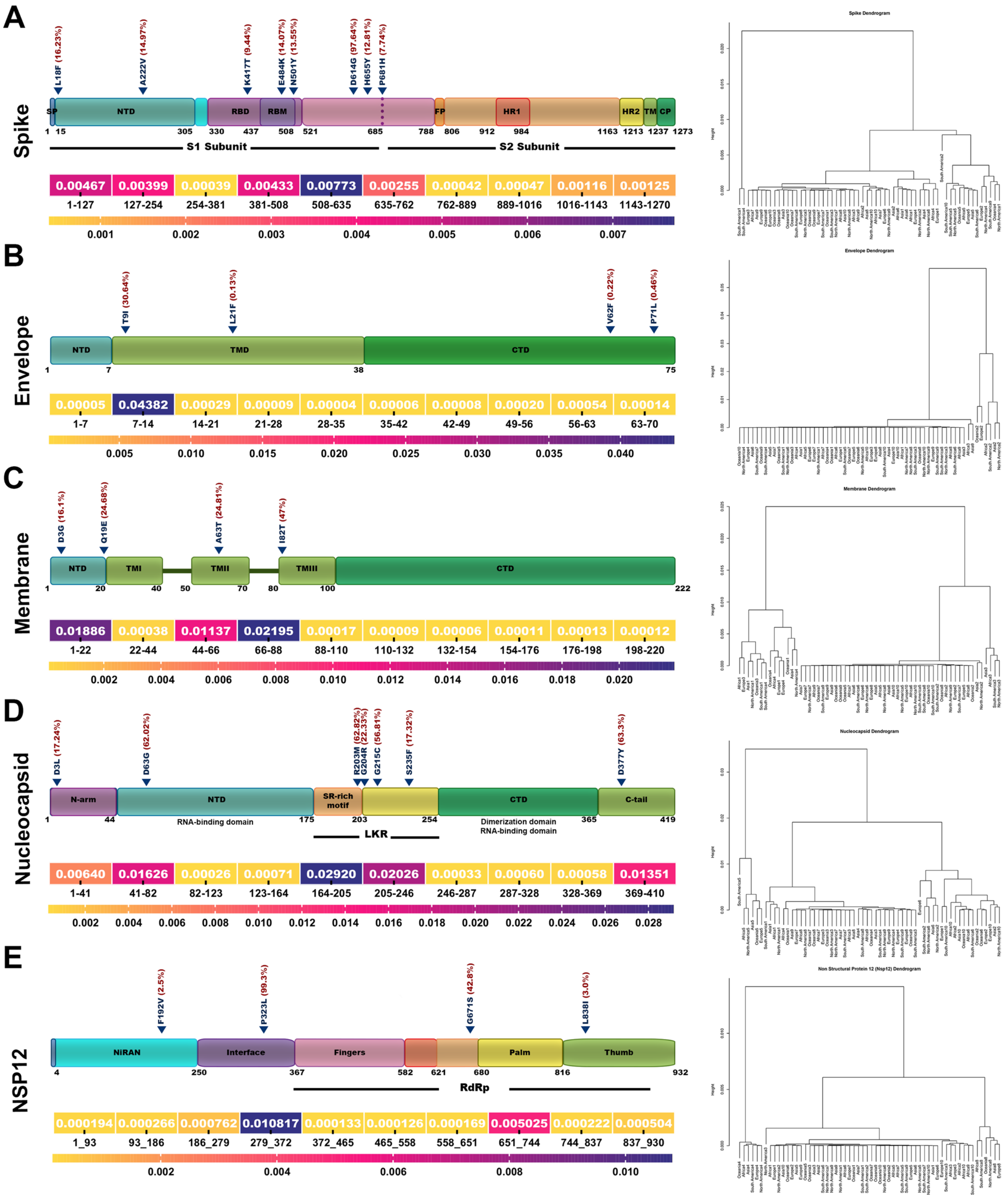

Fig. 5 Clustering analysis of SARS-CoV-2 proteins. Heat map and dendrogram illustration of SARS-CoV-2 proteins based on the frequency of mutations in ten different regions of SARS-CoV-2 genomes. **A** Spike. **B** Envelope. **C** Membrane, **D** Nucleocapsid and **E** NSP12. NTD, N-terminal domain; RBD, receptor-binding domain; FP, fusion peptide; HR, heptapeptide repeat sequence; TM, transmembrane; CT, cytoplasmic tail. LKR, serine-arginine (SR) rich-linker region; RBM, receptor binding motif; CP, cytoplasm domain; TMD α-helical transmembrane domain

studies suggested that RBD of CoV in the highly mutable region may not be an ideal drug target [29, 94]. A receptor-binding domain (RBD, 319–541 residues) recognizes the receptor ACE2 specifically. RBD is a critical target for antiviral compounds and antibodies. The most non-conserved region in the S was aa 508–635. Interestingly, region 603–634 of the S protein of SARS is a major immunodominant epitope in S protein [26]. The C662–C671 epitope has also been targeted by neutralizing antibodies [41, 42]. S protein is the well-studied structural protein in the SARS-CoV-2 that mediates human ACE2 receptor binding and is responsible for entry into a cell and endosomal escape. Numerous studies have demonstrated the high frequency of mutations in the spike SARS-CoV2 as variants of concern (VOC) [48]. The study of van Dorp et al. revealed that about 80% of SARS-CoV-2 genome mutations occur in the spike protein [85]. S protein has the highest mean aa change/deletion frequency per sequence in the study of Troyano-Hernáez [82].

In our study, aa 7–14 had the highest frequency of mutation (4.38%) in the E gene. T9I (30.6%) was the most prevalent mutation in the TMD region of the E protein. Hierarchical cluster analysis classified the E protein into two major clusters. The cluster 2 contained aa 7–14 that enriched with T9I mutation (Fig. 5B). SARS-CoV-2 E protein is 228 nucleotides long and has a variety of functions such as viral assembly, replication, propagation, and pathogenesis [64].

In the M protein, the highest mutation frequency was observed in aa 66–88 (2.19%) and aa 1–22 (1.88%). The most common mutations found in this region were I82T (47%), D3G (16.1%), and Q19E (24.68%). M protein classified in two major clusters through a hierarchical clustering. The cluster one sub-classified into two other clusters that enriched aa 1–22 and aa 66–88 that contained D3G, Q19E, and I82T (Fig. 5C). M protein comprises 223 amino acids and performs various functions, including virion formation and assembly. A previous study declared that M protein is conserved across ß-coronaviruses [9]. The sequence identities and sequence similarities between the M proteins sequence of SARS-CoV-2 and SARS-CoV were 90.5% and 96.40%, respectively [50].

In the N protein, the highest frequency of mutation was reported in aa 164–246 (2.92%), aa 205–246 (2.02%), and aa 41–48 (1.62%). These regions harbor several common mutations such as R203M (62.82%), D63G(62.02%), G215C (56.81%), G204R (22.33%), and S223F (17.32%). The result of hierarchical cluster analysis revealed that N protein classified into the three major clusters. The cluster 1 contained aa164-205 that enriched with R203M and G204R mutations (Fig. 5D). The N protein of SARS-CoV-2 has 419 amino acids, with 45.6 kDa positively charged unstable hydrophobic protein, and plays a pivotal role in transcription and replication; interaction with M during viral assembly, N protein is involved in the packing of RNA, the release of virus particles, and the formation of the ribonucleoprotein core. Yu et al. analyzed 5,167,111 N proteins and reported low mutation rates in their amino acid sequences. [96]. In the study of Troyano-Hernáez et al., the mean aa change/deletion frequency per sequence of N was 3.79 [82]. N gene is one of the most non-conserved genes in the SARS-CoV-2 [78]. In the N protein, aa 164–246 was the non-conserved region that harbored common mutations such as R203M, G204R, G215C, and S223F. This region is located in the serine-arginine (SR) rich-linker region (LKR) (aa 175–254) that forms a phosphorylation-dependent binding domain and is responsible for oligomerization, phosphoregulation and RNA and protein binding [13, 91, 93]. It has been reported that the SR-linker was the most variable region within the N protein [81]. It has been reported that R203M, G204R, G215C, and S223F mutations in this region could have an important biological impact and increase the infectivity, fitness, and pathogenicity of SARS-CoV-2 [93]. The N gene of SARS-CoV-2 is the target of many diagnostic assays for COVID-19. Miller et al. reported that two point mutations in the N gene, a C to T mutation at position 29197 and a C to T mutation at position 29200, negatively impact the SARS-CoV-2 detection by the Cepheid Xpert Xpress SARS-CoV-2 assay [52]. N gene mutation C29200T was also reported in Hasan et al. (84) study. The C29197T mutation results in N-gene target failure in the Xpert Omni SARS-CoV-2 assays from Cepheid (Sunnyvale, CA) [47].

In the NSP12 protein, the highest mutation frequency was observed in aa 279–372. The most common mutation P323L (99.3%) in SARS-CoV-2 genomes was observed in this region. The hierarchical cluster analysis classified NSP12 into two major clusters. The cluster 1 contained aa 279–372 that enriched with P323L mutations (Fig. 5E). The SARS-CoV-2 NSP12 is a key component of the viral replication [28]. P323L is a key mutation in the NSP12 and has been associated with a high mutation rate and severity of COVID-19 [11, 58].

Effect of common SARS-CoV-2 mutations on protein stability

Understanding the stability changes in SARS-CoV-2 proteins is essential for predicting virus infectivity. Furthermore, it could be generating novel insights for updating the COVID vaccines. Changes in Gibbs free energy of unfolding (ΔΔG) between the wild type and mutant protein could predict the effect of the stability of protein structure [59].

In the current study, structured-based analysis results by Dynamut2 and MAESTROweb indicate that the P323L mutation decreases the stability of NSP12. While DynaMut and SDM results showed that, this is a stabilizing mutation (Fig. 6D, Table 2). The topology-based mutation predictor (TML-MP) in the study of Wang et al. suggested that P323L destabilizes the NSP12 [88]. On the other hand, the study of Kim et al. MicroScale Thermophoresis analysis revealed that the NSP12 P323L mutation stabilized the NSP12-NSP7-NSP8 complex interaction [37]. Periwal et al. also suggesetd that the P323L had a stabilizing effect relative to the wild type protein [61].

Our results indicate that all the methods predicted that the Spike D614G and nucleoprotein D63G are stabilizing mutations (Fig. 6A, C, Table 2). The results of Gellenoncourt et al. suggested that the D614G mutation stabilized S1/S2 association and enabled the selection of mutations that increased S1/S2 cleavage [24]. According to DynaMut estimations in the study of Chakraborty et al. D614G is stable mutation [12]. In the study of Aljindan et al., CUPSAT, SDM 2.0,and DUET analysis tools predict

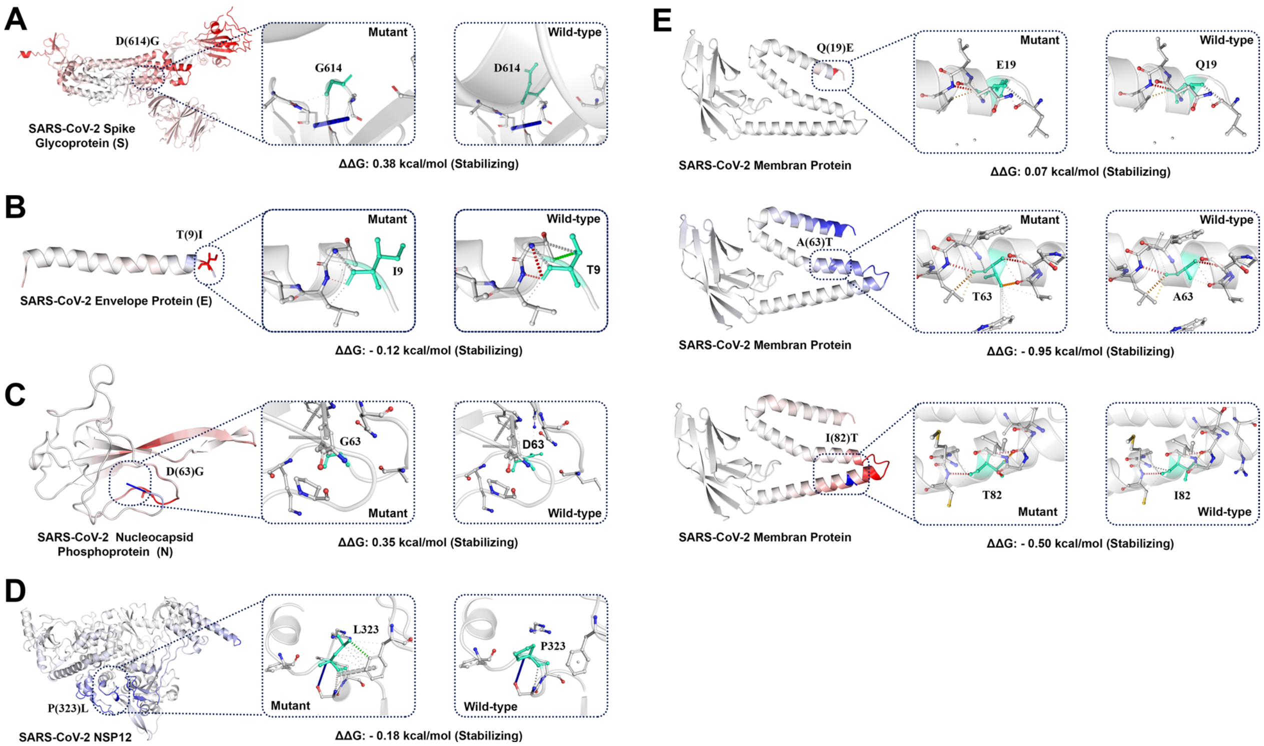
A
D(614)G
Mutant
G614
Wild-type
D614
SARS-CoV-2 Spike Glycoprotein (S)
ΔΔG: 0.38 kcal/mol (Stabilizing)
B
T(9)I
Mutant
I9
Wild-type
T9
SARS-CoV-2 Envelope Protein (E)
ΔΔG: - 0.12 kcal/mol (Stabilizing)
C
G63
D(63)G
D63
Mutant
Wild-type
SARS-CoV-2 Nucleocapsid Phosphoprotein (N)
ΔΔG: 0.35 kcal/mol (Stabilizing)
D
L323
P323
P(323)L
Mutant
Wild-type
SARS-CoV-2 NSP12
ΔΔG: - 0.18 kcal/mol (Stabilizing)
E
Q(19)E
Mutant
E19
Wild-type
Q19
SARS-CoV-2 Membran Protein
ΔΔG: 0.07 kcal/mol (Stabilizing)
A(63)T
Mutant
T63
Wild-type
A63
SARS-CoV-2 Membran Protein
ΔΔG: - 0.95 kcal/mol (Stabilizing)
I(82)T
T82
Mutant
I82
Wild-type
SARS-CoV-2 Membran Protein
ΔΔG: - 0.50 kcal/mol (Stabilizing)

Fig. 6 Dynamut prediction of molecular flexibility and destabilizing effect of common SARS-CoV-2 mutations. The protein rigidification and structural flexibility are highlighted in blue and red color, respectively. Light green represents wild-type and mutant residues of protein**. A** Spike mutation D614G**. B** Envelope mutation T9I**. C** Nucleoprotein mutation D63G. **D** NSP12 mutation 9323L**. E** Membrane mutations A63T, I82T and Q19E

Table 2 Predicted output from DynaMut, Dynamut2, SDM and MAESTROweb for the stability of SARS-CoV-2 proteins

Protein	Mutation	Protein structure	Dynamut	Dynamut2	SDM	MAESTROweb
NSP12	P323L	7C2K	1.532	– 0.18	1.18	– 0.082
Spike	D614G	6VXX	0.299	0.38	2.5	0.101
Envelope	T9I	7K3G	0.214	– 0.12	1.1	2.902
Membrane	A63T	8CTK	0.050	– 0.95	– 1.53	1.351
Membrane	I82T	8CTK	– 0.460	– 0.5	– 1.5	0.485
Membrane	Q19E	8CTK	– 0.120	0.07	0.4	– 0.481
Nucleoprotein	D63G	6VYO	0.007	0.35	0.87	1.101

increasing stability of the spike protein [4]. Based on the results of FoldX and ROSETTA outputs, Mahmoudi Gomari et al. suggested that the D614G mutation increases the stability of spike protein [49]. Plante et al. measured the decay of infectivity of D614 and G614 viruses over different times and found out the D614G mutation may increase the stability of SARS-CoV-2 [62].

For the membrane protein, Dynamut2 and SDM predict membrane Q19E mutation as a stabilizing mutation. DynaMut and MAESTROweb analysis showed Q19E mutation destabilized structure of the membrane protein. In addition, Dynamut2 and SDM results showed that membrane A63T is destabilizing mutation, while Dynamut and MAESTROweb results indicate that, these are stabilizing mutations. Membrane I82T mutation destabilize protein structure according to Dynamut, Dynamut2 and SDM analysis, however MAESTROweb results showed that this mutation stable membrane structure. (Fig. 6E, Table 2). The Omicron variant of concern is the most mutated SARS-CoV-2 variant (N = 65, including 16 deletions and 3 insertions) and is characterized by several mutations in a membrane such as Q19E, and A63T [5].

Dynamut, SDM and MAESTROweb analysis results indicate that envelope T9I mutation stabilized envelope structure; however, Dynamut2 results showed decreased stability of the encoded proteins (Fig. 6C, Table 2). The T9I is one of the envelope mutations in the Omicron variant, and the DynaMut prediction outcome revealed that this mutation had a stabilizing effect [55, 72]. In support of our in silico findings, the validation of key results using experimental approaches is warranted the impact of changes in protein sequence on protein stability.

Conclusion

COVID-19 is one of the most significant global health catastrophes, causing more than 6.56 million deaths globally. SARS-CoV-2 has a relatively high dynamic mutation rate, and large-scale genome-sequencing efforts have provided a pattern for the global spread and diversification of SARS-CoV-2. Thanks to the GISAID database, we could access viral genomes from all over the world in the study.

Early diagnosis of SARS-CoV-2 infection is essential for controlling and treating COVID-19 patients. WHO recommends RT-PCR and other NAT assays, which are widely applied in different countries. However, these methods' false-negative results are a significant challenge to controlling the pandemic. Conserved regions in the SARS-CoV-2 genome could be considered promising targets for diagnostic tools and strengthen the detection sensitivity to reduce false-negative results.

Vaccinations resulted in a decline in the risk of COVID-19 infection and hospitalizations worldwide. However, numerous studies have demonstrated that the efficacy of vaccines against infection decreases over time. Therefore, molecular surveillance programs are critical to guide the development of vaccines based on molecular change and novel emerging SARS-CoV-2 variants.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-023-03996-w.

Additional file 1. The mutation frequency of genes in samples grouped by continents.

Additional file 2. Each gene's length is divided into ten sections, and the frequency of mutations in each region is categorized by continent. These data facilitate the segregation of protected regions from susceptible ones.

Acknowledgements

The authors thank all of the researchers who have shared genome data openly via the Global Initiative on Sharing All Influenza Data (GISAID).

Author contributions

KR, MM, MMS, ST, and BMo contributed to data collection; YD, MHA, BMa, and MM, contributed to study design; KR and MM designed workflow, code and data analysis; BMa, K.R, and MM contributed to data visualization; MHA and BMa wrote the manuscript; ST and MM corrected the manuscript and provided useful comments; MM, KR, and BMa monitored the accuracy of additional data; BMa designed graphical contents; YD have edited and supervised the work. All authors read and approved the final manuscript.

Funding

This work was supported by the NIH Grants; (2U54CA143727, 5P30GM114737, 5P20GM103466, 5U54MD007601, 5P30CA071789, P20GM139753, and U54GM138062).

Availability of data and materials

The dataset(s) supporting the conclusions of this article are available in the SARS2Mutant repository, [http://sars2mutant.com/].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 6 December 2022 Accepted: 15 February 2023
Published online: 25 February 2023

References

1. WHO. Coronavirus disease (COVID-19) pandemic, 2022. https://www.who.int/emergencies/diseases/novelcoronavirus-2019.
2. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5(1):1–9.
3. Alafeef M, Dighe K, Moitra P, Pan D. Rapid, ultrasensitive, and quantitative detection of SARS-CoV-2 using antisense oligonucleotides directed electrochemical biosensor chip. ACS Nano. 2020;14(12):17028–45.

4. Aljindan RY, Al-Subaie AM, Al-Ohali AI, Kamaraj B. Investigation of nonsynonymous mutations in the spike protein of SARS-CoV-2 and its interaction with the ACE2 receptor by molecular docking and MM/GBSA approach. Comput Biol Med. 2021;135:104654.
5. Alkhatib M, Salpini R, Carioti L, Ambrosio FA, D'Anna S, Duca L, Costa G, Bellocchi MC, Piermatteo L, Artese A. Update on SARS-CoV-2 Omicron Variant of Concern and Its Peculiar Mutational Profile. Microbiol Spectr. 2022;10(2):e02732-e2721.
6. Angelini MM, Akhlaghpour M, Neuman BW, Buchmeier MJ. Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles. MBio. 2013;4(4):e00524-e513.
7. Barauna VG, Singh MN, Barbosa LL, Marcarini WD, Vassallo PF, Mill JG, Ribeiro-Rodrigues R, Campos LC, Warnke PH, Martin FL. Ultrarapid on-site detection of SARS-CoV-2 infection using simple ATR-FTIR spectroscopy and an analysis algorithm: high sensitivity and specificity. Anal Chem. 2021;93(5):2950–8.
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res. 2000;28(1):235–42.
9. Bianchi M, Benvenuto D, Giovanetti M, Angeletti S, Ciccozzi M, Pascarella S. Sars-CoV-2 envelope and membrane proteins: structural differences linked to virus characteristics? BioMed Res Int. 2020. https://doi.org/10.1155/2020/4389089.
10. Biswal M, Diggs S, Xu D, Khudaverdyan N, Lu J, Fang J, Blaha G, Hai R, Song J. Two conserved oligomer interfaces of NSP7 and NSP8 underpin the dynamic assembly of SARS-CoV-2 RdRP. Nucleic Acids Res. 2021;49(10):5956–66.
11. Biswas SK, Mudi SR. Spike protein D614G and RdRp P323L: the SARS-CoV-2 mutations associated with severity of COVID-19. Genom Inform. 2020;18(4):e44.
12. Chakraborty C, Bhattacharya M, Sharma AR, Mallik B. Omicron (B. 1.1. 529)-a new heavily mutated variant: mapped location and probable properties of its mutations with an emphasis on S-glycoprotein. Int J Biol Macromol. 2022;219:980–97.
13. Chang C-K, Chen C-MM, Chiang M-H, Hsu Y-L, Huang T-H. Transient oligomerization of the SARS-CoV N protein–implication for virus ribonucleoprotein packaging. PLoS ONE. 2013;8(5):e65045.
14. Chen J, Malone B, Llewellyn E, Grasso M, Shelton PM, Olinares PDB, Maruthi K, Eng ET, Vatandaslar H, Chait BT. Structural basis for helicase-polymerase coupling in the SARS-CoV-2 replication-transcription complex. Cell. 2020;182(6):1560-1573. e1513.
15. Chen Y, Liu Q, Guo D. Emerging coronaviruses: genome structure, replication, and pathogenesis. J Med Virol. 2020;92(4):418–23.
16. Chen Z, Azman AS, Chen X, Zou J, Tian Y, Sun R, Xu X, Wu Y, Lu W, Ge S. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. Nat Genet. 2022;54(4):499–507.
17. Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, Bleicker T, Brünink S, Schneider J, Schmidt ML. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Eurosurveillance. 2020;25(3):2000045.
18. Decroly E, Imbert I, Coutard B, Bouvet M, Selisko B, Alvarez K, Gorbalenya AE, Snijder EJ, Canard B. Coronavirus nonstructural protein 16 is a cap-0 binding enzyme possessing (nucleoside-2′ O)-methyltransferase activity. J Virol. 2008;82(16):8071–84.
19. DeGrace MM, Ghedin E, Frieman MB, Krammer F, Grifoni A, Alisoltani A, Alter G, Amara RR, Baric RS, Barouch DH. Defining the risk of SARS-CoV-2 variants on immune protection. Nature. 2022;605(7911):640–52.
20. Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X, Scherbakova S, Graham RL, Baric RS, Stockwell TB. Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. PLoS Pathog. 2010;6(5): e1000896.
21. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Chall. 2017;1(1):33–46.
22. Fooladinezhad H, Shahidi M, Mahmanzar M, Mahdavi B, Tokhanbigli S, Sisakht MM, Moradi B, Ganjalikhany MR, Rahimian K, Ganjalikhani-Hakemi M. SARS-CoV-2 NSP3, NSP4 and NSP6 mutations and epistasis during the pandemic in the world: evolutionary trends and natural selections in six continents. medRxiv. 2022;5:536.
23. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, Wang T, Sun Q, Ming Z, Zhang L. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science. 2020;368(6492):779–82.
24. Gellenoncourt S, Saunders N, Robinot R, Auguste L, Rajah MM, Kervevan J, Jeger-Madiot R, Staropoli I, Planchais C, Mouquet H. The spike-stabilizing D614G mutation interacts with S1/S2 cleavage site mutations to promote the infectious potential of SARS-CoV-2 variants. bioRxiv. 2022. https://doi.org/10.1128/jvi.01301-22.
25. Guo G, Ye L, Pan K, Chen Y, Xing D, Yan K, Chen Z, Ding N, Li W, Huang H. New insights of emerging SARS-CoV-2: epidemiology, etiology, clinical features, clinical treatment, and prevention. Front Cell Dev Biol. 2020;8:410.
26. He Y, Zhou Y, Wu H, Luo B, Chen J, Li W, Jiang S. Identification of immunodominant sites on the spike protein of severe acute respiratory syndrome (SARS) coronavirus: implication for developing SARS diagnostics and vaccines. J Immunol. 2004;173(6):4050–7.
27. Herlocher ML, Elias S, Truscon R, Harrison S, Mindell D, Simon C, Monto AS. Ferrets as a transmission model for influenza: sequence changes in HA1 of type A (H3N2) virus. J Infect Dis. 2001;184(5):542–6.
28. Hillen HS, Kokic G, Farnung L, Dienemann C, Tegunov D, Cramer P. Structure of replicating SARS-CoV-2 polymerase. Nature. 2020;584(7819):154–6.
29. Huang Y, Yang C, Xu X-F, Xu W, Liu S-W. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol Sin. 2020;41(9):1141–9.
30. Ishige T, Murata S, Taniguchi T, Miyabe A, Kitamura K, Kawasaki K, Nishimura M, Igari H, Matsushita K. Highly sensitive detection of SARS-CoV-2 RNA by multiplex rRT-PCR for molecular diagnosis of COVID-19 by clinical laboratories. Clin Chim Acta. 2020;507:139–42.
31. Jiang S, Hillyer C, Du L. Neutralizing antibodies against SARS-CoV-2 and other human coronaviruses. Trends Immunol. 2020;41(5):355–9.
32. Jungreis I, Sealfon R, Kellis M. SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. Nat Commun. 2021;12(1):1–20.
33. Kaushal N, Gupta Y, Goyal M, Khaiboullina SF, Baranwal M, Verma SC. Mutational frequencies of SARS-CoV-2 genome during the beginning months of the outbreak in USA. Pathogens. 2020;9(7):565.
34. Keep S, Oade MS, Lidzbarski-Silvestre F, Bentley K, Stevenson-Leggett P, Freimanis GL, Tennakoon C, Sanderson N, Hammond JA, Jones RC. Multiple novel non-canonically transcribed sub-genomic mRNAs produced by avian coronavirus infectious bronchitis virus. J Gen Virol. 2020;101(10):1103.
35. Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, Akite N, Ho J, Lee RT, Yeo W. GISAID's role in pandemic response. China CDC Weekly. 2021;3(49):1049.
36. Kim C, Mahasenan KV, Bhardwaj A, Wiest O, Chang M, Mobashery S. Production of proteins of the SARS-CoV-2 proteome for drug discovery. ACS Omega. 2021;6(30):19983–94.
37. Kim, S.-M., E.-H. Kim, M. A. B. Casel, Y.-I. Kim, R. Sun, M.-J. Kwack, J.-S. Yoo, M.-A. Yu, K.-M. Yu and S.-G. Jang. SARS-CoV-2 variants show temperature-dependent enhanced polymerase activity in the upper respiratory tract and high transmissibility. bioRxiv: 2022.2009. 2027.509689. 2002.
38. Kirchdoerfer RN, Ward AB. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. Nat Commun. 2019;10(1):1–9.
39. Laimer J, Hiebl-Flach J, Lengauer D, Lackner P. MAESTROweb: a web server for structure-based protein stability prediction. Bioinformatics. 2016;32(9):1414–6.
40. Lei J, Kusov Y, Hilgenfeld R. Nsp3 of coronaviruses: structures and functions of a large multi-domain protein. Antiviral Res. 2018;149:58–74.
41. Li X, Wang W, Zhao X, Zai J, Zhao Q, Li Y, Chaillon A. Transmission dynamics and evolutionary history of 2019-nCoV. J Med Virol. 2020;92(5):501–11.
42. Li Y, Lai D-Y, Zhang H-N, Jiang H-W, Tian X, Ma M-L, Qi H, Meng Q-F, Guo S-J, Wu Y. Linear epitopes of SARS-CoV-2 spike protein elicit neutralizing antibodies in COVID-19 patients. Cell Mol Immunol. 2020;17(10):1095–7.
43. Liang Y, Wang M-L, Chien C-S, Yarmishyn AA, Yang Y-P, Lai W-Y, Luo Y-H, Lin Y-T, Chen Y-J, Chang P-C. Highlight of immune pathogenic response and hematopathologic effect in SARS-CoV, MERS-CoV, and SARS-Cov-2 infection. Front Immunol. 2020;11:1022.
44. Lin S, Chen H, Chen Z, Yang F, Ye F, Zheng Y, Yang J, Lin X, Sun H, Wang L. Crystal structure of SARS-CoV-2 nsp10 bound to nsp14-ExoN domain reveals an exoribonuclease with both structural and functional integrity. Nucleic Acids Res. 2021;49(9):5382–92.

45. Lin S, Chen H, Ye F, Chen Z, Yang F, Zheng Y, Cao Y, Qiao J, Yang S, Lu G. Crystal structure of SARS-CoV-2 nsp10/nsp16 2′-O-methylase and its implication on antiviral drug design. Signal Transduct Target Ther. 2020;5(1):1–4.
46. Lippi, G. and B. M. Henry. The landscape of SARS-CoV-2 genomic mutations. 2002.
47. Lopez D, Roberts J, Bourgeois M, Kootstra J, Minnick S, Black A, Mauss J, Flores N. Infection clusters can elevate risk of diagnostic target failure for detection of SARS-CoV-2. PLoS ONE. 2022;17(2):e0264008.
48. Magazine N, Zhang T, Wu Y, McGee MC, Veggiani G, Huang W. Mutations and evolution of the SARS-CoV-2 spike protein. Viruses. 2022;14(3):640.
49. Mahmoudi Gomari M, Rostami N, Omidi-Ardali H, Arab SS. Insight into molecular characteristics of SARS-CoV-2 spike protein following D614G point mutation, a molecular dynamics study. J Biomol Struct Dyn. 2022;40(12):5634–42.
50. Mahtarin R, Islam S, Islam MJ, Ullah MO, Ali MA, Halim MA. Structure and dynamics of membrane protein in SARS-CoV-2. J Biomol Struct Dyn. 2022;40(10):4725–38.
51. Mansky LM. Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line. AIDS Res Hum Retroviruses. 1996;12(4):307–14.
52. Miller S, Lee T, Merritt A, Pryce T, Levy A, Speers D. Single-point mutations in the N gene of SARS-CoV-2 adversely impact detection by a commercial dual target diagnostic assay. Microbiol Spectr. 2021;9(3):e01494-e1421.
53. Minskaia E, Hertzig T, Gorbalenya AE, Campanacci V, Cambillau C, Canard B, Ziebuhr J. Discovery of an RNA virus 3′→ 5′ exoribonuclease that is critically involved in coronavirus RNA synthesis. Proc Natl Acad Sci. 2006;103(13):5108–13.
54. Motayo BO, Oluwasemowo OO, Olusola BA, Akinduti PA, Arege OT, Obafemi YD, Faneye AO, Isibor PO, Aworunse OS, Oranusi SU. Evolution and genetic diversity of SARS-CoV-2 in Africa using whole genome sequences. Int J Infect Dis. 2021;103:282–7.
55. Mou K, Abdalla M, Wei DQ, Khan MT, Lodhi MS, Darwish DB, Sharaf M, Tu X. Emerging mutations in envelope protein of SARS-CoV-2 and their effect on thermodynamic properties. Inform Med Unlocked. 2021;25:100675.
56. Obermeyer F, Jankowiak M, Barkas N, Schaffner SF, Pyle JD, Yurkovetskiy L, Bosso M, Park DJ, Babadi M, MacInnis BL. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. Science. 2022;376(6599):1327–32.
57. Otto SP, Day T, Arino J, Colijn C, Dushoff J, Li M, Mechai S, Van Domselaar G, Wu J, Earn DJ. The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. Curr Biol. 2021;31(14):R918–29.
58. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo RC. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med. 2020;18:1–9.
59. Pan Q, Nguyen TB, Ascher DB, Pires DE. Systematic evaluation of computational tools to predict the effects of mutations on protein stability in the absence of experimental structures. Brief Bioinform. 2022. https://doi.org/10.1093/bib/bbac025.
60. Pandurangan AP, Ochoa-Montaño B, Ascher DB, Blundell TL. SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res. 2017;45(W1):W229–35.
61. Periwal N, Rathod SB, Sarma S, Johar GS, Jain A, Barnwal RP, Srivastava KR, Kaur B, Arora P, Sood V. Time series analysis of SARS-CoV-2 genomes and correlations among highly prevalent mutations. Microbiol Spectr. 2022;10(5):e01219-01222.
62. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, Zhang X, Muruato AE, Zou J, Fontes-Garfias CR. Spike mutation D614G alters SARS-CoV-2 fitness. Nature. 2021;592(7852):116–21.
63. Rabi FA, Al Zoubi MS, Kasasbeh GA, Salameh DM, Al-Nasser AD. SARS-CoV-2 and coronavirus disease 2019: what we know so far. Pathogens. 2020;9(3):231.
64. Rahman MS, Hoque MN, Islam MR, Islam I, Mishu ID, Rahaman MM, Sultana M, Hossain MA. Mutational insights into the envelope protein of SARS-CoV-2. Gene Rep. 2021;22: 100997.
65. Reshamwala SM, Likhite V, Degani MS, Deb SS, Noronha SB. Mutations in SARS-CoV-2 nsp7 and nsp8 proteins and their predicted impact on replication/transcription complex structure. J Med Virol. 2021;93(7):4616–9.
66. Robins WP, Mekalanos JJ. Covariance predicts conserved protein residue interactions important to the emergence and continued evolution of SARS-CoV-2 as a human pathogen. Biorxiv. 2022;17:e0270276.
67. Rochman ND, Wolf YI, Faure G, Mutz P, Zhang F, Koonin EV. Ongoing global and regional adaptive evolution of SARS-CoV-2. Proc Natl Acad Sci. 2021;118(29):e2104241118.
68. Rodrigues CH, Pires DE, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res. 2018;46(W1):W350–5.
69. Rodrigues CH, Pires DE, Ascher DB. DynaMut2: assessing changes in stability and flexibility upon single and multiple point missense mutations. Protein Sci. 2021;30(1):60–9.
70. Saha O, Islam I, Shatadru RN, Rakhi NN, Hossain M, Rahaman M. Temporal landscape of mutational frequencies in SARS-CoV-2 genomes of Bangladesh: possible implications from the ongoing outbreak in Bangladesh. Virus Genes. 2021;57(5):413–25.
71. Salemi M, Fitch WM, Ciccozzi M, Ruiz-Alvarez MJ, Rezza G, Lewis MJ. Severe acute respiratory syndrome coronavirus sequence characteristics and evolutionary rate estimate from maximum likelihood analysis. J Virol. 2004;78(3):1602–3.
72. Shah M, Woo HG. Omicron: a heavily mutated SARS-CoV-2 variant exhibits stronger binding to ACE2 and potently escapes approved COVID-19 therapeutic antibodies. Front Immunol. 2022;12:6031.
73. Shen S, Zhang Z, He F. The phylogenetic relationship within SARS-CoV-2s: an expanding basal clade. Mol Phylogenet Evol. 2021;157:107017.
74. Shu T, Huang M, Wu D, Ren Y, Zhang X, Han Y, Mu J, Wang R, Qiu Y, Zhang D-Y. SARS-coronavirus-2 Nsp13 possesses NTPase and RNA helicase activities that can be inhibited by bismuth salts. Virolog Sinica. 2020;35(3):321–9.
75. Shu, Y. and J. McCauley (2017). Eurosurveillance 2017, https://doi.org/10.2807/1560-7917, ES.
76. Smith EC, Blanc H, Vignuzzi M, Denison MR. Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. PLoS Pathog. 2013;9(8):e1003565.
77. Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, Gorbalenya AE. Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. J Mol Biol. 2003;331(5):991–1004.
78. Sun C, Chen L, Yang J, Luo C, Zhang Y, Li J, Yang J, Zhang J, Xie L. SARS-CoV-2 and SARS-CoV spike-RBD structure and receptor binding comparison and potential implications on neutralizing antibody and vaccine development. Biorxiv. 2020;25:2000044.
79. Swaminathan S, Lineburg KE, Ambalathingal GR, Crooks P, Grant EJ, Mohan SV, Raju J, Panikkar A, Le Texier L, Tong ZWM. Limited recognition of highly conserved regions of SARS-CoV-2. Microbiol Spectr. 2022;10(1):e02780-e2721.
80. Te Velthuis AJ, Van Den Worm SH, Snijder EJ. The SARS-coronavirus nsp7+ nsp8 complex is a unique multimeric RNA polymerase capable of both de novo initiation and primer extension. Nucleic Acids Res. 2012;40(4):1737–47.
81. Troyano-Hernáez P, Reinosa R, Holguín Á. Evolution of SARS-CoV-2 envelope, membrane, nucleocapsid, and spike structural proteins from the beginning of the pandemic to september 2020: a global and regional approach by epidemiological week. Viruses. 2021;13(2):243.
82. Troyano-Hernáez P, Reinosa R, Holguín Á. Evolution of SARS-CoV-2 in Spain during the First Two years of the pandemic: circulating variants, amino acid conservation, and genetic variability in structural, non-structural, and accessory proteins. Int J Mol Sci. 2022;23(12):6394.
83. Ulloa S, Bravo C, Parra B, Ramirez E, Acevedo A, Fasce R, Fernandez J. A simple method for SARS-CoV-2 detection by rRT-PCR without the use of a commercial RNA extraction kit. J Virol Methods. 2020;285:113960.
84. Ullrich S, Nitsche C. The SARS-CoV-2 main protease as drug target. Bioorg Med Chem Lett. 2020;30:127377.
85. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CC, Boshier FA. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol. 2020;83:104351.

86. Van Egeren D, Novokhodko A, Stoddard M, Tran U, Zetter B, Rogers M, Pentelute BL, Carlson JM, Hixon M, Joseph-McCarthy D. Risk of rapid evolutionary escape from biomedical interventions targeting SARS-CoV-2 spike protein. PLoS ONE. 2021;16(4):e0250780.
87. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell. 2020;181(2):281-292. e286.
88. Wang R, Chen J, Gao K, Hozumi Y, Yin C, Wei G-W. Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. Communications biology. 2021;4(1):1–14.
89. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, Graham BS, McLellan JS. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 2020;367(6483):1260–3.
90. Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, Meng J, Zhu Z, Zhang Z, Wang J. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell Host Microbe. 2020;27(3):325–8.
91. Wu C, Qavi AJ, Hachim A, Kavian N, Cole AR, Moyle AB, Wagner ND, Sweeney-Gibbons J, Rohrs HW, Gross ML. Characterization of SARS-CoV-2 nucleocapsid protein reveals multiple functional consequences of the C-terminal domain. Iscience. 2021;24(6):102681.
92. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265–9.
93. Wu H, Xing N, Meng K, Fu B, Xue W, Dong P, Tang W, Xiao Y, Liu G, Luo H. Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. Cell Host Microbe. 2021;29(12):1788-1801. e1786.
94. Xia S, Yan L, Xu W, Agrawal AS, Algaissi A, Tseng C-TK, Wang Q, Du L, Tan W, Wilson IA. A pan-coronavirus fusion inhibitor targeting the HR1 domain of human coronavirus spike. Sci Adv. 2019;5(4):eaav4580.
95. Yin W, Mao C, Luan X, Shen D-D, Shen Q, Su H, Wang X, Zhou F, Zhao W, Gao M. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. Science. 2020;368(6498):1499–504.
96. Yu J, Qin Z, Liu X, He X, Yao J, Zhou X, Wen K, Yu N, Wu Q, Xiao W. High-specificity targets in SARS-CoV-2 N protein for serological detection and distinction from SARS-CoV. Comput Biol Med. 2022;143:105272.
97. Zhai Y, Sun F, Li X, Pang H, Xu X, Bartlam M, Rao Z. Insights into SARS-CoV transcription and replication from the structure of the nsp7–nsp8 hexadecamer. Nat Struct Mol Biol. 2005;12(11):980–6.
98. Zhang C, Yang M. Newly emerged antiviral strategies for SARS-CoV-2: from deciphering viral protein structural function to the development of vaccines, antibodies, and small molecules. Int J Mol Sci. 2022;23(11):6083.
99. Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, Becker S, Rox K, Hilgenfeld R. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. Science. 2020;368(6489):409–12.
100. Zhang S, Guo M, Duan L, Wu F, Hu G, Wang Z, Huang Q, Liao T, Xu J, Ma Y. Development and validation of a risk factor-based system to predict short-term survival in adult hospitalized patients with COVID-19: a multicenter, retrospective, cohort study. Crit Care. 2020;24(1):1–13.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.