SARS-CoV-2 NSP-12 mutations survey during the pandemic in the world

Sogol Mazhari
Shahid Beheshti University
Helia Alavifard
Islamic Azad University
Karim Rahimian
Tarbiat Modares University
Zohreh Karimi
Kharazmi University
Mohammadamin Mahmanzar
Tarbiat Modares University
Mahsa Mollapour Sisakht
Erasmus University Medical Center
Masoud Bitaraf
Tehran University of Medical Sciences
Ehsan Arefian (Arefian@ut.ac.ir)
University of Tehran

Research Article

Keywords: SARS-CoV-2, RNA-dependent RNA polymerase, prevalence, mutation

Posted Date: September 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-877078/v1

Abstract

Background

SARS-CoV-2 belongs to the *beta coronavirus* family responsible for coronavirus disease 2019 (COVID-19), a novel severe acute respiratory syndrome pandemic. The infection first emerged in Wuhan, China, and rapidly spread worldwide. The ongoing outbreak has posed an urgent worldwide health threat due to the rapid transmittable potential and high mortality rate. Due to the critical role of none structural protein – 12 (NSP-12) in COVID-19. This study tries to investigate the link between genotype-phenotype NSP-12 variation and the prevalence of this disease.

Methods

We analyzed approximately 2 million Nsp12 of SARS-CoV-2 protein sequence from January 2020 until June 2021. Python programming language was utilized to preprocess and apply inclusion criteria on FASTA files to prepare a list of suitable samples for clustering samples. NSP-12 regions were aligned to the reference sequence to compare and identify mutation patterns, categorized based on frequency and continent.

Results

The rate of NSP-12 mutation in divided geographical areas was different. Based on continental studies, the P227L and G671S mutations have multiplied over time and in European and Asian societies in recent months. According to biochemical studies, the occurrence of G671S mutation increases the stability of the protein.

Conclusion

We concluded that NSP-12 P227L and G671S mutations in SARS-CoV-2 are increased in recent months. Further studies will be required to investigate whether these mutations impact the severity of the disease.

Introduction

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), causing Coronavirus disease 2019 (COVID-19)[1], belongs to the *Coronaviridae* family of viruses, which has one of the largest single plus-strand RNA genomes among all known RNA viruses with approximately 30 kb length [2]. COVID-19, as a disease with potential respiratory involvement, was first detected in Wuhan, China, in December 2019. However, with variable geographical prevalence, it has afflicted over 200 countries and is considered a significant threat to global health[3,4]. The number of infected people and the mortality rate caused by

COVID-19 is variable in different societies. Despite extensive research explaining the possibilities for the difference, such as various age distribution, host genetic background, and virus genomic type, the exact reasons are still unclear[5–7]. One plausible explanation might be the mutation rate within the SARS-CoV-2 genome[8]. For instance, the D614G mutation of spike protein was significantly associated with SARS-CoV-2 mortality rates in the 28 countries[9]. The S194L mutation in n nucleocapsid (N) protein is notably correlated with the mortality rate in Western India[10].

The SARS-CoV-2 genome comprises at least six open reading frames (ORFs). The two large ORFs (ORF1a/b), which contain about two-thirds of the whole genome, are located at the 5' terminus. The ORF1a/b encodes 16 non-structural proteins (NSP1– NSP16) that participate in the transcription and replication [11]. Based on collected data from the literature, mutation in SARS-CoV-2 RNA polymerase machinery might considerably impact the virus's virulence and mortality rate caused by the virus[12]. Among the NSP proteins participating in the RNA replication machinery, RNA-dependent RNA polymerase (RdRp, NSP-12) is the core enzyme involved in this process[13]. NSP-12 acts as the main component of replication machinery of SARS-CoV-2 and comprises three parts; RNA dependent RNA polymerase (RdRp), interface, and nidovirus RdRp-associated nucleotidyltransferase (NiRAN)[14]. The biochemical investigation demonstrated that Nsp12 is linked to Nsp7/8 by subunit to perform efficient RNA synthesis. Moreover, Nsp 14 associates with Nsp 12/7/8 complex by binding to a subunit involved in the final mRNA cap synthesis and proofreading. Additionally, Nsp12 directly interacts with helicase, which results in enhancing the helicase activity[15]. The limited proofreading activity of SARS-CoV-2 RdRp introduces an actual mutation rate (approximately 4 x 10^4) in the process of virus replication[16,17].

On the other hand, the conserved presence of RdRp in the RNA virus families with structural similarities and the absence of homologs in the host genomic structure make it an ideal target for drug repurposing to diminish the SARS-CoV-2s clinical consequences[18,19]. Some FDA-approved antiviral drugs that act as NSP-12 inhibitors are Remdesivir[20], Favipiravir[21], and Ribavirin[22] that showed promising results by reducing the clinical consequences of SARS-CoV-2 infection and further triggering of immunopathogenic responses[23].

To date, vaccination is the optimal global solution against Covid19 to make the immunity. Most vaccines are developed against spike protein, a structural protein on the virus surface, as the entrance mediating ligand connecting virus to host cell via Angiotensin-converting enzyme 2 (ACE-2) receptor[24]. Thus, not only humoral immunity but also cell-mediated immunity plays a pivotal role in vaccine-induced immunity. The host cell-mediated immunity responses are associated with SARS-CoV-2 non-structural proteins (NSPs) such as NSP-4, ORF3s, ORF7a, and particularly NSP-12[25].

Considering the importance mentioned roles of NSP-12 and the therapeutic potential proposed against it, identifying and investigating the characteristics of mutations in the NSP-12 protein structure in different geographical distributions and the evolutionary process of the virus might be considered as effective

approaches to study the epidemiology of COVID-19 and the molecular basis of region-specific SARS-CoV-2 vaccine design.

The current study aims to understand the evolutionary trend in NSP-12 of SARS-CoV-2 by investigating the amino acid mutation patterns and their specificities divided on continents from the beginning of the pandemic to June 2021. Furthermore, it discusses the relation between different continental prevalence and mutations that can provide valuable information on predicting the future transmission dynamics of SARS-CoV-2, development of DNA-based diagnosis, and vaccine for COVID-19.

Results

Numbers and incidence of mutations in NSP-12 AAS based on geographical areas

First, the incidence of mutations on the NSP-12 protein structure was statistically investigated in order to identify the potentially essential mutations. As of now, the total number of 1,759,792 sequences was examined in terms of the number of AAS mutations. The global distribution depicts that the overall significant mutation rate of the NSP-12 region may result in new clades, which in turn may have a notable impact on the mortality rate, drug resistance or vaccine escape, and severity of the disease across different geographical distributions (Fig. 1).

In particular, the statistics showed that no mutation was detected in 1.36% of AASs. 74.84% AASs contained one mutation, 20.2% of AASs comprised two mutations. In comparison, 3.38% of sequences carried three mutations, and 0.23% of AASs represented more than four mutations in their AASs. Next, mutations were sorted into six geographical regions; North America, South America, Europe, Asia, Oceania, and Africa (Fig. 2. A).

North America data with 553,149 AASs included no mutations in 1.15% of sequences, one mutation in 71.94%, two mutations in 24.03%, three mutations in 1.87% of AASs, and 0.42% of AASs showed more than four mutations in their AASs. South America data included 25,130 AASs demonstrated no mutations in 0.88% of sequences, one mutation in 84.86%, two mutations in 12.85%, three mutations in 0.99% of AASs, and 0.13% of AASs illustrated more than four mutations in their AASs. In 1,037,450 AASs of Europe, the mutation was not detected in 1.12% of sequences, one mutation was detected in 76.26%, two mutations in 17.82%, three mutations in 4.34% of AASs, and 0.27% of AASs showed more than four mutations in their AASs. Data regarding Asia with 110,085 AASs included no mutations in 2.1% of sequences, one mutation in 71.68%, two mutations in 22.64%, three mutations in 3.34% of AASs, and 0.24% of AASs showed more than four mutations in their AASs. Oceania data with 17,595 AASs comprised no mutations in 0.76% of sequences, one mutation in 78.12%, two mutations in 13.65%, three mutations in 0.05% of AASs, and 7.41% AASs demonstrated more than four mutations in their AASs. Finally, data related to Africa with 15612 AASs revealed no mutations in 2.75% of sequences, one

mutation in 79.24%, two mutations in 12.76%, three mutations in 4.80% of AASs, and 0.35% of AASs represented more than four mutations in their AASs (Fig. 2. A).

Given that at least one mutation has occurred in most NSP-12 sequences in all geographical areas, our data indicates NSP-12 as a mutation hotspot in line with previous studies[26,27].

Counting the number of mutations that occur is not enough to investigate the impact of mutations alone, as some modifications may appear multiple times. In contrast, others may occur on only a few samples. For this purpose, a heat map was drawn to investigate the frequency of mutations of the NSP-12 protein section. The results demonstrated that the highest mutations frequency occurred in the region of 301 to 400 AA (0.7847 frequency) and then in the areas of 201 to 300 AA (0.0634 frequency) and 601 to 700 AA (0.0370), respectively (Fig. 2.B). The mutation frequency was detected based on the number of mutations in each section relative to the total AASs.

Mutation's Specificities According To Geographical Areas

To study the NSP-12 AASs mutations in more detail, the location of mutations in the protein structure and their frequency were investigated between January 2020 and June 2021. Table 1 describes the first five frequent mutations regardless of geographical distribution. The complete list of mutations and their frequencies is attached to the additional file 1 A to G.

Table 1
NSP-12 first five frequent mutations globally from January 2020 until June 2021.

Rank	Residue	Frequency	Total frequency
Top 1	P(323)L	0.98366	1,731,040
Top 2	P(227)L	0.061411	108,071
Top 3	G(671)S	0.028901	50,860
Top 4	V(776)L	0.018056	31,775
Top 5	A(185)S	0.017245	30,348

According to our findings, the most frequent mutation belongs to P323 and up to June 2021. The most dominant mutation is the approximate 0.98366 frequency rate (1,731,040 times in 1,759,792 AASs). P323 resides in the interface domain of the NSP-12 protein, which was previously shown to be associated with the stabilization of the protein structure. A recent *silico* study based on the virtual molecular docking investigation nominated potential drugs as SARS-CoV-2 RdRp inhibitor, including Simeprevir and Filibuvir[28]. The docking site of the desired drugs is located within a hydrophobic cleft which includes phenylalanine at the 326th position close to the P323 mutation site. Hence, this mutation may interfere with the affinity of RdRp with these antiviral drugs[29].

P227 mutation appeared recently and may increase in prevalence during the latest global peak, located at the N-terminal extension domain that adopts a nidovirus RdRp-associated nucleotidyltransferase (NiRAN) structure (residues D60-R249)[30].

G671S is now considered a fixed mutation of SARS-CoV-2 Delta variant emergence from India, Dec 2020 (https://viralzone.expasy.org/9556). A variant with a significant increase in transmissibility, severity of disease, and the potential to escape neutralization by antibodies[31].

It is noteworthy that some studies demonstrated the increase in the prevalence of substitutions A185S and V776L mutations, which suggests the co-occurrence of these mutations [32,33]. Moreover, A185s mutation may have a notable impact on the NSP-12 protein structure by preserving the secondary structure of the protein[29].

The statistical occurrence of these top five mutations base on the continents is listed in Table 2. Interestingly, not all of these mutations are present among the continents as the top five mutations. Among these mutations, the p323 mutation was present in all continents (North America (0.9862 frequency), South America (0.9896 frequency), Europe (0.9877 frequency), Asia (0.9491 frequency), Oceania (0.9098 frequency), and Africa (0.9394 frequency) as the mutation with the highest incidence rate. P227 mutation has been observed as one of the top mutations in North America (0.1292 Frequency), Europe (0.0327 Frequency), South America (0.0193 Frequency), and Africa (0.0331 Frequency). G671 mutation has been remarked in North America (0.008 Frequency), Asia (0.0454 Frequency), and Europe (0.0396 Frequency) among the top five mutations. Finally, V776 mutation in North America (0.0085 Frequency), Europe (0.0254 Frequency), and Africa (0.0228 Frequency), and A185 mutation in Europe (0.0262 Frequency) and America (0.0258 Frequency) are observed among the top five mutations.

Table 2
The incidence of the global NSP-12 top-five mutations is based on the continents.

Residue	Variant frequency					
	North America	South America	Europe	Asia	Oceania	Africa
P(323)L	0.9862	0.9896	0.9877	0.9491	0.9098	0.9394
P(227)L	0.1292	0.0193	0.0327	0.128	0.0055	0.0331
G(671)S	0.0080	0.0009	0.0396	0.0454	0.0091	0.0048
V(776)L	0.0085	0.0038	0.0254	0.0013	0.0026	0.0228
A(185)S	0.0042	0.0046	0.0262	0.0018	0.0022	0.0258

In North America, the A97 (0.0044 Frequency) mutation ranks fifth according to the highest mutation rates. In the 97rd AA position, the alanine amino acid is predominantly substituted by valine with a larger side chain. A previous study demonstrated that this mutation has a negative impact on the packaging of the NSP-12[29]. This mutation was mainly detected in the mild and asymptomatic samples[34].

The pattern observed in South America shows that the K91 (0.0063 Frequency) and I548 (0.0053 Frequency) mutations are placed in the fourth and fifth ranks, respectively, which have not had a significant prevalence in other geographical areas. In Europe, the pattern of the top five mutations is consistent with the global pattern, which may be due to the presence of the largest number of database sequences (1,037,450 AASs). In Asia, the A423 (0.0826 Frequency), the A97 (0.0148 Frequency), and the M666 (0.0139 Frequency) mutations rank the second, fourth, and fifth mutations respectively. [35]. The substitution of methionine with isoleucine results in increasing the flexibility of the NSP-12 structure (http://biosig.unimelb.edu.au/covid3d/mutation/QHD43415_11/I/M666I/A).

Oceania shows almost a different pattern of top five high ranks variants. The K718 (0.0478 Frequency) mutation ranks second, the N215 (0.0109 Frequency) mutation ranks third, the A97 (0.0102 Frequency) mutation ranks fourth, and the K267 (0.0101 Frequency) mutation ranks fifth. Finally, the African continent differs from the global T85 (0.008 Frequency) mutation pattern, which ranks fifth of the top five most frequent mutations. The top five mutations of each continent and which amino acids they have been substituted are shown in Fig. 3. The complete list of mutations and their frequencies is attached to the additional files 2 A to G.

Due to differences in the emergence of RdRp variants in SARS-CoV-2 in different continents. This can pose a tremendous provocation to the effectiveness of antiviral therapies. Thus, investigating the evolutionary patterns and spread dynamics of the SARS-CoV-2 NSP-12 variant is of enormous importance.

Evolutionary assessment of top five mutation’s incidence according to time and geographical regions

To determine the appearance of each mutation, we analyzed AASs from each geographic area over time by classifying them according to the month of sample collection from December 30, 2019, till June 30, 2021, as indicated in the GISAID database. This process helps identify mutations that caused the global peak and have increased in frequency over time through stabilizing and beneficiary qualities.

The detailed distribution of the top five high mutation rates of NSP-12 variants from the world and each continent is provided by the month of sample collection and illustrated in Fig. 4.

Next, we investigate the continuation of mutations with a prevalence rate higher than 0.1 per AASs collected each month. P323L mutation began to be observed at the beginning of the pandemic., its prevalence fluctuated in all geographical regions. After that, the incidence increased exponentially. In the last month of the study, June 2020, it was observed with 0.99 frequency rate in June.

The P227L mutation and G671 mutation gained in prevalence from their appearance on February 2021 and April 2021, respectively, presented by 0.01 and 0.77 frequency of the AASs collected worldwide on June 2021.

In particular, P227L mutation in North America has been on the rise since December 2020. In April 2021, the rate of 0.25 frequency is based on the collected samples per month, but in the last month of the study, its prevalence has been decreased to 0.23 frequency. The growing exponential trend of G671S mutation in the last months of the study (June 2021) is very significant, as observed with 0.19 and 0.81 and 0.57 frequency in North America, Europe, and Asia, respectively.

Assessment of P227L and G671S mutations on dynamicity and flexibility of NSP-12

To reveal the effect of P227L and G671S mutations on the tertiary Structure of NSP-12, we have done protein modeling applying DynaMut website. We calculated the alteration in vibrational entropy energy ($\Delta\Delta S_{vib}$ENcom) between the wild and mutant types. Our data demonstrated that the mutation at P227L decreases molecular flexibility on the protein structure by value – 0.174 kcal.mol^{-1}K^{-1}, however, substituted serine AA on the 671 residues (G671S mutation) with $\Delta\Delta S_{vib}$ENcom 0.080 kcal.mol^{-1}K^{-1} increased molecular flexibility of NSP-12 protein structure.

Furthermore, the investigation on the changes in the intramolecular interactions caused by P227L mutation revealed that this mutation might affect the interaction with the residues that are closed to wild-type proline. The substitution of leucine alters the side chain, resulting in the alteration of intramolecular bonds in the pocket. These amino acids residue are shown in Fig. 5. A. Moreover, substituting glycine (none polar amino acid) to serine (polar amino acid) has changed intramolecular interactions among amino acids adjacent to 671rd residue by Ionic interactions and water-mediated weak hydrogen bonds Fig. 5. B.

Discussion

As days go by, the number of COVID-19 victims increases. Its related mortality rate varies among different regions, and This is attributed to several factors, including genetic basis, community age, and viral mutations[36]. Since universal vaccination is currently considered the ultimate key to disease elimination, the emergent genetic variants might undermine the efficacy of the desired therapeutic interventions-ongoing effort to develop effective medicines. Recently, location-dependent species such as Delta shows the weak sides of vaccination[37]. A recent investigation demonstrated that although vaccination provides proper immune response against Delta variant but shows less efficacy rather than the older strain of SARS-CoV-2s, particularly after the recipient of the first dose[38]. Hence, it is essential to carefully assess virus structure and monitor its AAS to recognize structural changing mutations [39].

Generally, mutations are more frequent among RNA viruses than DNA viruses, primarily due to a less competent proofreading system. Notably, the virus mutagenic capability is related to the fidelity of RdRp, which shows the significance of mutation in RdRp. Interestingly, it was reported that viral strains with RdRp mutations exhibited a mutation rate three times higher than those without a RdRp mutation[40].

Rapid generation of new SARS-CoV-2 variants is observed due to the individual infection of different ages and genetic compositions.

Consequently, due to the vital role of NSP-12, this study was designed and conducted comprehensively to include all the mutations that occurred between January 2020 and June 2021. Furthermore, regional patterns of mutations were tracked based on the following six regions; Africa, North and South America, Asia, Europe, and Oceania.

Not only RdRp is a critical element in viral replication, but it also is the target site of many drugs used in the management of COVID-19[41]. Remdesivir, Sofosbuvir, Ribavirin, Favipiravir, Baloxavir marboxil, Galidesivir, Pimodivir, and Beclabuvir are all studied in COVID-19 and have shown variable efficacy. An important point in using drugs is that they should target the highly constrained regions of the NSP-12 to prevent drug resistance [42]. Some changing structural mutations are associated with drug resistance[43]. For instance, mutations F480L, V557L, and D484Y in RdRp protein may lead to resistance to Remdesivir[44,45]. Therefore, it may need to use two to three different drugs (combination therapy) against different protein targets the disease in the treatment[46].

The phenomenon of natural selection is a significant factor in determining the fate of mutations in the viral genome[47]. Typically, mutations that result in adaptation will remain, boosting the potential of resistance to the host's immune system, increasing proliferation, and facilitating its transmission. Based on our data, some less frequent mutations such as A185S, V776L, V720I, A423V, and E254D are found in RdRp but have shown decreased prevalence over time, possibly due to unfavorable outcomes in terms of evolution.

In the present study, according to the results of evolutionary patterns of tope five mutations based on the monthly trend, P227L and G671S, due to the increase in incidence at the end of this study can be of great importance.

P227L is the second frequent mutation and affects the RNA binding domain on the protein surface, located in the NiRAN domain. Substitution of Proline by Leucine can damage the structural integrity of the protein[48]. Proline illustrated a unique property. Its side-chain cyclizes back onto the backbone amide position. It contributes to the secondary structure formation because of its bulky pyrrolidine ring that places steric constraints on the conformation of the preceding residue. Leucine is a hydrophobic aliphatic AA. Its embedment in 227 locations leads to alterations in the secondary structure of the protein. Proline is associated with focal flexibility of protein, which affects complex global structure-dynamics by decreasing the flexibility of the NSP-12 protein. So we hypothesized that its substitution could lead to instability in co-factors binding (NSP-7 and NSP-8), ultimately altering the proofreading complex and losing the structural integrity provided by proline.

G671 is located at the RdRp domain in Thr680, making a hydrogen bond with incoming NTP. Hence, its substitution with serine with increasing the number of intramolecular interactions with closed residues

like Thr680 may result in its replication fidelity and have a notable impact on NSP-12 sensitivity to promising drugs Remdesivir and Favipiravir.

One of the limitations of the present study is that we focus on the AASs alone without considering the nucleotide sequences, which makes some aspects of developing variants, including codon bias, not be considered[49]. Another limitation that should be considered is that the sample reporting rate varies among regions with more reports from Europe and North America and the least ones. Hence, there may be functionally significant variants present without extensive sequencing of samples globally that have not yet been discovered.

Our findings support the idea that the occurrence of several mutations in the coronavirus genome can result in milder or more severe patient outcomes. The functional attribution of P227L and G671S mutations reported in our study needs to be carried out to discover the exact role of these mutations in disease severity. It could lead to appropriate therapeutic targeting of the covid-19 virus.

Conclusion

Altogether, our data strongly suggest that SARS-CoV-2 is acquiring mutations as it is spreading to new locations. Most likely, these mutations are helping SARS-CoV-2 to adapt better inside hosts and in new geographical areas. One of the mutations identified in our study (P227L and G671S) might have functional consequences that need to be addressed in future studies.

Materials And Methods

Sequence source

This study evaluated whole data regarding NSP-12 of SARS-CoV-2 amino-acid (AA) sequences (AASs). AAS of the Wuhan-2019 virus with access number ‘EPI_ISL_402124’ was adopted as the reference sequence. All AASs were compared to this sample. AASs of NSP-12 region (with 931 AA) of 1,952,292 samples from January 2020 until June 2021 were extracted from GISAID (www.gisaid.org) database. We have access to this database with Erasmus Medical Center's permission. NSP-12 is located on the ORF1ab sequence and comprises amino acids between 4393-Ser and 5324-Gln.

Sequence analyses and processing

Python 3.8.0 software was utilized to preprocess FASTA files, extract NSP-12 from other genes and perform sequence alignment and mutation analysis. Each difference between sample and reference was recognized as a mutation, and the location and substituted AA were reported. Non-human samples (such as bat and pangolin), those with less or more than 931 AAs and samples containing non-specified AAs (reported as X) were omitted. Ultimately, 1,759,792 samples were included in the study. The whole process was optimized by applying ‘Numpy’ and ‘Pandas’ libraries.

The identifying algorithm for detecting mutants is as follows:

Since all sequences have equal lengths, the following algorithm used ‘Refseq,’ and ‘seq’ refer to reference sequence sample sequence, respectively.

For refitem, seqitem in zip (refseq, seq)

If (refitem! =seqitem)

Report a new mutant

After extracting NSP-12, each sample's continent name and geographical coordinates were obtained and reported using pycountry-convert 0.5.8 software and ‘Titlecase’ library in Python to draw global prevalence maps of mutations. Finally, global maps were drawn using Matlab 2021 ‘Geobubble’ package. The data refining procedure is shown in Figure 6.

Secondary protein structure and dynamic prediction

NSP-12 protein modeling was performed by DynaMut web server (http://biosig.unimelb.edu.au/dynamut/) to analyze the protein's mutational structure and its molecular flexibility. For this purpose, the wild type of protein sequence (Wuhan) was imported. Next, the P227L mutation sequence was uploaded. The protein structure stability parameters including vibrational entropy, atomic fluctuations, and deformation energies by applying mutation in the desired region.

Statistical analysis

Data normalization and comparison charts outlining were conducted through R 4.0.3 and Microsoft Power BI. To better compare the data of each continent, the normalized frequency of each region was reported. For this purpose, the number of mutations was divided by the number of sequences on that continent comparable in equal proportions.

List Of Abbreviations

coronavirus disease 2019 (COVID-19), None structural protein -12 (NSP-12), Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), Open reading frames (ORFs), RNA-dependent RNA polymerase (RdRp), Nidovirus RdRp-associated nucleotidyltransferase (NiRAN), Amino-acid (AA), Amino- acid sequences (AASs).

Declarations

Acknowledgments: Not applicable.

Authors' contributions:

Sogol Mazhari: carried out the experiments and major contributor writing the manuscript.

Helia Alavifard: carried out the experiments and major contributor writing the manuscript.

Karim Rahimian: contributed to the design study of the project and interpretation of the main results and contributed to preparing and interpreting Additional data 1,2(A-G).

Zohreh Karimi: contributed to interpreting the main results and preparing the figures and contributed to organizing and arranging Additional data 1,2(A-G).

Mohammadamin Mahmanzar: conceived and planned the experiments, supervised the project, major contributor writing the manuscript, and monitored the accuracy of Additional data 1,2(A-G).

Mahsa Mollapour Sisakht: contributed to the interpretation of the results.

Masoud Bitaraf: aided in interpreting the results.

Ehsan Arefian: Experimental design and advice for improving the study, final edit, and proofread draft for submission and monitored the accuracy of Additional data 1,2(A-G).

Additional information:

Competing interests: The authors declare no competing interests.

Funding: Not applicable.

Ethics approval: Not applicable.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: Not applicable.

References

1. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature*, **579**, 265–269 (2020).
2. Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in China, 2019(2020).
3. Rosenkrantz, L., Schuurman, N., Bell, N. & Amram, O. J. H. & Place. The need for GIScience in mapping COVID-19. **67**,102389(2021).

4. Qin, C., Ziwei, M. P. L. Z. M., Tao, S. Y. M. Y. & Ke, P. C. X. M. P. & Shang, M. M. P. K. J. C. I. D. Dysregulation of immune response in patients with COVID-19 in Wuhan, China; Clinical Infectious Diseases; Oxford Academic.
5. Ovsyannikova, I. G., Haralambieva, I. H., Crooke, S. N., Poland, G. A. & Kennedy, R. B. The role of host genetics in the immune response to SARS-CoV-2 and COVID-19 susceptibility and severity. *Immunological reviews*, **296**, 205–219 (2020).
6. Rastogi, M., Pandey, N., Shukla, A. & Singh, S. K. SARS coronavirus 2: from genome to infectome. *Respir. Res*, **21**, 1–15 (2020).
7. Bonanad, C. *et al.* The effect of age on mortality in patients with COVID-19: a meta-analysis with 611,583 subjects. *Journal of the American Medical Directors Association*, **21**, 915–918 (2020).
8. Sanjuán, R. & Domingo-Calap, P. J. C. & sciences, m. l. *Mechanisms of viral mutation*, **73**, 4433–4448 (2016).
9. Toyoshima, Y., Nemoto, K., Matsumoto, S., Nakamura, Y. & Kiyotani, K. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. *Journal of human genetics*, **65**, 1075–1082 (2020).
10. Joshi, M. *et al.* Genomic variations in SARS-CoV-2 genomes from Gujarat: Underlying role of variants in disease epidemiology.*Frontiers in genetics***12** (2021).
11. Naqvi, A. A. T. *et al.* Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, **1866**, 165878 (2020).
12. Becerra-Flores, M. & Cardozo, T. SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. *International journal of clinical practice*, **74**, e13525 (2020).
13. Gao, Y. *et al.* Structure of the RNA-dependent RNA polymerase from COVID-19 virus. **368**,779–782(2020).
14. Ilmjärv, S. *et al.* Epidemiologically most successful SARS-CoV-2 variant: concurrent mutations in RNA-dependent RNA polymerase and spike protein. (2020).
15. Wang, Q. *et al.* Structural basis for RNA replication by the SARS-CoV-2 polymerase., **182**, 417–428 (2020).
16. Venkataraman, S., Prasad, B. V. & Selvarajan, R. J. V. RNA dependent RNA polymerases: insights from structure. *function and evolution*, **10**, 76 (2018).
17. Salemi, M. *et al.* Severe acute respiratory syndrome coronavirus sequence characteristics and evolutionary rate estimate from maximum likelihood analysis. *Journal of virology*, **78**, 1602–1603 (2004).
18. Pushpakom, S. *et al.* Drug repurposing: progress, challenges and recommendations. **18**,41–58(2019).
19. de Farias, S. T., dos Santos Junior, A. P., Rêgo, T. G. & José, M. V. J. F. i. g. Origin and evolution of RNA-dependent RNA polymerase. 8, 125(2017).

20. Beigel, J. H. *et al.* Remdesivir for the treatment of Covid-19—preliminary report(2020).
21. Seneviratne, S. L. *et al. Favipiravir in COVID-19*, **19**, 143–145 (2020).
22. Khalili, J. S., Zhu, H., Mak, N. S. A., Yan, Y. & Zhu, Y. J. J. o. m. v. Novel coronavirus treatment with ribavirin: Groundwork for an evaluation concerning COVID-19. **92**,740–746(2020).
23. Vicenti, I., Zazzi, M. & Saladini, F. J. E. O. o. T. P. SARS-CoV-2 RNA-dependent RNA polymerase as a therapeutic target for COVID-19. **31**,325–337(2021).
24. Frampton, D. *et al.* Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B. 1.1. 7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study. (2021).
25. Grifoni, A. *et al.* Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. **181**,1489–1501. e1415 (2020).
26. Wang, R., Hozumi, Y., Yin, C. & Wei, G. W. Decoding SARS-CoV-2 transmission and evolution and ramifications for COVID-19 diagnosis, vaccine, and medicine. *Journal of chemical information and modeling*, **60**, 5853–5865 (2020).
27. Pachetti, M. *et al.* Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of translational medicine*, **18**, 1–9 (2020).
28. Ruan, Z. *et al.* Potential inhibitors targeting RNA-dependent RNA polymerase activity (NSP12) of SARS-CoV-2. (2020).
29. Chand, G. B., Banerjee, A. & Azad, G. K. Identification of novel mutations in RNA-dependent RNA polymerases of SARS-CoV-2 and their implications on its protein structure. *PeerJ*, **8**, e9492 (2020).
30. Lehmann, K. C. *et al.* Discovery of an essential nucleotidylating activity associated with a newly delineated conserved domain in the RNA polymerase-containing protein of all nidoviruses. *Nucleic acids research*, **43**, 8416–8434 (2015).
31. Campbell, F. *et al.* Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Eurosurveillance*, **26**, 2100509 (2021).
32. Showers, W. M., Leach, S. M., Kechris, K. & Strong, M. Analysis of SARS-CoV-2 Mutations Over Time Reveals Increasing Prevalence of Variants in the Spike Protein and RNA-Dependent RNA Polymerase. bioRxiv(2021).
33. Laha, S. & Chatterjee, R. Country specific mutational profile of SARS-CoV-2 in pre-and post-international travel ban: Effect on vaccine efficacy.*medRxiv*(2021).
34. Biswas, N. *et al.* Structural and Drug Screening Analysis of the Non-structural Proteins of Severe Acute Respiratory Syndrome Coronavirus 2 Virus Extracted From Indian Coronavirus Disease 2019 Patients.*Frontiers in genetics***12** (2021).
35. Hillen, H. S. Structure and function of SARS-CoV-2 polymerase.*Current Opinion in Virology*(2021).
36. Zhou, B., Kojima, S., Kawamoto, A. & Fukushima, M. J. J. o. m. v. COVID-19 pathogenesis, prognostic factors, and treatment strategy. *Urgent recommendations*, **93**, 2694–2704 (2021).
37. Lopez Bernal, J. *et al.* Effectiveness of Covid-19 vaccines against the B. 1.617. 2 (delta) variant.*New England Journal of Medicine*(2021).

38. Planas, D. *et al.* Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization.*Nature*,1–7(2021).
39. Alazab, M. *et al.* COVID-19 prediction and detection using deep learning. **12**,168–181(2020).
40. Pachetti, M. *et al.* Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant(2020).
41. Vilar, S. & Isom, D. G. J. B. One year of SARS-CoV-2: How much has the virus changed? **10**,91(2021).
42. Tu, Y. F. *et al.* A review of SARS-CoV-2 and the ongoing clinical trials. *International journal of molecular sciences*, **21**, 2657 (2020).
43. EA, R. J. L. S. Remdesivir, sofosbuvir, galidesivir, and tenofovir against SARS-CoV-2 RNA dependent RNA polymerase (RdRp): a molecular docking study. (2020).
44. Menéndez-Arias, L. & Richman, D. D. J. C. o. i. v. Editorial overview: antivirals and resistance: advances and challenges ahead. 8, iv(2014).
45. Agostini, M. L. *et al.* Coronavirus susceptibility to the antiviral remdesivir (GS-5734) is mediated by the viral polymerase and the proofreading exoribonuclease. **9** (2018).
46. Mason, S., Devincenzo, J. P., Toovey, S., Wu, J. Z. & Whitley, R. J. *J. A. r. Comparison of antiviral resistance across acute and chronic viral infections*, **158**, 103–112 (2018).
47. Sanjuán, R. & Domingo-Calap, P. Genetic diversity and evolution of viral populations. Reference Module in Life Sciences(2019).
48. Chand, G. B., Banerjee, A. & Azad, G. K. J. P. Identification of novel mutations in RNA-dependent RNA polymerases of SARS-CoV-2 and their implications on its protein structure. **8**,e9492(2020).
49. Plotkin, J. B. & Dushoff, J. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. Proceedings of the National Academy of Sciences 100, 7152–7157(2003).

Figures

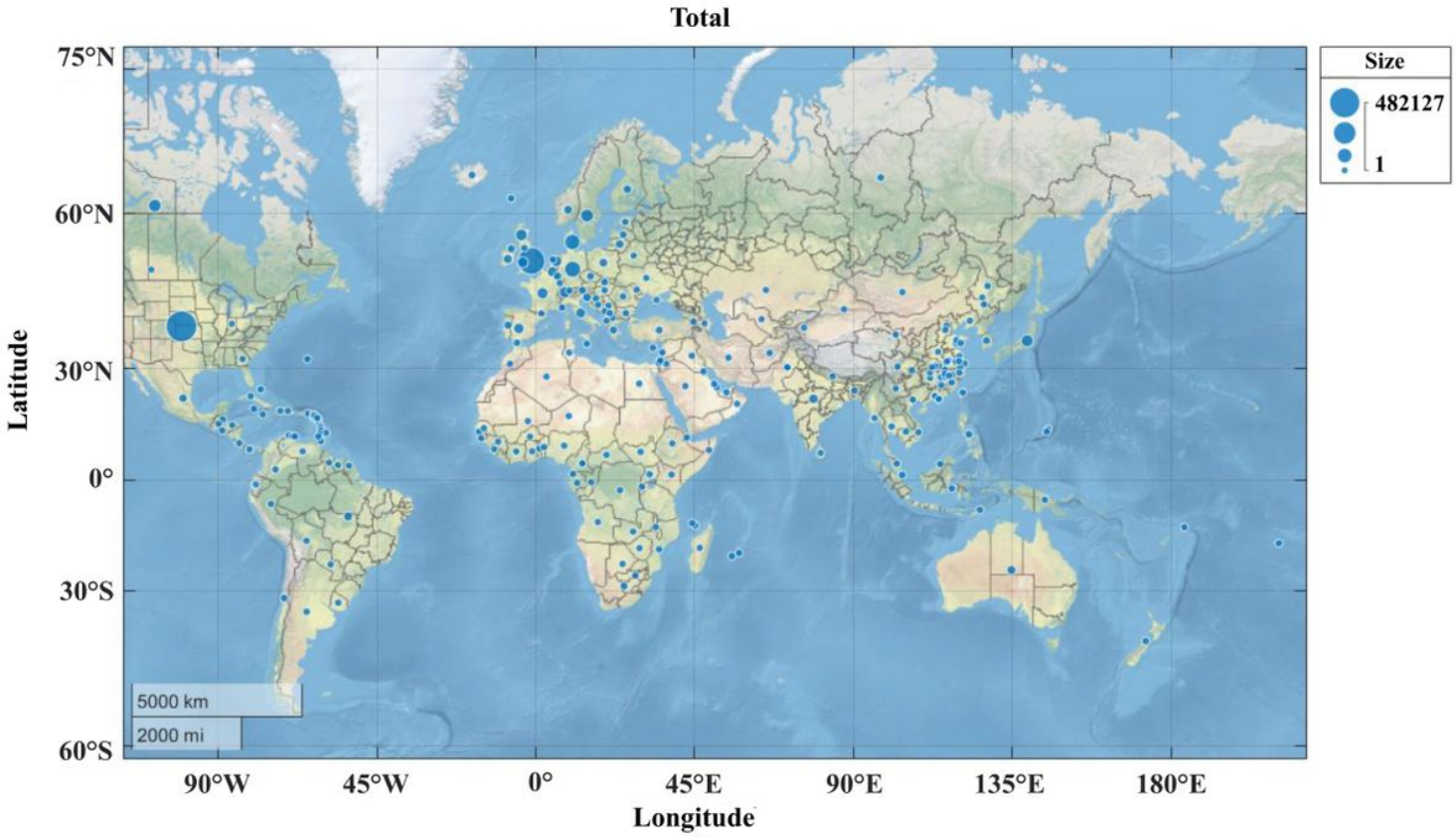

Figure 1

Tracking Sars-Cov-2 NSP-12 distribution as of January 2020. Each circle illustrated two elements, including identifying the sequence's geographical area and the number of mutations. The larger size of the circle represented the more significant number of sequences presented in that area.

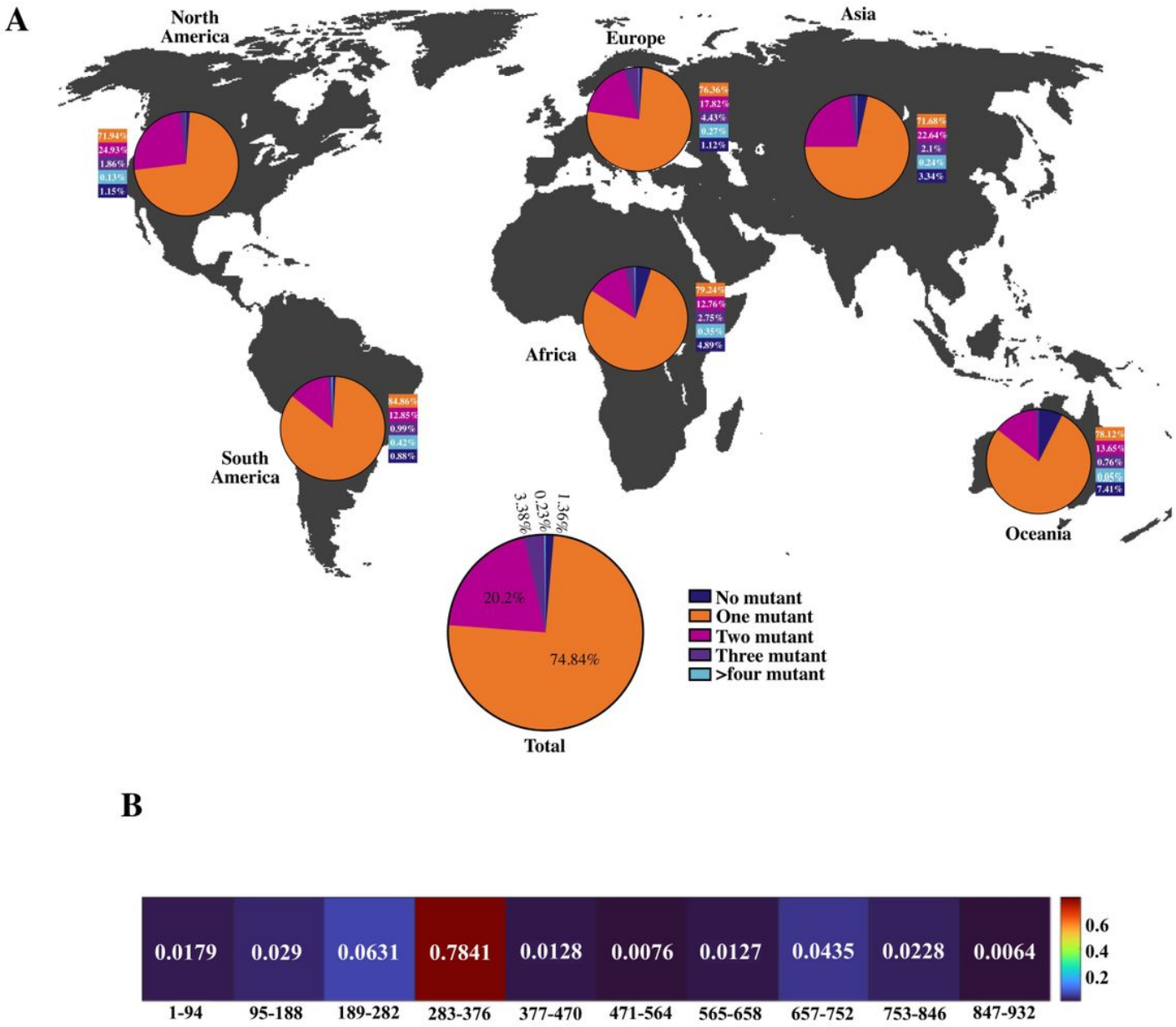

Figure 2

A) Pie chart plot of the number of mutations in NSP-12 of SARS-COV-2 as of December 2019. The dark blue, orange, purple, blue, and light blue colors represent zero, one, two, three, and four or more mutations. B) The heat map indicates the rate of mutation per 100 amino acids of the NSP-12 protein. The highest frequency rate occurred in the 301 to 400 amino acid sequence of the NSP-12 protein.

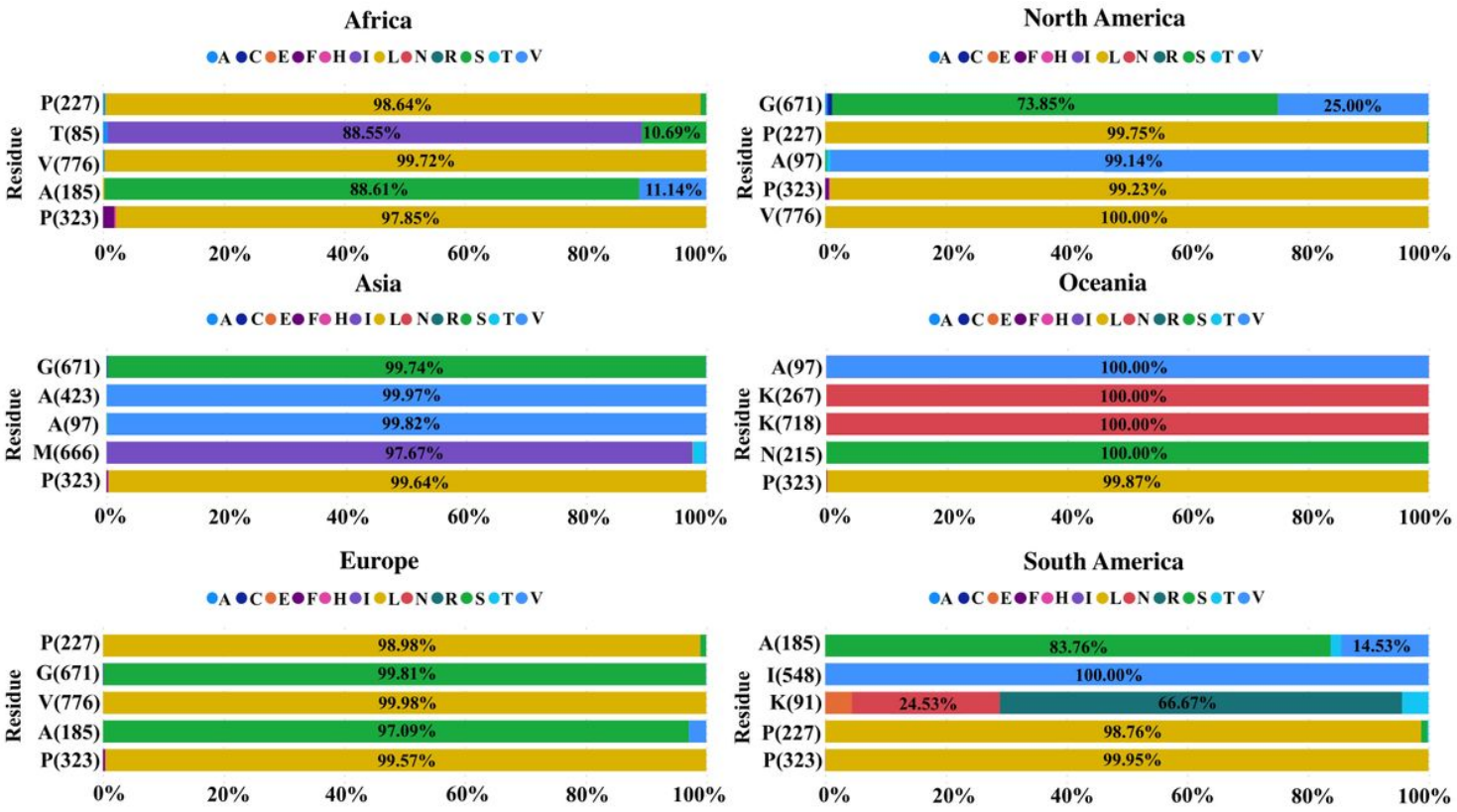

Figure 3

NSP-12 top-five mutation with the highest frequency worldwide and geographic areas including North America, South America, Europe, Asia, Oceania, and Africa. The position of altered amino acids and substituted ones is shown differently based on the frequency percentage of substituted AA. The mutation frequency was estimated for each of them by normalizing the number of genomes carrying a given mutation in a desired geographic area.

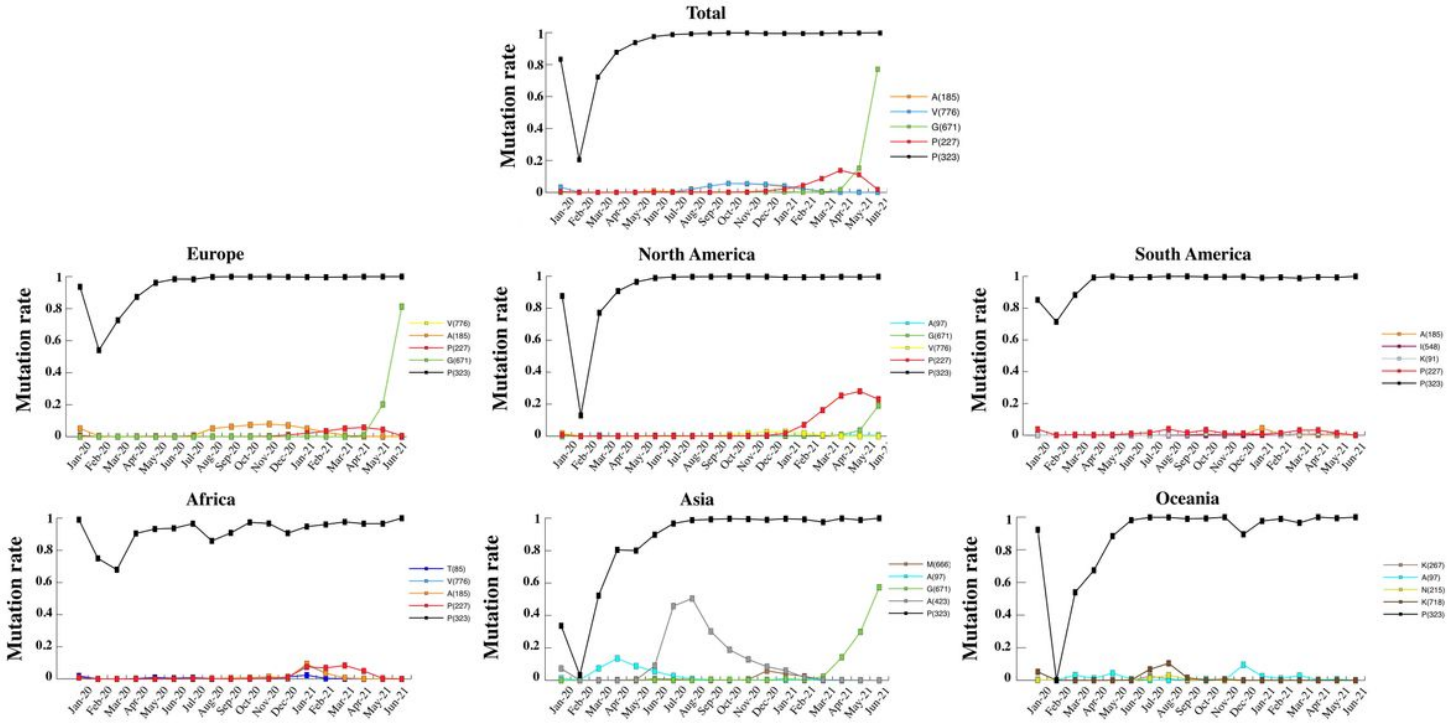

Figure 4

Plots time evolution trajectories of top five high-rate mutations of NSP-12 worldwide and different geographic areas including North America, South America, Europe, Asia, Oceania, and Africa. According to the month of sample collection, data is computed as the number of AASs having a given mutation over the total number of AASs according to the month of sample collection.

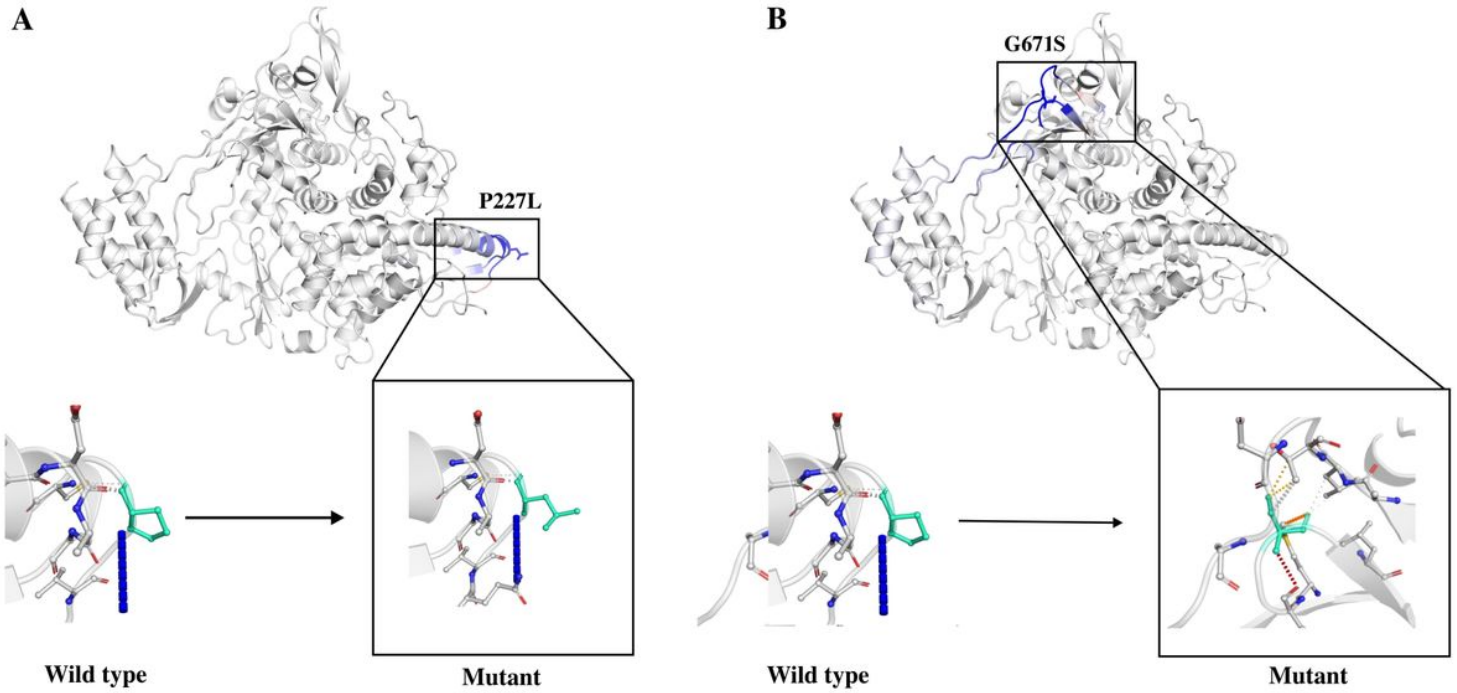

Figure 5

Impact of P227L (A) and G671S (B) mutations on structural dynamics of RdRp protein. The blue color represents a rigidification of the Structure of mutant NSP-12 Structure. Wild-type and substituted residues are colored in light-green. They are also represented as sticks alongside the surrounding residues involved in any interaction.

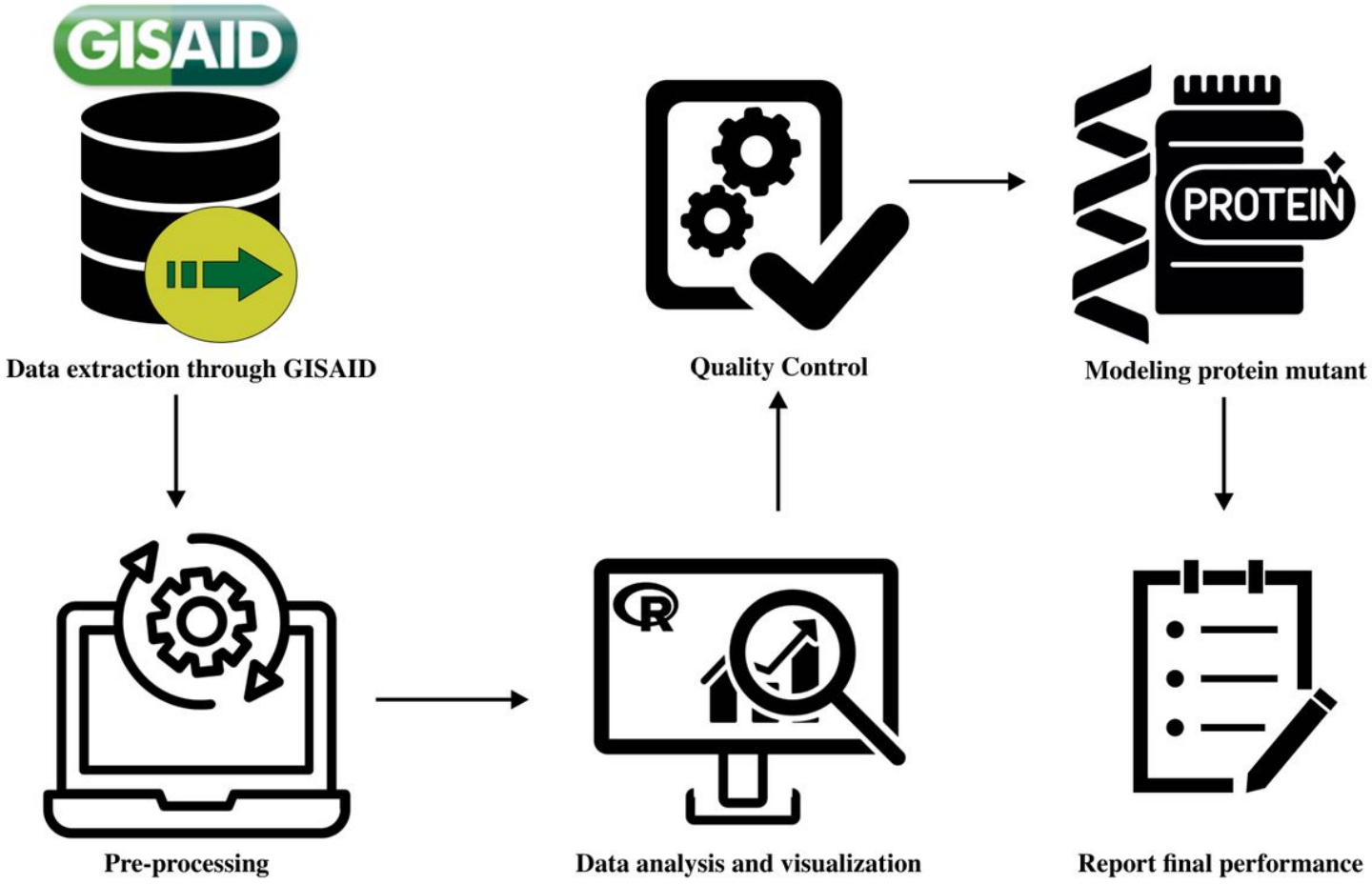

Figure 6

Data processing workflow used to demonstrate and validate NSP-12 AAs mutations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additonalfile1A.Africa.xlsx
- Additonalfile1B.Asia.xlsx
- Additonalfile1C.Europe.xlsx
- Additonalfile1D.NorthAmerica.xlsx
- Additonalfile1F.SouthAmerica.xlsx
- Additonalfile1G.Total.xlsx

- Additionalfile1and2legends.docx
- Additonalfile2A.Africa.xlsx
- Additonalfile2B.Asia.xlsx
- Additonalfile2C.Europe.xlsx
- Additonalfile2D.NorthAmrica.xlsx
- Additonalfile2E.Oceania.xlsx
- Additonalfile2F.SouthAmerica.xlsx
- Additonalfile2G.Total.xlsx