Received: 16 May 2022 | Revised: 27 November 2022 | Accepted: 27 December 2022

DOI: 10.1002/hsr2.1052

ORIGINAL RESEARCH

Health Science Reports Open Access WILEY

Characterization and phylogenetic analysis of Iranian SARS-CoV-2 genomes: A phylogenomic study

Nasrin Aliabadi | Marzieh Jamaliduost | Gholamreza Pouladfar | Nahid H. Marandi | Mazyar Ziyaeyan

Department of Clinical Virology, Clinical Microbiology Research Center, Shiraz University of Medical Sciences, Namazi Hospital, Shiraz, Iran

Correspondence
Mazyar Ziyaeyan, Department of Clinical Virology, Clinical Microbiology Research Center, Shiraz University of Medical Sciences, Namazi Hospital, Shiraz, Iran.
Email: ziyaeyanm@sums.ac.ir

Abstract

Background and Aim: Characterization of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) based on analyzing the evolution and mutations of viruses is crucial for tracking viral infections, potential mutants, and other pathogens. The purpose was to study the complete sequences of SARS-CoV-2 to reveal genetic distance and mutation rate among different provinces of Iran.

Methods: As of March 2020–April 2021, a total of 131 SARS-CoV-2 whole genome sequences submitted from Tehran and 133 SARS-CoV-2 full-length sequences from 24 cities with high coverage submitted to EpiCoV GISAID database were analyzed to infer clades and mutation annotation compared with the wild-type variant Wuhan-Hu-1.

Results: The results of variant annotation were revealed 11,204 and 9468 distinct genomes were identified among the samples from different cities and Tehran, respectively. The phylogenetic analysis of genomic sequences showed the presence of eight GISAID clades, namely GH, GR, O, GRY, G, GK, L, and GV, and six Nextstrain clades; that is, 19A, 20A, 20B, 20I (alpha, V1), 20H (Beta, V2), and 21I (Delta) in Iran. The GH (GISAID clade), 20A (Nextstrain clade), and B.1 (Pango lineage) were predominant in Iran. Notably, analysis of the spike protein revealed D614G mutation (S_D614G) in 56% of the sequences. Also, the delta variant of the coronavirus, the super-infectious strain that was first identified among the sequences submitted from the southern cities of the country such as Zahedan, Yazd and Bushehr, and most likely from these places to other cities of Iran as well has expanded.

Conclusions: Our results indicate that most of the circulated viruses in Iran in the early time of the pandemic had collected in eight GISAID clades. Therefore, a continuous and extensive genome sequence analysis would be necessary to understand the genomic epidemiology of SARS-CoV-2 in Iran.

KEYWORDS

clade, genome sequence, Iran, mutation, phylogenetic analysis, SARS-CoV-2, spike protein

1 | BACKGROUND

Human coronaviruses (HCoVs) are associated with upper respiratory tract infections (RTIs). In late December 2019, China's Centers for Disease Control notified suspected cases of pneumonia in Wuhan, Hubei Province, China due to a novel coronavirus (COV) referred to as severe acute respiratory syndrome COV-2 (SARS-CoV-2).[1,2] The genome of this new virus was revealed and placed in the public records on the Global Initiative on Sharing All Influenza Data (GISAID) (https://www.who.int/emergencies). SARS-CoV-2, as a highly contagious viral disease, spread across the continents and eventually led to the global epidemic of COVID-19, with at least 212 million cases and more than 4.4 million deaths worldwide at the time of writing this manuscript (August 24, 2021).[3] Iran was one of the first countries to experience a significant outbreak of the virus. Iran, after China and Italy, reported the largest number of COVID-19 cases (n = 14,991) in the Middle East region up to March 2020.[4] In this country, the largest number of cases was reported in Tehran (1945) followed by Qom (712) and Mazandaran (633). Qom, Semnan, and Markazi provinces also reported 27.2–34.9 cases/100,000 population. As per the population of GIS-based maps, the virus was spread from northern central provinces such as Tehran and Qom.[5] Until October 2021, 5.88 million cases and 126,716 deaths were recorded in Iran (https://www.worldometers.info/coronavirus/country/iran/).

COVID-19 vaccines have been found to provide minimal protection against the new Variants Of Concern (VOCs). On the other hand, there is a high chance of viral mutations due to the spread of the virus among the Iranian population as well as the creation of many infections. Transmission can be even easier as more mutations build up.[6] Generally, most viral mutations have small effects on the ability of the virus to cause infection and disease.[7] Nonetheless, depending on where the mutations are located in the virus genome, they may affect such properties as transmission or severity.[8] SARS-CoV-2 is transmitted by spike glycoprotein, which binds to the human receptor Angiotensin-Converting Enzyme 2 (ACE2). Studies have shown that in the VOCs, most mutations in the binding (S1) and fusion (S2) domains have been identified as important contributors to the worrying phenotype of the disease.[9,10] By classifying the variants, it is possible to identify each new variant and determine how the mutation affects the spread of the disease.

Given that Iran has the highest incidence and mortality rates in the Middle East, it is important to identify the types of variants circulating in the country. Analysis of the common types in Middle Eastern countries is required for the development of a vaccine that can treat variations in the area.[4,11] This analysis can be extremely helpful for understanding the whole genome and identifying the regional clades. It has been hypothesized in the current study that the types of SARS-CoV-2 sequences available from different provinces of Iran show an increase in mutation compared to the dominant variants in the world. Bioinformatic tools and samples published by GISAID can be used to determine the compositions of variants in Iran. In this study, we analyzed the complete sequences of SARS-CoV-2 to reveal genetic distance and mutation rate among different provinces of Iran to determine the variety of SARS-CoV-2 variants present in Iranian populations and to learn more about the circulation of the virus in Iran.

2 | METHODS

2.1 | Sample source

The submitted full-length genomes with an average length of more than 29,000 base pairs from Iranian provinces were obtained from the GISAID from March 8, 2020 to April 28, 2021. The whole genome sequence IDs have been listed in Tables S1 and S2.

2.2 | Sample size

A total of 131 samples from Tehran and 134 samples from other provinces as well as Wuhan's reference sequence NC_045512 were downloaded from GISAID.

2.3 | Sample selection

The samples from the provinces of Iran were obtained using such filters as high coverage, complete genome, location, and annotations of SARS-CoV-2 on the GISAID website. The samples obtained between March 8, 2020 and April 28, 2021 were analyzed. The provinces and cities of Iran on GISAID included Abadan, Ahvaz, Alborz, Arak, Ardbil, Babol, Birjand, Bushehr, Esfahan, Gilan, Gorgan, Hamedan, Hormozgan, Karaj, Kashmar, Kerman, Qazvin, Qom, Rasht, Sari, Shiraz, Tabriz, Tehran, Urmia, Yazd, and Zahedan. Most of the samples belonged to Ahvaz, Tehran, Sari, and Birjand (more than 10 samples for each). All the sequences were downloaded from GISAID, concatenated into two multi-sequence files related to Iranian cities and Tehran, and saved in the FASTA format.

2.4 | Multiple sequence alignment

A Multiple Sequence Alignment (MSA) was performed using the samples obtained from the GISAID website via Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/)[12] and Fast Fourier Transformation (MAFFT) online service for large alignment files sizing more than 4 MB (https://mafft.cbrc.jp/alignment). The Clustal Omega algorithm was used to separately align the whole genome of the SARS-CoV-2 samples for each cit. Generally, 4 MB is the maximum capacity of online tools. Thus, the maximum number of samples was utilized on the MAFFT (https://mafft.cbrc.jp/alignment/server/add_fragments). The concatenated files were uploaded on the online tool on the ebi website. In the second step, PEARSON/FASTA were selected as the output parameters. All the other options remained in the default mode. The output was a generated alignment file that contained all the sequences with gaps marked with "-". The sequence output file was obtained from both websites in the FASTA format.

2.5 | Genomic variant identification

The process for identifying the variants from the sequence data was developed by the alignment FASTA file and extraction of single nucleotide polymorphism (SNP) using SNP-identification tools.[13] SNPs were identified as input through a FASTA file. Then, sequence alignment into a Variant Call Format (VCF) was extracted from the program to provide a clear mapping of SNPs. In the VCF file, it is easy to check the SNPs and genotype locations in each sample. It shows the rows and columns related to each unique variant as well as the genotype at the given site.

2.6 | Genomic variant annotation

SNP-Eff[10] has been used as the annotation information to define the variants and genes overlap. The SNP-Eff can also predict the exact effect of the sequence variance on the protein structure and function. SNP-Eff supplies integration into the Galaxy, as third-party tools, resulted in a web-based interface for bioinformatic analysis pipelines. In the present study, the VCF-format data set was uploaded in the Galaxy as the platform. The Wuhan-Hu-1, as the reference strain (GenBank accession number NC_045512.2), was downloaded to create the SNP-Eff database. The genomic variants annotation was done using the "SNP-Eff build." The custom parameters included setting the upstream/downstream length at 5000 bases and setting the base splice site (donor and acceptor) size at 2 bases. After the analysis, the output was the annotation data in the form of VCF and HTML report files. CoVsurver was also utilized to analyze the mutations in the GISAID database.

2.7 | Phylogenetic analysis of SARS-CoV-2 data

Phylogenetic analysis was carried out by the Bayesian Evolutionary Analysis Sample Trees (BEAST) package, v1.10.4. Bayesian analysis of the molecular sequences was performed using Monte Carlo Markov Chains (MCMC).[14] An epidemic analysis was also done following the recommended approach for reconstructing the evolutionary dynamics. The purpose of this work was to obtain an estimate of the origin and spread of the epidemic in Iran. To perform the analysis, use was made of an interactive graphical application (BEAUTi) to design the analysis and generate the control file.

3 | RESULTS

According to the Nextclade quality criteria, the sequences that were suspected to sequencing errors (265/1 sequences) were excluded. A total of 131 SARS-CoV-2 whole genome sequences submitted from Tehran and 133 SARS-CoV-2 full-length sequences from 24 cities were analyzed. MSA and variant calling were performed separately and successfully on the alignment files related to Tehran and the 24 cities. Once these processes were completely performed, variant annotation was done as a crucial step in the analysis of the genome. Totally, 11,204 and 9468 distinct genomes were identified among the samples from different cities and Tehran, respectively (Table 1). The nucleotide changes in the 264 samples have been presented in Table 2. Accordingly, most changes (649 changes) were related to C>T mutations.

TABLE 1 The frequency of the mutation types found in the Iranian SARS-CoV-2 isolates' genomes

	Cities	Tehran
Downstream gene variants	4494	4253
Intergenic region	64	70
Missense variants	1154	958
Start lost	1	0
Stop gained	8	6
Synonymous variants	857	721
Upstream gene variants	4626	3460

Abbreviations: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

TABLE 2 Detecting the Iranian SARS-CoV-2 SNPs of the 264 samples

Reference nucleotide	Mutated nucleotide	Numbers
C	T	649
G	T	273
T	C	163
A	G	170

Abbreviation: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; SNPs, single nucleotide polymorphisms.

Using CoV-GLUE (http://cov-glue.cvr.gla.ac.uk), the genomic distributions of the SNPs in the Iranian cities (≥25 cities) were 1397G>A, 3037C>T, 14408C>T, 23403A>G, 14408C>T, 25563G>T, 18877C>T, and 28881G>C (Figure 1). The most prevalent co-mutations found in both file sequences from the 24 cities and Tehran were 3037C>T, 14408C>T, and 23403A>G with more than 90 variations (Table 3).

According to the CoVsurver analysis (gisaid.org/covsurver), the isolated strains categorized into two groups of Tehran and different cities of Iran showed 28.2% of the mutations in the spike glycoprotein, 39% at the ORF1ab gene, and 32.7% at the end of the genomes (Figure 2).

In addition to the mutations found in all SARS-CoV-2 sequences, some less common mutations and some new mutations were identified. Novel mutations were detected in nsp2, nsp5, nsp6, and nsp13 gene regions. A novel amino acid replacement I78N in nsp5 (EPI_ISL_2795653) was also detected in one of the isolated strains from Babol. Surprisingly, the several novel amino acid replacements

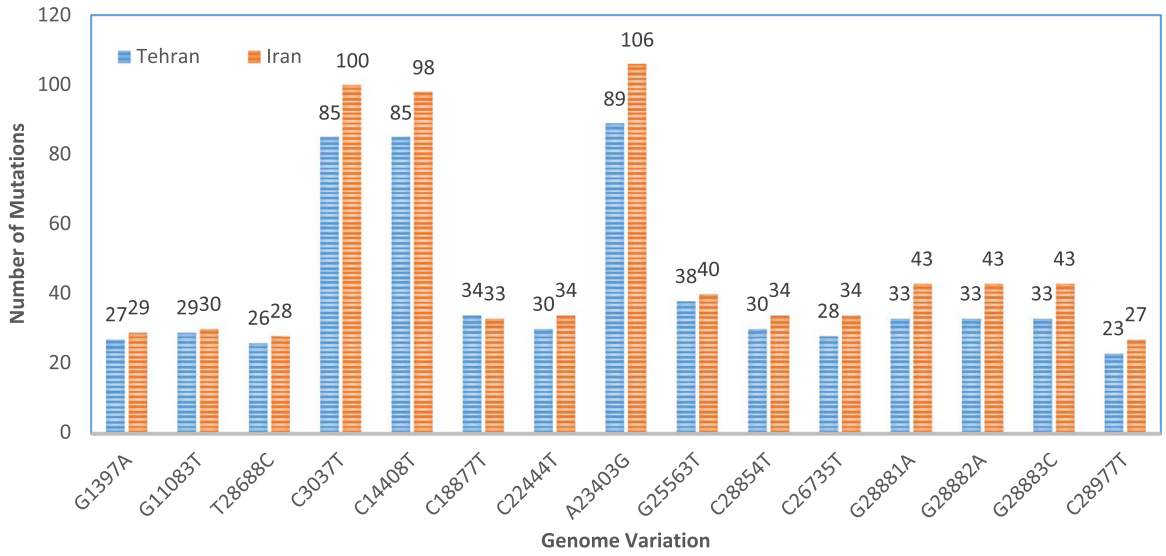

FIGURE 1 Distribution of the most common variations in the 264 sequences from the 24 cities and Tehran.

TABLE 3 Pangolin lineage distribution of the circulating strains of SARS-CoV-2 in Iran

Lineage (Tehran's samples)	Number	Lineage (cities' samples)	Number
B.1.1.413	27	B.1.1	23
B.1.1.7	21	B.1.1.7	19
B.4	17	B.4	18
B.1.36	15	B.1.36	17
B.1.36.7	12	B.1	15
B.1	12	B.1.36.7	11
B.1.1	10	B.1.617.2	10
B.1.533	3	B.1.9.5	4
B.4.8	3	B.4.8	3
B.1.438.1	2	B	2
B.1.9.5	2	B.1.351	2
B.1.1.317	2	B.1.22	1
B.1.1.326	1	B.1.36.9	1
A.23.1	1	B.1.1.316	1
B.1.36.9	1	B.1.1.317	1
B.1.229	1	B.1.36.8	1
B.1.210	1	B.1.243	1
B.1.468	1	B.1.1.194	1
		AY.4	1
Total	131		133

S232P/S in nsp2, D1047D/V, T1797T/S, and V1885A in nsp3, F251L/F in nsp6, and T416V, N423K/N, and L417V in nsp13 were unique to a single sample from Hormozgan (ID: EPI_ISL_1533610). These mutations were not found in other sequences stored in the GISAID database. Mutations found in these genetic regions might have important functional implications that have to be evaluated in the context of vaccine development and therapeutic options.

According to the GISAID nomenclature system, clustering the samples based on the GISAID scheme indicated major GISAID clades. Most of the samples submitted to GISAID related to the two categories of cities and Tehran belonged to November 2020–February 2021 (Figure 3A). In the 264 sequences related to the cities of Iran and Tehran, seven out of the nine clades classified as per the different GISAID clades for SARS-CoV-2 were detected in Iran: GH (74), GR (62), O (52), GRY (41), G (20), GK (11), L (2), and GV (1). Among these clades, GH (27.5%) was predominant in the sequences related to the cities and Tehran (Figure 3B). Analysis of the individual sequences in each division also revealed the dominance of the GR clade in most cities and provinces. Overall, four clades; that is, GH, GRY, GR, and O, were identified at the beginning of the epidemic from February to November 2020. In May 2020, three separate isolated samples were placed on the GV and L clades and were soon disappeared. Most clades were introduced from November 2020 to February 2021 and were widely circulating in the country.

The maximum likelihood phylogenetic tree by Nextstrain online (https://clades.nextstrain.org) showed six and five clades in the samples obtained from the 24 cities of Iran and Tehran, respectively (Figure 1S a and b). To investigate the prediction of the transmission routes and possible SARS-CoV-2 divergence events, the time-scaled MCC tree

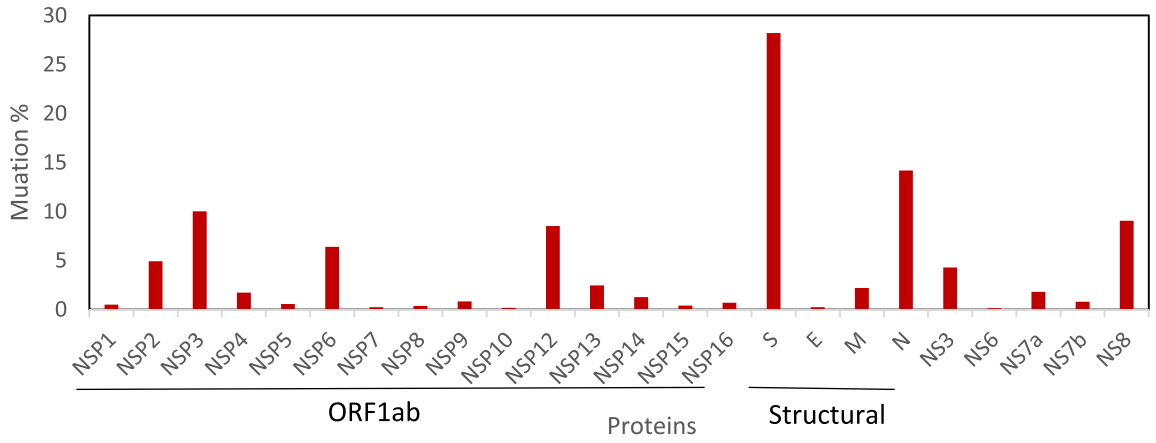

FIGURE 2 Mutational profile of the Iranian SARS-CoV-2 genome. The total number of mutations detected by each encoded protein. SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

(maximum clade credibility) was delineated from the sequence datasets of Iranian cities by whole genomes from Iranian isolates, with strain Wuhan-Hu-1 (NCBI accession number NC_045512.2) being considered the reference strain.

The results of the phylogenetic analysis revealed the fast divergence rate of SARS-CoV-2 into six distinct transmission clusters (Figure 4). Cluster computational analysis showed that the first divergent event, as the most recent common ancestor of the first stage of the outbreak, was merged into clusters one to three from March to August 2020. Cluster I (19A) was composed of almost a larger number of viruses from Alborz, Birjand, Qom, Gilan, and Khuzestan. Cluster II (20A) consisted of more diverse viruses across Iran including Karaj, Esfahan, Shiraz, Hamedan, Gilan, Sari, Qazvin, and Hormozgan, with the majority of the viruses being from Sari and Qazvin. Cluster VI (21A) was dominated by viruses from Yazd, Bushehr, and Zahedan. This was the latest cluster representing the most recently diverged lineage in Iran's population as a common ancestor with the longest branch length up to NC_045512.

Results of Pangolin COVID-19 Lineage Assigner Phylogenetic Assignment of Named Global Outbreak Lineages (https://pangolin.cog-uk.io) revealed a variety of viruses have been found throughout the country possibly due to travels (Table 3). In both sample files related to different cities and Tehran, nearly 20 lineages were observed with a combination of both A and B lineage viruses from March 2020 to April 2021. In the cities, the most frequent lineage was B.1.1 (23 sequences) followed by B.1.1.7 (19 sequences) and B.4 (18 sequences). Among the 131 samples from Tehran, 21 lineages were detected, with B.1.1.413 (27 sequences) and B.1.1.7 (21 sequences) having the highest frequency among the sequences. In this context, the identification of some lineages, such as A.23.1 and AY.4 was noteworthy.

4 | DISCUSSION

In this study, 264 whole genome sequences of SARS-COV-2 from 25 cities in Iran that were available at GISAID were analyzed from March 2020 to April 2021. Genomic variations, phylogenetics, and evolution were assessed through comparison with the Wuhan-Hu-1 (NC_045512) reference strain. These findings presented the genetic background of the molecular epidemiological trends of SARS-CoV-2 in Iran.

As revealed in Iran in February 2020, there were many ups and downs over the past 2 years. The results of the phylogenetic analysis indicated genetic diversity among the strains submitted from Iran using the GISAID,[15] BEAUti, BEAST, and Figtree Software packages (https://beast.community/programs), presenting several strains of SARS-COV-2 among Iranian populations.

During the COVID-19 pandemic, the frequency of mutations in the genome increased with more than 3000 unique point mutations of viruses isolated around the world. The MCC tree showed six clusters of Iranian strains on a time scale without defining time. However, from the first 6 months of 2021, it formed a distinct cluster called the VI cluster in the present study. SNPs and nucleotide distance matrices were the causes of multiple clusters. However, since this study was not conducted in a specific time frame to identify a regional cluster, several strains with a risk of rapid evolution were supposed to circulate across the country. GISAID,[15] Nextstrain,[16] and lineage[17] nomenclature systems indicated several lineages and clades among Iranian virus populations distinguished by the mutation of amino acid sequences. In a study conducted in the early stage of the epidemic in Asia, 11083G>T 1397G>A, 28688T>C, and 29742G>T mutations were prevalent belonged to clade O, while in our study, C3037>T, C14408>T, and A23403>G mutations related to clade GH and GR clades were more frequent.[18] Taken together, all of these evidence indicate the circulated Iran viruses in the early pandemic had evolved and had differences with the early ancestors. In this regard, more than 27.5% of the Iranian strains belonged to the GH clade. It should be noted that the GH and GR clades are two main branches of the G clade, newly introduced as GRY clades. S_D614G mutation is a common distinguishing feature in these three clades. One of the most common mutations that may increase the SARS infection is the S_D614G mutation, which has been observed in 95.12% of the global sequences.[19] The frequency of this mutation was 56% in the 264 samples from Iran.

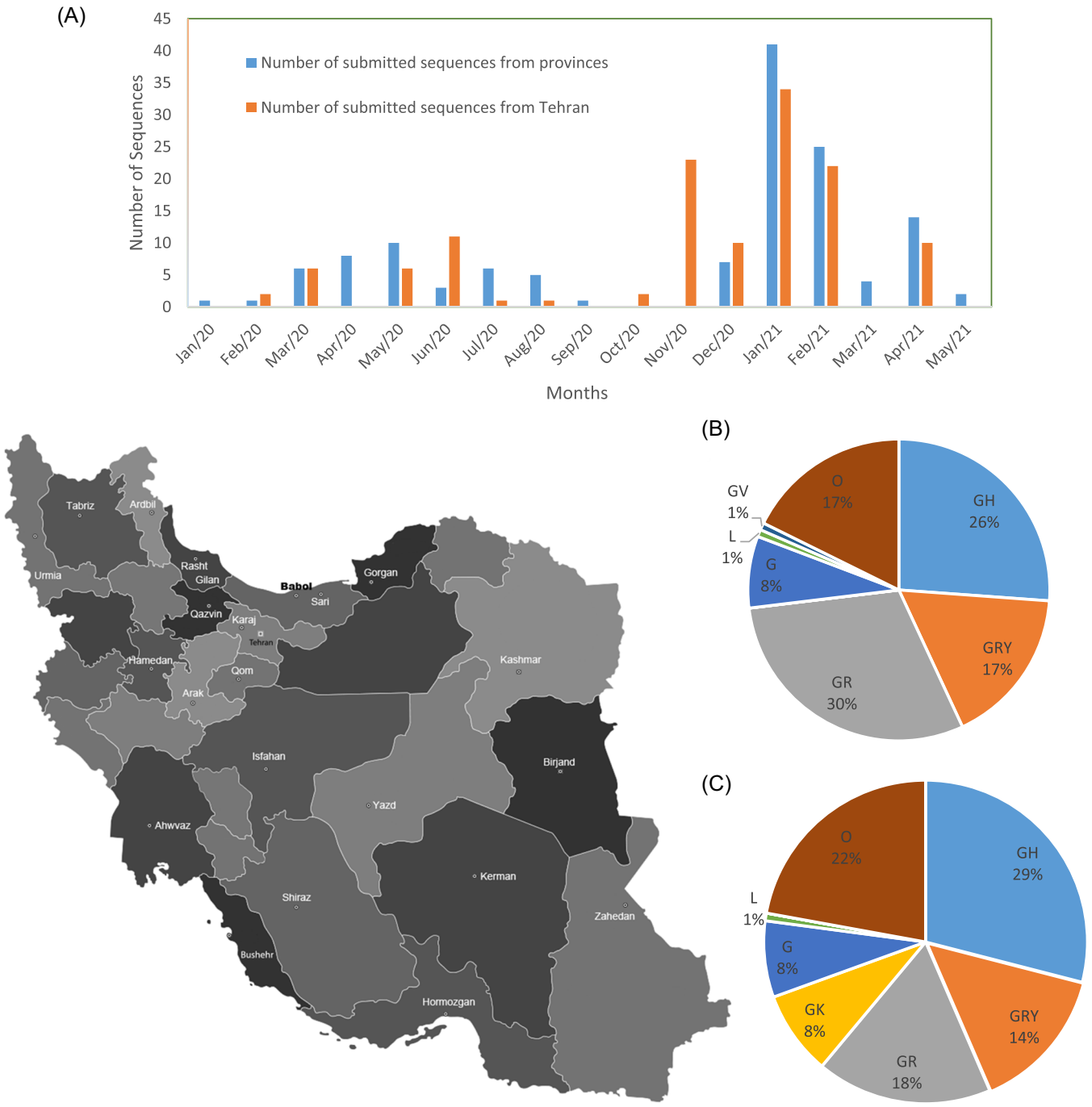

FIGURE 3 GISAID distribution of SARS-CoV-2 genomes in Iran. (A) The number of sequences submitted to the GISAID database monthly. (B) The proportionate prevalence of the eight clades of SARS-CoV-2 detected in different provinces and cities of Iran on GISAID (included Abadan, Ahvaz, Alborz, Arak, Ardbil, Babol, Birjand, Bushehr, Esfahan, Gilan, Gorgan, Hamedan, Hormozgan, Karaj, Kashmar, Kerman, Qazvin, Qom, Rasht, Sari, Shiraz, Tabriz, Tehran, Urmia, Yazd, and Zahedan). (C) The proportionate distribution of the seven clades of SARS-CoV-2 in Tehran. GISAID, Global Initiative on Sharing All Influenza Data; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

Glycoprotein S plays an important role in viral infection and virus pathogenesis. Spike protein contains two functional subunits called S1 and S2. The S1 subunit includes two parts: the N-Terminal Domain (NTD) and the Receptor Binding Domain (RBD). The S1 subunit binds to the host cell receptor, human ACE-2, on epithelial cells. This triggers the cleavage of ACE2, which plays a key role in the infection process. The S2 subunit plays a pivotal role in mediating viral fusion with and invasion of the host cell including the Fusion Peptide (FP), Heptad Repeat 1 (HR1), central helix, binding domain, Heptad Repeat 2 (HR2), Transmembrane Domain (TM), and Cytoplasmic Domain (CD).[20,21] According to the present study findings, clade GR with the N_G204R mutation in the nucleocapsid protein

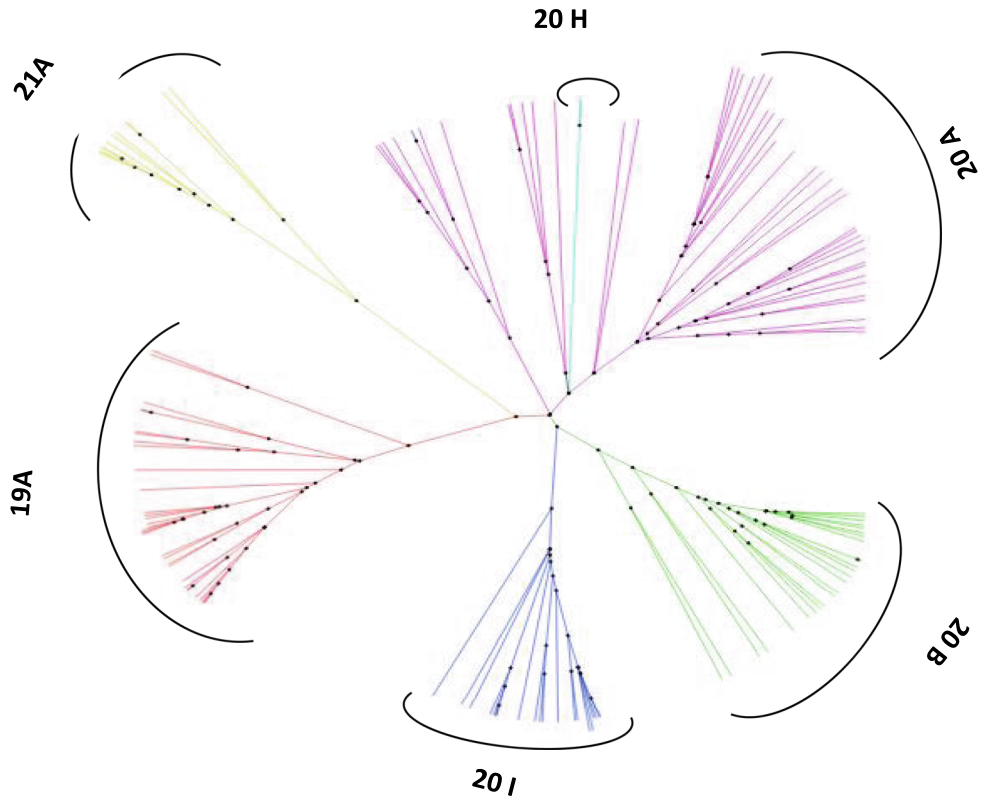

FIGURE 4 A time-scaled MCC phylogenetic tree of the SARS-CoV-2 sequences in Iran. The phylogenetic tree was built by an approximately maximum likelihood method on the full genomes of 133 SARS-CV-2 from the cities of Iran. The six lineage groups were classified into six clusters: Cluster I, II, and VI (NC_045512 was the reference). 19A, red; 20A, purple; 20B, green; 20I (Alpha, V1), dark blue; 20H (Beta, V2), light blue; 21I (Delta), yellow. MCC< maximum clade credibility; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

along with the extensive S_D614G mutation accounted for approximately 3% of the mutations in the Iranian sequences. Similarly, the Nextstrain clade 20B and the B.1.1 and B.1.1.413 lineages were very common in the cities' and Tehran's sequences. Similarly, a study that investigated SARS-CoV-2 sequences collected in the Eastern Mediterranean Region found that more than 65.8% of the viruses belong to clades 20A, 20B, and 20C (GISAID clades GR, GH, G and GV).[22] However, the emergence of B.1.1.7/GRY (UK type/S_N501Y mutation) lineage in Iran in late 2020 became of much concern. Among the three types of SARS-CoV-2; i.e., B.1.1.7 (UK variant), B.1.351 (South African variant), and P.1 (Brazilian variant), B.1.351 was detected only among the sequences of Hormozgan and B.1.1.7 was identified in 39 out of the 264 sequences. In a study in Pakistan done from May 01 to June 09, 2021, it was shown that beta and delta strains are present in the cities of Karachi and Islamabad, and the same variants were identified in Hormozgan, a city bordering Pakistan and other southern cities of the country. For the first time in Iran, we have also found one significant mutation E484Q with delta variant (B.1.617.2) in Hormozgan and Yazd, which was most likely transferred to Iran through a passenger from Pakistan.[23] In August 2021, the emergence and spread of SARS-CoV-2, lineage A (A.23.1), with multiple changes in the virus protein and genome were reported in Uganda and Rwanda. The identification of some lineages such as A.23.1 and AY.4 was noteworthy. In almost the same way, Algeria and North African Countries revealed six dominant variants, including B.1, the Delta variants (AY.X, B.1.617.2), C.36, B.1.1.7 and B.1.1. Clades GR, GH and GK were the most frequently identified among the analyzed genomes.[24] These lineages presented distinctive features that were highly infectious and highly transmissible. S_E484K in RBD was yet another important mutation with regard to the immune escape from the immune system,[25,26] as previously found in three VOCs and also observed in both Hormozgan strains in the present study. This mutation helped the virus slip but did not decline the immune-neutralizing activity during recovery or post-vaccination.[25,27] Iranian strains showed other mutations in S protein like L5F, Q675H, and P681H/P681R that were previously found to be linked with increased infectivity under experimental setups and L452R, F490S, S477N, and S151I that escaped two or three of four serum tests.[28] L452R appeared worldwide between December 2020 and February 2021, suggesting that this amino acid replacement was likely the result of viral adaptation due to greater immunity in the population.[29] Among the strains studied in Iran, this mutation was only found in Yazd, Zahedan, and Bushehr, while other cities might not have recorded a sequence in GISAID during the Delta outbreak. In addition, some strains in Iran had 69del, V70del, and Y144del or Y145del, which were shown to be twice as infectious compared to the wild type[30] with reduced sensitivity to

convalescent sera. Furthermore, F157del and R158del were observed in Yazd sequences. Previous studies suggested that mutations in the NTD might affect the high infectivity of the Delta variant. In addition, A475V, V483A, and F490L mutations might affect resistance to some neutralizing antibodies.[31] Many other mutations, such as missense and synonymous mutations like ORF1ab/nsp12_P323L, NS3_Q57H, NS8_R52I, N_S194L, N_R203K, and N_G204R were also observed in the relevant proteins of the Iranian virus populations and have been found to be common globally. The NSP12_P323L mutation exerted a significant effect on protein folding and aggregation.[32] There were several other low-frequency mutations in other parts of the world, and some new mutations occurred in Iran (Table S1). Such mutations seem to be uncommon and may be the result of the viral adaptation to the host's genetic status, environmental conditions, or unidentified causes requiring further investigations.

This survey had some limitations, including inconsistencies in the samples submitted to the GISAID website from different cities and provinces in the past 2 years. In other words, the number of samples from different months was not the same.

5 | CONCLUSION

In summary, this study demonstrated the genomic variation and evolutionary dynamics contributing to the identification and distribution of the SARS-CoV-2 strains and clades in Iran from March 2020 to April 2021. Out of these eight clades and 21 lineages, GR clades and B.1.1 lineage were widespread, which might be the main reason for infecting the community. The potential for viral infection has been accelerated by the widespread distribution of many strains or clades as well as by the simultaneous spread of SARSCoV2 strains, which has been argued to have caused the actual viral storm in such densely populated areas. Yet, the higher rates of infectivity and mortality in Iran compared to many other countries might be attributed to differences in clade pathogenicity or unknown factors. Therefore, comparative genomics may help understand the etiology and pathogenicity of the virus.

AUTHOR CONTRIBUTIONS

Nasrin Aliabadi: Conceptualization; data curation; formal analysis; methodology; software; writing – original draft; writing – review & editing. **Marzieh Jamaliduost**: Investigation; project administration; validation; visualization. **Gholamreza Pouladfar**: Validation; visualization; writing – original draft. **Nahid Heydari Marandi**: Data curation; methodology; software. **Mazyar Ziyaeyan**: Formal analysis; investigation; validation; visualization.

ACKNOWLEDGMENTS

The authors would like to thank Ms. A. Keivanshekouh at the Research Consultation Center (RCC) of Shiraz University of Medical Sciences for her invaluable assistance in editing the manuscript. This research is financially supported by Professor Alborzi Clinical Microbiology Research Center and Shiraz University of Medical Sciences. The funder had no contribution to the study design, data acquisition, analysis, publication decision, and manuscript preparation.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The datasets used in the study are not publicly available due to protect the participants' anonymity but are available on reasonable request.

ETHICS STATEMENT

Under IR, the study design and protocols were ethically approved by the Department of Medical Ethics and Philosophy of Health, Shiraz University of Medical Sciences, Iran.SUMS.REC.1397.901.

TRANSPARENCY STATEMENT

The lead author Mazyar Ziyaeyan affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

ORCID

Nasrin Aliabadi http://orcid.org/0000-0002-2202-4148
Marzieh Jamaliduost http://orcid.org/0000-0002-7034-1236

REFERENCES

1. Owusu M, Annan A, Corman VM, et al. Human coronaviruses associated with upper respiratory tract infections in three rural areas of Ghana. *PLoS One*. 2014;9(7):e99782.
2. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China. *N Eng J Med*. 2020;382:727-733.
3. Mohapatra RK, Pintilie L, Kandi V, et al. The recent challenges of highly contagious COVID-19, causing respiratory infections: symptoms, diagnosis, transmission, possible vaccines, animal models, and immunotherapy. *Chem Biol Drug Des*. 2020;96(5):1187-1208.
4. Bindayna KM, Crinion S. Variant analysis of SARS-CoV-2 genomes in the Middle East. *Microb Pathog*. 2021;153:104741.
5. Arab-Mazar Z, Sah R, Rabaan AA, Dhama K, Rodriguez-Morales AJ. Mapping the incidence of the COVID-19 hotspot in Iran - implications for travellers. *Travel Med Infect Dis*. 2020;34:101630-101631.
6. Moya A, Holmes EC, González-Candelas F. The population genetics and evolutionary epidemiology of RNA viruses. *Nat Rev Microbiol*. 2004;2(4):279-288.
7. Lauring AS, Hodcroft EB. Genetic variants of SARS-CoV-2-what do they mean? *JAMA*. 2021;325(6):529-531.
8. Banada P, Green R, Banik S, et al. A simple RT-PCR melting temperature assay to rapidly screen for widely circulating SARS-CoV-2 variants [published online ahead of print September 20, 2021]. *medRxiv*. 2021.
9. Iacobucci G. COVID-19: new UK variant may be linked to increased death rate, early data indicate. *BMJ*. 2021;372:n230.
10. Giron CC, Laaksonen A, Barroso da Silva FL. Up state of the SARS-COV-2 spike homotrimer favors an increased virulence for new variants. *Front Med Technol*. 2021;3(29):694347.

11. Basheer A, Zahoor I. Genomic epidemiology of SARS-CoV-2 divulge B.1, B.1.36, and B.1.1.7 as the most dominant lineages in first, second, and third wave of SARS-CoV-2 infections in Pakistan [published online ahead of print July 28, 2021]. *medRxiv*. 2021.
12. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*, 7:539.
13. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012;6(2):80-92.
14. Suchard MA, Lemey P, Baele G, et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol*. 2018;4(1):vey016.
15. Shu Y and McCauley J. GISAID: global initiative on sharing all influenza data - from vision to reality. *Euro Surveill*. 2017;22(13):30494.
16. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018;34(23):4121-4123.
17. Rambaut A, Holmes EC, O'Toole Á, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*. 2020;5(11):1403-1407.
18. Moradi J, Moghoofei M, Doroudian M, Abiri R. Genomic characterization and phylogenetic analysis of SARS-CoV-2 during the early phase of the pandemic in Asia. *Preprints*. 2020;2020050100.
19. Conceicao C, Thakur N, Human S, et al. The SARS-CoV-2 Spike protein has a broad tropism for mammalian ACE2 proteins. *PLoS Biol*. 2020;18(12):e3001016.
20. Korber B, Fischer WM, Gnanakaran S, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*. 2020;182(4):812-827.
21. Shang JA-O, Wan Y, Luoq C, Li F. Cell entry mechanisms of SARS-CoV-2. *Biol Sci*, 117(21):11727-11734.
22. Umair M, Ikram A, Salman M, et al. Genomic surveillance reveals the detection of SARS-CoV-2 delta, beta, and gamma VOCs during the third wave in Pakistan. *J Med Virol*. 2022;94:1115-1129.
23. Ahmad SU, Hafeez Kiani B, Abrar M, et al. A comprehensive genomic study, mutation screening, phylogenetic and statistical analysis of SARS-CoV-2 and its variant omicron among different countries. *Journal of Infection and Public Health*. 2022;15(8):878-891.
24. Menasria T, Aguilera M. Genomic diversity of SARS-CoV-2 in Algeria and north African countries: what we know so far and what we expect? *Microorganisms*. 2022;10(2):467. doi:10.3390/microorganisms10020467
25. Starr TN, Greaney AJ, Hilton SK, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*. 2020;182(5):1295-1310.e20.
26. Wise J. Covid-19: the E484K mutation and the risks it poses. *BMJ*. 2021;372:n359.
27. Jangra S, Ye C, Rathnasinghe R, et al. The E484K mutation in the SARS-CoV-2 spike protein reduces but does not abolish neutralizing activity of human convalescent and post-vaccination sera [published online ahead of print January 29, 2021]. *medRxiv*. 2021. doi:10.1101/2021.01.26.21250543
28. Li Q, Wu J, Nie J, et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell*. 2020;182(5):1284-1294.
29. Harvey WT, Carabelli AM, Jackson B, et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol*. 2021;19(7):409-424.
30. Kemp SA, Collier DA, Datir RP, et al. SARS-CoV-2 evolution during treatment of chronic infection. *Nature*. 2021;592(7853):277-282.
31. Wang L, Wang L, Zhuang H. Profiling and characterization of SARS-CoV-2 mutants' infectivity and antigenicity. *Signal Transduct Target Ther*. 2020;5(1):185.
32. Maitra A, Sarkar MC, Raheja H, et al. Mutations in SARS-CoV-2 viral RNA identified in Eastern India: possible implications for the ongoing outbreak in India and impact on viral structure and host susceptibility. *J Biosci*. 2020;45(1):76.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Aliabadi N, Jamaliduost M, Pouladfar G, Marandi NH, Ziyaeyan M. Characterization and phylogenetic analysis of Iranian SARS-CoV-2 genomes: a phylogenomic study. *Health Sci Rep*. 2023;6:e1052. doi:10.1002/hsr2.1052