Received: 17 June 2021 | Accepted: 6 July 2021

DOI: 10.1002/jmv.27188

RESEARCH ARTICLE

JOURNAL OF MEDICAL VIROLOGY WILEY

Genomic characterization of SARS-CoV-2 isolates from patients in Turkey reveals the presence of novel mutations in spike and nsp12 proteins

Erdem Sahin[1] | Gulendam Bozdayi[1] | Selin Yigit[1] | Hager Muftah[1] | Murat Dizbay[2] | Ozlem G. Tunccan[2] | Isil Fidan[1] | Kayhan Caglar[1]

[1]Division of Medical Virology, Department of Medical Microbiology, Faculty of Medicine, Gazi University, Ankara, Turkey

[2]Department of Infectious Diseases and Medical Microbiology, Faculty of Medicine, Gazi University, Ankara, Turkey

Correspondence

Gulendam Bozdayi, Application and Research Hospital of Gazi University, Faculty of Medicine, Medical Virology and Covid Laboratory, Block C, Floor 5, Besevler, Ankara 06500, Turkey.
Email: gulendam@gazi.edu.tr

Funding information

Gazi University Scientific Research Projects Unit, Grant Number: 01/2020-20

Abstract

Novel mutations have been emerging in the genome of severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2); consequently, the evolving of more virulent and treatment resistance strains have the potential to increase transmissibility and mortality rates. The characterization of full-length SARS-CoV-2 genomes is critical for understanding the origin and transmission pathways of the virus, as well as identifying mutations that affect the transmissibility and pathogenicity of the virus. We present an analysis of the mutation pattern and clade distribution of full-length SARS-CoV-2 genome sequences obtained from specimens tested at Gazi University Medical Virology Laboratory. Viral RNA was extracted from nasopharyngeal specimens. Next-generation sequencing libraries were prepared and sequenced on Illumina iSeq 100 platform. Raw sequencing data were processed to obtain full-length genome sequences and variant calling was performed to analyze amino acid changes. Clade distribution was determined to understand the phylogenetic background in relation to global data. A total of 293 distinct mutations were identified, of which 152 missense, 124 synonymous, 12 noncoding, and 5 deletions. The most frequent mutations were P323L (nsp12), D614G (ORF2/S), and 2421C>T (5′-untranslated region) found simultaneously in all sequences. Novel mutations were found in nsp12 (V111A, H133R, Y453C, M626K) and ORF2/S (R995G, V1068L). Nine different Pangolin lineages were detected. The most frequently assigned lineage was B.1.1 (17 sequences), followed by B.1 (7 sequences) and B.1.1.36 (3 sequences). Sequence information is essential for revealing genomic diversity. Mutations might have significant functional implications and analysis of these mutations provides valuable information for therapeutic and vaccine development studies. Our findings point to the introduction of the virus into Turkey through various sources and the subsequent spread of several key variants.

KEYWORDS

full-length genome, mutation analysis, next-generation sequencing, phylogenetic tree, SARS-CoV-2, Turkey

1 | INTRODUCTION

Coronaviruses are a large family of RNA viruses, which can cause mild to severe respiratory infections in humans.[1] The severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2) is a newly discovered virus that causes the respiratory disease called the coronavirus disease 2019 (COVID-19) and was first reported in Wuhan City, China, in mid-December 2019.[2] It was declared as a pandemic by the World Health Organization on March 11, 2020. The first case of SARS-CoV-2 infection in Turkey was identified on March 11, 2020, in a person with recent travel history to Europe, and the first death was announced on March 17, 2020. Since the first confirmed case, the number of cases has risen rapidly both globally and in Turkey, and it continues to rise.[3] According to data by the Turkish Ministry of Health, over 4 million cases have been reported in Turkey as of June 2021, including over 40 000 deaths.[4] Significant public health measures have been implemented all over the country to have control over the number of cases. Following the onset of the COVID-19 pandemic, virology laboratories were reorganized to ensure early diagnosis of the infection. In this context, various diagnostic techniques such as real-time polymerase chain reaction (PCR), rapid antibody and enzyme-linked immunosorbent assay tests have been used in the laboratories for the rapid and accurate diagnosis of COVID-19.[5]

SARS-CoV-2 is an RNA virus that frequently develops mutations. New mutations have been emerging in the genome of SARS-CoV-2, and evolving of more virulent and treatment-resistant strains has the potential to increase transmissibility and mortality rates.[6] The molecular characterization of full-length SARS-CoV-2 sequences is necessary to understand the origin, transmission pathways and to identify the mutations that affect the transmissibility and pathogenicity of the virus. In addition, the analysis of genomic sequence data will provide valuable information for therapeutic and vaccine development studies.

Throughout the period of the COVID-19 pandemic, genetic variants of SARS-CoV-2 have emerged and spread around the world. Genetic differences between viruses have been compared to identify variants and their relation to each other to understand the transmission trends. The availability of virus genome sequences and online sequence sharing tools allowed for close monitoring of SARS-CoV-2 molecular epidemiology (Table 1). Several studies have been conducted during the pandemic to reveal the genomic characterization and phylogenetic relationships of SARS-CoV-2 genomes circulating in Turkey. The first SARS-CoV-2 genome sequence shared from Turkey was made available via Global Initiative on Sharing All Influenza Data (GISAID) on March 25, 2020, and over 4000 sequences from Turkey had been released by May 31, 2021, in GISAID (Figure 1). Molecular characterization and phylogenetic analysis of the first SARS-CoV-2 full-length genomes isolated in Turkey revealed virus introductions from Europe into Turkey.[7] SARS-CoV-2 genomes isolated and sequenced in Turkey revealed that the virus was introduced to the country earlier than the first reported case.[8] Since the start of the pandemic, the virus has undergone multiple nucleotide substitutions, including silent and missense mutations, according to the genomic characterization investigations of SARS-CoV-2 genomes in Turkey.[9,10] According to nomenclature systems, SARS-CoV-2 genomes isolated in Turkey are found in most lineages, and data on virus genome diversity in Turkey reflects multiple introductions from various sources and subsequent local adaptation of the virus.[8,10]

The first positive case in our virology laboratory was detected on March 23, 2020. Since the beginning of the pandemic, over 80 000 samples have been tested and more than 12 000 of the samples were found positive in our laboratory. In this study, the mutation profile of full-length genome sequence data obtained from specimens tested positive for SARS-CoV-2 in our laboratory is analyzed. We present an overview of the mutation pattern and the phylogenetic analysis of the genomes. We determine the clade distribution according to the three nomenclature systems (Pango lineage, Nextstrain, and GISAID) to understand the phylogenetic context and genomic diversity of the isolates in relation to global data.

2 | MATERIALS AND METHOD

2.1 | Sample collection and storage

The study included 20 male and 25 female patients, a total of 45 samples with average ages of 38.1 (between ages of 1 and 83), tested positive for SARS-CoV-2 with C_t values less than 20, which were

TABLE 1 Current variants of concern (VOC) which are monitored and characterized are listed

Pango lineage[a]–Nextstrain[b] name	First detected country	Sequences isolated in Turkey and their VOC distribution, *n* (%)
B.1.1.7–20I/501Y.V1	UK	571 (51%)
B.1.351–20H/501.V2	South Africa	527 (47%)
B.1.427–20C/S:452R + B.1.429–20C/S:452R	USA (California)	2 (0.2%)
P.1–20J/501Y.V3	Japan/Brazil	22 (2%)

Note: Current prevalence of VOC in Turkey (May 31, 2021).

[a] A proposal for a dynamic nomenclature for SARS-CoV-2 lineages to aid genomic epidemiology.[11]

[b] Nextstrain has organized the variants into clades defined by specific signature mutations. There are currently 11 major clades defined by Nextstrain.[12]

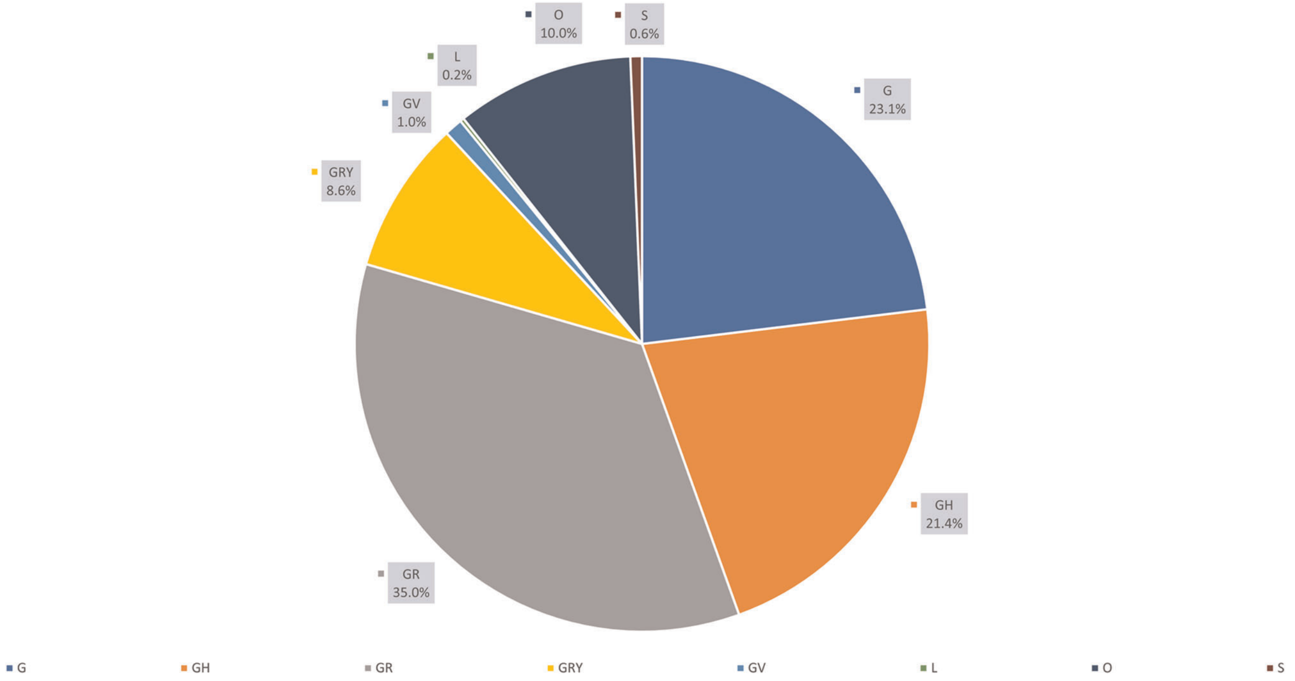

FIGURE 1 Global Initiative on Sharing All Influenza Data (GISAID) clade classification of severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2) sequences from Turkey. GISAID developed a nomenclature system for major clades based on marker mutations within eight high-level phylogenetic groups. The SARS-CoV-2 sequences isolated in Turkey and deposited in GISAID have a diverse clade distribution

submitted to the Medical Virology Laboratory of Gazi University. Samples used in the study were initially obtained from suspected individuals corresponding to the case description for SARS-CoV-2 infection for routine diagnostic testing. The specimens were collected in vNAT (Bioeksen) transfer tubes and transported to the laboratory where they were stored at 4°C. Samples were processed using Bio-Speedy SARS-CoV-2 (2019-nCoV) qPCR Detection Kit v5.3 (Bioeksen) and stored at −80°C until sequencing.

Nasopharyngeal swab samples were obtained from the patients with suspected COVID-19 infection and tested positive for SARS-CoV-2 at the Medical Virology Laboratory of Gazi University. We decided to choose the samples with high viral loads with C_t values lower than 20 since the viral nucleic acid extraction buffer in vNAT transfer tubes causes RNA degradation and nucleic acid loss and also a high viral load is required for a good performance in the next-generation sequencing (NGS). We only had vNAT transfer tubes with viral nucleic acid extraction buffer causing nucleic acid loss during the long-term storage of the samples. Therefore, we selected recent specimens handled in our laboratory that were assumed to contain high viral loads. A high viral load containing a good quality nucleic acid is required to obtain significant results in the NGS. We thought that we would have a greater opportunity of obtaining high-quality PCR products and get better NGS results when high viral load containing samples are used. Another reason for selecting the samples with high viral loads is that there is some evidence of the association between high viral loads and increased infectivity[13] and disease severity.[14] We did not aim to compare and analyze the mutations between different age groups, gender, or patients with different clinical manifestations. This study focused on determining whether or not a mutation is occurring in the samples of the patients applied to our hospital. Considering all of these, we anticipated we could detect mutations in the patient samples with high viral loads.

2.2 | SARS-CoV-2 RT-PCR; SARS-CoV-2 nucleic acid isolation and amplification

Viral RNA was extracted using EZ1 Virus Mini Kit v2.0 on the automated EZ1 Advanced (Qiagen). A sample volume of 400 µl was used and the viral RNA was eluted in 60 µl. For the quality control of the extraction process, 2 µl of Genesig Easy RNA Internal Extraction Control was added to each sample. Following the extraction, real-time RT-PCR targeting the RdRp region of the SARS-CoV-2 genome was performed using Primer Design Genesig COVID-19 (Genesig). Rotor-Gene Q (Qiagen) was used in fluorescence channel Cycling Green for the detection of amplicons. Rotor-Gene Software was used to evaluate the amplification curves.

2.3 | Next-generation sequencing

NGS libraries were constructed using CleanPlex SARS-CoV-2 Panel and Dual-Indexed PCR Primers for Illumina Set A (Paragon Genomics). The library construction workflow consists of four steps. PCR cleaning and nucleic acid separation were applied at the end of each step using Sera-Mag Select (Cytiva) magnetic purification beads. A final elution volume of 10 µl was obtained at the end of each step.

2.4 | RT-PCR reaction

Purified viral RNA was converted into cDNA by reverse transcription reaction. The reaction was performed using 5 µl of viral nucleic acid extract. The extraction material, 6 µl of nuclease-free water and the 3 µl of RT primer mix were combined. The reaction mixture was incubated for 5 min at 65°C on a thermal cycler. Following the primer annealing, 1 µl of RT enzyme and 5 µl RT reaction buffer were added into the reaction mixture and the final mixture was incubated for 10 min at 8°C then for 80 min at 42°C on a thermal cycler. Following the RT reaction, 2 µl of stop buffer was added to each tube. The RT-PCR materials were cleaned using the method of 2.2X bead-based purification.

2.5 | The multiplex PCR reaction

The entire viral genome was amplified with a multiplexed target enrichment strategy in two separated primer pools containing a total of 343 amplicon fragments. The primer pools span the entire genome of SARS-CoV-2 ranging from 116 to 192 bp. The multiplex PCR (mPCR) reaction was performed in 10 µl volume reactions, consisting of 2 µl of nuclease-free water, 2 µl of 5X mPCR mix, 5 µl of purified reaction material from the previous step, and 1 µl of 10X SARS-CoV-2 primer pools for each pool. A thermal cycler protocol was programmed for initial denaturation for 10 min at 94 C, followed by 10 cycles of denaturation for 15 s at 98°C, annealing/extension for 5 min at 60°C. Following the PCR amplification, 2 µl of stop buffer was added to each tube. The equal volumes of the postamplification reaction materials for each sample were pooled into a microcentrifuge tube. The mPCR materials were cleaned using the method of 1.3X bead-based purification.

2.6 | The digestion reaction

The third step was the digestion reaction of the amplified DNA fragments, which performs background cleaning by eliminating nonspecific PCR products. The digestion reaction mixture was performed in 20 µl volumes, consisting of 7 µl of nuclease-free water, 2 µl of reagent buffer, 1 µl of digestion reagent, and 10 µl of purified sample from the previous step. The final reaction mixtures were incubated for 10 min at 37°C on a thermal cycler. Following the incubation, 2 µl of stop buffer was added to each tube. The postdigestion reaction materials were cleaned using the method of 1.3X bead-based purification.

2.7 | Indexing

The index sequences were added to each amplified DNA fragment using the combination of i5 and i7 primers designed for the Illumina sequencing platforms. The indexing reaction mixture was performed in 40 µl, consisting of 18 µl of nuclease-free water, 8 µl of 5X PCR Mix, 10 µl of purified sample from the previous step, 2 µl of each unique indexed PCR primers referred to i5 and i7. A thermal cycler protocol was programmed for initial denaturation for 10 min at 95°C, followed by 10 cycles of denaturation for 15 s at 98°C, annealing/extension for 75 s at 60°C. PCR amplicons were cleaned using the method of 1X bead-based purification.

2.8 | Quantification and normalization

The quantity of indexed libraries was measured in ng/µl unit on Qubit 3.0 Fluorometer (Thermo Fisher Scientific). The quantitation was performed using iQuant Broad Range dsDNA Quantitation Kit (ABP Biosciences). The average size of the libraries was determined by the method of gel electrophoresis visualization. The presence of a peak at about 250 bp was observed. Using the instructions and the formula recommended by Illumina, concentrations in ng/µl were converted to nM values. Each library was normalized and diluted to a final loading concentration of 100 pM. The equal volumes of each normalized

library combined with unique indexes were pooled into a microcentrifuge tube.

2.9 | Sequencing

The sequencing run was created with Local Run Manager software installed on the iSeq. 100 instrument. Using the Generate FASTQ Analysis module with the custom library preparation kit options, the number of sequencing cycles was set to 151 bp and the paired-end sequencing option was determined. The pooled libraries were loaded onto iSeq 100 system using iSeq. 100 i1 Reagent v1 (Illumina).

2.10 | Data analysis

The removal of adapter sequences from demultiplexed FASTQ files was performed with the Cutadapt v3.0 tool.[15] Further analysis of the trimmed raw sequencing reads was done using Genome Detective Virus Tool. Genome Detective Virus Tool is a web-based software that evaluates the read quality, removes low-quality reads, performs read mapping, assembles the reads, and generates consensus sequences for NGS data.[16] Nextclade tool was used to check for sequencing errors and the quality of the assembled sequences. Nextclade tool utilizes different metrics to analyze the quality of consensus sequences. Assembled sequences containing a high volume of missing data, high divergence, and too many ambiguous nucleotides have been removed from further analysis.[17] Variant calling and the generation of the variant database from the assembled sequences were performed using the Malvirus tool.[18] The annotation and the filtering of the variants was done using Snpeff[19] and SnpSift[20] tools against a database created from the reference sequence (NC_045512.2). Further analysis and the manual control of nucleic acid and amino

TABLE 2 Number of mutations observed in coding regions

		Substitution			
Genome segment	Missense	Synonymous		Total of substitutions	Deletion
ORF1ab	*nsp1*	3	5	8	–
	nsp2	16	9	25	–
	nsp3	17	18	35	1
	nsp4	5	5	10	–
	nsp5/3CLpro	5	6	11	–
	nsp6	3	4	7	–
	nsp7	2	2	4	–
	nsp8	2	5	7	–
	nsp9	1	2	3	–
	nsp10	–	4	4	–
	nsp12/RdRp	10	8	18	–
	nsp13/Hel	11	3	14	–
	nsp14/EXoN	9	5	14	–
	nsp15/endoRNAse	5	7	12	–
	nsp16/2′-O-MTase	4	3	7	–
Total of ORF1ab mutations		93	86	179	1
ORF2/S		25	15	40	4
ORF3a		11	5	16	–
M		1	6	7	–
ORF6		–	1	1	–
ORF7		2	2	4	–
ORF8		3	2	5	–
N		15	6	21	–
ORF10		2	1	3	–
Total of mutations in coding regions		152	124	276	5
Number of mutations observed in noncoding regions					
Genome segment		Substitution			Deletion
Noncoding regions	*5′-UTR*	3			–
	3′-UTR	7			–
	Intergenic	2			–
Total of mutations in noncoding regions		11			–

Abbreviations: 2′-O-MTase, 2′-O-methyltransferase; 3CLpro, 3-chymotrypsin-like cysteine protease; endoRNAse, endoribonuclease; EXon, exonuclease; Hel, helicase; M, matrix; N, nucleocapsid; nsp, nonstructural protein; ORF, open reading frame; RdRp, RNA-dependent RNA polymerase; S, spike; UTR, untranslated region.

acid changes were performed using web-based CoV-Glue[21] and CoVsurver[22] tools. All sequencing data used in this publication are available at NCBI's Sequence Read Archive (Bioproject accession ID: PRJNA687366) and GISAID's EpiCoV Database (Supporting Information Material).

2.11 | Phylogenetic tree construction

Consensus sequences were aligned with the reference sequence using Mafft. Following the multiple sequence alignment, noncoding regions including the 5′ and 3′ ends of the sequences were trimmed without losing key nucleotide positions. PhyloSuite[23] was used to conduct, manage and streamline the analyses with the help of several plug-in programs. ModelFinder[24] was used to select the best-fit model using the BIC criterion. Bayesian inference phylogenies were inferred using MrBayes[25] 3.2.6 under the GTR + I + F model (2 parallel runs, 5000002 generations), in which the initial 25% of sampled data were discarded as burn-in. Pango lineage,[11] Nextstrain[12] clade, and GISAID[26] clade assignments were carried out using the online tools of the respective nomenclature systems.

2.12 | Ethical statement

Official permission to conduct the study is obtained from the Ministry of Health and the study is approved by the Ethics Board of Gazi University, Faculty of Medicine (Decision number: 341 Date: May 22, 2020). All research participants gave their informed consent in writing before inclusion in the study.

3 | RESULTS

Sequences with suspected underlying sequencing errors (10/45 sequences) according to Nextclade quality metrics were excluded from the analysis. In total, 35 SARS-CoV-2 genome sequences were analyzed.

3.1 | Mutational dimension

In total, 595 mutations with 292 distinct mutations were found. Of 293 distinct mutations, 240 were observed once in the sequences (237 substitution mutations and 3 deletions). The 293 distinct mutations consist of 152 missense mutations, 124 synonymous mutations, 12 mutations in the noncoding regions, and 5 deletions (Table 2).

The majority of the mutations (179 mutations) were located in the ORF1ab gene region, followed by the ORF2/S (40 mutations) and N (21 mutations) gene regions. Out of the 152 missense mutations, 93 mutations are found in ORF1ab, the longest ORF occupying two-thirds of the entire genome. ORF1ab is expressed into a polyprotein and 16 nonstructural proteins (nsps) are subsequently cleaved from the polyprotein. Of these proteins, nsp3 has the largest number of missense mutations among ORF1ab proteins (17 mutations), followed by nsp2 (16 mutations). Of the nsp3 missense mutations, T133I was the most common (17 sequences) followed by H342Y (2 sequences) and P1326L (2 sequences). Missense mutations have also been detected in nonstructural protein RNA-dependent polymerase (RdRp), such as P323L (35 sequences), Q822H (5 sequences).

Mutations with a recurrence of more than 11 sequences are shown in Table 3. The most common mutations were the missense mutations P323L in nsp12 and D614G in ORF2/S gene regions, 241C>T in the 5′-UTR noncoding region and were always found simultaneously in all 35 sequences. These mutations were followed by the synonymous mutation 3037C>T in the nsp3 gene region (34 sequences). In addition to mutations found in all sequences, several less frequent mutations and some novel mutations were identified.

Novel mutations are detected in nsp12 and S gene regions. While the H133R (ORF1ab, nsp12) and M626K (ORF1ab, nsp12) amino acid mutations were unique to sample 12, V111A (ORF1ab, nsp12) and Y453R (ORF1ab, nsp12) were found in Samples 1 and 39, respectively. The novel ORF2/S gene mutations, V1068L and R995G, were found in Samples 34 and 39, respectively (Table 4). These mutations could not be found in any other sequences deposited in the GISAID database. Mutations found in these gene regions may have important functional implications that need to be discussed in the context of vaccine development and therapeutic options.

In addition to substitution mutations, 5 deletion sites were found in ORF1ab/nsp3 (1 deletion, nucleotide positions 6653–6671) and ORF2/S (4 deletions, nucleotide positions 21765–21770 [ΔH69/ΔV70], 22665–22705, 21991–21993 [ΔY144], 21990–21992) gene regions. Deletions at nucleotide positions 21765–21770 (ΔH69/ΔV70; Sample

TABLE 3 Mutations with a recurrence of more than 11 sequences

Genome region	Protein	Amino acid position and change	Nucleotide position	Reference nucleotide	Mutated nucleotide	Type of mutation	Number of samples
5′-UTR	–	–	241	C	T		35
ORF1b	nsp12	P323L	14408	C	T	Missense	35
ORF2	S	D614G	23403	A	G	Missense	35
ORF1a	nsp3	–	3037	C	T	Synonymous	34
ORF9	N	G204L	28883	G	C	Missense	21
ORF9	N	R203K	28881	G	A	Missense	20
ORF9	N	–	28882	G	A	Synonymous	20
ORF1b	nsp15	–	19839	T	C	Synonymous	12

Abbreviations: nsp, nonstructural protein; ORF, open reading frame; S, spike; UTR, untranslated region.

23), 21990–21992; Sample 38) and 21991–21993 (ΔY144; Sample 10, 11, 27, 34, 40) of spike protein were already reported. Deletions at nucleotide positions 22665–22705 (Sample 6, 12) of spike protein and 6653–6671 (Sample 11) of ORF1ab/nsp3 are novel deletions, which have not been reported in the GISAID database.

3.2 | GISAID and Nextstrain clade distribution

Three nomenclature schemes are illustrated on the tree (Figure 2). Clustering the sequences according to the GISAID scheme shows three major GISAID clades. Out of 35 sequences, 9 sequences in G-clade, 19 sequences in GR-clade, and 6 sequences in GH-clade were classified. These sequences were grouped into two main Nextstrain clades. The 20B clade is found to be the most prevalent (21 sequences).

TABLE 4 Novel mutations detected in the sequences

Nucleotide position and change	Amino acid position and change (sample id)	Genome region	Protein
13772T>C	V111A (1)	ORF1ab	nsp12
13838A>G	H133R (12)	ORF1ab	nsp12
14798A>G	Y453C (39)	ORF1ab	nsp12
15317T>A	M626K (12)	ORF1ab	nsp12
24545A>G	R995G (39)	ORF2	S
24763G>C	V1068L (34)	ORF2	S

Abbreviations: nsp, nonstructural protein; ORF, open reading frame; S, spike.

3.3 | Pango lineage distribution

A total of nine Pango lineages were observed, five of which were observed in more than one sample. The most frequently detected lineage group was B.1.1 (17 sequences), followed by B.1 (7 sequences) and B.1.36.1 (36 sequences). Lineages assigned to the sequences and the most common countries they have been observed are shown in Table 5.

4 | DISCUSSION

Since the beginning of the pandemic, the SARS-CoV-2 genome has been rapidly evolving and there has been evidence that the occurring mutations in the genome have an impact on the virulence of the virus. Thanks to the widespread availability of NGS tools and the online sharing of genome information, mutations, and variants are routinely tracked and a large number of full-length genome

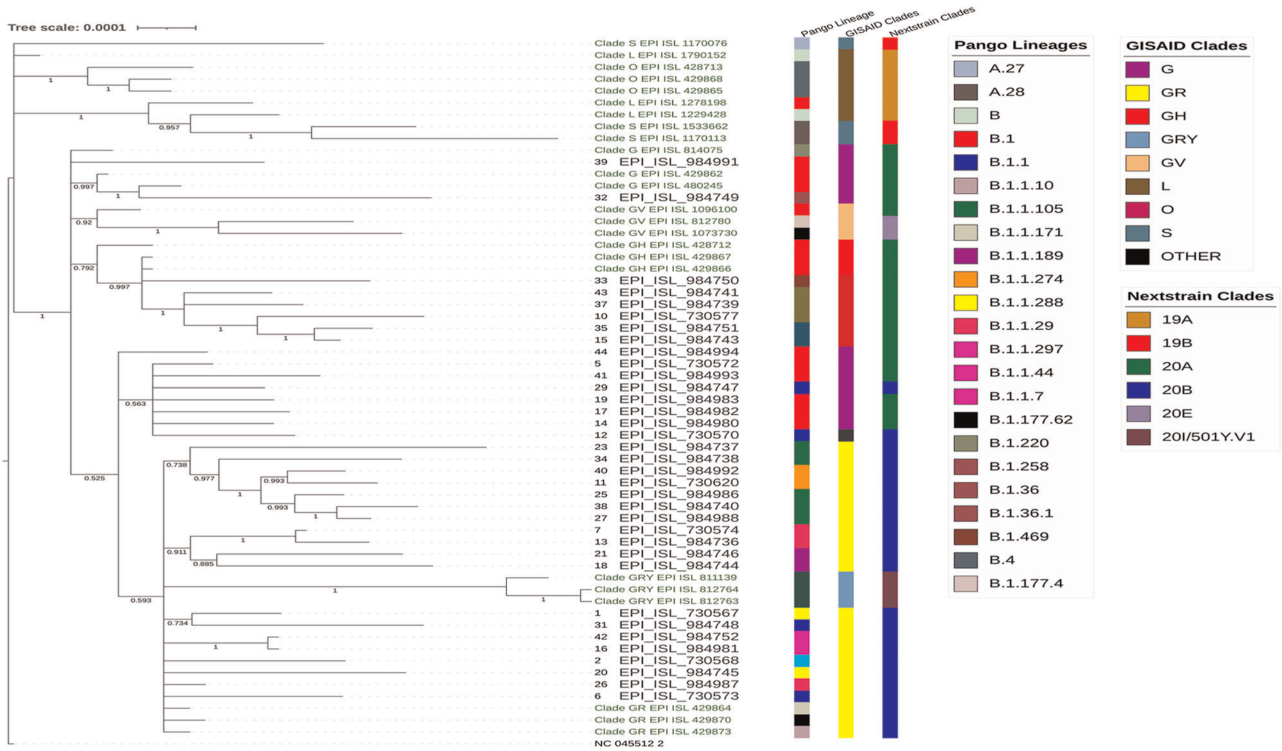

FIGURE 2 Bayesian phylogenetic inference of 35 SARS-CoV-2 genomes and clade representative sequences from Turkey (Their clades and Global Initiative on Sharing All Influenza Data [GISAID] accession number are shown in green color). The numbers along the branches mark the posterior probability values. The reference genome NC_045512.2 was used to outgroup root the tree. Three nomenclature methods were used to assign sequences to the clusters. The clusters are represented as colored strips near the GISAID accession numbers of the sequences and are shown in different colors on the tree

TABLE 5 Distribution of SARS-CoV-2 lineages among our sequences

Pango lineage	Sequence name	Most common countries
B.1	5, 14, 17, 19, 41, 39, 44	USA, UK, Spain
B.1.1	2, 6, 7, 11, 12, 13, 16, 23, 25, 26, 27, 29, 31, 34, 38, 42, 40	UK, USA, Russia
B.1.1.189	18, 21	UK, Denmark, Turkey
B.1.1.288	1	Denmark, UK, Turkey
B.1.1.521	20	Germany, Switzerland, Denmark
B.1.258	32	UK, Denmark, Switzerland
B.1.36	37, 43	Canada, UK, India
B.1.36.1	10, 15, 35	Switzerland, UK, Denmark
B.1.469	33	UK, Denmark, Turkey

information from various countries affected by the pandemic are available.

Throughout the pandemic, there has been an increased frequency of mutations in the genome, with over 3000 unique point mutations were discovered in viral isolates from around the world.[27] Similar to global findings, SARS-CoV-2 isolates from Turkey were proposed to have an elevated mutation rate in a study conducted on 166 genomes.[28] Previous studies discovered that the S and RdRp gene regions in the genome have the highest mutation rates.[27,29,30] Similar to previous studies conducted globally and in Turkey, the majority of the mutations in our sequences are missense mutations, frequently located in gene regions related to the expression of enzymes and cofactors, involving in the replication of the SARS-CoV-2 genome.

Even though the protein structure alterations and the functional analysis of the mutations are not included in our study, we discussed the possible effects of the mutations frequently found in the sequences, depending on their locations in the genome. Based on the mutation analysis, all sequences included D614G mutation in the spike glycoprotein (23403A>G) and P323L mutation (14408C>T) in the nsp12/RdRp. Previous research has also found that the co-occurrence of these mutations is high in sequences obtained from Turkey.[8,28,31] The D614G mutation gradually became dominant throughout the world during the pandemic. The D614G mutation is found to be associated with higher viral loads in patients.[32] The spike protein is responsible for the attachment and entrance of the virus to the host cell. S protein mutations and their effects on virulence should be closely monitored and evaluated, as this protein is the main target in the current vaccine development studies.[33]

3037C>T mutation in the nsp3 gene region has appeared in 34 of 35 analyzed sequences, which corresponds to 97.2% of all sequences and is almost always accompanied by D614G and P323L mutations. A similar scenario was observed in another study which was also carried out in Turkey.[9] In a recent study comparing mutation profiles according to disease severity, D614G and P323L mutations in SARS-CoV-2 are found to be correlated to severe COVID-19 cases.[34]

One of the most globally common 5′-UTR mutations in the SARS-CoV-2 genome is the 241C>T and it was also present in all of our sequences. Studies in some viral genomes have reported that variations in UTRs may affect the activity of viral RNA folding and packaging.[35] Although 241C>T mutation has an uncertain significance, the results of a study indicated that the mutation causes a decrease in the replication rates which results in a reduction of the mortality rates.[36] The mutation 241C>T has been reported to be co-occurred with three other mutations (3037C>T, nsp3; 14408C>T/P323L, RdRp), and 23403A>G/D614G, S).[37] These co-occurring mutations have been identified as one of the major clades of SARS-CoV-2.[30]

We found novel mutations in the nsp12 gene region, which are not detected in any other sequences deposited in the GISAID database. Nsp12 gene region encodes the viral protein RdRp and it is responsible for the replication of viral RNA. This protein is shown to be bound and blocked by the antiviral drug remdesivir which is predicted to be effective against SARS-CoV-2 by in vitro analysis.[38] The mutations occurring in RdRp have also been demonstrated in the influenza A virus to confer resistance against favipiravir.[39] Mutations especially in critical residues for drug efficacy can lead to loss of binding affinity of drugs and evolving of treatment resistance strains. P214L was one of the most common mutations in the RdRp gene region in other studies.[40] However, this mutation has not been seen in our sequences. Analyses of RdRp mutations have been proven to cause altering in the mutation rates of other genes[41] and this could lead to the increasing number of mutations in the whole genome of the virus.[42]

We report two novel mutations in the spike protein sequences of two isolates (24545A>G; R995G, Sample 39 and 24763G>C; V1068L, Sample 34). The two mutations are located in the S2 domain of the protein. As the S2 domain is characterized as a viral fusion peptide, mutations in the S2 domain can contribute to the stabilization of membrane fusion and increase infectivity through enhanced fusion activity.[43]

Other common mutations included three nucleotide changes at positions 28881, 28882, and 28883 (GGG to AAC), which were found in 20 sequences (57.1%). These changes affect two consecutive codons and result in two amino acid changes (R203K, G204L) in the nucleocapsid (N) protein. Nucleocapsid protein has a critical role in the transcription of viral RNA and replication of the virus.[44] The N protein may have an effect on the immune response[45] and would be considered for vaccine development. This block mutation is expected to affect the pathogenicity of the SARS-CoV-2.

Several deletion mutations have been found in our sequences, which may be sequencing errors. Further confirmation of these deletion mutations is required by molecular and sequencing analysis. The ΔH69/V70 (21765–21770) and ΔY144 (21991–21993) deletions in combination with several other mutations have recently been described as a newly emerging variant from the United Kingdom (Lineage B.1.1.7/VOC‐202012/01, Variant of Concern 202012/01). The ΔH69/V70 and ΔY144 deletions in our sequences have been detected independently and not in combination with other mutations of the variant. Although the impact of deletion ΔH69/ΔV70 is not clear yet, it has been associated with infections in immunocompromised patients.[46,47]

Based on the phylogenetic tree, the clustering of the sequences matches mainly the previously established GISAID clades. Currently, the G clade and its subclades, GH, GR, GV, and GRY are the most prevalent clades globally. The GRY (B.1.1.7) clade is currently the most prominent representative of the viral population in the world. Generally, the GH clade is prevalently present in North America, whereas GR has been mostly observed in Europe. Currently, the most prevalent clade in Turkey is GR, accounting for more than 35% of sequences submitted in the GISAID database.[26]

Pango lineages are identifiers of spreading lineages defined by a phylogenetic framework and often represent distinct introductions of viral variants into new territories or regions. Our findings identified nine Pango lineages of SARS‐CoV‐2 among our sequences, suggesting an introduction of different sources in connection with our samples and their subsequent spread.[11] The most prevalent lineage assigned to the sequences was B.1.1, which has been reported in over 130 countries. Although most of the identified lineages originated from Europe and North America, some of them originated from different geographical locations (Russia, India). B.1.1.189 lineage was represented in 5.7% of our sequences (5 sequences). Overall, a total of 175 sequences from Turkey (34%) have been observed worldwide in B.1.1.189 lineage and the lineage was designated as a European lineage. Global lineages were further subdivided into sublineages to identify ongoing transmission. Lineages are subject to change and must be recalculated for all genomes with each reanalysis.[48]

Mutations are known to have an effect on virulence which creates variants of SARS‐CoV‐2 that stand out for their increased infectious properties. Mutations can increase the replication capabilities and cause the rapid spread of the virus. Such a variant may become dominant in the population in a short period of time. Therefore, monitoring of mutations and their roles in virulence‐related conditions, such as affinity to host cell receptors, transmissibility, high viral loads, and lethal effects, should be thoroughly investigated. The genomic characterization of the sequences is also important to monitor the movement of the virus between individuals and across geographical areas. Existing mutations that have been detected allow the genome sequences to be classified into separate groups and different variants can be further characterized by the most recent detected mutations and detection of the most recent mutations enable new variants to be identified. Genomic epidemiological studies involving the sequencing of the viral genome may help to understand the pathways of the transmission and interpreting the potential prevention strategies.

ACKNOWLEDGMENT

This study was supported by research grant from Scientific Research Projects Unit of Gazi University (Grant ID 01/2020‐20).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

Erdem Sahin carried out the sequencing, analyzed sequencing data, wrote the manuscript with support from Gulendam Bozdayi, Kayhan Caglar, and Isıl Fidan. Selin Yigit contributed to the sample preparation and performed the RT‐qPCR tests. Gulendam Bozdayi is head of the project and assisted the acquisition of the financial support for the project (Gazi University Scientific Research Project [01/2020‐20]) leading to this study and publication. Hager Muftah helped to run sequencing. Kayhan Caglar and Isil Fidan provided stylistic and grammatical revisions to manuscript. Murat Dizbay and Ozlem G. Tunccan performed the clinical specimen collection. All authors provided critical feedback and helped shape the research, analysis and manuscript.

ORCID

Erdem Sahin https://orcid.org/0000-0002-1389-3253
Gulendam Bozdayi https://orcid.org/0000-0002-6036-6819

REFERENCES

1. Yin Y, Wunderink RG. MERS, SARS and other coronaviruses as causes of pneumonia. *Respirology*. 2018;23(2):130‐137. https://doi.org/10.1111/resp.13196
2. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265‐269. https://doi.org/10.1038/s41586-020-2008-3
3. Coronavirus ‐ Worldometer. https://www.worldometers.info/coronavirus/. Accessed June 17, 2021.
4. Covid19. https://covid19.saglik.gov.tr/. Accessed June 17, 2021.
5. Demirbilek Y, Pehlivantürk G, Özgüler ZÖ, Alp Meşe E. COVID‐19 outbreak control, example of ministry of health of Turkey. *Turk J Med Sci*. 2020;50:489‐494. https://doi.org/10.3906/sag-2004-187
6. Toyoshima Y, Nemoto K, Matsumoto S, Nakamura Y, Kiyotani K. SARS‐CoV‐2 genomic variations associated with mortality rate of COVID‐19. *J Hum Genet*. 2020;65(12):1075‐1082. https://doi.org/10.1038/s10038-020-0808-9
7. Karacan I, Akgun TK, Agaoglu NB, et al. The origin of SARS‐CoV‐2 in Istanbul: Sequencing findings from the epicenter of the pandemic in Turkey. *North Clin Istanb*. 2020;7(3):203‐209. https://doi.org/10.14744/nci.2020.90532
8. Adebali O, Bircan A, Çirci D, et al. Phylogenetic analysis of SARS‐CoV‐2 genomes in Turkey. *Turk J Biol*. 2020;44(3):146‐156. https://doi.org/10.3906/biy-2005-35

9. Demir A, Benvenuto D, Abacioglu H, Angeletti S, Ciccozzi M. Identification of the nucleotide substitutions in 62 SARS-CoV-2 sequences from Turkey. *Turk J Biol*. 2020;44(3):178-184.
10. Ergünay K, Kaya M, Serdar M, Akyön Y, Yılmaz E. A cross-sectional overview of SARS-CoV-2 genome variations in Turkey. *Research Square*. 2021. https://doi.org/10.21203/rs.3.rs-472330/v1
11. 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. https://doi.org/10.1038/s41564-020-0770-5
12. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018;34(23):4121-4123. https://doi.org/10.1093/bioinformatics/bty407
13. McEllistrem MC, Clancy CJ, Buehrle DJ, et al. SARS-CoV-2 is associated with high viral loads in asymptomatic and recently symptomatic healthcare workers. *PLoS One*. 2021;16(3):e0248347. https://doi.org/10.1371/journal.pone.0248347
14. Fajnzylber J, Regan J, Coxen K, et al. SARS-CoV-2 viral load is associated with increased disease severity and mortality. *Nat Commun*. 2020;11(1):5493. https://doi.org/10.1038/s41467-020-19057-5
15. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 2011;17:10-12. https://doi.org/10.14806/ej.17.1.200
16. Cleemput S, Dumon W, Fonseca V, et al. Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics*. 2020;36(11):3552-3555. https://doi.org/10.1093/bioinformatics/btaa145
17. Nextclade. https://clades.nextstrain.org/. Accessed April 21, 2021.
18. Ciccolella S, Denti L, Bonizzoni P, et al. MALVIRUS: an integrated web application for viral variant calling. *bioRxiv*. 2020. https://doi.org/10.1101/2020.05.05.076992
19. Cingolani P, Platts A, Wang le L, 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. https://doi.org/10.4161/fly.19695
20. Cingolani P, Patel VM, Coon M, et al. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front Genet*. 2012;3:35. https://doi.org/10.3389/fgene.2012.00035
21. Singer J, Gifford R, Cotten M, Robertson D. CoV-GLUE: a web application for tracking SARS-CoV-2 genomic variation. *Preprints*. 2020. https://doi.org/10.20944/preprints202006.0225.v1
22. GISAID - CoVsurver mutations App. https://www.gisaid.org/epiflu-applications/covsurver-mutations-app/. Accessed April 21, 2021.
23. Zhang D, Gao F, Jakovlić I, et al. PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol Ecol Resour*. 2020;20(1):348-355. https://doi.org/10.1111/1755-0998.13096
24. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017;14(6):587-589. https://doi.org/10.1038/nmeth.4285
25. Ronquist F, Teslenko M, van der Mark P, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 2012;61(3):539-542. https://doi.org/10.1093/sysbio/sys029
26. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Challenges*. 2017;1(1):33-46. https://doi.org/10.1002/gch2.1018
27. Flores-Alanis A, Cruz-Rangel A, Rodríguez-Gómez F, et al. Molecular epidemiology surveillance of SARS-CoV-2: mutations and genetic diversity one year after emerging. *Pathogens*. 2021;10(2):184. https://doi.org/10.3390/pathogens10020184
28. Eskier D, Akalp E, Dalan Ö, Karakülah G, Oktay Y. Current mutatome of SARS-CoV-2 in Turkey reveals mutations of interest. *Turk J Biol*. 2021;45(1):104-113.
29. Koçhan N, Eskier D, Suner A, Karakülah G, Oktay Y. Different selection dynamics of S and RdRp between SARS-CoV-2 genomes with and without the dominant mutations. *Infect Genet Evol*. 2021;91:104796. https://doi.org/10.1016/j.meegid.2021.104796
30. Mercatelli D, Giorgi FM. Geographic and genomic distribution of SARS-CoV-2 mutations. *Front Microbiol*. 2020;11:1800. https://doi.org/10.3389/fmicb.2020.01800
31. Hanifehnezhad A, Kehribar EŞ, Öztop S, et al. Characterization of local SARS-CoV-2 isolates and pathogenicity in IFNAR$^{-/-}$ mice. *Heliyon*. 2020;6(9):05116. https://doi.org/10.1016/j.heliyon.2020.e05116
32. 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 e19. https://doi.org/10.1016/j.cell.2020.06.043
33. Singh PK, Kulsum U, Rufai SB, Mudliar SR, Singh S. Mutations in SARS-CoV-2 Leading to antigenic variations in spike protein: a challenge in vaccine development. *J Lab Physicians*. 2020;12(02):154-160. https://doi.org/10.1055/s-0040-1715790
34. Biswas SK, Mudi SR. Spike protein D614G and RdRp P323L: the SARS-CoV-2 mutations associated with severity of COVID-19. *Genomics Inform*. 2020;18(4):e44. https://doi.org/10.5808/gi.2020.18.4.e44
35. Thompson SR. Tricks an IRES uses to enslave ribosomes. *Trends Microbiol*. 2012;20(11):558-566. https://doi.org/10.1016/j.tim.2012.08.002
36. Chaudhari A, Chaudhari M, Mahera S, et al. In-silico analysis reveals lower transcription efficiency of C241T variant of SARS-CoV-2 with host replication factors MADP1 and HNRNP-1. *bioRxiv*. 2020. https://doi.org/10.1101/2020.11.22.393009
37. Yin C. Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics*. 2020;112(5):3588-3596. https://doi.org/10.1016/j.ygeno.2020.04.016
38. Pruijssers AJ, George AS, Schäfer A, et al. Remdesivir inhibits SARS-CoV-2 in human lung cells and chimeric SARS-CoV expressing the SARS-CoV-2 RNA polymerase in mice. *Cell Rep*. 2020;32(3):107940. https://doi.org/10.1016/j.celrep.2020.107940
39. Goldhill DH, Te Velthuis AJW, Fletcher RA, et al. The mechanism of resistance to favipiravir in influenza. *Proc Natl Acad Sci U S A*. 2018;115(45):11613-11618. https://doi.org/10.1073/pnas.1811345115
40. Kim JS, Jang JH, Kim JM, Chung YS, Yoo CK, Han MG. Genome-wide identification and characterization of point mutations in the SARS-CoV-2 genome. *Osong Public Health Res Perspect*. 2020;11(3):101-111. https://doi.org/10.24171/j.phrp.2020.11.3.05
41. Eskier D, Karakülah G, Suner A, Oktay Y. RdRp mutations are associated with SARS-CoV-2 genome evolution. *PeerJ*. 2020;8:8. https://doi.org/10.7717/peerj.9587
42. Pachetti M, Marini B, Benedetti F, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J Transl Med*. 2020;18(1):179. https://doi.org/10.1186/s12967-020-02344-6
43. Xia X. Domains and functions of spike protein in SARS-Cov-2 in the context of vaccine design. *Viruses*. 2021;13(1):109. https://doi.org/10.3390/v13010109
44. McBride R, Van Zyl M, Fielding BC. The coronavirus nucleocapsid is a multifunctional protein. *Viruses*. 2014;6(8):2991-3018. https://doi.org/10.3390/v6082991
45. Shah VK, Firmal P, Alam A, Ganguly D, Chattopadhyay S. Overview of immune response during SARS-CoV-2 infection: lessons from the past.

Front Immunol. 2020;11:1949. https://doi.org/10.3389/fimmu.2020.01949

46. Kemp SA, Meng B, Ferriera I, et al. Recurrent emergence and transmission of a SARS-CoV-2 spike deletion H69/V70. *bioRxiv*. 2021. https://doi.org/10.1101/2020.12.14.422555
47. Kemp SA, Collier DA, Datir R, et al. Neutralising antibodies in spike mediated SARS-CoV-2 adaptation. *medRxiv*. 2020. https://doi.org/10.1101/2020.12.05.20241927
48. Page AJ, Mather AE, Le-Viet T, et al. Large scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management. *medRxiv*. 2020. https://doi.org/10.1101/2020.09.28.20201475

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Sahin E, Bozdayi G, Yigit S, et al. Genomic characterization of SARS-CoV-2 isolates from patients in Turkey reveals the presence of novel mutations in spike and nsp12 proteins. *J Med Virol*. 2021;93:6016-6026. https://doi.org/10.1002/jmv.27188