Future
MICROBIOLOGY

Research Article

For reprint orders, please contact: reprints@futuremedicine.com

Molecular characterization, phylogenetic and variation analyses of SARS-CoV-2 strains in Turkey

Murat Karamese*,[1], Didem Ozgur[1] & Emin E Tutuncu[2]

[1]Department of Medical Microbiology, Kafkas University, Kars, 36100, Turkey
[2]Department of Infectious Diseases & Clinical Microbiology, Kafkas University, Kars, 36100, Turkey
*Author for correspondence: Tel.: +90 553 810 1112; murat_karamese@hotmail.com

Aims: We present the sequence and single-nucleotide polymorphism (SNP) analysis for 47 complete genomes for SARS-CoV-2 isolates on Turkish patients. **Methods:** The Illumina MiSeq platform was used for sequencing the libraries. The SNPs were detected by using Genome Analysis Toolkit – HaplotypeCaller v.3.8.0 and were inspected on GenomeBrowse v2.1.2. **Results:** All viral genome sequences of our isolates were located in lineage B under the different clusters, such as B.1 (n = 3), B.1.1 (n = 28) and B.1.9 (n = 16). According to the Global Initiative on Sharing All Influenza Data nomenclature, all of our complete genomes were placed in G, GR and GH clades. In our study, 549 total and 53 unique SNPs were detected. **Conclusion:** The results indicate that the SARS-CoV-2 sequences of our isolates have great similarity with all Turkish and European sequences.

First draft submitted: 4 May 2021; Accepted for publication: 1 September 2021; Published online: 7 October 2021

Keywords: complete genome sequencing • COVID-19 • phylogenetic analysis • SARS-CoV-2 • SNP analysis

A new virus-associated disease, COVID-19 by SARS-CoV-2, was first declared in late December 2019 in Wuhan, China [1]. As of this writing, there have been more than 173 million confirmed cases and more than 3.7 million deaths reported worldwide [2]. The etiological agent of this disease, SARS-CoV-2, has an ssRNA (single stranded) genome and its length is approximately 29,890 base pairs according to the data obtained from NCBI GenBank (GenBank NC_045512.2) [3]. After the confirmation of human-to-human transmission and extensive global spread, WHO declared COVID-19 as a pandemic on 11 March 2020 [4].

Understanding of the transmission patterns and evolution of SARS-CoV-2 virus is crucial and necessary for creating better drug and vaccine designs for disease control and prevention [4–6]. The analysis of genetic sequence data from a pathogen is known as an important tool in infectious disease epidemiology [7,8]. There are nearly 1.9 million pieces of SARS-CoV-2 virus data available on the database of Global Initiative on Sharing All Influenza Data (GISAID) [9]. The availability of these genomic data has been helping hundreds of researchers to analyze the genomic diversity of SARS-CoV-2 virus. First, Rambaut *et al.* [8] have defined the lineage A and B (both of them include sublineages such as A.1, B.1 and A.1.1) to help with tracking and understanding the patterns and determinants of the global spread of SARS-CoV-2. Then, on 4 July 2020, GISAID published the new clade and lineage nomenclature aids in genomic epidemiology studies of SARS-CoV-2 viruses. They developed a nomenclature system based on marker mutations in phylogenetic groups and named them S, L, V, G, GH and GR [10].

Next-generation sequencing (NGS) is a term that describes a DNA-sequencing technology. This technology is a reliable and powerful high-throughput tool and can be used to perform a sequencing of the whole genomes or constrained to specific areas of interest. The most important advantages of the NGS technique are capturing a broader spectrum of mutations than Sanger sequencing and being more sensitive to detection of mosaic mutations [11,12].

In this study, we performed NGS analysis for 47 complete genomes for SARS-CoV-2 isolates in Turkish patients. To identify their genetic similarity, phylogenetic analysis was performed by comparing the worldwide SARS-CoV-2

Future
Medicine

10.2217/fmb-2021-0118 © 2021 Future Medicine Ltd ISSN 1746-0913

sequences, selected from GISAID, with the complete genomes from Turkish isolates. In addition, we focused on the variation analysis to show the mutations on SARS-CoV-2 genomes.

The obtained data were one of the first data both in the country and all around the world that contained 47 full-genome sequences of SARS-CoV-2. Similar studies with more data should be performed to fight against COVID-19. Additionally, the scientific field needs new similar data until an effective vaccine or drug is developed.

Methods

This study was approved by both the Republic of Turkey Ministry of Health COVID-19 Scientific Research Evaluation Commission (approval date: 02/05/2020; number: 2020-05-02T16_13_50) and the Local Ethics Committee of Kafkas University Faculty of Medicine (approval date: 06/05/2020 number: 80576354-050-99/130).

Of the patients who tested positive for SARS-CoV-2, the 63 samples with the highest viral loads (assessed by real-time PCR by Biospeedy SARS-CoV-2 qPCR Kit targeting the RdRp gene) were selected for NGS-based assays. All samples were the first sample of patients (first positivity). After NGS analysis, we used 47 of those samples for complete genome sequencing and variation analyses.

The library preparation was performed using CleanPlex® SARS-CoV-2 Panel (Paragon Genomics Inc, IA, USA) with isolated RNA samples following the manufacturer instructions. Then, Illumina MiSeq (Illumina Inc., CA, USA) platform with a 2× 150-cycle kit was used for sequencing the libraries. The quality of the raw data was examined by FastQC v.0.11.5, and low-quality bases and primers were trimmed using Trimmomatic (version 0.32). The raw reads were aligned to the known SARS-CoV-2 genome (GenBank Accession: MN908947.3) using the Burrows-Wheeler aligner (version 0.7.1). The consensus sequences were aligned via multiple sequence alignment using the MAFFT v7.450 tool. The phylogenetic tree was constructed using Phylip v.3.6 with a neighbor-joining and composite likelihood method and 100 bootstrap replications. For phylogenetic analysis of our data, the GISAID database was used to collect SARS-CoV-2 complete genomes of different patients from all around the world (i.e., Belgium, Bosnian-Herzegovina, Czech Republic, Croatia, Cyprus, Finland, France, Germany, Hungary, Italy, Poland, Portugal, Spain, Russia, Australia, USA, Mexico, Costa Rica, Canada, Austria, Thailand, China, Kuwait, Senegal, Egypt, Japan, Chile, Nigeria, Algeria, Brazil, Ecuador, Mongolia, Colombia, India, Indonesia, Malaysia, Bangladesh, Kazakhstan and Turkey). Only complete genomes (28,000–30,000 base pairs) were analyzed. The SNPs were detected by using Genome Analysis Toolkit – HaplotypeCaller (GATK) v.3.8.0 and were inspected on GenomeBrowse v2.1.2 (GoldenHelix). The SNPs with low quality and a low variant fraction (%<60) were removed. The filtered SNPs and reference SARS-CoV-2 genome were used to generate the consensus sequence using bcftools v1.9.

Results

A database of 68 complete genomes of SARS-CoV-2 strains belonging to different countries randomly selected from the GISAID database were compared with our isolations. The reference SARS-CoV-2 strain, which was rooted from Wuhan (NCBI GenBank, NC_045512) was also selected for comparison. A total of 115 SARS-CoV-2 genomes were placed in the phylogenetic tree. The maximum likelihood phylogenetic tree in Figure 1 shows a main lineage including several sublineages. All viral genome sequences of our isolates were located in lineage B under the different clusters, such as B.1 (n = 3), B.1.1 (n = 28) and B.1.9 (n = 16). According to the GISAID nomenclature, all of our complete genomes were placed in G, GR and GH clades (Figure 1 & Supplementary Table 1).

In our study, 549 total and 53 unique SNPs were detected as shown in Supplementary Table 2. All 47 genomes exhibited different kinds of SNPs. The distinct SNPs consist of 274 missense, 225 synonymous and 50 noncoding alleles (Supplementary Table 2). The most common detected SNPs were c.1-25G>T (5′UTR), c.2772C>T (ORF1ab), c.14143C>T (ORF1ab), c.1841A>G (S) in all our genomes (n = 47), c.608G>A (N), c.609G>A (N), c.610G>C (N) in 28 samples, and c.2700G>T (S), c.310G>T (ORF7a) in 23 samples. Additionally, 31 SNPs were detected only once. Two hundred ninety-one (53%) SNPs were detected in ORF1ab, which is the longest ORF consisting of approximately 70% of the whole genome. The ORF1ab is cleaved into 16 nonstructural proteins (nsp). Among those, nsp3 and nsp2 had more SNPs in our study (n = 100 and n = 27, respectively). However, all noncoding mutations were detected in the 5′UTR region. All genomes had c.1-25C>T nucleotide variation, but only two genomes had additional noncoding mutations: c.1-56G>T and c.1-77G>T. In terms of base changing, the one most commonly detected (96%) was C>T. The detailed data about coding and noncoding mutations in our isolates is shown in Supplementary Table 2.

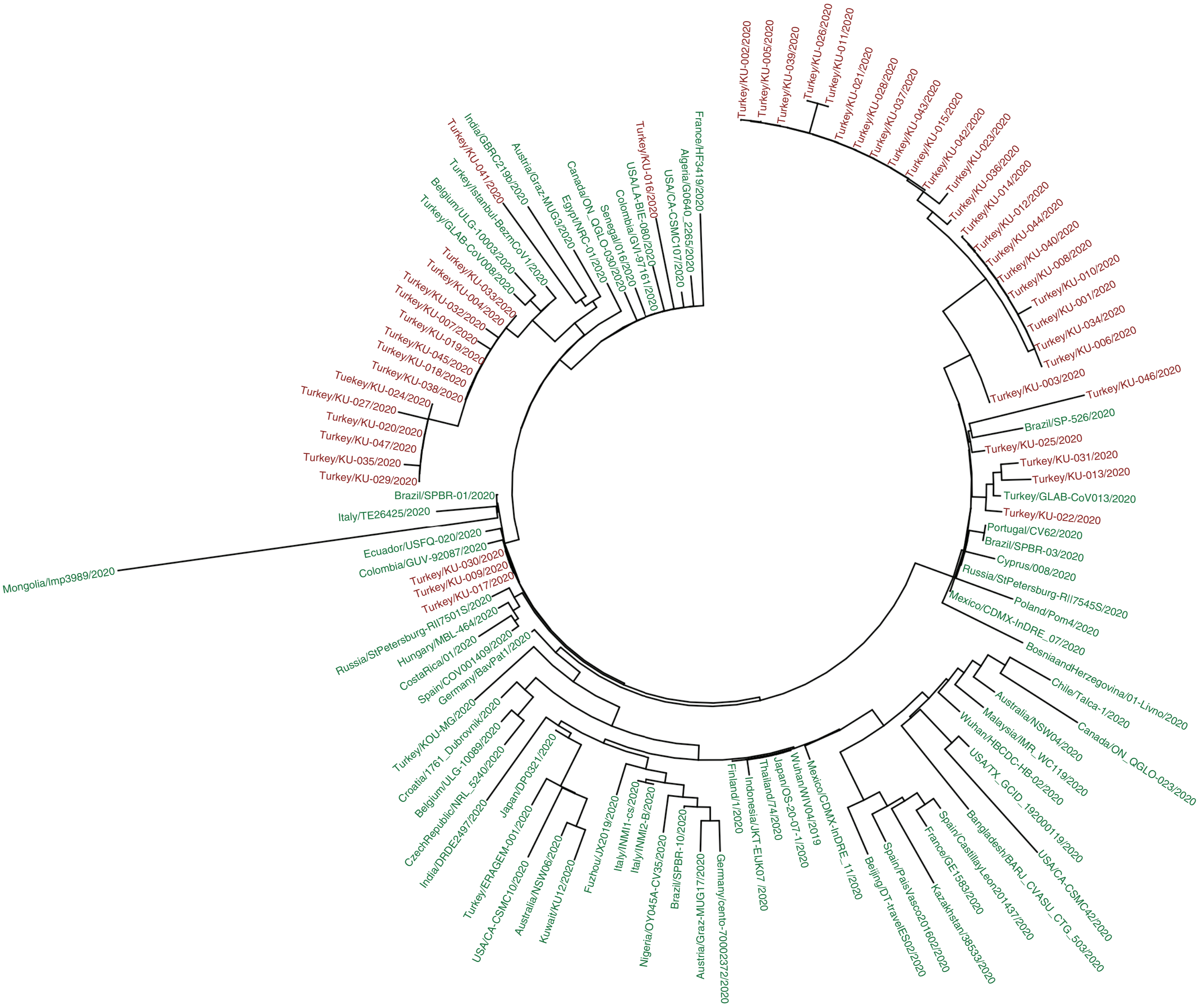

Figure 1. Subtree showing the SARS-CoV-2 strains containing also at least one strain in different countries. Our viruses are shown in red font color.

Discussion

The analysis of complete genome sequencing of new viruses is an important tool for epidemiology of infectious diseases, updating diagnostics and assessing viral evolution [7,8]. Complete genome data of a virus makes visible some epidemiological parameters including doubling time of an outbreak, reconstruction of transmission routes, and the identification of possible sources and animal reservoirs. It can also help the study of drug and vaccine design. In our study, we described 47 SARS-CoV-2 genomes and compared them with a dataset of 68 available SARS-CoV-2 genomes from different countries obtained from GISAID [9].

After the phylogenetic analysis, it was shown that all of our SARS-CoV-2 genomes are similar to the European strains. According to GISAID lineage, 47 of 47 of our SARS-CoV-2 genomes are in lineage B, which mostly consists of European strains. According to the marker mutations, those were placed in G, GR and GH clades. Moreover, most of European isolates were placed in G, GR and GH clades in GISAID [9].

In current literature, there are some studies that investigate the first isolates of their country to find the origin of those [3,13–17]. One study contains two complete genomes of a Chinese patient visiting Rome and an Italian patient, reporting that the Italian patient's sequence clustered with European sequences and the Chinese patient's sequence clustered with the Wuhan sequence (NC_045512) [18]. The same researchers compared the two genomes regarding concurrent evolution and accumulation of mutations and reported that four mutations were detected in the Italian sequence. Another study reported that because of the variations, the first isolates of India did not highly identify with the Wuhan sequence (NC_045512) [17]. Moreover, Bal *et al.* reported that the first three French sequences were located in the European clade instead of the reference Wuhan clade because of the three-nucleotide deletion in Orf1a at positions 1607–1609 [13]. However, some European isolates that were generally the first isolates of a related country belong to non-European clades. A German sequence was evaluated in a study performed by Zehender *et al.* [19] and reported that it belonged to lineage A because of traveling to Shanghai between 20 and 24 January.

So far, there is a total of 138 SARS-Cov-2 complete genome sequence data except for our 47 isolates in the GISAID database from Turkey [9]. It can be seen that all of those genomes belong to lineage B, including the first case (lineage B.4) in Turkey. All sequences (n = 47) belong to lineage B, similar to all other Turkish sequences. However, the first cases in Turkey belong to the L clade according to GISAID classification, whereas ours belong to G, GR and GH clades.

However, the length of the SARS-CoV-2 genome is approximately 30 Kb including 5′UTR and 3′UTR non-coding sequences, *Orf1ab*, *S*, *Orf3a*, *E*, *M*, *Orf6a*, *Orf7a*, *Orf7b*, *Orf8*, *N* and *Orf10* genes. In this study, we aim to detect the variations of our SARS-CoV-2 isolates.

According to the current literature, ORF genes have a crucial role in COVID-19 [20]. So, in our study, 549 total and 53 unique SNPs were detected from 47 SARS-CoV-2 isolates (Supplementary Table 2). A study performed by Khailany *et al.* [21] reported that 156 total and 116 unique SNPs were found in 95 SARS-CoV-2 isolates. The same study also pointed out that the most frequently observed base changes were C>T, the same as our study.

The most observed variations are found in positions 3036, 11083 and 13402 belonging to the *Orf1ab* gene; 28854 belonging to the *N* gene; 21707 and 21575 belonging to the *S* gene; and 28077 and 28144 belonging to the *Orf8* gene [4,17,22,23]. Our variations in related positions are the same as those. We mostly detected variations in positions 3037 and 11083 in the *Orf1ab* gene; 28166 in the *Orf8* gene; 28854 in the *N* gene and 24262 in the *S* gene. According to the global frequency of variations, most of our SNPs are novel; however, the studies on this issue are ongoing, and more detailed information will be presented in the near future. *Orf1ab* is the longest and most important gene among coronaviruses [20]. Most of the scientists detected variations in the *Orf1ab* gene. Thus, the mutations in this region may be significant concerning clinical features.

Conclusion

The results of the present study indicated that the SARS-CoV-2 sequences of our isolates have great similarity with all Turkish and European sequences. Further studies should be performed for better comparison of strains, after more complete genome sequences are released; however, these data may be useful to understand the dynamics of virus spread and may help further vaccine and treatment studies. The increase of SARS-CoV-2 cases all over the world is giving more genomes that may present some visibility of populace structure. This study showed the common and new variations in SARS-CoV-2 isolates. The fight against COVID-19 will last a long time until an effective vaccine or drug is developed. However, we believe that collecting and sharing any data about SARS-CoV-2 virus and COVID-19 will be effective and may help the related studies. At that point, we should carry on with detecting new variations.

Study limitations

Our study has some limitations: the number of national genomes available at the time of analysis and the number of detected variations. Additionally, not using a more robust methodology, such as maximum likelihood, is another limitation of the study. These limitations should be eliminated for further studies.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/suppl/10.2217/fmb-2021-0118

Acknowledgments

The authors would like to thank all patients, clinicians and laboratory technicians who contributed to this study.

Financial & competing interests disclosure

The current study was supported by Kafkas University, Scientific Research Project Council, with the following project number: 2021-TS-26. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

This study was approved by both the Republic of Turkey Ministry of Health COVID-19 Scientific Research Evaluation Commission (approval date: 02/05/2020; number: 2020-05-02T16_13_50) and the local ethics committee of Kafkas University Faculty of Medicine (approval date: 06/05/2020 number: 80576354-050-99/130).

Availability of data & material

The data that support the findings of this study are openly available in NCBI GenBank and GISAID at https://submit.ncbi.nlm.nih.gov/subs/ and https://www.gisaid.org/.

Summary points

- To understand the transmission patterns and evolution of SARS-CoV-2, it is crucial and necessary to create better drug and vaccine designs for disease control and prevention. The analysis of genetic sequence data from a pathogen is known as an important tool in infectious disease epidemiology.
- Forty-seven positive samples were used for complete genome sequencing and variation analyses. A database of 68 complete genomes of SARS-CoV-2 strains belonging to different countries randomly selected from the Global Initiative on Sharing All Influenza Data database were compared with our isolations.
- The results of present study indicated that the SARS-CoV-2 sequences of our isolates have great similarity with all Turkish and European sequences.
- The fight against COVID-19 will last a long time until an effective vaccine or drug is developed. However, we believe that collecting and sharing any data about the SARS-CoV-2 virus and COVID-19 will be effective and may help the related studies. At that point, we should carry on detecting the new variations.

References

1. Lu J, du Plessis L, Liu Z, Hill *et al.* Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* 181(5), 997–1003 (2020).
2. WHO. Coronavirus (COVID-19) dashboard (2021). https://covid19.who.int/
3. Castillo AE, Parra B, Tapia P *et al.* Phylogenetic analysis of the first four SARS-CoV-2 cases in Chile. *J. Med. Virol.* 92(9), 1562–1566 (2020).
4. Van Dorp L, Acman M, Richard D *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* 83, 104351 (2020).
5. Faria NR, Kraemer MUG, Hill SC *et al.* Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science* 361(6405), 894–899 (2018).
6. Ladner JT, Grubaugh ND, Pybus OG *et al.* Precision epidemiology for infectious disease control. *Nat. Med.* 25(2), 206–211 (2019).
7. Grenfell BT, Pybus OG, Gog JR *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303(5656), 327–332 (2004).
8. Rambaut A, Holmes EC, O'Toole A *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* 5, 1403–1407 (2020).
9. GISAID. EpiCoV (2020). www.epicov.org
10. GISAID. Clade and lineage nomenclature (2021). www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/
11. Behjati S, Tarpey PS. What is next generation sequencing? *Arch. Dis. Child Educ. Pract. Ed.* 98(6), 236–238 (2013).
12. Hangyu Z, Rujiao L, Cong Y *et al.* Advantage of next-generation sequencing in dynamic monitoring of circulating tumor DNA over droplet digital PCR in cetuximab treated colorectal cancer patients. *Trans. Oncol.* 12(3), 426–431 (2019).
13. Bal A, Destras G, Gaymard A *et al.* Molecular characterization of SARS-CoV-2 in the first COVID-19 cluster in France reveals an amino acid deletion in nsp2 (Asp268del). *Clin. Microbiol. Infect.* 26(7), 960–962 (2020).
14. Eden JS, Rockett R, Carter I *et al.* An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol.* 6(1), 27 (2020).

15. Park WB, Kwon NJ, Choi SJ *et al.* Virus isolation from the first patient with SARS-CoV-2 in Korea. *J. Korean Med. Sci.* 35(7), e84 (2020).

16. Tabibzadeh A, Zamani F, Laali A *et al.* SARS-CoV-2 molecular and phylogenetic analysis in COVID-19 patients: a preliminary report from Iran. *Infect. Genet. Evol.* 84, 104387 (2020).

17. Yadav PD, Potdar VA, Choudhary ML *et al.* Full-genome sequences of the first two SARS-CoV-2 viruses from India. *Indian J. Med. Res.* 151(2-3), 200–209 (2020).

18. Stefanelli P, Faggioni G, Lo Presti A *et al.* Whole genome and phylogenetic analysis of two SARS-CoV-2 strains isolated in Italy in January and February 2020: additional clues on multiple introductions and further circulation in Europe. *Euro. Surveill.* 25(13), 1 (2020).

19. Zehender G, Lai A, Bergna A *et al.* Genomic characterization and phylogenetic analysis of SARS-COV-2 in Italy. *J. Med. Virol.* 92(9), 1637–1640 (2020).

20. Wan Y, Shang J, Graham R *et al.* Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *J. Virol.* 94(7), 307 (2020).

21. Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. *Gene. Rep.* 19, 100682 (2020).

22. 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.* 18(1), 179 (2020).

23. Phan T. Genetic diversity and evolution of SARS-CoV-2. *Infect. Genet. Evol.* 81, 104260 (2020).