Current mutatome of SARS-CoV-2 in Turkey reveals mutations of interest

DOĞA ESKİER
EVREN AKALP
ÖZLEM DALAN
GÖKHAN KARAKÜL AH
YAVUZ OKTAY

Follow this and additional works at: https://journals.tubitak.gov.tr/biology

Recommended Citation
ESKİER, DOĞA; AKALP, EVREN; DALAN, ÖZLEM; KARAKÜL AH, GÖKHAN; and OKTAY, YAVUZ (2021) "Current mutatome of SARS-CoV-2 in Turkey reveals mutations of interest," Turkish Journal of Biology: Vol. 45: No. 1, Article 10. https://doi.org/10.3906/biy-2008-56
Available at: https://journals.tubitak.gov.tr/biology/vol45/iss1/10

This Article is brought to you for free and open access by TÜBİTAK Academic Journals. It has been accepted for inclusion in Turkish Journal of Biology by an authorized editor of TÜBİTAK Academic Journals. For more information, please contact academic.publications@tubitak.gov.tr.
Current mutatome of SARS-CoV-2 in Turkey reveals mutations of interest

Doğa ESKİER1,2, Evren AKALP1, Özlem DALAN1, Gökhan KARAKÜLAH1,2,*, Yavuz OKTAY1,2,3,*

1İzmir Biomedicine and Genome Center (IBG), İzmir, Turkey
2İzmir International Biomedicine and Genome Institute (iBG-İzmir), Dokuz Eylül University, İzmir, Turkey
3Department of Medical Biology, Faculty of Medicine, Dokuz Eylül University, İzmir, Turkey

Abstract: As the underlying pathogen for the COVID-19 pandemic that has affected tens of millions of lives worldwide, SARS-CoV-2 and its mutations are among the most urgent research topics worldwide. Mutations in the virus genome can complicate attempts at accurate testing or developing a working treatment for the disease. Furthermore, because the virus uses its own proteins to replicate its genome, rather than host proteins, mutations in the replication proteins can have cascading effects on the mutation load of the virus genome. Due to the global, rapidly developing nature of the COVID-19 pandemic, local demographics of the virus can be difficult to accurately analyze and track, disproportionate to the importance of such information. Here, we analyzed available, high-quality genome data of SARS-CoV-2 isolates from Turkey and identified their mutations, in comparison to the reference genome, to understand how the local mutatome compares to the global genomes. Our results indicate that viral genomes in Turkey has one of the highest mutation loads and certain mutations are remarkably frequent compared to global genomes. We also made the data on Turkey isolates available on an online database to facilitate further research on SARS-CoV-2 mutations in Turkey.

Key words: COVID-19, SARS-CoV-2, coronavirus, genome analysis, mutation profiling, Turkey, database

1. Introduction
Coronavirus disease 2019 (COVID-19) is an ongoing pandemic, characterized by long-term respiratory damage and slow onset fever, and caused by the SARS-CoV-2 betacoronavirus. The virus was first observed in human patients in late 2019, in the Wuhan province of China, and soon after showed the capacity for human-to-human transmission (Chan et al., 2020; Riou and Althaus, 2020). As of 17 August 2020 there are over 21 million confirmed cases and 767,158 recorded deaths1. In addition to the immediately apparent symptoms, its long term effect on humans are still topics of research (Kochi et al., 2020; Lee et al., 2020; Li et al., 2020; Zhu et al., 2020). Because COVID-19 is a highly transmissible disease with a capacity for asymptomatic transmission (Wong et al., 2020); understanding the disease, as well as its underlying pathogen and its routes of transmissions is a high priority topic. As a result, extant databases on viral pathogens, such as GISAID2 (Elbe and Buckland-Merrett, 2017) and Nextstrain3 (Hadfield et al., 2018) have become vital resources for researchers who seek to track the evolution of the virus during its transmissions. SARS-CoV-2 has a single-stranded RNA genome that codes for the proteins responsible for its own replication, many of which are produced via cleavage of the Orf1ab polyprotein, the largest gene on the genome. Therefore, mutations in the SARS-CoV-2 genome can lead to cascading effects by reducing the fidelity of subsequent replication cycles. Key proteins in the RNA replication complex include nsps 7, 8, and 12 (also known as RNA dependent RNA polymerase or RdRp), which together form the core polymerase complex (Kirchdoerfer and Ward 2019; Peng et al., 2020), as well as nsp14, a dual function protein which joins the larger replication complex as a 3’-5’ error-correcting exonuclease (Subissi et al., 2014; Romano et al., 2020). Our previous findings show that frequently observed mutations in both nsp12 and nsp14 and Nextstrain3 (Hadfield et al., 2018) have become vital resources for researchers who seek to track the evolution of the virus during its transmissions. SARS-CoV-2 has a single-stranded RNA genome that codes for the proteins responsible for its own replication, many of which are produced via cleavage of the Orf1ab polyprotein, the largest gene on the genome. Therefore, mutations in the SARS-CoV-2 genome can lead to cascading effects by reducing the fidelity of subsequent replication cycles. Key proteins in the RNA replication complex include nsps 7, 8, and 12 (also known as RNA dependent RNA polymerase or RdRp), which together form the core polymerase complex (Kirchdoerfer and Ward 2019; Peng et al., 2020), as well as nsp14, a dual function protein which joins the larger replication complex as a 3’-5’ error-correcting exonuclease (Subissi et al., 2014; Romano et al., 2020). Our previous findings show that frequently observed mutations in both nsp12 and nsp14 are associated with an increase in mutation density in the SARS-CoV-2 genome (Eskier et al., 2020a, 2020b, 2020c). In this study, we aimed to analyze the current mutatome of SARS-CoV-2 in Turkey, with three main

Received: 29.08.2020 • Accepted/Published Online: 30.12.2020 • Final Version: 09.02.2021

* Correspondence: yavuz.oktay@ibg.edu.tr; gokhan.karakulah@deu.edu.tr

1World Health Organization (1948). WHO timeline – COVID-19 [online]. Website https://www.who.int/news-room/detail/27-04-2020-who-timeline---covid-19 [accessed 17 August 2020].
2The GISAID Initiative (2008). GISAID [online]. Website https://www.gisaid.org/ [accessed 27 August 2020].
3Nextstrain (2015). Nextstrain [online]. Website https://www.nextstrain.org [accessed 27 August 2020].
questions in mind: (i) are there any key reoccurring mutations observed in a large number of isolates? (ii) how does the distribution of mutations among isolates compare to other regions in the world? and finally, (iii) are there any mutations observed in Turkey but not the rest of the world? We focused on the latter two questions in particular, with an emphasis on mutations of interest previously described in the literature. Our findings reveal the presence of three main clades of SARS-CoV-2 in Turkey, roughly analogous to 19A, 20A, and 20B as described in NextStrain, with a preponderance of high mutability variants (Eskier et al., 2020a, 2020b, 2020c) compared to international isolates. Furthermore, we identified several frequently recurring, previously uncharacterized variants in Turkey isolates not observed in isolates from other countries, which can serve as potential candidates for validation and study. Furthermore, we collected our analysis of Turkey isolates in a regularly maintained and updated database, which we hope will serve as a potential resource for future research on the local mutatome of SARS-CoV-2.

2. Materials and methods
2.1. Genome sequence filtering, retrieval, and preprocessing
SARS-CoV-2 isolate genome sequences and the corresponding metadata were obtained from the GISAID EpiCoV database on 28 July 2020\(^4\). These sequences were filtered for location to limit our database to isolates with the location “Europe/Turkey”, which resulted in 180 isolate sequences. We applied further quality filters, including selecting only isolates obtained from human hosts (excluding environmental samples and animal hosts), those sequenced for the full length of the genome (sequence size of 29 kb or greater), and those with high coverage for the reference genome (<1% N content, <0.05% unique mutations, no unverified indel mutations), which further narrowed down the list to 166 isolates. To ensure alignment accuracy, as characters that are not one of A, C, G, T, or N would not be aligned according to potential biological meanings of the alternative characters, all nonstandard unverified nucleotide masking was changed to N, using the Linux sed command, and the isolates were aligned against the SARS-CoV-2 reference genome using the MAFFT (v7.450) alignment software (Katoh et al., 2002). Variant sites in the isolates were annotated using snp-sites (2.5.1), bcftools (1.10.2)\(^5\), and ANNOVAR (release date 24 October 2019) software (Wang et al., 2010; Page et al., 2016), to identify whether a given mutation was synonymous or nonsynonymous. In addition, the \(^4\)The GISAID Initiative (2008). EpiCoV database [online]. Website https://www.epicov.org/epi3/frontend [accessed 28 July 2020].
\(^5\)GitHub (2007). Bcftools repository [online]. Website https://github.com/samtools/bcftools [accessed 20 August 2020].

2.2. Development of the database and user interfaces
The genome data is stored using the MariaDB 10.3.22 database installed on Debian Linux 10 operating system. For web application, the genome data is visualized on the map using jVectorMap with HTML 5 and Ajax web development techniques, using the Django 3.0.5. framework and Python 3.7.3 programming language. A modified version of TreeTime, an open-source phylogenetic analysis software, is used to create the phylogenetic tree (Sagulenko et al., 2018).

3. Results
3.1. The distributions of mutations across isolates in Turkey
Our analysis of the genome sequences of 166 isolates from Turkey revealed 258 distinct mutations across the isolates, 87 of which are observed in multiple isolates, and 43 of them are found in at least five isolates (hereafter referred to as recurring mutations). 19 of the 43 recurring mutations are nonsynonymous, 21 are synonymous, and 3 are found outside of coding regions. C>T transitions are the most common, comprising over half of the mutations, consistent with previous international findings on C>U hypermutations in SARS-CoV-2 (Simmonds, 2020). The most commonly seen mutations are 3037 C>T, 14408 C>T, and 23403 A>G, observed together in 139 of the isolates, with one singleton instance of 23403 A>G, also consistent with previous findings (Pachetti et al., 2020; Yin, 2020). Orf1ab mutations are the most common, comprising 23 of the recurring mutations, consistent with the size of the gene, as Orf1ab makes up two thirds of the SARS-CoV-2 genome. Orf9 (nucleocapsid; N) gene has the second highest number of recurring mutations (n = 7, however, 3 of them are block mutations of 28881–28883 trinucleotide), followed by Orf5 (membrane; M) and S genes (n = 5) (Table 1).

To identify which of the recurring mutations are stronger indicators of Turkey genotype, we compared their frequency in the isolate population from Turkey to frequencies in other geographical regions, using a metric of mutation instance per sequenced isolate. To eliminate the potential confounding effect of earlier isolates having a lower number of mutations on average, and different regions having started sequencing efforts in different timetables, we selected isolates sequenced after the day when each region of interest had at least ten isolates sequenced. As Turkey was the latest region to have the required number of isolates (19 March 2020), it was used as the filtering metric. Four of the recurring mutations were found only in Turkey isolates, and six more were not
Table 1. Recurring mutations in Turkey.

| Position | Reference | Variant | Frequency | Cities | Exonic function  | Gene  | Aminoacid change |
|----------|-----------|---------|-----------|--------|-----------------|-------|-----------------|
| 23403    | A         | G       | 140       | İstanbul (65), Karaman (1), Kastamonu (1), Nevşehir (2), Ankara (13), Kocaeli (5), Siirt (1), Akşaray (1), Sakarya (3), Afyon (1), Balıkesir (1), Konya (1), Denizli (2), Tekirdağ (1), Tokat (1), Kars (41) | Nonsynonymous SNV | S     | p.D614G         |
| 3037     | C         | T       | 139       | İstanbul (65), Karaman (1), Kastamonu (1), Nevşehir (2), Ankara (13), Kocaeli (4), Siirt (1), Akşaray (1), Sakarya (3), Afyon (1), Balıkesir (1), Konya (1), Denizli (2), Tekirdağ (1), Tokat (1), Kars (41) | Synonymous SNV ORF1a; ORF1ab | ORF1ab | p.F924F         |
| 14408    | C         | T       | 139       | İstanbul (66), Nevşehir (1), Kocaeli (3), Ankara (6), Denizli (1), Ağrı (1), Tekirdağ (1), Tokat (1), Kars (41) | Nonsynonymous SNV ORF1ab | ORF1ab | p.P4715L        |
| 241      | C         | T       | 121       | İstanbul (40), Sakarya (2), Kocaeli (3), Ankara (3), Kars (25) | Nonsynonymous SNV ORF9 | ORF9  | p.R203K         |
| 28881    | G         | A       | 73        | İstanbul (40), Sakarya (2), Kocaeli (3), Ankara (3), Kars (25) | Synonymous SNV ORF9 | ORF9  | p.R203R         |
| 28882    | G         | A       | 73        | İstanbul (40), Sakarya (2), Kocaeli (3), Ankara (3), Kars (25) | Synonymous SNV ORF9 | ORF9  | p.G204R         |
| 28883    | G         | C       | 73        | İstanbul (40), Sakarya (2), Kocaeli (3), Ankara (3), Kars (25) | Nonsynonymous SNV ORF9 | ORF9  | p.G204R         |
| 25563    | G         | T       | 61        | İstanbul (24), Karaman (1), Kastamonu (1), Nevşehir (2), Ankara (8), Kocaeli (1), Siirt (1), Akşaray (1), Sakarya (1), Afyon (1), Balıkesir (1), Konya (1), Tekirdağ (1), Tokat (1), Kars (16) | Nonsynonymous SNV ORF3a | ORF3a | p.Q57H          |
| 18877    | C         | T       | 58        | İstanbul (24), Kastamonu (1), Nevşehir (2), Ankara (7), Kocaeli (1), Siirt (1), Akşaray (1), Sakarya (1), Afyon (1), Balıkesir (1), Konya (1), Tekirdağ (1), Tokat (1), Kars (15) | Synonymous SNV ORF1ab | ORF1ab | p.L6205L        |
| 7765     | C         | T       | 35        | İstanbul (18), Siirt (1), Nevşehir (1), Tekirdağ (1), Kars (14) | Synonymous SNV ORF1ab | ORF1ab | p.S2500S        |
| 17690    | C         | T       | 35        | İstanbul (18), Siirt (1), Nevşehir (1), Tekirdağ (1), Kars (14) | Nonsynonymous SNV ORF1ab | ORF1ab | p.S8509L        |
| 11083    | G         | T       | 31        | Kayseri (1), İstanbul (11), Karaman (1), Ankara (6), Balıkesir (1), Çanakkale (1), Eskişehir (1), Kocaeli (4), Mardin (1), Ağrı (1), Kars (3) | Nonsynonymous SNV ORF1ab | ORF1ab | p.L3606F        |
| 29742    | G         | T       | 24        | Kayseri (1), Ankara (5), Kocaeli (5), Balıkesir (1), Çanakkale (1), Eskişehir (1), Mardin (1), İstanbul (8), Ağrı (1) | Nonsynonymous SNV ORF10; ORF9 | ORF10; ORF9 | p.V378I         |
| 1397     | G         | A       | 23        | Kayseri (1), Ankara (5), Balıkesir (1), Çanakkale (1), Eskişehir (1), Kocaeli (4), Mardin (1), İstanbul (8), Ağrı (1) | Nonsynonymous SNV ORF1ab | ORF1ab | p.V378I         |
| 12809    | C         | T       | 23        | İstanbul (20), Kars (3) | Nonsynonymous SNV ORF1ab | ORF1ab | p.L4182F        |
| 28688    | T         | C       | 23        | Kayseri (1), Ankara (5), Balıkesir (1), Çanakkale (1), Eskişehir (1), Kocaeli (4), Mardin (1), İstanbul (8), Ağrı (1) | Synonymous SNV ORF9 | ORF9  | p.L139L         |
| 27703    | G         | T       | 20        | Kars (20) | Nonsynonymous SNV ORF7a | ORF7a | p.V104F         |
| Position | Reference | Variant | Frequency | Cities                                      | Exonic function | Gene      | Aminoacid change |
|----------|-----------|---------|-----------|--------------------------------------------|----------------|----------|-----------------|
| 24262    | G         | T       | 20        | Kars (20)                                  | Nonsynonymous SNV | S        | p.M900I         |
| 313      | C         | T       | 19        | Kars (19)                                  | Synonymous SNV   | ORF1a; ORF1ab | p.L166L         |
| 2509     | C         | T       | 19        | Kars (19)                                  | Synonymous SNV   | ORF1a; ORF1ab | p.P748P         |
| 13620    | C         | T       | 19        | Kars (19)                                  | Synonymous SNV   | ORF1ab    | p.D4452D        |
| 14724    | C         | T       | 19        | Kars (19)                                  | Synonymous SNV   | ORF1ab    | p.F4820F        |
| 19839    | T         | C       | 18        | Ankara (2), İstanbul (15), Kars (1)        | Synonymous SNV   | ORF1ab    | p.N6525N        |
| 26735    | C         | T       | 15        | Kastamonu (1), Nevşehir (1), Ankara (5), Kocaeli (1), Aksaray (1), Sakarya (1), Afyon (1), Balikesir (1), Konya (1), İstanbul (2) | Synonymous SNV | ORF5      | p.Y71Y          |
| 8326     | C         | T       | 14        | Kars (14)                                  | Synonymous SNV   | ORF1a; ORF1ab | p.D2687D        |
| 2113     | C         | T       | 13        | İstanbul (11), Nevşehir (1), Tekirdağ (1)  | Synonymous SNV   | ORF1a; ORF1ab | p.I616I         |
| 884      | C         | T       | 12        | Balikesir (1), Çanakkale (1), Eskişehir (1), Kocaeli (3), Ankara (3), İstanbul (3) | Nonsynonymous SNV | ORF1a; ORF1ab | p.R207C         |
| 8653     | G         | T       | 12        | Balikesir (1), Çanakkale (1), Eskişehir (1), Kocaeli (3), Ankara (3), İstanbul (3) | Nonsynonymous SNV | ORF1a; ORF1ab | p.M2796I        |
| 26549    | C         | T       | 12        | Kocaeli (2), İstanbul (5), Ankara (3), Tokat (1), Denizli (1) | Synonymous SNV | ORF5      | p.T9T           |
| 5015     | G         | A       | 11        | Kars (11)                                  | Nonsynonymous SNV | ORF1a; ORF1ab | p.V1584M        |
| 28854    | C         | T       | 10        | Ankara (3), Aksaray (1), Sakarya (1), Konya (1), İstanbul (3), Kars (1) | Nonsynonymous SNV | ORF9      | p.S194L         |
| 228      | C         | T       | 9         | Kocaeli (1), İstanbul (4), Ankara (3), Denizli (1) |                      | ORF1a; ORF1ab |                 |
| 22444    | C         | T       | 9         | Ankara (3), Aksaray (1), Sakarya (1), Konya (1), İstanbul (3) | Synonymous SNV | S         | p.D294D         |
| 9514     | A         | G       | 8         | Kocaeli (1), İstanbul (4), Ağrı (1), Ankara (2) | Synonymous SNV   | ORF1a; ORF1ab | p.L3083L        |
| 26720    | G         | C       | 8         | Kocaeli (1), İstanbul (4), Ağrı (1), Ankara (2) | Synonymous SNV   | ORF5      | p.V666V         |
| 9479     | G         | T       | 7         | Kocaeli (1), İstanbul (4), Ağrı (1), Ankara (1) | Nonsynonymous SNV | ORF1a; ORF1ab | p.G3072C        |
| 28835    | T         | C       | 7         | Kocaeli (1), İstanbul (4), Ağrı (1), Ankara (1) | Nonsynonymous SNV | ORF9      | p.S188P         |
| 5736     | C         | T       | 6         | Ankara (3), Denizli (1), İstanbul (2)       | Nonsynonymous SNV | ORF1a; ORF1ab | p.A1824V        |
| 16428    | C         | T       | 6         | Kars (6)                                   | Synonymous SNV   | ORF1ab    | p.Y3388Y        |
| 25611    | C         | A       | 6         | İstanbul (6)                               | Synonymous SNV   | ORF3a     | p.L73L          |
| 28857    | G         | T       | 6         | Kars (6)                                   | Nonsynonymous SNV | ORF9      | p.R1951         |
| 10702    | C         | T       | 5         | Eskişehir (1), Ankara (1), İstanbul (3)     | Synonymous SNV   | ORF1a; ORF1ab | p.D3479D        |
| 20268    | A         | G       | 5         | Ankara (2), Denizli (1), Kocaeli (1), İstanbul (1) | Synonymous SNV   | ORF1ab    | p.L6668L        |
recurring mutations in other regions. Therefore, we focused the comparisons on the remaining 33 recurring mutations. We identified the percentages of these mutations in each region (Africa, Asia, Europe, North America, Oceania, South America), as well as worldwide totals, excluding isolates from Turkey, where applicable, and compared them to the corresponding percentages in Turkey (Table 2).

Table 2. Frequencies of SARS-CoV-2 mutations in different geographical regions.

| Mutation | Africa | Asia | Europe | North America | Oceania | South America | Turkey | Worldwide |
|----------|--------|------|--------|--------------|---------|---------------|--------|-----------|
| 23403A>G | 88.73% (252) | 58.15% (835) | 82.34% (10484) | 78.15% (4470) | 67.07% (894) | 93.48% (258) | 84.34% (140) | 78.94% (17193) |
| 3037C>T | 79.58% (226) | 58.84% (845) | 82.03% (10444) | 78.25% (4476) | 66.99% (893) | 93.48% (258) | 83.73% (139) | 78.7% (17142) |
| 14408C>T | 87.68% (249) | 58.57% (841) | 82.32% (10481) | 78.36% (4482) | 67.14% (895) | 93.48% (258) | 83.73% (139) | 79% (17206) |
| 241C>T | 88.73% (252) | 58.29% (837) | 82.24% (10471) | 77.36% (4425) | 51.24% (683) | 93.48% (258) | 72.89% (121) | 77.71% (16926) |
| 20268A>G | 3.17% (9) | 6.48% (93) | 0.02% (2) | 0% (0) | 0% (0) | 0% (0) | 5.42% (9) | 0.03% (6) |
| 9479G>T | 0.35% (1) | 0.14% (2) | 0.07% (9) | 0.05% (3) | 0.15% (2) | 0% (0) | 4.22% (7) | 0.08% (17) |
| 16428C>T | 0% (0) | 0% (0) | 0.01% (1) | 0.12% (7) | 0% (0) | 0.36% (1) | 3.61% (6) | 0.04% (9) |
| 28857A>G | 0% (0) | 0.07% (1) | 0.09% (12) | 0.02% (1) | 0% (0) | 0% (0) | 3.61% (6) | 0.06% (14) |
| 20268A>G | 3.17% (9) | 1.04% (15) | 8.18% (1042) | 1.03% (59) | 4.13% (55) | 90.66% (25) | 3.01% (5) | 5.53% (1205) |
The 8326 C > T mutation was first isolated from a patient in Taipei on March 19 and only two other cases were reported (UK and Denmark) before the first case in Turkey (Kars) on April 29. Thirteen more cases with 8326 C>T in the same city implicates local transmission, and with 8 cases reported on the most recent update (July 15), this particular mutation is a candidate for even further spread, assuming that the isolates sequenced were randomly selected across the infected population, instead of all being selected from a known cluster of patients. The 12809 C>T mutation was first reported in an isolate from Washington/USA collected on March 14, which spread to five more states by the end of the month; however, only 4 more cases were reported afterwards, the last being one on April 27. The first isolate from Turkey with 12809 C>T mutation was collected on April 13 in İstanbul (EPI_ISL_480230). Being present in only three other countries with one isolate each (UK, India, Australia), in addition to the USA, this mutation was very likely an introduction from the USA. Since then, 20 more cases in İstanbul and 3 more cases in Kars were reported, all in April and May. 13620 C>T mutation was first reported in an isolate in Italy on March 5 and later in South Africa, USA, Denmark, Belgium, Luxembourg and Singapore by April 6. However, only three more cases were reported for the rest of April: two in Italy on April 12, and one in USA on April 28. The first case with the same mutation in Turkey was reported in Kars on May 17 and has been reported in 18 more cases, all in the same city. Initial phylogenetic analysis does not support introduction of this mutation from abroad, however, limited sampling makes it difficult to reach a definitive conclusion. 13620 C>T, 14724 C>T, and 27703 G>T mutations are linked in SARS-CoV-2 genomes from Turkey, all from a single city (Kars), suggesting a founder effect and local transmission. It bears noting that each of the four mutations exclusively found as recurring in Turkey are limited to a single batch of isolates obtained by a single center, therefore pending verification.

Afterwards, we sought to understand how the mutation load of the isolates in Turkey compare to distributions in other regions. Using our previous date filter, we calculated the number of single nucleotide variants (SNVs) per isolate in each region (Table 3). Turkey had the highest number of SNVs per isolate, followed by South America. In comparison, Africa, another region which started sequencing efforts later than the other regions, had a mean SNV number lower than that of Asia, the region with the earliest sequences available, implying that the mutation numbers are strongly influenced by other factors in addition to the date of introduction of the virus to the region. We also compared the number of SNVs per isolate in each region per gene, normalized by kilobase of gene region (Table 4). Turkey had the most SNVs of any region in Orf1ab, M, and Orf7a genes, with Orf7a having more than three times as many SNVs as any other region.

### Database implementation
Data regarding Turkey isolates are available as a database comprising an interactive phylogenetic tree of the isolates, a geographical heatmap of sequenced isolates, and tables for both the mutatome of individual isolates, and summaries of the mutations observed in the isolates (Figure). The phylogenetic tree can be viewed both in real time and divergence time, and colored according to nucleotide of interest, location, or sequencing date. The tables are generated using the sequencing metadata available from GISAID as well as ANNOVAR variant annotation tables. We aim to regularly validate and update the database as new sequences are made available.

### Discussion
COVID-19 has been causing tremendous challenges for clinicians, healthcare systems, societies, and governments. Future plans include implementation of Nextstrain clade and branch information in the phylogenetic tree to aid the user in comparisons with international sequencing data.

| Region         | Orf1ab | S    | Orf3 | E    | M    | Orf6 | Orf7a | Orf7b | Orf8 | N    | Orf10 |
|----------------|--------|------|------|------|------|------|-------|-------|------|------|-------|
| Turkey         | 0.25   | 0.32 | 0.49 | 0.06 | 0.38 | 0.03 | 0.44  | 0.05  | 0.11 | 1.45 | 0.06  |
| Worldwide      | 0.19   | 0.29 | 0.62 | 0.08 | 0.13 | 0.11 | 0.11  | 0.10  | 0.38 | 0.84 | 0.10  |
| Africa         | 0.21   | 0.40 | 0.25 | 0.06 | 0.15 | 0.00 | 0.06  | 0.00  | 0.23 | 0.75 | 0.12  |
| Asia           | 0.20   | 0.31 | 0.48 | 0.11 | 0.31 | 0.10 | 0.13  | 0.12  | 0.36 | 0.83 | 0.10  |
| Europe         | 0.18   | 0.29 | 0.48 | 0.09 | 0.13 | 0.09 | 0.12  | 0.09  | 0.15 | 1.10 | 0.11  |
| North America  | 0.20   | 0.26 | 0.97 | 0.06 | 0.09 | 0.12 | 0.08  | 0.12  | 0.84 | 0.27 | 0.11  |
| Oceania        | 0.19   | 0.26 | 0.66 | 0.05 | 0.13 | 0.08 | 0.07  | 0.05  | 0.62 | 0.75 | 0.06  |
| South America  | 0.16   | 0.33 | 0.59 | 0.08 | 0.06 | 1.68 | 0.11  | 0.14  | 0.11 | 1.46 | 0.06  |

Table 3. Mean number of variants per isolate in different geographical regions.

The database is freely accessible at http://covid19.ibg.edu.tr.
provide a valuable and practical resource for researchers is a challenging task. We believe that our database will be closely monitored, with tens of millions of mutations expected. New mutations that affect viral behavior may arise, supporting a role for the S D614G mutation in increased infectivity and likely in transmissibility of SARS-CoV-2 (Daniloski et al., 2020; Korber et al., 2020). It is possible that some mutations have the capacity to impact on the replication and transmission cycle of the virus, a higher mutation density in these genes can lead to an accelerated mutation rate. Of note, the 18877 C>T mutation in nsp14, the 3’-5’ exonuclease responsible for error correction during genomic replication, has the second highest frequency in Turkey of any country. Our previous study (Eskier et al., 2020a) shows a strong correlation between increased mutation density and the 18877 C>T mutation, which might be a potential reason for Turkey’s increased SNV average per isolate.

Two groups of mutations we identified that is worth further attention are the 3037 C>T, 14408 C>T, 23403 A>G haplotype, and the 28881–28883 block mutation. Both of these groups of mutations are found almost exclusively in Turkey, and worldwide. In both cases, Turkey has a higher incidence of mutations in these groups than worldwide averages, and four of the major regions (Asia, Europe, North America, Oceania). We previously found that the 14408 C>T and 23403 A>G mutations, when occurring together, are strongly associated with increased mutation density over time (Eskier et al., 2020a), and the prevalence of both these mutations and the 18877 C>T mutation in Turkey isolates may further contribute to a variant-rich mutation landscape (Eskier et al., 2020b).

Our findings show the viral isolates in Turkey have accumulated a higher number of mutations compared to other regions on average, even after normalizing for the isolates sequenced earlier during the pandemic having accumulated fewer mutations. Furthermore, it has more mutations in the Orf1ab gene, which produces the polyprotein that is cleaved into the mature peptides responsible for viral replication, than any other region. In addition, it has the third highest number of mutations in the S gene, which is responsible for the viral infection of the cells. As these two genes have the highest potential impact on the replication and transmission cycle of the virus, a higher mutation density in these genes can lead to an accelerated mutation rate. Of note, the 18877 C>T mutation in nsp14, the 3’-5’ exonuclease responsible for error correction during genomic replication, has the second highest frequency in Turkey of any country. Our previous study (Eskier et al., 2020a) shows a strong correlation between increased mutation density and the 18877 C>T mutation, which might be a potential reason for Turkey’s increased SNV average per isolate.

Table 4. SNV densities of SARS-CoV-2 genes in different geographical regions.

| Region       | Mean number of variants per isolate |
|--------------|-------------------------------------|
| Turkey       | 10.18                               |
| Worldwide    | 8.01                                |
| Africa       | 8.47                                |
| Asia         | 8.53                                |
| Europe       | 7.88                                |
| North America| 8.20                                |
| Oceania      | 7.54                                |
| South America| 9.32                                |

and has required development of novel approaches to fight the pandemic. With an unpredictable future course for the ongoing pandemic, close monitoring and characterization of mutations has emerged as top priorities for better understanding of possible genotype-phenotype relations, and therefore better management of healthcare efforts.

Mutations in any viral infection, especially those that have crossed interspecies barriers, have to be considered in the context of natural selection. As the evolution of a virus will likely affect its fitness in a new host, any attempts against such an infection have to consider the causal relationships between genomic variances and the spread of the virus. Previous studies suggest that the selective pressure on mutations in SARS-CoV-2 in human hosts are largely confined to modest positive selection, with very little purifying selection, due to the short span of the pandemic, and that most of the positive selection have occurred in previous hosts (MacLean et al., 2020). Therefore, any investigation of the mutations will need to consider most of the mutations have to be beneficial or neutral to create true strains of the virus. A comprehensive analysis by Jungreis et al. (2020) showed that SARS-CoV-2 mutations are excluded from the evolutionarily conserved amino acid residues and nucleotides, and the authors concluded both synonymous and nonsynonymous mutations are under purifying selection. Therefore, not only the nonsynonymous mutations, but also the synonymous ones should be considered as potentially functional.

Many studies already provided lines of evidence that supports a role for the S D614G mutation in increased infectivity and likely in transmissibility of SARS-CoV-2 (Daniloski et al., 2020; Korber et al., 2020). It is possible that new mutations that affect viral behavior may arise, and therefore emergence and spreading of such mutations should be monitored closely. However, with tens of millions affected worldwide, monitoring of every single mutation is a challenging task. We believe that our database will provide a valuable and practical resource for researchers in Turkey, as well as in other countries, to track the spread of SARS-CoV-2 mutations in Turkey.

Our findings show the viral isolates in Turkey have accumulated a higher number of mutations compared to other regions on average, even after normalizing for the isolates sequenced earlier during the pandemic having accumulated fewer mutations. Furthermore, it has more mutations in the Orf1ab gene, which produces the polyprotein that is cleaved into the mature peptides responsible for viral replication, than any other region. In addition, it has the third highest number of mutations in the S gene, which is responsible for the viral infection of the cells. As these two genes have the highest potential impact on the replication and transmission cycle of the virus, a higher mutation density in these genes can lead to an accelerated mutation rate. Of note, the 18877 C>T mutation in nsp14, the 3’-5’ exonuclease responsible for error correction during genomic replication, has the second highest frequency in Turkey of any country. Our previous study (Eskier et al., 2020a) shows a strong correlation between increased mutation density and the 18877 C>T mutation, which might be a potential reason for Turkey’s increased SNV average per isolate.
This phylogenetic tree displays the phylogenetic relationships between isolates sampled and sequenced in Turkey, and can be viewed in both phylogenetic distance and time of sequencing distance modes. It is modeled on and uses data obtained via the TreeTime phylogenetic analysis network.

Pavel Sagulenko, Vadim Puller, Richard A Neher, TreeTime: Maximum-likelihood phylogenetic analysis, Virus Evolution, Volume 4, Issue 1, January 2018, vex042, https://doi.org/10.1093/virology/vex042

Toggle tree-time selects whether the distances of each branch from the root are based on date of sampling (tree-time selected), or on phylogenetic distance (tree-time deselected). Color by dropdown menu allows each branch and node to be colored according to the date of interest, and includes Numdate (date given numerically with year followed by the percent of year passed), nucleotide (typed in the input box Pos), province of sequencing, full virus name, country of sequencing, GISAID accession ID, continent of sequencing, date of sequencing given in YYYY-MM-DD, and the length of connecting branch to the upstream node.

This map of Turkey shows most current information regarding identified and sequenced SARS-CoV-2 isolates, with cities colored based on the number of sequenced isolates. You can hover your cursor over a city to see the number of isolates, and select a city from the pull down menu to see detailed information on the isolates sequenced in that city, including identified mutations and their effect on the peptide sequence, if any.

Number of isolates: 166

**Figure.** Snapshot of SARS-CoV-2 genome map of Turkey database. Due to size constraints, tables showing information on individual isolates, or summaries of individual variants, are not included.
available in large batches by the centers, which further introduces bias to the samples by potentially generating sequencing or assembly artifacts to the sequences. Unless verified by multiple centers, in multiple batches, or by other experimental methods, caution is required when studying these mutations. As more genomes are sequenced, a more clear picture of the SARS-CoV-2 mutatome in Turkey will emerge and we will likely be able to draw more solid conclusions.

Finally, it should be noted that mutational profiles of viral genomes may determine whether infected patients will develop lasting immunity and remain protected from re-infection. Although exposure to SARS-CoV-2 protected rhesus macaques from re-infection with the same strain of virus (Deng et al., 2020), there are questions still remaining to be answered related to whether each recovered patient will have lasting immunity. Recent news within days reported that four patients from Hong Kong, Belgium, the Netherlands, and USA, who had earlier recovered from COVID-19 has been reinfected, with a different strain of SARS-CoV-2 than the original infection (Tillett et al., 2021). In support of this observation, an earlier study reported that convalescent plasma from some of the COVID-19 patients showed reduced neutralizing activity against pseudoviruses with D614G mutation in

References

Ayub M (2020). Reporting two SARS-CoV-2 strains based on a unique trinucleotide-bloc mutation and their potential pathogenic difference. Preprints. : 10.20944/preprints202004.0337.v1.

Chan JF-W, Yuan S, Kok K-H, To KK-W, Chu H et al. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. The Lancet 395: 514-523. doi: 10.1016/S0140-6736(20)30154-9.

Daniloski Z, Jordan TX, Ilmain JK, Guo X, Bhabha G et al. (2020). The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. bioRxiv 2020. doi: 10.1101/2020.06.14.151357.

Deng W, Bao L, Liu J, Xiao C, Liu J et al. (2020). Primary exposure to SARS-CoV-2 protects against reinfection in rhesus macaques. Science 369: 818-823. doi: 10.1126/science.abc5343.

Elbe S, Buckland-Merrett G (2017). Data, disease and diplomacy: GISAID’s innovative contribution to global health. Global Challenges 1: 33-46. doi: 10.1002/gch2.1018.

Eskier D, Suner A, Karakülal G, Oktay Y (2020a). Mutation density changes in SARS-CoV-2 are related to the pandemic stage but to a lesser extent in the dominant strain with mutations in spike and RdRp. PeerJ – Life & Environment 8: e9703. doi: 10.7717/peerj.9703.

Eskier D, Suner A, Oktay Y, Karakülal G (2020b). Mutations of SARS-CoV-2 nsp14 exhibit strong association with increased genome-wide mutation load. bioRxiv 2020. doi: 10.1101/2020.08.12.248732.

Eskier D, Karakülal G, Suner A, Oktay Y (2020c). RdRp mutations are associated with SARS-CoV-2 genome evolution. PeerJ – Life & Environment 8: e9587. doi: 10.7717/peerj.9587.

Grifoni A, Weiskopf D, Ramirez SL, Mateus J, Dan JM et al. (2020). Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell 181: 1489-1501. doi: 10.1016/j.cell.2020.05.015.

Hachim A, Kavian N, Cohen CA, Chin AW, Chu DKW et al. (2020). ORF8 and ORF3b antibodies are accurate serological markers of early and late SARS-CoV-2 infection. Nature Immunology 21: 1293-1301. doi: 10.1038/s41590-020-0773-7.

Hadfield J, Megill C, Bell SM, Huddleston J, Potter B et al. (2018). Nexstrain: real-time tracking of pathogen evolution. Bioinformatics 34: 4121-4123. doi: 10.1093/bioinformatics/bty407.

ESKİER et al. / Turk J Biol
culture environment (Hue et al., 2020). We do not have a clear understanding of the viral determinants of lasting immunity to SARS-CoV-2, however, it seems that certain viral proteins may be more critical than others, based on analyses of patient plasma samples. Grifoni et al. (2020) suggested that M, Spike and N proteins are the major determinants of CD4+ response, with additional responses to nsp3, nsp4, ORF3a and ORF8. Hachim et al. (2020) showed that ORF8, ORF3b and N proteins of SARS-CoV-2 elicited the strongest specific antibody responses in infected patients. It is plausible that certain mutations within these proteins affect the immune response, however, it remains to be explored whether any of the mutations common or more frequently seen in Turkish isolates have any effect on the immune response.

Acknowledgment/Disclaimers

The authors would like to extend their thanks to Izmir Biomedicine and Genome Center (IBG) COVID-19 platform IBG-COVID-19 for their support in implementing the study and the Scientific and Technological Research Council of Turkey (TÜBİTAK) for their financial support of IBG-COVID-19. Yavuz Oktay is supported by the Turkish Academy of Sciences Young Investigator Program (TÜBA-GEİBİP).

Conflict of interest

The authors have no conflicts of interest to disclose.
Hue S, Beldi-Ferchiou A, Bendib I, Surenaud M, Fourati S et al. (2020). Uncontrolled innate and impaired adaptive immune responses in patients with COVID-19 acute respiratory distress syndrome. American Journal of Respiratory and Critical Care Medicine 202: 1509-1519. doi: 10.1164/rccm.202005-1885OC.

Jungraei I, Sealfon R, Kellis M (2020). Sarbecovirus comparative genomics elucidates gene content of SARS-CoV-2 and functional impact of COVID-19 pandemic mutations. bioRxiv 2020. doi: 10.1101/2020.06.02.130955.

Katoh K, Misawa K, Kuma K, Miyata T (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30: 3059-3066. doi: 10.1093/nar/gkf436.

Kirchdoerfer RN, Ward AB (2019). Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. Nature Communications 10: 2342. doi: 10.1038/s41467-019-10280-3.

Kochi AN, Tagliari AP, Forleo GB, Fassini GM, Tondo C (2020). Cardiac and arrhythmic complications in patients with COVID-19. Journal of Cardiovascular Electrophysiology 31: 1003-1008. doi: 10.1111/jce.14479.

Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J et al. (2020). Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182: 812-827. e19. doi: 10.1016/j.cell.2020.06.043.

Lee J-C, Huh T-I, Huang Y-H (2020). Gastrointestinal and liver manifestations in patients with COVID-19. Journal of the Chinese Medical Association 83: 521-523. doi: 10.1097/JCMA.0000000000000319.

Li Y-C, Bai W-Z, Hashikawa T (2020). The neuroinvasive potential of SARS-CoV2 may play a role in the respiratory failure of COVID-19 patients. Journal of Medical Virology 92: 552-555. doi: 10.1002/jmv.25728.

MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF et al. (2020). Natural selection in the evolution of SARS-CoV-2 in bats, not humans, created a highly capable human pathogen. bioRxiv 2020. doi: 10.1101/2020.05.28.122366.

McBride R, Van Zyl M, Fielding BC (2014). The coronavirus nucleocapsid is a multifunctional protein. Viruses 6: 2991-3018. doi: 10.3390/v6082991.

Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E et al. (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. Journal of Translational Medicine 18: 179. doi: 10.1186/s12976-020-02344-6.

Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T et al. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-Fasta alignments. Microbial Genomics 2 (4): e000056. doi: 10.1099/mgen.0.000056.

Peng Q, Peng R, Yuan B, Zhao J, Wang M et al. (2020). Structural and biochemical characterization of the nsp12-nsp7-nsp8 core polymerase complex from SARS-CoV-2. Cell Reports 31: 107774. doi: 10.1016/j.celrep.2020.107774.

Riou J, Althaus CL (2020). Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin 25. doi: 10.2807/1560-7917.ES.2020.25.4.2000058.

Romano M, Ruggiero A, Squeglia F, Maga G, Berisio R (2020). A structural view of SARS-CoV-2 RNA replication machinery: RNA synthesis, proofreading and final capping. Cells 9. doi: 10.3390/cells9051267.

Sagulenko P, Puller V, Neher RA (2018). TreeTime: Maximum-likelihood phylogenetic analysis. Virus Evolution 4. doi: 10.1093/ve/vev042.

Simmonds P (2020). Rampant C>U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses – causes and consequences for their short and long evolutionary trajectories. bioRxiv 2020. doi: 10.1101/2020.05.01.072330.

Subissi L, Posthuma CC, Collet A, Zevenhoven-Dobbe JC, Gorbalenya AE et al. (2014). One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. Proceedings of the National Academy of Sciences 111: E3900-E3909. doi: 10.1073/pnas.1323705111.

Tillett RL, Sevinsky JR, Hartley PD, Kerwin H, Crawford N et al. (2021). Genomic evidence for re-infection with SARS-CoV-2: a case study. The Lancet Infectious Diseases 21: 52-58. doi: 10.1016/S1473-3099(20)30764-7.

Tylor S, Andonov A, Cutts T, Cao J, Grudesky E et al. (2009). The SR-rich motif in SARS-CoV nucleocapsid protein is important for virus replication. Canadian Journal of Microbiology 55: 254-260. doi: 10.1139/w08-139.

Wang K, Li M, Hakonarson H (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research 38: e164-e164. doi: 10.1093/nar/gkq603.

Wong J, Jamaludin SA, Alikhan MF, Chaw L (2020). Asymptomatic transmission of SARS-CoV-2 and implications for mass gatherings. Influenza and Other Respiratory Viruses 14 (5): 596-598. doi: 10.1111/irv.12767.

Yin C (2020). Genotyping coronavirus SARS-CoV-2: methods and implications. Genomics 112 (5): 3588-3596. doi: 10.1016/j.ygeno.2020.04.016.

Zhu H, Rhee J-W, Cheng P, Waliany S, Chang A et al. (2020). Cardiovascular complications in patients with COVID-19: consequences of viral toxicities and host immune response. Current Cardiology Reports 22: 32. doi: 10.1007/s11886-020-01292-3.