



Research article

Complete genome sequencing and molecular characterization of SARS-CoV-2 from COVID-19 cases in Alborz province in Iran



Amir Kaffashi^{a,*}, Jiabin Huang^b, Amir Bairami^c, Mohammad Hossein Fallah Mehrabadi^a, Somayeh Yaslianifard^d, Mohsen Bashashati^a, S Reza Banihashemi^a, Fatemeh Soleimanifar^e, Mohsen Lotfi^a, Morteza Taghizadeh^a, Alireza Soleimani^f, Akbar Khorasani^a, Farzaneh Moshiri^e, Sayed-Hamidreza Mozhgani^{d,**}

^a Agricultural Research Education and Extension Organization (AREEO), Razi Vaccine and Serum Research Institute, Karaj, Iran

^b Institute for Medical Microbiology, Virology and Hygiene, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

^c Department of Medical Parasitology and Mycology, Faculty of Medicine, Alborz University of Medical Sciences, Karaj, Iran

^d The Department of Microbiology, School of Medicine, Alborz University of Medical Sciences, Karaj, Iran

^e Department of Medical Biotechnology, School of Medicine, Alborz University of Medical Sciences, Karaj, Iran

^f Department of Infectious Disease, Imam Ali Hospital, Alborz University of Medical Sciences, Karaj, Iran

ARTICLE INFO

Keywords:

COVID-19

Phylogenetic

Lineage

Clade

SARS-CoV-2

Single nucleotide variation

ABSTRACT

Iran was among countries which was hard hit at the early stage of the coronavirus disease 2019 (COVID-19) pandemic and dealt with the second wave of the pandemic in May and June 2020; however, there are a very limited number of complete genome sequences of acute respiratory syndrome coronavirus 2 (SARS-CoV-2) from Iran. In this study, complete genome sequences of the virus in the samples obtained from three patients in Alborz province in May and June 2020 were generated and analyzed using bioinformatic methods. The sequenced genomes were positioned in a cluster with B.4 lineage along with the sequences from other countries namely, United Arab Emirates and Oman. There were seven single nucleotide variations (SNVs) in common in all samples and only one of the sequenced genomes showed the D614G amino acid substitution. Three SNVs, 1397 G > A, 28688T > C, 29742 G > T, which had already been reported in February, were found with high frequency in all the sequenced genomes in this study, implying that viral diversity reflected in the early stages of viral transmission in Iran were established in the second wave. Considering the importance of molecular epidemiology in response to ongoing pandemic, there is an urgent need for more complete genome sequencing and comprehensive analyses to gain insight into the transmission, adaptation and evolution of the virus in Iran.

1. Introduction

The World Health Organization (WHO) declared the coronavirus disease 2019 (COVID-19) as a global pandemic on March 11, 2020 [1]. Iran was among the first countries where it was hardly hit at the early stage of the pandemic [2]. The first COVID-19 case in Iran was confirmed on 19 February 2020, in Qom, a city in central Iran, resulting in death [3]. The virus spread rapidly and the country soon ranked third in the world with the highest number of reported COVID-19 cases after China, and Italy, up to March 16, 2020 [4]. The epidemic curve of the disease reached a peak in Iran and then started to decrease sharply following restrictive measures and lockdown applied by the authorities. When such

measures were relaxed from early in May, the shape of the curve became bimodal because of a sharp increase in the number of COVID-19 cases peaking again at the beginning of June [5].

The epidemiological data will provide necessary information for decision making if combined with real time whole genome sequencing data [6]. The genomic data related to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the global initiative on sharing all influenza data (GISAID) (<https://www.gisaid.org>) have been growing dramatically since the first sequence of SARS-CoV-2 was deposited in January 2020 [7]. While there are over 2,370,000 complete genome sequences of SARS-CoV-2 in GISAID as of July 20, 2021, the number of complete genome sequences from Iran is less than 400 sequences, which

* Corresponding author.

** Corresponding author.

E-mail addresses: a.kaffashi@rvsri.ac.ir (A. Kaffashi), hamidrezamozhgani@gmail.com (S.-H. Mozhgani).

<https://doi.org/10.1016/j.heliyon.2021.e08027>

Received 19 May 2021; Received in revised form 5 July 2021; Accepted 15 September 2021

2405-8440/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

clearly shows the necessity of much more complete genome sequencing of the virus in the country.

This study aims to address the above-mentioned issue and report complete genome sequencing of the three SARS-CoV-2 samples obtained from three patients in Alborz province during the second wave of COVID-19 in May and June 2020. Alborz province is adjacent to Tehran province wherein the capital and the largest city in the country is located and there are large number of people travel between these provinces every day (Figure 1). Alborz province is the among provinces with the highest population density in Iran and the high incidence rate of COVID-19 [8]. The samples from the second wave of COVID-19 were used because it is more likely to see common or distinct features among the sequences in an outbreak. In addition to obtaining knowledge of circulating lineages and their relationships to each other and other sequences in Iran and the world, the variant analysis in this study helps to identify important mutations in the generated genome sequences and to compare them with those in other studies.

2. Material and methods

2.1. Collection of samples

This study was approved by the Research Ethics Committee of Alborz University of Medical Sciences (ID Number IR.ABZUMS.REC.1399.169). This study also complies with all regulations and confirmation that informed consent was obtained. Three RNA samples were used in this study. Two samples (AK-SARS-27 and AK-SARS-48) were originated from throat swabs of 57 and 25 year-old men who have attended Alborz province local health center in Chahar-Bagh on 15 June 2020 because of the clinical presentations of the COVID-19 disease and were confirmed as SARS-CoV-2 positive by RT-PCR assay. The third sample (AK-SARS-A7) was originated from a throat swab of a 43 year-old man, who attended Imam Ali Hospital in Karaj, the capital of Alborz province, on 13 May 2020, tested positive by RT-PCR assay and passaged four times in the Vero cell lines in the Razi Vaccine and Serum Research Institute.

2.2. Whole genome sequencing

The viral RNA was subjected to RT-PCR using OneStep Ahead RT-PCR kit (Qiagen, Hilden, Germany) and 26 primer pairs covering the entire genome of SARS-CoV-2 which have been previously described [9]. PCR amplicons of each sample pooled together and purified using High Pure PCR Product Purification kit (Roche, Mannheim, Germany). Library was constructed using TruSeq Nano DNA Kit and the sequencing was performed on an Illumina NovaSeq6000 using the 150 bp paired-end sequencing protocol. Library preparation and sequencing were carried out at Macrogen, INC, (Seoul, South Korea).

2.3. Assembly

Sequence raw data were subjected to quality control using FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc>). Adapter sequences of the reads and bases with a score of less than Q30 were trimmed and any reads shorter than 40 nt removed using Trimmomatic v0.36 [10]. The high-quality paired-end reads were mapped to the reference Wuhan-Hu-1 sequence (GenBank accession number NC_045512.2) using the Burrows-Wheeler Aligner (BWA) package v0.7.17-4 with MEM algorithm and -M -R options [11]. The aligned sequences were converted to a fastq files using the Samtofastq tool available in Picard tools (<http://broadinstitute.github.io/picard>). The tool FLASH [12] was used to stitch the paired-end reads since about one-half of the reads overlap the read generated from the opposite end of the same DNA fragment. The resulting stitched sequences together with the remaining paired-end reads were used as input data for the SPAdes assembler v3.13.0 [13], resulting in three single, positive-stranded RNAs.

2.4. Phylogenetic analysis

The sequence data related to Iran and the countries more likely affecting Iran were downloaded from the GSAID with following filters: complete genomes, high coverage and the collection dates between



Figure 1. Map of Iran with highlighted Alborz province.



1/12/2019 and 1/7/2020 [Armenia 3, Bahrain 35, China 853, Iran 64, Iraq 28, Kuwait 8, Oman 302, Pakistan 39, Qatar 392, Saudi Arabia 736, Turkey 350, United Arab Emirates (UAE) 1355]. A 3 percent of the sequence data was subsampled using a program in the BBMap suite (<https://sourceforge.net/projects/bbmap>) and the assembled sequences in this study and the reference genome (NC_045512.2) were then added to it to build a dataset of 125 sequences. Multiple sequence alignments on the

dataset was performed using MAFFT v7.471 [14] and phylogenetic tree was built by running IQ-Tree using the ultrafast method with 1000 replicates and with the ModelFinder [15], visualized and annotated using FigTree v1.4.4 (<http://tree.bio.ed.ac.uk/software/figtree>). Lineages were assigned using Phylogenetic Assignment of Named Global Outbreak LINeages (Pangolin) COVID-19 Lineage Assigner at <https://pangolin.cog-uk.io> on the basis of previously described methodology [16].

Table 1. Number of SNVs, frequencies (>1%) and depth of the sequenced genomes.

	Total number of SNVs	Number of SNVs with frequency $\geq 1\%$	Average depth (range) x
AK-SARS-27	48	18	26945 (2502–47383)
AK-SARS-48	634	45	41307 (57–75665)
AK-SARS-A7	78	32	34720.82 (1820–74782)

2.5. Variant analysis

The PCR duplicates of the mapped reads (described above) were flagged with Samblaster v0.1.24 [17]. The resultant SAM file was then converted to a BAM file, sorted and indexed using the SAMtools package v1.10 [18]. The BAM file was visualized with Integrative Genomics Viewer (IGV) [19] and its quality was checked. The VCF file containing variants were generated using both FreeBayes v1.3.2 [20] and Lofreq v2.1.5 [21] and reference sequence and were annotated using SnpEff v4.5covid19 [22] available in the Galaxy web platform [23] at usegalaxy.org.

3. Results

3.1. Whole-genome sequencing and assembly

As a result of whole-genome sequencing of samples AK-SARS-27, AK-SARS-48 and AK-SARS-A7, total numbers of 14,260,854, 22,944,020 and 23,134,292 2×151 paired-end reads were generated, respectively. The de novo genome assembly of the aligned reads to the reference genome resulted in three single, positive-stranded RNAs of 29,850 bp, 29,843 bp and 29,828 bp with the mean coverage depths of 32331.2x, 53492.5x and 52873.8x, for the samples AK-SARS-27, AK-SARS-48 and AK-SARS-A7, respectively.

3.2. Phylogenetic analysis

Maximum likelihood phylogenetic analysis showed that the sequences in this study form a distinct cluster with other Iranian sequences (EPI_ISL_2692885, EPI_ISL_437512, EPI_ISL_594185) and other sequences from UAE (EPI_ISL_698182, EPI_ISL_435139, EPI_ISL_469279) and Oman (EPI_ISL_457706) (Figure 2). The cluster (turquoise color in Figure 2) belonged to B.4 lineage according to Pangolin COVID-19 lineage assigner and to O clade in GISAID and it is likely seeded by the Chinese sequence (EPI_ISL_454908). The genome sequences AK-SARS-48 of this study and EPI_ISL_698182 from UAE in the cluster have become descendant lineages B.4.8 and B.4.7, respectively.

3.3. Variant analysis

Total number of single nucleotide variations (SNVs) in the samples AK-SARS-27, AK-SARS-48 and AK-SARS-A7 were 48, 78 and 634,

respectively, and the depth and frequency of SNVs for each sample were different (Table 1).

Seven same SNVs, positioned in ORF1ab, N gene and 3'-untranslated regions of the genome were found in all samples. Two SNVs (1397 G > A and 11083 G > T) of four in ORF1ab corresponded to a missense change with high frequency. The remaining two SNVs, (13129A > G) and (20446A > T) in ORF1ab result in a synonymous change and a stop codon with very low frequency, respectively. There was only one SNV (28688T > C) in the N gene which corresponded to a synonymous change in amino acid with 100% frequency. In the 3'-untranslated region, there were two SNVs 29742 and 29844 with average frequency of 100% and 1% (Table 2).

The SNVs in the spike region of all samples had high frequency. The sample AK-SARS-27 had only one SNV (23403A > G) which corresponded to the missense change D614G. Sample AK-SARS-48 had 4 SNVs. One of them (23525C > T) leads to a missense and the rest (22042T > C, 22708 G > T, 25318C > A) result in synonymous changes. There were 3 SNVs in the Sample AK-SARS-A7. One SNV (22042T > C) was the same as one in AK-SARS-48 and two of them (22350A > V and 23997C > G) corresponded to missense changes. The latter SNVs had lower frequency of 0.83 and 0.51, respectively, in comparison with the frequency of 100% seen in other SNVs in the spike region (Table 3).

4. Discussion

The genome sequences of this study were positioned in a cluster which all belonged to O clade in GISAID. It has been reported the SARS-CoV-2 isolates of O clade are substantial in Iran, Sri Lanka and Malaysia [24]. The cluster included genome sequences from B.4 lineage and its descendant lineages B.4.7 and B.4.8. According to PANGO lineages website (<https://cov-lineages.org/index.html>), B.4 is described as an Iranian lineage which has global exports and B.4.7 as UAE and Africa lineage. B.4 lineage has been reported to be dominant lineage in Iran from start of pandemic to the end of June [25]. This is consistent with our findings since the samples in this study were collected in May and June. One of the sequences in this study had the descendant B.4.8 lineage which shows that the viruses in lineage B.4 might have undergone further evolution. B.4.8 lineage has also occasional international appearances in other countries such as the United Kingdom, Canada, Portugal and Denmark (PANGO lineages website). Overall, it may not be possible to deduce the origin of transmission because of lack of access to more complete genome sequences from Iran and absence of travel history of all the sequences in the cluster.

This study attempted to address genetic diversity of the sequenced genomes in a sub-consensus level as quasispecies within infected patients. The SNVs like other studies [26, 27] were scattered in the ORF1ab, S and N genes of SARS-CoV2 in all samples. The genetic diversity in the sample AK-SARS-48 was much higher than other samples. Apart from recurrent sequence biases, hypermutability and artifacts [27] in the sequenced genomes, other epidemiological variables may play a role in the sequence diversity [26]. The increase in genetic diversity has already been associated with the age and viral load [28, 29]. The reported Ct value of the sample 48 was 2 logs higher than sample 27 (17 vs 19), however, the association between age and genetic diversity was not seen in this study.

Table 2. The SNVs in common in all samples compared to reference genome (GenBank accession number NC_045512.2).

Nucleotide position in reference	Codon in Reference	Amino acid in reference	Alternative codon	Alternative Amino acid	Genome location
1397	GTA	Valine	ATA	Isoleucine	NSP2
11083	TTG	Leucine	TTT	Phenylalanine	NSP6
13129	GGA	Glycine	GGG	Glycine	NSP10
20446	AAA	Lysine	TAA	stop	NSP15
28688	TTG	Leucine	CTG	Leucine	N
29742	NA	NA	NA	NA	3'-UTR
29844	NA	NA	NA	NA	3'-UTR

Table 3. The SNVs in the spike region of the genome of the samples.

Sample ID	Nucleotide position in reference	Codon in Reference	Amino acid in reference	Alternative codon	Alternative Amino acid
AK-SARS-27	23403	GAT	Asparatic acid	GGT	Glycine
AK-SARS-48	22042	TAT	Tyrosine	TAC	Tyrosine
AK-SARS-48	22708	GTG	Valine	GTT	Valine
AK-SARS-48	23525	CAT	Histidine	TAT	Tyrosine
AK-SARS-48	25318	TTC	Serine	TTA	Serine
AK-SARS-A7	22042	TAT	Tyrosine	TAC	Tyrosine
AK-SARS-A7	22350	GCA	Alanine	GTA	Valine
AK-SARS-A7	23997	CCA	Proline	CGA	Arginine

Seven SNVs have been shared in all samples. It has already been reported that viral genomes of all patients who returned to Australia, New Zealand and Canada from Iran in February 2020 form a distinct monophyletic group defined by the three SNVs (1397 G > A, 28688T > C, 29742 G > T) [30]. The fact that these SNVs with high frequency were in common in the sequenced genomes in this study, demonstrates that viral diversity reflected in the early stages of viral transmission in Iran were established next four month. However, this trend may not have continued, as one study in India has reported that the mutational probability of 11083 G > T which was high in the beginning, decreased to zero in August and September 2020, suggesting that this mutation was not selected in the virus population for further propagation because of lesser efficiency in infection or fitness disadvantage [31].

Only one of the sequenced genomes had the SNV 23403A > G which results in D614G amino acid substitution, but it was not accompanied by three other mutations which are always co-occurring in the same genome. Mutation D614G which causes a change in spike region outside of its receptor binding domain became the most predominant form in pandemic from May 2020 globally. There are reports that this mutation is associated with higher infectivity and viral load in the upper respiratory tract [32], but this was not supported by other studies [33]. The small size of SARS-CoV-2 genome sequences from Iran hinders the robust analysis of D614G and other mutations at a national level. However, the available data from limited sequences from Iran in GSAID shows that frequency of mutation D614G increased from 1 (AK-SARS-27 in this study) in June 2020 to 16 sequences in March 2021 (data checked on March 14, 2021).

This study provided the phylogenetic and variant analyses of three complete genome sequences in Alborz province in Iran. One potential limitation is the low number of generated sequences; however, this study will contribute to our understanding of transmission, adaptation and evolution of the virus where the sufficient data is not available. There is an urgent need for more complete genome sequencing of SARS-CoV-2 in Iran and more comprehensive analyses in response to ongoing pandemic.

5. Conclusions

The results of this study revealed that the lineage B.4, as a previously reported dominant lineage in Iran, is still present during the second wave

of COVID-19 in Alborz province and might have undergone further evolution as a descendant B.4.8 lineage. The diversity reported in the early stages of viral transmission became established in the second wave as three mutations in the viral genomes of COVID-19 patients who had travel history to Iran in February 2020 was also found in the viral genome sequences in this study. The mutation D614G which became the most predominant form in pandemic from May 2020 globally was only found in one of the genome sequences of this study.

Declarations

Author contribution statement

Amir Kaffashi: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Jiabin Huang: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Amir Bairami, Mohammad Hossein Fallah Mehrabadi, Somayeh Yaslianifard, S Reza Banihashemi, Fatemeh Soleimanifar, Alireza Soleimani, Farzaneh Moshiri, Sayed-Hamidreza Mozhgani: Contributed reagents, materials, analysis tools or data.

Mohsen Bashashati, Akbar Khorasani: Performed the experiments; Contributed reagents, materials, analysis tools or data.

Mohsen Lotfi, Morteza Taghizadeh: Performed the experiments.

Funding statement

This work was supported under a joint project of Razi Vaccine and Serum Research Institute and Alborz University of Medical Sciences, Iran.

Data availability statement

Data associated with this study has been deposited at Global initiative on sharing all influenza data (GSAID) under the accession numbers EPI_ISL_1398364, EPI_ISL_1398925, EPI_ISL_1398937, and at NCBI Sequence Read Archive (SRA) under the accession numbers PRJNA694865, SAMN17575756, SAMN17575757, SAMN17575758.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

- [1] D. Cucinotta, M. Vanelli, WHO declares COVID-19 a pandemic, *Acta Biomed.* 91 (2020) 157–160.
- [2] P. Venkatesan, COVID-19 in Iran: round 2, *Lancet Infect. Dis.* 20 (2020) 784.
- [3] M.R. Ghadir, A. Ebraheh, J. Khodadadi, et al., The COVID-19 outbreak in Iran; the first patient with a definite diagnosis, *Arch. Iran. Med.* 23 (2020) 503–504.
- [4] Z. Arab-Mazar, R. Sah, A.A. Rabaan, K. Dhama, A.J. Rodriguez-Morales, Mapping the incidence of the COVID-19 hotspot in Iran - implications for travellers, *Trav. Med. Infect. Dis.* 34 (2020) 101630.
- [5] A. Doosti-Irani, E. Mostafavi, M. Nazempour, M.A. Mansournia, A.A. Haghdoust, Challenges for management of the COVID-19 epidemic in Iran, *Glob. Epidemiol.* 2 (2020) 100035.
- [6] B.B. Oude Munnink, D.F. Nieuwenhuijse, M. Stein, et al., Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in The Netherlands, *Nat. Med.* 26 (2020) 1405–1410.
- [7] P. Zhou, X.L. Yang, X.G. Wang, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, *Nature* 579 (2020) 270–273.
- [8] M. Dadar, Y. Fakhri, G. Björklund, Y. Shahali, The association between the incidence of COVID-19 and the distance from the virus epicenter in Iran, *Arch. Virol.* 165 (2020) 2555–2560.
- [9] F. Wu, S. Zhao, B. Yu, et al., A new coronavirus associated with human respiratory disease in China, *Nature* 579 (2020) 265–269.

- [10] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics* 30 (2014) 2114–2120.
- [11] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics* 25 (2009) 1754–1760.
- [12] T. Magoč, S.L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies, *Bioinformatics* 27 (2011) 2957–2963.
- [13] A. Bankevich, S. Nurk, D. Antipov, et al., SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, *J. Comput. Biol.* 19 (2012) 455–477.
- [14] K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Res.* 30 (2002) 3059–3066.
- [15] J. Trifinopoulos, L.T. Nguyen, A. von Haeseler, B.Q. Minh, W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis, *Nucleic Acids Res.* 44 (2016) W232–235.
- [16] A. Rambaut, E.C. Holmes, O'Toole Á, V. Hill, J.T. McCrone, C. Ruis, L. du Plessis, O.G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology, *Nat. Microbiol.* 5 (2020) 1403–1407.
- [17] G.G. Faust, I.M. Hall, SAMBLASTER: fast duplicate marking and structural variant read extraction, *Bioinformatics* 30 (2014) 2503–2505.
- [18] B. Korber, W.M. Fischer, S. Gnanakaran, et al., Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus, *Cell* 182 (2020) 812–827, e819.
- [19] J. Kuipers, A.A. Batavia, K.P. Jablonski, et al., Within-patient Genetic Diversity of SARS-CoV-2, *bioRxiv*, 2020, 2020.2010.2012.335919.
- [20] E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing, 2012.
- [21] A. Wilm, P.P. Aw, D. Bertrand, et al., LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets, *Nucleic Acids Res.* 40 (2012) 11189–11201.
- [22] P. Cingolani, A. Platts, le L. Wang, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D.M. Ruden, 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* 6 (2012) 80–92.
- [23] E. Afgan, D. Baker, B. Batut, et al., The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update, *Nucleic Acids Res.* 46 (2018) W537–w544.
- [24] A. Sengupta, S.S. Hassan, P.P. Choudhury, Clade GR and clade GH isolates of SARS-CoV-2 in Asia show highest amount of SNPs, *Infect. Genet. Evol.* 89 (2021) 104724.
- [25] Z. Fattahi, M. Mohseni, K. Jalalvand, et al., SARS-CoV-2 outbreak in Iran: the dynamics of the epidemic and evidence on two independent introductions, *Transbound Emerg. Dis.* (2021).
- [26] A. Armero, N. Berthet, J.C. Avarre, Intra-Host diversity of SARS-cov-2 should not be neglected: case of the state of Victoria, Australia, *Viruses* 13 (2021).
- [27] Y. Wang, D. Wang, L. Zhang, et al., Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients, *Genome Med.* 13 (2021) 30.
- [28] H.A. Al Khatib, F.M. Benslimane, I.E. Elbashir, P.V. Coyle, M.A. Al Maslamani, A. Al-Khal, A.A. Al Thani, H.M. Yassine, Within-Host diversity of SARS-CoV-2 in COVID-19 patients with variable disease severities, *Front. Cell Infect. Microbiol.* 10 (2020) 575613.
- [29] J.T. Robinson, H. Thorvaldsdóttir, A.M. Wenger, A. Zehir, J.P. Mesirov, Variant review with the integrative genomics viewer, *Cancer Res.* 77 (2017) e31–e34.
- [30] J.S. Eden, R. Rockett, I. Carter, et al., An emergent clade of SARS-CoV-2 linked to returned travellers from Iran, *Virus Evol.* 6 (2020) veaa027.
- [31] B.S. Hudson, V. Kolte, A. Khan, G. Sharma, Dynamic tracking of variant frequencies depicts the evolution of mutation sites amongst SARS-CoV-2 genomes from India, *J. Med. Virol.* 93 (2021) 2534–2537.
- [32] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools, *Bioinformatics* 25 (2009) 2078–2079.
- [33] L. van Dorp, D. Richard, C.C.S. Tan, L.P. Shaw, M. Acman, F. Balloux, No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2, *Nat. Commun.* 11 (2020) 5986.