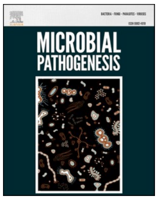
Contents lists available at ScienceDirect

Microbial Pathogenesis

journal homepage: www.elsevier.com/locate/micpath



Variant analysis of SARS-CoV-2 genomes in the Middle East

Khalid Mubarak Bindayna [a,*], Shane Crinion [b]

[a] Department of Microbiology, Immunology, and Infectious Diseases, College of Medicine and Medical Sciences, Arabian Gulf University, Manama, Bahrain
[b] School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, University Road, Galway, Ireland

ARTICLE INFO

Keywords:
Covid 19
Sequencing
Variant

ABSTRACT

Background: Coronavirus (COVID-19) was introduced into society in late 2019 and has now reached over 88 million cases and 1.9 million deaths. The Middle East has a death toll of ~80,000 and over 35000 of these are in Iran, which has over 1.2 million confirmed cases. We expect that Iranian cases caused outbreaks in the neighbouring countries and that variant mapping and phylogenetic analysis can be used to prove this. We also aim to analyse the variants of severe acute respiratory syndrome coronavirus-2 (SARS -CoV-2) to characterise the common genome variants and provide useful data in the global effort to prevent further spread of COVID-19.
Methods: The approach uses bioinformatics approaches including multiple sequence alignment, variant calling and annotation and phylogenetic analysis to identify the genomic variants found in the region. The approach uses 122 samples from the 13 countries of the Middle East sourced from the Global Initiative on Sharing All Influenza Data (GISAID).
Findings: We identified 2200 distinct genome variants including 129 downstream gene variants, 298 frame shift variants, 789 missense variants, 1 start lost, 13 start gained, 1 stop lost, 249 synonymous variants and 720 upstream gene variants. The most common, high impact variants were 10818delTinsG, 2772delCinsC, 14159delCinsC and 2789delAinsA. These high impact variant ultimately results in 36 number of mutations on spike glycoprotein. Variant alignment and phylogenetic tree generation indicates that samples from Iran likely introduced COVID-19 to the rest of the Middle East.
Interpretation: The phylogenetic and variant analysis provides unique insight into mutation types in genomes. Initial introduction of COVID-19 was most likely due to Iranian transmission. Some countries show evidence of novel mutations and unique strains. Increased time in small populations is likely to contribute to more unique genomes. This study provides more in depth analysis of the variants affecting in the region than any other study.

1. Introduction

On January 9h 2020, the China Centre for Disease Control reported that 15 of 59 suspected cases of pneumonia were due to a novel human coronavirus (CoV), now known as Severe Acute Respiratory Syndrome CoV 2 (SARS-CoV-2) [1]. The genome for this novel virus was then made publicly available on the *Global Initiative on Sharing All Influenza Data (GISAID)* the next day. SARS-CoV-2 is an easily spreadable virus which would evolve into a global pandemic of at least 88 million cases and 1.9 million deaths [2]. One of the first countries to experience a significant outbreak was Iran. The country reported its first confirmed case on 19th February 2020 from a merchant in Qom who travelled from China [3]·. Many of the first countries with infections in the Middle East were linked to travellers from Iran including Lebanon, Kuwait, Bahrain, Iraq, Oman and UAE. COVID-19 continued to spread to the remaining Middle Eastern countries with a death toll of over 50,000 people according to health authorities. This number is expected to be an underestimation due to countries effected by war including Libya, Syria and Yemen. Needless to say, there have been devastating effects to the region and the real effects are expected to be unreported [4].

Researchers are racing to develop a vaccine that can provide viral immunity and avoid additional deaths. SAR-CoV-2 is transmitted using the spike protein which binds to human angiotensin-converting enzyme 2 (ACE2) receptor; the virus is easily transmittable due to mutations in the receptor-binding (S1) and fusion (S2) domain of the strain [5]. Transmission could be made even easier if more mutations accumulate. Although mutations are rare, they can create new strains and it is not guaranteed that the current leading vaccine trials will be effective as SARS-CoV-2 continues to mutate [6]. By categorizing variants, we can identify any new strains and how the mutations are likely to affect

[*] Corresponding author.
E-mail addresses: bindayna@agu.edu.bh (K.M. Bindayna), s.crinion1@nuigalway.ie (S. Crinion).

https://doi.org/10.1016/j.micpath.2021.104741
Received 13 October 2020; Received in revised form 8 January 2021; Accepted 8 January 2021
Available online 13 February 2021

spread. As the Middle East is often under reported, it is important to characterise the variants of strains that are commonly present. Analysis of the common variants in the Middle East is essential to develop a vaccine that treats the strains in the region. This analysis helps understand the viral genome landscape and identify clades of the region.

2. Objectives

- Our hypothesis is that variants found in SARS-CoV-2 genomes from Middle Eastern samples will indicate delivery from Iran. We will use bioinformatics tools and publicly available samples to explore the composition of strains within each country. We expect that many strains will show evidence of Iranian origin.
- The aim is to explore the structure of Middle Eastern genome strains using multiple sequence alignment, tree generation and variant prediction (and others). If we explore the structure and common variants of SARS-CoV-2 strains in these populations, we expect to learn more about how the virus spread.

3. Methods

Sample Source: We obtained the publicly available data from the Global Initiative on Sharing All Influenza Data (GISAID) [7].

Sample Size: 122 Middle Eastern samples, Wuhan reference sequence NC_045512 and 5 recent Wuhan samples.

Sample Selection: Samples were selected from the Middle East by using filtering available on the GISAID website. Only complete genome samples were used. The countries considered were Afghanistan, Bahrain, Cyprus, Egypt, Iraq, Iran, Israel, Jordan, Kuwait, Lebanon, Libya, Oman, Qatar, Saudi Arabia, Sudan, Syria, Turkey, United Arab Emirates (UAE) and Yemen. Iran had only 7 samples available after filtering for low coverage. Cyprus, Kuwait, Lebanon all had 8 samples available after filtering. No samples were available on the database from Afghanistan, Iraq, Libya, Sudan, Syria or Yemen. 10 samples were taken from all other countries. Samples were also filtered to high quality when possible. 10 samples was selected as the optimum number to cover all possible countries and remain within alignment file limit of size 4 Mb (maximum size for Clustal Omega tool). In countries with 10 samples, to prevent sample sourcing from same outbreak, the 5 earliest and 5 most recent samples were taken. All samples were downloaded from GISAID and then concatenated into a single multi-sample file and saved in FASTA format.

Multiple sequence alignment: Using the collected samples, multiple sequence alignment (MSA) was performed using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) [8]. The Clustal Omega online tool was used to perform the alignment (found at: https://www.ebi.ac.uk/Tools/msa/clustalo/). The online tools allows up to 4000 sequences or a maximum file size of 4 MB, therefore the maximum number of samples was used. The concatenated file of samples was uploaded to the online tool. For step 2, the output parameters selected were PEARSON/FASTA. All other options were kept at the default option. The output file generated is an alignment file; the file consists of all sequences with gaps denoted by '-'. The output file format is also a FASTA file.

Variant identification: Variant calling was performed using the alignment FASTA file and the SNP extraction tool snp-sites [9] (https://github.com/sanger-pathogens/snp-sites). These tools identify the SNP sites by taking a multi-sample FASTA file as input. The program then restructures the data as a variant call format (VCF) file. The VCF file provides a clear mapping of SNPs from the aligned sequences – this allows us to easily identify the SNP location and the genotype for each sample at a given locus. In the outputted VCF file, the rows correspond with each unique variant and the column provides the genotype at the given site.

Variant annotation: SNP-eff [10] has been used to perform the variant annotation information such as the variant definition and the overlapping gene (https://pcingola.github.io/SnpEff/SnpEff.html). SnpEff could also predict the effect of the variants. SNPeff is integrated into the Galaxy web-based tool for bioinformatics analysis (usegalaxy.org) [11]. VCF file is uploaded in the Galaxy platform. SnpEff database has been created by downloading the Wuhan reference NC_045512.2 from NCBI. The variants annotation have been performed by the "SnpEff build". The custom parameters include the setup of 5000 bases for Upstream/Downstream length and 2 bases as set size for splice sites (donors and acceptor) in bases. Once the analysis was executed, the annotation data is outputted as an annotated VCF and a HTML report file. We have also analyzed the mutations using CoVsurver in the GISAID database [7].

Data visualisation: Once the annotated VCF was generated, the VCF was imported to R for extraction of the variant annotation information. The annotated data was imported, manipulated and plotted using R v3.6.2 [12]. The dplyr v0.8.4 package was used to summarise and align the data [13]. The ggplot2 package was used to align the identified variants and visualise the types of mutations that re-occur [14]. Variant position along the SARS-CoV-2 genome is indicated in the X axis. The y-axis indicates the sample name and the right y-axis represents the country of origin for each sample. This plot is used to compare the genome in different populations.

Phylogenetic analysis: Phylogenetic analysis was performed using BEAST (Bayesian Evolutionary Analysis Sample Trees) v1.10.4, to perform Bayesian analysis of molecular sequences using Monte Carlo Markov Chains (MCMC) [15]. The analysis followed the approach recommended to reconstruct the evolutionary dynamics of an epidemic. The aim of this is to obtain an estimate of the origin of the epidemic in the region and understand how it spread through the Middle East. To undertake the analysis, we opened BEAUTi, the graphical application used to analyse the control file. Although it requests a NEXUS file, the FASTA file can also be used. The data was uploaded using the Import Data option and appeared under the Partitions section. BEAUTi confirmed that 30851 sites are present in the uploaded data. The default options are selected for site model and clock model. Next, we specified the individual virus dates by selecting the "Tips" panel and selecting the "Use tip dates" option. A tab delimited file was uploaded which specified the upload date. This information was extracted from the names as they were downloaded from GISAID. Next, we set the substitution model by selecting the "Sites" tab and selected the default options of HKY model, the default Estimated base frequencies and select Gamma as Site Heterogeneity Model. Next, the molecular clock was selected under the "Clock" tab as a strict clock since we know that the frequency of mutation is low. The tree options are elected under the "Tree Prior" tab as "Random starting tree" for the tree model and "Coalescent: Exponential Growth", a model that assumes a finite but constant population size and predicts that all alleles will be removed from the population individually. This provides additional predictions on the reproductive rate. In the "Priors" tab, select the scale as 100 for prior distribution which models the expected growth for a pandemic. The operators require no changes from the default. The MCMC option for chain length is set as 100,000 and sampling frequency to 100.

Tree visualisation: Finally, we summarised the tree using the TreeAnnotator tool, an additional package as part of BEAST. We first select the file generated using BEAST and outputted the tree file. Then the output NEXUS file was imported to FigTree program to display. Once we opened FigTree to display, we re-ordered the order by

Increasing value and then switched on Branch Labels. We switched on Node Bars and selected the 95% highest posterior density (HPD) credible intervals for the node heights. We plotted a time scale by turning on the Scale Axis and then setting the Time Scale section for Offset as 2020.7, the latest date of collection for our samples.

4. Results

Sequence alignment and variant calling were completed successfully.

Once these were complete, variation annotation was performed. We identified 2200 distinct genome variants which are recorded in Table 1. The most common, high impact variants were 10818delTinsG, 2772delCinsC, 14159delCinsC and 2789delAinsA. The frequency of each unique variant type can be found in Table 2, which outlines the locus of all SNPs with over 50 instances.

CoVsurver [7] analysis indicated that these strains have overall 2.828% mutations on spike glycoproteins. Total number of 36 mutations have been identified. Strain EPI_ISL_507007, isolated from Iran, has maximum number of unique mutations. These mutations include NSP1_W161L, NSP1_V116A, NSP1_V106G, NSP1_V54A, NSP2_I393T, NSP2_N92S, NSP2_T153L, NSP2_L506S, NSP2_E309A, NSP2_A159D, NSP2_C136S, NSP3_K1838R, NSP3_R1341L, NSP3_K1596R, NSP3_R1345L, NSP3_A886D, NSP3_M1865T, NSP3_I468T, NSP3_K140Q, NSP3_A1279D, NSP3_A1105G, NSP3_G1389D, NSP3_N1778S, NSP3_L781S, NSP3_G1944V, NSP3_L1523H, NSP3_K1715R, NSP3_C296R, NSP3_R558P, NSP3_V1673D, NSP3_E378V, NSP3_S674Y, NSP3_S721 N, NSP4_P274L, NSP4_L321P, NSP4_A48D, NSP4_L243P, NSP4_L329H, NSP4_L176Q, NSP4_Q488L, NSP4_P168Q, NSP5_G174V, NSP5_E166V, NSP5_N84S, NSP5_S62Y, NSP5_P9L, NSP5_T111 N, NSP5_P52L, NSP6_C221Y, NSP6_A119G, NSP6_W140L, NSP6_S53Y, NSP6_C68Y, NSP6_F70Y, NSP6_F184S, NSP6_H64 N, NSP6_G188D, NSP6_V101G, NSP6_G48D, NSP6_F220S, NSP6_P87L, NSP7_L13S, NSP8_D134E, NSP9_P57H, NSP10_V119A, NSP10_E135A, NSP10_R134H, NSP11_Q5P, NSP12_E144D, NSP13_D160E, Spike_G1246A, Spike_R1185H, Spike_L368P, Spike_S974P, Spike_G268D, Spike_R190S, Spike_A411D, Spike_G798A, Spike_A672D, Spike_V1230E, Spike_Q774R, Spike_H146R, Spike_P337R, Spike_Q607L, E_I33T, M_P59T, M_K14E, NS7a_F114I, NS8_V62 M, N_Q390L, N_P20H, N_R10Q, N_R149L.

Other than that EPI_ISL_514753, isolated from Iran, also have goof nuber of unique mutations like NSP1_W161L, NSP2_E309A, NSP2_C136S, NSP3_R1341L, NSP3_K232T, NSP3_N1778S, NSP3_E229A, NSP3_C296R, NSP6_H64 N, Spike_D808G, Spike_H146R, Spike_N1192S. But only on the basis of spike proteins EPI_ISL_427420 strain which is isolated from Qatar has highest number of mutations on the spike protein. On the other hand EPI_ISL_514306, isolated from Israel has highest number of known mutantion on overall protain structure. Table 3 has summerized top 5 strains that have maximum number of known mutations. These strains have been isolated from Israel(3), United Arab Emirates(1), Bahrain(1), Iran(1) and Egypt(1). A detailed account of these mutations are reported in the supplementary information 1.

Strains	Known mutations
EPI_ISL_463740	NSP1_D75E, NSP3_P153L, NSP8_M129I, NSP14_F233L, NS8_V62L, NS8_L84S, N_A208G, N_T393I
EPI_ISL_486889	NSP3_P109L, NSP3_A994D, NSP4_S481L, NSP12_P323L, NSP14_P412S, Spike_D614G, N_G204R, N_R203K
EPI_ISL_507007	NSP2_A361V, NSP3_T1482I, NSP4_T73I, NSP8_R51H, NSP12_P323L, Spike_D614G, Spike_S1147L, NS3_Q57H, N_S194L
EPI_ISL_514303	NSP2_T85I, NSP6_L37F, NSP12_P323L, NSP16_P236S, Spike_D614G, Spike_T95I, NS3_Q57H, NS3_G44V, N_G25C
EPI_ISL_514305	NSP12_M666I, NSP12_P323L, NSP14_E204D, Spike_D614G, NS3_W131C, NS3_K75 N, N_S193I, N_G204R, N_R203K
EPI_ISL_514306	NSP2_T85I, NSP7_S25L, NSP12_M666I, NSP12_P323L, NSP14_A320V, Spike_D614G, NS3_Q57H, NS3_K75 N, N_S193I, N_G243C, N_G204R, N_R203K
EPI_ISL_479733	NSP3_T428I, NSP5_G15S, NSP8_T148I, NSP12_P323L, Spike_D614G, Spike_Q677H, N_G212V, N_R203K

(continued on next column)

(continued)

Table 1
The frequency of each type of mutation found in the data.

Annotation	Count
Downstream gene variants	129
Frame shift variants	298
Missense variants	789
Start lost	1
Start gained	13
Stop lost	1
Synonymous variant	249
Upstream gene variant	720

Table 2
The frequency of mutations at each locus with >50 hits.

Position	Count
241	74
3037	74
24351	74
11083	73
14408	72
8	64
21	59
22	59
23	57

The results on SNP analysis were then used to generate the dotplot of variant alignment (Fig. 1). The dotplot successfully indicated a pattern in variants that could not be easily identified from the alignment or annotation files. The alignment includes samples in facets based on their country. The alignment shows a pattern in variants that occur between each country. For example, this is prominent in Qatar, Jordan and Oman where the pattern makes the country distinctive from the variants plotted for other countries. In addition, the phylogenetic tree generated branching indicative of an Iranian origin (Fig. 2). All variants are in relation to Wuhan reference sequence NC_045512.2. The Wuhan samples are in the top facet and show low mutation frequency in comparison to the reference. The following samples show a greater accumulation of mutations. One Iranian sample has many more variants than orders. The Saudi Arabia also appears to have a low mutation rate which may be due to their early contraction of the virus. Oman samples also indicate evidence of a distinctive strain. Many cases feature a missense variant at 3037, 14408 and 24351 base pairs.

5. Discussion

The aim of this study was to identify whether COVID-19 was introduced to the Middle East from Iran and also to explore the genomic composition in the region.

Our study performs sequence alignment to compare all sequences against the reference genome. Once this is complete, the annotated variants were extracted to generate a plot mapping variants, grouping samples by country. The plot as seen in Fig. 1, shows clear distinctive patterns within countries that are not obvious from the generated alignment and annotation files. The variants found in different regions, ordered by their first reported case, are from United Arab Emirates (UAE), Egypt, Iran, Israel, Lebanon, Bahrain, Kuwait, Oman, Qatar, Jordan, Saudi Arabia, Turkey and Cyprus. Global travel plays an important role in the spreading of SARS-Cov2 in middle east where Dubai in the United Arab Emirates acts as a travel hub, as reported recently [23]. The variants at position 241, 3037, 24351, 11083 appear in many Middle Eastern countries but do not occur in Wuhan samples. This variant characterization may be useful in the fight against COVID-19 and the development of treatments. Identifying unique variants to a region may explain why treatment is working for some and not others, should the mutations have an effect on the delivery or the severity of the virus. On the basis of SNP analysis 10818delTinsG, 2772delCinsC, 14159delCinsC and 2789delAinsA are identified as high impact variants.

The Iranian samples appear more diverse and interestingly do not

Table 3
Catalogue of sample accession ID by country.

	Region	ID
1	Bahrain	EPI_ISL_487274
2	Bahrain	EPI_ISL_486889
3	Bahrain	EPI_ISL_483545
4	Bahrain	EPI_ISL_510528
5	Bahrain	EPI_ISL_483542
6	Bahrain	EPI_ISL_483548
7	Bahrain	EPI_ISL_483547
8	Bahrain	EPI_ISL_483543
9	Bahrain	EPI_ISL_485401
10	Bahrain	EPI_ISL_487273
11	Cyprus	EPI_ISL_463742
12	Cyprus	EPI_ISL_463743
13	Cyprus	EPI_ISL_463744
14	Cyprus	EPI_ISL_463748
15	Cyprus	EPI_ISL_463741
16	Cyprus	EPI_ISL_463745
17	Cyprus	EPI_ISL_463746
18	Cyprus	EPI_ISL_463747
19	Egypt	EPI_ISL_482761
20	Egypt	EPI_ISL_479735
21	Egypt	EPI_ISL_479733
22	Egypt	EPI_ISL_479734
23	Egypt	EPI_ISL_510532
24	Egypt	EPI_ISL_430819
25	Egypt	EPI_ISL_430820
26	Egypt	EPI_ISL_479732
27	Egypt	EPI_ISL_482759
28	Egypt	EPI_ISL_482760
29	Iran	EPI_ISL_424349
30	Iran	EPI_ISL_445088
31	Iran	EPI_ISL_507007
32	Iran	EPI_ISL_514753
33	Iran	EPI_ISL_437512
34	Iran	EPI_ISL_442044
35	Iran	EPI_ISL_442523
36	Israel	EPI_ISL_435291
37	Israel	EPI_ISL_514303
38	Israel	EPI_ISL_514305
39	Israel	EPI_ISL_514302
40	Israel	EPI_ISL_435286
41	Israel	EPI_ISL_435289
42	Israel	EPI_ISL_419211
43	Israel	EPI_ISL_435284
44	Israel	EPI_ISL_514306
45	Israel	EPI_ISL_514301
46	Jordan	EPI_ISL_430012
47	Jordan	EPI_ISL_430002
48	Jordan	EPI_ISL_430003
49	Jordan	EPI_ISL_430009
50	Jordan	EPI_ISL_429993
51	Jordan	EPI_ISL_450188
52	Jordan	EPI_ISL_434516
53	Jordan	EPI_ISL_429997
54	Jordan	EPI_ISL_450186
55	Jordan	EPI_ISL_450187
56	Kuwait	EPI_ISL_421652
57	Kuwait	EPI_ISL_422427
58	Kuwait	EPI_ISL_416543
59	Kuwait	EPI_ISL_416458
60	Kuwait	EPI_ISL_416541
61	Kuwait	EPI_ISL_416542
62	Kuwait	EPI_ISL_422426
63	Kuwait	EPI_ISL_422424
64	Lebanon	EPI_ISL_498551
65	Lebanon	EPI_ISL_498552
66	Lebanon	EPI_ISL_498554
67	Lebanon	EPI_ISL_450512
68	Lebanon	EPI_ISL_450515
69	Lebanon	EPI_ISL_450511
70	Lebanon	EPI_ISL_450508
71	Lebanon	EPI_ISL_450509
72	Oman	EPI_ISL_492023
73	Oman	EPI_ISL_492026
74	Oman	EPI_ISL_492024
75	Oman	EPI_ISL_492065
76	Oman	EPI_ISL_492025
77	Oman	EPI_ISL_457706
78	Oman	EPI_ISL_457704
79	Oman	EPI_ISL_457937
80	Oman	EPI_ISL_457701
81	Oman	EPI_ISL_457974
82	Qatar	EPI_ISL_427404
83	Qatar	EPI_ISL_427420
84	Qatar	EPI_ISL_427407
85	Qatar	EPI_ISL_427406
86	Qatar	EPI_ISL_427419
87	Qatar	EPI_ISL_427418
88	Qatar	EPI_ISL_427405
89	Qatar	EPI_ISL_427408
90	Qatar	EPI_ISL_427417
91	Qatar	EPI_ISL_427416
92	SaudiArabia	EPI_ISL_512924
93	SaudiArabia	EPI_ISL_512926
94	SaudiArabia	EPI_ISL_489996
95	SaudiArabia	EPI_ISL_489998
96	SaudiArabia	EPI_ISL_490000
97	SaudiArabia	EPI_ISL_489999
98	SaudiArabia	EPI_ISL_489997
99	SaudiArabia	EPI_ISL_512927
100	SaudiArabia	EPI_ISL_512922
101	SaudiArabia	EPI_ISL_512923
102	Turkey	EPI_ISL_495421
103	Turkey	EPI_ISL_495445
104	Turkey	EPI_ISL_495433
105	Turkey	EPI_ISL_429868
106	Turkey	EPI_ISL_429867
107	Turkey	EPI_ISL_429866
108	Turkey	EPI_ISL_428712
109	Turkey	EPI_ISL_495436
110	Turkey	EPI_ISL_495429
111	Turkey	EPI_ISL_424366
112	UnitedArabEmirates	EPI_ISL_469277
113	UnitedArabEmirates	EPI_ISL_469279
114	UnitedArabEmirates	EPI_ISL_435126
115	UnitedArabEmirates	EPI_ISL_435121
116	UnitedArabEmirates	EPI_ISL_435131
117	UnitedArabEmirates	EPI_ISL_463740
118	UnitedArabEmirates	EPI_ISL_469281
119	UnitedArabEmirates	EPI_ISL_469280
120	UnitedArabEmirates	EPI_ISL_435137
121	UnitedArabEmirates	EPI_ISL_435134
122	Wuhan	EPI_ISL_402123
123	Wuhan	EPI_ISL_454949
124	Wuhan	EPI_ISL_454948
125	Wuhan	EPI_ISL_406798
126	Wuhan	EPI_ISL_454951
127	Wuhan	EPI_ISL_403931

share the mutation at 14408 which is evident in most samples [18]. Iranian sample EPI_ISL_507007 appears to have a high frequency of variants that is not seen in others. THE GISAID detection system indicates no faults were found with sample EPI_ISL_507007 and report a full sequence match so it was not removed. It is noteworthy that the Iran samples appear to have a lower average SNP frequency than other countries. This may indicate that the virus transmitted to Iran earlier than other countries, as we expected. Israel, Bahrain, Kuwait, Oman, Saudi Arabia, Jordan and Turkey all share similar variant mapping. Qatar shows an usual mapping with a high frequency of frameshift variants. This may indicate a diverse, new strain is circulating in the country. Cyprus has little diversity in the variant mapping which is surprising given its late date for first reported cases. CoVsurver analysis indicates various unique and known mutations have been identified in most of the samples from middle east in terms of amino acids change. Iran strain EPI_ISL_507007 has maximum number of unique mutations. Where as Israel isolate EPI_ISL_514306 has highest number of known mutant. Similarly, on the basis of mutations on the spike glycoprotein,

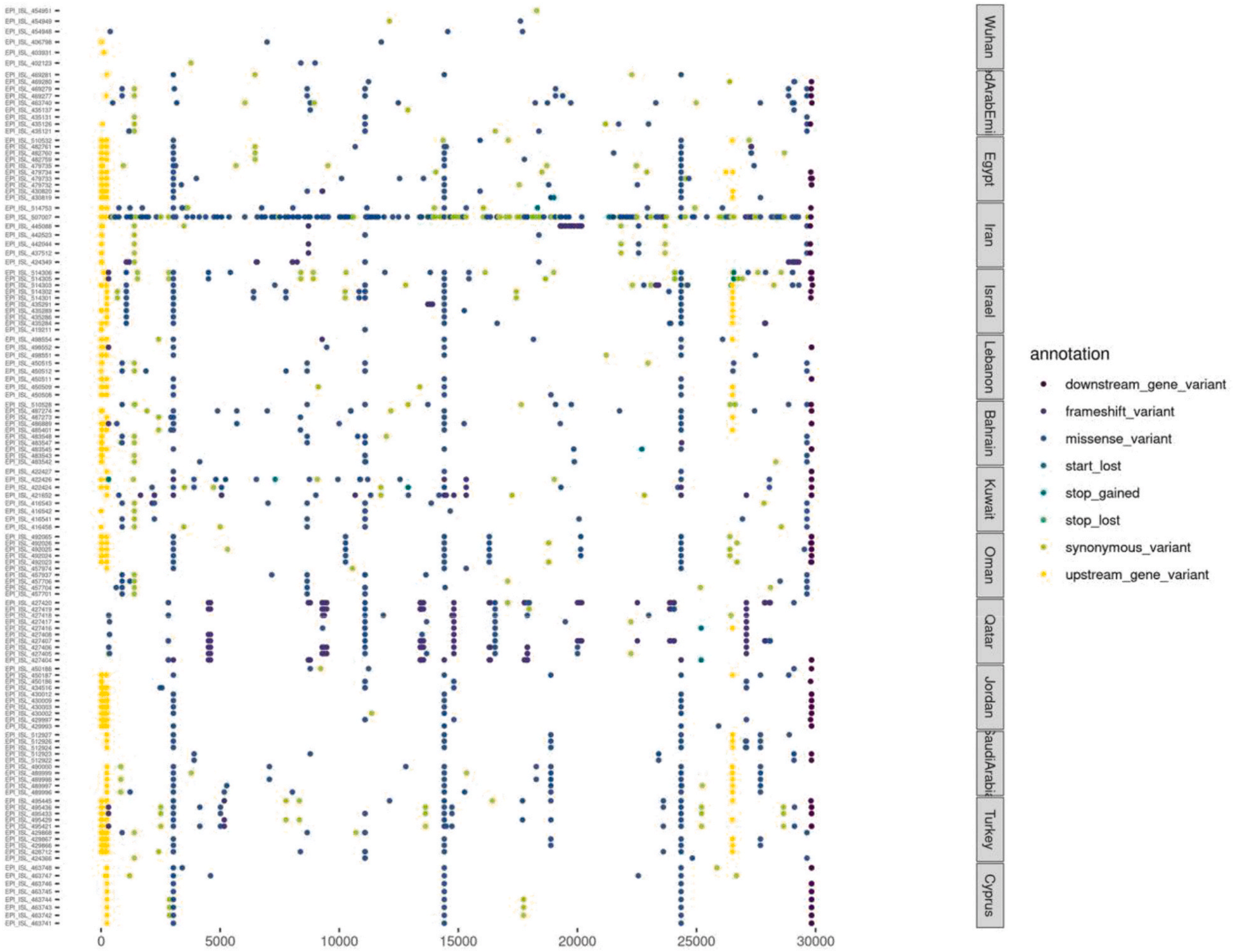


Fig. 1. Dotplot of variants per sample by country. All samples included in the above plot are grouped by country of origin. The left x-axis denotes the accession ID and the right x-axis denotes the country of origin. The y-axis is the position along the SARS-CoV-2 genome. The order of countries is based on date of first reported of COVID -19.

Quatar strain EPI_ISL_427420 has highest number of mutations.

Another interesting point is that time-varied samples were taken for countries with 10 samples. We see not indication that there are distinct groups within countries. This further indicates the mutation frequency is low. It also indicates that there is more variation in the genomic composition in samples from different countries than differences found in samples from different collection times. Smaller populations can cause greater accumulation of variants through genetic drift. This may occur given local lockdowns and travel restrictions that have been enforced worldwide which requires in depth planning [20]. It is possible that these genomic strains with new mutations may create a situation where the countries develop a deathly strain that is not prominent in other parts of the world. This could result in a situation where a country is disproportionately affected by accumulating deaths or an inefficient vaccine.

Phylogenetic trees help in understanding the evolutionary relationships between groups. In the present context, we use them to identify the earliest strains and to track the spread of COVID-19 across the Middle East. The tree shows that UAE samples are distinguished and form one clade. This correlates with their early intervention and lockdown and subsequently appears to have resulted in a unique genome. Gómez-Carballa et al. Indicates the effect of world wide lockdown on SARS-CoV2 genome variations in presence of super spreaders [21,22].

Samples from Qatar also form the majority of 1 clade, with many of the Wuhan samples, indicating that they are similar to the Wuhan samples and show little distinction. Egypt also becomes a distinct branch earlier than most samples. These examples are indicative of the global response – the lockdown of each country and prevention of spreading has resulted in SARS-CoV-2 strains of great similarity within each country. If lockdowns were not enforced, it is likely that these clades would be less distinguisable as mutations are spread between countries. Though the lockdowns has other effects on various socio-econimical aspects [17,19]. As we expected, 2 of the highest branches points attach to Iranian samples, further implementing Iran in the initial spread across the Middle East. The phylogenetic tree therefore indicates what we suggest in our hypothesis – most samples originate from the Iranian sample. This is not surprising given the vast number of cases and early crisis state of the country. However, it is useful to see that the variant analysis shows what we suspect at the genome level. A related study also came to this conclusion by using contact tracing from cases related to religious events in the city of Qom, Iran [16].

CRediT authorship contribution statement

Khalid Mubarak Bindayna: Conceptualization, Data curation, Curation, Writing - original draft, Writing - review & editing, Visualization. **Shane Crinion:** Methodology, Software, Writing - original draft, and sharing writing a draft.

Declaration of competing interest

The authors declare that they have no known competing financial

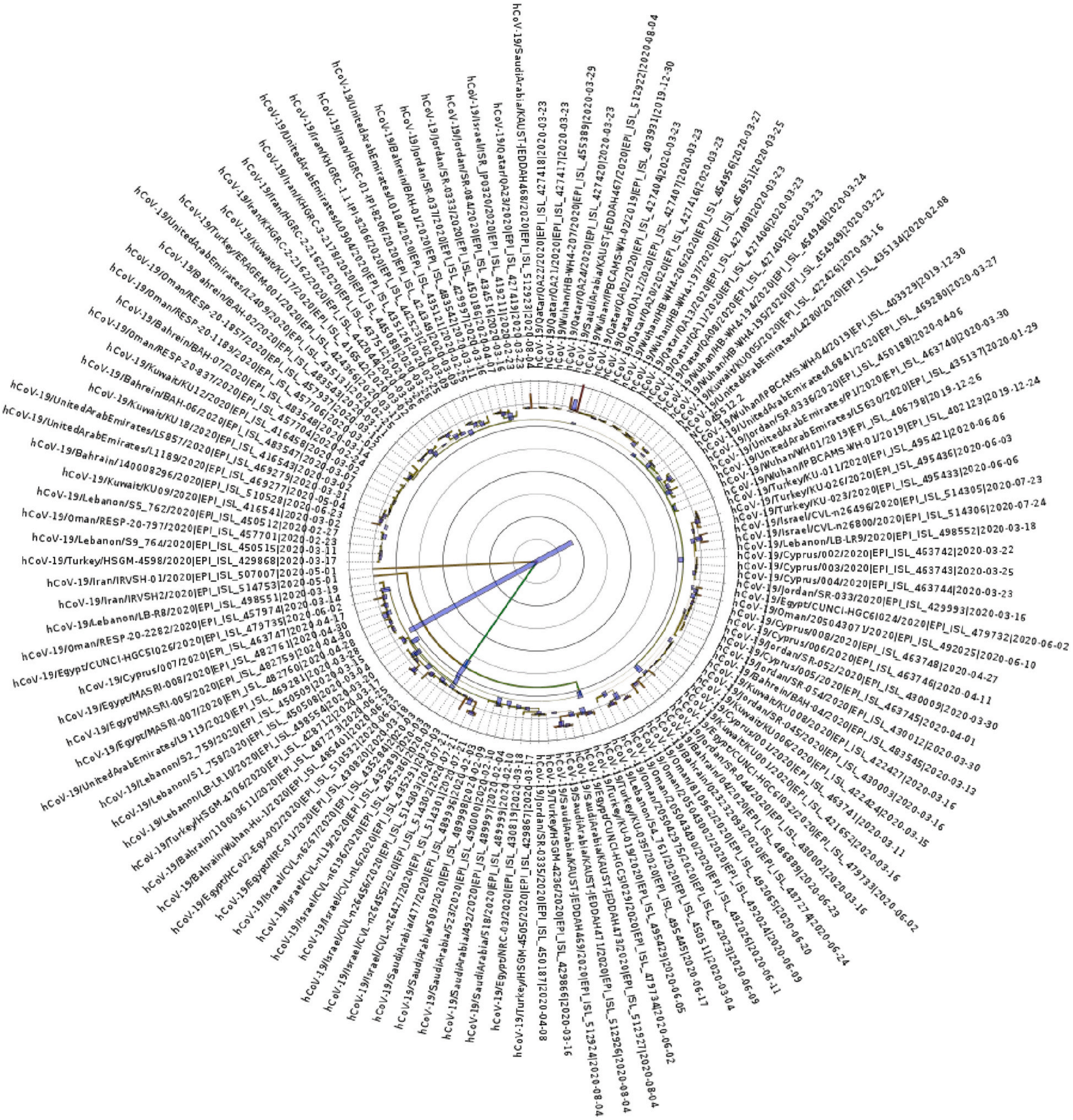

hCoV-19/Lebanon/LB-LR9/2020|EPI_ISL_498552|2020-03-18
hCoV-19/Cyprus/002/2020|EPI_ISL_463742|2020-03-22
hCoV-19/Cyprus/003/2020|EPI_ISL_463743|2020-03-25
hCoV-19/Cyprus/004/2020|EPI_ISL_463744|2020-03-23
hCoV-19/Jordan/SR-033/2020|EPI_ISL_429993|2020-03-16
hCoV-19/Lebanon/S9_764/2020|EPI_ISL_457701|2020-02-23
hCoV-19/Turkey/HSGM-4598/2020|EPI_ISL_429868|2020-03-17
hCoV-19/Iran/IRVSH-01/2020|EPI_ISL_507007|2020-05-01
hCoV-19/Iran/IRVSH2/2020|EPI_ISL_514753|2020-05-01

Fig. 2. Phylogenetic tree for the SARS-CoV-2 genomes. Contains all 128. samples (including reference NC_045512). The colours represent the greatest height in association. The blue lines represent the 95% HDP for each region. The highest branching point is with Iran samples EPI_ISL_507007 and EPI_ISL_514753 with length of 4.319 and 0.8752 respectively.

interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: We have no conflict of interest. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgements

The authors are grateful for the timely sequencing and release of genomes to make this study possible and to Dr. Anusha C P for her comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.micpath.2021.104741.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] European Centre for Disease Prevention and Control (ECDC), Risk Assessment: Outbreak of Acute Respiratory Syndrome Associated with a Novel Coronavirus,

ECDC, Stockholm, 2020. Wuhan, China; first update 2020 [updated 22 January 2020].

[2] COVID-19 situation reports [Internet]. [cited 2020 September 4]. Available from, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

[3] Mapping the incidence of the COVID-19 hotspot in Iran – implications for travellers, Trav. Med. Infect. Dis. 34 (2020 March-April) 101630, https://doi.org/10.1016/j.tmaid.2020.101630. Published online 2020 Mar 14.

[4] G.N. Dhabaan, W.A. Al-Soneidar, N.N. Al-Hebshi, Challenges to testing COVID-19 in conflict zones: Yemen as an example, J Glob Health 10 (1) (2020), 010375, https://doi.org/10.7189/jogh.10.010375.

[5] J.A. Jaimes, N.M. André, J.S. Chappie, J.K. Millet, G.R. Whittaker, Phylogenetic analysis and structural modeling of SARS-CoV-2 spike protein reveals an evolutionary distinct and proteolytically sensitive activation loop, J. Mol. Biol. 432 (10) (2020 May 1) 3309–3325.

[6] X. Tang, C. Wu, X. Li, Y. Song, X. Yao, X. Wu, et al., On the origin and continuing evolution of SARS-CoV-2 [Internet]. [cited 2020 Jun 9]. Available from, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/.

[7] Y. Shu, J. McCauley, GISAID: global initiative on sharing all influenza data–from vision to reality, Euro Surveill. 22 (13) (2017) 30494.

[8] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, D.G. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol. Syst. Biol. 7 (2011) 539, https://doi.org/10.1038/msb.2011.75.

[9] SNP-sites, Rapid efficient extraction of SNPs from multi-FASTA alignments, Microb. Genom. 2 (4) (2016). Andrew J. Page, Ben Taylor, Aidan J. Delaney, Jorge Soares, Torsten Seemann, Jacqueline A. Keane, Simon R. Harris.

[10] P. Cingolani, A. Platts, L.L. Wang, M. Coon, T. Nguyen, L. Wang, et al., A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3, Fly (Austin) 6 (2) (2012) 80–92.

[11] E. Afgan, D. Baker, B. Bér, B. Batut, M. Van Den Beek, D. Bouvier, et al., The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update [Internet], Nucleic Acids Res. 2;46 (W1) (2018) W537–W544. Available from, https://galaxyproject.org.

[12] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013. URL, http://www.R-project.org/.

[13] Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, Dplyr: a grammar of data manipulation, R package version v 0.8.4, https://CRAN.R-project.org/package=dplyr, 2018.

[14] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2016, ISBN 978-3-319-24277-4. https://ggplot2.tidyverse.org.

[15] M.A. Suchard, P. Lemey, G. Baele, D.L. Ayres, A.J. Drummond, A. Rambaut, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10, Virus Evolution 4 (2018) vey016, https://doi.org/10.1093/ve/vey016.

[16] N. Al-Rousan, H. Al-Najjar, Is visiting Qom spread CoVID-19 epidemic in the Middle East? Eur. Rev. Med. Pharmacol. Sci. 24 (2020) https://doi.org/10.26355/eurrev_202005_21376, 10 5813-5818.

[17] L. Cheikh Ismail, T.M. Osaili, M.N. Mohamad, A. Al Marzouqi, A.H. Jarrar, D. O. Abu Jamous, A.S. Al Dhaheri, Eating habits and lifestyle during COVID-19 lockdown in the United Arab Emirates: a cross-sectional study, Nutrients 12 (11) (2020) 3314.

[18] R. Salimi, R. Gomar, B. Heshmati, The COVID-19 outbreak in Iran, J. Global Health 10 (1) (2020).

[19] A.S. Alqasemi, M.E. Hereher, G. Kaplan, A.M.F. Al-Quraishi, H. Saibi, Impact of COVID-19 lockdown upon the air quality and surface urban heat island intensity over the United Arab Emirates, Sci. Total Environ. (2020) 144330.

[20] F.E. Alvarez, D. Argente, F. Lippi, *A Simple Planning Problem for Covid-19 Lockdown* (No. W26981), National Bureau of Economic Research, 2020.

[21] A. Gómez-Carballa, X. Bello, J. Pardo-Seco, F. Martinón-Torres, A. Salas, Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders, Genome Res. 30 (10) (2020) 1434–1448.

[22] M. Pachetti, B. Marini, F. Giudici, F. Benedetti, S. Angeletti, M. Ciccozzi, D. Zella, Impact of lockdown on Covid-19 case fatality rate and viral mutations spread in 7 countries in Europe and North America, J. Transl. Med. 18 (1) (2020) 1–7.

[23] A. Abou Tayoun, T. Loney, H. Khansaheb, S. Ramaswamy, D. Harilal, Z.O. Deesi, A. Alsheikh-Ali, Multiple early introductions of SARS-CoV-2 into a global travel hub in the Middle East, Sci. Rep. 10 (1) (2020) 1–7.