Genomic characterization and phylogenetic analysis of the first SARS-CoV-2 variants introduced in Lebanon

Rita Feghali, Georgi Merhi, Aurelia Kwasiborski, Véronique Hourdel, Nada Ghosn, Sima Tokajian

Genomic characterization and phylogenetic analysis of the first SARS-CoV-2 variants introduced in Lebanon

Rita Feghali[1,*], Georgi Merhi[2,*], Aurelia Kwasiborski[3], Veronique Hourdel[3], Nada Ghosn[4] and Sima Tokajian[2]

[1] Department of Laboratory Medicine, Rafik Hariri University Hospital, Beirut, Lebanon
[2] Department of Natural Sciences, Lebanese American University, Byblos, Lebanon
[3] Laboratory: Environment and Infectious Risks, Pasteur Institute, Paris, France
[4] Epidemiological Surveillance Unit, Ministry of Public Health, Beirut, Lebanon
* These authors contributed equally to this work.

ABSTRACT

Background: In December 2019, the COVID-19 pandemic initially erupted from a cluster of pneumonia cases of unknown origin in the city of Wuhan, China. Presently, it has almost reached 94 million cases worldwide. Lebanon on the brink of economic collapse and its healthcare system thrown into turmoil, has previously managed to cope with the initial SARS-CoV-2 wave. In this study, we sequenced 11 viral genomes from positive cases isolated between 2 February 2020 and 15 March 2020.

Methods: Sequencing data was quality controlled, consensus sequences generated, and a maximum-likelihood tree was generated with IQTREE v2. Genetic lineages were assigned with Pangolin v1.1.14 and single nucleotide variants (SNVs) were called from read files and manually curated from consensus sequence alignment through JalView v2.11 and the genomic mutational interference with molecular diagnostic tools was assessed with the CoV-GLUE pipeline. Phylogenetic analysis of whole genome sequences confirmed a multiple introduction scenario due to international travel.

Results: Three major lineages were identified to be circulating in Lebanon in the studied period. The B.1 (20A clade) was the most prominent, followed by the B.4 lineage (19A clade) and the B.1.1 lineage (20B clade). SNV analysis showed 15 novel mutations from which only one was observed in the spike region.

Subjects Genomics, Microbiology, Molecular Biology, Virology, Infectious Diseases
Keywords COVID-19, SASR-CoV-2, Lebanon, SNV analysis, B.1 (20A clade), B.4 (19A clade)

INTRODUCTION

In December 2019, unknown cases of pneumonia were detected in the city of Wuhan, Hubei province, China. Infected individuals exhibited symptoms similar to that of severe acute respiratory syndrome (SARS) (*Li et al., 2020*). Deep sequencing identified the causative agent as a novel β-coronavirus, named nCoV-2019, later, renamed as SARS-CoV-2 (*Gralinski & Menachery, 2020*). Since then, the World Health Organization (WHO) has classified the spread of the virus as a global pandemic (*Astuti & Ysrafil, 2020*) and as of 18 January 2021, there was a total of 93,194,922 confirmed cases with

Submitted 23 September 2020
Accepted 5 February 2021
Published 16 March 2021

Corresponding author
Sima Tokajian, stokajian@lau.edu.lb

Academic editor
Alexander Bolshoy

Additional Information and Declarations can be found on page 11

DOI **10.7717/peerj.11015**

2,014,729 deaths (https://covid19.who.int/). COVID-19's common clinical symptoms include, but not limited to, fever, development of a dry cough, myalgia/fatigue, dyspnea, headaches and pneumonia (*Zhou et al., 2020*; *Huang et al., 2020*).

Molecular characterization of the SARS-CoV-2 genome showed 79.6% sequence similarity to SARS-CoV and 96% to the RaTG13 bat-CoV (*Zhou et al., 2020*) supporting that bats (*Rhinolophus affinis*) may have acted as a reservoir for the SARS-CoV-2 progenitor (*Andersen et al., 2020*). Interestingly, analysis of SARS-CoV-2's structural protein ORF3a revealed highly conserved protein domains within homologs from civet, pangolin and SARS-like bat-CoVs (*Issa et al., 2020*).

Despite the difficult economic situation and the waning infrastructure of its healthcare sector, Lebanon has successfully managed the initial COVID-19 wave. As of 21 February 2020, there has been 249,158 confirmed cases with 1,866 total deaths, with 1.4% of cases being attributed to exposure outside Lebanon and 98.6% due to local community spread of the SARS-CoV-2 virus (https://www.moph.gov.lb/en/Pages/2/193/esu#/en/Pages/2/24870/novel-coronavirus-2019-).

In this study, we performed a comprehensive genomic analysis of 11 SARS-CoV-2 isolates recovered from Lebanese individuals during the first phase of the pandemic. We also looked at their phylogeny based on location and date of exposure. Finally, we determined the single nucleotide variants (SNVs) and amino acid changes and compared our results with worldwide disseminating SARS-CoV-2 variants to assess epidemiological relatedness.

MATERIALS AND METHODS

COVID-19 response in Lebanon and genomic sequencing

The 11 cases (Table 1) undertaken in this study were detected and then selected randomly between 21 February 2020 and 15 March 2020 and designated as S1–S11. The clinical data were collected as part of the quarantine monitoring measures at the Rafik al Hariri University Hospital with the support of the Lebanese Ministry of Public Health. Required written informed consent was not obtained for the first studied cases due to the pressing need for data collection in the early stages of the outbreak. This study was approved by the Institutional Review Board (IRB) of the Lebanese American University (IRB #:AU.SAS.ST2.19/May/2020).

RNA was extracted from the specimens using the Qiagen QIAamp Viral RNA Mini kit (QIAGEN, Hilden, Germany) by following the manufacturer's instructions. An qRT-PCR corresponding to the Charité protocol (published on 17 January 2020) was used for detection of SARS-CoV-2. The assay relies on a first-line E gene screening, followed by a confirmatory assay using the RNA dependent RNA polymerase (*RdRp*) gene and a synthetic RNA positive control (Charité virology institute—Universitätsmedizin Berlin, Berlin, Germany). A 25 µl reaction was set up containing 1 µl of forward primer (10 µM), 1 µl of reverse primer (10 µM), 0.5 µl of probe (10 µM), 5 µl (100 ng/µl) of extracted RNA, 12.5 µl of 2X reaction buffer and 1 µl of reverse Transcriptase/Taq Polymerase mixture

Table 1 Clinical characteristics of Lebanese patients with COVID-19 between February and March 2020.

Accession number—GISAID	Sample	Gender	Age (years)	Location of exposure	Sample type	Ct*	Patient status	Signs and symptoms
EPI_ISL_450508	S1	Male	56	Egypt	Sputum/PBS	14.78	Hospitalized-deceased	Early stages: flu like symptoms Followed by severe dyspnea and severe ARDS† Chest X-ray: patchy bilateral upper lobe consolidation
EPI_ISL_450509	S2	Male	63	Local—community acquired	Nasopharyngeal VTM	26.15	Hospitalized-released	Asymptomatic
EPI_ISL_450510	S3	Male	19	Iran	Nasopharyngeal VTM	37.5	Hospitalized-released	Headache, abdominal pain and diarrhea
EPI_ISL_450511	S4	Female	33	United Kingdom	Nasopharyngeal VTM	34.67	Hospitalized-released	Rhinorrhea and headache
EPI_ISL_450512	S5	Male	42	Iran	Nasopharyngeal VTM	16.11	Hospitalized-released	Asymptomatic
EPI_ISL_450513	S6	Male	–	France	Nasopharyngeal VTM	33.7	Hospitalized-released	Asymptomatic
EPI_ISL_454420	S7	Female	41	Iran	Nasopharyngeal VTM	20.24	Hospitalized-released	Sore throat and rhinorrhea
EPI_ISL_450514	S8	Female	74	Local—community acquired	Nasopharyngeal VTM	33.8	Hospitalized-released	Dyspnea
EPI_ISL_450515	S9	Female	25	United Kingdom	Nasopharyngeal VTM	34	Hospitalized-released	Asymptomatic
EPI_ISL_450516	S10	Male	36	Local—community acquired	Sputum/PBS	–	Hospitalized-released	Asymptomatic
EPI_ISL_450517	S11	Male	22	Italy	Nasopharyngeal VTM	–	Hospitalized-released	Early stages: dysuria, fever and flu-like symptoms Followed by severe dyspnea and ARDS Chest CT: bilateral infiltrates with ground glass appearance

Notes:
* CT, RT-PCR Cycle Threshold.
† ARDS, acute respiratory distress syndrome.

(Invitrogen, New York, NY, USA) provided with the Superscript III One step RT-PCR system (*Corman et al., 2020*).

Thermal cycling was performed at 55 °C for 10 min for reverse transcription, followed by 95 °C for 3 min and then 45 cycles of 95 °C for 15 s and 58 °C for 30 s. cDNA amplification was done through the tiled amplification approach and following ARCTIC's network recommended protocol (*Quick, 2020*).

Sequencing libraries were prepared from tiled amplicons using the Miseq reagent kit v3 (Illumina, San Diego, CA, USA) and sequenced on an Illumina MiSeq system using 250 bp paired-end reads and following the manufacturer's instructions. Sequences were randomly labeled from S1 to S11 based on their loading order onto the sequencer.

Consensus sequences

Raw sequencing reads were input into FastQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for quality assessment and subsequent quality control was performed through FQCleaner v3.0 (*Criscuolo & Brisse, 2013*) with a 28 quality score threshold and a minimum read length of 30. Quantified reads were mapped to the Wuhan-Hu-1 reference genome (MN908947.3) using BWA v0.7.17 with the "MEM" alignment algorithm (https://arxiv.org/abs/1303.3997). Consensus sequences were generated for all 11 SARS-CoV-2 isolates using SAMtools v1.9 (http://www.htslib.org/) and vcftools v0.1.16 (https://vcftools.github.io/index.html).

Phylogenetic analysis

Six high quality ($N < 0.5\%$; Length > 29,000 bp) genomes (S1, S2, S4, S5, S8, S9) were selected for phylogenetic analysis. Reference sequences ($n = 55$) were chosen based on several criteria including phylogenetic placement of reference genomes, collection dates, history of exposure (travel), GISAID's BLAST feature within the EpiCoV^{TM} browser (https://www.gisaid.org) and overall genome quality/completeness to avoid sequence-based bias (Table S1).

Sequences were aligned with MAFFT v7.467 (*Katoh & Standley, 2013*). The resulting alignment was used in masking terminal regions and gaps with Nextstrain's custom Python script (https://github.com/nextstrain/ncov/tree/master/scripts). The alignment was input into ModelFinder to assess the best fit substitution model (*Kalyaanamoorthy et al., 2017*). A maximum likelihood (ML) tree was generated with IQ-TREE v2 (*Minh et al., 2020*) using the TIM2+F substitution model. Boot-strap support was established through 1,000 iterations for Ultra-Fast Bootstrapping (UFBoot) and SH-like approximate likelihood ratio test (SH-aLRT) (*Guindon et al., 2010*). The consensus tree was visualized with the interactive tree of life v4 (IToL; https://itol.embl.de/) (*Letunic & Bork, 2019*).

We assigned genetic lineages based on three commonly used systems including the recently proposed dynamic classification using Pangolin tool v1.1.14 (https://github.com/hCoV-2019/pangolin) (*Rambaut et al., 2020*), the Nextstrain classification and GISAID's internal classification.

Comparative genome and spike (S) protein analyses

SAM files from read alignment for all samples were converted into BAM files with SAMtools v1.9 (http://www.htslib.org/). Using bcftools v1.9 (https://samtools.github.io/bcftools/), variants were called and extracted through the "*mpileup*" and "*call*" commands with ploidy set to 1 and invoking the multiallelic-caller through the "-m" flag. Obtained variants were filtered with the "varfilter" command using the custom perl script *vcfutils.pl* (http://www.htslib.org/).

Consensus genomes were aligned against Wuhan-Hu-1 (MN908947.3) with MAFFT v7.467 (*Katoh & Standley, 2013*) and the resulting alignment was input in SNP-sites v2.5.1 (*Page et al., 2016*) to extract all identified SNPs (Table 2). The alignment was visualized (Fig. S1) with JalView v2.11 (*Waterhouse et al., 2009*) and polymorphic sites were manually curated (positions, with low coverage, defined by N strings were omitted).

Table 2 SNP distribution in SARS-CoV-2 genome sequences.

Gene	High depth (DP > 15) nucleotide mutations	Low depth (DP < 15) nucleotide mutations	Isolates
Nsp2	G1397A	C884T, C1093T, C1884T	S5, S9
Nsp3	C3037T, A7766G	C6078T, C6198A, A6281G, A6282T, C6285A, C7528T,	S1, S2, S4, S5, S7, S8, S10, S11
Nsp4	C9118T	G8653T, C8655T, A8658G, A8897T, T9860C, T9861G	S2, S5, S7, S8, S9, S10, S11
Nsp5	–	C10074T, A10075T	S7, S11
Nsp6	G11083T	–	S5, S9
Nsp8	–	A12297T	S10
Nsp10	C13381T	–	S2, S8, S11
Nsp12	C14408T	G14369T, C14703T, C14724T, C14802T, C14993T	S1, S2, S4, S7, S8, S10, S11
Nsp13	–	G16301T	S7
Nsp14	C18877T	G18670T	S1, S7
Nsp15	–	A19499C	S7
Spike (S) gene	A23403G	G22021T, T22092G	S1, S2, S4, S7, S8, S9, S10, S11
ORF3a	G25563T, C25578T, C25609T	C25611T, A25965G	S1, S2, S5, S8, S9, S10, S11
ORF7a	–	C27643T	S11
None-coding region	–	G27788A, T27789C	S5, S9
N gene	T28688C, G28881A, G28882A, G28883C	C28354T	S4, S5, S8, S9
None-coding region	–	G29543T	S7
3′ UTR	G29742T	–	S5, S7, S9

For added stringency, sequences were input into the CoV-GLUE analysis pipeline (http://cov-glue.cvr.gla.ac.uk/) where all SNPs and amino acid variations were identified for all genomes. Potential interference with all available diagnostic assays for SARS-CoV-2 was also investigated.

Genome annotation was performed with Prokka v1.14.6 (*Seemann, 2014*). Subsequently, spike (S) protein amino acid sequences were extracted, aligned with MAFFT v.7467 (*Katoh & Standley, 2013*) and visualized with JalView v2.11 (*Waterhouse et al., 2009*).

RESULTS

Clinical characteristics of patients

The first SARS-CoV-2 positive case was documented in Lebanon on 21 February 2020. By 15 March 2020, Lebanon had a total of 108 positive cases (https://www.moph.gov.lb/maps/covid19.php). Among the eleven patients, five were clinically asymptomatic (n = 5) and four exhibited mild symptoms (n = 4). The remaining two patients displayed a severe form of COVID-19 (Fig. 1). All patients were hospitalized, including asymptomatic carriers as a form of quarantined isolation to slow community spread (Fig. 1). Four patients out of eleven (n = 4) were female and the median age was 38.5 years and the range was 19–74 (Table 1).

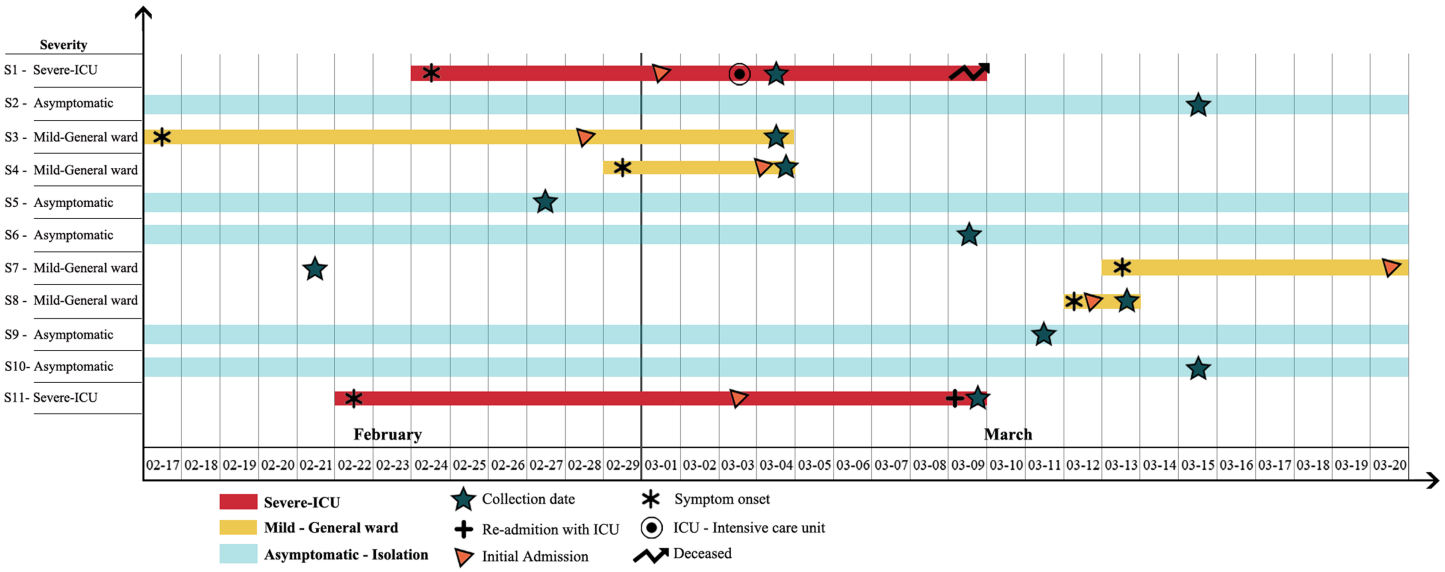

Figure 1 Timeline of symptom onset, illness severity, SARS-CoV-2 RNA sample collection date, hospitalization, and ICU admission of the 11 COVID-19 patients between February and March 2020 in Lebanon. All patients were admitted to the hospital with varying illness severity. Light blue bars represent asymptomatic patients. Gold bars indicate patients displaying mild symptoms while red bars denote patients with severe cases. Days are numbered sequentially from the 17th of February until the 20th of March. The star symbol indicates the RNA sample collection date while the asterisk (*) symbol denotes the date of symptom onset. Orange colored triangles mark the initial hospitalization date. The target symbol indicates admission into the intensive care unit (ICU) and the plus (+) sign is unique to patient S11 where it indicates his admission into the ICU in a different healthcare facility. The forked arrow symbol represents a patient's death. Full-size DOI: 10.7717/peerj.11015/fig-1

Dates of symptom onset varied between the 17th of February and 13th of March (Fig. 1). Travel history differed between patients and included countries such as Iran, France, Italy and the United Kingdom (Table 1), which was consistent with multiple introduction incidences. The initial signs and symptoms were headache (n = 2), rhinorrhea (n = 2) and flu-like symptoms (n = 2). Over the course of illness, a patient reported abdominal pain and diarrhea while another only suffered from dyspnea (Table 1). The two patients with severe COVID-19, initially developed flu-like symptoms followed by severe dyspnea and acute respiratory distress syndrome (Fig. 1).

Phylogenetic analysis

The SARS-CoV-2 isolates recovered from Lebanon clustered, and according to Nextstrain's classification, in three distinct clades namely: 19A (93% bootstrap support), 20A (81% bootstrap support) and 20B (98% bootstrap support) (Fig. 2). S5 (EPI_ISL_450512) and S9 (EPI_ISL_450515) were grouped closely to sequences from India (EPI_ISL_435106, EPI_ISL_421667, EPI_ISL_435101) and Kuwait (EPI_ISL_416458) in clade19A. Interestingly, S9 displayed less phylogenetic divergence than S5 based on the distance observed in the phylogenetic tree and polymorphic sites differences. Both patients had different exposure histories (Table 1).

S1 (EPI_ISL_450508), S2 (EPI_ISL_450509) and S8 (EPI_ISL_450514) clustered within clade 20A. S2 and S8 were closely related to one recovered from Egypt (EPI_ISL_430820). However, S2 was phylogenetically more related to the isolate from Egypt than S8, with both being linked to local community transfer (Fig. 2; Table 1). S1 was recovered from a

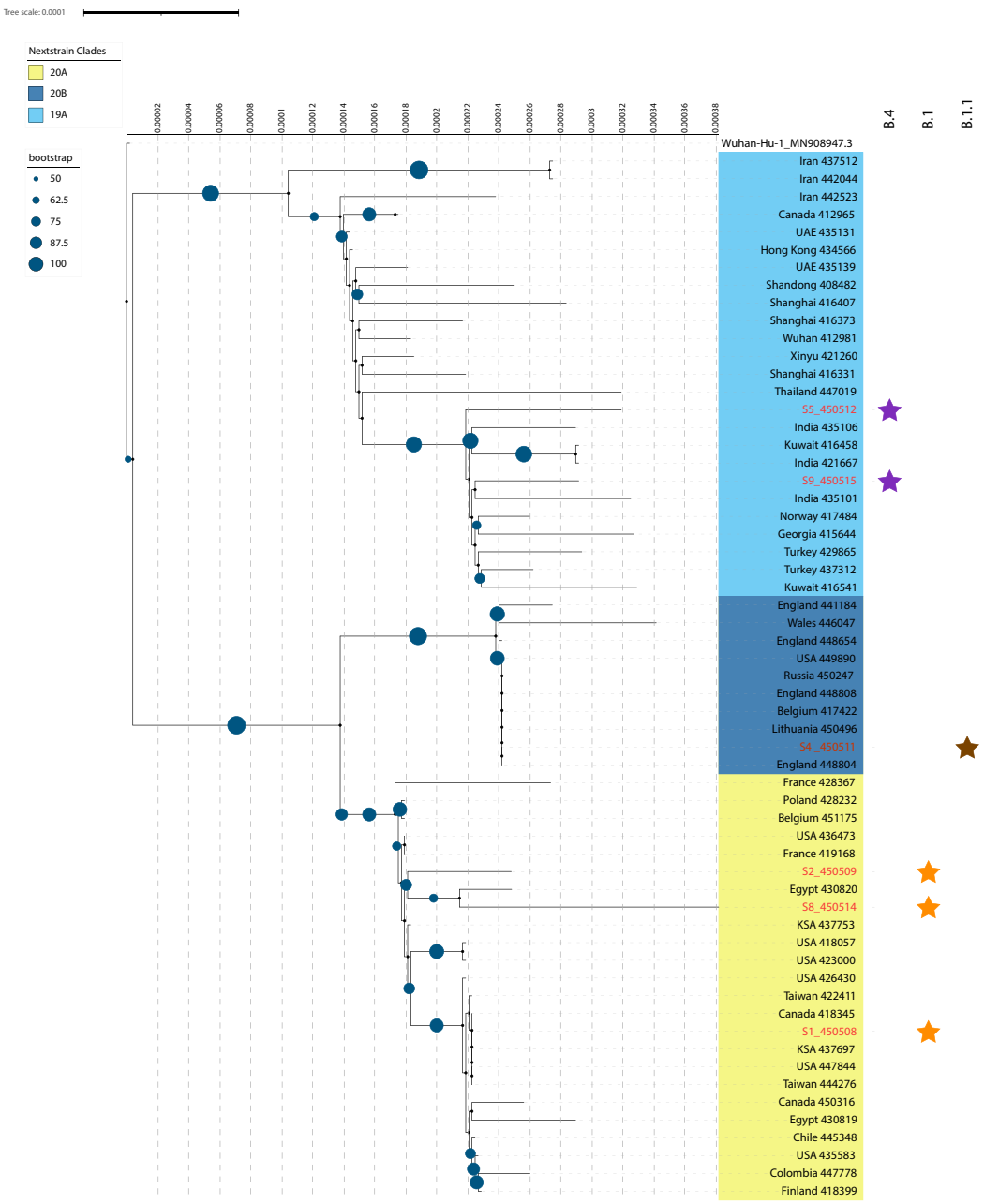

Figure 2 A maximum-likelihood (ML) phylogenetic tree of SARS-CoV-2 genomic sequences isolated from Lebanon. Only six genomes, with labels colored red, were selected for the phylogenetic tree to prevent any bias due to N strings in genomic sequences. Bootstrap values are represented by dark blue circles on the tree branches. Nextstrain clades were defined through different clade colors with light blue denoting the 19A clade, dark blue representing the 20B clade and gold designating the 20A clade. Pangolin lineages are displayed by star symbols. Purple stars represent the B.4 lineage, brown denotes the B.1.1 lineage and orange-colored stars stand for the B.1 lineage. Tree scale connotes for raw branch length and an internal scale system was added for additional stringency.

Full-size DOI: 10.7717/peerj.11015/fig-2

Table 3 Novel amino acid mutations in SARS-CoV-2 genomes.

Sample	Total number of SNVs	Genome coverage (%)	Novel amino acid changes as of time of submission	Pangolin lineage assignment	SH-alrt (%)	UFbootstrap (%)
S1	5	>95	None	B.1	100	100
S2	6	>95	None	B.1	100	100
S3	n/a	<95	n/a	B.4	100	100
S4	6	>95	None	B.1.1	100	99
S5	8	>95	nsp3: I1683V	B.4	100	100
S6	n/a	<95	n/a	B.1	100	100
S7	16	>95	nsp12: C310F; nsp13: R22I; nsp14: D211Y, H487P	B.1	100	100
S8	10	>95	nsp3: S1160Y; nsp4: M33Ffs	B.1	100	100
S9	7	>95	None	B.4	100	100
S10	18	>95	nsp3: N1188V, T1189N; nsp4: L436R; nsp8: Q69L; ORF7b: C12R	B.1	100	100
S11	19	>95	nsp4: N115Y; nsp12: S518L; S protein: M177R	B.1	100	100

patient with travel history to Egypt and clustered close to sequences recovered from Saudi Arabia (EPI_ISL_437697), USA (EPI_ISL_447844) and Taiwan (EPI_ISL_444276) (Table 1). Additionally, S1 showed lesser evolutionary distance than S2 and S8 (Fig. 2) indicating less genomic diversity. S4 (EPI_ISL_450511) on the other hand, was of the same superclade as that of S1, S2 and S8, but clustered under 20B (Fig. 2). It also showed proximity to sequences recovered from Europe and more so from England (EPI_ISL_448804), Belgium (EPI_ISL_417422) and Lithuania (EPI_ISL_450496), which was in accordance with the patient's travel history.

We also assigned the lineages using pangolin v1.1.14, and the results (Table 3) obtained were consistent (Fig. 2). S5 and S9 were assigned to the B.4 lineage alongside S3 (Table 3). S4 and S6 were assigned to lineage B.1.1, while the remaining isolates clustered under the B.1 lineage representing clade 20A (Fig. 2). The GISAID nomenclature could be also correlated with the lineage assignments. S5 and S9 were included in the O clade with all the others fitting under the super G clade: S4 and S6: GR clade and S1, S2, S7, S8, S10 and S11: GH clade.

SNVs in sequenced genomes and the S protein

The obtained consensus genomes of S3 and S6 showed low coverage values being 70.2% and 92.6%, respectively. Accordingly, polymorphism could not be confirmed and as such were excluded from downstream analysis. Table 2 shows all detected SNP (high and low depth) sites across the 11 aligned genomes. The genetic variation within the aligned genomes was relatively low, with a minimum of 5 and a maximum of 19 SNPs (median: 11 SNPs). In S1, S2, S7, S8, S10 and S11 three amino acid (AA) changes were detected within: ORF1ab within the non-structural protein 12 (nsp12) P323L, spike (S) protein D614G and ORF3a's Q57H (Table 2). Only P323L and D614G AA changes were detected in S4 with two other additional mutations in the N protein (R203K and G204R). S5 and S9 shared four common AA changes in nsp2 (R27C, V198I), nsp4 (M33I) and nsp6

(L37F). Interestingly, S5 had an additional mutation in nsp2 (A360V) and a novel mutation in nps3 (I1683V), while S9 displayed an AA change in the spike (S) protein at position 153 (M153I).

We also detected in S7, S8, S10 and S11 novel amino acid changes in multiple coding regions within the SARS-CoV-2 genome (Table 3). Furthermore, S8 had a novel frame-shift (fs) deletion at nucleotide position 8,651, leading to a change in nsp4 at position 33 and as a result replacing a methionine residue by the aromatic residue phenylalanine (M33Ffs).

Analysis of the polymorphic sites in the context of diagnostics revealed multiple hot-spots where accumulating nucleotide polymorphisms could interfere with the diagnostic schemes based on the ARCTIC network amplicon sequencing primers. In particular, changes such as: C13381T, G28881A, G28882A and G28883C in S2, S4, S6, S8 and S11 may interfere with the specificity of the primers and probes designed by the Chinese Center for Disease Control and Prevention (China CDC) and used for the detection of 2019-nCoV through targeting the ORF1ab encoded polymerase and N protein.

We also compared the S protein within the 11 genomes through an intra-isolate alignment. The D614G amino acid change was detected in all the isolates except S3, S5 and S9. This mutation, however, would not cause changes in the receptor binding domain (RBD-Spike: 455–505) or in the polybasic cleavage site unique (PCS-Spike: 681–686) to SARS-CoV-2. Finally, M153I and M177R mutations were only observed in isolates S9 and S11, respectively.

DISCUSSION

We aimed in this study at studying the SARS-CoV-2 isolates recovered at the early stages of the COVID-19 outbreak in Lebanon. To that end, we sequenced eleven genomes, including patient zero, collected from 21 February 2020 to 15 March 2020. We investigated the phylogenetic and epidemiological relatedness of the genomes and found three lineages circulating since the start of the local outbreak. Furthermore, several novel amino acid mutations in ORF1ab encoding for non-structural proteins and other structural proteins were detected.

Our results showed that S3 and S5 belonged to the B.4 lineage (O clade on GISAID) with a travel history linked to Iran (Table 1). This is consistent with previous reports where distinct clades were attributed to returnees from Iran, with all the genomes clustering under the B.4 lineage (O clade) (*Eden et al., 2020*; *Potdar et al., 2020*). Moreover, sequences that clustered close to S5 in Fig. 2 were also associated with possible exposure and travel history to Iran. It is noteworthy, that the first COVID-19 case in Iran was officially reported on 19 February (https://covid19.who.int/region/emro/country/ir) whereas the first case (S7) in Lebanon was on 21 February and linked to a patient returning from the city of Qom, Iran (https://www.moph.gov.lb/en/Media/view/27426/coronavirus-disease-health-strategic-preparedness-and-response-plan-), and which clustered under the B.1 lineage (GH clade).

S4 and S9 were however, both linked to patients with travel history to the United Kingdom (UK), but clustered in two different lineages, B.1.1 and B.4, respectively. Detailed analysis of the European sub-clusters by Nextstrain (*Hadfield et al., 2018*) revealed that the UK outbreak was largely rooted in B.1 and B.1.1 lineages which was in agreement with the analysis from GISAID's EpiCoV™ database (*Shu & McCauley, 2017*). The outbreak in Europe was mainly linked to isolates clustering under clade G and its variants GH/GR. Available data suggested that sequences linked to travel history to the UK represented the early stages of the outbreak with the B.4 lineage being actively circulating at the time.

The predominant lineage in this study was B.1 (clade GH) (S1, S2, S7, S8, S10, S11). Figure 2 shows that S2 and S8 were closely related to SARS-CoV-2 recovered in Egypt (EPI_ISL_430820) and somewhat distant from S1. S1 was linked to travel history to Egypt, while S2 and S8 to local community spread suggesting potential multiple introductions especially with the observed discrepancies in the number of SNPs in S8 compared to S1 (Table 3).

S protein amino acid changes revealed three variants among the sequenced isolates. Previously, *Bhattacharyya et al. (2020)* suggested the widespread dominance of SARS-CoV-2 with D614G/A23403G substitution in Europe and North America (Table 2). The delC allele in *TMPRSS2*, common in Europe and North America, facilitated the entry of the 614G subtype into host cells, thus accelerating the spread of 614G subtype (*Bhattacharyya et al., 2020*).

A novel S protein mutation (M177R) was detected in the S1 subunit in one of the isolates in this study (S11) (Table 3), but at this point it is not clear whether it has any implications on its interaction with the ACE2 receptor. Additionally, the *RdRp* mutation P323L (C14408T), generally detected in isolates recovered from Europe and North America (*Pachetti et al., 2020*), was also detected in the sequenced isolates from Lebanon, and was consistent with the metadata further showing multiple introduction points from Europe (UK, Italy, France). Mutations in the *RdRp* are of interest with the encoded polymerase being an important target for current therapeutic polymerase inhibitors (*Pachetti et al., 2020*), and with our data revealing several novel mutations in nsp12 (Table 3).

CONCLUSIONS

The estimated mutation rate driving the genomic global diversity of the virus has been determined at approximately 6×10^{-4} nucleotides/genome/year (*Van Dorp et al., 2020*). The accumulated genetic diversity in the form of SNVs in the SARS-CoV-2 genomes throughout the COVID-19 pandemic serves as a tool to assess and quantify the genomic diversity, evolutionary distribution, and epidemiological linkage of the virus (*Yang et al., 2020*). With daily confirmed cases on an exponential rise in Lebanon (https://www.moph.gov.lb/en/Pages/2/193/esu#/en/Pages/2/24870/novel-coronavirus-2019-) and the rapid emergence of novel variants of concern (*Volz et al., 2021*), further sequencing efforts are urgently needed to assess the spread and phylogenomic characteristics of SARS-CoV-2 in Lebanon and keep track of the emerging variants which is much needed to mitigate the spread, and for vaccine development and efficacy. This study offers a comprehensive

genomic snapshot of the earlier stages of the local outbreak. Importantly, our analysis highlights the viral lineages and genomic mutations identified at the root of the Lebanese outbreak which are also reflective of the situation in neighboring regions that lack genomic and epidemiological data.

ACKNOWLEDGEMENTS

We thankfully acknowledge the personnel and laboratories who have generated and submitted sequences to the GISAID's EpiCoV™ database. This study does not declare ownership of these sequences. The openly available data was used to compare our results within an international framework and to provide further information about the phylogenomic status and the spreadability of the SARS-CoV-2 in countries with little to non-existing epidemiological and genomic data. We also acknowledge Dr. Guillain Mikaty, Dr. Valerie Caro and Dr. Jean-Claude Manuguerra from Institut Pasteur, for their technical assistance and support in the sequencing of the SARS-CoV-2 samples.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The sequencing of the SARS-CoV-2 samples was financed by the MediLabSecure project, founded by the European Commission (DEVCO: IFS/2018/402-247). This study was also funded by the Strategic Research Review Committee (Grant #SRRC-R-2019-38) at the Lebanese American University and by the National Council for Scientific Research (Grant #00993). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
MediLabSecure: SARS-CoV-2.
European Commission (DEVCO): IFS/2018/402-247.
Strategic Research Review Committee: #SRRC-R-2019-38.
Lebanese American University and by the National Council for Scientific Research: #00993.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Rita Feghali conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Georgi Merhi conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

- Aurelia Kwasiborski conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Veronique Hourdel conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Nada Ghosn conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Sima Tokajian conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Human Ethics

The following information was supplied relating to ethical approvals (i.e., approving body and any reference numbers):

Patient clinical data were collected as part of the quarantine monitoring measures with at the Rafik al Hariri University Hospital (RHUH) with the support of the Lebanese Ministry of Public Health (MoPH). Required written informed consent was waived due to the pressing need for data collection in the early stages of the outbreak. This study was approved by the Institutional Review Board (IRB) of the Lebanese American University IRB#LAU.SAS.ST2.19/May/2020.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

Data is available at GSAID: EPI_ISL_450508, EPI_ISL_450509, EPI_ISL_450510, EPI_ISL_450511, EPI_ISL_450512, EPI_ISL_450513, EPI_ISL_454420, EPI_ISL_450514, EPI_ISL_450515, EPI_ISL_450516, EPI_ISL_450517.

Accessing the GISAID databases requires users to fill out a registration form and accepting GISAID's terms of use and database access agreement. Upon filing the request and subsequent review, GISAID will provide the user with unique credentials in an activation email, enabling access to the EpiCoV and EpiFlu databases and all the data included. All genomic data are identifiable based on a unique accession ID known as the EPI accession numbers.

Data Availability

The following information was supplied regarding data availability:

Data is available at GSAID: EPI_ISL_450508, EPI_ISL_450509, EPI_ISL_450510, EPI_ISL_450511, EPI_ISL_450512, EPI_ISL_450513, EPI_ISL_454420, EPI_ISL_450514, EPI_ISL_450515, EPI_ISL_450516, EPI_ISL_450517.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.11015#supplemental-information.

REFERENCES

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of SARS-CoV-2. *Nature Medicine* **26(4)**:1–3 DOI 10.1038/s41591-020-0820-9.

Astuti I, Ysrafil. 2020. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): an overview of viral structure and host response. *Diabetes & Metabolic Syndrome* **14(4)**:407–412 DOI 10.1016/j.dsx.2020.04.020.

Bhattacharyya C, Das C, Ghosh A, Singh AK, Mukherjee S, Majumder PP, Basu A, Biswas NK. 2020. Global spread of SARS-CoV-2 subtype with spike protein mutation D614G is shaped by human genomic variations that regulate expression of TMPRSS2 and MX1 genes. *bioRxiv* DOI 10.1101/2020.05.04.075911.

Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, Bleicker T, Brünink S, Schneider J, Schmidt ML, Mulders DG, Haagmans BL, van der Veer B, van den Brink S, Wijsman L, Goderski G, Romette J-L, Ellis J, Zambon M, Peiris M, Goossens H, Reusken C, Koopmans MP, Drosten C. 2020. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance* **25(3)**:2431 DOI 10.2807/1560-7917.ES.2020.25.3.2000045.

Criscuolo A, Brisse S. 2013. AlienTrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics* **102(5–6)**:500–506 DOI 10.1016/j.ygeno.2013.07.011.

Eden J-S, Rockett R, Carter I, Rahman H, de Ligt J, Hadfield J, Storey M, Ren X, Tulloch R, Basile K, Wells J, Byun R, Gilroy N, O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S, Sorrell TC, Holmes EC, Dwyer DE, Kok J, Donovan L, Kumar S, Tran T, Ko D, Ngo C, Sivaruban T, Timms V, Lam C, Gall M, Gray K-A, Sadsad R, Arnott A. 2020. An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evolution* **6**:veaa027 DOI 10.1093/ve/veaa027.

Gralinski LE, Menachery VD. 2020. Return of the coronavirus: 2019-nCoV. *Viruses* **12(2)**:135 DOI 10.3390/v12020135.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59(3)**:307–321 DOI 10.1093/sysbio/syq010.

Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34(23)**:4121–4123 DOI 10.1093/bioinformatics/bty407.

Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z, Yu T, Xia J, Wei Y, Wu W, Xie X, Yin W, Li H, Liu M, Xiao Y, Gao H, Guo L, Xie J, Wang G, Jiang R, Gao Z, Jin Q, Wang J, Cao B. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395(10223)**:497–506 DOI 10.1016/S0140-6736(20)30183-5.

Issa E, Merhi G, Panossian B, Salloum T, Tokajian S. 2020. SARS-CoV-2 and ORF3a: nonsynonymous mutations, functional domains, and viral pathogenesis. *mSystems* **5**:e00266-20 DOI 10.1128/mSystems.00266-20.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14(6)**:587–589 DOI 10.1038/nmeth.4285.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30(4)**:772–780 DOI 10.1093/molbev/mst010.

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* **47(W1)**:W256–W259 DOI 10.1093/nar/gkz239.

Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, Xing X, Xiang N, Wu Y, Li C, Chen Q, Li D, Liu T, Zhao J, Liu M, Tu W, Chen C, Jin L, Yang R, Wang Q, Zhou S, Wang R, Liu H, Luo Y, Liu Y, Shao G, Li H, Tao Z, Yang Y, Deng Z, Liu B, Ma Z, Zhang Y, Shi G, Lam TTY, Wu JT, Gao GF, Cowling BJ, Yang B, Leung GM, Feng Z. 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine* **382(13)**:1199–1207 DOI 10.1056/NEJMoa2001316.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* **37(5)**:1530–1534 DOI 10.1093/molbev/msaa015.

Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo RC, Zella D, Ippodrino R. 2020. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of Translational Medicine* **18(1)**:179 DOI 10.1186/s12967-020-02344-6.

Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics* **2(4)**:e000056 DOI 10.1099/mgen.0.000056.

Potdar V, Cherian SS, Deshpande GR, Ullas PT, Yadav PD, Choudhary ML, Gughe R, Vipat V, Jadhav S, Patil S, Nyayanit D, Majumdar T, Walimbe A, Gaikwad S, Dighe H, Shete-Aich A, Mohandas S, Chowdhury D, Sapkal G, Basu A, Gupta N, Gangakhedkar RR, Giri S, Dar L, Jain A, Malhotra B, Abraham P, Team NIC. 2020. Genomic analysis of SARS-CoV-2 strains among Indians returning from Italy, Iran & China, & Italian tourists in India. *Indian Journal of Medical Research* **151**:255 DOI 10.4103/ijmr.IJMR_1058_20.

Quick J. 2020. nCoV-2019 sequencing protocol. *protocols.io. Available at dx.doi.org/10.17504/protocols.io.bdp7i5rn*.

Rambaut A, Holmes EC, Hill V, O'Toole Á, McCrone JT, Ruis C, Du Plessis L, Pybus OG. 2020. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *bioRxiv* DOI 10.1101/2020.04.17.046086.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30(14)**:2068–2069 DOI 10.1093/bioinformatics/btu153.

Shu Y, McCauley J. 2017. GISAID: global initiative on sharing all influenza data—from vision to reality. *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* **22(13)**:30494 DOI 10.2807/1560-7917.ES.2017.22.13.30494.

Van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, Ortiz AT, Balloux F. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution* **83(1)**:104351 DOI 10.1016/j.meegid.2020.104351.

Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, Hinsley WR, Laydon DJ, Dabrera G, O'Toole Á, Amato R, Ragonnet-Cronin M, Harrison I, Jackson B, Ariani CV, Boyd O, Loman N, McCrone JT, Gonçalves S, Jorgensen D, Myers R, Hill V, Jackson DK, Gaythorpe K, Groves N, Sillitoe J, Kwiatkowski DP, Flaxman S, Ratmann O, Bhatt S, Hopkins S, Gandy A, Rambaut A, Ferguson NM, COG-UK. 2021. Transmission of SARS-CoV-2 lineage B.1.1.7 in England: insights from linking epidemiological and genetic data. *medRxiv* DOI 10.1101/2020.12.30.20249034.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25(9)**:1189–1191 DOI 10.1093/bioinformatics/btp033.

Yang H-C, Chen C, Wang J-H, Liao H-C, Yang C-T, Chen C-W, Lin Y-C, Kao C-H, Liao JC. 2020. Genomic, geographic and temporal distributions of SARS-CoV-2 mutations. *bioRxiv* DOI 10.1101/2020.04.22.055863.

Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L, Chen H-D, Chen J, Luo Y, Guo H, Jiang R-D, Liu M-Q, Chen Y, Shen X-R, Wang X, Zheng X-S, Zhao K, Chen Q-J, Deng F, Liu L-L, Yan B, Zhan F-X, Wang Y-Y, Xiao G-F, Shi Z-L. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579(7798)**:1–4 DOI 10.1038/s41586-020-2012-7.