PLOS ONE

Genomic epidemiology of SARS-CoV-2 importation and early circulation in Israel

--Manuscript Draft--

Manuscript Number:	PONE-D-20-27397
Article Type:	Research Article
Full Title:	Genomic epidemiology of SARS-CoV-2 importation and early circulation in Israel
Short Title:	SARS-CoV-2 genomic epidemiology in Israel
Corresponding Author:	Neta Zuckerman Ministry of Health Ramat Gan, Israel ISRAEL
Keywords:	SARS-CoV-2, clades, mutations, genomic epidemiology
Abstract:	Severe acute respiratory disease coronavirus 2 (SARS-CoV-2) which causes corona virus disease (COVID-19) was first identified in Wuhan, China in December 2019 and has since led to a global pandemic. Importations of SARS-CoV-2 to Israel in late February from multiple countries initiated a rapid outbreak across the country. In this study, SARS-CoV-2 whole genomes were sequenced from 59 imported samples with a recorded country of importation and 101 early circulating samples in February to mid-March 2020 and analyzed to infer clades and mutational patterns with additional sequences identified Israel available in public databases. Recorded importations in February to mid-March, mostly from Europe, led to multiple transmissions in all districts in Israel. Although all SARS-CoV-2 defined clades were imported, clade 20C became the dominating clade in the circulating samples. Identification of novel, frequently altered mutated positions correlating with clade-defining positions provide data for surveillance of this evolving pandemic and spread of specific clades of this virus. SARS-CoV-2 continues to spread and mutate in Israel and across the globe. With economy and travel resuming, surveillance of clades and accumulating mutations is crucial for understanding its evolution and spread patterns and may aid in decision making concerning public health issues.
Order of Authors:	Neta S Zuckerman
	Efrat Bucris
	Yaron Drori
	Oran Erster
	Danit Sofer
	Rakefet Pando
	Ella Mendelson
	Orna Mor
	Michal Mandelboim
Additional Information:	
Question	**Response**
Financial Disclosure Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed requirements. View published research articles from *PLOS ONE* for specific examples.	The authors received no specific funding for this work

This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate. **Unfunded studies** Enter: *The author(s) received no specific funding for this work.* **Funded studies** Enter a statement with the following details: • Initials of the authors who received each award • Grant numbers awarded to each author • The full name of each funder • URL of each funder website • Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript? • **NO** - Include this sentence at the end of your statement: *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.* • **YES** - Specify the role(s) played. * typeset	
Competing Interests Use the instructions below to enter a competing interest statement for this submission. On behalf of all authors, disclose any competing interests that could be perceived to bias this work—acknowledging all financial support and any other relevant financial or non-financial competing interests. This statement **will appear in the published article** if the submission is accepted. Please make sure it is accurate. View published research articles from *PLOS ONE* for specific examples.	the authors have declared that no competing interests exist

NO authors have competing interests Enter: *The authors have declared that no competing interests exist.* **Authors with competing interests** Enter competing interest details beginning with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests: [insert competing interests here]* * typeset	
Ethics Statement Enter an ethics statement for this submission. This statement is required if the study involved: • Human participants • Human specimens or tissue • Vertebrate animals or cephalopods • Vertebrate embryos or tissues • Field research Write "N/A" if the submission does not require an ethics statement. General guidance is provided below. Consult the submission guidelines for detailed instructions. **Make sure that all information entered here is included in the Methods section of the manuscript.**	The study has been approved by the Sheba Medical Center Helsinki committee (#7045-20-smc). This is a retrospective study of archived samples, where sample names were anonymized; institutional Helsinki committee waived the requirement for informed consent.

Format for specific study types **Human Subject Research (involving human participants and/or tissue)** • Give the name of the institutional review board or ethics committee that approved the study • Include the approval number and/or a statement indicating approval of this research • Indicate the form of consent obtained (written/oral) or the reason that consent was not obtained (e.g. the data were analyzed anonymously) **Animal Research (involving vertebrate animals, embryos or tissues)** • Provide the name of the Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board that reviewed the study protocol, and indicate whether they approved this research or granted a formal waiver of ethical approval • Include an approval number if one was obtained • If the study involved *non-human primates*, add *additional details* about animal welfare and steps taken to ameliorate suffering • If anesthesia, euthanasia, or any kind of animal sacrifice is part of the study, include briefly which substances and/or methods were applied **Field Research** Include the following details if this study involves the collection of plant, animal, or other materials from a natural setting: • Field permit number • Name of the institution or relevant body that granted permission	
Data Availability Authors are required to make all data underlying the findings described fully available, without restriction, and from the time of publication. PLOS allows rare exceptions to address legal and ethical concerns. See the PLOS Data Policy and FAQ for detailed information.	Yes - all data are fully available without restriction

A Data Availability Statement describing where the data can be found is required at submission. Your answers to this question constitute the Data Availability Statement and **will be published in the article**, if accepted. **Important:** Stating 'data available on request from the author' is not sufficient. If your data are only available upon request, select 'No' for the first question and explain your exceptional situation in the text box. Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	
Describe where the data may be found in full sentences. If you are copying our sample text, replace any instances of XXX with the appropriate details. • If the data are **held or will be held in a public repository**, include URLs, accession numbers or DOIs. If this information will only be available after acceptance, indicate this by ticking the box below. For example: *All XXX files are available from the XXX database (accession number(s) XXX, XXX.*). • If the data are all contained **within the manuscript and/or Supporting Information files**, enter the following: *All relevant data are within the manuscript and its Supporting Information files.* • If neither of these applies but you are able to provide **details of access elsewhere**, with or without limitations, please do so. For example: *Data cannot be shared publicly because of [XXX]. Data are available from the XXX Institutional Data Access / Ethics Committee (contact via XXX) for researchers who meet the criteria for access to confidential data.* *The data underlying the results presented in the study are available from (include the name of the third party*	all sequences used and generated in this study are submitted and available in GISAID

and contact information or URL). • This text is appropriate if the data are owned by a third party and authors do not have permission to share the data. * typeset	
Additional data availability information:	

Manuscript

Click here to access/download;Manuscript;Strains_Manuscript_plos.docx

Genomic epidemiology of SARS-CoV-2 importation and early circulation in Israel

Neta S. Zuckerman[1], Efrat Bucris[1], Yaron Drori[1,2], Oran Erster[1], Danit Sofer[1], Rakefet Pando[1,3], Ella Mendelson[1,2], Orna Mor*[1,2], Michal Mandelboim*[1,2]

1 Central Virology Laboratory, Ministry of Health, Chaim Sheba Medical Center, Ramat Gan, Israel

2 School of Public Health, Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel

3 Israel Center for Disease Control, Israel Ministry of Health, Chaim Sheba Medical Center, Ramat Gan 5265601, Israel

* equal contribution

Corresponding author: Neta S. Zuckerman

Address: Central Virology Laboratory, Sheba Medical Center, Tel-Hashomer, 52621, Israel

Telephone number: +972-3-5302341

Email: Neta.Zuckerman@sheba.health.gov.il

Short title: SARS-CoV-2 genomic epidemiology in Israel

Abstract

Severe acute respiratory disease coronavirus 2 (SARS-CoV-2) which causes corona virus disease (COVID-19) was first identified in Wuhan, China in December 2019 and has since led to a global pandemic. Importations of SARS-CoV-2 to Israel in late February from multiple countries initiated a rapid outbreak across the country. In this study, SARS-CoV-2 whole genomes were sequenced from 59 imported samples with a recorded country of importation and 101 early circulating samples in February to mid-March 2020 and analyzed to infer clades and mutational patterns with additional sequences identified Israel available in public databases. Recorded importations in February to mid-March, mostly from Europe, led to multiple transmissions in all districts in Israel. Although all SARS-CoV-2 defined clades were imported, clade 20C became the dominating clade in the circulating samples. Identification of novel, frequently altered mutated positions correlating with clade-defining positions provide data for surveillance of this evolving pandemic and spread of specific clades of this virus. SARS-CoV-2 continues to spread and mutate in Israel and across the globe. With economy and travel resuming, surveillance of clades and accumulating mutations is crucial for understanding its evolution and spread patterns and may aid in decision making concerning public health issues.

Keywords: SARS-CoV-2, clades, mutations, genomic epidemiology, Israel

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first identified in Wuhan, China in December 2019 [1] and has since rapidly spread, infecting over 20 million people worldwide to this day. SARS-CoV-2 causes corona viral disease (COVID-19) and was declared a pandemic by the world health organization on March 2020 [2]. Currently, there is no vaccine or approved effective therapeutic treatments [3].

Major SARS-CoV-2 clades have been characterized based on whole viral genome sequencing data, with over 80,000 sequences currently deposited from countries worldwide in the global initiative on sharing all influenza data (GISAID) database [4]. The main nomenclature systems of SARS-CoV-2 clades include Nextstrain, who name a new major clade when it reaches a frequency of 20% globally by using a year-letter genetic clade naming [5], and GISAID, who use the statistical distribution of genome distances in phylogenetic clusters and name the clades by the actual letters of the defining marker mutations of each cluster [4]. According to Nextstrain's nomenclature system [5], five globally circulating SARS-CoV-2 clades are currently defined – 19A (the root clade) and 19B, that originated in Asia and are still widespread there, and clades 20A, B and C now dominate global infections and are widespread in Europe [6][7]. The 20 clades (G clades by GISAID nomenclature) have emerged in Europe in mid-January, and bear the D614G mutation (refers to the mutation in the amino acid sequence; A23403G refers to the nucleotide sequence) in the spike protein which bind the human ACE2 receptor [8]. This mutation has recently been associated with high viral loads and increased infectivity but not with patient hospitalization status [6], although recent reports argue that this variant is related to COVID-19 mortality [9][10]. Additional mutations within the SARS-CoV-2 genome are being monitored as potential emerging-clades (e.g. C18877T emerging from clade 20, C13730T emerging from clade 19) via Nextstrain's global genomic epidemiology analysis [5] and may become a major clade once they reach sufficient global frequency/spread.

SARS-CoV-2 started to spread in Israel in late February through early March 2020, where multiple importation events of SARS-CoV-2 into Israel from countries worldwide initiated a rapid outbreak across the country with >88,000 infected individuals and ~700 deaths by August 2020. Prompted by recent escalations in the daily number of infected individuals in Israel, in this study we sequenced 160 SARS-CoV-2 complete genomes from imported and early circulating samples. Along epidemiological data including country of importation and district of residence and additional Israel-based sequences from the same time frame available in GISAID, we thoroughly investigated mutation patterns to characterize the origins of viral evolution and spread patterns of SARS-CoV-2 in Israel.

Materials and Methods

Sample collection, nucleic acid extraction and viral genome quantification by real-time PCR (q-PCR)

Starting with the first imported cases into Israel in February and until mid-March 2020, all individuals entering Israel suspected to have contracted SARS-CoV-2 were exclusively diagnosed in Israel's Central Virology Laboratory (ICVL) via real-time PCR. Viral genomes were extracted from 200 µL respiratory samples with the MagNA PURE 96 (Roche, Mannheim, Germany), according to the manufacturer instructions and qRT-PCR reactions using primers corresponding to the SARS-CoV-2 envelope (E) gene were performed as previously described [11]. All samples were tested for the human RNAseP gene, which served as a housekeeping gene. The Quantitative reverse transcription PCR (qRT-PCR) reactions were performed in 25 µL

Ambion Ag-Path Master Mix (Life Technologies, Carlsbad, CA, USA) using TaqMan Chemistry on the ABI 7500 instrument. Nucleic extraction samples from SARS-CoV-2 positive samples were taken for further molecular analysis.

Ethics statement: The study has been approved by the Sheba Medical Center Helsinki committee. This is a retrospective study of archived samples, where sample names were anonymized; institutional Helsinki committee waived the requirement for informed consent.

Specific amplification of SARS-CoV-2 from clinical samples

RNA in extracted nucleic acids was reverse transcribed to single strand cDNA using SuperScript IV (ThermoFisher Scientific, Waltham, MA, USA) as per manufacturer's instructions. SARS-CoV-2 specific primers designed to capture SARS-CoV-2 whole genome (version 1—total 218 primers, divided into two primer pools designed by Josh Quick from ARTIC Network) were used to generate double strand cDNA and amplify it via PCR using Q5 Hot Start DNA Polymerase (NEB) [12]. Briefly, each sample underwent two PCR reactions with primer pool 1 or 2 and 5X Q5 reaction buffer, 19 mM dNTPs and nuclease-free water. Resulting DNA was combined and quantified with Qubit dsDNA BR Assay kit (ThermoFisher Scientific) as per manufacturer's instructions and 1ng of amplicon DNA in 5 μL per sample was taken into library preparation.

Library preparation and sequencing

Libraries were prepared using NexteraXT library preparation kit and NexteraXT index kit V2 as per manufacturer's instructions (Illumina, San Diego, CA, USA). Libraries were purified with AMPure XP magnetic beads (Beckman Coulter, Brea, CA, USA) and library concentration was measured by Qubit dsDNA HS Assay kit (Thermo Fisher Scientific, Waltham, MA, USA). Library validation and mean fragment size was determined by TapeStation 4200 via DNA HS

D1000 kit (Agilent, Santa Clara, CA, USA). The mean fragment size was ~400 bp, as expected. The library mean fragment size and concentration molarity was calculated and each library was diluted to 4 nM. Libraries were pooled, denatured and diluted to 10pM and sequenced on MiSeq with V3 2X300 bp run kit (Illumina). Sequences are available in GISAID.

Bioinformatics analyses

Fastq files were subjected to quality control using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC [13] and low-quality sequences were filtered using trimmomatic [14]. To obtain a consensus sequence per sample, paired-end fastq files were combined for each sample via Unix cat command. SARS-CoV-2 reference genome was downloaded from the national center for biotechnology information (NCBI) (NC_045512.2) and indexed using Burrows-Wheeler aligner (BWA) [15]. Combined fastq files were mapped to the indexed reference genome using BWA mem [15]. SAMtools suite [16] was used to convert sam to bam files, remove duplicates and filter unmapped reads. Bam files were sorted, indexed and subjected to quality control using SAMtools suite. Coverage and depth of sequencing was calculated from sorted bam files using a custom python script. A consensus sequence was constructed for each sample using SAMtools mpileup and bcf tools [17] and converted to a fasta file using seqtk (https://github.com/lh3/seqtk). Resulting consensus sequences were further analyzed together with additional sequences identified in Israel from late March to late April (n=211) available in GISAID [4]. Using Augur pipeline [5], sequences were aligned to SARS-CoV-2 reference genome (NC_045512.2) using MAFFT [18], and a time-resolved phylogenetic tree was constructed with IQ-Tree [19] and TreeTime [20] under the GTR substitution model and visualized with auspice [5]. Clade nomenclature was attained from Nextstrain [5].

Additional bioinformatic analyses such as translation from nucleotide to amino acid sequences, comparison of differences across sequences and sample clustering were carried out using R and Bioconductor packages Seqinr [21], HDMD (https://CRAN.R-project.org/package=HDMD) and ggplot2 [22]. Classification to amino acid groups was set according to physiochemical attributes determined by Atchley et al. [23].

Results

SARS-CoV-2 genomic epidemiology of imported and early circulating viruses

On February 21, 2020, two Israeli citizens infected with SARS-CoV-2 from the Diamond Princess cruise ship anchoring in Japan were brought to designated SARS-CoV-2 quarantine facilities in Israel. The first non-controlled imported case of SARS-CoV-2 into Israel from Europe (Italy) was diagnosed in February 27, 2020, followed by additional importations, mostly from other European countries but also from countries worldwide until early March, when air traffic was largely suspended. At that time, SARS-CoV-2 suspected individuals were exclusively diagnosed by the central virology laboratory, such that all epidemiologically-verified importations were recorded and samples were retained. Here, we sequenced complete SARS-CoV-2 genomes from all imported cases identified in late February to mid-March (n=59) and from circulating viruses from individuals diagnosed between mid-March and April (n=101) using SARS-CoV-2 whole genome capture (ARTIC network V3 primers, https://artic.network/ncov-2019) and next generation sequencing (Illumina). Results were analyzed together with additional sequences identified in Israel from late March to late April (n=211) available in GISAID [4]. Sequences were aligned to SARS-CoV-2 reference genome (NC_045512.2) using MAFFT [18]. A time-resolved phylogenetic tree was constructed using the augur toolchain [5], utilizing IQ-

Tree [19] with the GTR substitution model, TreeTime [20] and visualized with auspice [5]. Additional mutation analyses were carried out using R and Bioconductor.

The phylogenetic tree, depicting imported and circulating cases, shows that importation events from Europe, United States, Asia and Africa (Egypt) in late February to mid-March led to multiple transmission chains in Israel (Figure 1A). All districts in Israel were affected, with the highest number of importations occurring into the Central and Tel-Aviv districts (Figure 1B).

SARS-CoV-2 imported and circulating clades

To define imported and circulating clades in Israel, we applied the Nextstrain nomenclature (https://github.com/nextstrain/ncov/blob/master/defaults/clades.tsv), that includes the originating clade 19A and its derivation 19B, and the emerging clades 20 and its derivations 20B and 20C with the spike mutation D614G [6] that had widely spread in Europe since mid-February [5]. All five clades were imported into Israel during late February to mid-March (Figure 2A, n=59). Clades 19A and 19B constituted 40% of the clades imported into Israel with relatively equal representation (~20% each). Clade 19A included the two Diamond Princess samples imported from Japan, and clade 19B was almost exclusively imported from Spain (11/12 of 19B importations) (Figure 2B). Clade 20 constituted 60% of imported cases and included the first importation from Italy (clade 20B) (Figure 2B). Clade 20C, which was equally represented as clades 20A and 20B in the imported population became the dominant circulating clade in Israel (51%), whereas clade 19B diminished in the circulating population (Figure 2C). Within Israel, the Jerusalem and Tel Aviv districts had the highest SARS-CoV-2 incidence and the Haifa district the lowest during the early spread. Clade 20, specifically 20C, was the dominant clade in most districts (Figure 2C).

SARS-CoV-2 mutation patterns

To further explore patterns in viral evolution, we identified positions along the SARS-CoV-2 genome that were frequently altered across the Israeli sequences compared to the reference genome. Correlations of these positions revealed novel positions that were altered in the Israeli sequences, in addition to the known clade-defining positions (Figure 3A). The novel positions were associated with defined clades via Nextstrain auspice visualization tool [5]. Clusters of positive correlations were observed between mutated positions within each of the 19 and 20 clades, whereas negative correlations were observed between mutated positions associated with clades 19 and positions in clade 20, suggesting distinct linkage of these positions to either clade. Interestingly, negative correlations were observed between the clade 20B mutated positions (313, 28881, 28882, 28883) and some of the positions in clades A/C (e.g. 1059, 25563, 11916), which may hint that the clade B positions are strongly linked to one another (Figure 3A). Visualization with Nextstrain global analysis (https://nextstrain.org/ncov/global) showed that these mutated positions are not specifically unique to Israel and were observed in several SARS-CoV-2 genome sequences worldwide. To assess the impact of all these alterations, the resultant amino acid (AA) substitutions were classified into silent (S) or replacement (R), and in the latter case, the change in the physiochemical attributes of the AA (classified by Atchley [23]) was also assessed. R mutations were observed in higher frequency (20/29 mutations) compared to S (9/29), most of which led to a change in the AA attribute group (12/20 R mutations). Many of the AA group exchanges involved a change between the aliphatic (non-polar, hydrophobic) and hydroxylated (polar, uncharged) AA groups. Finally, over half of the mutations observed (15/29) were C-to-T, suggesting viral restriction by host APOBEC mechanism in these positions, as previously observed [24].

Discussion

Since its first importation into Israel late February 2020, SARS-CoV-2 had expeditiously spread in Israel. The first importations occurred from Japan and Europe (Italy), however the spread in the population is more likely to have been initiated by first importations from Europe, as the importations from Japan (Diamond Princess passengers) were planned and controlled in specialized treatment facilities. Sequencing and analyses of SARS-CoV-2 complete genomes from imported and circulating samples revealed that although several clades were initially imported into Israel in late February to mid-March, clade 20 quickly became dominant, similar to observations across Europe. Clade 20 (including 20A, B and C), also known as clade G by GISAID nomenclature [4], is an emerging clade that has gained prominence in Europe in early March followed by expansion into North America and Asia, where its hallmark mutation, D614G (a23403g in the nucleotide sequence), has been recently shown to increase infectivity [6]. Specifically, clade 20C, a dominant clade in North America [5] that was observed in 51% of circulating samples in Israel, may have been reinforced in gaining prominence in Israel by additional importations from the United States in late March, in addition to its naturally higher infectivity compared to clade 19.

Frequently mutated positions were identified in the Israeli samples, some of which correlated with known clade-defining mutations and observed also in sequences worldwide. Most of these mutations were R mutations that caused a change in the AA attribute group, which have a greater chance to affect the protein. It is important to closely observe the emerging mutated positions throughout this continuous pandemic as some may gain evolutionary advantage and affect larger portions of the population. This might have an impact on the specificity of diagnostic tests such as real time PCR and even vaccine design targeting these positions.

SARS-CoV-2 is still spreading in Israel and across the globe. Surveillance of SARS-CoV-2 genomes is crucial for understanding its evolution and spread patterns and may aid in decision making concerning public health issues.

Acknowledgments: none

Figure legends

Figure 1. SARS-CoV-2 genomic epidemiology of samples imported and circulating in Israel. **(A)** time-resolved phylogenetic tree representing 372 imported and early circulating samples sequenced in Israel. Samples are colored according to their origin: Israel circulating samples in yellow, and imported samples from Europe, USA, Asia (China-reference sequence, Japan) and Africa (Egypt) in red, green, blue and purple, respectively. SARS-CoV-2 clades are noted by each relevant branch. **(B)** Distribution of imported and circulating samples across districts in Israel.

Figure 2. SARS-CoV-2 imported and circulating clades. **(A)** Distribution of SARS-CoV-2 clades from first diagnosed sample in late February through early circulation in Israel. **(B)** Distribution and frequency of clades in imported samples in late February to mid-March, by country of origin. **(C)** Distribution and frequency of clades in early circulating samples (mid-March to late April), by district.

Figure 3. SARS-CoV-2 frequently observed mutations. 29 mutations along the SARS-CoV-2 genome occurred in >2% of the 372 Israeli sequences. **(A)** Correlation table of the frequently observed mutations. Positive/negative correlations are denoted in blue/red respectively. Known clade-defining mutations are underlined and clade association is noted. **(B)** Listed for each frequently observed mutation its position, gene, % frequency in Israeli sequences, nucleotide substitution, whether it’s an R or S mutation, and in case of an R mutation, the originating and altered AA group. Known clade-defining mutations are highlighted in grey.

References

[1] N. Zhu *et al.*, “A Novel Coronavirus from Patients with Pneumonia in China, 2019,” *N. Engl. J. Med.*, vol. 382, no. 8, pp. 727–733, Feb. 2020, doi: 10.1056/NEJMoa2001017.

[2] “WHO Director-General’s opening remarks at the media briefing on COVID-19 - 11 March 2020.” who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (accessed Aug. 15, 2020).

[3] C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao, “A novel coronavirus outbreak of global health concern,” *Lancet*, vol. 395, no. 10223, pp. 470–473, Feb. 2020, doi: 10.1016/S0140-6736(20)30185-9.

[4] S. Elbe and G. Buckland-Merrett, “Data, disease and diplomacy: GISAID’s innovative contribution to global health,” *Glob. Challenges*, vol. 1, no. 1, pp. 33–46, Jan. 2017, doi: 10.1002/gch2.1018.

[5] J. Hadfield *et al.*, “Nextstrain: real-time tracking of pathogen evolution,” *Bioinformatics*, vol. 34, no. 23, pp. 4121–4123, Dec. 2018, doi: 10.1093/bioinformatics/bty407.

[6] B. Korber *et al.*, “Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus,” *Cell*, Jul. 2020, doi: 10.1016/j.cell.2020.06.043.

[7] E. Alm *et al.*, “Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020,” *Eurosurveillance*, vol. 25, no. 32, Aug. 2020, doi: 10.2807/1560-7917.ES.2020.25.32.2001410.

[8] J. Lan *et al.*, “Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor.,” *Nature*, vol. 581, no. 7807, pp. 215–220, 2020, doi: 10.1038/s41586-020-2180-5.

[9] Y. Toyoshima, K. Nemoto, S. Matsumoto, Y. Nakamura, and K. Kiyotani, “SARS-CoV-2 genomic variations associated with mortality rate of COVID-19,” *J. Hum. Genet.*, Jul. 2020, doi: 10.1038/s10038-020-0808-9.

[10] M. Becerra-Flores and T. Cardozo, “SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate,” *Int. J. Clin. Pract.*, vol. 74, no. 8, Aug. 2020, doi: 10.1111/ijcp.13525.

[11] V. M. Corman *et al.*, “Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR.,” *Euro Surveill.*, vol. 25, no. 3, 2020, doi: 10.2807/1560-7917.ES.2020.25.3.2000045.

[12] “Artic network, SARS-CoV-2.” https://artic.network/ncov-2019 (accessed Aug. 15, 2020).

[13] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, “MultiQC: Summarize analysis results for multiple tools and samples in a single report,” *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, 2016, doi: 10.1093/bioinformatics/btw354.

[14] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimmer for Illumina sequence data.,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–20, Aug. 2014, doi: 10.1093/bioinformatics/btu170.

[15] H. Li, “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM,” Mar. 2013, [Online]. Available: http://arxiv.org/abs/1303.3997.

[16] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools.,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–9, Aug. 2009, doi: 10.1093/bioinformatics/btp352.

[17] V. Narasimhan, P. Danecek, A. Scally, Y. Xue, C. Tyler-Smith, and R. Durbin, “BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data,” *Bioinformatics*, vol. 32, no. 11, pp. 1749–1751, 2016, doi: 10.1093/bioinformatics/btw044.

[18] K. Katoh, “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform,” *Nucleic Acids Res.*, vol. 30, no. 14, pp. 3059–3066, Jul. 2002, doi: 10.1093/nar/gkf436.

[19] L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh, “IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.,” *Mol. Biol. Evol.*, vol. 32, no. 1, pp. 268–74, Jan. 2015, doi: 10.1093/molbev/msu300.

[20] P. Sagulenko, V. Puller, and R. A. Neher, “TreeTime: Maximum-likelihood phylodynamic analysis,” *Virus Evol.*, vol. 4, no. 1, Jan. 2018, doi: 10.1093/ve/vex042.

[21] C. Delphine and L. Jean R., “SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis,” in *Structural Approaches to Sequence Evolution*, Springer: Berlin/Heidelberg, Germany, 2007, pp. 207–232.

[22] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[23] W. R. Atchley, W. Terhalle, and A. Dress, “Positional Dependence, Cliques, and Predictive Motifs in the bHLH Protein Domain,” *J. Mol. Evol.*, vol. 48, no. 5, pp. 501–516, May 1999, doi: 10.1007/PL00006494.

[24] S. Di Giorgio, F. Martignano, M. G. Torcia, G. Mattiuz, and S. G. Conticello, “Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2,” *Sci. Adv.*, vol. 6, no. 25, p. eabb5813, Jun. 2020, doi: 10.1126/sciadv.abb5813.

Figure 1

Click here to access/download;Figure;Figure1.pdf

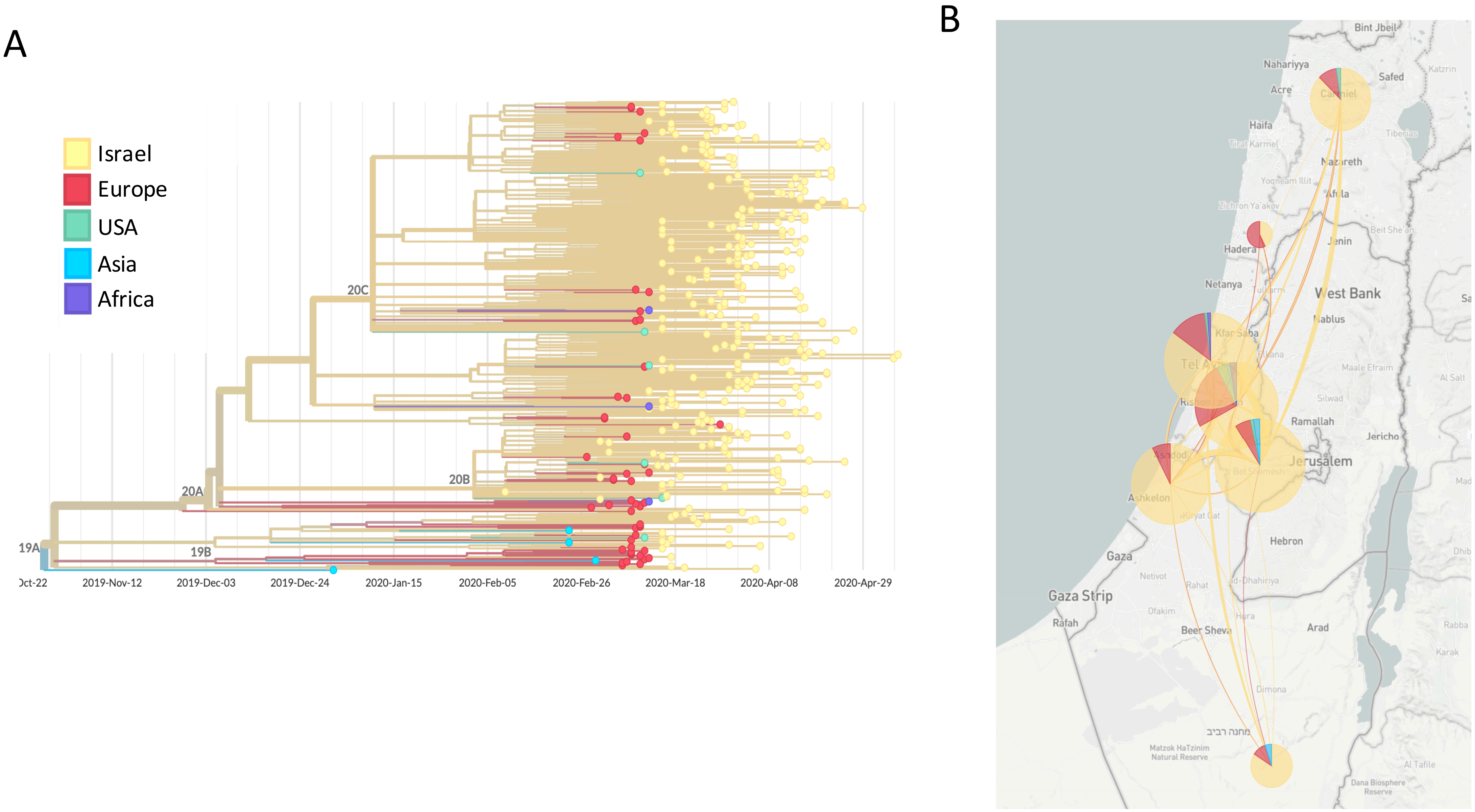


Figure 2

Click here to access/download;Figure;Figure2.pdf

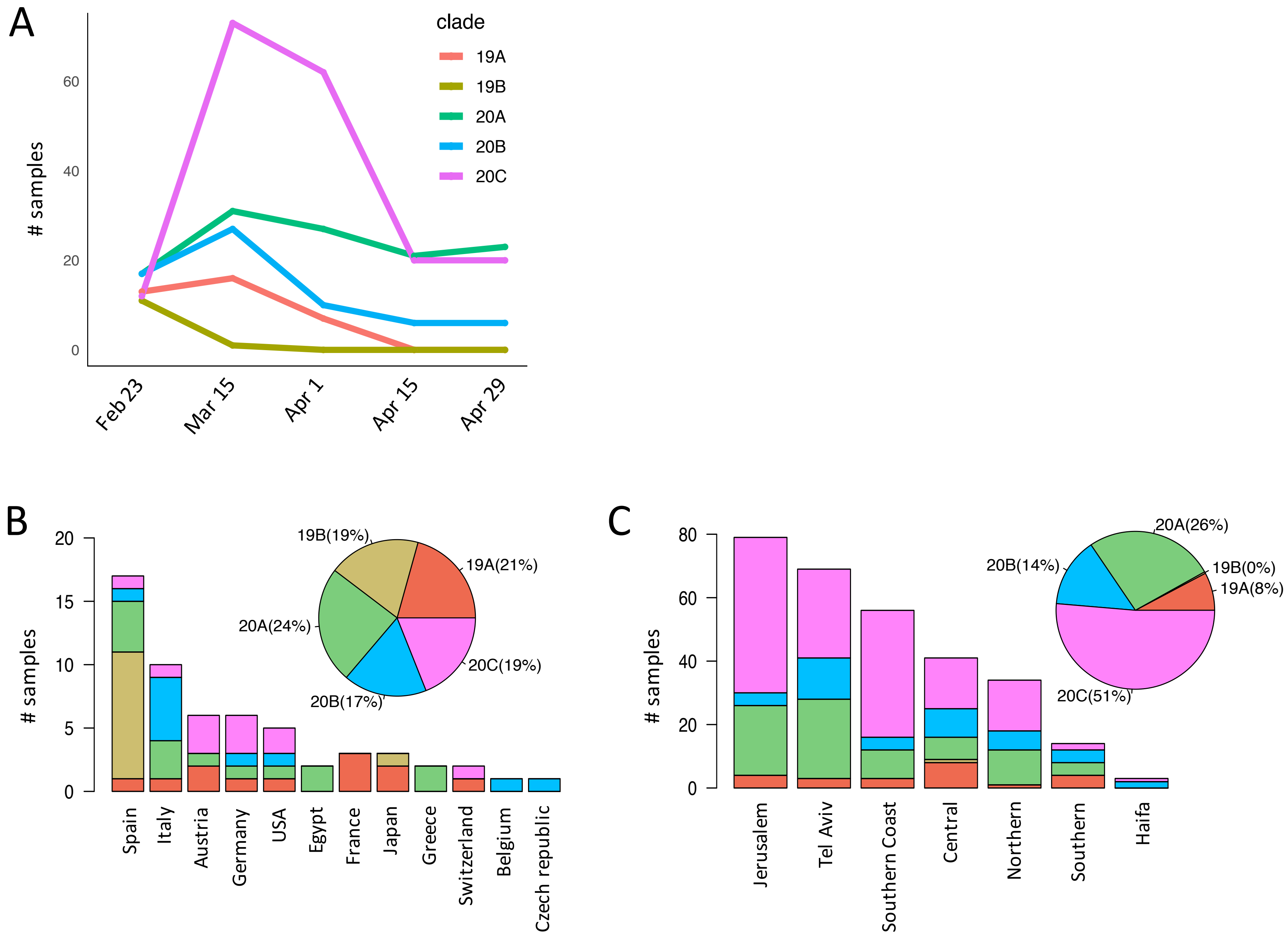
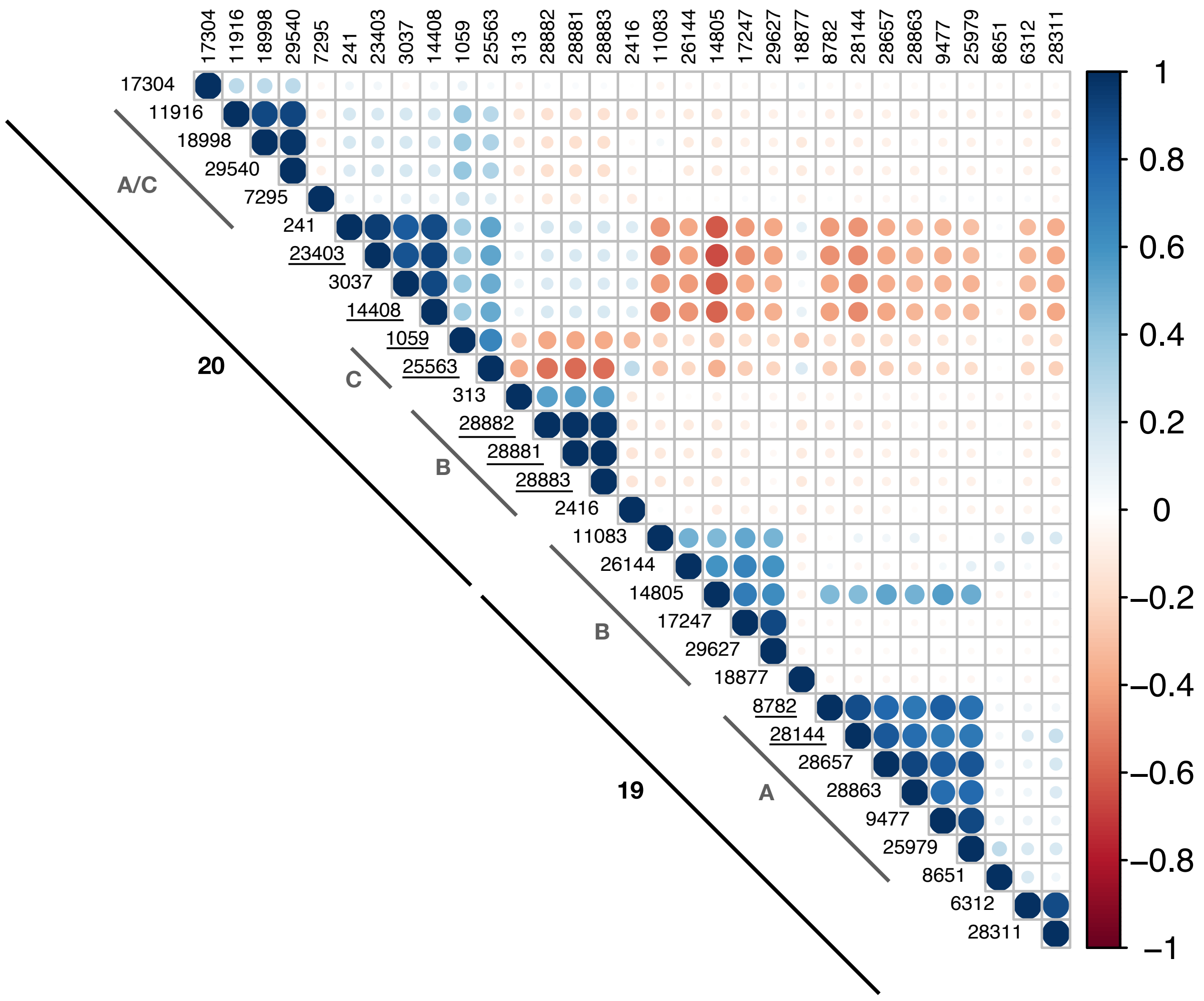


Figure 3

Click here to access/download;Figure;Figure3.pdf

A



B

position	gene	%	Ref	Alt	R/S	AA group
241	5'UTR	83.0	c	t		
313	Nsp1	6.5	c	t	S	
1059	Nsp2	46.0	c	t	R	Aliphatic (I) , Hydroxylated (T)
2416	Nsp3	8.9	c	t	S	
3037	Nsp3	85.1	c	t	S	
6312	Nsp3	2.3	c	a	R	Basic (K) , Hydroxylated (T)
7295	Nsp4	4.4	g	t	R	Hydroxylated (S) , Aliphatic (A)
8651	Nsp4	2.1	a	c	R	Aliphatic (M/L)
8782	Nsp4	3.1	c	t	S	
9477	Nsp4	2.1	t	a	R	Aromatic (F/Y)
11083	Nsp6	7.8	g	t	R	Aromatic (F) , Aliphatic (L)
11916	Nsp7	15.9	c	t	R	Aliphatic (L) , Hydroxylated (S)
14408	Nsp12	86.9	c	t	S	
14805	Nsp12	5.7	c	t	R	Aliphatic (I) , Hydroxylated (T)
17247	Nsp13	2.9	t	c	S	
17304	Nsp13	2.3	c	t	S	
18998	Nsp14	14.4	c	t	R	Aliphatic (A/V)
23403	Spike	85.1	a	g	R	Acidic (D) , Aliphatic (G)
25563	Orf3a	63.4	g	t	R	Aminic (Q) , Basic (H)
25979	Orf3a	2.3	g	t	R	Aliphatic (V/G)
26144	Orf3a	3.9	g	t		
28144	Orf8	3.7	t	c	R	Hydroxylated (S) , Aliphatic (L)
28311	Nucap	2.6	c	t	R	Aliphatic (L) , Proline (P)
28657	Nucap	2.6	c	t	S	
28863	Nucap	2.6	c	t	R	Aliphatic (L) , Hydroxylated (S)
28881	Nucap	14.9	g	a	R	Basic (K/R)
28882	Nucap	14.9	g	a	R	Basic (K/R)
28883	Nucap	15.1	g	c	R	Basic (R) , Aliphatic (G)
29627	Orf10	2.6	c	t	R	Cysteine (C) , Basic (R)