Whole genome and phylogenetic analysis of SARS-CoV-2 strains from the index and early patients with COVID-19 in Dubai, United Arab Emirates, 29 January to 18 March 2020

Ahmad Abou Tayoun[1,2 ‡], Tom Loney[2 ‡], Hamda Khansaheb[3], Sathishkumar Ramaswamy[1], Divinlal Harilal[1], Zulfa Omar Deesi[4], Rupa Murthy Varghese[4], Hanan Al Suwaidi[2], Abdulmajeed Alkhajeh[3], Laila Mohamed AlDabal[5], Mohammed Uddin[2,6], Rifat Hamoudi[7], Rabih Halwani[7], Abiola Senok[2], Qutayba Hamid[8], Norbert Nowotny[2,9*], Alawi Alsheikh-Ali[2*]

[1]Al Jalila Genomics Center, Al Jalila Children's Hospital, Dubai, United Arab Emirates.

[2]College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, United Arab Emirates.

[3]Medical Education & Research Department, Dubai Health Authority, Dubai, United Arab Emirates.

[4]Microbiology and Infection Control Unit, Pathology and Genetics Department, Latifa Women and Children Hospital, Dubai Health Authority, Dubai, United Arab Emirates.

[5]Medical Affairs Department, Rashid Hospital, Dubai Health Authority, Dubai, United Arab Emirates.

[6]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON, Canada

[7]Clinical Sciences Department, College of Medicine, University of Sharjah, Sharjah, United Arab Emirates.

[8]College of Medicine, University of Sharjah, Sharjah, United Arab Emirates.

[9]Institute of Virology, University of Veterinary Medicine Vienna, Vienna, Austria.

*Corresponding authors: Norbert Nowotny, Norbert.Nowotny@vetmeduni.ac.at; Alawi Alsheikh-Ali, Alawi.Alsheikhali@mbru.ac.ae.

‡ Joint first authorship.

Author Contributions

AAA, AT, NN, and TL conceived the study and drafted the protocol. All authors provided critical input into the protocol. AAA, AAK, and HK coordinated the ethical approval and sample retrieval. RV and ZD conducted the RT-qPCR analysis. AT, SR, and DH performed the whole genome sequencing and phylogenetic analysis. HK performed data extraction from the medical records and TL completed data analysis for the manuscript. AT and TL drafted the manuscript, AAA and NN refined it before AAK, AS, DH HA, HK, MU, QH, RH, RHA, RV, SR, and ZD, provided comments and feedback on the first draft. All authors read and approved the final version of the manuscript.

Competing Interest Statement: Authors have no conflicts of interest to disclose.

This file includes:

Main Text
Figures 1 to 2

Word Count: 2,356

Reference Count: 8

Abstract

Whole genome and phylogenetic analysis of SARS-CoV-2 strains from the index and early patients with COVID-19 in Dubai (United Arab Emirates; UAE) showed multiple spatiotemporal introductions from Asia, Europe, and the Middle East. At least one type A introduction and several type B sub-clusters can be distinguished during the early phases of the outbreak in the UAE. Our findings can be used to further understand the global transmission network of SARS-CoV-2.

Introduction

In December 2019, several cases of a new respiratory illness (now called COVID-19) were reported in the city of Wuhan (Hubei Province, China) and in January 2020 it was confirmed these infections were caused by a novel coronavirus subsequently named SARS-CoV-2 [1-2]. On 12 March 2020, the ongoing SARS-CoV-2 outbreak was declared as a pandemic by the World Health Organization (WHO) [3]. To date (03 May 2020), there have been over 3.4 million laboratory-confirmed cases of COVID-19 and more than ~244,000 deaths in 187 countries [4].

Dubai in the United Arab Emirates (UAE) is a cosmopolitan metropolis that has become a popular tourist destination and home to one of the busiest airport hubs in the world connecting the east with the west [4]. Currently, the UAE has reported 14,163 confirmed cases and 126 deaths associated with COVID-19 (0.9% case fatality; 03 May 2020) [5]. In view of Dubai's important tourism and travel connections, we attempted to characterize the full-genome sequence of SARS-CoV-2 strains from the index and early patients with COVID-19 in Dubai to gain a deeper understanding of the molecular epidemiology of the outbreak in Asia, Europe, and the Middle East.

Results

Phylogenetic analysis

To understand early viral transmission in Dubai in the global context, we performed phylogenetic analysis using the 25 novel viral genomes we sequenced from early patients in the UAE (**Supplemental Table 1**) in this study (**Methods**) along with 157 largely complete SARS-CoV-2 genomes deposited in GISAID from different countries between December 2019 and early March 2020 (**Supplemental Table 2)** [6]. We used an open-source bioinformatics package from Nextstrain.org which consists of two main tools, 'Augur' and 'Auspice" [8]. The 'Augur' package performs multiple sequence alignment and infers a maximum likelihood phylogenetic tree decorated with input data labels such as date, origin, and sample name. 'Auspice' is a visualization tool for data presentation (see **Figures 1 and 2**).

The 157 non-UAE viral genomes were distributed among three central virus types (A, B, and C) as previously described [6] (**Figure 1**).

In the UAE, only one viral genome obtained from L5630, a family member of the early Chinese index patient, belonged to the A type. Although we did not obtain full viral genome sequences from the other members of that Chinese family, we expect that all had a similar strain to L5630. Interestingly, our data do not suggest any transmission of this A type virus at least among the earliest patients (**Figures 1** and **2**) included in this study which is consistent with the reported early detection and isolation of this family. This finding also suggests a secondary source for the ongoing local transmission.

The remaining 24 viral genomes belonged to the B type, though were distributed over two major sub-clusters and possibly four other earlier introductions (**Figures 1** and **2**). The latter introduction(s) consisted of four Asians, two residents (L4280, L6599) and two tourists (L4184, L9766), and one Czech resident (L1014) working as an airline cabin crew with travel history to Austria. Consistent with its Asian predominance and the fewer mutations relative to the Wuhan reference genome (**Figure 2**), several viral strains submitted in Asia clustered very closely to this group (**Figure 1**). L4280 was the first sequenced patient without travel history and became infected after transporting a work colleague with severe/critical symptoms to hospital (L0826; not sequenced; no travel history). Patient L0826 reported symptom onset on 22 January suggesting that community-based transmission started in the UAE in early-to-mid January. L6599 was an Indian expatriate living with three other expatriates from the Philippines and Sri Lanka (L3715, L2771, L8480), for whom full viral genome sequences could not be obtained but are likely to belong to this sub-cluster.

One of the major B sub-clusters consisted of five cases with travel history to Iran (L2409, L6627, L0904, L0184, and L4682) and one Indian resident (L0231) and one Indian tourist (L0068) (**Figure 2)**. The origin of this sub-cluster is most likely Iranian. Based on mutation divergence within this sub-cluster (**Figure 2**), it is highly likely that L2409 is the origin of virus transmission in all but L4682 who interestingly was the only individual with severe clinical presentation in this group. Furthermore, a SARS-CoV-2 genome submitted by the University of Sydney (GISAID ID: EPI_ISL_412975) on 28 February 2020 differed by only two mutations from that of L2409 and both this Iranian male tourist and the Australian male had a recent travel history to Iran. Individuals with travel history to Iran around this time frame (L8386, L6867, and L3280) are also very likely to belong to this sub-cluster.

The other major B sub-cluster consisted of 12 individuals with extensive travel history to Europe (**Figure 1**, **Supplemental Figure 1** and **Supplemental Table 1**) highlighting a very likely strong European origin of transmission in this sub-cluster as supported by close identity to at least two European viral strains (**Figure 1**). Of note, a SARS-CoV-2 genome submitted from Mexico (GISAID ID: EPI_ISL_412972) was 100% identical to that of an Italian expatriate working in the UAE (L0881), while another submitted in Germany (GISAID ID: EPI_ISL_412912) differed by a single mutation. All three individuals had a recent travel history to Italy. Interestingly, based on mutation divergence, the viral strains sampled from L3779, L0879, and L9768 are direct descendants of that from L0881 (**Figure** 2). Of note, L1758, L0484, and L2185 were infected with the exact same viral strain (**Figure 2**) suggesting a common direct source of transmission.

In aggregate, we identified 70 variants relative to the reference GenBank SARS-CoV-2 sequence NC_045512.2. The majority of these variants were missense (n=41) with the most frequent nucleotide change being C>T (n=33), and more than half (38/70) were localized in the *ORF1ab* gene (**Supplemental Table 3**). Notably, 17 out of the 70 variants were novel as they were not identified in the Chinese National Center for Bioinformation Database (https://bigd.big.ac.cn/ncov/variation/annotation; last accessed April 29, 2020). Of those, 12 were coding missense variants distributed across the *ORF1ab* (n=8), *S* (n=1), *ORF7a* (n=1), and the *N* (n=2) genes. Assessing the biological significance of these variants is critical for determining the efficacy of any future therapies and/or vaccines.

Discussion

Our findings suggest multiple spatiotemporal introductions of SARS-CoV-2 into the UAE from Asia, Europe, and the Middle East. The new mutations and novel amino acid changes found in the viruses identified in Dubai warrant further investigation to explore whether they influence viral characteristics, especially pathogenicity, or provide important information for vaccine development. Limitations include the sociodemographic and travel history data available in the medical records and the inability to conduct full whole genome sequencing on half of the samples due to low viral

load, although we were able to deduce the origin of transmission in those individuals based on travel history. Regardless, this study contributes important molecular epidemiological data that can be used to further understand the global transmission network of SARS-CoV-2.

Materials and Methods

Patients and whole viral genome sequencing

This study was approved by the Dubai Scientific Research Ethics Committee - Dubai Health Authority (approval number #DSREC-04/2020_02). The electronic medical records of all patients with laboratory confirmed SARS-CoV-2 from 29 January to 18 March 2020 (N=49) were reviewed and sociodemographic and clinical data extracted using the WHO case report form. In brief, the index patient with COVID-19 in the UAE was a female Chinese tourist (aged 63 years) travelling from Wuhan with her daughter-in-law and two grandchildren to visit her son working in Dubai. The Chinese family arrived in Dubai on 16 January 2020 and the grandmother, son, daughter-in-law, and one grandchild all tested positive on the 29 January 2020 (**Supplemental Table 1**). Over the next seven weeks, there were multiple new travel-related introductions of the virus into the UAE from European and Iranian tourists and residents, and a cluster of cases amongst residents with no travel history (**Supplemental Table 1**). Of the 49 COVID-19 patients in Dubai, 22 had a travel history from Europe including Austria, France, Germany, Italy, Ireland, Norway, and the United Kingdom (**Supplemental Table 1**). Majority of patients (88%) were asymptomatic or had mild symptoms and only four required intensive care with invasive ventilation.

All 49 patients tested positive for SARS-CoV-2 by RT-qPCR in the central Dubai Health Authority virology laboratory where viral RNA was extracted from nasopharyngeal swabs following the QIAamp Viral RNA Mini or the EZ1 DSP Virus Kits (Qiagen, Hilden, Germany). Extracted genomic RNA from all samples were then subjected to shotgun metagenomic sequencing using the Illumina (San Diego, CA, USA) next-generation sequencing (NGS) platform at Al Jalila Children's Specialty Hospital, Dubai, UAE. Approximately 1000ng of input RNA from each patient sample was used to prepare sequencing-ready libraries using the TruSeq Stranded Total RNA Library Prep Gold kit from Illumina (San Diego, CA, USA) following manufacturer's instructions. Libraries were sequenced using the Illumina SP Reagent kit (2 X 150 cycles) and the NovaSeq platform (San Diego, CA, USA). Sample L5630 underwent a target enrichment approach where double stranded DNA (synthesized using the QuantiTect Reverse Transcription Kit from Qiagen, Hilden, Germany) was amplified using 26 overlapping primer sets covering most of SARS-CoV-2 genome [7]. PCR products were then sheared by ultra-sonication (Covaris LE220-plus series, MA, USA) and prepared for sequencing using the SureSelectXT Library Preparation kit (Agilent, CA, USA). This library was sequenced using the MiSeq Micro Reagent Kit, V2 (2 X 150 cycles) on the MiSeq platform (Illumina, San Diego, CA, USA). High quality sequencing reads, obtained through shotgun or targeted enrichment, were trimmed and then aligned to the reference SARS-CoV-2 genome from Wuhan, China (GenBank accession number: NC_045512.2) using a custom-made bioinformatics pipeline (**Supplemental Figure 1**). Assembled genomes with at least 20X average coverage across most nucleotide positions (56-29,797; [6]) were used for subsequent phylogenetic analysis (**Supplemental Table 4**). A total of 25 viral genomes (24 by shotgun and 1 by target enrichment) met this inclusion criterion and were submitted to the Global Initiative on Sharing All Influenza Data (GISAID) database under accession numbers listed in **Supplemental Table 2**. The 25 SARS-CoV-2 genomes were obtained from cases with disease onsets in late January (n=1), early February (n=1), late February (n=6), early March (n=8), and late March (n=9).

References

1. Ashour HM, Elkhatib WF, Rahman MM, Elshabrawy HA. Insights into the Recent 2019 Novel Coronavirus (SARS-CoV-2) in Light of Past Human Coronavirus Outbreaks. *Pathogens*. 9(3):E186. doi: 10.3390/pathogens9030186 (2020).

2. Uddin M, Mustafa F, Rizvi TA, Loney T, Al Suwaidi H, Al Marzouqi A, et al. SARS-CoV-2/COVID-19: Viral Genomics, Epidemiology, Vaccines, and Therapeutic Interventions. Viruses (in revision); Preprint, 2020040005 (doi: 10.20944/preprints202004.0005.v1) (2020).
3. World Health Organization (WHO). Coronavirus disease 2019 (COVID-19) Situation Report – 52. Geneva: WHO; 12 Mar 2020. Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200312-sitrep-52-covid-19.pdf?sfvrsn=e2bfc9c0_4
4. Johns Hopkins Center for Systems Sciences and Engineering. COVID19 Dashboard. Available from: https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
5. Loney T, Aw TC, Handysides DG, Ali R, Blair I, Grivna M, et al. An analysis of the health status of the United Arab Emirates: the 'Big 4' public health issues. Glob Health Action. 2013;6:20100.
6. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A*. 117(17):9241-9243 (2020).
7. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 579(7798):265-269 (2020).
8. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 34(23):4121-4123 (2018).

Figures

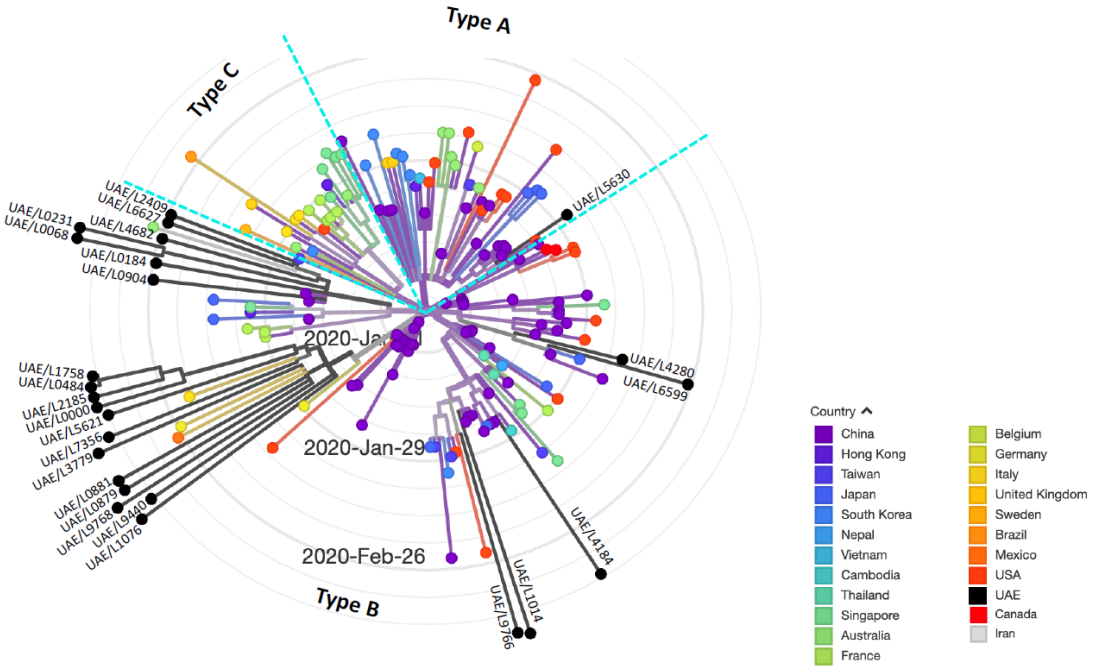

Figure 1. A radial phylogenetic tree generated using 'Augur' and 'Auspice' which are open-source tools (nexstrain.org) used for multiple sequence alignment, maximum likelihood phylogenetic tree construction, and data visualization [see reference 8]. A total of 182 SARS-CoV-2 genomes (157 obtained from GISAID and 25 genomes in this study) were used for this analysis. Filled circles represent viral strains color-coded by country of origin. Branch length for each strain represents the date for each sample obtained from GISAID (n=157), and dates of SARS-CoV-2 test results

confirmation for the UAE samples (n=25). Dashed black lines mark the boundaries of SARS-CoV-2 Types A, B, and C as previously described [6]. With the exception of L5630 which is an A type, all other early UAE strains belonged to the B type (**Figure 2** and main text).

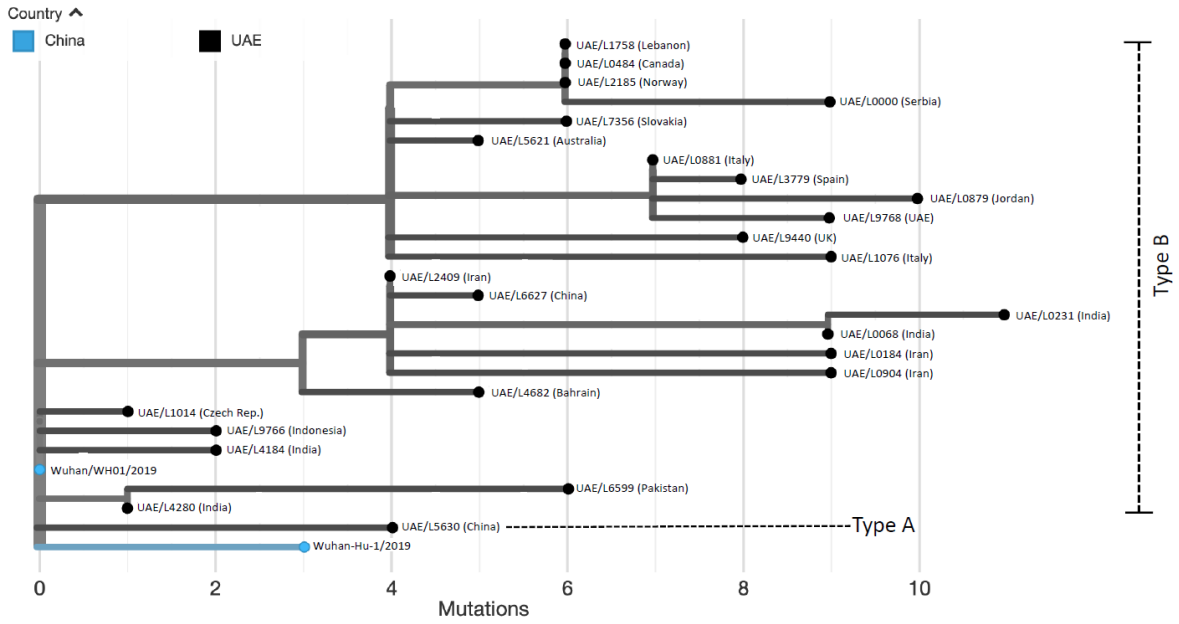

Figure 2. A rectangular phylogenetic tree including all 25 UAE viral strains with full genomes. Analysis was performed using open-source 'Augur' and 'Auspice' [8] for multiple sequence alignment, maximum likelihood phylogenetic tree construction, and data visualization. The two Wuhan genomes (Wuhan-Hu-1/2019, GISAID ID: EPI_ISL_402125 and Wuhan/WH01/2019, GISAID ID: EPI_ISL_406798) were used as reference genomes (blue filled circles). UAE viral strains (black filled circles) were labeled with sample ID and Nationality (in brackets). Branch lengths mark divergence from the reference Wuhan SARS-CoV-2 genome (GenBank accession number: NC_045512.2) in mutations numbers. Dashed black lines mark the UAE strains with SARS-CoV-2 Types A (only L5630) or B (all other strains). The B strains distribute over two major sub-clusters and 4 other minor introductions (see main text). CZ = Czech Republic, UAE = United Arab Emirates.

Supplementary materials for this manuscript include the following:

Supplemental Tables 1 to 4
Supplemental Figure 1

Supplemental Figure 1. Analysis pipeline starting with sequencing raw data to phylogenetic tree construction and visualization. Input and output file types for each programme are shown in brackets. R1, sequencing read 1; R2, sequencing read 2.