SARS-CoV-2 Whole Genome Amplification and Sequencing for Effective Population-Based Surveillance and Control of Viral Transmission

Divinlal Harilal[1‡], Sathishkumar Ramaswamy[1‡], Tom Loney[2], Hanan Al Suwaidi[2], Hamda Khansaheb[3], Abdulmajeed Alkhaja[3], Rupa Varghese[4], Zulfa Deesi[4], Norbert Nowotny[2,5], Alawi Alsheikh-Ali[2], Ahmad Abou Tayoun[1,2*]

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

*Corresponding Author: Ahmad Abou Tayoun, Ahmad.Tayoun@ajch.ae

[‡]These authors contributed equally to this work.

Abstract

Background

With the gradual reopening of economies and resumption of social life, robust surveillance mechanisms should be implemented to control the ongoing COVID-19 pandemic. Unlike RT-qPCR, SARS-CoV-2 whole genome sequencing (cWGS) has the added advantage of identifying cryptic origins of the virus, and the extent of community-based transmissions versus new viral introductions, which can in turn influence public health policy decisions. However, the practical and cost considerations of cWGS should be addressed before it is widely implemented.

Methods

We performed shotgun transcriptome sequencing using RNA extracted from nasopharyngeal swabs of patients with COVID-19, and compared it to targeted SARS-CoV-2 genome amplification and sequencing with respect to virus detection, scalability, and cost-effectiveness. To track virus origin, we used open-source multiple sequence alignment and phylogenetic tools to compare the assembled SARS-CoV-2 genomes to publicly available sequences.

Results

We found considerable improvement in whole genome sequencing data quality and viral detection using amplicon-based target enrichment of SARS-CoV-2. With enrichment, more than 99% of the sequencing reads mapped to the viral genome compared to an average of 0.63% without enrichment. Consequently, an increase in genome coverage was obtained using substantially less sequencing data, enabling higher scalability and sizable cost reductions. We also demonstrated how SARS-CoV-2 genome sequences can be used to determine their possible origin through phylogenetic analysis including other viral strains.

Conclusions

SARS-CoV-2 whole genome sequencing is a practical, cost-effective, and powerful approach for population-based surveillance and control of viral transmission in the next phase of the COVID-19 pandemic.

The COVID-19 pandemic continues to inflict devastating human life losses (1), and has imposed major social changes and costly global economic shut downs (2). With the accumulating financial burdens and unemployment rates, several governments are sketching out plans for slowly re-opening the economy and reviving social life and economic activity. However, robust population-based surveillance systems are essential to track viral transmission during the re-opening process.

While reverse transcriptase real time PCR (RT-qPCR) targeting SARS-CoV-2 RNA can be effective in identifying infected individuals for isolation and contact tracing, it is not useful in determining which viral strains are circulating in the community: autochthonous versus imported ones. It is important to know the origin of the strains, which in turn influences public health policy decisions. In addition, it is vital to identify super-spreader events as they can be influenced by the virus strain (3). SARS-CoV-2 whole genome sequencing (cWGS), on the other hand, can detect the virus and can delineate its origins through phylogenetic analysis (4, 5) in combination with other local and international viral strains, especially given the accumulation of thousands of viral sequences from countries all over the world (6) (**Figure 1**). However, practical considerations, such as cost, scalability, and data storage, should first be investigated to assess the feasibility of implementing cWGS as a population-based surveillance tool.

Materials and Methods

Human subjects and ethics approval

All patients had laboratory-confirmed COVID-19 based on positive RT-qPCR assay for SARS-CoV-2 in the Centralized Dubai Health Authority (DHA) virology laboratory. This study was approved by the Dubai Scientific Research Ethics Committee - Dubai Health Authority (approval number #DSREC-04/2020_02).

RNA extraction and SARS-CoV-2 detection

Viral RNA was extracted from nasopharyngeal swabs of patients with COVID-19 using the EZ1 DSP Virus Kit (Qiagen), optimized for viral and bacterial nucleic acids extractions from

human specimens using magnetic bead technology. SARS-CoV-2 positive results were confirmed using a RT-qPCR assay, originally designed by the US Centers for Disease Control and Prevention (CDC), which is currently provided by Integrated DNA Technologies (IDT). This assay consists of oligonucleotide primers and dual-labelled hydrolysis (TaqMan®) probes (5'FAM/3'Black Hole Quencher) specific for two regions (N1 and N2) of the virus nucleocapsid (N) gene. An additional primer/probe set is also included to detect the human RNase P gene (RP) as an extraction control. The reverse transcription and amplification steps are performed using the TaqPath™ 1-Step RT-qPCR Master Mix (ThermoFisher) following manufacturer's instructions. A sample was considered positive if the cycle threshold (Ct) values were less than 40 for each of the SARS-CoV-2 targets (N1 and N2) and the extraction control (RP). To estimate the viral load relative to human RNA, we calculated the ΔCt value for each target as follows: $\Delta Ct = Ct_{Nn} - Ct_{RP}$, where N_n is either N1 or N2. The inverse of the average N1 and N2 target ΔCt values (i.e. – ΔCt) was used to estimate the relative viral load which is inversely correlated with Ct value.

Shotgun transcriptome SARS-CoV-2 sequencing

RNA libraries from all samples were prepared for shotgun transcriptomic sequencing using the TruSeq Stranded Total RNA Library kit from Illumina, following manufacturer's instructions. RNA specific fluorescent dye is used to quantify extracted RNA using the Qubit RNA XR assay kit and the Qubit Flourometer system (ThermoFisher). Then, 1µg of input RNA from each patient sample was depleted for human ribosomal RNA, and the remaining RNA underwent fragmentation, reverse transcription (using the SuperScript II Reverse Transcriptase Kit from Invitrogen), adaptor ligation, and amplification. Libraries were then sequenced using the NovaSeq SP Reagent kit (2 X 150 cycles) from Illumina.

Targeted amplification and sequencing of SARS-CoV-2 genome

RNA extracted (approximately 1µg) from patient nasopharyngeal swabs was used for double stranded cDNA synthesis using the QuantiTect Reverse Transcription Kit (Qiagen) according to manufacturer's protocol. This cDNA was then evenly distributed into 26 PCR reactions for SARS-CoV-2 whole genome amplification using 26 overlapping primer sets covering most of its genome (**Figure 2A** and online **Supplemental Table 1**). The SARS-CoV-2 primer sets used in this study were modified from Wu *et al* (7) by adding M13 tails to enable sequencing by

Sanger, if needed (online **Supplemental Table 1**). PCR amplification was performed using the Platinum™ SuperFi™ PCR Master Mix (ThermoFisher) and a thermal protocol consisting of an initial denaturation at 98°C for 60 seconds, followed by 27 cycles of denaturation (98°C for 17 seconds), annealing (57°C for 20 seconds), and extension (72°C for 1 minute and 53 seconds). A final extension at 72°C for 10 minutes was applied before retrieving the final PCR products. Amplification was confirmed by running 2µl from each reaction on a 2% agarose gel.

All PCR products were then purified using Agencourt AMPure XP beads (Beckman Coulter), quantified by NanoDrop (ThermoFisher), diluted to the same concentration, and then pooled into one tube for next steps.

A minimum of 200-800ng of the pooled PCR products in 55µl were then sheared by ultra-sonication (Covaris LE220-plus series) to generate a target fragment size of 250-750bp using the following parameters: 20% duty factor, peak power of 150 watts, 900 cycles per burst, 320 seconds treatment time, an average power of 30 watts, and 20°C bath temperature. Target fragmentation were confirmed by the TapeStation automated electrophoresis system TapeStation (Agilent) (**Figure 2A**). Subsequently, the fragmented product was purified and then processed to generate sequencing-ready libraries using the SureSelectXT Library Preparation kit (Agilent) following manufacturer instructions. Indexed libraries from multiple patients were pooled and sequenced (2 X 150 cycles) using the MiSeq or the NovaSeq systems (Illumina). A step-by-step SARS-CoV-2 target enrichment and sequencing protocol is provided in the online **Appendix**.

Bioinformatics analysis and SARS-CoV-2 genome assembly

Demultiplexed Fastq reads, obtained through shotgun or target enrichment sequencing, were generated from raw sequencing base call files using BCL2Fastq v2.20.0, and then mapped to the reference Wuhan genome (GenBank accession number: NC_045512.2) by Burrow-Wheeler Aligner, BWA v0.7.17. Alignment statistics, such as coverage and mapped reads, were generated using Picard 2.18.17. Variant calling was performed by GATK v3.8-1-0, and was followed by SARS-CoV-2 genome assembly using BCFtools v.1.3.1 (**Figure 2B**).

All tools used in this study are freely accessible. For laboratories without bioinformatics support, several publicly accessible, end-to-end bioinformatics pipelines (INSaFlu and

Genome Detective) (8,9), composed of the above tools, can be used to generate viral sequences from raw Fastq data.

For downstream analysis, a general quality control metric was implemented to ensure assembled SARS-CoV-2 genomes had at least 20X average coverage (sequencing reads >Q30) across most nucleotide positions (56-29,797).

For target enrichment and shotgun sequencing comparisons in **Table 1**, we used data from 7 samples (UAE/P1/2020, L0287, L1189, L4711, L5857, L6841, L9119) generated using both methods. Data from patient L5630 were not included in this analysis since, unlike the above 7 samples, appreciably more sequencing data was allocated for the target enriched sample in this patient (online **Supplemental Table 2**) which can overestimate the efficiency of the target enrichment protocol.

All new SARS-CoV-2 sequences (n=7) generated in this study were submitted to GISAID (Global Initiative on Sharing All Influenza Data) under accession IDs: EPI_ISL_463740 and EPI_ISL_469276 to EPI_ISL_469281.

Phylogenetic analysis

We used Nexstrain (10), which consists of Augur v6.4.3 pipeline for multiple sequence alignment (*MAFFT* v7.455) (11) and phylogenetic tree construction (*IQtree* v1.6.12) (12). Tree topology was assessed using the fast bootstrapping function with 1,000 replicates. Tree visualization and annotations were performed in *FigTree* v1.4.4 (13).

Results

SARS-CoV-2 whole genome sequencing

Shotgun transcriptome sequencing was used to fully sequence SARS-CoV-2 RNA extracted from patients (n=17) who tested positive for the virus (4). Analysis of the sequencing data showed that this approach required, on average, 4.71Gb of data per sample yielding 31.7 million total reads, of which approximately 0.63% of the reads (~199,000 reads) mapped to the SARS-CoV-2 genome with an average coverage of ~173x (**Table 1**). This is attributed to the fact that most of the shotgun data (~99%) is allocated to the human transcriptome while a minority of the reads align to the SARS-CoV-2 genome (**Table 1**). In addition to cost and storage considerations discussed below, this approach is not highly sensitive for detecting

SARS-CoV-2 genomes in general and specifically in samples with low viral abundance. In fact, despite high viral abundance relative to human RNA, most samples had less than 100x sequencing coverage across the SARS-CoV-2 genome. Samples with seemingly very low viral loads failed to yield full SARS-CoV-2 genome sequence using this approach (**Table 1** and **Figure 3A**).

To enrich viral sequences and minimize sequencing cost and data storage issues addressed below, we describe an alternative approach where the entire SARS-CoV-2 genome is first amplified using 26 overlapping primer sets each yielding around 1.5kb long inserts (**Figures 2A**, **3B** and online **Supplemental Table 1**). All inserts were then pooled and fragmented to 250-750bp inserts which were then prepared for short read next generation sequencing (**Figures 2A** and **3C**).

RNA extracted from eight COVID-19 patients, which was first sequenced by shotgun transcriptome (**Table 1** and **Supplemental Table 2**), was sequenced using the enrichment protocol. As expected, we observed significant enhancement in virus detection using this protocol where, on average, 99.3% of the reads now mapped to the SARS-CoV-2 genome leading to tenfold increase in coverage relative to shotgun transcriptome (avg. 440x versus 45.5x, respectively) despite generating two hundred fold less sequencing data (avg. 0.02Gb versus 4.28Gb, respectively, **Table 2** and **Figure 3D**).

Cost, data storage and scalability

On average, 37x coverage per 1Gb of sequencing data was generated using shotgun sequencing (**Table 1**) compared to ~23,000x per 1Gb using target enrichment (**Table 2**) suggesting the latter method is more cost effective and is highly scalable. We calculate the cost of SARS-CoV-2 full genome sequencing to be ~$87 per sample when sequencing 96 samples in a batch at 400x using the target enrichment method. The number of samples in a batch can be doubled (196) while maintaining a low cost (~$104) and a very high coverage of 40,000x per sample (online **Supplemental Table 3**). On the other hand, the cost of sequencing one sample at a much lower coverage (50x) using the shotgun method is $403, while increasing sequencing coverage more than doubled the cost ($1735 at 100x and $1060 at 200x) (online **Supplemental Table 3**). However, using higher throughput sequencing can significantly lower the cost of shotgun sequencing to $232 for 62 samples in a batch at 200x

per sample. Nonetheless, using a similar throughput, the per sample cost of enrichment sequencing is $108 for 196 samples in a batch where each sample receives considerably more coverage (~40,000x) (online **Supplemental Table 3**). Therefore, target enrichment sequencing is still more cost-effective and scalable than shotgun transcriptome sequencing even at higher sequencing throughputs.

Another factor impeding scalability of the shotgun approach is data storage. Even with higher throughput sequencing (NovaSeq SP flowcell), shotgun sequencing requires an allocation of 1TB of data for ~250 sequenced samples. On the other hand, with 1TB of data, a total of around 80,000 samples can be sequenced using the enrichment method and the MiSeq Micro flowcell (online **Supplemental Table 3**). Therefore, long term data storage allocations, and cost, are substantially higher, and perhaps formidable, when using the shotgun sequencing approach.

Genomic surveillance of SARS-CoV-2 origin

To illustrate the utility of SARS-CoV-2 whole genome sequencing, we tracked the origin(s) of the virus in seven patients (UAE/P1/2020, L0287, L1189, L4711, L5857, L6841, L9119) by comparing their assembled sequences to virus strains (n=25) identified during the early phase of the pandemic, between January 29 and March 18 2020, in the UAE (4). All seven patient samples were collected between March 28 and April 5 2020, and are therefore good candidates to determine whether transmissions were community-based due to the previously documented 25 strains or were independent external introductions.

Multiple sequence alignment and phylogenetic analysis (**Figures 2B** and **4**) showed that the new isolates from patients L0287, L4711, L1189, and L5857 clustered with earlier strains of Iranian origin (or clade A3), while that from patient L9119 belonged to the early European (or clade A2a) cluster (**Figure 4**). This information suggests that transmissions for all those five patients were most likely community-based, which we then confirmed from patient medical records where no recent travel history was reported by any of those individuals.

SARS-CoV-2 isolates from patients P1/UAE/2020 and L6841 were, on the other hand, closer to the earliest Asian strains, which are more diverse due to fewer but distinct mutations (**Figure 4**). Hence, with the available sequencing data it is challenging to ascertain whether the P1/UAE/2020 and L6841 transmissions were community-based or due to early

independent introductions. However, patient L6841 did not have any recent travel history before symptoms onset, suggesting the case for community-based transmission related to an Asian strain. On the other hand, travel history in patient P1/UAE/2020 was not known and the corresponding isolate appeared to match closely to five other strains from the United States and Taiwan (**Figure 4**). Therefore, transmission in patient P1/UAE/2020 was unlikely to be community-based from the early 25 strains (4), but rather due to an independent travel-related introduction of the virus.

Discussion

Genomics-based SARS-CoV-2 population-based surveillance is a powerful tool for controlling viral transmission during the next phase of the pandemic. Therefore, it is important to devise efficient methods for SARS-CoV-2 genome sequencing for downstream phylogenetic analysis and virus origin tracking. Towards this goal, we describe a cost-effective, robust, and highly scalable target enrichment sequencing approach, and provide an example to demonstrate its utility in characterizing transmission origin.

Our target enrichment protocol is amplicon-based for which oligonucleotide primers can be easily ordered by any molecular laboratory. Next generation sequencing (NGS) has also become largely accessible to most labs, and in our protocol we show that highly affordable, low throughput sequencers, such as the Illumina MiSeq system, can be used efficiently to sequence up to 96 samples at 400x coverage each at a cost of $87 per sample (**Table 3**). This cost is likely comparable to RT-qPCR testing for the virus. Other low throughput, highly affordable semiconductor sequencers can also be used with this protocol (14).

One possible limitation is the use of ultra-sonication for fragmentation of PCR products after SARS-CoV-2 whole genome amplification. Several labs might lack sonication systems due to accessibility and affordability issues. In such situations, our protocol can be easily modified to use enzymatic fragmentation instead provided by commercial kits, such as the Agilent SureSelect^{QXT} kit. Furthermore, we have added M13 tails to all our primer sets making them amenable to Sanger sequencing for those labs not equipped with NGS. However, with this approach, manual analysis of sequencing data limits scalability of the approach.

Upon sequence generation, the bioinformatics analysis can be performed using open source scripts. Labs without bioinformatics expertise or support can use online tools (INSaFlu and

Genome Detective) (8,9) which can take raw sequencing (Fastq) files to assemble viral genomes, and to perform multiple sequence alignment and phylogenetic analysis for virus origin tracking. In addition, the described approach does not require large data storage or computational investment as shown by our cost, data, and scalability calculations (**Supplemental Table 3**).

Our phylogenetic analysis demonstrates how SARS-CoV-2 genomic sequencing can be used to track origins of virus transmission. However, data should be carefully interpreted, and should be combined with other epidemiological information (such as travel history) to avoid inaccurate conclusions. The major limitation facing genomic-based SARS-CoV-2 surveillance includes the lack of virus sequencing data representing most strains in any country. Nonetheless, SARS-CoV-2 strains are continuously being sequenced by government, private, and academic entities all over the world, and the sequencing data is being shared publicly. This proliferation of sequencing datasets will empower genomic-based surveillance of the virus, and the availability of cost effective sequencing options, like the one described in this study, will contribute to democratizing SARS-CoV-2 sequencing and data sharing.

In summary, we show that SARS-CoV-2 whole genome sequencing is a highly feasible and effective tool for tracking virus transmission. Genomic data can be used to determine community-based versus imported transmissions, which can then inform the most appropriate public health decisions to control the pandemic.

Author Declaration: A version of this paper was previously posted as a preprint on bioRxiv as https://www.biorxiv.org/content/10.1101/2020.06.06.138339v1.

Acknowledgments Authors would like to thank members of the Microbiology Laboratory, Latifa Women and Children Hospital, Dubai Health Authority and Al Jalila Children's Specialty Hospital Genomics Center for supporting SARS-CoV-2 diagnostic testing and for arranging samples used in this study.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 4 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and*

interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved.

S. Ramaswamy, statistical analysis; H. Alsuwaidi, provision of study material or patients; R. Varghese, provision of study material or patients; Z. Deesi, provision of study material or patients.

Authors' Disclosures or Potential Conflicts of Interest: *No authors declared any potential conflicts of interest.*

Role of Sponsor: No sponsor was declared.

References

1. Johns Hopkins Center for Systems Sciences and Engineering. COVID19 Dashboard. https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6 [Accessed 7/14/2020]
2. Uddin M, Mustafa F, Rizvi T, Loney T, Al Suwaidi H, Al-Marzouqi A, Eldin A, et al. SARS-CoV-2/COVID-19: Viral Genomics, Epidemiology, Vaccines, and Therapeutic Interventions. Viruses 2020; 12: 526.
3. Frieden TR, Lee CT. Identifying and interrupting superspreading events — Implications for control of severe acute respiratory syndrome coronavirus 2. Emerg Infect Dis 2020; 26:1059-66.
4. Abou Tayoun A, Loney T, Khansaheb H, Ramaswamy S, Harilal D, Deesi Z, Varghese R, et al. Genomic surveillance and phylogenetic analysis reveal multiple introductions of SARS-CoV-2 into a global travel hub in the Middle East. Preprint at:

https://www.biorxiv.org/content/10.1101/2020.05.06.080606v3 [Accessed 7/14/2020]

5. Butler D, Mozsary C, Meydan C, Dnako D, Foox J, Rosiene J, Shaiber A, et al. Shotgun transcriptome and isothermal profiling of SARS-CoV-2 infection reveals unique host responses, viral diversification, and drug interactions. Preprint at: https://www.biorxiv.org/content/10.1101/2020.04.20.048066v5 [Accessed 7/14/2020]
6. www.nextstrain.org [Accessed 7/14/2020]
7. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020; 579:265-9.
8. Borges V, Pinheiro M, Pechirra P, Guiomar R, Gomes J. InSaFLU: an automated open web-based bioinformatics suite “from reads” for influenza whole-genome-sequencing-based surveillance. Genome Med 2018; 10: 46.
9. Vilsker M, Moosa Y, Nooij S, Fonseca V, Ghysens Y, Dumon K, Pauwels R, et al. Genome detective: An automated system for virus identification from high-throughput sequencing data. Bioinformatics 2019; 35: 871-3.
10. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34:4121-3.
11. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. Nucleic Acids Res 2002;;30:3059-66.
12. Chernomor O, von Haeseler A, Quang Minh B. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Systematic Biology 2016;65:997-1008 (2016).
13. Rambaut. A. FigTree 1.4.2 Software. Institute of Evolutionary Biology, Univ. Edinburg.
14. Abou Tayoun A, Tunkey, C, Pugh T, Ross T, Shah M, Lee C, Harkins T, et al. A comprehensive assay for CFTR mutational analysis using next-generation sequencing. Clin Chem 2013; 59:1481-8.

Table 1. RT-PCR and transcriptome sequencing statistics for COVID-19 patients

Sample ID	RT-PCR	Shotgun Transcriptome Sequencing data				
	Viral RNA Abundance (-ΔCt)	Data size (Gb)	Total Reads	Reads Aligned*	% Reads aligned*	Mean Coverage*
L8205**	-0.995	4.80	32,117,004	124,373	0.38	10.13
L4280	3.852	5.10	33,921,438	69,460	0.20	20.94
L0826**	-12.065	3.21	21,383,276	57,143	0.26	2.77
L2771**	-6.196	5.70	38,114,908	131,449	0.34	5.27
L9440	3.258	4.20	28,015,394	107,206	0.38	38.39
L1758	10.741	4.90	32,649,592	937,403	2.87	2106.37
L0000	2.823	5.05	33,700,572	208,432	0.62	43.23
L3779	5.345	4.49	29,912,422	395,560	1.32	320.98
L5630**	-1.648	4.57	30,462,036	70,058	0.23	4.16
L4184	1.635	3.77	25,150,950	215,797	0.86	31.14
L0287**	9.096	5.40	35,991,646	129,076	0.36	30.07
L1189**	7.344	5.70	37,968,228	72,599	0.19	7.85
L4711**	8.307	4.91	32,762,088	140,388	0.43	13.63
L5857**	10.824	3.86	25,757,828	162,345	0.63	49.44
L6841**	7.571	4.96	33,051,790	264,298	0.80	12.85
L9119**	9.221	5.36	35,719,244	97,709	0.27	19.10
UAE/P1/2020**	3.628	4.13	27,542,314	126,756	0.46	227.00
Average	3.691	4.71	31,667,401	198,956	0.63	173.14

*Statistics with respect to the SARS-CoV-2 genome

**Sequencing Data generated in this study. For the remaining samples, data were generated in the Abou Tayoun et al study

Table 2. Comparison of sequencing statistics between target enrichment and shotgun transcriptome for COVID-19 patients*

	Target Enrichment	**Shotgun Transcriptome**
Total Reads	112,165	32,684,734
Reads Aligned**	111,046	141,882
% of Reads Aligned**	99.3	0.45
Data Size (Gb)**	0.02	4.28
Mean Target Coverage	439.51	45.53
% >20x**	100	48

*Sequencing data per sample (n=7) details in online **Supplemental Table 2**, and explained in Methods. **Statistics with respect to the SARS-CoV-2 genome

Figure Legends

Figure 1. SARS-CoV-2 whole genome sequencing-based surveillance. A schematic illustrating how SARS-CoV-2 whole genome sequencing (cWGS) can be used as a surveillance tool to uncover community-based versus international/travel-related introductions. Mutations are represented by colored dots or circles on SARS-CoV-2 genomes (black bars) within each patient with COVID-19. A population of viral genomes in a community can be used as a reference set (circled blue) for future analysis when new cases (circled orange) emerge. Two scenarios are represented for the new case: the first represents community transmission while the second represents external introduction. The strain representing community transmission has two mutations, one of which (blue) has been identified in a strain from a previous patient in this community, while the second is a new mutation (brick red), arising as part of the virus evolution. The strain with a single novel mutation (green) not seen previously in this population represents a new introduction.

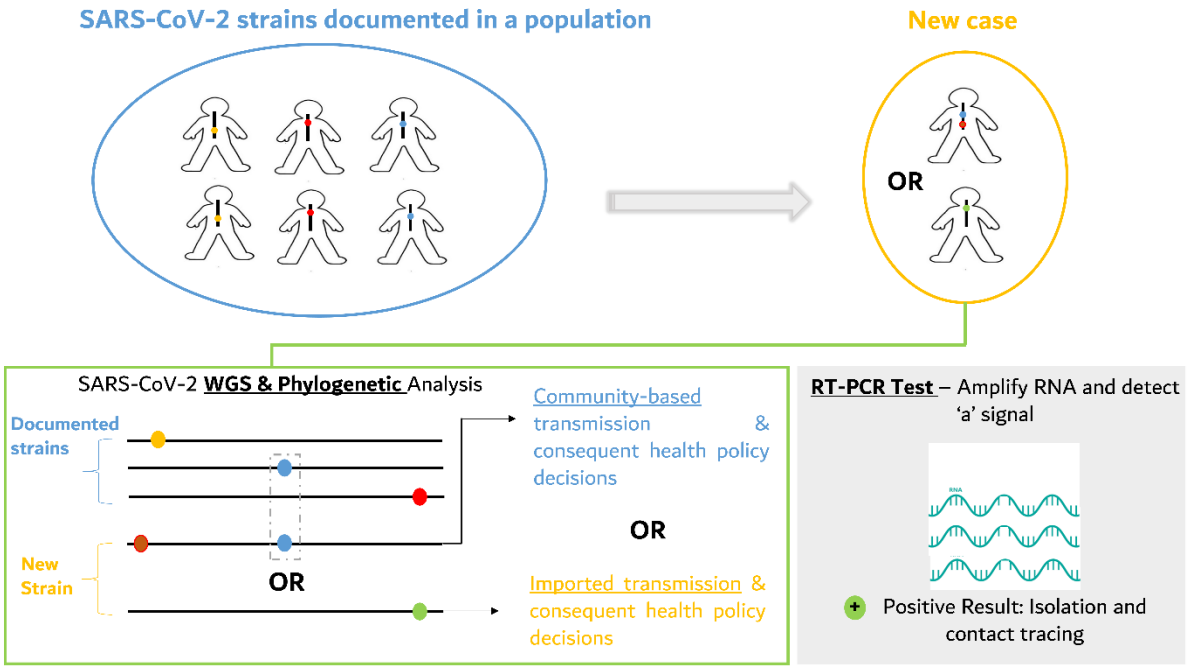

Figure 2. Whole genome amplification, sequencing, and phylogenetic analysis of SARS-CoV-2 genome. A) Wet bench steps describing SARS-CoV-2 genome enrichment and sequencing. B) Bioinformatics and computational steps for sequence alignment, variant calling, SARS-CoV-2 genome assembly, multiple sequence alignment and phylogenetic analysis. All steps are described in detail in Methods.

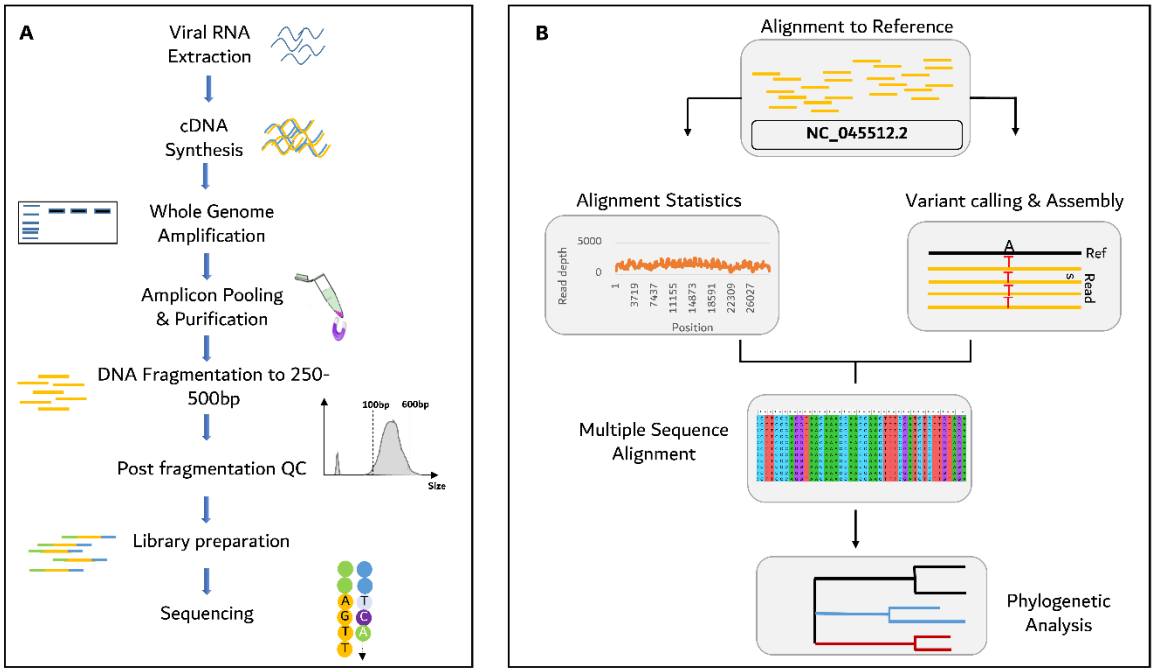

Figure 3. SARS-CoV-2 RNA detection, targeted enrichment, and full sequencing. A) Relationship between the RT-qPCR cycling threshold and sequencing coverage over the SARS-CoV-2 genome. −ΔCt is calculated as an estimate of viral load relative to human RNA (see Methods). Red circles represent lowest −ΔCt values (and lowest relative viral abundance) from samples with very low sequencing coverage. Sequencing data were generated by the shotgun method. B) an agarose gel showing the overlapping 26 PCR products (approximately 1.5kb) covering the SARS-CoV-2 genome. C) An electrophoretic graph showing a major peak between 250-700bp corresponding to fragmented PCR products in B which was pooled and sheared by ultra-sonication. D) SARS-CoV-2 sequence coverage comparison using target enrichment and shotgun sequencing methods in one sample (P1/UAE/2020). *top,* sequencing coverage across the SARS-CoV-2 genomic positions using shotgun transcriptome sequencing (average coverage ~200x); *bottom*, sequencing coverage across SARS-CoV-2 genome (from same patient sample P1/UAE/2020) using target enrichment (average coverage ~1,400x). E) an example of *de novo* assembly of the viral genome isolated from patient P1/UAE/2020 shows clear overlap with the SARS-CoV-2 reference genome.

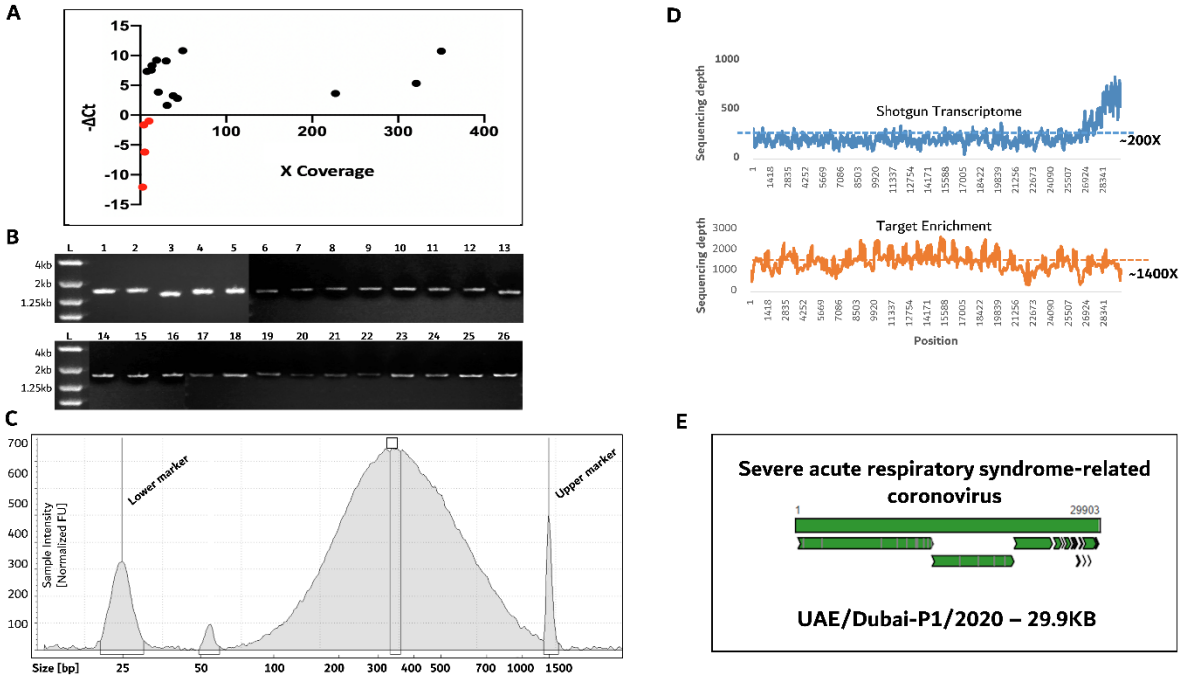

Figure 4. Phylogenetic relationships of SARS-CoV-2 isolates from new patients in this study and previous 'early' patients in the UAE, and other countries. A maximum likelihood phylogeny of 37 SARS-CoV-2 genomes (7 obtained in this study (online **Supplemental Table 2**), 5 downloaded from GISAID database (https://www.epicov.org/), and 25 genomes from early patients in UAE (4)). Bootstrap values >70% supporting major branches are shown. The previous European, Iranian, and Asian clusters are highlighted. The 5 non-UAE isolates were selected based on a BLAST search against GISAID database (last accessed 11 May 2020) and high similarity to the P1/UAE/2020 isolate (all red branches). Scale bar represents number of nucleotide substitutions per site. UAE = United Arab Emirates. GISAID = Global Initiative on Sharing All Influenza Data.