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], Rupa Varghese[3], Zulfa Deesi[3], Norbert Nowotny[2,4], 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]Microbiology and Infection Control Unit, Pathology and Genetics Department, Latifa Women and Children Hospital, Dubai Health Authority, Dubai, United Arab Emirates.

[4]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 resuming social life, robust surveillance mechanisms should be implemented to control the ongoing COVID-19 pandemic. Unlike RT-PCR, 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, practical considerations of cWGS should be addressed before it can be 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 full 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 show a significant improvement in whole genome sequencing data quality and viral detection using amplicon-based target enrichment of SARS-CoV-2. With enrichment, more than 95% of the sequencing reads mapped to the viral genome compared to an average of 0.7% without enrichment. Consequently, a dramatic increase in genome coverage was obtained using significantly less sequencing data, enabling higher scalability and significant cost reductions. We also demonstrate how this SARS-CoV-2 genome sequence can be used to determine its 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 enforced significant social changes and 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 RT-PCR targeting SARS-CoV-2 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, and – if imported – it is important to know the origin of the strains, which in turn influences public health policy decisions. In addition, super-spreader events are very important to identify 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 (www.nextstrain.org) (**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. Here we show that cWGS is cost-effective, and is highly scalable when using a target enrichment sequencing method, and we also demonstrate its utility in tracking the origin of SARS-CoV-2 transmission.

Materials and Methods

Human subjects and ethics approval

All patients were confirmed to have COVID-19 based on positive RT-PCR assay for SARS-CoV-2 in the central 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, Hilden, Germany). SARS-CoV-2 positive results were confirmed

using a RT-PCR assay, originally designed by the US Centers for Disease Control and Prevention (CDC), and is currently provided by Integrated DNA Technologies (IDT, IA, USA). 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, MA, USA) 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, we first accounted for extraction and amplification efficiencies by calculating the ΔCt value for each target as follows: $\Delta Ct = Ct_{Nn} - Ct_{RP}$, where N_n is either N1 or N2. The average of the N1 and N2 target ΔCt values was then negated to reach a relative estimate of 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 (San Diego, CA, USA), following manufacturer's instructions. Briefly, 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, Carisbad, USA), adaptor ligation, and amplification. Libraries were then sequenced using the NovaSeq SP Reagent kit (2 X 150 cycles) from Illumina (San Diego, CA, USA).

Targeted amplification and sequencing of SARS-CoV-2 genome

RNA extracted (~1µg) from patient nasopharyngeal swabs was used for double stranded cDNA synthesis using the QuantiTect Reverse Transcription Kit (Qiagen, Hilden, Germany) 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 2** and **Supplemental Table 1**). The SARS-CoV-2 primer sets used in this study were modified from Wu *et al* (6) by adding M13 tails to enable sequencing by Sanger, if needed (**Supplemental Table 1**). PCR amplification was performed using the Platinum™ SuperFi™ PCR Master Mix (ThermoFisher, MA, USA) 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 (**Figure 2**).

All PCR products were then purified using Agencourt AMPure XP beads (Beckman Coulter, CA, USA), quantified by NanoDrop (ThermoFisher, MA, USA), 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, MA, USA) 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 was confirmed by the TapeStation automated electrophoresis system (Agilent, CA, USA) (**Figure 2A**). Subsequently, the fragmented product is purified and then processed to generate sequencing-ready libraries using the SureSelectXT Library Preparation kit (Agilent, CA, USA) following manufacturer's instructions. Indexed libraries from multiple patients were pooled and sequenced (2 X 150 cycles) using the MiSeq or the NovaSeq systems (Illumina, San Diego, CA, USA). A step-by-step SARS-CoV-2 target enrichment and sequencing protocol is provided in **Appendix I**.

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: https://insaflu.insa.pt/; Genome Detective:

https://www.genomedetective.com/app/typingtool/virus/) (7,8), 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 have at least 20X average coverage (sequencing reads >Q30) across most nucleotide positions (56-29,797).

Phylogenetic analysis

We used Nexstrain (9), which consists of Augur v6.4.3 pipeline for multiple sequence alignment (*MAFFT* v7.455) (10) and phylogenetic tree construction (*IQtree* v1.6.12) (11). Tree topology was assessed using the fast bootstrapping function with 1,000 replicates. Tree visualization and annotations were performed in *FigTree* v1.4.4 (12).

Results

SARS-CoV-2 whole genome sequencing

Shotgun transcriptome sequencing was used to fully sequence SARS-CoV-2 RNA extracted from patients who tested positive for the virus (4). Analysis of the sequencing data showed that this approach required, on average, 4.5Gb of data per sample yielding 30.5 million total reads, of which approximately 1% of the reads (~231,000 reads) mapped to the SARS-CoV-2 genome with an average coverage of 255x (**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 samples with low viral abundance. In fact, viral abundance seemed to correlate with sequencing coverage such that 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 for viral sequences and minimize sequencing cost, 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 **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 two COVID-19 patients (P1/UAE/2020 and L5630), which were first sequenced by shotgun transcriptome (**Table 1**), were sequenced using the enrichment protocol. As expected, we observed significant enhancement in virus detection using this protocol where 95-98% of the reads now mapped to the SARS-CoV-2 genome leading to several fold increase in coverage relative to shotgun transcriptome despite generating less sequencing data (**Table 2** and **Figure 3D**).

Cost, data storage and scalability

On average, 56x coverage per 1Gb of sequencing data was generated using shotgun sequencing (**Table 1**) compared to ~24,000x 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 (**Table 3**). On the other hand, the cost of sequencing one sample at 50x coverage using the shotgun method is $403, while increasing sequencing coverage more than doubled the cost ($1735 at 100x and $1060 at 200x) (**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 around 40,000x coverage (**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 (**Table 3**). Therefore, long term data storage allocations, and cost, are significantly higher, and perhaps formidable, when using the shotgun sequencing approach.

SARS-CoV-2 origin

To illustrate the utility of SARS-CoV-2 whole genome sequencing, we tracked the origin of the virus in patient P1/UAE/2020 by comparing its sequence to virus strains (n=25) identified during the early phase of the pandemic, between January 29 and March 18 2020, in the UAE (4). P1/UAE/2020 patient sample was collected in the first week of April 2020, and is therefore a good candidate to determine if transmission was community based or was an independent external introduction.

Multiple sequence alignment and phylogenetic analysis (**Figure 2B**) using SARS-CoV-2 sequences from P1/UAE/2020 and early patients in the UAE showed that the isolate from patient P1/UAE/2020 did not belong to any of the previously described clusters (clades A2a, A3, and B2) (4). Rather, it belonged to the ancestral 'S' type based on its genotypic profile (**Supplemental Table 2**), and appeared to match closely to 5 other strains from the United Stated and Taiwan (**Figure 4**). Therefore, the transmission in patient P1/UAE/2020 was unlikely to be community-based transmission from the early 25 strains, 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-PCR testing for the virus. Other low throughput, highly affordable semiconductor sequencers can also be used with this protocol (13).

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: https://insaflu.insa.pt/; Genome Detective: https://www.genomedetective.com/app/typingtool/virus/) (7,8) 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 significant data storage or computational investment as shown by our cost, data, and scalability calculations (**Table 3**).

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.

Disclosures. Authors do not have any conflicts of interests to disclose.

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

References

1. Johns Hopkins Center for Systems Sciences and Engineering. COVID19 Dashboard. https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
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(5): 526.

3. Zhang Y, Li Y, Wang L, Li M, Zhou X. Evaluating Transmission Heterogeneity and Super-Spreading Event of COVID-19 in a Metropolis of China. Int J Environ Res Public Health. 2020; 17(10):E3705.
4. Abou Tayoun A, Loney T, Khansaheb H, Ramaswamy S, Harilal D, Deesi Z, Varghese R, et al. Whole genome sequencing 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. Preprint at: https://www.biorxiv.org/content/10.1101/2020.05.06.080606v1
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
6. 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(7798):265-269.
7. 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.
8. 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. Bioinfromatics 2019; 35(5): 871-873.
9. 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).
10. 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(14):3059-66.
11. Chernomor O, von Haeseler A, Quang Minh B. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Systematic Biology 2016; 65(6):997-1008.
12. Rambaut. A. FigTree 1.4.2 Software. *Institute of Evolutionary Biology, Univ. Edinburg*.
13. 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(10):1481-8.

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 coloured 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.

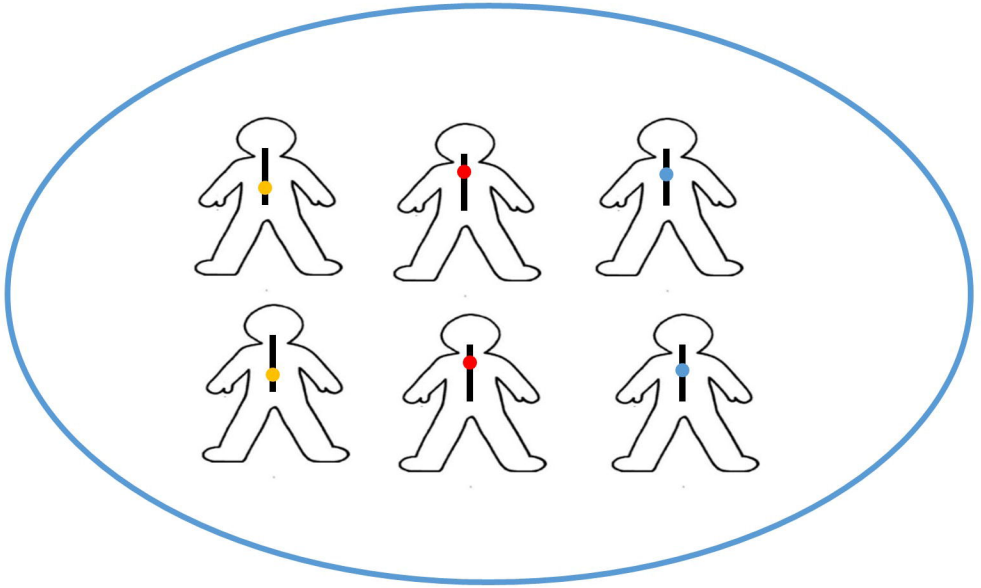
Figure 2. Whole genome amplification, sequencing, and phylogenetic analysis of SARS-CoV-2 genome. A, Wet bench steps describing SARS-CoV-2 genome enrichment and sequencing. B, Bioinformatics and computational steps for sequence alignment, variant calling, SARS-CoV-2 genome assembly, multiple sequence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

Figure 3. SARS-CoV-2 RNA detection, targeted enrichment, and full sequencing. A, Relationship between the RT-PCR cycling threshold and sequencing coverage over the SARS-CoV-2 genome. $-\Delta Ct$ is calculated as an estimate of viral load (see Methods). Coverage increases with viral abundance. Red circles represent lowest $-\Delta Ct$ values (and lowest 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 (~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, *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

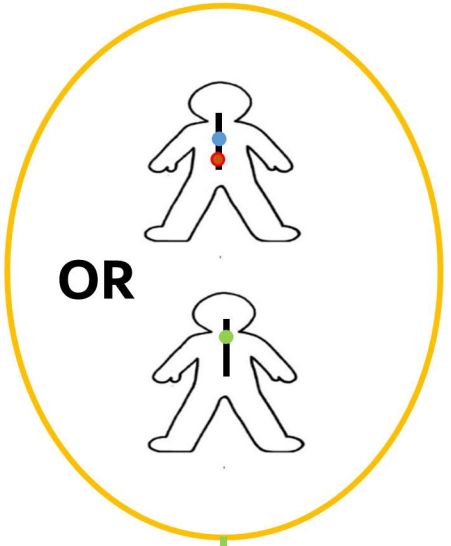
P1/UAE/2020) using target enrichment (average coverage ~1,400x). E, *De novo* assembly of the viral genome isolated from patient P1/UAE/2020 shows clear overlap with the SARS-CoV-2 reference genome.

Figure 4. Phylogenetic relationships of SARS-CoV-2 isolates from patient P1/UAE/2020 and early patients in the UAE, and other countries. A maximum likelihood phylogeny of 31 SARS-CoV-2 genomes (1 obtained from P1/UAE/2020, 5 downloaded from GISAID database (https://www.epicov.org/), and 25 genomes from early patients in UAE (2)). Bootstrap values >70% supporting major branches are shown. 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. Scale bar represents number of nucleotide substitutions per site. UAE = United Arab Emirates. GISAID = Global Initiative on Sharing All Influenza Data.

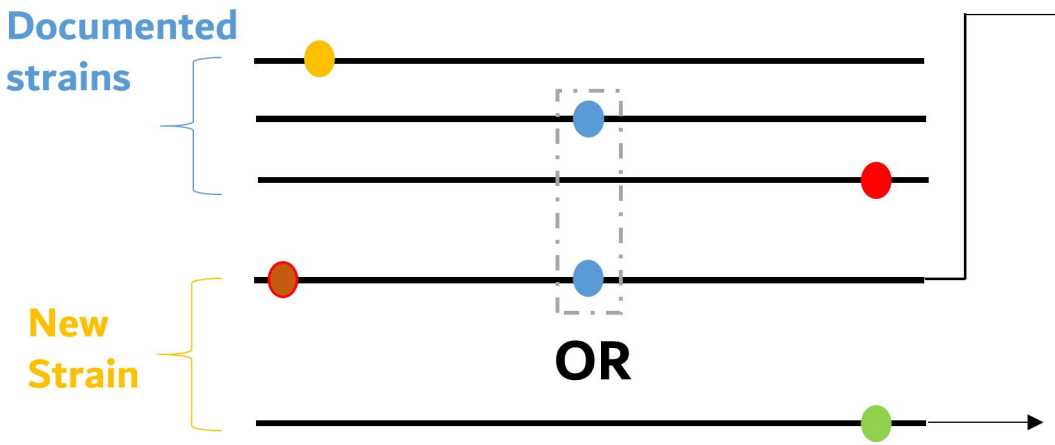
SARS-CoV-2 strains documented in a population



New case



SARS-CoV-2 **WGS & Phylogenetic** Analysis

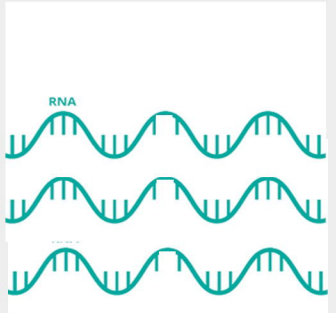


Community-based transmission & consequent health policy decisions

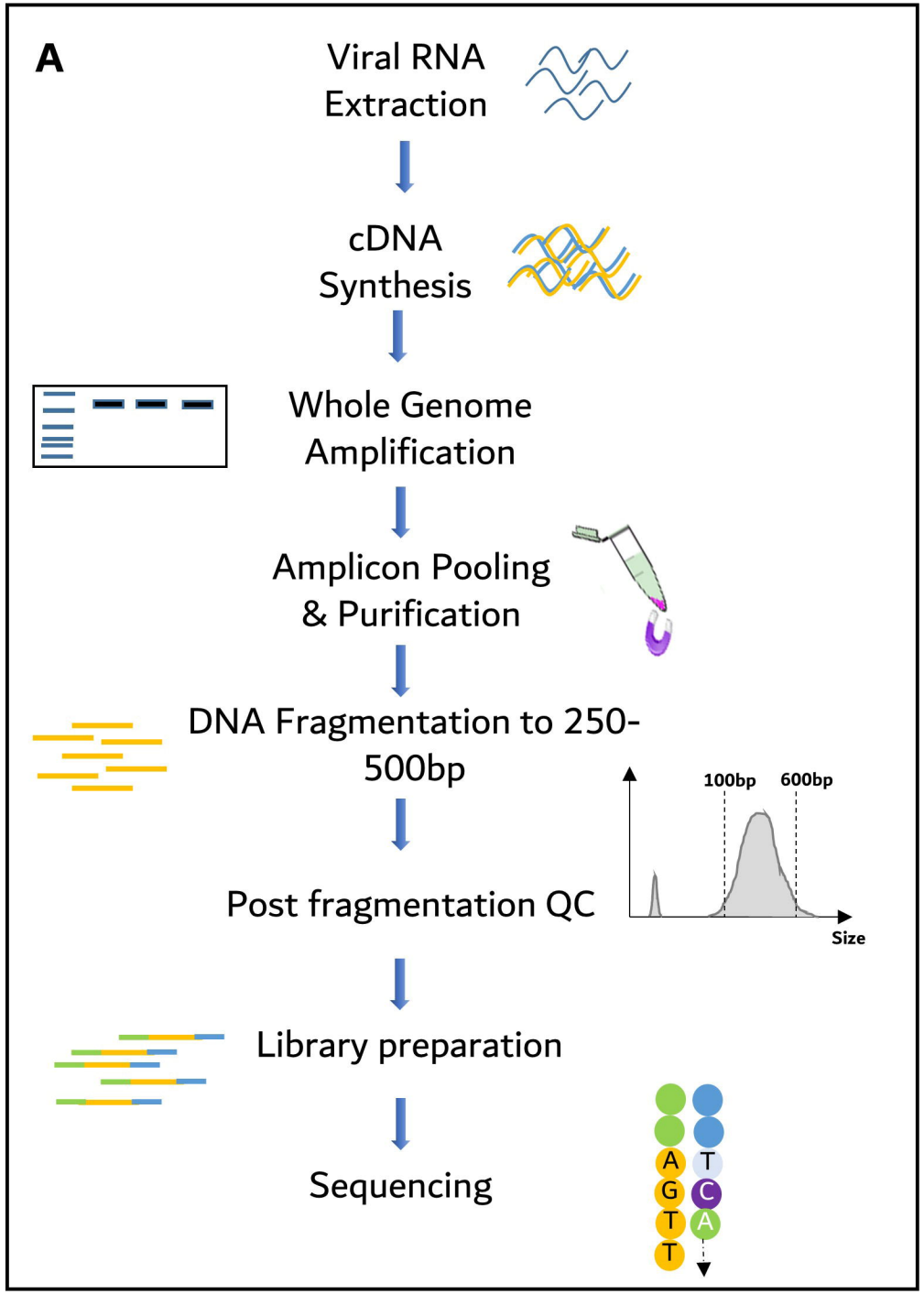
OR

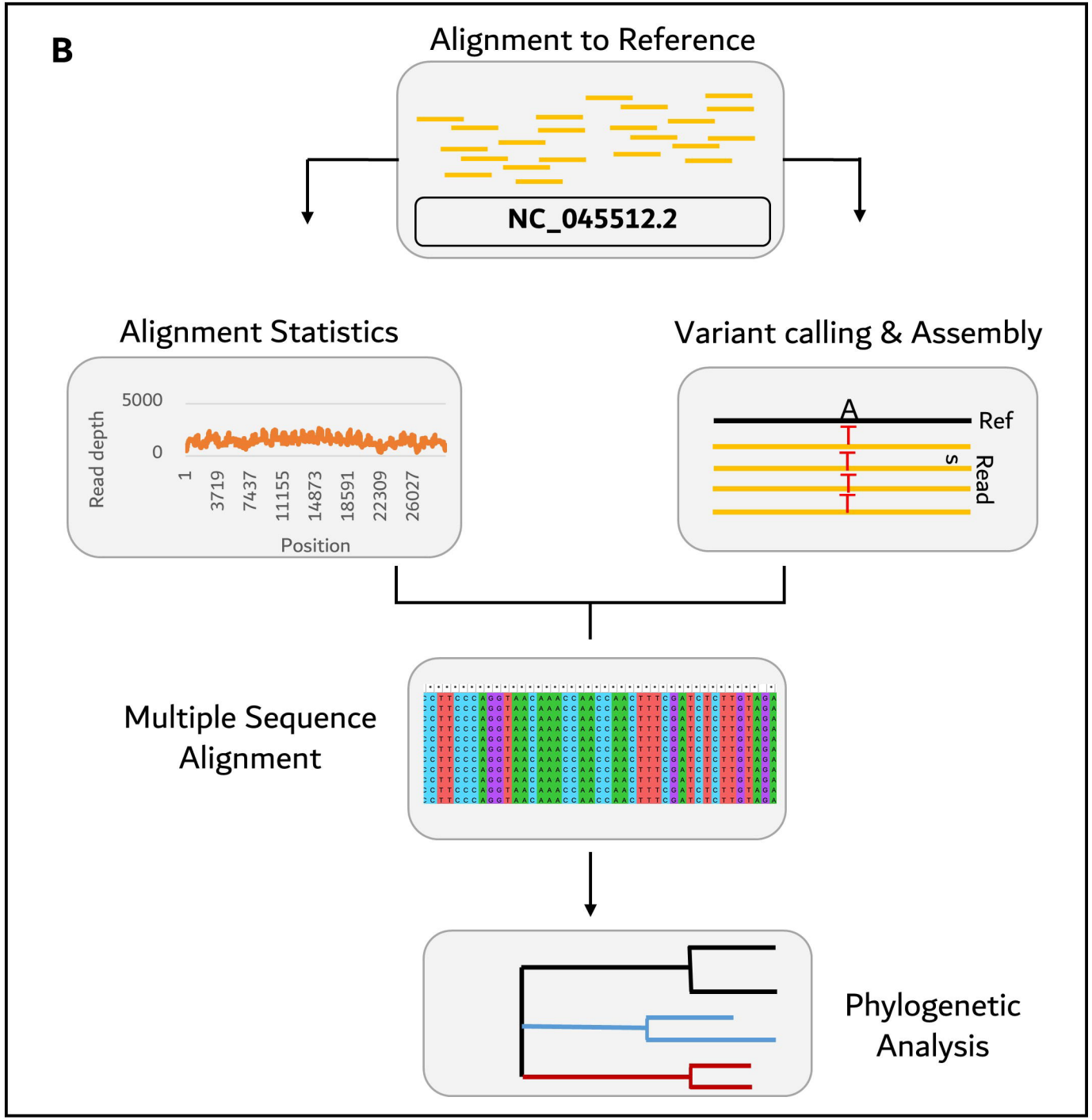
Imported transmission & consequent health policy decisions

RT-PCR Test – Amplify RNA and detect 'a' signal

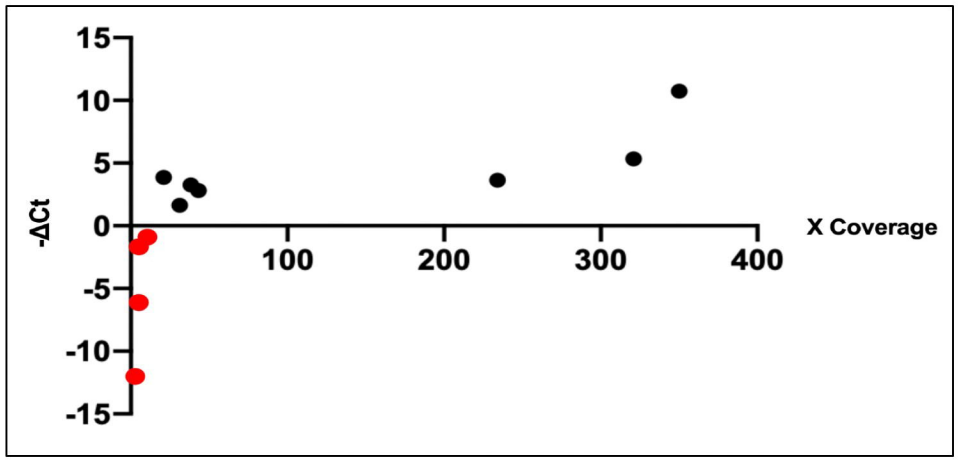


+ Positive Result: Isolation and contact tracing

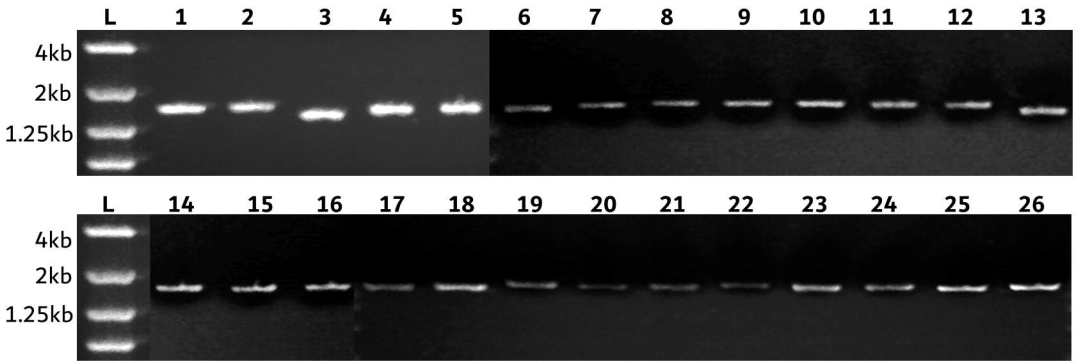

A
Viral RNA Extraction
cDNA Synthesis
Whole Genome Amplification
Amplicon Pooling & Purification
DNA Fragmentation to 250-500bp
100bp
600bp
Size
Post fragmentation QC
Library preparation
Sequencing
A
T
G
C
T
A
T


B
Alignment to Reference
NC_045512.2
Alignment Statistics
Variant calling & Assembly
5000
0
Read depth
1
3719
7437
11155
14873
18591
22309
26027
Position
A
Ref
Reads
Multiple Sequence Alignment
Phylogenetic Analysis

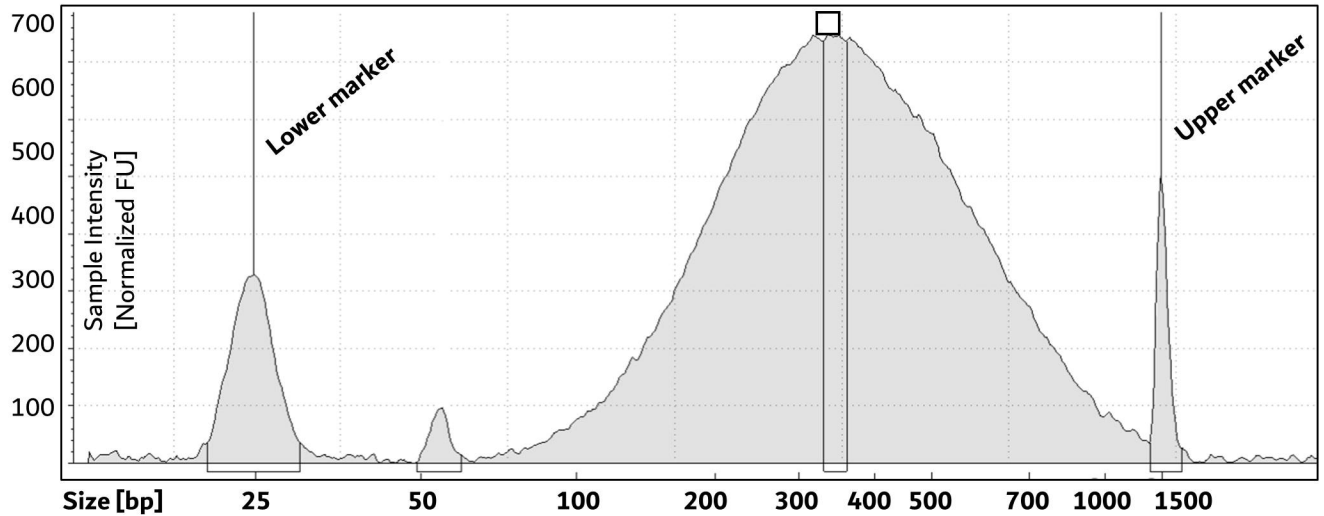
A


15
10
5
0
-5
-10
-15
-ΔCt
100
200
300
400
X Coverage

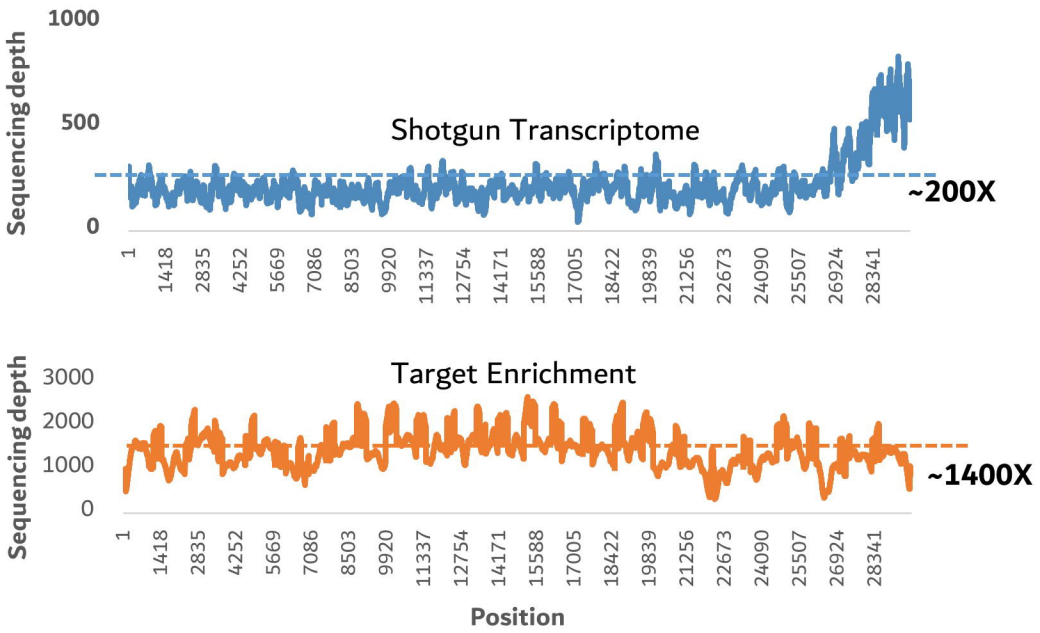
B


L 1 2 3 4 5 6 7 8 9 10 11 12 13
4kb
2kb
1.25kb
L 14 15 16 17 18 19 20 21 22 23 24 25 26
4kb
2kb
1.25kb

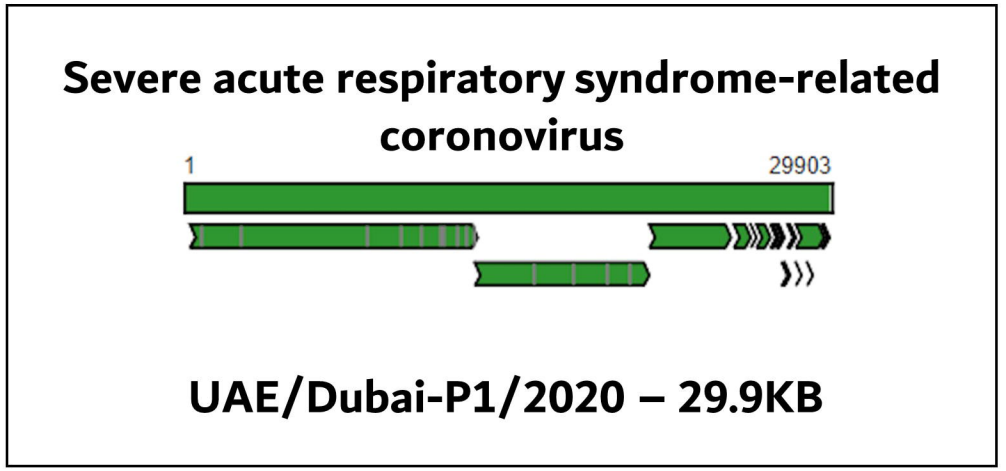
C


Lower marker
Upper marker
Sample Intensity
[Normalized FU]
700
600
500
400
300
200
100
Size [bp]
25
50
100
200
300
400
500
700
1000
1500

D


Shotgun Transcriptome
Sequencing depth
1000
500
0
~200X
Target Enrichment
Sequencing depth
3000
2000
1000
0
~1400X
1 1418 2835 4252 5669 7086 8503 9920 11337 12754 14171 15588 17005 18422 19839 21256 22673 24090 25507 26924 28341
Position

E


Severe acute respiratory syndrome-related coronovirus
1
29903
UAE/Dubai-P1/2020 – 29.9KB

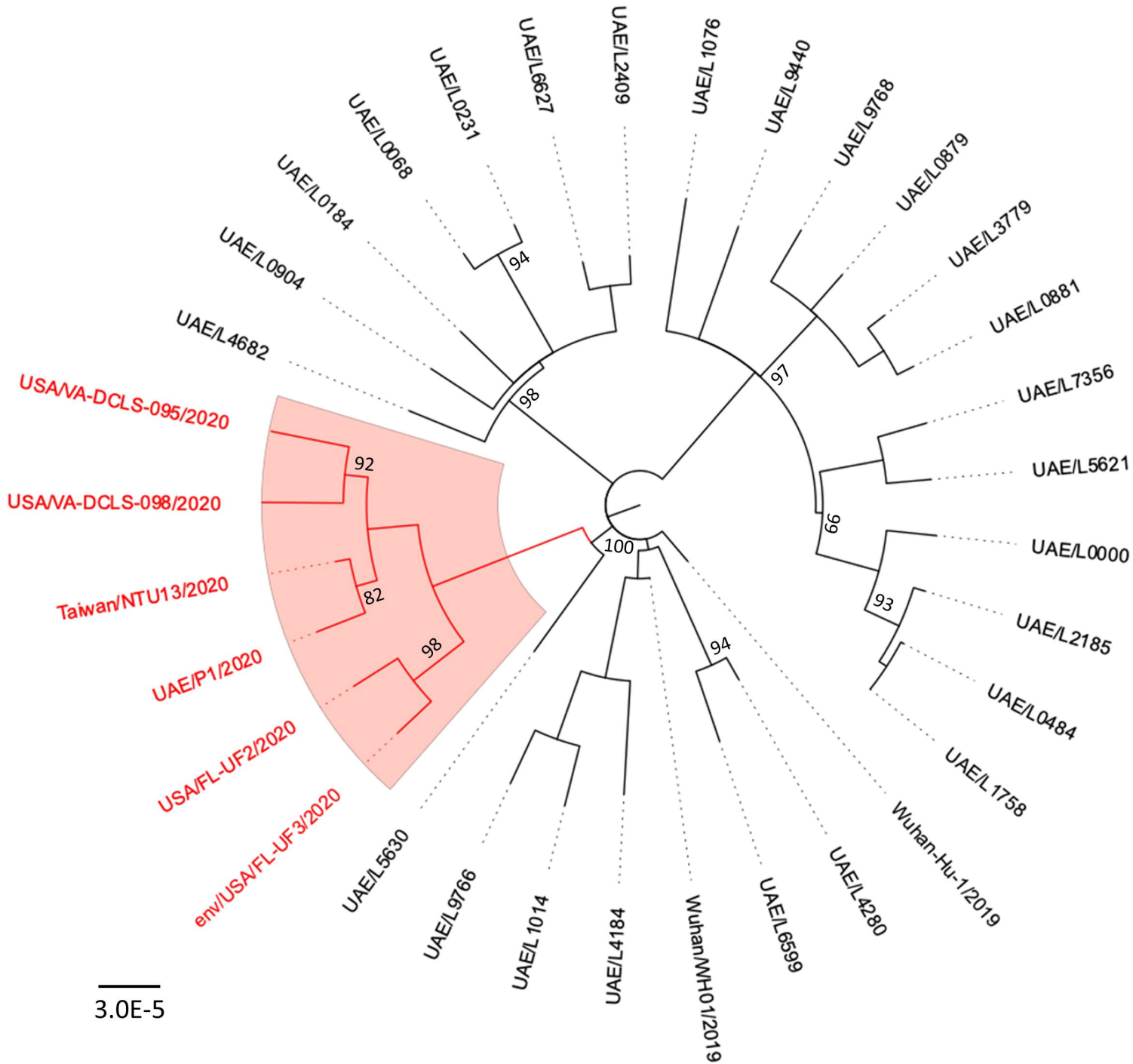

UAE/L2409
UAE/L1076
UAE/L9440
UAE/L9768
UAE/L0879
UAE/L3779
UAE/L0881
UAE/L7356
UAE/L5621
UAE/L0000
UAE/L2185
UAE/L0484
UAE/L1758
Wuhan-Hu-1/2019
UAE/L4280
UAE/L6599
Wuhan/WH01/2019
UAE/L4184
UAE/L1014
UAE/L9766
UAE/L5630
env/USA/FL-UF3/2020
USA/FL-UF2/2020
UAE/P1/2020
Taiwan/NTU13/2020
USA/VA-DCLS-098/2020
USA/VA-DCLS-095/2020
UAE/L4682
UAE/L0904
UAE/L0184
UAE/L0068
UAE/L0231
UAE/L6627
94
98
97
99
100
93
94
92
82
98
3.0E-5

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	0.3872	10.135
L4280	3.852	5.10	33,921,438	69,460	0.20	20.94
L0826**	-12.065	3.21	21,383,276	57,143	0.2672	2.774
L2771**	-6.196	5.70	38,114,908	131,449	0.3449	5.274
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	4.161
L4184	1.635	3.77	25,150,950	215,797	0.86	31.14
UAE/P1/2020**	3.628	4.13	27,542,314	126,756	0.46	227.00
Average	0.943	4.54	30,542,759	231,688	0.72	255.49

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

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

Gray shaded regions highlight samples with low coverage failing to generate full SARS-CoV-2 RNA sequence.

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

	UAE/P1/2020		L5630	
	Target Enrichment	**Shotgun Transcriptome**	**Target Enrichment**	**Shotgun Transcriptome**
Total Reads	298,069	27,542,314	7,027,150	30,462,036
Data Size	60Mb	4.1Gb	1.05Gb	4.57Gb
Reads Aligned*	292,750	126,756	6,731,841	70,058
% of Reads Aligned*	98.21	0.46	95.8	0.23
Mean Target Coverage*	1464	227	25339.89	4.161
% >5X*	100	100	100	97.1
% >10X*	100	100	100	26.14
% >20X*	100	100	100	2.3
% >30X*	100	99.95	100	0
% >40X*	100	99.8	100	0
% >50X*	100	99.63	100	0
% >100X*	100	77.54	100	0

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

Table 3. Cost of SARS-CoV-2 whole genome sequencing using Enrichment or shotgun sequencing at different throughputs.

Enrichment Method				
	MiSeq V2 300 cycles micro flowcell (1.2 Gb)	MiniSeq Midi kit 300 cycles (2.4 Gb)	MiSeq V2 300 cycles (4.5 Gb)	NovaSeq_SP flowcell 300 cycles (250 Gb)
No of Samples / Run	96	96	96	196
Avg: X Coverage	400	800	1500	40,000
Price USD	$ 87.37	$ 91.07	$ 94.21	$ 108.30
Shotgun Metagenome				
	MiSeq V2 300 cycles micro flowcell (1.2 Gb)	MiniSeq Midi kit 300 cycles (2.4 Gb)	MiSeq V2 300 cycles (4.5 Gb)	NovaSeq_SP flowcell 300 cycles (250 Gb)
No of Samples / Run	1	1	1	62
Avg: X Coverage	50	100	200	200
Price USD	$ 403.13	$ 1,735.30	$ 1,059.91	$ 232.48