Whole Genome Sequence Analysis of SARS-CoV-2 Strains Circulating in Malaysia During First Wave and Early Second Wave of Infections.

Zarina Mohd Zawawi (zarina.zawawi@moh.gov.my)
Virology Unit, Infectious Disease Research Center, Institute for Medical Research, National Institutes of Health Complex, Ministry of Health Malaysia, Selangor, Malaysia

Jeyanthi Suppiah
Virology Unit, Infectious Disease Research Center, Institute for Medical Research, National Institutes of Health Complex, Ministry of Health Malaysia, Selangor, Malaysia

Jeevanathan Kalyanasundram
Virology Unit, Infectious Disease Research Center, Institute for Medical Research, National Institutes of Health Complex, Ministry of Health Malaysia, Selangor, Malaysia

Muhammad Afif Azizan
Virology Unit, Infectious Disease Research Center, Institute for Medical Research, National Institutes of Health Complex, Ministry of Health Malaysia, Selangor, Malaysia

Shuhaila Mat-Sharani
Molecular Pathology Unit, Cancer Research Centre, Institute for Medical Research, National Institutes of Health Complex, Ministry of Health Malaysia, Selangor, Malaysia

Hamidah Akmal Hisham
Molecular Pathology Unit, Cancer Research Centre, Institute for Medical Research, National Institutes of Health Complex, Ministry of Health Malaysia, Selangor, Malaysia

Tan Lu Ping
Molecular Pathology Unit, Cancer Research Centre, Institute for Medical Research, National Institutes of Health Complex, Ministry of Health Malaysia, Selangor, Malaysia

Manisya Zauri Ab Wahid
Virology Unit, Infectious Disease Research Center, Institute for Medical Research, National Institutes of Health Complex, Ministry of Health Malaysia, Selangor, Malaysia

Tengku Rogayah Tengku Abdul Rashid
Virology Unit, Infectious Disease Research Center, Institute for Medical Research, National Institutes of Health Complex, Ministry of Health Malaysia, Selangor, Malaysia

Rozainanee Mohd Zain
Virology Unit, Infectious Disease Research Center, Institute for Medical Research, National Institutes of Health Complex, Ministry of Health Malaysia, Selangor, Malaysia

Norazah Ahmad

Infectious Disease Research Center, Institute for Medical Research, National Institutes of Health Complex, Ministry of Health Malaysia, Selangor, Malaysia

Ravindran Thayan

Virology Unit, Infectious Disease Research Center, Institute for Medical Research, National Institutes of Health Complex, Ministry of Health Malaysia, Selangor, Malaysia

---

Research Article

Keywords: COVID-19, SARS-CoV-2, whole genome sequencing, variants, phylogeny

Posted Date: September 21st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-81152/v1

Abstract

Background: Since December 2019, the outbreak of COVID-19 has raised a great public health concern globally. Here, we report the whole genome sequencing analysis of SARS-CoV-2 strains in Malaysia isolated from six patients diagnosed with COVID-19.

Methods: The SARS-CoV-2 viral RNA extracted from clinical specimens and isolates were subjected to whole genome sequencing using NextSeq 500 platform. The sequencing data were assembled to full genome sequences using Megahit and phylogenetic tree was constructed using Mega X software.

Results: Six full genome sequences of SARS-CoV-2 comprising of strains from 1st wave (25th January 2020) and 2nd wave (27th February 2020) infection were obtained. Downstream analysis demonstrated diversity among the Malaysian strains with several synonymous and non-synonymous mutations in four of the six cases, affecting the genes M, orf1ab, and *S* of the SARS-CoV-2 virus. The phylogenetic analysis revealed viral genome sequences of Malaysian SARS-CoV-2 strains clustered under the ancestral Type B.

Conclusion: This study comprehended the SARS-CoV-2 virus evolution during its circulation in Malaysia. Continuous monitoring and analysis of the whole genome sequences of confirmed cases would be crucial to further understand the genetic evolution of the virus.

Background

Coronavirus disease 2019 (COVID-19) emerged in December 2019 as a new pandemic form of life-threatening infection caused by a novel coronavirus, namely SARS-CoV-2 [1]. As of 14th May 2020, there were 4.2 million cases of SARS-CoV-2 with 294,046 deaths reported worldwide (WHO) [2]. In Malaysia, the total number of cases was 6,819 cases with 5,351 recovered, and 112 deaths as of 14th May 2020 [3]. The most common clinical manifestation for SARS-CoV-2 were fever, cough, fatigue and expectoration [1,4].

Coronaviruses are enveloped viruses with a positive-sense, single-stranded RNA viruses belonging to the family *Coronaviridae*. To date, four coronavirus genera have been identified which are *Alphacoronavirus, Betacoronavirus, Gammacoronavirus* and *Deltacoronavirus* [5]. These viruses generally infect animals, including birds and mammals [6]. Similar to SARS (Severe Acute Respiratory Syndrome) and MERS (Middle East Respiratory Syndrome), SARS-CoV-2 is a zoonotic coronavirus that belongs to the genus *Betacoronavirus.* It has a genome size varying from 29.8 kb to 29.9 kb that encodes for multiple structural and non-structural proteins [7]. The structural proteins include the spike (S) protein, the envelope (E) protein, the membrane (M) protein, and the nucleocapsid (N) protein [8]. Briefly, S protein guides the entry of the virus into host cells, E protein plays a role in production and maturation of the virus, M protein will determine the shape of virus and N protein involves in viral replication [9,10].

Malaysia experienced the first wave of SARS-CoV-2 infection in late January 25th, 2020 when three cases were confirmed positive for COVID-19 through contact tracing of the index case identified in Singapore.

The second wave of SARS-CoV-2 infection started on 27th February 2020 after Malaysia reported no new cases for 11 days. Initially, there were more confirmed positive imported cases originating from China travellers. Subsequently, clustered and confirmed cases without a history of travel to China increased as the outbreak progressed. Hence there is a need to look into the molecular epidemiology of the SARS-CoV-2 to comprehend the evolution of the virus and to compare with those circulating elsewhere. Whole genome sequencing was carried out by the Institute for Medical Research, Malaysia to compare the genomic evolution of SARS-CoV-2 strains circulating in Malaysia during the first wave and early second wave of infections.

Methods

Sample selection and viral cultivation

Nasopharyngeal swab (NPS) and oropharyngeal swab (OPS) specimens from COVID-19 suspected patients that were sent for routine diagnosis to the Virology Unit, Institute for Medical Research, Malaysia, were selected for viral cultivation. Selection criteria were based on confirmed positive cases by COVID-19 Real-Time RT-PCR from the first wave and second wave clusters. Specimens were inoculated into Vero E6 cells in a biosafety Level-3 facility according to the WHO laboratory biosafety guidelines and monitored for cytopathic effect (CPE).

Viral RNA Extraction

Virus isolates were harvested from passage 1. Viral RNAs were extracted from these isolates using QIAamp Viral RNA Mini kit (QIAGEN, Hilden, Germany) according to the manufacturer's instruction and confirmed for presence of SARS-Cov-2 genome by Real-Time RT-PCR for E gene and RdRp (Berlin WHO) [11]. Additionally, viral RNAs were also re-extracted from retrospective original specimens from confirmed COVID-19 cases that showed Cq <20.

Next Generation Sequencing (NGS) and data analysis

NGS library was constructed using TruSeq Stranded Total RNA Gold library prep kit (Illumina, USA) from five clinical specimens and five viral isolates which passed quality control assessment for RNA concentration by Qubit™ RNA HS Assay (Thermo Fisher, USA). Sequencing was performed on the NextSeq 500 platform (Illumina USA) and subjected to 160 million pair reads that were evaluated with FastQC [12, 13]. These raw reads were then parsed through quality filtration with remaining TruSeq Illumina adaptor sequences. Any low quality, low complexity and unpaired reads were removed using Trimmomatic [14] with the option: LEADING:3 TRAILING:3 MINLEN:30 and Qpred 33. *De novo* genome assembly was conducted using Megahit [15] with default parameter. The genome sequence built from each sample was blast to reference viral genome SARS-CoV-2 (NC_045512.2) and mapped using Hisat2 program [16]. Gene prediction was done with Vgas (http://cefg.uestc.cn/vgas/) [17] followed by identification of single nucleotide variants (SNVs) using samtools [18], GATK [19] and Lofreq [20]. The SNVs were annotated to the reference strain using snpEfff [21] and effects were predicted via snpSift [22].

Phylogenetic analysis

The SARS-CoV-2 full length genome sequences were subjected to phylogenetic analysis. A dataset of 50 SARS-CoV-2 complete genomes from different countries was retrieved from GISAID (https://www.gisaid.org/, last access 16 March 2020). Sequence alignment was performed using Multiple Sequence Comparison by Log-Expectation (MUSCLE) software and the phylogenetic tree was constructed by neighbour joining method (bootstrap 1000x) using Molecular Evolutionary Genetics Analysis (MEGA, v7).

Ethical consideration

This study used the retrospective specimens and this study does not require approval from human subject's ethics review committee.

Results

General information of selected COVID-19 cases

Ten confirmed cases were subjected to whole genome sequencing study. Of these, only six were successfully sequenced to full genome; comprising two clinical specimens and four viral isolates; representing the first and early second wave of SARS-CoV-2 outbreak in Malaysia. As shown in Table 1, cases number 5 and 19 were from the first wave of the COVID-19 outbreak whereas the cases number 27, 26, 70 and 62 were from the second wave. Based on contact tracing, case number 62 was identified as a Malaysian male who had closed contact with positive case number 33 (sample not analysed in this study) who was the second-generation cluster of case number 26, suggesting a local transmission chain in Malaysia.

Table 1: Segregation by case number and history of patients.

Wave of infection	Accession No	Specimen type	Case No	Gender/Age	History
1st (25/01/2020)	EPI_ISL_430443	Isolate	5	F/36	Chinese tourist from Wuhan, closed contact with index case in Singapore.
	EPI_ISL_430444	Isolate	19	F/38	Chinese tourist from Wuhan, closed contact with two positive cases.
2nd (27/02/2020)	EPI_ISL_430441	Isolate	27	F/19	Malaysian student nurse, closed contact with one positive case.
	EPI_ISL_430442	Isolate	26	M/52	Malaysian with history of travelling to Shanghai.
	EPI_ISL_430440	Clinical	70	F/50	Malaysian, closed contact with positive case.
	EPI_ISL_430439	Clinical	62	M/49	Malaysian, closed contact with positive case from second generation cluster.

Table 1 shows the general information of selected COVID-19 cases involved in this study, segregated by wave of infections and case number. F: Female; M: Male.

Whole genome sequencing analysis

After removing reads mapped to human genome, it was found that there were about 4 to 10 million of reads that passed quality control for each sample. De novo assembly resulted in about 29.8kb genome sequence for each sample, with coverage between 439X to 1166X and average GC content of about 45% (Supplementary Table 1). The genome sequence assembled from each sample had 99-100% similarity to reference viral genome SARS-CoV-2 (NC_045512.2). Our findings revealed 5 non-synonymous variants detected in four of the six cases, affecting the genes M (membrane glycoprotein), orf1ab (orf1ab polyprotein), and S (spike protein) of the SARS-CoV-2 virus (Table 2).

Table 2: The variant analysis of Malaysian SARS-CoV-2 strains.

Accession No	Nucleotide variation	Gene	Amino acid change	Mutation type
EPI_ISL_430443	3163T>A	orf1ab	-	Synonymous mutation
	16272T>C	orf1ab	-	Synonymous mutation
	27147G>C	M	D209H	Missense
EPI_ISL_430444	1758C>T	orf1ab	A498V	Missense
	10604C>T	orf1ab	P3447S	Missense
EPI_ISL_430442	23583-23597del15bp	S	In-frame del [QTQTN] 268-272	Deletion
EPI_ISL_430439	4A>T	-	-	UTR
	11752C>T	orf1ab	-	Synonymous mutation
	19170C>A	orf1ab	F6302L	Missense

Phylogenetic analysis

The phylogenetic tree in Fig 1 showed a main clade containing four groups. The viral genome sequences of Malaysian SARS-CoV-2 strains clustered under one identical group; the ancestral Type B together with the reference genome sequences of Wuhan-Hu-1, Shanghai and other strains from different countries. Only cases from first wave of outbreak were seen to be closely related to Wuhan strains while others Malaysian SARS-CoV-2 strains were segregated towards other strains in the same group.

Discussion

The current circulating SARS-CoV-2 was reported to diverse into three main variants which are A, B and C, characterized through whole genome sequencing [23]. Whilst type A and C are discovered to be the dominant types outside of East Asia, type B is commonly detected to be circulating in East Asia. However, derived B types which mutated from ancestral B-type enabled these subtypes to transmit outside of East Asia. In comparison to A-type, B-type genomic sequence showed two synonymous mutation which are T8782C and C28144T. On the other hand, Type C is reported to have nonsynonymous mutation G26144T compared to its parental type B [23].

In this study, the SARS-CoV-2 virus from clinical specimens and culture isolates were successfully sequenced to whole genome using the Illumina NexSeq platform. The nucleotides identity reached up to 99.9% similarities to 93 full genomes of SARS-CoV-2 and the sequence homology analysis showed that all sequenced samples belonged to ancestral Type B variant. The two cases in first wave were closely related to the Wuhan strains whereas the other four cases were dispersed from the Wuhan strain and branched close to the strains of Shenzhen and Hangzhou. Although analysis of larger sample size is needed, our preliminary finding supports the evidence of Forster et al (2020) that the East Asians monopolized the ancestral B type.

The variant analysis of Malaysian SARS-CoV-2 strains was studied. Intriguingly, it was found that SARS-CoV-2 virus from case number 26 (EPI_ISL_430442) had 15 nucleotides in frame deletion in the spike protein that none of the other strains had. This in-frame deletion has been previously characterised to result in the loss of five amino acids (QTQTN) flanking the polybasic cleavage site of the spike protein and hypothesized to be passage oriented [24]. However, one study demonstrated that the deletion also occurred in SARS-CoV-2 extracted from clinical samples [25]. The spike protein of coronaviruses plays pivotal role in viral infectivity, transmissibility and, antigenicity. Therefore, the genetic characterization of the spike protein in SARS-CoV-2 would shed light on its evolution. The viral isolate from case 26 was passaged once in Vero E6 cells, which could have promoted such deletion. The prevalence of this mutation among clinical samples, warrants further investigation. Notably, based on epidemiological mapping, the nature of case 26 that was reported to have generated first-generation and second-generation clusters after attending a meeting highlights its high transmission ability which could have been caused by the deletion in the spike protein of the virus. However, it is unclear if this deletion contributes to severity of the disease.

Conclusion

In this study, six SARS-CoV-2 strains isolated from Malaysian COVID-19 patients and Chinese tourists were sequenced and analysed. Compared to the genome sequence of SARS-CoV-2 from Wuhan, five non-synonymous variants were identified in four of the six sequences from Malaysia. The phylogenetic analysis consistently grouped all Malaysian SARS-CoV-2 strains in ancestral Type B together with majority of China's strains. This study provides the first whole genome sequencing data on SARS-CoV-2 strains comparing first and early second wave strains circulating in Malaysia.

Abbreviations

SARS: Severe acute respiratory syndrome; SARS-CoV-2: Novel human coronavirus; MERS: Middle East Respiratory Syndrome; S: Spike protein; E: E protein; M: Membrane protein; N: nucleocapsid protein; orf1ab: orf1ab polyprotein; NPS: nasopharyngeal swab; OPS: oropharyngeal swab; RT-PCR: Reverse transcriptase polymerase chain reaction; WHO: World Health Organization; CPE: cytopathic effect; RNA: ribonucleic acid; RdRp: RNA-dependent RNA polymerase; COVID-19: Coronavirus disease 2019; Cq: quantification cycle; SNV: single nucleotide variant; GISAID: Global Initiative on Sharing Avian Influenza Data; MUSCLE: Multiple Sequence Comparison by Log-Expectation; MEGA: Molecular Evolutionary Genetics Analysis

Declarations

Acknowledgments

The authors would like to thank the Director General, Ministry of Health, Malaysia for permission to publish the work.

Authors' contributions

ZMZ, JS, JK, MAA & MZAW were involved in the conception, design of the studies, performed experiments and acquired the data. ZMZ, JS, SMS, HAH & TLP were involved in the bioinformatic analysis. JS, TGR, RT, RMZ & NA delivered intellectual input and co-edited the paper. All authors have read and approved the final version of the manuscript.

Funding

The study was funded by Ministry of Health, Malaysia (NMRR-20-884-54816). The funder had no role in study design, data collection and analysis or preparation of the manuscript.

Availability of data and materials

The data that support the findings of this study are available from the corresponding author upon request.

Ethics approval and consent to participate

The study has been approved by Medical Research & Ethics Committee (MREC), Ministry of Health, Malaysia.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Tanu S. A review of Coronavirus Disease-2019 (COVID-19). Indian J Pediatr. 2020;87(4): 281-286.
2. Coronavirus disease (CIOVID-2019) situation report. Geneva. World Health Organization; (https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200514-covid-19-sitrep-115.pdf?sfvrsn=3fce8d3c_4, assessed 15.5.2020).
3. Ministry of Health Malaysia (MOH) – COVID-19 media – http://www.moh.gov.my/index.php/pages/view/2019-ncov-wuhan-kenyataan-akhbar.
4. Jiang F, Deng L, Zhang L, Cai Y, Cheung CW, Xia Z. Review of the Clinical Characteristics of Coronavirus Disease 2019 (COVID-19). J Gen Intern Med. 2020;4:1-5.
5. Phan M, Ngo-Tri T, Hong-Anh P, Baker S, Kellam P, Cotten M. Identification and characterization of Coronaviridae genomes from Vietnamese bats and rats based on conserved protein domains. Virus Evol. 2018;4(2):vey035.

6. Saif LJ. ANIMAL CORONAVIRUSES: LESSONS FOR SARS. In: Institute of Medicine (US) Forum on Microbial Threats; Knobler S, Mahmoud A, Lemon S, et al., editors. Learning from SARS: Preparing for the Next Disease Outbreak: Workshop Summary. Washington (DC): National Academies Press (US); 2004. Available from: https://www.ncbi.nlm.nih.gov/books/NBK92442/
7. Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. Gene Rep. 2020;19:100682.
8. Pal M, Berhanu G, Desalegn C, Kandi V. Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2): An Update. Cureus. 2020;12(3):e7423.
9. Tortorici MA, Veesler D. Structural insights into coronavirus entry. Adv. Virus Res. 2019;105:93-116.
10. Astuti I, Ysrafil. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response. Diabetes Metab Syndr. 2020;14(4):407–412.
11. Corman VM, Landt O, Kaiser M, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill. 2020;25(3):2000045.
12. Brown J, Pirrung M, McCue LA. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics. 2017;33(19):3137–3139.
13. Ward CM, To TH, Pederson SM. ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files. Bioinformatics. 2020;36(8):2587-2588.
14. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-
15. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674-
16. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907-915.
17. Zhang KY, Gao YZ, Du MZ, Liu S, Dong C, Guo F-B. Vgas: A Viral Genome Annotation System. Front Microbiol. 2019;10:184.
18. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. Genome Project Data Processing. The Sequence Alignment/Map format and SAMtools. 2009;25(16): 2078-2079.
19. Brouard JS, Schenkel F, Marete A, Bissonnette N. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. J Anim Sci Biotechnol. 2019;10:44
20. Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012;40(22):11189-11201.
21. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." Fly (Austin). 2012;6(2): 80-92.

22. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012;3:35.

23. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proceedings of the National Academy of Sciences of the United States of America. 2020;117(17):9241–9243.

24. Lau SY, Wang P, Mok BWY, Zhang AJ, Chu H, Lee ACY, et al. Attenuated SARS-CoV-2 variants with deletions at the S1/S2 junction, Emerg. Microbes Infect. 2020;9:1:837-842

25. Liu Z, Zheng H, Yuan R, Li M, Lin H, Peng J, et al. Identification of a common deletion in the spike protein of SARS-CoV-2. BioRxiv. 2020;03.31.015941.

Figures

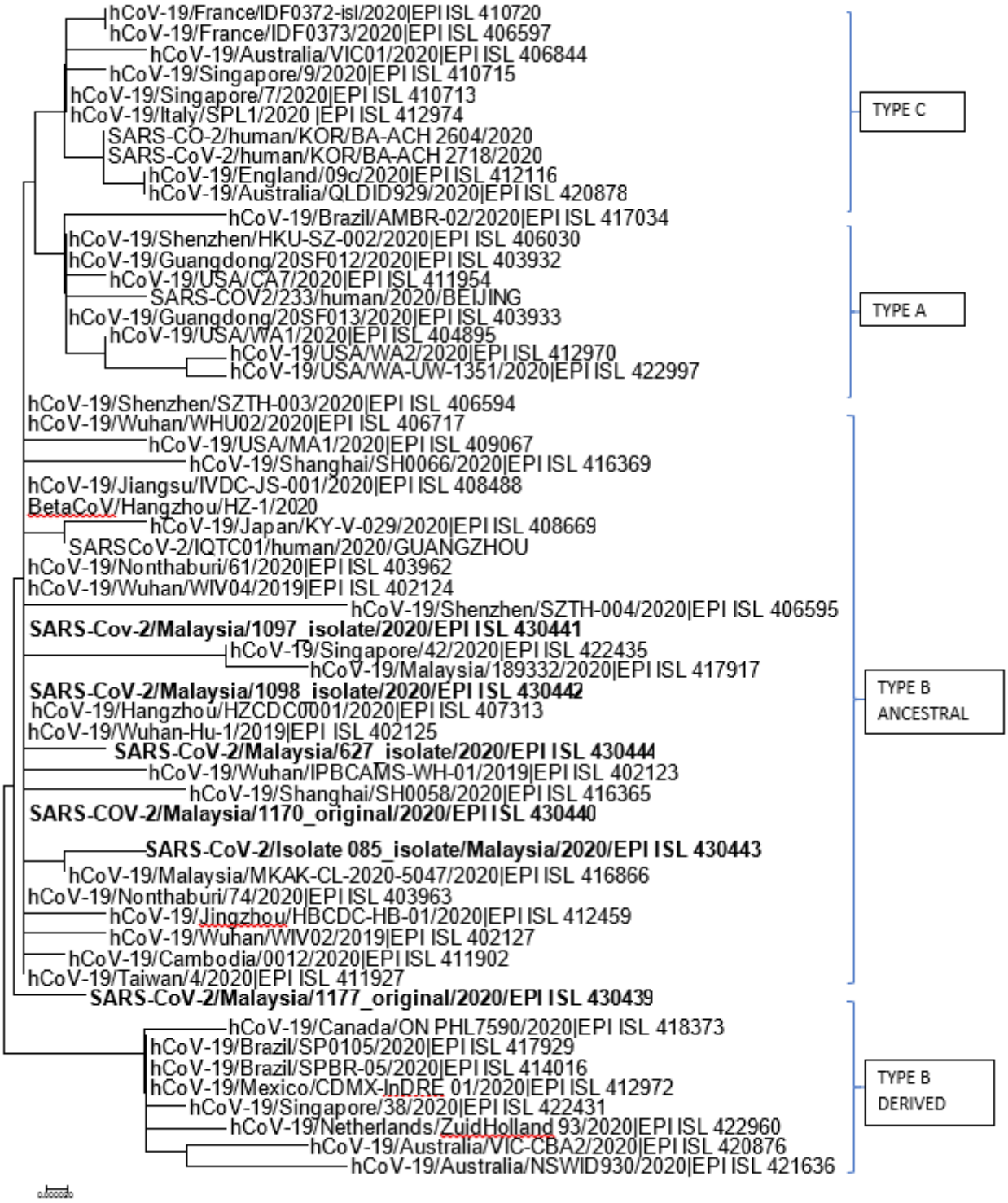

Figure 1

Phylogeny of full genome sequences of SARS-CoV-2. All six Malaysian strains reported in this study are in bold font. The phylogenetic tree was edited with FigTree v1.4.4 software.A

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- WGSAnalysisofMalaysianSARSCoV2supplementalInfo.docx