Indonesian Journal of Biotechnology

VOLUME 27(1), 2022, 19-28 | RESEARCH ARTICLE

The design of Indonesian SARS-CoV-2 primers based on phylogenomic analysis of the SARS-CoV-2 clades

Tsania Taskia Nabila[1], Ata Rofita Wasiati[1], Afif Pranaya Jati[2], Annisa Khumaira[1,*]

[1]Biotechnology Study Program, Faculty of Science and Technology, Universitas 'Aisyiyah Yogyakarta, Jl. Ringroad Barat No.63, Nogotirto, Gamping, Sleman, Daerah Istimewa Yogyakarta 55592, Indonesia
[2]Masyarakat Bioinformatika dan Biodiversitas Indonesia (MABBI), Ruang 613, Lantai 6, Program Studi Bioteknologi, Universitas Esa Unggul, Jl. Arjuna Utara No. 9, Jakarta Barat 11510, Indonesia
*Corresponding author: annisakhumaira@unisayogya.ac.id

SUBMITTED 20 May 2021 **REVISED** 7 September 2021 **ACCEPTED** 2 November 2021

ABSTRACT Molecular detection needs to be augmented for COVID-19 detection in Indonesia using the PCR method with primer-based gene analysis. This is necessary because the RNA of the SARS-CoV-2 virus, the causative infectious agent of the pandemic, has been mutated. Therefore, this study aimed to develop a primer design for determining SARS-CoV-2 clades in Indonesia using phylogenomic analysis. Data were obtained from 38 GISAID (Global Initiative on Sharing All Influenza Data) viruses and the relationships were analyzed using maximum likelihood (ML) phylogenomic analysis with a substitution model of generalized time-reversible (GTR) to construct the tree topology. The results showed that the five types of SARS-CoVs-2 clades in Indonesia were L, G, GH, GR, and O. It also indicated that the GH region had the highest rate of clade at 50%, with the S clade affecting its formation. Furthermore, the genome sequences of the GH type used to design its primer were based on three genes, namely *RdRp*, *S*, and *N*. The *RdRp* and *N* genes were found to be conserved and hardy mutants, while the *S* gene occurred repeatedly. Several previous studies have stated that the designed primers produced missense mutations compared to another *in silico*. Therefore, three sets of primers were achieved from the GC contents and clamps, Tm range, and structural secondary indicator standards.

KEYWORDS maximum likelihood method; phylogenomic analysis; primer design; SARS-CoV-2

1. Introduction

The first outbreak of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) occurred in Wuhan, China, in December 2019, rapidly spreading to other parts of the world. Since its inception, over 5% of infected patients have suffered severe pneumonia and are likely to have multi-organ dysfunction with a mortality rate of 1.4%. Therefore, studies need to be conducted on accurate COVID-19 molecular detection, which is found to be a challenge for many clinical laboratories (Tang et al. 2020).

The coronavirus diseases belong to the Coronaviridae family and the Nidovirales order, comprising the alpha, beta, gamma, and delta variants. The term "Corona" describes the crown-like spikes on the surface of the virus, which are approximately 65-125 nm in diameter with a single-stranded RNA ranging between 26–32 kbs in lengths (Ansori et al. 2020). In 2003 and 2019, the first and second outbreaks of the viruses (SARS-CoV-1 and 2) occurred and infected approximately 8,098 and 27.89 million people, respectively (Shereen et al. 2020), leading to SARS-CoV-2 being found to be more infectious. Moreover, the structural proteins of this virus are encoded by four genes, namely the envelope (*E*), membrane (*M*), nucleocapsid (*N*), and spike (*S*). In addition, the *ORF1ab* is the largest gene in the SARS-CoV-2 virus (Ansori et al. 2020).

The first Indonesian case of SARS-CoV-2 infection was reported on March 2, 2020, in two Depok, DKI Jakarta residents that had interacted with some Japanese citizens. This gradually began to emerge and was identified in various provinces, leading to a high mortality rate. As of May 9, 2020, Indonesia confirmed and declared a total of 13,645 COVID-19 cases, as well as 959 deaths and 2,607 curable patients (Turista et al. 2020).

According to Djalante et al. (2020), the COVID-19 molecular detection in Indonesia was limited due to its reliance on a rapid serological test, which had several advantages capable of detecting SARS-CoV-2 antigens, such as (1) cost efficiency, (2) quick data output, and (3) the ability of a novice to handle the process without adequate knowledge, as opposed to other molecular techniques. One of the major disadvantages associated with this test was its low sensitivity and production of false-negative results.

Therefore, the molecular detection of the COVID-19 pandemic needs to be augmented in Indonesia using the reverse transcription-quantitative polymerase chain reaction (RT-qPCR) from a nasopharyngeal swab. The target of some genes to prohibit the production of inaccurate result swabs is also needed (Scohy et al. 2020). Furthermore, primers play an essential role in the amplification of PCR, leading to the need for proper design and validation in wet laboratory studies.

Indonesia is the fourth most populous country in the world with expansive geography. According Djalante et al. (2020), the COVID-19 pandemic affected this country through further mutations for extended periods due to its environmental conditions. These mutations quickly occurred in the RNA viruses than the DNA, based on the single-stranded mutants mutating faster than the double-stranded forms (Sanjuán and Domingo-Calap 2016). Moreover, several RT–qPCR assays have been used by clinical, study, and public health laboratories, with the test reliability also dependent on the primer, specifically detecting its target. This is because the SARS-CoV-2-specific primers should be 100% specific for the virus, leading to the sole amplification of the viral sequences. Furthermore, this amplification is reportedly detected in the exponential phase, indicating that the PCR amplicon is being synthesized (Bustin and Nolan 2020).

Primers are the single and most critical component of a reliable RT-qPCR assay due to their properties controlling the exquisite specificity and sensitivity of this robust method. Also, several viral genomic regions are targeted by this assay, with multiple primer designs focusing on the same genes, including the *RdRP* (RNA-dependent RNA polymerase gene), *E*, and *N* groups (Bustin and Nolan 2020). The *S* gene is also usually used for the target due to its capability to mutate at any time. For instance, four mutations (N501Y, 69-70del, K417N, and E484K) were successfully detected by Vega-Magaña et al. (2021) through a low-cost RT-qPCR assay. However, the lack of designed primers and the failure to optimize reaction conditions potentially led to false-negative results. In addition, the variable RNA sequences within the *ORF1ab* and *N* genes provided negative or low sensitivity results (Wang et al. 2020).

The SARS-CoV-2 in various countries presently varies due to mutation. According to Mercatelli and Giorgi (2020) and Hamed et al. (2021), the L clade was the reference cluster based on the GISAID database. It was also differentiated into three major groups, namely G, V, and S clades. The G clade has a variation in the spike region of aspartic acid to glycine, at a position of 614, such as D614G. The variations of the V clade at L37F and G251V is further found in the NSP6 and NS3 respectively, while the S group has a variation on ORF8 at L84S. Furthermore, the derivatives of clade G include the GR, GV, and GH, which specifically has variations at D614G and NS3-Q57H positions. According to Korber et al. (2020) and Githinji et al. (2021), the D614G mutation spread early during the pandemic and became the dominant global variant. The GR and GV also had variations at N-G204R and S-A222V positions, respectively. Meanwhile, the genomes that did not belong to the three major groups had the "O clade" designation (Hamed et al. 2021).

The SARS-CoV-2 viruses are presently found to vary in several Indonesian regions due to mutation. Therefore, the phylogenomic analysis should be carried out to properly analyze the relationships between all these viruses and their epidemiology towards acquiring specific primers with the ability to bind the gene targets. In this study, the primer candidates for COVID-19 molecular detection are designed based on the phylogenomic relationships, using the ML method. The results showed that the Indonesian SARS-CoV-2 clades were separated from the Wuhan reference sequence. It also showed that three pairs of primers specifically bonded to the viral sequences, based on the in-silico test. In subsequent tests, this should be applied to a larger amount of virally extracted clinical RNA samples to verify the specificity and sensitivity of the study. This is useful in obtaining the primer sets representing the distribution of all the SARS-CoV-2 clades in Indonesia.

2. Materials and Methods

2.1. Genome sequences data mining of SARS-CoV-2 Indonesia

The SARS-CoV-2 genomes in Indonesia were obtained from the GISAID platform (www.gisaid.org). A total of 38 complete genomes (±29 kbps) located in various Indonesian provinces were downloaded in FASTA format on September 10, 2020, to conduct the ML-based phylogenomic analysis. Furthermore, the study established a sequence with Reference Genome: NC 045512.2, using the Wuhan SARS-CoV-2 downloaded from the NCBI database (ncbi.nlm.nih.gov). The clade produced from the analysis was further evaluated, with one of the genomes selected to design its primers (accession ID: hCoV-19/Indonesia/JB-TFRIC19-R49544/2020) found to be targeting the *RdRp, S,* and *N* genes.

2.2. ML phylogenomic analysis

This method was carried out in the Virus Pathogen Resource (ViPR) web tool (www.vipr.org), using the PhyML algorithm at the Generate Phylogenetic Tree menu. The PhyML was associated with a phylogenetic topology and the application of a substitution model to the nucleotide sequences. Furthermore, the algorithm was applied to different datasets, namely (1) less than 100 long sequences and (2) 100 to 1,000 small or medium length sequences (Guindon et al. 2005). An Interactive Tree of Life software, iTOL (https://itol.embl.de), containing the XML file format, was downloaded and used for further illustrative purposes. In addition, the SARS-CoV-2 Wuhan sequence was re-rooted on the phylogenomic tree, as the relationship between all genomes and their clades was analyzed and determined.

2.3. Gene annotation

This method was conducted in the Virus Pathogen Resource (ViPR) web tool (www.vipr.org). The process was carried out by uploading a FASTA format genome sequence on the Workbench menu through the VIGOR4 Genome Annotator. In addition, the FASTA format of *RdRp, S,* and *N* was obtained to acquire each primer.

2.4. Primer design

This method was carried out in the Virus Pathogen Resource (ViPR) web tool (www.vipr.org), at the PCR Primer Design menu. For the S primer set design, a single mutation variant of 23403A>G (p.D614G) was specifically selected from a dominant clade through the phylogenomic analysis. A total of five primer pairs were analyzed using the Oligonucleotide Properties Calculator (OligoCalc) web tool (http://biotools.nubic.northwestern.edu/OligoCalc.html). Also, appropriate primer pairs were used to determine the parameter standards.

2.5. In silico PCR simulation and validation

The specificity of the primers was simulated and validated using the genome browser (https://genome.ucsc.edu/cgi-bin/hgPcr) of the UCSC (University of California Santa Cruz). The validation of the PCR primers was determined using the MSA (Multiple Sequence Alignment) of the reference genome (SARS-CoV-2 Wuhan), and the other sequence samples. This method was carried out through the ViPR software, while the clustal file format was downloaded and visualized using the Jalview tool.

3. Results and Discussion

3.1. ML phylogenomic analysis

The phylogenomic analysis is useful for conceptualizing, visualizing, and analyzing biological genealogies, with its trees often designed to determine the probable evolutionary relationships between species. It also used input data in the form of genome sequences. Furthermore, several computational methods are observed for phylogenetic trees, based on distance (e.g. Unweighted-Pair Group Method With Arithmetic Means (UPGMA) and Neighbor-Joining (NJ)) and character (including ML and maximum parsimony (MP)) sequences, respectively. The Distance-Based Method is carried out by grouping the Operational Taxonomy Units (OTU) due to the similarities per gene site and time. The level of similarity between the two OTUs is often found to be high, indicating that the genetic distance is not very far. This method assumed that the same mutation rate between sequences was inaccurate. To overcome this problem, the Sequence Character-Based Method was designed, focusing more on the level of similarity between sequences based on the rate of evolution (Van de Peer and Salemi 2012).

One of the methods of phylogenomic analysis is ML, which was based on the evolutionary relationship of several sequences, according to Selberg et al. (2021). The probability estimation of this method was also calculated at each sequence alignment position, leading to the determination of mutational presence or absence. Using the alignment results on the phylogenomic tree, the ancestor amino acid sequences and the location of specific mutations were found to occur.

The ML method reconstructed the ancestor across all tree nodes, as the calculations of the algorithm included (1) determining the branch lengths and optimal substitution models from several types of provided tree topologies and (2) determining a tree topology that provides the highest likelihood score (similar), based on the substitution model parameters (Dhar and Minin 2016). This branch length indicated the significance of the topology and counted the mutations in each sequence. Moreover, the ML method accurately characterized the tested sequences using the highest probability number model (Selberg et al. 2021).

The ML is a technique comprising quantitative probabilities and mutational events. It is based on evolutionary sequences and uses the probability rate as a mathematical model to construct the phylogenomic tree topology. In addition, the probability rate of this method is estimated in numerous alignment positions while considering the level of mutational events (Makarenkov et al. 2006).

The substitution model of the ML was used to describe the rates of change of fixed mutations among sequences. This constituted the basis of evolutionary analysis and genetic data at the molecular level. With the emergence of other methods, numerous substitution models are still presently existing to copy the DNA process and evolution of the complex sequence datasets. This model reportedly used mathematical equations comprising probability and statistics (e.g., the Log-likelihood value is represented as a probability). Therefore, a more significant negative value leads to a higher probability of selecting phylogenomic topology construction (Arenas 2015). Furthermore, the Generalized Time Reversible (GTR) is one of the complex substitution models with better real data, comprising different rates and nucleotide frequencies (Tavaré 1986). Figure 1 shows the ML phylogenomic tree.

Based on the available GISAID sequences on September 10, 2020, this study initially determined the phylogenomic analysis of 38 Indonesian SARS-CoV-2 genomes from mid-March to July 25, 2020. Afterward, the sequences initially isolated within East Java in April 2020, which emerged as a distinct phylogenomic clade, were subsequently analyzed. Compared to other provinces, the Wuhan SARS-CoV-2 is a group located in East Java. In DKI Jakarta, it was closely related to the virus within the Special Region of Yogyakarta, with varying isolated clades of one month and sixteen days. This group contained O clade, indicating that the virus had unknown mutation sites known as ”Another”.

Based on further studies, the strains were distinguished by a derived missense mutation in the spike protein (S-protein) encoding gene, leading to an amino acid change from aspartate to a glycine residue, at a position 614 (D614G). This result showed that the D614G mutation in

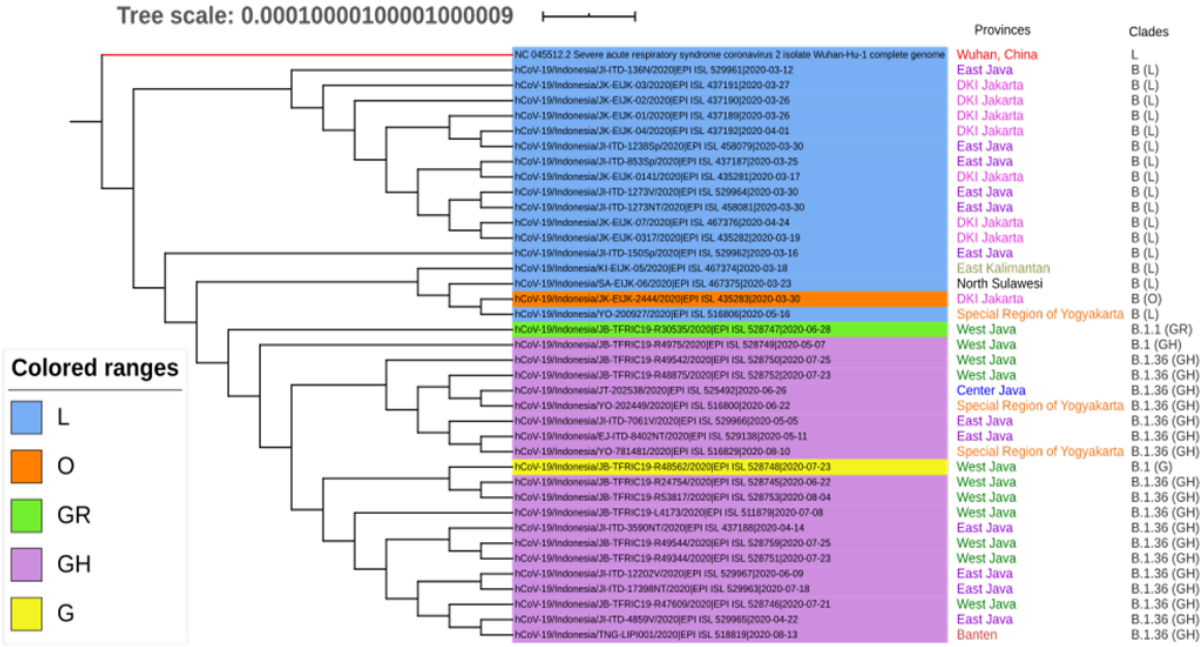


FIGURE 1 ML phylogenomic tree of SARS-CoVs-2 in each province in Indonesia.

Indonesia occurred in the GH clade, with the highest frequency of 19/38 (50%). According to Figure 1, there were five clades of SARS-CoVs-2 located in Indonesia, namely L, G, GH, GR, and O. The compositions of GH, GR, G, and O were 26.3% respectively, while L was 44.7%. In addition, the Wuhan sequence was closely related to the East Java virus (12 March 2020 EPI ISL 529961).

3.2. Primer design

According to the phylogenomic analysis, the GH clade was known as the newly isolated sample obtained from a patient. This led to the designation of a genome sequence with a newer sample isolation date through the primers. Furthermore, the S-site played an essential role in clade formation, while GH, G, and GR were mutated on the D614G region. The study of Eaaswarkhanth et al. (2020), stated that D614G changed the nucleotide confirmation and was the most infectious SARS-CoV-2. The *RdRp* and *N* encoding genes were also used for primer design due to being more conserved than other groups. These *RdRp* genes were located in the *ORF1ab* region, based on being more conserved in RNA virus and ideal for understanding evolutionary patterns. Due to the high rate of error during the copying process, proofreading the exonuclease activity was impossible (Venkataraman et al. 2018).

The S-region is a gene that encodes the S-protein, which is often integrated over the virus surface. It is also known as the most mutative site, compared to other SARS-CoV-2 encoding genes. This was due to its ability to mediate the attachment of the virus to the surface receptors of the host and the fusion between the viral and cell membranes (Boopathi et al. 2020). SARS-CoV-2 is known to enter the cells through the bond of this part to the host receptor, namely the angiotensin-converting enzyme (ACE2), which is present in many human organs and tissues (Turista et al. 2020). Therefore, the S-glycoprotein played a vital role in binding to the host cell receptors. It was also the major target for neutralizing antibodies (Ansori et al. 2020).

The N-gene is a helical ribonucleoprotein comprising positive-strand RNA genomes and monomers of a single N protein-bound (Kuo and Masters 2003). This coated the viral RNA genome and played an essential role in its replication and transcription (Boopathi et al. 2020). According to Wurm et al. (2001), N-protein induced a cell cycle delay in the G2/M phase. Furthermore, the quality of each primer was checked using the OligoCalc web tool, with the sets used to appropriately target all genes towards determining the parameter standard, as shown in Table 1.

The functions of the *In Silico* PCR menu were observed by analyzing a sequence database with a pair of primers. Moreover, Table 2 showed that the product's size or amplicon lengths were 232, 256, and 260 bp for *RdRp*, *S*, and *N* genes, respectively. The location of each gene was also amplified in vitro. Therefore, all the primer sets in this study successfully passed the simulation process. The search further conducted a sequence output file in FASTA format, containing all the database groups and primer pairs.

The FASTA body was capitalized in areas where the primer sequence matched the database, while the lower case was used elsewhere. The primer sets were also validated using the MSA method to align the reference genome and target products. According to Figures 2, 3, and 4, all primer sets targeted *RdRp*, *S*, and *N*, respectively, to specifically bind and complement the genome regions. Also, a primer set of the *RdRp* gene was used to identify and amplify the mutation site in all viral clades.

TABLE 1 Primer sets and their specifications targeted *RdRp*, *S*, and *N* encoding genes compared with Li et al. (2020) and Davi et al. (2021) references.

Primers	Parameter								
	Nucleotide Length	Primer Length (bp)	GC Content(%)	GC Clamp	Melting Temperature (Tm) °C	Repetitive Sequence	Self-complementarity	3' Self complementarity	Hairpins
Primers Result									
Forward Primer RdRp 5' → 3' TTGTGTATGCTGCTGACCCT	14241-14261	20	50	G=0, C=3	59.308 (Ta= 54.308)	None	None	None	None
Reverse Primer RdRp 5' → 3' GCAGCATTACCATCCTGAGC	14503-14523	20	55	G=2, C=1	59.05v(Ta= 54.05)	None	None	None	None
Forward Primer S 5' → 3' ACTGATGCTGTCCGTGATCC	1717-1737	20	55	G=1, C=2	59.823 (Ta = 54.823)	None	None	None	None
Reverse Primer S 5' → 3' TGTTGACATGTTCAGCCCCT	1953-1973	20	50	G=0, C=4	59.522 (Ta = 54.522)	None	None	None	None
Forward Primer N 5' → 3' TTGGTTCACCGCTCTCACTC	153-173	20	55	G=0, C=3	59.966 (Ta = 54.966)	None	None	None	None
Reverse Primer N 5' → 3' CTCCCTCAGTTGCAACCCAT	393-413	20	55	G=0, C=3	59.961 (Ta = 54.961)	None	None	None	None
Primer Sequences Reference from Li et al. (2020)									
Forward Primer RdRp 5' → 3' CAAGTG GGGTAAGGCTAG ACTTT	14961-14983	23	48	G=0, C=1	55.3 (Ta= 50.3)	None	None	None	5' CAAGTGGGGTAAG-GCTAGACTTT 3'
Reverse Primer RdRp 5' → 3' ACTTAGGATAATCCCAACCCAT	15283-15304	22	41	G=0, C=3	51.1 (Ta= 46.1)	None	None	None	None
Forward Primer S (Ref. Seq. MN938384) 5' → 3' CCTACTAAAT-TAAATGATCTCTGCTTTACT	22712-22741	30	55	G=1, C=2	59.823 (Ta = 54.823)	None	None	None	5' GCTGGTGCT-GCAGCTTATTA 3'
Reverse Primer S (Ref. Seq. MN975262) 5' → 3' CAAGCTATAACGCAGCCTGTA	22849-22869	21	48	G=1, C=1	52.4 (Ta = 47.4)	None	None	None	None
Forward Primer N (Ref. Seq. MN908947) 5' → 3' GGGGAACTTCTCCTGCTAGAAT	28881-28902	22	50	G=1, C=0	54.8 (Ta = 49.8)	None	5' GGGGAACTTCTC-CTGCTAGAAT 3' 3' TAAGATCGTC-CTCTTCAAGGGG 5' 5' GGGGAACTTCTC-CTGCTAGAAT 3' 3' TAAGATCGTC-CTCTTCAAGGGG 5'	None	None
Reverse Primer N 5' → 3' CTCCCTCAGTTGCAACCCAT	28958-28979	22	45	G=0, C=3	53 (Ta = 48)	None	None	None	None
Primer Sequences Reference from Davi et al. (2021)									
Germany, 17 January 2020									
Forward Primer RdRp RdRP SARSr-F2 GTG ARATGGTCATGT GTGGCGG		22	57.14	G=4, C=1	63.25 (Ta= 58.25)	None	None	None	None
Reverse Primer RdRp RdRP SARSr-R1 CAR GCAATGTA TTAAASACACTATTA		26	25	G=0, C=0	54.25 (Ta= 49.25)	None	None	None	None
Paris (2 March 2020)									
Forward Primer RdRp nCoV IP2-12669Fw ATGAGCTTAGTCCTGTTG		18	44.44	G=2, C=0	51.11 (Ta= 46.11)	None	None	None	None
Reverse Primer RdRp nCoV IP2-12759Rv CTCCCTTTGTTGTGTTGT		18	44.44	G=2, C=0	52.57 (Ta= 47.57)	None	None	None	None
Forward Primer RdRp nCoV IP4-14059Fw GGTAACTGGTATGATTTCG		20	42.11	G=1, C=1	50.65(Ta= 45.65)	None	None	None	None
Reverse Primer RdRp nCoV IP4-14146Rv CTGGTCAAGGTTAATATAGG		19	40	G=2, C=0	49.98 (Ta= 44.98)	None	None	None	None
Japan (24 January 2020)									
Forward Primer S WuhanCoV-spk1-f TTGTTTGCAAAATTCAAGACTCAC		24	33.33	G=0, C=3	58.02 (Ta= 53.02)	None	None	None	None
Reverse Primer S WuhanCoV-spk1-f TGTTGTGGTGTCATAAAAATTCCTT		25	32	G=0, C=2	56.98 (Ta= 51.98)	None	None	None	None
Forward Primer S_NIID WH-1 F24381 TCAAGACTCACTTTCTTCCAC		21	42.86	G=0, C=3	55.48 (Ta= 50.48)	None	None	None	5' TCAAGACT-CACTTTCTTC-CAC 3'
Reverse Primer S_NIID_WH-1_R24873 ATTTGAAACAAAGACACCTTCAC		23	34.78	G=0, C=2	56.13 (Ta= 51.13)	None	None	None	5' ATTTGAAA-CAAAGACAC-CTTCAC 3' 5' ATTTGAAA-CAAAGACAC-CTTCAC 3'
Forward Primer S_NIID_WH-1_Seq_f3F24383 AAGACTCACTTTCTTCCACAG		21	42.86	G=1, C=2	55.47 (Ta= 50.47)	None	None	None	5' AAGACT-CACTTTCTTC-CACAG 3'
Reverse Primer_S_NIID_WH-1_Seq_R24865 CAAAGACACCTTCACGAGG		19	52.63	G=3, C=1	55.88 (Ta= 50.88)	None	None	None	None

TABLE 1 (continued).

Forward Primer N_NIID_2019-nCOV_N_F2 AAATTTTGGGGACCAGGAAC	20	45	G=2, C=1	56.09 (Ta= 51.09)	None	None	None	None
Reverse Primer N_NIID_2019-nCOV_N_R2 TGGCAGCTGTGTAGGTCAAC	20	55	G=0, C=2	60.25 (Ta= 55.25)	None	None	None	None
Thailand (23 January 2020)								
Forward Primer N_WH-NIC N-F CGTTTGGTGGACCCTCAGAT	20	55	G=1, C=1	59.68 (Ta= 55.68)	None	None	None	None
Reverse Primer N_WH-NIC N-R CCCCACTGCGTTCTCCATT	19	57.89	G=0, C=2	60 (Ta= 55)	None	None	None	None
Germany (17 January 2020)								
Forward Primer N N_Sarbeco_F1 CACATTGGCACCCGCAATC	19	57.89	G=0, C=1	60.15(Ta= 55.15)	None	None	None	None
Reverse Primer N N_Sarbeco_R1 GAGGAACGAGAAGAGGCTTG	20	55	G=2, C=1	58 (Ta= 53)	None	None	None	None
Hong Kong (23 January 2020)								
Forward Primer N HKU-NF 5' → 3' TAATCAGACAAGGAACTGATTA	22	31.82	G=1, C=0	52.27 (Ta= 47.27)	None	5' TAATCAGACAAG-GAACTGATTA 3' 3' ATTAGTCAAG-GAACAGACTAAT 5' 5' TAATCAGACAAG-GAACTGATTA 3' 3' ATTAGTCAAG-GAACAGACTAAT 5'	None	5' TAATCAGA-CAAGGAACT-GATTA 3'
Reverse Primer HKU-NR 5' → 3' CGAAGGTGTGACTTCCATG	19	52.63	G=1, C=2	55.95 (Ta= 50.95)	None	None	None	5' CGAAGGTGT-GACTTCCATG 3'
China (24 January 2020)								
Forward Primer N_F 5' → 3' GGGGAACTTCTCCTGCTAGAAT	22	50	G=1, C=0	59.23 (Ta= 54.23)	None	5' GGGGAACTTCTC-CTGCTAGAAT 3' 3' TAAGATCGTC-CTCTTCAAGGGG 5' 5' GGGGAACTTCTC-CTGCTAGAAT 3' 3' TAAGATCGTC-CTCTTCAAGGGG 5'	None	None
Reverse Primer N_R 5' → 3' CAGACATTTTGCTCTCAAGCTG	22	45.45	G=2, C=1	58.18 (Ta= 53.18)	None	None	None	None
USA (24 January 2020)								
Forward Primer N_2019-nCoV_N1-F_ GACCCCAAAATCAGCGAAAT	20	45	G=1, C=0	56.67 (Ta= 51.67)	None	None	None	None
Reverse Primer N_2019-nCoV_N1-R TCTCTGGGTTACTGCCAGTTGAAT	24	45.83	G=1, C=0	60.8 (Ta= 55.8)	None	None	None	5' TCTCTGGGT-TACT-GCCAGTTGAAT 3'
Forward Primer N_2019-nCoV_N2-F TTACAAACATTGGCCGCAAA	20	40	G=1, C=1	57.11 (Ta= 52.11)	None	None	None	None
Reverse Primer N_2019-nCoV_N2-R GCGCGACATTCCGAAGAA	18	55.56	G=1, C=0	58.53 (Ta= 53.53)	None	None	None	None
Forward Primer N_2019-nCoV_N3-F GGGAGCCTTGAATACACCAAAA	22	45.45	G=0, C=1	58.84 (Ta= 53.84)	None	None	None	None
Reverse Primer_N_2019-nCoV_N3-R TGTAGCACGATTGCAGCATTG	21	47.62	G=1, C=1	59.87 (Ta= 54.87)	None	None	None	None

Based on Figure 2, a white mark indicated that the designed *RdRp* primer set was used to detect a C-to-T mutation at nucleotides of 14,636. This slightly corresponded to Korber et al. (2020), which stated that "The D614G change was often accompanied by three reactions, i.e., C-to-T mutations at the 5' UTR (position 241 relative to the Wuhan reference sequence), as well as the 3,037 and 14,408 positions, therefore leading to an amino acid alteration in the RNA-dependent RNA polymerase (*RdRp* P323L)". According to Figure 3, the primer set of the S gene was marked with white color to identify the D614G mutation site, where adenine mutated into guanine at nucleotides of 23,401. This was in line with Korber et al. (2020), which stated that "The Spike D614G amino acid change was caused by an A-to-G nucleotide mutation at position 23,403 in the Wuhan reference strain. It was also the only region that met the threshold criterion and was initially identified in early March 2020". Moreover, the mutation site made a difference in distinguishing the SARS-CoV-2 virus in Wuhan and Indonesia, with the N primer set targeting the consensus sequence.

Numerous clades of SARS-CoV-2 have globally emerged with the L-clade, known as the first spread in January-February (Umair et al. 2021). This study led to a specific theory, which stated that the *S* primer set was used as a mutation tracker, indicating that its inability to bind the gene template during in vitro experiment was held, leading to the formation of new transformations. Further-

TABLE 2 *In silico* simulation result of each primer targeted *RdRp*, *S*, and *N* encoding genes.

Primer sets	Genes Target	Size product (bp)	Target product
Forward Primer 5'→3' TTGTGTATGCTGCTGACCCT	RdRp	232	>NC_045512v2:14555+14786 232bp TTGTGTATGCTGCTGACCCT GCAGCATTACCATCCTGAGC TTGTGTATGCTGCTGACCCTgctatgcacgctgcttctggtaatctatta ctagataaacgcactacgtgcttttcagtagctgcacttactaacaatgt
Reverse Primer 5' →3' GCAGCATTACCATCCTGAGC	RdRp	232	tgcttttcaaactgtcaaacccggtaattttaacaaagacttctatgact ttgctgtgtctaagggtttctttaaggaaggaagttctgttgaattaaaa cacttcttctttGCTCAGGATGGTAATGCTGC
Forward Primer 5'→ 3' ACTGATGCTGTCCGTGATCC	S	256	>NC_045512v2:23279+23534 256bp ACTGATGCTGTCCGTGATCC TGTTGACATGTTCAGCCCCT ACTGATGCTGTCCGTGATCCacagacacttgagattcttgacattacacc atgttcttttggtggtgtcagtgttataacaccaggaacaaatacttcta accaggttgctgttctttatcaggatgttaactgcacagaagtccctgtt
Reverse Primer 5' → 3' TGTTGACATGTTCAGCCCCT	S	256	gctattcatgcagatcaacttactcctacttggcgtgtttattctacagg ttctaatgtttttcaaacacgtgcaggctgtttaatAGGGGCTGAACATG TCAACA
Forward Primer 5' → 3' TTGGTTCACCGCTCTCACTC	N	260	>NC_045512v2:28426+28685 260bp TTGGTTCACCGCTCTCACTC CTCCCTCAGTTGCAACCCAT TTGGTTCACCGCTCTCACTCaacatggcaaggaagaccttaaattccctc gaggacaaggcgttccaattaacaccaatagcagtccagatgaccaaatt
Reverse Primer 5' → 3' CTCCCTCAGTTGCAACCCAT	N	260	ggctactaccgaagagctaccagacgaattcgtggtggtgacggtaaaat gaaagatctcagtccaagatggtatttctactacctaggaactgggccag aagctggacttccctatggtgctaacaaagacggcatcatATGGGTTGCA ACTGAGGGAG

more, a mutation tracker was used to indicate when the qRT PCR method was carried out in Indonesia. Therefore, the *RdRp* and *N* were obtained when the *S* encoding gene was not detected, with the virus being categorized as a new clade. Meanwhile, the virus was categorized as L clade when all of the gene targets were detected due to the designed S primer set being sequentially similar with the Reference Genome SARS-CoV-2 Wuhan NC 045512.2 (Figure 3). This was supported by Korber et al. (2020), which reported that the D614G GISAID G clade carried three

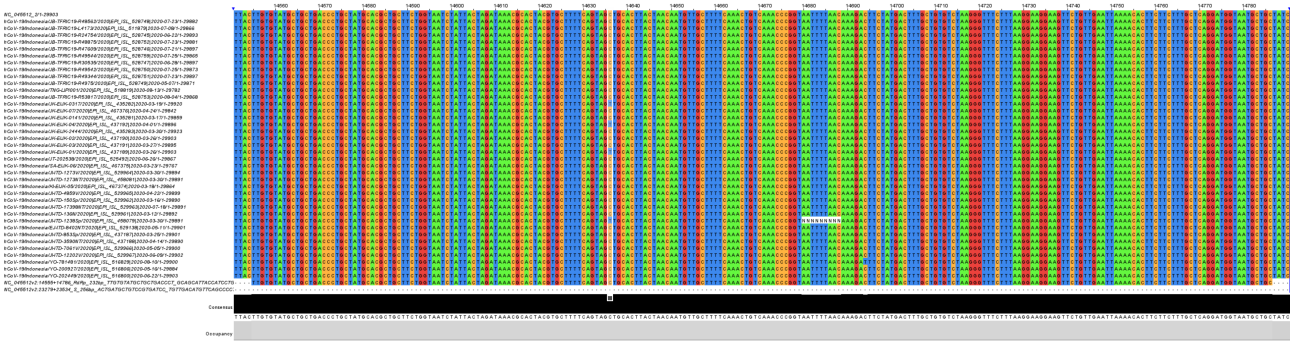


FIGURE 2 MSA result of primer set targeting *RdRp* encoding genes. The *RdRp* primer set is presented on below. Yellow, green, red, and blue color indicate cytosine, adenine, guanine, and thymine nucleotide, respectively. A white mark indicates that the designed *RdRp* primer set can detect a C-to-T mutation at position 14,636 of nucleotides.

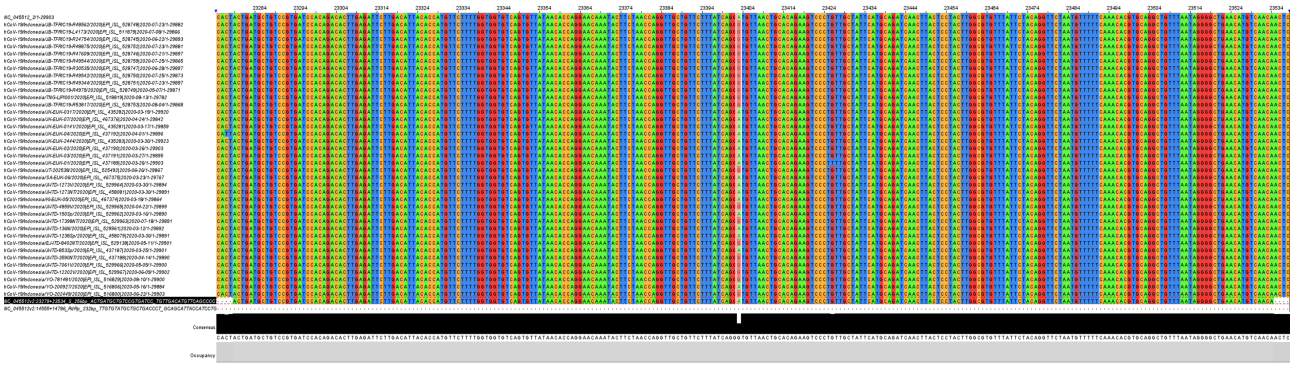


FIGURE 3 MSA result of primer set targeting *S* encoding genes. The *S* primer set is presented on below. The D614G mutation site is marked with white highlight in which adenine mutates into guanine. Yellow, green, red, and blue color indicate cytosine, adenine, guanine, and thymine nucleotide, respectively

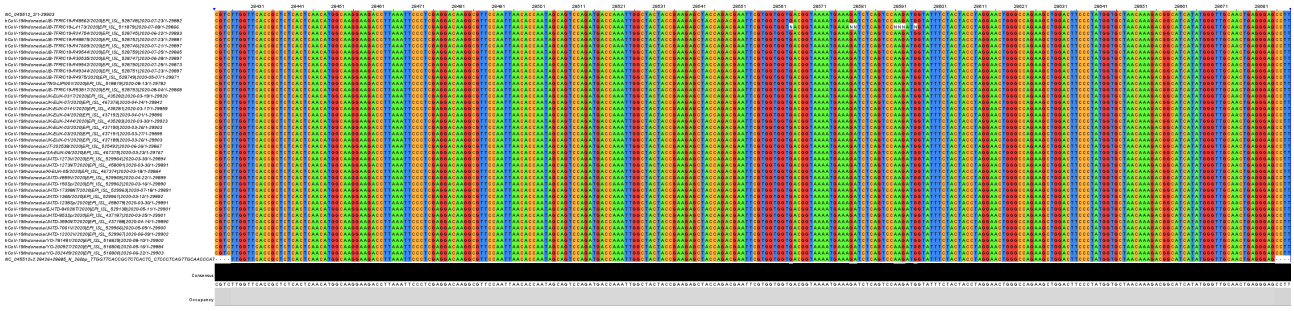
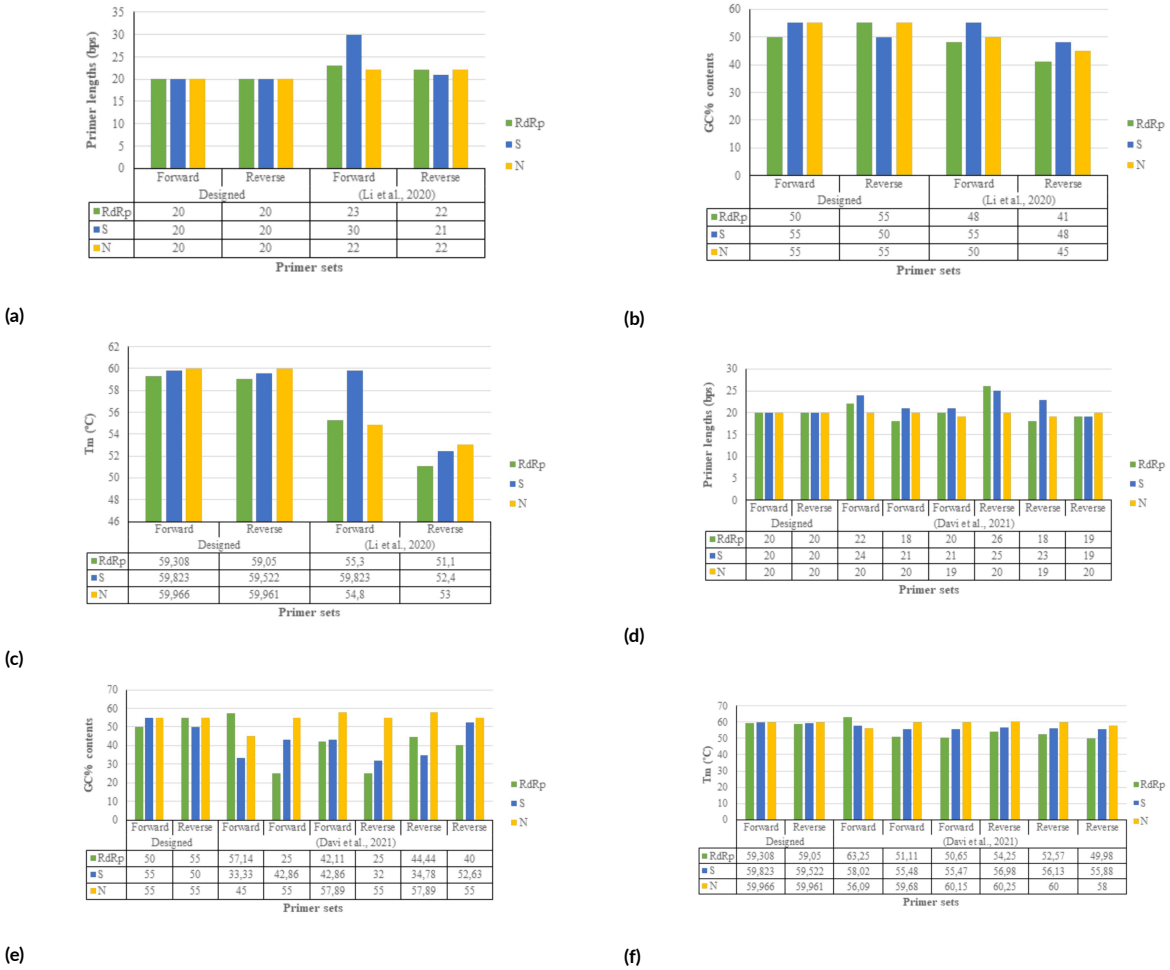


FIGURE 4 MSA result of primer set targeting *N* encoding genes. The *N* primer set is presented on below. It can be targeted exactly to the consensus sequence. Yellow, green, red, and blue color indicate cytosine, adenine, guanine, and thymine nucleotide, respectively

of the four mutational determinants. These three clades occurred in the *RdRp* protein (nucleotide C14408T producing a P323L amino acid change), with one each at the Spike (nucleotide A23403G resulting in the D614G amino acid change) and silent (C3037T) regions, respectively. This was in line with Islam et al. (2021), which concluded that the positive amplification of the wild-type-targeting primers was determined as the L clade. It also indicated that other types were determined based on the selected sequence target sites.

These results were compared with the studies of Li et al. (2020) and Davi et al. (2021), which contained



(a)	Designed		(Li et al., 2020)	
	Forward	Reverse	Forward	Reverse
RdRp	20	20	23	22
S	20	20	30	21
N	20	20	22	22

(b)	Designed		(Li et al., 2020)	
	Forward	Reverse	Forward	Reverse
RdRp	50	55	48	41
S	55	50	55	48
N	55	55	50	45

(c)	Designed		(Li et al., 2020)	
	Forward	Reverse	Forward	Reverse
RdRp	59,308	59,05	55,3	51,1
S	59,823	59,522	59,823	52,4
N	59,966	59,961	54,8	53

(d)	Designed		(Davi et al., 2021)					
	Forward	Reverse	Forward	Forward	Forward	Reverse	Reverse	Reverse
RdRp	20	20	22	18	20	26	18	19
S	20	20	24	21	21	25	23	19
N	20	20	20	20	19	20	19	20

(e)	Designed		(Davi et al., 2021)					
	Forward	Reverse	Forward	Forward	Forward	Reverse	Reverse	Reverse
RdRp	50	55	57,14	25	42,11	25	44,44	40
S	55	50	33,33	42,86	42,86	32	34,78	52,63
N	55	55	45	55	57,89	55	57,89	55

(f)	Designed		(Davi et al., 2021)					
	Forward	Reverse	Forward	Forward	Forward	Reverse	Reverse	Reverse
RdRp	59,308	59,05	63,25	51,11	50,65	54,25	52,57	49,98
S	59,823	59,522	58,02	55,48	55,47	56,98	56,13	55,88
N	59,966	59,961	56,09	59,68	60,15	60,25	60	58

FIGURE 5 The comparison of primer parameters visualized in histograms. (a) The histogram of lengths between designed primers and Li et al. (2020). (b) The histogram of GC content, indicating that the value on Li et al. (2020) were more various than the designed primers. (c) The histogram of Tm, indicating that the values at designed primers were almost similar to each other. (d) The histogram of lengths between designed primer sets and Davi et al. (2021), indicating that the structures did not vary. (e) The histogram of GC% content indicating that some primers had very low compositions, leading to the inability to strongly bind to the target. (f) The histogram of Tm parameter indicating that the designed primers had similar range (approximately 59˚C), leading to easier PCR reaction settings than Davi et al. (2021).

the targeting genes of a SARS-CoV-2 genome located in China and other countries. The targeted *RdRp*, *S*, and *N* genes with specifications were shown in Table 1, while Figure 5 was used to illustrate the associated histograms. These results showed that some of the primer parameters and probes listed in the previous studies of Li et al. (2020) and Davi et al. (2021) should be enhanced. They also indicated that the sets used in this study did not have secondary structures. However, these parameters had similar primer lengths (Figure 5a and 5d), with varying GC contents between 50-55% (Figure 5b). According to Li et al. (2020), the Tm range of each primer set failed to significantly diverge *in silico*. This indicated that the primer sets contained a hairpin structure at *RdRp* and *S*, with a self-complementary design at N. There was also a low GC content value at *RdRp* reverse value (41%), indicating its inappropriateness to the standard parameter, as well as the ability to split the primer and gene targets. In addition, the primer lengths were more varied at S (30 bp), indicating that it is miss-priming. According to Davi et al. (2021), some primer sequences had self-complementarity and hairpin structures. Most of them also had very low GC % content (about 25-33.33%), leading to miss-priming or inability to perfectly bind to the targets. Based on the parameter comparisons, this study obtained adequate and standard indicators through the *in silico*. At a low-cost SYBR Green, a non-specific probe qRT PCR detection method was obtained, as previously tested and stated by Pereira-Gómez et al. (2021).

4. Conclusions

There were five clades of Indonesian SARS-CoV-2 virus in this study, with the primer sets obtained from *RdRp*, *S*, and *N* genes. The ML also showed that the GH clade had the highest value, with the designed *S* primer set being a mutation tracker. Moreover, all the designed sets of this study were better than the previous reports, as observed from the secondary structure indicator standard of primers.

Acknowledgments

This study was supported by the Faculty of Science and Technology, Biotechnology Study Program, Universitas' Aisyiyah Yogyakarta, Indonesia.

Authors' contributions

TTN, APJ designed the study. TTN, ARW, AK funding acquisition. TTN, ARW, APJ carried out the data analysis. TTN, ARW, APJ, AK wrote the manuscript. AK, APJ reviewed the manuscript. All authors read and approved the final version of the manuscript

Competing interests

The authors declare no competing interest.

References

Ansori ANM, Kharisma VD, Antonius Y, Tacharina MR, Rantam FA. 2020. Immunobioinformatics analysis and phylogenetic tree construction of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Indonesia: spike glycoprotein gene. J. Teknol. Lab. 9(1). doi:10.29238/teknolabjournal.v9i1.221.

Arenas M. 2015. Trends in substitution models of molecular evolution. Front. Genet. 6(OCT). doi:10.3389/fgene.2015.00319.

Boopathi S, Poma AB, Kolandaivel P. 2020. Novel 2019 coronavirus structure, mechanism of action, antiviral drug promises and rule out against its treatment. doi:10.1080/07391102.2020.1758788.

Bustin SA, Nolan T. 2020. RT-QPCR testing of SARS-COV-2: A primer. Int. J. Mol. Sci. 21(8). doi:10.3390/ijms21083004.

Davi MJP, Jeronimo SM, Lima JP, Lanza DC. 2021. Design and in silico validation of polymerase chain reaction primers to detect severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Sci. Rep. 11(1). doi:10.1038/s41598-021-91817-9.

Dhar A, Minin V. 2016. Maximum Likelihood Phylogenetic Inference. Oxford: Academic Press. doi:https://doi.org/10.1016/B978-0-12-800049-6.00207-9. URL https://www.sciencedirect.com/science/article/pii/B9780128000496002079.

Djalante R, Lassa J, Setiamarga D, Sudjatma A, Indrawan M, Haryanto B, Mahfud C, Sinapoy MS, Djalante S, Rafliana I, Gunawan LA, Surtiari GAK, Warsilah H. 2020. Review and analysis of current responses to COVID-19 in Indonesia: Period of January to March 2020. Prog. Disaster Sci. 6. doi:10.1016/j.pdisas.2020.100091.

Eaaswarkhanth M, Al Madhoun A, Al-Mulla F. 2020. Could the D614G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality? Int. J. Infect. Dis. 96. doi:10.1016/j.ijid.2020.05.071.

Githinji G, de Laurent ZR, Mohammed KS, Omuoyo DO, Macharia PM, Morobe JM, Otieno E, Kinyanjui SM, Agweyu A, Maitha E, Kitole B, Suleiman T, Mwakinangu M, Nyambu J, Otieno J, Salim B, Kasera K, Kiiru J, Aman R, Barasa E, Warimwe G, Bejon P, Tsofa B, Ochola-Oyier LI, Nokes DJ, Agoti CN. 2021. Tracking the introduction and spread of SARS-CoV-2 in coastal Kenya. Nat. Commun. 12(1). doi:10.1038/s41467-021-25137-x.

Guindon S, Lethiec F, Duroux P, Gascuel O. 2005. PHYML Online - A web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. 33(SUPPL. 2). doi:10.1093/nar/gki352.

Hamed SM, Elkhatib WF, Khairalla AS, Noreddin AM. 2021. Global dynamics of SARS-CoV-2 clades and their relation to COVID-19 epidemiology. Sci. Rep. 11(1). doi:10.1038/s41598-021-87713-x.

Islam MT, Alam ARU, Sakib N, Hasan MS, Chakrovarty

T, Tawyabur M, Islam OK, Al-Emran HM, Jahid MIK, Anwar Hossain M. 2021. A rapid and cost-effective multiplex ARMS-PCR method for the simultaneous genotyping of the circulating SARS-CoV-2 phylogenetic clades. J. Med. Virol. 93(5). doi:10.1002/jmv.26818.

Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans CM, Freeman TM, de Silva TI, Angyal A, Brown RL, Carrilero L, Green LR, Groves DC, Johnson KJ, Keeley AJ, Lindsey BB, Parsons PJ, Raza M, Rowland-Jones S, Smith N, Tucker RM, Wang D, Wyles MD, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire EO, Montefiori DC. 2020. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell 182(4). doi:10.1016/j.cell.2020.06.043.

Kuo L, Masters PS. 2003. The Small Envelope Protein E Is Not Essential for Murine Coronavirus Replication. J. Virol. 77(8). doi:10.1128/jvi.77.8.4597-4608.2003.

Li D, Zhang J, Li J. 2020. Primer design for quantitative real-time PCR for the emerging Coronavirus SARS-CoV-2. Theranostics 10(16):7150–7162. doi:10.7150/thno.47649.

Makarenkov V, Kevorkov D, Legendre P. 2006. Phylogenetic Network Construction Approaches, volume 6. Elsevier. doi:https://doi.org/10.1016/S1874-5334(06)80006-7.

Mercatelli D, Giorgi FM. 2020. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Front. Microbiol. 11. doi:10.3389/fmicb.2020.01800.

Pereira-Gómez M, Fajardo Á, Echeverría N, López-Tort F, Perbolianachis P, Costábile A, Aldunate F, Moreno P, Moratorio G. 2021. Evaluation of SYBR Green real time PCR for detecting SARS-CoV-2 from clinical samples. J. Virol. Methods 289. doi:10.1016/j.jviromet.2020.114035.

Sanjuán R, Domingo-Calap P. 2016. Mechanisms of viral mutation. doi:10.1007/s00018-016-2299-6.

Scohy A, Anantharajah A, Bodéus M, Kabamba-Mukadi B, Verroken A, Rodriguez-Villalobos H. 2020. Low performance of rapid antigen detection test as frontline testing for COVID-19 diagnosis. J. Clin. Virol. 129. doi:10.1016/j.jcv.2020.104455.

Selberg AG, Gaucher EA, Liberles DA. 2021. Ancestral Sequence Reconstruction: From Chemical Paleogenetics to Maximum Likelihood Algorithms and Beyond. doi:10.1007/s00239-021-09993-1.

Shereen MA, Khan S, Kazmi A, Bashir N, Siddique R. 2020. Covid-19 infection: origin, transmission, and characteristics of human coronaviruses. Https://doi.org/10.1016/j.jare.2020.03.005 65. World Health Organization (WHO).2020. Advice on the use of masks in the c. J. Adv. Res. 24.

Tang YW, Schmitz JE, Persing DH, Stratton CW. 2020. Laboratory diagnosis of COVID-19: Current issues and challenges. doi:10.1128/JCM.00512-20.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences.

Turista DDR, Islamy A, Kharisma VD, Ansori ANM. 2020. Distribution of COVID-19 and phylogenetic tree construction of sars-CoV-2 in Indonesia. J. Pure Appl. Microbiol. 14. doi:10.22207/JPAM.14.SPL1.42.

Umair M, Ikram A, Salman M, Khurshid A, Alam M, Badar N, Suleman R, Tahir F, Sharif S, Montgomery J, Whitmer S, Klena J. 2021. Whole-genome sequencing of SARS-CoV-2 reveals the detection of G614 variant in Pakistan. PLoS One 16(3 March). doi:10.1371/journal.pone.0248371.

Van de Peer Y, Salemi M. 2012. Phylogenetic inference based on distance methods. Cambridge University Press. doi:10.1017/cbo9780511819049.007.

Vega-Magaña N, Sánchez-Sánchez R, Hernández-Bello J, Venancio-Landeros AA, Peña-Rodríguez M, Vega-Zepeda RA, Galindo-Ornelas B, Díaz-Sánchez M, García-Chagollán M, Macedo-Ojeda G, García-González OP, Muñoz-Valle JF. 2021. RT-qPCR Assays for Rapid Detection of the N501Y, 69-70del, K417N, and E484K SARS-CoV-2 Mutations: A Screening Strategy to Identify Variants With Clinical Impact. Front. Cell. Infect. Microbiol. 11. doi:10.3389/fcimb.2021.672562.

Venkataraman S, Prasad BV, Selvarajan R. 2018. RNA dependent RNA polymerases: Insights from structure, function and evolution. doi:10.3390/v10020076.

Wang Y, Kang H, Liu X, Tong Z. 2020. Combination of RT-qPCR testing and clinical features for diagnosis of COVID-19 facilitates management of SARS-CoV-2 outbreak. doi:10.1002/jmv.25721.

Wurm T, Chen H, Hodgson T, Britton P, Brooks G, Hiscox JA. 2001. Localization to the Nucleolus Is a Common Feature of Coronavirus Nucleoproteins, and the Protein May Disrupt Host Cell Division. J. Virol. 75(19). doi:10.1128/jvi.75.19.9345-9356.2001.