Genes & Genomics (2022) 44:899–912
https://doi.org/10.1007/s13258-022-01264-7

Online ISSN 2092-9293
Print ISSN 1976-9571

RESEARCH ARTICLE

Identification of conserved regions from 230,163 SARS-CoV-2 genomes and their use in diagnostic PCR primer design

Haeyoung Jeong[1] · Siseok Lee[2] · Junsang Ko[2] · Minsu Ko[2] · Hwi Won Seo[1]

Received: 3 March 2022 / Accepted: 3 May 2022 / Published online: 2 June 2022

Abstract

Background As the rapidly evolving characteristic of SARS-CoV-2 could result in false negative diagnosis, the use of as much sequence data as possible is key to the identification of conserved viral sequences. However, multiple alignment of massive genome sequences is computationally intensive.

Objective To extract conserved sequences from SARS-CoV-2 genomes for the design of diagnostic PCR primers using a bioinformatics approach that can handle massive genomic sequences efficiently.

Methods A total of 230,163 full-length viral genomes were retrieved from the NCBI SARS-CoV-2 Resources and GISAID EpiCoV database. This number was reduced to 14.11% following removal of 5′-/3′-untranslated regions and sequence dereplication. Fast, reference-based, multiple sequence alignments identified conserved sequences and specific primer sets were designed against these regions using a conventional tool. Primer sets chosen among the candidates were evaluated by in silico PCR and RT-qPCR.

Results Out of 17 conserved sequences (totaling 4.3 kb), two primer sets targeting the nsp2 and ORF3a genes were picked that exhibited > 99.9% in silico amplification coverage against the original dataset (230,163 genomes) when a 5% mismatch between the primers and target was allowed. In addition, the primer sets successfully detected nine SARS-CoV-2 variant RNA samples (Alpha, Beta, Gamma, Delta, Epsilon, Zeta, Eta, Iota, and Kappa) in experimental RT-qPCR validations.

Conclusion In addition to the RdRp, E, N, and S genes that are targeted commonly, our approach can be used to identify novel primer targets in SARS-CoV-2 and should be a priority strategy in the event of novel SARS-CoV-2 variants or other pandemic outbreaks.

Keywords SARS-CoV-2 · Multiple sequence alignment · RT-qPCR

Introduction

The ongoing global COVID-19 pandemic, caused by a positive-sense, single-stranded RNA virus named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has disrupted everyday life beyond public health and the economy in the two years since its initial discovery in China. Together with vaccines and therapeutics, diagnostics are key tools for controlling the COVID-19 pandemic. PCR-based molecular diagnoses have superior sensitivity and specificity and are therefore favored over antigen tests (Lisboa Bastos et al. 2020). Reverse-transcription quantitative PCR (RT-qPCR) is used more widely than other molecular diagnostic methods such as digital PCR or loop-mediated isothermal amplification (LAMP). As various RT-qPCR primer–probe sets developed by various groups across the world perform differently, a comparative study has recommended using a combination of sets from different institutions (Jung et al. 2020). Despite its narrow dynamic range and high cost, droplet digital PCR has a relatively low dependency on the primer–probe sets and is therefore a more reliable method of quantifying viral nucleic acid targets than RT-qPCR (Park et al. 2021).

Since SARS-CoV-2 is a rapidly evolving virus, emerging variants harboring nucleotide changes in the primer-binding

✉ Haeyoung Jeong
hyjeong@kribb.re.kr

✉ Hwi Won Seo
seohw01@kribb.re.kr

[1] Infectious Disease Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 34141, Republic of Korea

[2] NanoHelix Co., Ltd. 43-15, Daejeon 34014, Republic of Korea

sites might evade detection, resulting in low diagnostic sensitivity. The ORF1ab, E, and N genes are conserved among members of the *Sarbecovirus* subgenus and have been used as the major targets for COVID-19 diagnoses (Li et al. 2020). Conserved sequences (CSs), which are usually identified from multiple sequence alignments (MSAs), are invaluable resources for successful PCR primer design because they can cover multiple variants. To increase the sensitivity of PCR analyses, degenerate primers can also be used to amplify multiple SARS-CoV-2 genotypes (Li et al. 2020), with MSA as the starting point. However, the rapid growth of publicly available SARS-CoV-2 genome sequences has made it difficult to implement such a straightforward strategy. As of December 10, 2021, the NCBI SARS-CoV-2 Resources website (https://www.ncbi.nlm.nih.gov/sars-cov-2/) lists 2,641,217 nucleotide records, a number that has increased by almost 60-fold over a 1-year period. Moreover, the Global Initiative for Sharing All Influenza Data (GISAID) (Shu and McCauley 2017) EpiCoV database (https://www.epicov.org/) holds approximately 6,000,000 viral sequence submissions. Several primer design software packages can take a MSA as input (Perini et al. 2020; Yoon and Leitner 2015), but these packages are not able to deal with massive alignment data. MRPrimerV (Kim et al. 2017), a database containing PCR primer pairs for the detection of 1,818 RNA virus, cannot be used for SARS-CoV-2 diagnoses because it has not been updated since 2016, despite the recent update of the primer design engine from MapReduce to a GPU-based pipeline (Bae et al. 2021).

MSA, a fundamental step in biological sequence analysis, is computationally demanding. Standard MSA methods such as MUSCLE (Edgar 2004), MAFFT (Katoh and Standley 2013), and Clustal (Sievers and Higgins 2018) scale poorly with increasing numbers of input sequences, such that alignment of hundreds of thousands of viral genome sequences is not feasible on a Linux desktop computer. A recent study reported the extraction of CSs from Clustal Omega-generated MSAs of SARS-CoV-2 genomes (Davi et al. 2021), but only 2,341 complete sequences were available at that time. Reference-guided MSA tools such as VIRULIGN (Libin et al. 2019) and ViralMSA (Moshiri 2021) run much faster than the standard methods and can be used for a large-scale viral genome analysis. Qu et al. (2022) described a different strategy for PCR primer design that does not depend on massive MSA of SARS-CoV-2 genomes. In this approach, viral genomes are aligned to a reference sequence to identify mutations, primer pairs are designed using the PRIMER3 software package (Untergasser et al. 2012) with the 'PRIMER_LOWERCASE_MASKING = 1' tag to avoid mismatches at the 3'-ends by using a soft-masked template sequence at the mutated sites, and then the coverages and specificities of the designed primer sets are re-evaluated using MFEprimer (Wang et al. 2019). Since this approach does not take into account the frequency of mutations at a given site, usable primers that can anneal to most of the viral genome dataset might be rejected at the design step.

Pretreatment of viral genome sequences can reduce the effective data size prior to MSA. For example, filters can be applied to retrieve complete high-quality viral genomes from the database and reject inadequate sequences that might lead to spurious analysis results. More importantly, sequence dereplication is required because multiple viral isolates originating from a local outbreak can have identical genome sequences, leading to redundant submissions and a concomitant increase in the available data. Indeed, dereplication is a popular sequence pretreatment technique that is used widely in 16S rRNA-based microbial profiling and (meta)genomics analyses to reduce the data size and/or identify representative members of a sequence cluster.

In this study, we identified 17 CSs (> 150 bp) from a reference-guided MSA of 32,483 SARS-CoV-2 genomes that were obtained via dereplication of 230,163 full-length viral sequences. Subsequently, we designed primer sets based on the CSs. For experimental validation, two primer sets were chosen among the candidates, and nine major variants were identified successfully using real-time quantitative PCR.

Materials and methods

Sequence data collection and manipulation

Most of the SARS-CoV-2 genome sequences (218,799) were retrieved from the NCBI SARS-CoV-2 Resources database on July 9, 2021 (collection period: December 31, 2020 to July 1, 2021; dataset: 'NCBI-all'). The following filters were applied: 'ambiguous character: max 290' and 'RefSeq genome/nucleotide completeness: complete.' On the same day, all viral sequences submitted from South Korea (4931) were downloaded from the GISAID EpiCoV database, applying the 'complete,' 'high coverage,' and 'low coverage excl' filters (dataset: 'GISAID-S. Korea'). Complete Delta variant sequences (6621; AY.3, AY.3.1, AY.4, AY.5, AY.6, AY.7, AY.9, AY.10, AY.12, AY.25, and B.1.617.2; dataset: 'NCBI-Delta') were retrieved from the NCBI SARS-CoV-2 Resources database on September 2, 2021, with the same filters activated regardless of the collection period. Sequences from these three datasets were chunked into 10,000-seq units and aligned with the 29,490 bp 'core region' of the SARS-CoV-2 reference genome sequence (RefSeq NC_045512.2) using the 'nucmer -maxmatch' command (Kurtz et al. 2004). The core region of the reference sequence was defined as the region spanning nucleotides 266 (start codon of ORF1ab) through 29,674 (stop codon of ORF10) and was therefore devoid of 5′- and 3′-untranslated regions. Because untranslated regions often have extra

sequence variations or low-quality nucleotides, they were trimmed off all sequences based on the nucmer alignment coordinates against the core region of the reference. Based on the sequence accessions, 188 redundant Delta variants were removed from the first dataset. The number of total sequences at this stage was 230,163, comprising 218,611 'NCBI-non-Delta,' 4,931 'GISAID-S. Korea,' and 6621 'NCBI-Delta' sequences.

Dereplication of viral genome sequences

The three trimmed datasets were dereplicated separately using the 'vsearch –derep_fulllength' command (Rognes et al. 2016). In principle, singletons were removed from the major dataset ('NCBI-non-Delta'), because they might have resulted from sequencing errors, but were not removed from the remaining two datasets (Korean isolates and Delta variants). The finalized datasets were combined to produce a single dataset comprising 32,483 sequences (85.8% reduction in data size). Pango lineage was assigned to each genome sequence using Pangolin (https://cov-lineages.org/resources/pangolin.html) v3.1.11 with Pango-designation v1.2.123, followed by hierarchical visualization using Krona Tools (Ondov et al. 2011).

MSA and CS extraction

Using NC_045512.2 as the reference, MSAs of the 32,483 sequences were carried out using ViralMSA v1.1.16 (Moshiri 2021). To extract CS position information, ClipKIT v1.1.5 (Steenwyk et al. 2020) was then applied to the aligned FASTA file with the '-m kpic-smart-gap' option that retains the union of parsimony informative and constant sites. From the log file, the positions of contiguous nucleotides longer than 150 bp were calculated using a custom Perl script. Conservations per column, defined as the percentages of the most frequent bases at a given position (taking gapped sequences into account), were calculated directly from the MSA using BioPerl (Stajich et al. 2002) Bio::AlignIO. For better visualization, less conserved reference positions (conservation < 99%) were masked with lowercase bases as described previously (Qu et al. 2022), and CSs were extracted from the masked reference and CS position information. The normalized Shannon entropy of each nucleotide position was calculated from the MSA using the ANDES tool (Li et al. 2010). Consensus sequences were also extracted from the MSA using the EMBOSS 'cons' program (Rice et al. 2000) and were compared with corresponding CSs.

Primer design and in silico validation

Primer and probe candidates were designed for each CS using Primer3 (Untergasser et al. 2012). The specificities of the primers and probes were assessed by Primer-BLAST (Ye et al. 2012) comparisons against the complete RefSeq RNA databases for *Homo sapiens* (taxid: 9606), bacteria (taxid: 2), fungi (taxid: 4751), and Apicomplexa (taxid: 5794). OligoAnalyzer online software (Integrated DNA Technologies, Coralville, IA, USA) was used to check for primer dimers and secondary structure formation.

The EMBOSS 'PrimerSearch' program (Rice et al. 2000) was used to check primer pairs against the three SARS-CoV-2 genome datasets used in this study, with 0%, 5%, and 10% mismatches between the primers and templates allowed. For comparison, nucleotide sequences from 16 primer sets that are either used widely worldwide or have been developed by domestic researchers were included in the in silico analysis (Corman et al. 2020; Won et al. 2020).

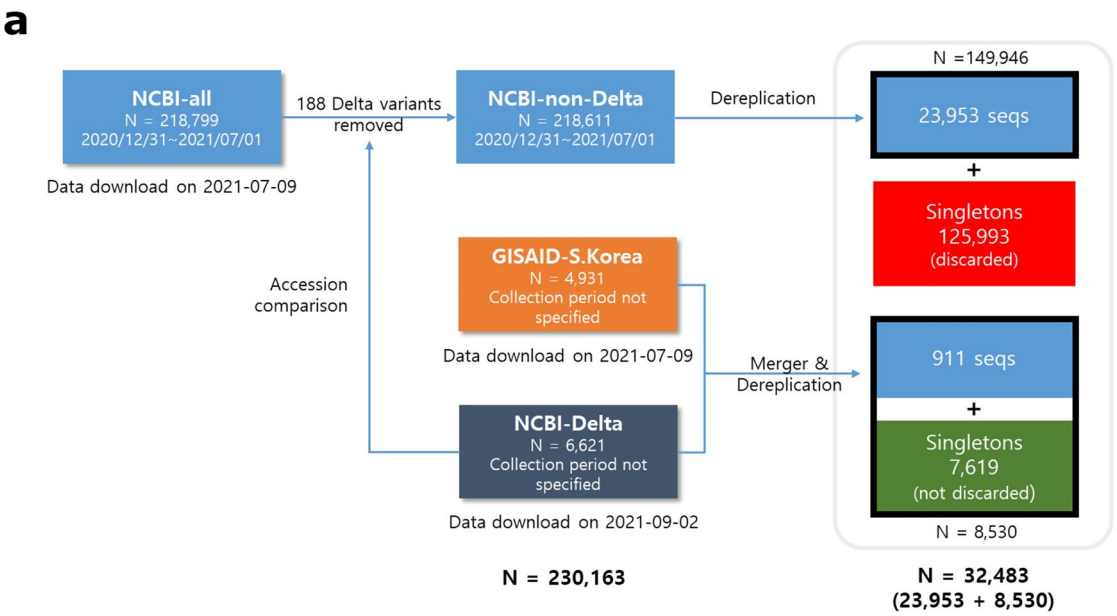
Experimental validation using RT-qPCR

All primers and probes were synthesized by Bioneer Inc. (Daejeon, Republic of Korea). The One-Step RT-qPCR Kit v.6 (Nanohelix, Daejeon, Republic of Korea) was used according to the manufacturer's instructions. The RT-qPCR reaction mixture (20 µL) included the following reagents: 4 µL of 5 × buffer mix, 2 µL of enzyme mix, 1.5 µL of primer/probe mix, 9.5 µL of nuclease-free water, and 3 µL of template RNA. Multiplex RT-qPCR was performed by serial dilutions of SARS-CoV-2 RNA templates (10^2, 10^3, and 10^4 copies per reaction) using specific primers for nsp2 and ORF3a targets and probes labeled with JOE and Texas Red, respectively on a CFX96 Real-Time PCR Detection System (Bio-Rad Laboratories, Hercules, CA, USA), with cycling conditions as follows: 50 °C for 10 min, 95 °C for 3 min, and then 42 cycles of 95 °C for 1 s and 60 °C for 5 s. All SARS-CoV-2 variant RNA samples used in this study were obtained from the National Culture Collection for Pathogens (NCCP; Cheongju-si, Chungcheongbuk-do, Republic of Korea).

Results

Collection and pretreatment of SARS-CoV-2 genome sequences

The viral genome sequences used in this study were collected initially as three separate datasets ('NCBI-all,' 'GISAID-S. Korea,' and 'NCBI-Delta') (Fig. 1a). The first dataset, generated by searching the NCBI SARS-CoV-2 Resources database, included 218,799 sequences and



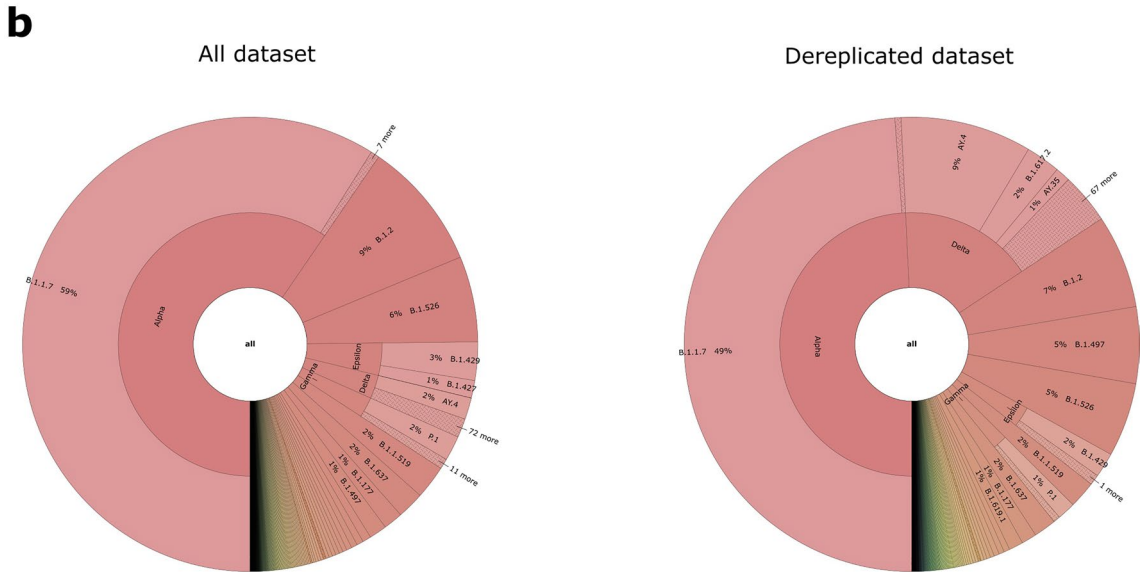


Fig. 1 The workflow for collection and manipulation of SARS-CoV-2 genomes. **a** Data collection and dereplication process. 32,483 complete viral sequences in the thick-lined boxes (right) were taken as the final dataset. **b** The Pango lineage distributions of the 'all' dataset (comprising the initial 'NCBI-non-Delta,' 'GISAID-S. Korea,' and 'NCBI-Delta' datasets; N = 230,163) and the dereplicated dataset (N = 32,483). Note that singletons were removed from the NCBI-non-Delta dataset after dereplication

accounted for 54.5% of all SARS-CoV-2 genomes (with filters activated) available on the data collection date. Because only 30 Korean isolates were included in the first dataset, the second dataset was generated by downloading 4931 South Korean viral sequences from the GISAID EpiCoV database, with no date restriction. There might be redundant Korean isolate records that had been submitted to both NCBI and GISAID, but they could not be removed due to the absence of information linking two datasets.

At the time of data download and analysis, the Delta variant accounted for the majority of COVID-19 cases worldwide.[1] However, only 185 genomes from the 'NCBI-all'

[1] CNBC News. WHO says delta variant accounts for 99% of Covid cases around the world. https://www.cnbc.com/2021/11/16/who-says-delta-variant-accounts-for-99percent-of-covid-cases-around-the-world.html.

dataset were assigned a Delta annotation (Pango lineage B.1.617.2; no AY lineages). Because 230 of the 218,799 records in the 'NCBI-all' dataset were probably recently submitted sequences and lacked a Pango lineage designation, Pangolin (v3.1.11) was executed over the same dataset. This approach increased the number of sequences belonging to Delta variants only slightly, to 193. Consequently, we generated the third dataset ('NCBI-Delta') by collecting all Delta variant genomes (6621) from the NCBI SARS-CoV-2 Resources NCBI database at a later date (September 2, 2021; about 2 months after the 'NCBI-all' dataset was generated). With a sequence quality filter applied, we identified 17,188 genomes that were deposited during the 2 months, of which 6436 (37.4%) were Delta variants. Based on the sequence accessions, 188 Delta variants were removed from the 'NCBI-all' dataset to generate the 'NCBI-non-Delta' dataset (comprising 218,611 sequences).

Dereplication reduces the total amount of sequence data markedly, enabling subsequent analysis processes to be performed more efficiently. In our current analysis, dereplication reduced the number of 'NCBI-non-Delta' sequences from 218,611 to 149,946 (31.4% reduction), of which 125,993 sequences were singletons. The representative sequence of the largest cluster (size = 366) was MZ336928.1, belonging to the B.1.1.7 lineage (Alpha variant). Since 57.6% of the total sequences in the 'NCBI-non-Delta' dataset were singletons that might have originated from sequencing errors or very rare variants, we discarded them, leaving 23,953 sequences in the dereplicated 'NCBI-non-Delta' dataset. In addition, the combined total of 11,552 sequences in the 'GISAID-S. Korea' and 'NCBI-Delta' datasets was reduced to 8530 sequences (26.2% reduction) after dereplication. Singletons, which accounted for 66.0% of this combined dataset, were not removed to preserve original data of importance (domestic isolates and Delta variants) as much as possible. Therefore, the combined dereplicated dataset comprised a total of 32,483 sequences (23,953 + 8530). Thirteen sequences were found to be identical to the reference after trimming, the representative ones being MZ093199.1 (size = 8, from the first dataset) and EPI_ISL426169_2020-02–05 (size = 5, from the combined second and third datasets). Figure 1b shows a comparison of the Pango lineage distributions before and after the dereplication and differential removal of singletons. This treatment, applied separately to datasets, slightly affected Pango lineage composition of the sequences. As expected, the proportion of Delta variants increased to approximately 16% after dereplication and singleton removal.

MSAs and extraction of CS candidates

Reference-guided MSAs of the 32,483 dereplicated sequences using the ViralMSA tool took less than 1 min on a Linux server with two Xeon E5-2640 @2.5 GHz CPUs (using 16 threads out of 24) and 128 GB memory. Considering Davi et al. (2021) reported that a supercomputer was required for the alignment of only 2,341 full viral genome sequences using Clustal Omega, ViralMSA would be the best choice for a PC-level computer. When the MAFFT (Katoh and Standley 2013) v7.490 program was used with the FFT-NS-2 option, it took more than 3 days to complete the alignment on the same computer.

Initially, we tried to use the well-known alignment trimmer trimAl (Capella-Gutierrez et al. 2009) to extract positional information on CSs from the FASTA-format MSA file (928 MB). In principle, alignment trimming/editing software packages do not simply select constant sites or CSs only, because they are not very informative for subsequent analyses such as phylogenetic tree generation. Thus, CS 'candidates' would be the correct term to describe the output of post-MSA process and its manipulation. Running trimAl with the '-automated1' option (a heuristic trimming method), as described previously (Davi et al. 2021), discarded only a few sites from the original alignments and generated six subsequences covering 29,384 bp of the original 29,903 bp, suggesting that parameter optimization was required. However, as a single-threaded program, the prolonged execution time for a typical trimAl run (~ 16 h) made it unfeasible to tweak the parameters to prioritize conserved sites. Furthermore, the proportion of identical nucleotides at a given site cannot be translated simply into trimAl parameters such as a residue similarity score, nor can it be inferred easily from the result files. Consequently, ClipKIT (Steenwyk et al. 2020) was chosen for this purpose because it runs much faster than trimAl and produces a comprehensive log file that contains information about all positions in the input MSA.

Although it is not impossible to extract constant sites through the manipulation of ClipKIT log files, extraction of constant sites literally yielded only fragmentary sequences, such that primer design was impractical. When we checked the ClipKIT log file for contiguous constant sites (gappyness < 0.001) to find out 'true' CSs that showed constant nucleotides across all genomes, the longest segment was only 14 bp long. Thus, ClipKIT was executed with one of the recommended trimming mode ('-m kpic-smart-gap' option), and based on the kept (non-discarded) site information, nucleotide sequences of 17 CS regions longer than 150 bp (max 503 bp; average 260.6 bp; total 4431 bp) were extracted from the soft-masked reference sequence (Table 1). When the CSs were compared to the corresponding *consensus* sequences, most were identical, and only two regions exhibited differences of 1 bp (Table 1, rightmost column). Applying the '-m kpic-gappy-g 0.05' or '-m kpic-gappy' (default gappyness threshold: 0.9)' option for ClipKIT execution did not produce substantially different results

Table 1 List of 17 conserved sequences identified from multiple sequence alignments of 32,483 SARS-CoV-2 genomes

ID	Position (length in bp)	Sequence	Gene (product)	Identity with consensus sequence
CS_1	2576–2836 (261)	AGTGAAGCTGTTGAAGCTCCAtTGGTTGGTACACCAGTTTGTATTA ACGGGCTTATGTTG CTCGAAATCAAAGACACAGAAAAGTACTGTGCCCTTGCACCTAAT ATGATGGTAACAAAC AATACCTTCACACTCAAAGGCGGTGCACCAACAAAGGTTACTTTT GGTGATGAcACTGTG ATAGAAGTGCAAGGTTACAAGAGTGTGAATATCACTTTTGAACTT GATGAAAGGATTGAT AAAGTACTTAATGAGAAGTGC	ORF1ab (nsp2..nsp3)	261/261 (100%)
CS_2	4508–4670 (163)	GTGGTTGATTATGGTGCTAGATTTTACTTTTACACCAGTAAAACA ACTGTAGCGTCACTT ATCAACACACTTAACGATCTAAATGAAACTCTTGTTACAATGCCA CTTGGCTATGTAACA CATGGCTTAAATTTGGAAGAAGCTGCTCGGTATATGAGATCTC	ORF1ab (nsp3)	163/163 (100%)
CS_3	6830–7332 (503)	AGAATTAAAGCATCTATGCCGACTACTATAGCAAAGAATACTGTT AAGAGTGTCGGTAAA TTTTGTCTAGAGGCTTCATTTAATTATTTGAAGTCACCTAATTTTTCT AAACTGATAAAT ATTAtAATTTGGTTTTTACTATTAAGTGTTTGCCTAGGTTCTTTAAT CTACTCAACCGCT GCTTTAGGTGTTTTAATGTCTAATTTAGGCATgCCTTCTTACTGTACT GGTTACAGAGAA GGCTATTTGAACTCTAcTAATGTCACTATTGCAACCTACTGTACTGG TTCTATAcCTTGT AGTGTTTGTCTTAGTGGTTTAGATTCTTTAGACACCTATCCTTCTTTA GAAACTATACAA ATTACCATTTCaTCTTTTAAATGGGATTTAACTGCTTTTGGCTTAGTT GCAGAGTGGTTT TTGGCATATATTCTTTTCACTAGGTTTTTCTATGTACTTGGATTGGCT GcAATCATGCAA TTGTTTTTCAGCTATTTTGCAGT	ORF1ab (nsp3)	502/503 (99%)
CS_4	8628–8903 (276)	ATTTAATAACACCTGTTCATGTCATGTCTAAACATACTGACTTTT CAAGTGAAATCATAG GATACAAGGCTATTGATGGTGGTGTCACTCGTGACATAGCATCTA CAGATACTTGTTTTG CTAACAAACATGCTGATTTTGACACATGGTTTAGCCAGCGTGGTG GTAGTTATACTAATG AcAAAGCTTGCCCATTGATTGCTGCAGTCATAACAAGAGAAGTG GGTTTTGTCGTGCCTG GTTTGCCTGGCACGATATTACGCACAACTAATGGTG	ORF1ab (nsp4)	276/276 (100%)

Table 1 (Continued)

ID	Position (length in bp)	Sequence	Gene (product)	Identity with consensus sequence
CS_5	16,509–16,758 (250)	TTTATATAAAAATACATGTGTTGGTAGCGATAATGTTACTGACTT TAATGCAATTGCAAC ATGTGACTGGACAAATGCTGGTGATTACATTTTAGCTAACACCTG TACTGAAAGACTCAA GCTTTTTGCAGCAGAAACGCTcAAAGCTACTGAGGAGACATTTAAA CTGTCTTATGGTAT TGCTACTGTACGTGAAGTGCTGTCTGACAGAGAATTACATCTTTC ATGGGAAGTTGGTAA ACCTAGACCA	ORF1ab (RdRp..helicase)	250/250 (100%)
CS_6	17,716–17,993 (278)	GGCGTGGTAAGAGAATTCCTTACACGTAACCCTGCTTGGAGAAAA-GcTGTCTTTATTTCA CCTTATAATTCACAGAATGCTGTAGCCTCAAAGATTTTGGGACTA CCAACTCAAACTGTT GATTCATCACAGGGCTCAGAATATGACTATGTCATATTCACTCAA ACCACTGAAACAGCT CACTCTTGTAATGTAAACAGATTTAATGTTGCTATTACCAGAGCA AAAGTAGGCATACTT TGCATAATGTCTGATAGAGACCTTTATGACAAGTTGCA	ORF1ab (helicase)	278/278 (100%)
CS_7	19,253–19,570 (318)	TGCTATCTAACCTTAACTTGCCTGGTTGTGATGGTGGCAGTttgtatg-taaataaacatg cattccacacaccagcttttgataaaagtgcttttgttaatttaaaacaattaccatttt tctattactctgacagtccatgtgagtctcatggaaaacaagtagtgtcagatatagatt atgtaccactaaagtctgctacgtgtataacacgttgcaatttaggtggtgctgtctgta gacatcatgctaatgagtacagattgtatctcgatgcttataacatgatgatctcagctg gctttagctTGTGGGTTT	ORF1ab (3′-to-5′ exonuclease)	318/318 (100%)
CS_8	19,923–20,235 (313)	TTGTTCTATGACTGACATAGCCAAGAAACCAACTGAAACgATTTGT GCACCACTCACTGT CTTTTTTGATGGTAGAGTTGATGGTCAAGTAGACTTATTTAGAAA TGCCCGTAATGGTGT TCTTATTACAGAAGGTAGTGTTAAAGGTTTACAACCATCTGTAGG TCCCAAACAAGCTAG TCTTAATGGAGTCACATTAATTGGAGAAGCCGTAAAAACACAGTT CAATTATTATAAGAA AGTTGATGGTGTTGTCCAACAATTACCTGAAACTTACTTTACTCA GAGTAGAAATTTACA AGAATTTAAACCC	ORF1ab (endoRNAse)	313/313 (100%)

Table 1 (Continued)

ID	Position (length in bp)	Sequence	Gene (product)	Identity with consensus sequence
CS_9	20,238–20,483 (246)	GAGTCAAATGGAAATTGATTTCTTaGAATTaGCTATGgATGAATTCA TTGAACGGTATAA ATTAGAAGGCTATGCCTTCGAACATATCGTTTATGGAGATTTTAG TCATAGTCAGTTAGG TGGTTTACATCTACTGATTGGACTAGCTAAACGTTTTAaGGAATC ACCTTTTGAATTAGA AGATTTTATTCCTATGGACAGTACAGTTAAAAACTATTTCATAAC AGATGCGCAAACAGG TTCATC	ORF1ab (endoRNAse)	246/246 (100%)
CS_10	21,169–21,376 (208)	ATAACAGAACATTCTTGGAATGCTGATCTTTATAAGCTCATG GGaCACTTCGCATGGTGG ACAGCCTTTGTTACTAATGTGAATGCgTCATCATCTGAAGCATTTTTA ATTGGATGTAAT TATCTTGGCAAACCAcGcGAACAAATAGATGGTTATGTCATGCAT GCAAATTACATATTT TGGAGGAATACAAATCCAATTCAGTTGT	ORF1ab (2′-*O*-ribose methyltransferase)	208/208 (100%)
CS_11	21,771–21,990 (220)	TCTCTGGGACCAATGGTACTAaGAGGTTTGaTAACCCTGTcCTACCA TTTAATGATGGTG TTTAtTTTGCTTCCAcTGAGAAGTcTAACATAATAAGAGGCTGGAT TTTTGGTACTACTT TAGATTCGAAGACCCAGTCCCTACTTATTGTTAATAacgCTACTA Atgttgttattaaag tctgtgaatttcaattttgtaatgatccatttttgggtgt	S (spike glycoprotein)	220/220 (100%)
CS_12	22,325–22,542 (218)	TCTTCAGGTTGGACAGCTGGTGCTGCAGCTTATTATGTGGGTTAT CTTCAACCTAGGACT TTTCTATTAAAATATAATGAAAATGGAACCATTACAGATGCTGTA GACTGTGCACTTGAC CCTCTCTCAGAAACAAAGTGTACGTTGAAATCCTTCACTGTAGAA AAAGGAATCTATCAA ACTTCTAACTTTAGAGTCCAACCAACAGAATCTATTGT	S (spike glycoprotein)	218/218 (100%)
CS_13	22,874–23,144 (271)	TCTAACAAtCTTGATTCTAAGGTTGGTGGTAATTATAATTACCtGTA TAGATTGTTTAGG AAGTCTAATCTCAAACCTTTTGAGAGAGATATTTCAACTGAAATC TATCAGGCCGGTAgC AcACCTTGTAATGGTGTTgAAGGTTTTAATTGTTACTTTCCTT TACAAtCATATGGTTTC CAACCCACTaATGGTGTTGGTTACCAACCATACAGAGTAGTAGTAC TTTCTTTTGAACTT CTACATGCACCAGCAACTGTTTGTGGACCTA	S (spike glycoprotein)	270/271 (99%)

Table 1 (Continued)

ID	Position (length in bp)	Sequence	Gene (product)	Identity with consensus sequence
CS_14	25,630–25,797 (168)	GTTTGCAACTTGCTGTTGTTGTTTGTAACAGTTTACTCACACCTT TTGCTCGTTGCTGCT GGCCTTGAAGCCCcTTTTCTCTATCTTTATGCTTTAGTCTACTTCTT GCAGAGTATAAAC TTTGTAAGAATAATAATGAGGCTTTGGCTTTGCTGgAAATGCCGT TCC	ORF3a (ORF3a protein)	168/168 (100%)
CS_15	25,974–26,214 (241)	ATCTGGAGTAAAAGACTGTGTTGTATTACACAGTTACTTCACTTC AGACTATTACCAGCT GTACTCAACTCAATTGAGTACAGACACTGGTGTTGAACATGTTAC CTTCTTCATCTACAA TAAAATTGTTGATGAGcCTGAAGAACATGTCCAAATTCACACAATCG ACGGTTCAtCCGG AGTTGTTAATCCAGTAATGGAACCAATTTATGATGAACCGACGAC GACTACTAGCGTGCC T	ORF3a (ORF3a protein)	241/241 (100%)
CS_16	27,467–27,808 (342)	GAGGTACAACAGTACTTTTAAAAGAACCTTGCTCTTCTGGAACAT ACGAGGGCAATTCAC CATTTCATCCTCTAGCTGATAACAAATTTGCACTGACTTGCTTTA GCACTCAATTTGCTT TTGCTTGTCCTGACGGCGTAAAACACGTCTATCAGTTACGTGCCA GATCAGtTTCACCTA AACTGTTCATCAGACAAGAGGAAGTTCAAGAACTTTACTCTCCAA TTTTTCTTATTGTTG CGGCAATAGTGTTTATAACACTTTGCTTCACAcTCAAAAGAAAGA cAGAATGATTGAACT TTCATTAATTGACTTCTATTTGTGCTTTTTAGCCTTTCTGCT	ORF7a (ORF7a protein)	342/342 (100%)
CS17	29,357–29,511 (155)	ACATTcCCACCAACAGAGCCTAAAAAGGACAAAAAGAAGAAG GCTgATGAAACTCAAGCC TTACCGCAGAGACAGAAGAAACAGCAAACTGTGACTCTTCTTCCT GCTGCAGATTTGGAT GATTTCTCCAAACAATTGCAACAATCCATGAGCAG	N (nucleocapsid phosphoprotein)	155/155 (100%)

All information was derived from RefSeq NC_045512.2. Less conserved positions (conservation < 99%) are shown in lowercase

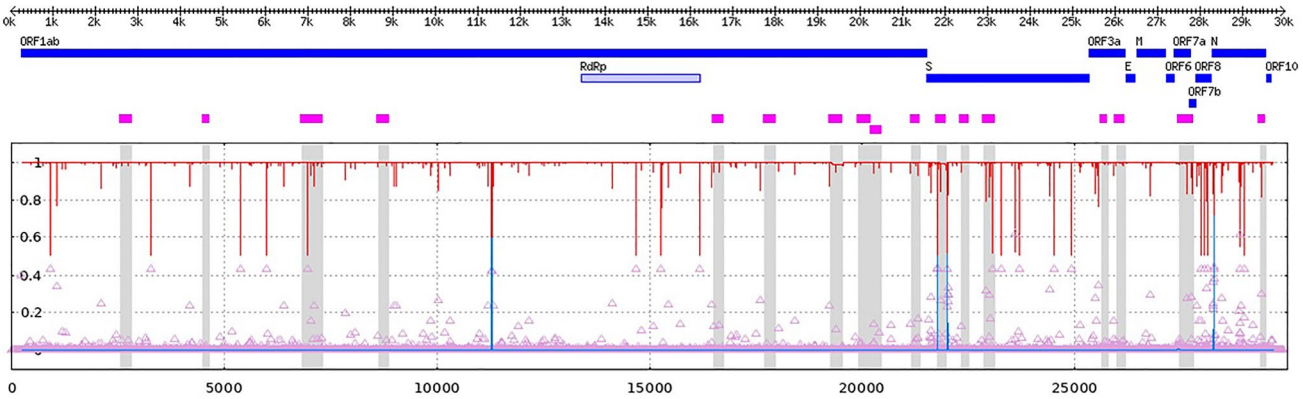


Fig. 2 The locations of the 17 conserved sequences (magenta blocks) on the reference SARS-CoV-2 genome map. The coding region of RdRp (RNA-dependent RNA polymerase) is shown beneath the ORF1ab coding region. The lower plot shows the conservation (red line), gappyness (blue line), and normalized Shannon entropy (plum triangles) of each nucleotide position. The conserved sequences are shown as gray shaded areas in the lower plot

(data not shown). Figure 2 shows the locations of the 17 CSs (CS_1 through CS_17) on the reference genome map; overall, 96.35% of sites displayed > 99% conservation.

Primer design and in silico/experimental validation

Eleven primer sets were designed initially and were tested against SARS-CoV-2 RNA (data not shown). Two primer sets, designed from CS_1 (NH-CS_1 in the nsp2 gene; 2576–2691) and CS_14 (NH-CS_14 in the ORF3a gene; 25,634–25,657), showed good amplification efficiency and were selected for in silico analysis. The sequence specificities of the NH-CS_1 and NH-CS_14 primer sets, as well as those of 16 other primer sets used in previous SARS-CoV-2 analyses, were checked against the genome datasets used in this study ('all' dataset, dereplicated dataset, and Delta dataset). Specifically, we examined the percentages of genomes in each dataset that would be identified by each set of primers, allowing a 0%, 5%, or 10% mismatch between the primer and template sequences (Table 2). Even at a 0% mismatch threshold, the performances of the NH-CS_1 and NH-CS_14 primer sets were at least on par with those of the 16 known primer sets. The NH-CS_1 and NH-CS_14 primer sets performed especially well for the Delta dataset. Because the MSA was derived from the dereplicated dataset, from which singletons were removed (although not for the Korean isolates or Delta variants), the percentage amplification coverages of the NH-CS_1 and NH-CS_14 primer sets against genomes in the 'all' dataset (230,163 genomes) were slightly lower (96.77% for NH-CS_1 and 98.79% for NH-CS_14) than those obtained for the other primer sets, whereas better performance (99.96%) was achieved if a 5% mismatch was allowed. Overall, these results demonstrate that dereplicated sequences can be useful for primer design without losing coverage of the primer sets against the original genome dataset. The amplification coverage of all primer sets against 6803 complete Omicron sequences (B.1.1.529), downloaded from the NCBI SARS-CoV-2 Resources on March 2, 2022, were also checked (Table 2).

Finally, we examined the abilities of the NH-CS_1 and NH-CS_14 primers (and corresponding probes) to detect all currently known nine SARS-CoV-2 variants, including Alpha, Beta, Gamma, Delta, Epsilon, Zeta, Eta, Iota, and Kappa variants (Fig. 3). No detectable signals were observed from NTC or human RNA samples. Based on the similarity of the Ct values (Table 3), the primer–probe sets were seemingly capable of detecting all variants with similar amplification efficiencies.

Discussion

When the number of available SARS-CoV-2 genomes was limited, choosing conserved target genes was a crucial factor for the design of optimal PCR primers (Li et al. 2020), and optimization of laboratory test protocols was important for effective detection (Won et al. 2020). Checking for abundant splicing variants (Park et al. 2020) and designing primers that can anneal to non-degenerate codons (encoding tryptophan) at the 3'-ends of sequences (Dong et al. 2021) were also suggested to ensure highly sensitive viral nucleic acid detection. On the contrary, our current study does not rely on previous knowledge of the structural aspects of conserved genes, and our approach can be scaled easily as the number of SARS-CoV-2 genomes increases.

The primer–probe sets validated here targeted the nsp2 and ORF3a genes; to our knowledge, these genes are rarely used as targets for the molecular diagnosis of SARS-CoV-2.

Table 2 The in silico amplification coverage of various primer sets against SARS-CoV-2 genome datasets. The percentages in the column headers represent the percent mismatches allowed between the primer and template sequences ('primersearch –mismatchpercent')

Name	Sequences (5'→3')	Size (bp)	All dataset (N=230163)			Dereplicated dataset (N=32483)			Delta variant only (N=6621)			Omicron variant only (N=6803)			Ref
			0%	5%	10%	0%	5%	10%	0%	5%	10%	0%	5%	10%	
NH-CS_1[a]	F: AGTGAAGCTGTTGAAGCTCCAT R: GTTACCATCATATTAGGTGCAAG	116	96.77%	99.96%	99.96%	97.71%	99.90%	99.90%	99.79%	100.00%	100.00%	99.15%	99.56%	99.60%	This work
NH-CS_14[b]	F: GCAACTTGCTGTTGTTGTTTGTAA R: GTTTATACTCTGCAAGAAGTAGAC	116	98.79%	99.96%	99.96%	98.96%	99.99%	99.99%	99.47%	100.00%	100.00%	98.85%	99.96%	99.96%	
CDC_N1	F: GACCCCAAAATCAGCGAAAT R: TCTGGTTACTGCCAGTTGAATCTG	72	98.20%	100.00%	100.00%	98.44%	100.00%	100.00%	99.18%	100.00%	100.00%	98.99%	100.00%	100.00%	CDC, USA[f]
CDC_N2	F: TTACAAACATTGGCCGCAAA R: GCGCGACATTCCGAAGAA	67	97.51%	99.23%	99.99%	93.29%	99.17%	99.99%	97.84%	98.55%	99.97%	89.28%	90.09%	99.99%	
CDC_N3	F: GGGAGCCTTGAATACACCAAAA R: TGTAGCACGATTGCAGCATTG	72	98.39%	99.99%	100.00%	98.51%	99.99%	100.00%	98.81%	100.00%	100.00%	99.12%	100.00%	100.00%	
NIID_2019-nCoV_N[c]	F: AAATTTTGGGGACCAGGAAC R: TGGCAGCTGTGTAGGTCAAC	158	0.00%	97.87%	99.98%	0.00%	98.35%	99.99%	0.00%	99.89%	100.00%	0.00%	99.32%	99.99%	NIID, JAPAN[g]
SARS-CoV-2_IBS_RdRP1	F: CATGTGTGGCGGTTCACTAT R: TGCATTAACATTGGCCGTGA	118	97.22%	99.99%	100.00%	86.28%	99.99%	100.00%	25.93%	99.94%	100.00%	99.29%	99.99%	99.99%	(Won et al. 2020)
SARS-CoV-2_IBS-S1	F: CTACATGCACCAGCAACTGT R: CACCTGTGCCTGTTAAACCA	100	98.94%	99.80%	99.81%	98.73%	99.47%	99.50%	97.12%	97.39%	97.52%	99.12%	99.81%	99.81%	
SARS-CoV-2_IBS_E1[d]	F: TTCGGAAGAGACAGGTACGTT R: CACACAATCGATGCGCAGTA	107	0.00%	99.97%	100.00%	0.00%	99.97%	99.99%	0.00%	99.95%	99.97%	0.00%	99.99%	100.00%	
SARS-CoV-2_IBS_N1	F: CAATGCTGCAATCGTGCTAC R: GTTGCGACTACGTGATGAGG	118	98.53%	99.99%	99.99%	98.86%	99.99%	100.00%	99.58%	100.00%	100.00%	98.44%	100.00%	100.00%	
SARS-CoV-2_IBS_RdRP2	F: AGAATAGAGCTCGCACCGTA R: CTCCTCTAGTGGCGGCTATT	102	93.90%	99.99%	99.99%	94.52%	100.00%	100.00%	99.79%	100.00%	100.00%	99.94%	100.00%	100.00%	
SARS-CoV-2_IBS_RdRP3	F: TCTGTGATGCCATGCGAAAT R: ACTACCTGGCGTGGTTTGTA	113	82.48%	99.94%	100.00%	85.64%	99.96%	100.00%	99.68%	100.00%	100.00%	99.50%	100.00%	100.00%	
SARS-CoV-2_IBS-S2	F: GCTGGTGCTGCAGCTTATTA R: AGGGTCAAGTGCACAGTCTA	108	97.01%	98.06%	98.14%	98.89%	99.75%	99.78%	99.34%	99.53%	99.59%	96.38%	97.18%	97.28%	
SARS-CoV-2_IBS_E2	F: TTCGGAAGAGACAGGTACGTTA R: AGCAGTACGCACACAATCG	116	99.62%	99.93%	100.00%	99.71%	99.94%	99.99%	99.82%	99.91%	99.98%	99.79%	99.97%	100.00%	
SARS-CoV-2_IBS-N2	F: GCTGCAATCGTGCTACAACT R: TGAACTGTTGCGACTACGTG	120	96.83%	99.96%	99.99%	97.73%	99.98%	100.00%	99.62%	100.00%	100.00%	98.40%	100.00%	100.00%	
RdRP_SARSr[e]	F: GTGARATGGTCATGTGTGGCGG R: CARATGTTAAASACACTATTAGCATA	100	0.00%	99.95%	100.00%	0.00%	99.93%	100.00%	0.00%	99.64%	100.00%	0.00%	99.99%	99.99%	(Corman et al. 2020)
E_Sarbeco	F: ACAGGTACGTTAATAGTTAATAGCGT R: ATATTGCAGCAGTACGCACACA	113	99.70%	99.99%	99.99%	99.78%	99.99%	99.99%	99.85%	99.98%	99.98%	99.75%	100.00%	100.00%	
N_Sarbeco	F: CACATTGGCACCCGCAATC R: GAGGAACGAGAAGAGGCTTG	128	98.56%	99.27%	99.99%	98.76%	99.37%	100.00%	99.44%	99.56%	100.00%	99.53%	99.66%	100.00%	

[a]Detection probe sequence: 5′-JOE-CGGGCTTATGTTGCTCGAAATCAA-BHQ1-3′

[b]Detection probe sequence: 5′-Texas Red-TGCTCGTTGCTGCTGGCCTTGAAG-BHQ2-3′

[c,d,e]Primers with a 1 bp mismatch to the reference sequence (NC_045512.2)

[f]Available from https://www.who.int/docs/default-source/coronaviruse/uscdcrt-pcr-panel-primer-probes.pdf?sfvrsn=fa29cb4b_2

[g]Available from https://www.niid.go.jp/niid/en/2013-03-15-04-39455-59/2483-disease-based/ka/corona-virus/2019-ncov/9334-ncov-vir3-2.html

However, Yip et al. (2020) developed a probe-free COVID-19-nsp2 assay system based on four specific regions > 50 bp in length (nsp2, two regions in S, and ORF8) that were identified by k-mer based comparison of 96 SARS-CoV-2 and 104 non-SARS-CoV-2 genomes. There are slight overlaps between the CS regions identified here and the targets identified by Yip et al. (data not shown). Although this approach might be able to maximize the specificities of the primers, they used only 200 genomes to identify four regions totaling 1188 bp; this region might be shortened if more genomes were used as inputs.

One shortcoming of the tool used here for MSA (ViralMSA) is its inability to handle insertions with respect to the reference genome sequence. The author that developed ViralMSA claimed that insertions usually lack phylogenetic or transmission clustering information among closely related viral isolates, and demonstrated that removal of insertions results in little impact on downstream analyses compared to other MSA software. Mercatelli and Giorgi (Mercatelli and Giorgi 2020) analyzed 48,635 complete SARS-CoV-2 genomes from across the world and found that single nucleotide transitions were the major mutational type. This finding suggests that ViralMSA can be used with minimum loss of accuracy for rapid MSA of SARS-CoV-2 genomes.

We did not incorporate Omicron in RT-qPCR experimental validation because it was only one of the variants of interest when experimental conditions were being optimized. Furthermore, Omicron RNA samples became very recently available domestically from the NCCP since late-December 2021. As soon as World Health Organization (WHO) designated Omicron variant (Pango lineage B.1.1.529) a variant of concern on November 26 2021,[2] it rapidly displaced Delta within several months, becoming the predominating one over major countries. Thus, the performance of primer sets designed from this study could be estimated from in silico PCR (Table 2) only. Although some other primer sets appear to outperform ours, we believe that > 99.5% coverage with 5% mismatches is sufficient to detect Omicron variants.

Considering the increasing availability of SARS-CoV-2 sequences in public databases, it will likely become more difficult to identify 'true' CSs using conventional sequence alignment methods. Regular sequence updates will find new polymorphic sites in pre-defined CS regions, resulting in fewer candidate regions. Our current study used a contemporary sequence alignment-based approach; however, dereplication was used to facilitate the analysis process, which would not otherwise have been feasible with the original dataset. Removal of singleton sequences that are identifiable

[2] World Health Organization. Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern. https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern.

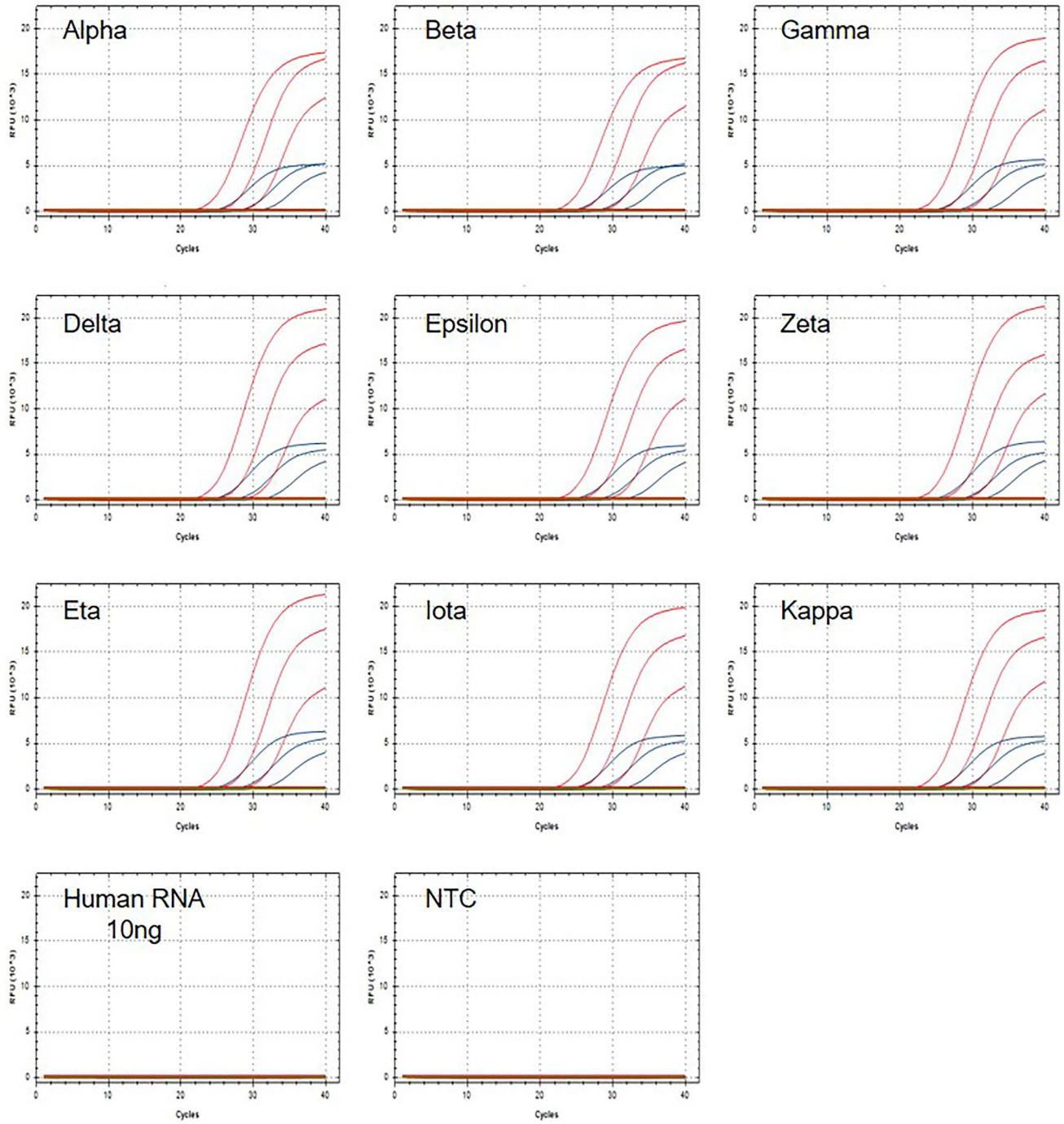


Fig. 3 Specific detection of nine SARS-CoV-2 variants (Alpha, Beta, Gamma, Delta, Epsilon, Zeta, Eta, Iota, and Kappa) using primer sets NH-CS_1 (blue) and NH-CS_14 (red). Multiplex RT-PCR was performed by serially diluted SARS-CoV-2 RNA templates (10^2 to 10^4 copies) using primers targeting ORF3a (red) and nsp2 (blue) genes

only after dereplication also contributed to the reduction of dataset size. The candidate CSs identified here, though not entirely composed of conserved sites, were still sufficient to design PCR primers covering most known SARS-CoV-2 sequences.

We expect that comprehensive primer databases that take target viruses, hosts, and non-target organisms into account

Table 3 Comparison of Ct values for RNA extracted from nine SARS-CoV-2 variants

Target	Copy number	Alpha	Beta	Gamma	Delta	Epsilon	Zeta	Eta	Iota	Kappa
nsp2 (JOE)	10^4	23.35	23.16	23.30	23.28	23.46	23.26	23.38	23.29	23.33
	10^3	26.75	26.74	26.34	26.42	26.83	26.87	26.42	26.39	26.31
	10^2	29.51	29.46	30.09	29.89	30.23	30.10	30.05	30.04	30.10
ORF3a (Texas Red)	10^4	21.71	21.67	22.01	21.77	22.16	22.03	21.90	21.79	21.94
	10^3	25.26	25.14	25.14	25.03	25.46	25.50	25.28	25.02	25.07
	10^2	28.30	28.47	28.58	28.57	29.18	28.96	28.95	28.39	28.54

will be available soon; these databases will require regular updates with large-scale collaborative efforts. Our current work demonstrates a fast and flexible approach to the development and evaluation of PCR primers for the detection of SARS-CoV-2. When new pandemics occur, machine learning-based methods (Lopez-Rincon et al. 2021; Randhawa et al. 2020) requiring neither biological knowledge nor massive amounts of sequence data can be the preferred choice.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13258-022-01264-7.

Acknowledgements We are grateful to the groups and laboratories that contributed SARS-CoV-2 genome sequences to GISAID (see Supplementary File for all contributors). This work was supported by the National Research Foundation of Korea (Grant Number 2021M3H4A4079381), funded by the Ministry of Science and ICT, and the KRIBB Research Initiative Program Republic of Korea.

Declarations

Conflict of interest The authors have no financial conflicts of interest to declare.

References

Bae J, Jeon H, Kim MS (2021) GPrimer: a fast GPU-based pipeline for primer design for qPCR experiments. BMC Bioinform 22:220

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973

Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, Bleicker T, Brunink S, Schneider J, Schmidt ML et al (2020) Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill. https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045

Davi MJP, Jeronimo SMB, Lima J, Lanza DCF (2021) Design and in silico validation of polymerase chain reaction primers to detect severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Sci Rep 11:12565

Dong H, Wang S, Zhang J, Zhang K, Zhang F, Wang H, Xie S, Hu W, Gu L (2021) Structure-based primer design minimizes the risk of PCR failure caused by SARS-CoV-2 mutations. Front Cell Infect Microbiol 11:741147

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinform 5:113

Jung Y, Park GS, Moon JH, Ku K, Beak SH, Lee CS, Kim S, Park EC, Park D, Lee JH et al (2020) Comparative analysis of primer-probe sets for RT-qPCR of COVID-19 causative virus (SARS-CoV-2). ACS Infect Dis 6:2513–2523

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780

Kim H, Kang N, An K, Kim D, Koo J, Kim MS (2017) MRPrimerV: a database of PCR primers for RNA virus detection. Nucleic Acids Res 45:D475–D481

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12

Li K, Venter E, Yooseph S, Stockwell TB, Eckerle LD, Denison MR, Spiro DJ, Methe BA (2010) ANDES: statistical tools for the ANalyses of DEep sequencing. BMC Res Notes 3:199

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

Libin PJK, Deforche K, Abecasis AB, Theys K (2019) VIRULIGN: fast codon-correct alignment and annotation of viral genomes. Bioinformatics 35:1763–1765

Lisboa Bastos M, Tavaziva G, Abidi SK, Campbell JR, Haraoui LP, Johnston JC, Lan Z, Law S, MacLean E, Trajman A et al (2020) Diagnostic accuracy of serological tests for covid-19: systematic review and meta-analysis. BMJ 370:m2516

Lopez-Rincon A, Tonda A, Mendoza-Maldonado L, Mulders D, Molenkamp R, Perez-Romero CA, Claassen E, Garssen J, Kraneveld AD (2021) Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning. Sci Rep 11:947

Mercatelli D, Giorgi FM (2020) Geographic and genomic distribution of SARS-CoV-2 mutations. Front Microbiol 11:1800

Moshiri N (2021) ViralMSA: massively scalable reference-guided multiple sequence alignment of viral genomes. Bioinformatics 37:714–716

Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a web browser. BMC Bioinform 12:385

Park M, Won J, Choi BY, Lee CJ (2020) Optimization of primer sets and detection protocols for SARS-CoV-2 of coronavirus disease 2019 (COVID-19) using PCR and real-time PCR. Exp Mol Med 52:963–977

Park C, Lee J, Hassan ZU, Ku KB, Kim SJ, Kim HG, Park EC, Park GS, Park D, Baek SH et al (2021) Comparison of digital PCR and quantitative PCR with various SARS-CoV-2 primer-probe sets. J Microbiol Biotechnol 31:358–367

Perini M, Piazza A, Panelli S, de Carlo D, Corbella M, Gona F, Vailati F, Marone P, Cirillo DM, Farina C et al. (2020) EasyPrimer: user-friendly tool for pan-PCR/HRM primers design Development of an HRM protocol on *wzi* gene for fast *Klebsiella pneumoniae* typing. Sci Rep 10:1307

Qu W, Li J, Cai H, Zhao D (2022) PCR primer design for the rapidly evolving SARS-CoV-2 genome. Methods Mol Biol 2392:185–197

Randhawa GS, Soltysiak MPM, El Roz H, de Souza CPE, Hill KA, Kari L (2020) Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. PLoS One 15:e0232391

Rice P, Longden I, Bleasby A (2000) EMBOSS: the european molecular biology open software suite. Trends Genet 16:276–277

Rognes T, Flouri T, Nichols B, Quince C, Mahe F (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584

Shu Y, McCauley J (2017) GISAID: global initiative on sharing all influenza data—from vision to reality. Euro Surveill 22:30494

Sievers F, Higgins DG (2018) Clustal Omega for making accurate alignments of many protein sequences. Protein Sci 27:135–145

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H et al (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12:1611–1618

Steenwyk JL, Buida TJ 3rd, Li Y, Shen XX, Rokas A (2020) ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biol 18:e3001007

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3–new capabilities and interfaces. Nucleic Acids Res 40:e115

Wang K, Li H, Xu Y, Shao Q, Yi J, Wang R, Cai W, Hang X, Zhang C, Cai H et al (2019) MFEprimer-3.0: quality control for PCR primers. Nucleic Acids Res 47:W610–W613

Won J, Lee S, Park M, Kim TY, Park MG, Choi BY, Kim D, Chang H, Kim VN, Lee CJ (2020) Development of a laboratory-safe and low-cost detection protocol for SARS-CoV-2 of the Coronavirus Disease 2019 (COVID-19). Exp Neurobiol 29:107–119

Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinform 13:134

Yip CC, Ho CC, Chan JF, To KK, Chan HS, Wong SC, Leung KH, Fung AY, Ng AC, Zou Z et al (2020) Development of a novel, genome subtraction-derived, SARS-CoV-2-specific COVID-19-nsp2 real-time RT-PCR assay and its evaluation using clinical specimens. Int J Mol Sci 21:2574

Yoon H, Leitner T (2015) PrimerDesign-M: a multiple-alignment based multiple-primer design tool for walking across variable genomes. Bioinformatics 31:1472–1474