 doi:10.20944/preprints202003.0392.v1

In silico comparative genomics of the severe acute respiratory syndrome coronavirus 2 (COVID-19) and designing of a cost-effective RFLP-PCR based mass screening diagnostic protocol

Aditya Singh*[1], Prateek Bhatia[1]

[1]Postgraduate Institute of Medical Education and Research, Chandigarh, India

*Corresponding author

Lab no. 4102, 4th Floor, Block A

Advanced Paediatrics Centre,

Postgraduate Institute of Medical Education and Research

Chandigarh, India

160012

aditya.onco@gmail.com

+91-98761-41188

Keywords: COVID-19, diagnosis, comparative genomics

 doi:10.20944/preprints202003.0392.v1

1. Abstract

SARS-CoV 2 also known as COVID-19 is a fast-spreading coronavirus related disease that emerged from China in December 2019 and has currently attained the status of a pandemic. There are currently no drugs/ vaccines against the same and limited diagnostic tests to identify the infection. Additionally, these tests are expensive and hence are exclusive for very highly suspected cases of the disease especially in developing countries. This is causing an under-diagnosis which is an alarming state of affairs, as even a single missed SARS-CoV 2 case would spread the disease exponentially and keep it in the community. Through this entirely *in silico* study, we've developed a cheaper and faster diagnostic method based on simple PCR and restriction enzyme digestion, commonly used in restriction fragment length polymorphism (RFLP) tests. Through comparative genomics, we found the closest neighbours of SARS-CoV 2 then found the highly conserved regions of the genome which weren't present in SARS-CoV 1, it's the closest neighbour. Then we found restriction sites for various enzymes followed by designing of PCR primers flanking those sites. We have found the primer pair to produce a 401 bp amplicon and when digested by *SwaI* enzyme, it produces two fragments of lengths 216 bp and 185 bp. As an internal control, GAPDH primers are pooled with the SARS-CoV 2 primers as the patient sample will also include human RNA mixed with the viral RNA. This primer pair gives an amplicon of 131 bp and hence a negative sample should show a single band of 131 bp while a positive digested sample will give three bands of 401 bp, 216 bp and 131 bp. The primers are specific to SARS-CoV 2 only and can additionally be used for SYBR green-based real-time quantification of viral load. The developed tests haven't yet been tested *in vitro* due to stressed-out working hours in the only pathogenic virus handling laboratory in our institute. Nonetheless, this study works as a head start for other laboratories to rapidly test the suggested protocols in vitro and make available a cheaper alternative test for SARS-CoV 2 which would especially be beneficial for the lower to middle-income countries.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2020 doi:10.20944/preprints202003.0392.v1

2. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV 2) also popularly known as the coronavirus disease 2019 (COVID-19) is a, as the name suggests, a respiratory disease closely related to the SARS-CoV 1 that had its first human outbreak back in the year of 2002. The new SARS-CoV 2 virus shares close similarity with the SARS-CoV 1 but is still untreatable by its medications [1]. The first case of the disease emerged from China back in December 2019 and since then it has been spreading rampantly across the world. One of the reasons for this is that infected people are contagious even before they develop any symptoms. There are limited options for diagnosis of the infection, especially in a resource constraint setting. In this article, we've performed an *in-silico* study of the sixteen published genomes of COVID-19 forty-two other coronavirus genomes. Through this study, we've deduced various components needed for conducting restriction fragment length polymorphism (RFLP) based technique to essentially confirm the presence of SARS-CoV 2. The developed primers are specific to the SARS-CoV 2 but we've also included a restriction digestion step which will give two specific bands of predetermined sizes as a confirmatory step for the presence of SARS-CoV 2 RNA in the sample. This was done to have an additional layer of confirmation before giving out the results. Current RFLP-PCR protocol can be used as a mass screening test as it requires a total of around five and a half hours to complete starting from the raw patient sample. Additionally, a large batch of samples can be simultaneously run considering a 96/384 block of PCR and a large gel electrophoresis system, which is usually available for testing in most of the basic virology laboratories. We couldn't validate the primers and enzyme pair *in vitro* because of current high load and stressed out the testing environment in the only certified pathogenic virus handling laboratory in our institution. We are giving out these *In-silico* prediction results so that any laboratory working on the virus or having such facilities could have a head start as the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2020 doi:10.20944/preprints202003.0392.v1

in-silico work is already completed. These developed primers can also be taken up for real-time PCR based tests.

3. Materials and Methods

3.1 Multiple genome alignment and phylogenetics

All the genomes used in this study were downloaded from the NCBI genome assembly database with reference IDs listed in the supplementary table S1. In addition to the FASTA files of all the genomes, GenBank files for the SARS-CoV 1 and SARS-CoV 2 were also downloaded. The sequences were aligned using MAFFT [2] with default parameters. The alignment file was used to calculate the phylogeny using the maximum-likelihood algorithm in MEGAX software [3]. According to the phylogeny, the closes neighbour of SARS-CoV 2 was SARS-CoV 1 and hence further studies were conducted with it as a reference sequence. A new alignment was created using the same protocol with MAFFT in which the sixteen genomes SARS-CoV 2 were aligned with the SARS-CoV 1 genome. Also, independently, all the sixteen SARS-CoV 2 genomes were aligned to find out a concordance. The further studies were proceeded on with the only complete genome of SARS-CoV 2 with NCBI reference ID NC_045512, as there were no significant differences in the genome sequences.

3.2 Designing of components of the diagnostic test

From the alignment of SARS-CoV 2 and SARS-CoV 1, we deduced the sequence regions which were unique in SARS-CoV 2 from SARS-CoV 1 and were conserved across all the sixteen genomes of SARS-CoV 2 using UGENE alignment viewer [4]. The region was then explored for the presence of restriction sites using SnapGene software (GSL Biotech LLC, IL, USA). Primers were designed flanking that restriction site using Primer3 [5] and verified across all sixteen SARS-CoV 2 and one SARS-CoV 1 genome using pcr.seqs package from MOTHUR [6]. The primer pairs were also verified to not have any targets on the human genome using UCSC *In-silico* PCR (https://genome.ucsc.edu/cgi-bin/hgPcr). As the extracted RNA

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2020 doi:10.20944/preprints202003.0392.v1

from human samples would also carry human RNA, a multiplex PCR protocol is designed with control GAPDH primers, forward 5'- GTCTCCTCTGACTTCAACAGCG-3', reverse 5'-ACCACCCTGTTGCTGTAGCCAA-3' with the SARS-CoV 2 primers after considering any possible dimerization using PrimerROC [7]. A restriction fragment length polymorphism (RFLP) based protocol was written based on the results of these tests. An agarose gel electrophoresis was simulated *in silico* using SnapGene software.

4. Results and Discussion

The phylogenetics of the 58 genomes resulted in the clustering of all sixteen SARS-CoV 2 genomes closely followed by SARS-CoV 1 genome. The resulting phylogenetic tree is illustrated in figure 1. The alignment of all the SARS-CoV 2 genomes displayed very minor differences and hence the only complete genome with NCBI accession number NC_045512 was chosen for further studies. The SARS-CoV 2 spike protein gene was found to be significantly different from the spike protein gene for SARS-CoV 1 with the former being 3,822 bp and the latter 3,768 bp in length. This region was additionally conserved across all the sixteen whole-genome sequences for SARS-CoV 2 and hence we selected this region to develop further tests. In this region, we found various restriction sites among which the ones giving a single result were further selected. We ended up with two choices for the restriction enzyme, *SwaI* and *BamHI* and selected *SwaI* site as the *BamHI* site was towards the end of the sequence and hence the reverse primer would have fallen in the next gene region, ORF3a, which shared similarity with the SARS-CoV 1. We selected around 500 bp flanking region around the restriction site of *SwaI* for finding primers using Primer3. The designed primers, forward 5'-TGTGTCTGGTAACTGTGATGTTG-3', reverse 5'-TGCAGCAGGATCCACAAGA-3' have a T_m of 58.87℃ and 58.93℃ respectively with the amplicon length of 401 bp. A detailed description of the primers and alignment as outputted by the Primer3 program can be referenced in supplementary file S2. Once the product is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2020 doi:10.20944/preprints202003.0392.v1

digested by *SwaI* and it is from SARS-COV 2, there would be two products of lengths 216 bp and 185 bp.

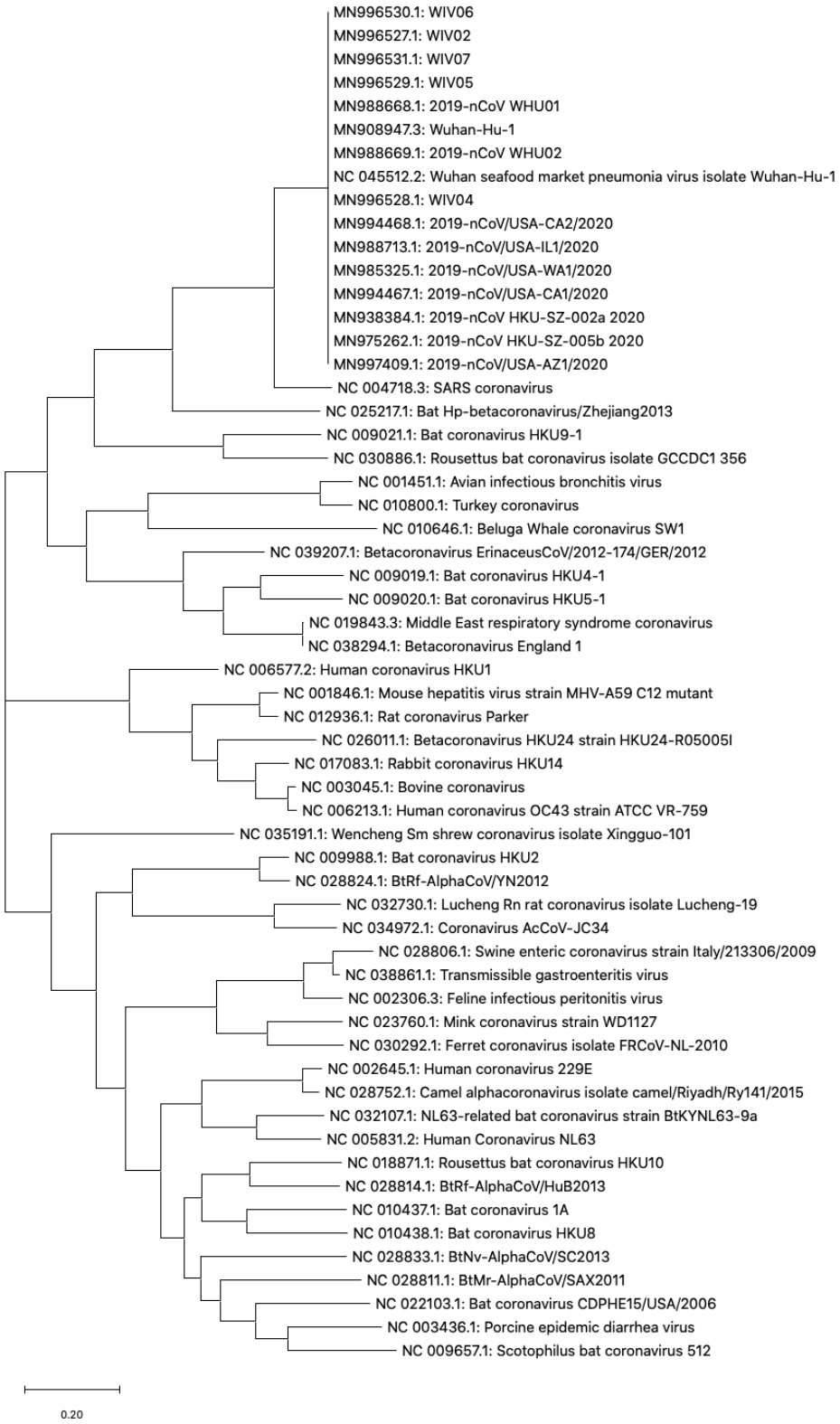

Figure 1: Phylogenetic tree for 58 genomes of various coronavirus

 doi:10.20944/preprints202003.0392.v1

The primers were found to not have any significant dimerization with the GAPDH primers given above. These primers were used to perform *in silico* PCR using MOTHUR for all the sixteen genomes of SARS-CoV 2 and the one genome of SARS-CoV 1 and the results showed that the primers were able to amplify the same sequence length of 401 bp across all SARS-CoV 2 genomes and they were specific as there were no products with SARS-CoV 1 genome (supplementary file S3). To perform this test, the first step would be to extract the total RNA from the patient sample which will include both viral as well as patient's RNA followed by generation of cDNA, then multiplex PCR using the two sets of primers, one for the internal control GAPDH and the other for SARS-CoV 2 spike gene region. Once we have the amplicons, we can simply run a 2.0% agarose gel electrophoresis and visualize the bands to come around 131 bp for GAPDH and 410 bp for viral spike gene, but to have a more confident result, we recommend digestion of the amplicons with *SwaI* enzyme for 15 minutes at 25°C and inactivation at 65°C for 20 minutes if using the enzyme form New England Biolabs cat#R0604 which comes in a fast-digesting format. The simulated agarose gel with and without digestion is illustrated in figure 2 wherein, a 100 bp ladder is loaded in lane MW, undigested/ digested PCR amplicon from a negative patient sample in lane 1, undigested PCR amplicon from a positive patient sample in lane 1 and *SwaI* digested amplicon from a positive patient sample in lane 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2020 doi:10.20944/preprints202003.0392.v1

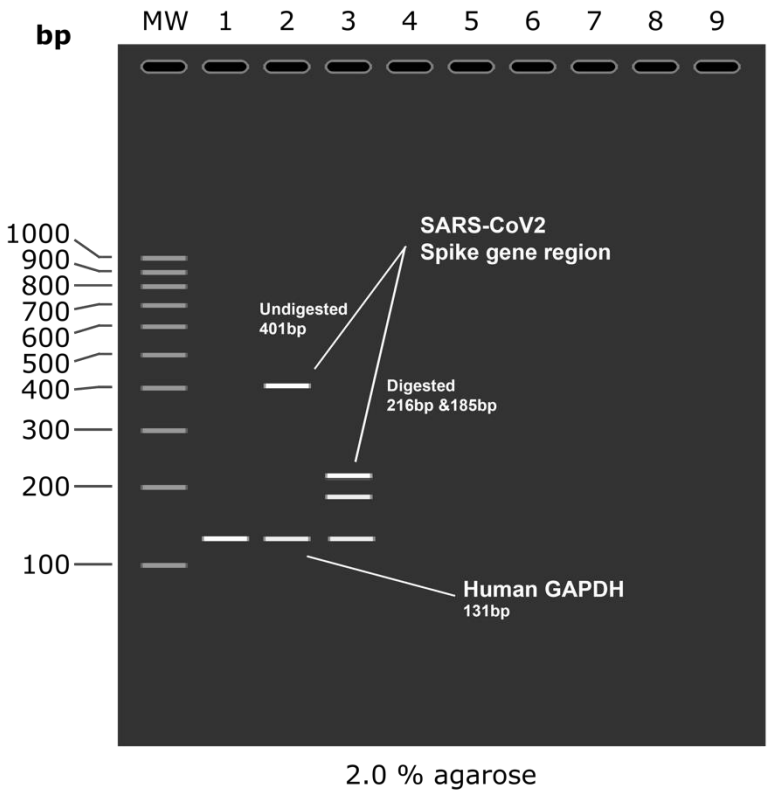

Figure 2: Simulated 2.0% agarose gel electrophoresis. MW: 100bp ladder; Lane 1: Digested PCR amplicon from virus negative sample; Lane 2: Undigested PCR amplicon; Lane 3: SwaI digested PCR amplicon

The laboratories are free to use their cDNA synthesis protocol hence we've not listed any specific one. Similarly, since we couldn't run the PCR in vitro, we haven't provided with exact annealing temperatures, still, we've provided with a protocol in table 1 as a probable PCR program. The primers can also be used to perform SYBR green real time PCR to quantify the viral load. Again, the laboratories are free to utilize their standardized protocol for the same and since we couldn't perform these tests *in vitro*, we haven't claimed any protocol. The whole workflow is illustrated in figure 3 for a better understanding.

Table 1: Multiplex PCR program for the given primers

S.No.	Temperature	Duration
1.	94°C	3 mins

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2020 doi:10.20944/preprints202003.0392.v1

2.	94°C	45 secs
3.	60°C*	60 secs
4.	72°C	90 secs
5.	Go to step 2	30x
6.	72°C	10 mins
7.	10°C	Infinite hold

*This temperature may vary and hence an *in vitro* standardization is required.

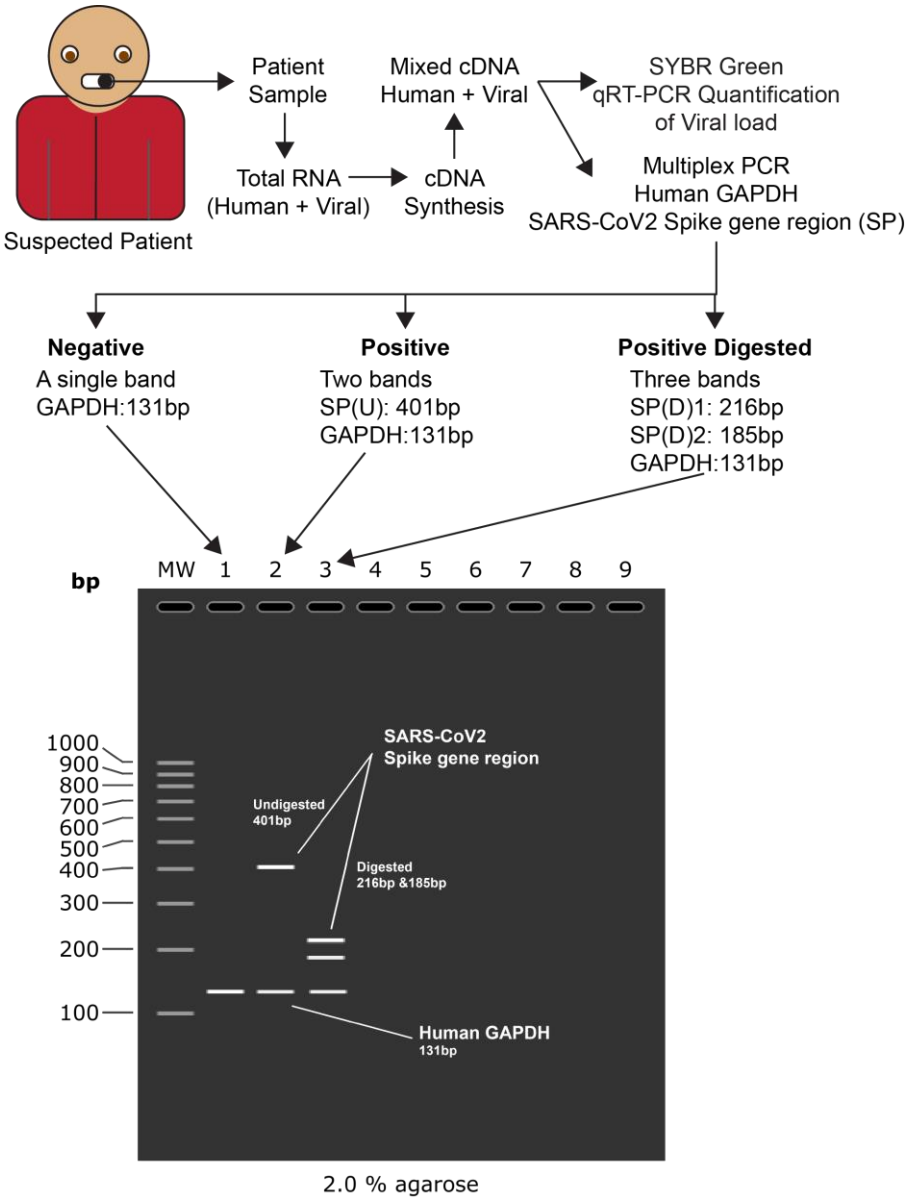

Figure 3: Workflow of the proposed diagnostic screening protocol

 doi:10.20944/preprints202003.0392.v1

5. Conclusion

The developed primers and enzyme pair will be vital for the diagnosis of SARS-CoV 2 with higher specificity as the test is a two-tier confirmatory one. This test is cheaper than performing TaqMan based real-time PCR, as the probes and mastermix for the same are very costly. Also, if required, an SYBR green-based real-time PCR can be conducted for quantifying the viral load using the same primer pair described in this manuscript. This study was conducted entirely *in silico* due to the overstrained testing facility in our institute as it caters to testing of samples from a large geographical region of Northern India. This study aimed to provide a head start to the laboratories with such facilities by performing all the *in silico* studies. The developed test is specifically aimed at the lower to middle-income countries wherein the cost of the test is a crucial deciding factor and where we believe "Mass Screening" is must to identify cases at earliest and prevent community spread. Through this study, we invite other laboratories with facilities to handle the virus to test the protocol *in vitro* and make it available for the community to benefit as soon as possible.

6. Acknowledgements

AS would like to thank the scientists who have made available the viral genomes at such fast pace.

7. Conflict of interest

The authors disclose no conflict of interests.

8. References

1 Cao, B. *et al.* A Trial of Lopinavir-Ritonavir in Adults Hospitalized with Severe Covid-19. *N Engl J Med*, doi:10.1056/NEJMoa2001282 (2020).

2 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

3 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547-1549, doi:10.1093/molbev/msy096 (2018).

4 Okonechnikov, K., Golosova, O., Fursov, M. & team, U. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**, 1166-1167, doi:10.1093/bioinformatics/bts091 (2012).

5 Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res* **40**, e115, doi:10.1093/nar/gks596 (2012).

6 Schloss, P. D. Reintroducing mothur: 10 Years Later. *Appl Environ Microbiol* **86**, doi:10.1128/AEM.02343-19 (2020).

7 Johnston, A. D., Lu, J., Ru, K. L., Korbie, D. & Trau, M. PrimerROC: accurate condition-independent dimer prediction using ROC analysis. *Sci Rep* **9**, 209, doi:10.1038/s41598-018-36612-9 (2019).

 doi:10.20944/preprints202003.0392.v1

Figure and Table legends:

Table 1: Multiplex PCR program for the given primers

Figure 1: Phylogenetic tree for 58 genomes of various coronavirus

Figure 2: Simulated 2.0% agarose gel electrophoresis. MW: 100bp ladder; Lane 1: Digested PCR amplicon from virus negative sample; Lane 2: Undigested PCR amplicon; Lane 3: SwaI digested PCR amplicon

Figure 3: Workflow of the proposed diagnostic screening protocol

Supplementary files legend:

Supplementary table S1: NCBI accession numbers and names of genomes in the study

Supplementary file S2: Primer3 result file showing alignment and properties of the primers

Supplementary file S3: FASTA file of all predicted PCR products across all genomes of SARS-CoV 2 and SARS-CoV 1.